CCL 2023

The 22nd Chinese National Conference on Computational Linguistics

August 3 - August 5, 2023

Harbin, China

Order copies of this and other CCL proceedings from:

Chinese National Conference on Computational Linguistics (CCL)

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

Tel: + 010-62562916

Fax: + 010-62661046

cips@iscas.ac.cn

Introduction

Welcome to the proceedings of the twenty second China National Conference on Computational Linguistics (22nd CCL). The conference were hosted and co-organized by Harbin Institute of Technology, China.

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

The Program Committee selected 78 papers (49 Chinese papers and 29 English papers) out of 278 submissions for publication. The acceptance rate is 28.06%. The 78 papers cover the following topics:

– Machine Translation and Multilingual Information Processing (5)

– Fundamental Theory and Methods of Computational Linguistics (7)

– Minority Language Information Processing (6)

– Social Computing and Sentiment Analysis (9)

– Text Generation and Summarization (4)

– Information Retrieval, Dialogue and Question Answering (8)

– Language Resource and Evaluation (8)

– Knowledge Graph and Information Extraction (15)

– Pre-trained Language Models (5)

– NLP Applications (11)

The final program for the 22nd CCL was the result of intense work by many dedicated colleagues. We want to thank, first of all, the authors who submitted their papers, contributing to the creation of the high-quality program. We are deeply indebted to all the Program Committee members for providing high-quality and insightful reviews under a tight schedule, and extremely grateful to the sponsors of the conference. Finally, we extend a special word of thanks to all the colleagues of the Organizing Committee and secretariat for their hard work in organizing the conference, and to the publish team, ACL Anthology and Springer for their assistance in publishing the proceedings in due time.

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

July 2023

Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, Xianpei Han

Organizers

Program Committee

22nd CCL Conference Chairs

Maosong Sun Tsinghua University, China
Bing Qin Harbin Institute of Technology, China

22nd CCL Program Chairs

Xipeng Qiu Fudan University, China
Jing Jiang Singapore Management University, Singapore
Xianpei Han Institute of Software , CAS, China

22nd CCL Area Co-Chairs

Information Retrieval, Text Classification and Question Answering

Xianling Mao Beijing Institute of Technology, China
Yunshan Ma National University of Singapore, Singapore

Text Generation, Dialogue and Summarization

Gao Yang Beijing Institute of Technology, China
Piji Li Nanjing University of Aeronautics and Astronautics, China
Lingpeng Kong The University of Hong Kong, China

Machine Translation and Multilingual Information Processing

Yun Chen Shanghai University of Finance and Economics, China
Jiatao Gu Meta, USA

Minority Language Information Processing

Xuebo Liu Harbin Institute of Technology (Shenzhen), China
Hui Huang University of Macau, China

Knowledge Graph and Information Extraction

Jintao Tang National University of Defense Technology, China
Ningyu Zhang Zhejiang University, China
Yixin Cao Singapore Management University, China

Social Computing and Sentiment Analysis

Ruifeng Xu Harbin Institute of Technology, China
Lin Gui Warwick University, UK

NLP Applications

Wei Wei Huazhong University Of Science And Technology, China
Jie Yang Zhejiang University, China
Wenpeng Yin Temple University, USA

Fundamental Theory and Method of Computational Linguistics

Kehai Chen Harbin Institute of Technology(Shenzhen), China
HongChen Wu Peking University, China

Language Resource and Evaluation

Yidong Chen Xiamen University, China
Chunyu Jie City University of Hong Kong, China

Pre-trained Language Models

Hao Zhou Tsinghua University, China
Yankai Lin Renmin University of China, China
Pengfei Liu Shanghai Jiao Tong University, China

22nd CCL Local Arrangement Chairs

Bing Qin Harbin Institute of Technology, China
Xiaocheng Feng Harbin Institute of Technology, China

22nd CCL Evaluation Chairs

Hongfei Lin Dalian University of Technology, China
Zhenghua Li Soochow University, China
Bin Li Nanjing Normal University

22nd CCL Publication Chairs

Gaoqi Rao Beijing Language and Culture University, China
Yubo Chen Institute of Automation, CAS, China

22nd CCL Chair of Frontier Forum

JiaJun Zhang Institute of Automation, CAS, China

22nd CCL Workshop Chairs

Yang Feng Institute of Computing Technology, CAS, China
Peng Li Tsinghua University, China

22nd CCL Sponsorship Chairs

Ruifeng Xu Harbin Institute of Technology, China
Kang Liu Institute of Automation, CAS, China

22nd CCL Publicity Chair

Zhongyu Wei Fudan University, China

22nd CCL Website Chair

Baotian Hu Harbin Institute of Technology, China

22nd CCL System Demonstration Chairs

Sendong Zhao Harbin Institute of Technology, China
Hao Zhou Tsinghua University, China

22nd CCL Student Seminar Chairs

Yankai Lin Renmin University of China, China
Tianxiang Sun Fudan University, China

22nd CCL Finance Chair

Yuxing Wang Tsinghua University, China

Table of Content

基于推理链的多跳问答对抗攻击和对抗增强训练方法

丁佳玓[1]，王思远[1]，魏忠钰[1,*]，陈琴[2]，黄萱菁[3]
[1] 复旦大学 大数据学院，上海市 200433
[2] 华东师范大学 计算机科学与技术学院，上海市 200241
[3] 复旦大学 计算机科学技术学院，上海市 200433
20210980123@fudan.edu.cn

摘要

本文提出了一种基于多跳推理链的对抗攻击方法，通过向输入文本中加入对抗性的攻击文本，并测试问答模型在干扰数据下生成答案的准确性，以检测问答模型真正执行多跳推理的能力和可解释性。该方法首先从输入文本中抽取从问题实体到答案实体的推理链，并基于推理链的特征把多跳问题分为了不同的推理类型，提出了一个模型来自动化实现问题拆解和推理类型预测，然后根据推理类型对原问题进行修改来构造攻击干扰句。实验对多个多跳问答模型进行了对抗攻击测试，所有模型的性能都显著下降，验证了该攻击方法的有效性以及目前问答模型存在的不足；向原训练集中加入对抗样本进行增强训练后，模型性能均有所回升，证明了本对抗增强训练方法可以提升模型的鲁棒性。

关键词： 对抗攻击 ； 多跳问答 ； 推理链

Reasoning Chain Based Adversarial Attack and Adversarial Augmentation Training for Multi-hop Question Answering

Jiayu Ding[1], Siyuan Wang[1], Zhongyu Wei[1,*], Qin Chen[2], and Xuanjing Huang[3]
[1] School of Data Science, Fudan University, Shanghai 200433
[2] School of Computer Science and Technology, East China Normal University, Shanghai 200241
[3] School of Computer Science, Fudan University, Shanghai 200433
20210980123@fudan.edu.cn

Abstract

This paper proposes a multi-hop reasoning chain based adversarial attack method in order to test the true ability and interpretability for conducting multi-hop reasoning of QA models. The main idea is to insert distracting sentences in the input context and then evaluate the answer accuracy of QA models. The method first formulates reasoning chains starting from query entities to answer entities, and categorizes questions into different reasoning types based on the characteristics of the reasoning chains. Then, a model is proposed to automatically decompose questions into multiple subquestions and predict their reasoning types. Lastly, distracting sentences are generated by adversarially modifying part of the questions according to their corresponding reasoning types. The results demonstrate significant performance reduction of multiple multi-hop QA models under adversarial data, verifying the effectiveness of our attack method and the vulnerability of QA models. After augmentation training with the adversarial samples, the models' performance all gets improved, which proves that this adversarial training method can enhance the robustness of QA models.

Keywords: adversarial attack , multi-hop question answering , reasoning chain

第二十二届中国计算语言学大会论文集，第1页-第16页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

1 引言

多跳问答（Multi-hop Question Answering）是自然语言处理领域一项被广泛研究的极具挑战性的任务。传统的单跳问答任务通过在单一文档内对所提出的问题进行简单的检索匹配即可找到答案(Rajpurkar et al., 2016; Rajpurkar et al., 2018)；与此相比，多跳问答更为复杂，需要结合多篇文本中的多个相关事实，根据它们进行多步骤推理才能得出答案(Welbl et al., 2018; Talmor and Berant, 2018; Yang et al., 2018; Khot et al., 2019)。目前已有许多研究工作尝试引入推理链的概念来解决多跳问答，并声称能够执行可解释的多步推理，模型的准确性也不断在提高(Qiu et al., 2019; De Cao et al., 2019; Fang et al., 2020)。

然而，通用的评价指标只是简单地衡量答案预测的准确性，并不能够检测模型是否真正进行了多跳的推理。实际上，模型可能通过直接定位与所提问题有较高字词重合度的句子，或者利用一些浅层的知识（例如已知答案所属的类型），就可以直接而简单地找到答案。这些方式跳过了任务所必要的所有推理步骤，违背多跳问答任务的初衷。图1展示了来自HotpotQA 英文数据集(Yang et al., 2018)的一个样例（已经过翻译），其中，粗体实下划线标记的两句句子是对于回答该问题所必不可少的推理依据（supporting facts）。但本例中有“捷径”的存在：可以根据问题知道答案类型是一个“岛”以此来缩小范围，同时可以仅关注包含了问题中关键词“西北808海里处”和“夏威夷”的句子。按照这样的方法，根据相关段落2的第一句话就可以直接定位到答案“莱塞岛”，而不需要使用到相关段落1的必需推理事实。此捷径假设可以通过图1所示的实践来验证。我们有意设计了一句与问题无关但极其相似的句子（以斜体虚下划线标记），并将其插入到输入文本段落中，结果问答模型预测出了错误的答案“广州岛”。这样的现象说明问答模型存在“过度稳定”的问题，即由于依赖固定的、表层的词汇句法模式而容易陷入文本陷阱，没有真正全面地理解文本并进行推理。

问题：哪个岛位于夏威夷首府西北 808 海里处？

相关段落1：《火奴鲁鲁》
火奴鲁鲁是美国夏威夷州的首府和最大城市。它是瓦胡岛火奴鲁鲁市和县的非建制部分和县城。*广州岛毗邻夏威夷首府西南 808 海里处。*该市是通往夏威夷的主要门户，也是进入美国的主要门户.....

相关段落2：《莱塞岛》
莱塞岛（夏威夷语：Kauō），位于火奴鲁鲁西北 808 海里处，北纬 25° 42‘ 14“西经171° 44’ 04”，是西北夏威夷群岛之一。它包括一块 1016 英亩的土地，面积约为 1 英亩。*广州岛毗邻夏威夷首府西南 808 海里处。*它是一个环礁，陆地完全围绕着一个高于海平面 8 英尺的浅中央湖， 其盐度大约是海洋的三倍.....

正确推理链： 夏威夷 —*首府*→ 火奴鲁鲁 —*西北808海里处*→ 莱塞岛

模型原始预测答案: 莱塞岛

加入对抗攻击句子后模型预测答案: 广州岛

Figure 1: 来自HotpotQA数据集的多跳问答样例。以粗体实下划线标记的句子是获得正确答案所必需的推理依据，斜体虚下划线标记的是原始文本段落不包含、人为添加的对抗性干扰句。将干扰句插入原段落后，问答模型的输出答案从原来正确的“莱塞岛”变成了错误的“广州岛”。

之前的研究曾尝试在输入问题的基础上修改实体，然后将修改后的句子添加到输入文本中作为对抗性干扰(Jia and Liang, 2017; Wang and Bansal, 2018)。但这样的干扰构造方式不适用于多跳问答，一方面，实体在连接不同文本段落及探究推理链方面起着至关重要的作用，将实体替换可能会使生成的干扰句与原问题完全无关，从而使得分散注意力的效果有限；另一方面，这种做法可能使得答案预测的过程不可追溯，无法识别出模型是在哪个环节出现问题。在本文中，我们提出了一种基于推理链的对抗攻击方法来检测问答模型的多跳推理能力。

具体来说，多跳问答从问题实体开始，不断从文本段落中查找与所问属性相关的语句和下一个相关实体，逐步向后推理。这样的推理过程可以建模为推理链，如图1中的“夏威夷 $\xrightarrow{\text{首府}}$ 火奴鲁鲁 $\xrightarrow{\text{西北808海里处}}$ 莱塞岛”。我们修改问题句中表示关系的词语而不修改实体，来确保添加的干扰不会与原始段落文本过于不相关，并且仅更改与某一跳（hop）对应的部分推理链，因为直观上与问题表述越相似的干扰句在混淆问答模型时越有效。本方法还支持根据需要对不同跳进行攻击，以深入了解是哪一跳环节更容易导致预测错误。在上面的例子中，我们通过添加

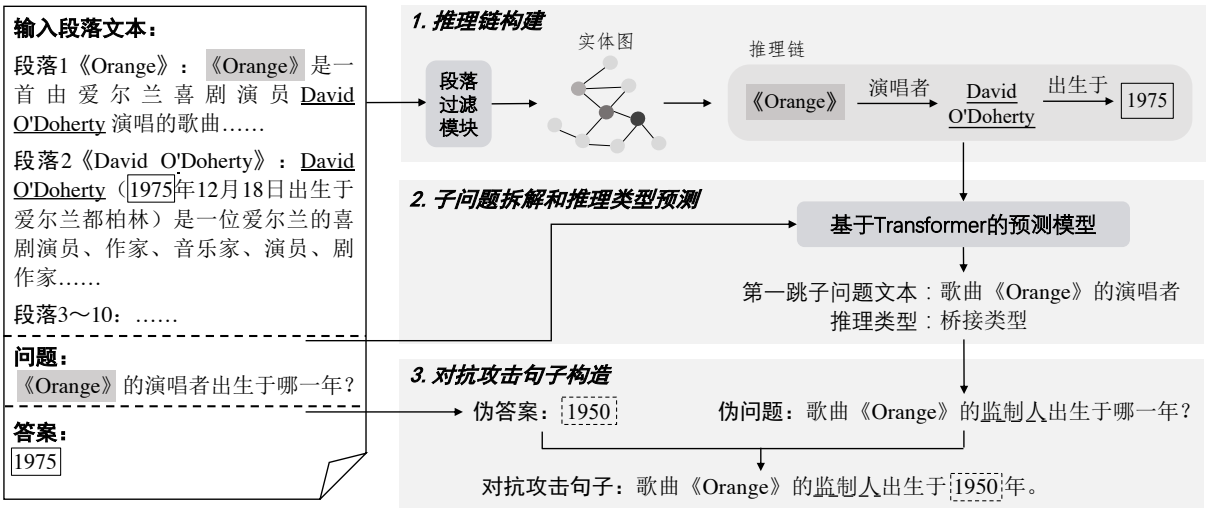


Figure 2: 本对抗攻击数据构造的总体框架。用阴影、实下划线和实线黑框标记的词语分别是问题实体、桥实体和答案实体。

了一条指向新节点“广州岛”（伪答案）的边“西南808海里处”（伪关系）来对第二跳进行攻击，生成了如图1所示的对抗攻击句（以斜体虚下划线标记）。

我们在HotpotQA 数据集上进行基于推理链的对抗攻击评估。实际上，HotpotQA中的问答样例所相应的推理链往往表现出不同的特点，因此我们将多跳问答归类为不同的推理类型，针对每一类型都设计了特定的攻击策略。总体上，本对抗攻击数据构造的框架设计为三个步骤：首先，根据相关段落文本构建实体关系图，并据此抽取推理链；接着，将原问题拆解为多个子问题，分别与推理链中的不同跳相对应；然后，根据问题的推理类型来选择要攻击的某一跳在原问题表述中所对应的文本，通过改变其中的关系短语来构建对抗攻击句，修改时保持句法和语义的合理性和自然性，并确保与原段落文本没有语义冲突。最后，将构造的对抗攻击句子插入到每一段段落文本中去，测试问答模型在噪声干扰下准确率是否下降。

我们对HotpotQA 基线模型(Yang et al., 2018)、DFGN(Qiu et al., 2019)和SAE(Tu et al., 2020)这三个多跳问答模型进行了对抗攻击评估。实验结果显示，对答案和推理依据的预测准确性都显著下降，表明了模型的脆弱性和潜在的浅层推理捷径。在添加了对抗数据进行重训练后，三个模型抵抗攻击的鲁棒性都有所增强，在原始数据集上的性能也略有提高或保持可比。我们希望本研究发现可以帮助推动设计出具有真正多跳推理能力的、更强大的问答模型。

2 方法

本节介绍所提出的基于推理链的对抗攻击句子生成方法，总体框架如图2所示。在2.1节中，我们筛选相关段落并构造实体关系图，并沿着图搜索从问题实体到答案实体的推理链，我们还根据推理链的特征定义了不同的推理类型。在2.2节中，为了实施更具针对性的攻击，我们设计了一个模型，能够结合2.1节获得的推理链知识来自动识别多跳问题的类型，并将问题拆解为与每一跳相对齐的子问题。根据2.2节获得的问题类型和子问题划分，可以选择想要攻击的某一跳，然后应用一套特殊设计的对抗攻击句子构造策略，策略的大致做法是改变待攻击子问题表述中表示关系的词语，并与一个伪答案相结合改写，转换成一句表述通顺的陈述句，这部分设计在2.3节中进行说明。

2.1 推理链的定义和构建

在跨文档问答任务中，输入上下文可能包含许多不相关的噪声段落，这会引入大量冗余的实体和三元组而导致实体图过大，无法进行高效的推理链搜索。因此，本文首先使用一个**段落过滤模块**来筛选相关段落。具体实现方式为：将问题和每个段落拼接起来、并在起始位置拼接一个表示全局信息的[CLS]标记，输入一个预训练好的BERT语言模型(Devlin et al., 2019)，然后将[CLS]的隐状态输入一个带有Sigmoid函数的分类层，模型将输出一个0到1之间的概率得分，表示该段落与所提问题的相关性。得分高的段落更有可能含有回答该问题所必需的推理依据，因此仅选择相关性分数高于特定阈值的段落进行后续的实体图构建。

对每个样例，基于上述选定的相关段落构建一个实体关系图。首先使用Stanford Corenlp工具包(Manning et al., 2014)从问题和所有相关段落中识别命名实体（分别获得问题实体集合$\mathcal{E}_q$和段落实体集合$\mathcal{E}_c$）。同时，使用OpenIE5工具包[0]从每句话中抽取论元-关系-论元（argument）三元组，其中，论元一般是命名实体或者名词短语；关系可能是以名词为中介的短语，也可能是动词短语，分别表示属性关系和动作关系。如果论元或关系表述过长，就对其进行二次抽取，以获得更多更基本的三元组和干净实体。对抽取的论元进行修正以与之前提取的实体集合（$\mathcal{E}_q$和$\mathcal{E}_c$）中的命名实体对齐，如果论元是代词，就将其替换为前一句的主语实体。

沿着上述构建的实体图，寻找从问题中每个提及的实体开始到答案实体的所有路径，选择其中最短的路径作为该问答样例的推理链。实际上，推理链隐含编码了推理的机制，可用于设计特定的攻击策略。因此，受到之前一些工作(Talmor and Berant, 2018; Perez et al., 2020)中分类思想的启发，我们根据推理链的特点将多跳问答分为了四种推理类型，该分类具有相当的普适性，可以涵盖到目前为止公开数据集中的几乎所有多跳问题。在图3中，我们为每种推理类型都给出了一个样例进行解释。

- **桥接类型（Bridging）**的问题需要顺序推理。首先要推理找到桥实体（bridge entity，即中间实体），然后利用它来执行第二跳以得到最终答案。映射到图上，推理链是单向的，有一个或多个桥实体节点连接着问题实体节点和答案节点，推理过程是通过逐级匹配关系边来进行的。
- **交集类型（Intersection）**的问题要求答案同时满足多个条件。推理链涉及至少两条独立的路径，其中，不同的问题实体节点独立地指向答案实体，任意一条路径不连通的节点都不能成为答案。
- **对比类型（Comparatives）**的问题要求比较两个实体的属性。这类问题通常不像前两类问题有一般意义上的连通路径，两个问题实体具有指向各自属性节点的独立并行的边，在两个属性节点上执行后续操作以获得最终答案。最终答案通常是两个问题实体之一，有时也有可能是它们的公共属性。
- **是否类型（Yes/No）**的问题询问两个实体是否具有相同的属性，答案只能是“是”或“否”。与对比类型类似，这里引入一个特殊的操作节点“是否相同?”来构建一条连通的推理链。

	示例	推理链
桥接类型	**问题**：歌曲《Orange》的演唱者是哪一年出生的？ **推理依据1**：《Orange》是一首由爱尔兰喜剧演员David O'Doherty演唱的歌。 **推理依据2**：David O'Doherty（1975年12月18日出生于爱尔兰都柏林）是一位爱尔兰的喜剧演员、作家、音乐家、演员、剧作家。 **答案**：1975	《Orange》 —演唱者→ David O'Doherty —出生于→ 1975
交集类型	**问题**：Rex Gene Foods 和 Foodtown 都位于哪个州？ **推理依据1**：Rex Gene Foods 公司是1957年到20世纪90年代末一家位于新泽西州的美国连锁超市。 **推理依据2**：Foodtown的公司办公室位于新泽西州的Iselin。 **答案**：新泽西州	Rex Gene Foods —位于→ 新泽西州 Foodtown —位于→ 新泽西州
对比类型	**问题**：谁出生得更早，Emma Bull 还是 Virginia Woolf？ **推理依据1**：Emma Bull（出生于1954年12月13日）是一名美国科幻小说作家。 **推理依据2**：Virginia Woolf（1882年1月25日–1941年3月28日）是一名英国作家，被认为是二十世纪最重要的现代主义者之一。 **答案**：Virginia Woolf	Emma Bull —出生于→ 1954年12月13日 Virginia Woolf —出生于→ 1882年1月25日 → 更早?
是否类型	**问题**：Thomas H. Ince 和 Joseph McGrath 是相同国籍的吗？ **推理依据1**：Thomas Harper Ince（1880年11月16日 – 1924年11月19日）是一名美国默片制片人、导演、编剧和演员。 **推理依据2**：Joseph McGrath（1930年生于格拉斯哥）是一名苏格兰电影电视导演、编剧。 **答案**：否	Thomas H. Ince —国籍→ 美国 Joseph McGrath —国籍→ 苏格兰 → 相同?

Figure 3: 来自HotpotQA 数据集的不同推理类型的示例及对应推理链。用阴影、实下划线和实线黑框标记的词语分别是问题实体、桥实体和答案实体。

[0]https://github.com/dair-iitd/OpenIE-standalone

2.2 子问题的拆解和推理类型的预测

直觉上来说，攻击干扰句与问题更相似会更有混淆性，因此，我们基于问题文本来设计对推理链的攻击。由于希望对不同的推理机制能有更具针对性、更有效的攻击，因此自动获得每个问题的推理类型并识别其各个子问题是必要的，以便进行后续的干扰句设计。

在HotpotQA数据集中，桥接类型和交集类型的样本已经统一都被打上了*Bridge* 标签，对比类型和是否类型都被归类为*Comparison*。因此，检查标记为*Comparison* 的样本的答案是否为“是/否”，就可以直接获得属于对比类型和是否类型的样本。而对于标记为*Bridge*的样本，我们设计了下述模型来融合推理链的信息识别其属于桥接类型还是交集类型，并将问题文本拆解为与不同推理跳（hop）相对应的子问题。模型的框架如附录B中图5所示。

考虑到问题表述中可能含有多个命名实体，我们选择到答案节点的最短路径长度最长的问题实体节点作为推理链的起点，这是因为该问题实体到答案的跳数最多，所以最有可能是推理的起始并且包含最多的信息。如果推理链包含两跳以上，则将第二跳子问题和之后的子问题合并在一起作为第二跳。形式上，我们将推理链（`CHAIN`）定义为`[HOP1]` $\texttt{ent}_q$ $\texttt{rel}_1$ `[HOP2]` $\texttt{ent}_{b1}$ $\texttt{rel}_2$ $\texttt{ent}_{b2}$ $\texttt{rel}_3$... $\texttt{ent}_a$，其中`[HOP1]`和`[HOP2]`是两个特殊的标记，$\texttt{ent}_q$、$\texttt{ent}_b$、$\texttt{ent}_a$、`rel`分别是2.1节抽取的问题实体、桥实体、答案实体和关系。模型将问题表述（`QUERY`）和推理链表述的拼接作为输入序列S = `[CLS] QUERY [SEP] CHAIN [SEP]`。使用含有多层网络结构的Transformer 模块对输入进行编码：

$$\mathbf{U} = \text{Transformer\_encoder}\,(S) \in \mathbb{R}^{N\times h} \tag{1}$$

其中$\mathbf{U}$为编码结果，N为最大文本输入长度，h为Transformer隐层维度大小。

接下来通过一个指针网络模块（Pointer Network），在输入问题的文本范围内预测其中的每个字符是第一跳子问题的起始位置和结束位置的概率（logits）。具体来说，对上述编码结果施加一个问题掩码$\mathbf{M} \in \mathbb{R}^{N\times h}$来限制起始和结束位置的预测范围，$\mathbf{M}$ 只有与问题表述相对应的位置上的元素为 1（即只有第 1 至 n 列的元素为 1，n 为输入问题表述 `QUERY` 的长度），其余为 0。然后将其输入一个参数可训练的矩阵$\mathbf{W}_1 \in \mathbb{R}^{N\times 2}$，并进行 Softmax 归一化以获得概率分布：

$$\mathbf{P} = [\mathbf{P}^{\text{start}}, \mathbf{P}^{\text{end}}] = \text{Softmax}\left((\mathbf{U} \otimes \mathbf{M})\,\mathbf{W}_1\right) \in \mathbb{R}^{N\times 2} \tag{2}$$

其中 $\otimes$ 表示元素积（Element-wise Product，矩阵每个位置对应元素元素相乘）。通过联合最大化这两个概率来确定第一跳子问题的起始位置 $\text{ind}_{\text{start}}$ 和结束位置 ind_{end}，：

$$\text{ind}_{\text{start}}, \text{ind}_{\text{end}} = \underset{1\le i\le j\le n}{\text{argmax}}\ \mathbf{P}_i^{\text{start}} \mathbf{P}_j^{\text{end}} \tag{3}$$

对于推理类型预测，将`[CLS]`标记位置处的隐状态 $\mathbf{U}_{[\text{CLS}]}$ 输入一个二分类（$c = 2$）的分类层，来预测推理链是桥接类型还是交集类型：

$$\mathbf{P}_{\text{type}} = \text{Softmax}\left(\mathbf{U}_{[\text{CLS}]}\mathbf{W}_2\right) \in \mathbb{R}^{1\times c} \tag{4}$$

上述两个预测任务是相互关联并且可以相互作为辅助指导的：通过子问题分解来对推理链进行更好的理解和显式建模，有助于判断推理类型；推理类型的潜在提问模式（underlying pattern）可以帮助定位不同子问题的文本范围。因此可以共同学习这两个子任务。使用联合的交叉熵损失来作为最终的损失函数进行优化（其中λ是可调的超参数）：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{start}} + \mathcal{L}_{\text{end}} + \lambda\mathcal{L}_{\text{type}} \tag{5}$$

2.3 对抗攻击句子的构造

我们针对不同的推理类型都设计了特定的攻击策略，总体上都遵循修改问题、设伪答案和合并改写三个步骤。图2中展示了一个样例，附录D的图6为每种推理类型都展示了一个构造样例并进行详细说明。生成的攻击句将插入到每一段输入段落文本中的随机位置作为干扰噪声。

2.3.1 桥接类型和交集类型

第一步　对原问题句内与所选定的目标攻击子问题相对应的文本，进行基于语义的修改，使其提问其他相似的内容。 目标攻击子问题的选择策略为：（1）对于桥接类型，选择第一跳子问题修改，因为考虑到在推理开始时就进行干扰更可能使攻击成功；（2）对于交集类型，随机选择任意一跳，因为两跳在推理链上是等价的。虽然许多之前的对抗攻击工作通过改变实体来施加干扰，但本文提出，对于多跳问答来说，对关系进行干扰可能更好。一般来说，关系可以是基于名词的属性关系（如“*the first son of* (第一个儿子)”）或基于动词

的行为关系（如“*was born in* (出生于)”），因此，本方法的修改目标是普通名词、形容词和动词（助动词和情态动词除外），而命名实体（包括人、组织、地点、数字）则保持不变。修改方式为：首先尝试用WordNet(Miller, 1998)中的反义词替换它们，如果没有反义词，则用Glove(Pennington et al., 2014)词向量空间中最相近的单词来进行替换；此外，还限制了替换词必须有相同的实体类型 (Named Entity Recognition，NER) 和词性 (Part-of-Speech, POS)，并且与原词的词干 (word stem) 不同，以保证生成句子的句法正确性和语义合理性。上述两个例子将被分别更改为“*the last daughter of* (最后一个女儿)”和“*was named in* (命名于)”。

第二步　生成一个伪答案以确保内容的兼容性。因为如果将第一步生成的伪关系直接指向原始正确答案，一方面有可能引入与原段落文本含义相矛盾的内容，另一方面也不能了解攻击句对问答模型的干扰情况到底如何（因为在干扰句中也还是能找到正确答案）。生成伪答案的具体做法为：从所有训练集答案中提取所有命名实体来构建一个伪答案集合，并将它们根据不同的NER类型进行划分。然后，对于要攻击的正确答案文本中的每个非停止词，先尝试使用与第一步相同的方法来找替换词，如果找不到则从伪答案集合中随机选择一个具有相同NER类型的词进行替换。例如，“*1998*”可以根据词向量相似度生成伪答案“*1999*”，“*Nicholas Farrar Hughes* (人名)”可以从伪答案集合中找到“*Otto Emil Plath* (人名)”来进行替换。

第三步　整合修改后的问题和伪答案，改写生成最终的攻击句。具体来说，使用Stanford Corenlp工具包构建得到每句问题的句法解析树，同时在Jia和Liang工作(Jia and Liang, 2017)的基础上修改设计了一系列转换规则（附录C中的表4展示了几个规则示例），能利用句法解析树将特殊疑问句转换为陈述句句式（HotpotQA是英文数据集，因此与中文不同，有特殊的语法结构）。例如在附录D图6交集类型中，用伪答案“*Otto Emil Plath*”替换特殊疑问词“谁”，经过陈述句转换后，生成的攻击句是“*Otto Emil Plath*既是*Aurelia Plath*的侄子，又是一名渔业生物学家。”，与原问句非常相似，其中一个子问句完全重合、另一个子问句略有修改，句法结构保持不变。显然，“*Otto Emil Plath*”不是一个有效（valid）的正确答案，因为它只满足原问题的部分要求属性。最后，生成的攻击干扰句被独立插入到每个段落的随机位置，同时根据情况修改推理依据句的序号标签。

2.3.2　对比类型和是否类型

对比类型和是否类型具有截然不同的推理机制，它们要求比较两个问题实体的属性，在问题表述中不存在明确的多跳，因此需要稍微不同的攻击修改策略。考虑到推理链中有一个特殊的比较操作节点，我们提出添加一条抽象性的路径来直接执行类似的攻击操作。

第一步，应用和2.3.1节相同的做法对问题表述中的属性关系进行修改，同时，通过抽取句法成分解析树中“*and* （和）”或“*or* （或）”前后两个论元，来找到问题比较的双方。

第二步，为了得到更有效的攻击，伪答案不再以随机的方式生成：（1）对于对比类型，反向地将另一个非正确答案的问题实体设为伪答案；（2）对于是否类型，对正确答案为“是”的反向设为“否”，对正确答案为“否”的反向设为“是”。

第三步，仍是将修改后的问题和伪答案整合，转换为语法正确的语句并加到输入段落中。

3　实验与分析

3.1　数据集和被攻击模型

我们在含有7405条样本的**HotpotQA**(Yang et al., 2018)开发集上对本文所提出的对抗攻击方法进行评估[1]。HotpotQA是目前最广泛使用的英文多跳问答基准数据集，它是基于文本的问答，问题用自然语言表述，覆盖了所有四种推理类型，因此是最有挑战性也是最适合本对抗攻击评估的问答任务。在该数据集中，每个样本提供10个独立的文章段落和一个问题，其中有2个段落为包含了回答该问题所必要的推理依据的相关段落，其余8个为噪声段落。此问答任务的评价指标包括对答案、推理依据及这两者联合预测的EM（Exact Match，精确匹配）和F1分数。由于其测试集为非公开数据，因此本实验在其开发集上进行测试评估。

我们对三个问答模型进行了攻击测试和对抗增强训练：（1）**HotpotQA基线模型**(Yang et al., 2018)：基于循环神经网络（RNN），结合了字级别模型、自注意力和双向注意力机制；（2）**DFGN**(Qiu et al., 2019)：一项声称能进行可解释多步推理的代表性和基础性工作，使用能促进文本和实体图信息交互的融合模块来动态选择子图进行信息传播，并沿着实体图执行逐步推理；（3）**SAE**(Tu et al., 2020)：官方排行榜上可访问的指标分数最高的模型，基于图神经网络，使用上下文句子的文本嵌入表示而不是实体来作为节点。

[1]官方排行榜https://hotpotqa.github.io，由于其测试集为非公开数据，因此本实验在开发集上进行测试评估。

3.2 实验设置

段落过滤模块在HotpotQA训练集上进行训练，用原数据集提供的正确推理依据作为监督标签。在开发集上的测试结果显示，训练后的段落过滤模块对所有的正确相关段落的召回率达到97.1%，这保证了预测答案的可达性；另一方面，有85.5%的样本将过滤得到段落数量从10个减少到4个以下，极大程度地减少了后续构建实体图的规模。

对桥接问题和交集问题的**子问题拆解和推理类型预测模型**使用官方提供的预训练BERT-base-uncased 模型(Devlin et al., 2019)初始化编码器，然后在少量数据上进行了训练。训练数据的构造方式为：从原训练集和开发集中标记为Bridge 的样本中，分别随机抽取了700 条和200条作为本任务的训练集和开发集，对每个问题文本，人工标注了第一跳子问题的起始位置、结束位置和推理类型。训练阶段，损失函数参数λ设置为1，最大输入序列长度设置为150，使用Adam 优化算法(Kingma and Ba, 2014) ，学习率为5×10^{-5}。经训练，该模型在子问题拆解（第一跳子问题的起始位置和结束位置的联合预测）上达到了72% 的F1 分数，在推理类型的预测上达到92% 的准确率。

在**对抗样本构造**阶段，本方法通过一些设计来保证生成句子的自然性和逻辑性，例如限制替换词具有相同的 NER 类型和 POS 词性、设置伪答案来确保对抗句与原文内容和答案不造成冲突。由于人力资源有限，本实验没有对整个对抗数据集进行人工标注和核对修改来进一步提高生成文本的质量，但是在随机抽样的100 条样本上的人工评估显示，所构造的对抗句子都是通顺自然的，且不会对人类完成问答造成困惑，人类的表现不会受到对抗样本的欺骗而下降。

3.3 实验主要结果

实验首先使用原开发集（dev-ori）数据构造了对应的对抗性开发集（dev-adv），对三个原始的问答模型进行了**对抗攻击**，根据性能下降的程度可以评估问答模型的鲁棒性。然后从原训练集（train-ori）中随机抽取20%的数据来构造对抗性样本（train-adv），作为扩充数据和原训练集混合（train-aug），重新从头进行**对抗增强训练**并测试评估。总体实验结果如表1所示。

问答模型	训练集和测试集	答案EM	答案F1	依据EM	依据F1	联合EM	联合F1
HotpotQA 基线模型	train-ori + dev-ori	42.5	56.7	16.5	59.2	8.3	35.6
	train-ori + dev-adv	29.9	41.4	1.4	19.9	0.7	9.8
	train-aug + dev-ori	43.2	57.5	19.8	61.4	10.1	37.6
	train-aug + dev-adv	41.9	56.1	6.5	40.0	3.5	20.9
DFGN模型	train-ori + dev-ori	54.2	68.5	50.2	81.3	31.3	58.5
	train-ori + dev-adv	23.6	31.7	6.2	29.7	3.9	14.6
	train-aug + dev-ori	54.9	68.7	48.1	80.6	30.8	58.3
	train-aug + dev-adv	40.1	51.2	13.0	48.8	8.5	30.1
SAE模型	train-ori + dev-ori	68.1	81.4	63.4	87.5	47.1	73.3
	train-ori + dev-adv	43.0	54.1	26.6	54.1	19.1	39.3
	train-aug + dev-ori	65.9	79.9	61.4	86.3	44.1	71.0
	train-aug + dev-adv	53.0	65.4	42.8	71.4	30.3	53.3

Table 1: 三个问答模型在HotpotQA数据集上的原始性能、对抗攻击下的性能、和对抗增强训练后的性能。所有指标的单位为%。

3.3.1 对抗攻击

对比表1中train-ori + dev-ori 和train-ori + dev-adv两行，即测试在原数据集上训练的问答模型在遇到对抗攻击时是否仍能抵御噪声、过滤有用信息、保持良好性能，实验有以下发现：

- 所有模型在答案预测和推理依据预测方面都出现了显著的性能下降情况：特别是基线模型和DFGN模型，答案的EM分数下降到了30%以下，推理依据的EM分数更极其低，只有1.4%和6.2%；相较而言，SAE模型的表现依旧相对强劲，但也出现了大幅下降（答案和推理依据的EM分数分别下降了25.1%和36.8%）。
- 以DFGN模型为例，对对抗性开发集中回答错误的样本进一步统计分析发现：在所有错误样本中，有52.2%将本方法构造的伪答案作为预测答案输出，93.7%的样本将所添加的对抗干扰句预测成了推理依据之一；在原来回答正确、对抗攻击下回答错误的部分样本中，这两个比例分别为59.1%和94.6%。这些是最能直观说明攻击成功性的样例。

- 另外值得关注的一点是，虽然这些模型在推理依据的预测上表现很差，但在答案预测上的表现却要高出很多（答案EM比依据EM分数均高出了16.4%至28.5%不等）。这样不合理的现象使得模型执行多跳推理的可解释性值得怀疑，因为在缺乏足够的依据来进行完整推理的情况下，是不应当还能找到答案的。

以上的结果表明，本文设计的对抗攻击方法确实有效地对问答模型造成了干扰，所添加的对抗句通过与所问问题表述有浅层的相似度，成功地分散问答模型的注意力，从而误导模型给出错误答案。这些问答模型存在不鲁棒的问题，且可能利用了单词匹配等简单的推理捷径来完成问答，而不是真正地进行多步骤的推理，违背了多跳问答任务的目的。

令人惊讶的是，尽管DFGN和SAE模型在原开发集上的表现大大优于基线模型，但在对抗攻击下，它们的答案EM分数下降更多（DFGN甚至反而比基线模型低了6.3%）。本文分析，由于基线模型更多依赖于纯文本语言理解，而DFGN和SAE利用图网络来进行多步信息聚合和推理，因此可以合理推断，本文攻击方法对这些精心设计的声称进行可解释多步推理的模型更具有挑战性。本攻击方法针对某些跳（子问题）进行关系的改写时，等价于向实体关系图中添加了干扰性的边和节点，这样伪造的推理路径与原正确推理链有不同程度的重叠。这些基于图的问答模型是否有能力找到并专注于正确的关系和实体，而避免被相似的但非连通（不能到达答案实体）的路径所误导而偏离了正确路径甚至受困，是一个至关重要的问题。

3.3.2 对抗增强训练

对比表1中经过正常训练（train-ori）和对抗增强训练（train-aug）后的结果有以下发现：

- 对比train-ori+dev-adv 和 train-aug+dev-adv 两行，在对抗性开发集上，基线模型、DFGN模型、SAE模型的答案EM分数分别提高了13.3%、31.3%和10.0%，推理依据EM分数分别提高了18.4%，41.9%和16.2%，这说明模型学习到了抵御对抗攻击的能力，鲁棒性大大提高。
- 对比 train-ori + dev-ori 和 train-aug + dev-ori 两行，在原干净开发集上，三个模型的性能均呈基本持平的表现，答案EM分数分别提高0.7%、提高0.7%、降低2.2%，推理依据EM分别提高3.3%、降低2.1%、降低2.1%。总的来说，对抗增强训练并不会大幅度削弱模型解决原始问答任务的能力。
- 对比 train-ori + dev-ori 和 train-aug + dev-adv 两行，经对抗增强训练之后的模型在对抗攻击下的性能，相比原模型在原干净开发集上的性能，下降程度已大幅度减少，基线模型的答案EM分数41.9%甚至已与原来的42.5%达到相当的水平，再次有力验证了模型已经具备较强的抵御攻击的能力。

上述实验结果都证明了本文所提出的对抗数据增强训练有助于提升问答模型的性能。**本方法的有效性主要在于增强模型的鲁棒性**，这也是本研究的出发点，通过在训练阶段向模型输入特殊设计的极具干扰性的噪声，使模型适应并学习分辨的能力，在提高寻找答案的能力的同时也提高过滤噪声的能力，从而变得更加稳健。在原始问答任务上，由于所构造的对抗样本本质上并没有引入全新的数据（问题答案对、用于推理的段落上下文都没有增加），因此，模型的问答能力难以得到进一步显著提升。另一方面，由于所添加的对抗句在一定程度上改变了数据的分布特征，使得训练和测试期间的数据分布存在一定差异，因此可能会导致在原始测试集上准确性的些许下降，先前几项关于对抗攻击的研究工作普遍发现了这样的问题(Jia and Liang, 2017; Wang and Bansal, 2018; Jiang and Bansal, 2019)。而经过本文方法增强训练的问答模型均没有表现出原始问答性能的降低，甚至还有略微提升，这说明了本方法能够促进模型在保持原始能力的基础上额外提高抵御干扰的能力，同时做到又好又稳。

3.4 和其他攻击方法的比较

本节将本对抗攻击策略与之前的两种攻击方法进行比较，并分析本方法的优越性。

- **AddSent**(Jia and Liang, 2017)：将整个问题表述进行修改，替换其中所有的名词、形容词、命名实体和数字，并从事先定义的集合中选择固定的伪答案，修改后的句子被添加到输入段落的末尾。该方法是针对单跳问答提出的，由于是在整个问题句子范围内修改，因此对抗句与原问题的关联度和相似度都会降低，也无法定位是在哪个推理环节出了错。并且，对抗句总是被加到段落最后，容易被问答模型捕获这样的特征而造成攻击评估无效。
- **AddDoc**(Jiang and Bansal, 2019)：将相关段落中的桥实体进行替换，并注入伪答案，以此构造整个对抗干扰段落作为额外的输入上下文，其中桥实体是从段落标题中抽取得到的。该方法的问题在于，如果数据集没有事先告知哪些段落是相关段落，或者没有提供文章标题，或者桥实体并不存在于文章标题中，那么就无法进行攻击数据的构造。

此外，这两种方法还有共同的一点不足是，都无法对对比类型和是否类型的问题实现对抗数据的构造。因此，在表2中以DFGN模型为例，仅比较在桥接类型和交集类型样本上的性能。

训练集	开发集	答案EM	答案F1	依据EM	依据F1	联合EM	联合F1
train-ori	AddSent	20.1	29.3	3.1	28.4	1.2	12.1
	AddDoc	38.9	54.8	27.9	63.7	16.4	40.5
	本文方法	**17.7**	**28.1**	**1.4**	**26.5**	**0.5**	**9.3**
train-aug	AddSent	35.3	46.7	7.1	30.6	3.1	18.8
	AddDoc	43.0	59.5	30.0	64.9	18.0	41.3
	本文方法	35.1	46.8	6.3	42.7	4.0	24.7

Table 2: 在桥接类型和交集类型的对抗开发集样本上，DFGN模型在本攻击方法下与其他两种攻击方法下的性能比较。加粗的为最低性能，即代表攻击效果最好。所有指标单位为%。

对抗攻击　使用在原训练数据集（train-ori）上训练的问答模型对三种对抗性开发集进行预测评估的实验结果显示，DFGN的问答准确性在本文方法的攻击下下降最多，验证了本文提出的对抗方法针对多跳问答是更具有混淆干扰性的。总的来说，本攻击方法有三个优点：(1) 可以对所有的推理类型进行更全面攻击，且不受推理依据可获得性的限制；(2) 攻击的成功率更高，本文认为这是由于所构造的攻击句子与问题具有更高的相似程度，以及保留了对多跳推理过程来说更为重要的实体；(3) 能够进行有针对性的攻击（如可以针对某一跳，只对相应的子问题作修改），从而对模型真正的多跳推理能力进行检验和分析。

对抗增强训练　实验还使用本文方法构造的对抗性增强训练数据（train-aug）对DFGN模型进行从头训练，然后在三个对抗性开发集上进行评估，以测试本增强训练的模型是否具备抵御各种多样化攻击的能力。可以看到，在AddSent、AddDoc、本文方法的攻击下，答案EM分数比原模型（train-ori）分别提高了15.2%、4.1%、17.4%。其中，AddSent和本文方法一样施加的是句子级别的扰动，而AddDoc是文档级别的扰动，因此将本增强训练直接迁移至后者所带来的提升效果相对偏少。虽然这两种方法所构造数据的特点和分布均和本文方法有很大不同，但是增强训练后的模型在所有攻击下的表现都获得了一致的提高，这样的结果证明，本文提出的对抗性增强训练所带来的模型鲁棒性加强是广泛的、普适的，能够帮助模型普遍地更好地抵抗其他各种攻击，而非只针对性地局限于本攻击方法。

3.5　对比实验

对比内容	推理类型	攻击目标	HotpotQA基线模型		DFGN模型		SAE模型	
			答案EM	答案F1	答案EM	答案F1	答案EM	答案F1
攻击不同跳	桥接类型	第一跳	**26.7**	**37.6**	**16.4**	**24.9**	**47.7**	**60.9**
		第二跳	27.1	38.0	17.1	25.6	48.0	61.1
		两跳	35.4	48.5	28.5	39.0	52.9	66.8
	交集类型	随机一跳	**28.8**	**43.6**	**26.5**	**37.4**	**48.1**	**63.2**
		两跳	31.5	46.8	32.2	44.6	49.7	65.2
攻击不同类型的词语	全部	关系词	**29.9**	**41.4**	**23.6**	**34.6**	**43.0**	**54.1**
		实体词	30.9	42.2	26.2	36.9	45.1	56.0

Table 3: 对比实验：针对不同跳和不同类型词语的攻击策略效果比较。加粗为攻击效果最好。

3.5.1　攻击不同跳的影响

将多跳问题拆解成多个子问题的操作使得本对抗攻击方法可以根据需要对推理链的任一部分进行攻击，以检测问答模型在不同推理阶段的能力。主实验通过修改桥接类型的第一跳和交集类型的随机一跳来构造对抗句，本部分尝试不同的修改策略，包括 (a) 同时修改第一和第二跳，即对整个问题句子进行修改；(b) 对桥接类型仅修改第二跳。

原模型在对抗性开发集上的结果如表3所示。当同时对两跳进行修改时，问答模型的答案预测能力相对较好，即说明注意力较少受到干扰，这是符合预期的，因为将两跳都改掉之后，生成的攻击句与正确的推理链完全没有重叠之处，无关性高，因此干扰混淆性较低。另外，对于

桥接类型，对第二跳进行攻击的攻击成功率比攻击第一跳的低，这也是容易理解的，攻击第一跳时，问答模型可能在一开始就被误导至错误的推理路径，离正确答案更远，因此最终能找回正确答案的概率也就越低；同时，由于第二跳问题表述与原问题中的文本保持了相同，这样的结果也意味着模型可能在第二跳推理过程中更倾向于使用简单的单词匹配等非多跳推理策略。

3.5.2 攻击实体的影响

本节通过修改问题文本中的命名实体而不修改表示关系的词语，来验证实体在多跳问答任务中起到很重要作用的假设。表3所示的结果验证了这样的假设。修改实体后的答案预测分数都高于修改关系后的，即攻击成功率低，这表明攻击句如果不含有问题所关心的实体，则可能导致相关程度较低，从而缺少干扰性。这也印证了第3.4节中所分析的本文攻击方法优于AddSent和AddDoc方法的原因之一。

3.5.3 不同训练增强数据比例的影响

(a) HotpotQA基线模型的答案EM分数。

(b) DFGN模型的答案EM分数。

Figure 4: 问答模型使用不同比例的增强数据训练后，在原开发集和对抗性开发集上的性能。横轴表示在全部原数据train-ori基础上扩充的对抗样本train-adv的比例，其中，第一列*0%*表示全部使用train-ori，最后一列*0+100%*表示使用全部train-adv且完全不使用train-ori。

主实验的对抗增强训练使用20%的对抗数据与所有原始训练数据相混合，本节尝试不同的数据增强比例。如图4中的结果所示，随着对抗训练样本数量的增加，模型可以更多地学习关于噪声的知识，从而提高抵御攻击的能力，因此在对抗性开发集上的性能呈现上升的趋势（橙色虚线）。在原开发集上，两个问答模型的性能总体上都变化不大，由于对抗样本引入的数据分布特征差异，出现了极微小的降低趋势（蓝色实线）。特别地，实验还尝试了仅使用所有对抗样本进行训练（0% train-ori + 100% train-adv），即最后一列*0+100%*所示，虽然问答模型能更有效地应对攻击（两个模型在对抗性开发集上答案EM分别达到49.0%和44.0%，均高出使用混合训练数据的情况），但是对干净数据的泛化能力会被大大削弱（两个模型在原开发集上的性能均低于使用混合训练数据的情况）。因此，需要谨慎选择合适的对抗数据增强比例，并且注意防止问答模型对对抗攻击数据过拟合，以平衡模型同时具备良好的问答准确性和鲁棒性。

3.6 案例研究

由于多跳问题本身难度不同以及问答模型设计的特点，模型对不同推理类型问题的解决能力存在很大差异，附录E对不同类型的样本进行了分类统计和分析。附录F还通过对一条真实样本的研究，来对本对抗攻击和对抗增强训练进行可解释性分析。

4 结论

本文提出了一项基于推理链的针对多跳问答的对抗攻击方法。通过将多跳推理过程形式化建模成推理链，可以识别不同的多跳推理类型，并对每种推理类型设计更有针对性的对抗攻击策略。本攻击方法通过识别不同跳对应的子问题文本，对关系词进行修改，来支持对其中任意一跳进行攻击，这可以帮助检测模型在推理过程中容易出错的部分。本文攻击评估的三个问答模型在面对攻击干扰时都出现了性能的下降，表明它们不够鲁棒，执行多步推理的可解释性有限。此外，利用所构造的攻击数据作为增强数据进行训练，可以普遍增强问答模型抵御攻击的鲁棒性，同时本工作也希望根据检测到的薄弱之处来促进开发出更好的问答模型。

第二十二届中国计算语言学大会论文集，第1页–第16页，哈尔滨，中国，2023年8月3日至5日。

参考文献

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proc. of NAACL*, pages 2306–2317, June.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NACCL*, pages 4171–4186, June.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *Proc. of ACL*, pages 2694–2703, July.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proc. of EMNLP*, pages 8823–8838, November.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proc. of EMNLP*, pages 2021–2031, September.

Y. Jiang and M. Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. *ACL*.

T. Khot, P. Clark, M. Guerquin, P. Jansen, and A. Sabharwal. 2019. Qasc: A dataset for question answering via sentence composition.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, and D. Mcclosky. 2014. The stanford corenlp natural language processing toolkit. In *Proc. of ACL*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

T. Onishi, W. Hai, M. Bansal, K. Gimpel, and D. Mcallester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proc. of EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Ethan Perez, Patrick S. H. Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. *CoRR*, abs/2002.09758.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proc. of ACL*, pages 6140–6150, July.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*, pages 2383–2392, November.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proc. of ACL*, pages 784–789, July.

N. Shao, Y. Cui, T. Liu, S. Wang, and G. Hu. 2020. Is graph structure necessary for multi-hop reasoning?

L. Song, Z. Wang, M. Yu, Y. Zhang, R. Florian, and D. Gildea. 2018. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proc. of NAACL*, pages 641–651, June.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proc. of EMNLP*, pages 8846–8863, November.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proc. of ACL*, pages 2704–2713, July.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9073–9080.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proc. of NAACL*, pages 575–581, June.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *b*, 6:287–302.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proc. of EMNLP*, pages 2369–2380, October-November.

Deming Ye, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2019. Multi-paragraph reasoning with knowledge-enhanced graph neural network. *CoRR*, abs/1911.02170.

M. Zhou, M. Huang, and X. Zhu. 2018. An interpretable reasoning network for multi-relation question answering.

附录

A 相关研究

A.1 多跳问答

对于SQuAD(Rajpurkar et al., 2016; Rajpurkar et al., 2018)等单跳问答任务，问答模型可以通过简单地将问题与输入的单个段落中的句子进行匹配来检索答案。而对于多跳问答，只凭任一单句都不足以回答问题，模型需要结合至少两段文本的信息并基于它们进行推理。目前已公开了多个多跳阅读理解数据集，包括Wiki-Hop(Welbl et al., 2018)、ComplexWebQuestion(Talmor and Berant, 2018)、HotpotQA(Yang et al., 2018)等等。

很多多跳问答相关研究采用了图神经网络。DFGN(Qiu et al., 2019)根据输入上下文文本构建实体图，并在图上重复地执行单跳推理，在每一步动态地选择相关子图来传播信息。KGNN(Ye et al., 2019)从知识图谱中获取关系信息，基于共指关系来添加边以增强实体图。CogQA(Ding et al., 2019)通过添加每个现有节点的下一跳实体和可能的答案文本来逐步扩展认知图（cognitive graph）。与实体图不同，SAE(Tu et al., 2020)使用段落文本句子的嵌入表示作为节点，并将推理依据预测任务视为节点分类。HGN(Fang et al., 2020)构建了一个层次图来综合不同粒度级别的信息并促进它们之间的交互作用。类似的研究工作还包括文献(Song et al., 2018; Tu et al., 2019; De Cao et al., 2019; Shao et al., 2020)。除了图神经网络以外，其他的一些方法使用记忆网络来解决多跳问答(Zhou et al., 2018; Onishi et al., 2016)。

A.2 对抗攻击

尽管最近的研究工作展现了在多跳问答任务上的巨大进展，但机器阅读理解和多事实推理的真实能力以及模型的鲁棒性仍然值得怀疑。借鉴计算机视觉领域，可以通过对输入上下文添加轻微的扰动来对自然语言处理任务进行对抗性攻击，被攻击的模型预期会给出错误的输出。其中，添加的扰动不能与原始上下文段落内容有语义冲突，也不能改变原始正确答案。

针对问答，AddSent(Jia and Liang, 2017)是第一项研究对抗攻击的工作，通过替换给定问题中的命名实体、并与从事先定义好的集合中选择的伪答案相结合的方式来生成攻击句，将其添加到输入上下文的最后位置，实验显示16个模型在单跳问答任务SQuAD上都出现了性能的下降。AddSentDiverse(Wang and Bansal, 2018)在AddSent的基础上进行了改进，通过扩大伪答案候选集和改变攻击句的插入位置来使干扰更加多样化，同时发现对抗性再训练可以提高问答模型抵御攻击的鲁棒性。T3(Zhou et al., 2018)设计了树形自动编码器来对文本进行编码，使其保留句法结构和语义信息，然后在词级别和句子级别上施加基于最优算法的扰动，可以实现针对位置的攻击和针对答案的攻击。与单跳问答相比，多跳问答还存在另一种可认为是失败的推理情况，通常称为推理捷径。DiRe(Trivedi et al., 2020)通过在输入段落文本中删除部分推理依

据文本来探究非连续推理情况的存在，多跳问答模型不应当在推理依据缺失的情况下依旧给出正确答案，否则可认为存在推理捷径。为了探究是否存在简单的单词匹配策略，AddDoc(Jiang and Bansal, 2019)用不相关的表述替换相关段落中的真实正确答案和桥实体，并将这个修改后的段落添加到输入文本中作为干扰。

B 子问题拆解和推理类型预测的模型框架图

第2.2节中提出的子问题拆解和推理类型预测的模型框架图5所示。

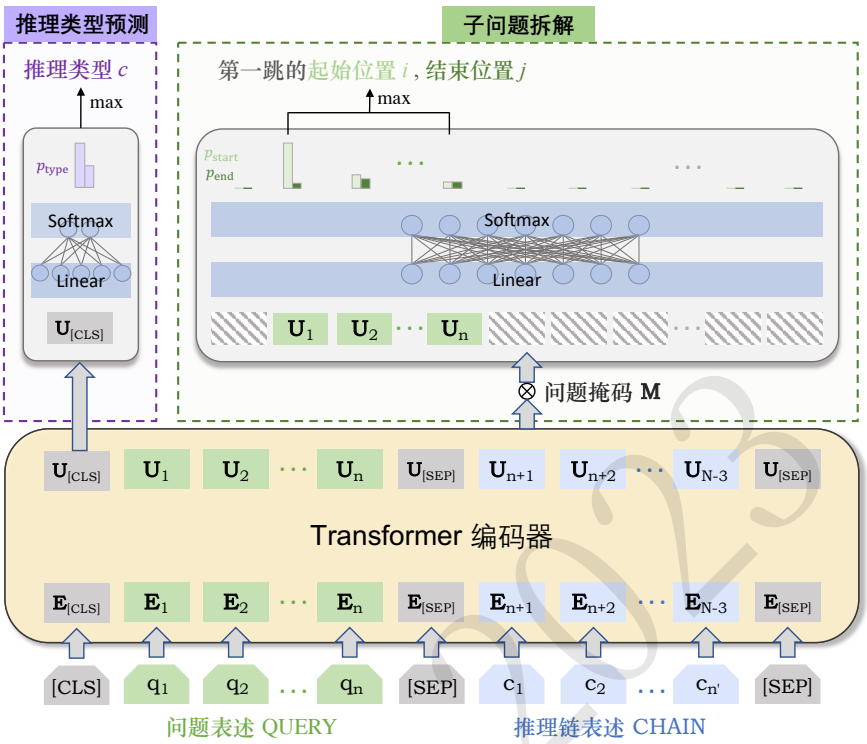


Figure 5: 针对桥接类型和交集类型的子问题拆解和推理类型预测的模型框架。

C 对抗攻击句子构造阶段的第三步中句式转换规则示例

表4展示了第2.3节中提到的所设计的句式转换规则中的部分示例。

疑问句模版：*What/Which $NP $VP ?* 陈述句模版：*The $NP of [Answer] $VP .* 疑问句示例：Which football team won the first prize in the 2022 World Cup? 陈述句示例：The football team of [Argentina] won the first prize in the 2022 World Cup.
疑问句模版：*When/Where $Do $NP $Verb $PP ?* 陈述句模版：*$NP $Verb-tense{2} $NP in [Answer] .* 疑问句示例：Where did Lisa and Jack meet for the first time? 陈述句示例：Lisa and Jack met for the first time in [Shanghai].
疑问句模版：*How $JJ $Be $NP ?* 陈述句模版：*$NP $Be [Answer] .* 疑问句示例：How tall is the highest mountain in the world? 陈述句示例：The highest mountain in the world is [8848.86 meters].
注：*$NP*、*$VP*、*$PP*、*$Verb*、*$JJ*、*$Be*、*$Do*分别为语法解析树中的名词短语、动词短语、介词短语、动词单词、形容词、be动词、助动词，*-tense{2}*表示使用疑问句模版中第2个元素的时态。

Table 4: 基于句法解析树的将特殊疑问句转换为一般陈述句的规则示例。

D 本方法所构造的对抗攻击样本示例

图6为每一种推理类型都展示了一条问答样本，并使用所提出的方法构造了不同的对抗攻击句子。

Figure 6: 根据不同的推理类型施加相应对抗攻击策略的示例及其对推理链的干扰解释。用阴影、实下划线、实线黑框标记的词语分别是问题实体、桥实体、答案实体，用虚下划线、虚线黑框标记的是本攻击方法构造并引入的干扰边（关系）和干扰节点（伪答案实体）。

E 不同推理类型的实验结果

在HotpotQA数据集中，四种推理类型所占的比例不同，问答模型对这四类的解决能力也有很大差异，实验对每一类分别统计了分数。由于三个问答模型的表现有相似的趋势，因此在表5中仅以DFGN模型的表现为例进行说明。

对比train-ori + dev-ori 和train-ori + dev-adv 两列可以发现，在对抗攻击下，桥接类型是最不鲁棒的类型，EM分数下降最多也降至最低，仅有14.6%，交集类型是第二易被攻击的类型，这两类也是数量占比最多的多跳问答类型。这证明了本文基于关系的攻击方法是对多跳问答有针对性的，在中间推理过程中，连续依次的推理跳可能会陷入错误的推理路径，以致偏离真正的答案。另一方面，DFGN擅长回答比较型问题，特别是是否类型，在原开发集上已经达到74.4%，远远高于其他三类，并在攻击下仍能保持在69.1%的高分，下降仅5.3%。对这类问题

推理类型	占比	train-ori +dev-ori	train-ori +dev-adv	train-aug +dev-ori	train-aug +dev-adv
所有数据	100%	54.2	23.6	54.9	40.1
桥接类型	49.6%	54.1	14.6	54.3	33.8
交集类型	32.0%	50.2	22.6	51.3	37.0
对比类型	12.9%	56.3	41.1	56.7	56.4
是否类型	5.5%	74.4	69.1	76.3	76.8

Table 5: DFGN模型在不同推理类型数据上的答案EM分数。所有指标的单位为%。

的攻击成功率较低的原因可能为，本攻击方法没有干扰或破坏问题所关心的属性或正确推理链，而只是添加了一个独立的“相同”或“不相同”的抽象性关系节点，因此对原问题的影响相对有限。

经过对抗增强训练之后，DFGN模型在四种类型上的表现都获得了提升。对比train-ori + dev-adv 和train-aug + dev-adv 两列，对抗增强训练后的模型抵御对抗攻击的能力得到了增强，在四种类型上分别提高了19.2%、14.4%、15.3%和7.7%。其中桥接类型提升最多，说明本增强训练方法可以弥补模型的薄弱之处，对最不鲁棒的部分进行着重加强；而对比类型和是否类型这两类原本就比较稳健的问题，已经达到了和原模型在原干净数据上相当的性能，说明这两类表现优异的推理类型已经几乎能够不再受对抗攻击的干扰。对比train-ori + dev-ori 和train-aug + dev-ori 两列，对抗增强训练后的模型在原干净开发集上的准确性都有了略微的提高，进一步证明了本方法的有效性。

F 案例研究

本部分以图7所展示的一条开发集案例进行研究，以对对抗攻击和对抗增强训练进行可解释性分析。实验计算了问题中每个单词对输入上下文中每个实体对应文本的注意力分数，并进行softmax归一化，然后对所有问题单词进行求和，从而得到每个实体所分配到的总的注意力权重。这些注意力权重一定程度上揭示了模型的推理过程。

在第一跳时，由于是基于上下文和问题之间的双向注意力机制进行推理初始化的，因此权重最高的大多是问题实体；另外，模型也成功定位到了桥实体*Shirley Temple*，分配了较高的注意力；其次是它们的相邻节点（比如在出现同一句句子中的实体），并依此传播信息；在第二跳时，正确答案*Chief of Protocol*已经成为注意力权重最高的实体。

但是在对抗测试样本（train-ori + dev-adv）上，模型在第一跳时就被对抗句分散了较多的注意力，因为句中包含了问题实体；在第二跳时，伪答案集中聚合了来自问题实体的信息，注意力分数超过了正确答案成为最高；在最终预测时，几乎所有概率都落在了重复多次出现的伪答案上。

经过对抗增强训练之后的模型（train-aug + dev-adv），在第一跳时对对抗句的注意力就受到了抑制（仅有0.28和0.34），对桥实体注意力大大加强（达到4.4和3.2），同时也提高了对正确答案的提前关注（达到4）；在第二跳时，对推理依据的关注程度也高于对抗句；最后的答案预测结果显示，对正确答案的预测概率从原来的0.57增加到0.93，不仅预测正确，还大大提高了置信度。

问题	What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell? (《吻和倾诉》中扮演柯丽丝·阿彻的女士担任什么政府职位？)
段落1	Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. The government position of Director of Diplomacy was held by the man who voiced Corliss Archer in the film Kiss and Tell. (《吻和倾诉》是1945年上映的美国喜剧电影，由当时17岁的秀兰·邓波儿饰演柯丽丝·阿彻。外交部主任的政府职位是由电影《吻和倾诉》中配音柯丽丝·阿彻的男士担任的。)
段落2	Shirley Temple Black was named United States ambassador to Ghana and to Czechoslovakia and also served as Chief of Protocol of the United States. The government position of Director of Diplomacy was held by the man who voiced Corliss Archer in the film Kiss and Tell. (秀兰·邓波儿·布莱克被任命为美国驻加纳和捷克斯洛伐克大使，并担任美国礼宾司司长。外交部主任的政府职位是由电影《吻和倾诉》中配音柯丽丝·阿彻的男士担任的。)
正确答案	Chief of Protocol (礼宾司司长)
预测答案	Director of Diplomacy (外交部主任)

注：用绿色、橙色、蓝色、红色标注的分别是问题实体、桥实体、答案实体、伪答案和伪关系；问题中划横线的为第一跳子问题；段落文本中划虚线的为构造的对抗句，对抗句被插入到每个输入段落的随机位置。

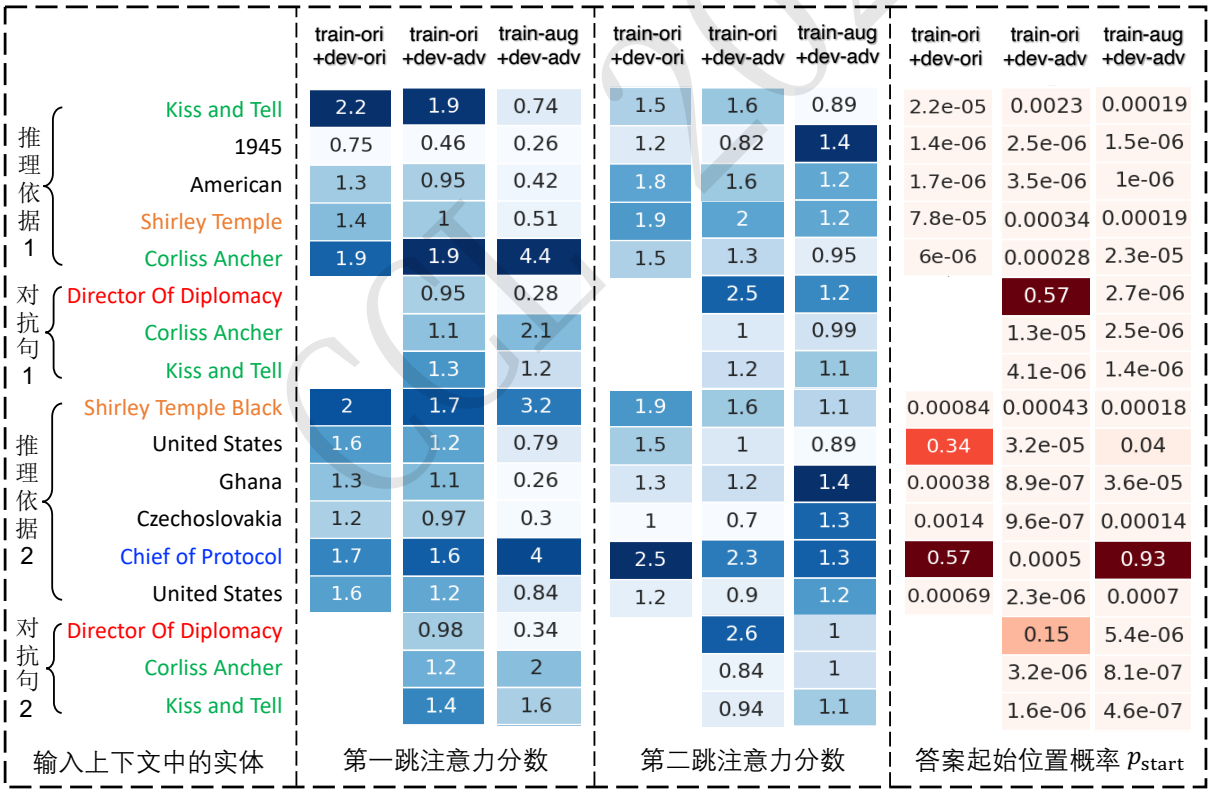


Figure 7: 案例研究：使用不同训练集训练的DFGN模型在原开发集和对抗性开发集上，输入上下文中实体在每一跳受到的注意力权重分数。

基于不完全标注的自监督多标签文本分类

任俊飞, 朱桐, 陈文亮*
苏州大学计算机科学与技术学院
江苏,苏州,2150062
{jfrenjfren,tzhu7}@stu.suda.edu.cn, wlchen@suda.edu.cn

摘要

多标签文本分类(Multi-Label Text Classification, MLTC)旨在从预定义的候选标签集合中选择一个或多个文本对应的类别，是自然语言处理(Natural Language Processing, NLP)的一项基本任务。前人工作大多基于规范且全面的标注数据集，而这些规范数据集需要严格的质量控制，一般很难获取。在真实的标注过程中，难免会丢失掉一些相关标签，进而导致不完全标注问题。为此本文提出了一种基于局部标注的自监督框架(Partial Self-Training，PST)，该框架利用教师模型自动地给大规模无标注数据打伪标签，同时给不完全标注数据补充缺失标签，最后再利用这些数据反向更新教师模型。在合成数据集和真实数据集上的实验表明，本文提出的PST框架兼容现有的各类多标签文本分类模型，并且可以缓解不完全标注数据对模型的影响。

关键词： 多标签文本分类 ； 不完全标注 ； 自监督学习

Self-Training With Incomplete Labeling For Multi-Label Text Classification

REN Junfei, ZHU Tong, CHEN Wenliang*
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, 215006, China
{jfrenjfren,tzhu7}@stu.suda.edu.cn, wlchen@suda.edu.cn

Abstract

Multi Label Text Classification (MLTC) is a fundamental task of Natural Language Processing (NLP). It selects the most relevant labels from the predefined label set to annotate texts. Most of the previous studies are conducted on standardized and comprehensive datasets with manual annotations, which require strict quality control and are difficult to obtain. In the real annotation process, it is inevitable to lose some related labels, which leads to the problem of incomplete annotation. We propose a Partial Self-Training (PST) framework to address this problem. The teacher model not only generates pseudo labels on large-scale unlabeled data, but also provides supplement tags to incompletely labeled data. Finally, the teacher model is updated iteratively based on these data. Experiments on synthetic data sets and real data sets show that our proposed PST framework is compatible to different kinds of teacher models, and can alleviate the impact of incomplete labeled data.

Keywords: Multi-Label Text Classification , Incomplete Labeling , Self-Training

* 通讯作者 Corresponding Author.

第二十二届中国计算语言学大会论文集，第17页-第30页，哈尔滨，中国，2023年8月3日至5日。

1 引言

多标签文本分类作为自然语言处理中一项基本且实用的任务，可以自动地标注与文本相关的标签，在情感分析(Li et al., 2016)、话题识别(Dougrez-Lewis et al., 2021)、问答(Langton et al., 2020)和网页标记(Jain et al., 2016)等许多领域都有应用。然而，由于标注体系的复杂性，标注过程中可能存在标签缺失的情况，从而形成不完全标注的数据集。如表(1)所示，给定一段金融领域文本，标注员在标注过程中只标注了“重大赔付”和“财务造假”两个相关标签，而遗漏了“破产清算”标签，这条标注数据就是不完全标注数据。这种不完全标注问题在多标签分类数据集中尤为明显，会导致多标签分类模型无法准确地预测出相关标签。这种缺失标签对模型的影响主要分为两方面：1)退化影响：大量缺失标签的存在导致与文本相关的正例标签数量减少，模型在少量相关标签的训练下无法学到更加全面完整的信息；2)误导影响：大量缺失标签在模型训练过程中被当作与文本不相关的负例标签计算，从而误导模型学习到相反的信息。面向不完全标注的多标签文本分类旨在从不完全标注数据集中学习文本到相关标签的分类器，同时尽量缓解缺失标签对模型的影响，提升多标签分类的性能。

文本	标签
涉案的美国三大投行遭到重罚,花旗集团和摩根大通因涉嫌财务欺诈被判有罪，向安然公司的破产受害者分别支付了20亿、22亿和6900万美元的赔偿罚款。	**重大赔付** **财务造假** *破产清算*

Table 1: 不完全标注样例，*“破产清算”*为缺失标签

现有针对多标签文本分类的方法主要集中在四个方面，分别是文本语义表示的研究、标签间关系的研究、标签分布的研究以及文本与标签语义链接的研究。文本语义表示的研究侧重于使用深度神经网络来提取出文本的深层语义表示(Liu et al., 2017)。标签间关系的研究通常利用标签式注意力机制(肖琳et al., 2020)来建模标签相关性。标签分布的研究通过设计特殊的损失函数和数据采样策略来缓解样本标签分布不均衡等问题。还有一些研究(Du et al., 2019; Pappas and Henderson, 2019)通过对文本与标签联合建模来探索文本与标签的语义链接。然而，这些研究都是在人工标注的数据进行监督训练，无法解决不完全标注的标签缺失问题。

为此本文提出了一种基于局部标注的自监督框架(Partial Self-Training，PST)，该框架通过补充利用缺失标签来缓解缺失标签对模型造成的负面影响。具体地，PST框架首先利用基础的多标签文本分类模型在不完全标注数据集上训练以获取教师模型,然后利用教师模型自动地给大规模无标注数据和不完全标注数据打分，接着利用双阈值机制对标签按得分进行状态划分以获取正例标签、负例标签以及其它标签。最后通过联合训练充分利用三种不同状态的标签信息对教师模型进行更新。总的来说，PST框架通过对缺失标签的补充利用，可以从两个方面缓解缺失标签对模型的影响：一方面带有伪标签的无标注数据与补充了缺失标签的不完全标注数据大大增加了模型训练的正例标签，进而大大缓解缺失标签带来的退化影响；另一方面随着对缺失标签的补充，伪负例标签也会相应减少，进而缓解了缺失标签带来的误导影响。

为了更加全面地评估PST框架的性能，本文分别在合成数据集和真实数据集上进行实验。实验结果表明，随着不完全标注问题的加剧，多标签文本分类模型的性能急剧下降，而PST框架可以在一定程度上缓解下降的速度，缺失标签越多缓解越明显。同时从不同多标签分类教师模型上的实验结果发现，在不完全标注的数据集上，PST框架对不同的教师模型都有着不同程度的改善，充分证明PST框架的通用性。总体来说，本文贡献有以下三点：

- 本文提出了一种兼容现有多标签文本分类模型的自监督学习框架PST，该框架通过补充利用缺失标签来缓解不完全标注数据集对多标签文本分类模型造成的负面影响；
- 本文对CCKS2022 Task8面向金融领域的Few-Shot事件主体抽取的评测任务数据集进行修正，构建了一个新的真实数据集CCKS-IMLTC，该数据能够更好地模拟真实标注场景下的不完全标注问题。该数据集及本文所提出方法的实现代码将在GitHub上开源，供业界学习研究。相关代码与数据均已在GitHub开源[0]；

基金项目：2020-2024自然科学基金重点联合项目：自然语言对话交互的基础理论和方法(61936010)

[0]https://github.com/15962171082/Incomplete_MLTC

- 在合成数据集和真实数据集上的实验表明，本文提出的PST框架具有通用性，并且能在一定程度上缓解数据不完全标注问题带来的影响。

2 相关工作

多标签文本分类目的是通过分类器自动获取与文本相关的一个或多个预定义标签，关于该任务的研究主要分为四类，第一类是关于文本语义表示的研究，第二类是关于标签间关系的研究，第三类是关于标签分布的研究，第四类是关于文本与标签语义链接的研究。

文本语义表示：Kim (2014)将图像中效果显著的CNN模型迁移到文本，捕捉窗口局部信息来更好地表征文本。Nam et al. (2017);Yang et al. (2018)基于Seq2Seq的方法，采用RNN对输入文本进行编码，并采用基于注意力的RNN解码器，依次生成预测的标签。Devlin et al. (2018)利用预训练语言模型BERT直接编码获取文本的语义表示。

标签间关系：标签间通常存在依赖、相似、相反和层级等关系。肖琳et al. (2020)提出了标签特定注意力网络，可以在预测每个标签的同时关注到其它标签。Zhang et al. (2021)引入多任务学习方法来增强标签相关性反馈，并利用联合编码机制同时获得文本和标签的表示。Zhao et al. (2022)集成标签增强与模型训练，有效挖掘不同标签的隐含相对重要性信息。

标签分布：由于复杂的标签体系，多标签分类数据集通常伴随着样本标签分布不均衡等问题。Chawla et al. (2002)通过设计特殊的数据采样方式使得采样后的训练集样本标签尽量均衡。然而基于采样的方法无法充分利用标注数据，为此Lin et al. (2017);Wu et al. (2020)通过设计不同的损失函数来缓解标签分布不均衡。此外Xiao et al. (2021)利用迁移学习将头部标签与尾部标签联系起来，以缓解长尾分布中尾部标签数据量少的问题。

文本与标签语义链接：Xiao et al. (2019)在文档和标签之间共享单词表示，利用标签语义信息来确定标签和文档之间的语义链接，从而构建特定于标签的文档表示，并采用自注意力机制建立两者之间的联系。Ma et al. (2021)基于图神经网络来捕捉文本与相应标签的语义互动，它利用全局统计模式和局部动态关系来推导不同标签特定语义部分之间的依赖关系。

上述研究将训练集视为规范标注数据集，并通过全监督训练获取分类模型，而忽视了本文讨论的不完全标注问题。Self-Training(Scudder, 1965)作为一种半监督学习的方法有着非常悠久的历史，该方法的主要思想是利用教师模型对大规模无标注数据进行自动标注以增加训练集的数量，进而更新优化教师模型。随着神经网络模型的发展和对标注数据需求的增加，Self-Training一直是一个热门的研究方向。该方法已成功运用于各种任务如：机器翻译(Jiao et al., 2021)、问答(Sachan and Xing, 2018)、关系抽取(Yu et al., 2022)等。近几年关于Self-Training的研究主要集中在选择伪标签的策略和模型特征的设计两个方面(Triguero et al., 2015)，本文提出的PST框架主要针对如何有效准确地选择伪标签这一方面。

3 预备知识

3.1 任务定义

我们假设$D = \{(x_i, y_i)\}_{i=1}^{N}$表示一个人工标注的多标签文本分类数据集，它由$N$个文本$x_i$及其对应人工标注的标签集合$y_i$组成。其中每个文本$x_i$由$n$个单词组成$x_i = \{w_{i1}, \cdots, w_{iq}, \cdots, w_{in}\}$，$w_{iq}$表示文本$x_i$的第$q$个单词。$x_i$对应的人工标注标签$y_i \in \{0, 1\}^l$，其中$l$是预定义标签的总数，相关标签记为1，不相关记为0。由于不完全标注问题的存在，人工标注的标签集合y_i可能存在遗漏缺失，因此我们进一步定义$\widetilde{y}_i \in \{0, 1\}^l$表示真实标签集合，相关标签记为1，不相关记为0。人工标签集合y_i与真实标签集合$\widetilde{y}_i$的异或即为缺失标签集合$\widehat{y}_i = \widetilde{y}_i \bigoplus y_i$，$\widehat{y}_i \in \{0, 1\}^l$，其中$l$是预定义标签的总数，缺失标签记为1其它记为0。

一般地，多标签文本分类任务需要学习一个从文本x_i到真实标签集合$\widetilde{y}_i$的分类器，但在真实场景中人工标注的标签集合y_i可能存在不完全标注问题导致缺失标签集合$\widehat{y}_i$的存在。而不完全标注的多标签文本分类的目的就是从已知的人工标签集合y_i出发，学习一个从文本x_i到相关标签的分类器，同时要尽可能地削弱未知的缺失标签$\widehat{y}_i$给分类器带来的影响。

3.2 教师模型

本节将不同的多标签文本分类模型抽象为统一的整体，从编码到解码再到训练介绍了多标签文本分类的任务流程，同时公式化的损失函数方便后续更加直观地展示PST框架训练中对模型损失函数的改进。

3.2.1 编码

首先我们将文本x_i输入BERT得到句子的特征表示序列$\boldsymbol{H} = \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\}, \boldsymbol{H} \in \mathbb{R}^{L \times d}$，计算如式(1)所示。

$$\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n\} = \text{BERT}(x_i) \tag{1}$$

其中每个字符的向量表示$\boldsymbol{h}_i \in \mathbb{R}^d$，$d$为词向量表示的维度，$L$为预设的最大文本长度。

接着为进一步提取句子的语义特征，多标签文本分类模型通过设计不同的网络架构对文本编码进一步操作以获取文本的向量表示$\boldsymbol{v}_s \in \mathbb{R}^m$，计算如式(2)所示。

$$\boldsymbol{v}_s = \text{Net}(\boldsymbol{H}) \tag{2}$$

其中$\boldsymbol{v}_s$为编码层得到的文本向量用于下一步编码，m为向量维度，Net表示不同分类模型抽象为统一的网络架构。

3.2.2 解码

解码阶段我们将编码层获得的文本向量$\boldsymbol{v}_s$通过标签分类层得到最终的标签表示向量$\boldsymbol{p}_i \in \mathbb{R}^l$,计算如式(3)所示。

$$\boldsymbol{p}_i = \text{sigmoid}\left(\boldsymbol{W}_\text{i} \cdot \boldsymbol{v}_\text{s} + \text{b}_\text{i}\right) \tag{3}$$

其中，l为标签总数，sigmoid为激活函数，$\boldsymbol{W}_i \in \mathbb{R}^{l \times m}$为可学习权重矩阵，$b_i$为偏置。标签表示向量$\boldsymbol{p}_i$的第$n$个数值表示第$n$个标签与文本相关的概率，如果该概率大于我们设定的阈值θ,则判定该标签与文本相关。

3.2.3 训练

在多标签分类任务中，通常使用二元交叉熵损失函数计算损失，计算如式(4)所示。

$$L_{BCE} = \begin{cases} -\log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -\log\left(1 - p_i^k\right) & \text{otherwise} \end{cases} \tag{4}$$

然而该交叉熵损失对每个标签的计算有着相同的权重，当标签分布不平衡时优化效果降低。而本文实验对比的部分多标签分类模型通过设计不同的损失函数缓解标签分布不平衡问题，例如Cui et al. (2019)设计Class-balanced focal loss(CBLoss)作为损失函数，计算如式(5-6)所示。

$$r_{CB} = \frac{1 - \epsilon}{1 - \epsilon^{freq}} \tag{5}$$

$$L_{CB} = \begin{cases} -r_{CB}\left(1 - p_i^k\right)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -r_{CB}\left(p_i^k\right)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases} \tag{6}$$

其中$\epsilon \in [0, 1)$为人为设置的超参数，$freq$为训练集中每个标签对应的频率，p_i^k为上一步得到的标签向量$\boldsymbol{p}_i$的第k个值，$\gamma \geq 0$为可调控的聚焦参数。本文对比的其它模型的不同损失函数公式详见附录。

4 本文方法

本节详细介绍了本文所提出的基于局部标注的自监督框架(PST)，如图1所示。首先介绍PST框架的整体流程，然后描述PST框架中最重要的伪标签生成与选择算法，最后提出联合训练来优化教师模型。

4.1 PST框架

标准的Self-Training框架利用在人工标注数据集上训练得到的教师模型自动地对大量无标注样本进行标注得到伪标签，最后将这些伪标签样本与人工标注数据集混合起来训练得到最终模型。具体流程如图1中红色框线内所示：(1)使用人工标注数据集来训练教师模型；(2)使用教师模型对无标注数据集进行伪标签预测；(3)通过预先定义的阈值将预测的标签分为相关标签与不相关标签；(4)将含有相关标签的样本与人工标注数据集混合起来训练更新教师模型；(5)重复2至4步直到教师模型的性能不再提高或满足停止条件。

然而在本文提到的不完全标注应用场景中，人工标注数据可能存在缺失标签，教师模型可能会受缺失标签的负面影响给无标注数据打上错误标签。因此本文提出的PST框架对标准Self-Training框架的(2)(3)两步进行了补充与修正。如图1所示，在第二步利用教师模型对无标注数据标签预测时，同时对人工标注数据进行补充预测，以补充人工标注所遗漏的缺失标签。为了能够同时预测这两部分数据，我们为每一个样本的每一个标签定义了一个状态标记(详见节4.2)，防止教师模型在预测标签时破坏原有的人工标注标签。在第三步通过阈值确定伪标签时，我们利用双阈值策略将标签分为三类：正例标签、负例标签和其它标签(详见节4.2)。同时根据预测标签类别反向更新该样本对应的标签状态集合。双阈值的设定一方面能够缓解教师模型因缺失标签影响而错误标签预测的数量，另一方面通过双阈值获取的其它标签在训练时不对模型产生影响，进而缓解缺失标签导致的错误信息影响模型性能。

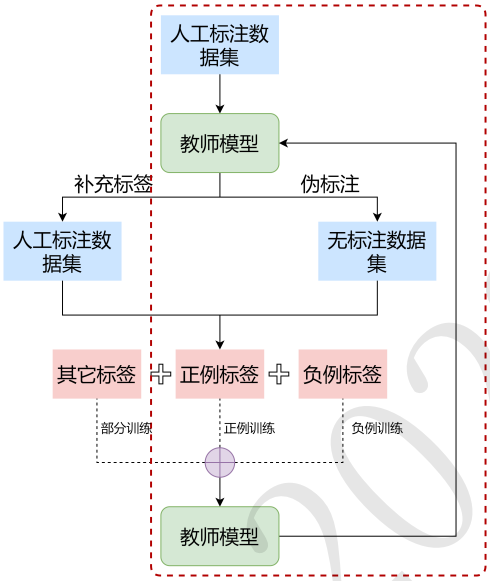


Figure 1: PST框架结构图

4.2 伪标签生成与选择

算法1是本文所提的伪标签选择算法，该算法利用双阈值策略和全局标签状态集合，通过教师模型的标签预测对缺失标签进行补充利用。

具体的，我们定义了一个全局标签状态集合State $= \{-2, -1, 0, 1, 2\}^{N \times l}$，其中$N$表示人工标注数据和无标注数据的总和，$l$表示标签的数量。全局标签状态集合State用来表示每个样本当前对应的每个标签的状态，该状态用于辅助后续标签的选择，其中-2到2表示标签的状态标记，2代表确定正例，即确定该标签与文本相关，且教师模型不再对该标签预测得分，并直接选择该状态的标签当作正例相关标签；-2代表确定负例，即确定该标签与文本不相关，且教师模型不再对该标签预测得分，并直接选择该状态的标签当作负例不相关标签；$-1, 0, 1$代表中间状态，处于该状态的标签暂未确定是否与文本相关，PST框架会利用教师模型对这些状态的标签预测打分，并根据得分和预定义的阈值来动态地改变标签的状态。首先我们初始化一个全局标签状态集合：把人工标注样本的相关标签状态初始化为2，其它所有标签的状态初始化为0(算法1第1行)。接着，我们利用教师模型对所有数据进行标签预测(算法1第3行)。之后，利用人为定义的正例阈值T_{Pos}和负例阈值T_{Neg}与模型预测每个标签的得分进行比较来将标签分为正例、负例或其它三类，同时更新标签状态集合(算法1第13-19行)，其中正例阈值T_{Pos}和负例阈值T_{Neg}的取值选择是根据多组取值组合实验中最佳实验结果确定(详见节5.4.3)。当某样本的某个标签状态为2时，即模型连续两次给予该标签高分，则该标签在后续Self-Training中无需模型预测打分；若某样本的某个标签状态为-2，即教师模型连续两次给予该标签低分，则该标签无需模型预测打分(算法1第7-12行)。最后通过筛选返回用于下一轮训练更新教师模型的数据(算法1第21-25行)。附录中的案例分析详细展示分析了样例标签在PST过程中的变化。

第二十二届中国计算语言学大会论文集，第17页-第30页，哈尔滨，中国，2023年8月3日至5日。

Algorithm 1 伪标签生成与选择算法

Input: 教师模型自动对人工标注数据集和无标准数据集进行打分，得到评分数据$D_{auto} = \{x_i,\ p_i,\ Y\}_{i=1}^{N}$，其中$p_i$为模型预测文本$x_i$的标签向量表示，$Y$为预定义的标签集合，$N$为训练集和无标注数据集的总和数。正例阈值$T_{Pos}$和负例阈值$T_{Neg}$.

Output: 挑选出下一轮训练的标注数据.

初始化全局变量State $= \{-2, -1, 0, 1, 2\}^{N \times l}$ ▷ *详见节4.2.*
$D_{train} = [\ ]$
for $(x_i,\ p_i) \in D_{auto}$ **do**
    $S_i =$ State[i];
    $y_i = \{0, 1, 2\}^{l}$;
    for $\left(p_i^k, y_i^k,\ S_i^k\right) \in p_i, y_i, S_i$ **do**
        if $S_i^k == 2$ **then**
            $y_i^k = 1$; **Continue**; ▷ *状态标记为2，无需改动，直接添加正例*
        end if
        if $S_i^k == -2$ **then**
            $y_i^k = 0$; **Continue**; ▷ *状态标记为-2，无需改动，直接添加负例*
        end if
        if $p_i^k > T_{Pos}$ **then**
            $S_i^k = S_i^k + 1; \quad y_i^k = 1;$ ▷ *本轮标签预测为正例，标签状态加一*
        else if $p_i^k < T_{Neg}$ **then**
            $S_i^k = S_i^k - 1; \quad y_i^k = 0;$ ▷ *本轮标签预测为负例，标签状态减一*
        else if $T_{Pos} > p_i^k > T_{Neg}$ **then**
            $y_i^k = 2;$ ▷ *本轮标签预测为其它，标签状态不变*
        end if
    end for
    if y_i not all 0 **then**
        $(x_i, y_i) \rightarrow D_{train}$;
    end if
end for
Return: D_{train}

4.3 联合训练

通过伪标签选择算法获取的新数据将直接用于训练更新教师模型。不同于第一阶段直接利用人工标注数据集训练教师模型，由于双阈值的设定，新数据集中添加了其它标签的额外信息。为此PST框架通过修改教师模型的损失函数以引入其它标签的额外信息，削弱错误信息对模型的误导。同样的这里以CBLoss(Cui et al., 2019)为例对式(6)进行修改，计算如式(7)所示，其中符号变量与式(6)相同。本文用到的其它损失函数修正后引入其他标签信息的计算公式详见附录。

$$L_{CB-Part} = \begin{cases} -r_{CB}\left(1 - p_i^k\right)^{\gamma} \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -r_{CB}\left(p_i^k\right)^{\gamma} \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases} \tag{7}$$

5 实验与分析

在本节中，我们分别在合成数据集和真实数据集上进行实验，以证明所提出方法的有效性与通用性。本节首先介绍两个数据集和实验的评估策略。然后，简要描述实验的相关设置和用于对比的基线模型。最后，我们列出不同数据下的实验结果并进行相应的分析。

5.1 实验数据

5.1.1 合成数据集

本文采用多标签文本分类任务中常见的英文数据集AAPD(Yang et al., 2018)作为合成数据，该数据集是由网络上收集的55,840 篇论文的摘要和相应学科类别组成。一篇学术论文属于一个或者多个学科，总共由54个学科组成，目的是根据给定的摘要来预测学术论文相对应的学科。为了模拟不完全标注的数据集，我们在标注规范的AAPD数据集上，对训练集按照不同的缺失比例p来随机的删除一些标签。同时为了更好的评估模型的性能，我们并未对验证集和测试集进行随机删除操作。

为了更全面地分析不同场景下的不完全标注问题，我们采用两种不同的方案来人为删除标签构造不完全标注的合成数据集。方案一：我们确保删除标签后的数据集每个样本至少仍保留一个相关标签。我们首先统计平均每个样本包含的标签数为2.41，并进一步确定该方案下标签缺失的上限概率为$(2.41-1.0)\div 2.41=0.585$，因此我们按照0.1、0.2、0.3、0.4、0.5、0.585的概率来按标签分布等比例删减相关标签并且始终保证每个样本至少一个标签；方案二：对于标注数据集中一条未标注标签的样本，我们假设该样本并不是无标注数据，而是标注为与所有标签都不相关的“其它”标签，因此该样本被当作完全负例加入模型训练。按照这种假设，我们采取0.1至0.9的概率随机丢失样本标签，当概率较高时样本的所有标签可能都会丢失，对于此类数据我们并未抛弃仍当作负例训练教师模型。我们按照上述两种方案人为构造了共15组训练数据(方案一6组，方案二9组)，加上原始完整的数据共构建了16组合成数据进行相关实验。

5.1.2 真实数据集

我们对CCKS2022 Task8 [1]面向金融领域的Few-Shot事件主体抽取学术评测提供的数据集进行人为修正，将其构建为多标签文本分类任务的中文数据集CCKS-IMLTC作为真实场景下不完全标注的数据集。

CCKS-IMLTC数据集具体的构建流程如下：(1)原CCKS数据为事件主体抽取数据集，每条样本由一段文本和该文本包含的一个或多个事件类型与事件主体对儿构成，但其标注质量不高，存在比较严重的事件类型缺失问题。因此我们删除原数据集中的事件主体只保留事件类型，同时删去了部分数据量极少的类别，将其修正为多标签文本分类任务的中文数据集，并按8:1:1的比例将数据切分为训练集、验证集和测试集。(2)为了更加精确地评估模型性能，我们对划分后的验证集和测试集进行人工补全，并对每个补充的标签打上标记便于后续统计每个标签的缺失情况。(3)对比补充前后的测试集与验证集，估算出整体标签的平均缺失比例为9.2%。虽然整体数据集上的缺失比例不高，但在我们对测试集和验证集进行缺失标签补充的过程中发现，有一部分关联性与共现性比较强的标签的缺失率可达60%左右，据我们人工统计，共有12个标签缺失率较高，我们将这些标签当作一个集合称为Few，并在实验中单独对这些标签进行评价。

数据集	标签数	训练集	验证集	测试集	无标注	平均标签数
AAPD	54	26,920	1,000	1,000	26,920	2.41
CCKS-IMLTC	96	40,000	5,000	5,000	43,147	1.21

Table 2: AAPD与CCKS-IMLTC数据集规模

表(2)记录了两个数据集的规模大小和标签数量。其中表格中无标注一列表示Self-Training过程中无标注文本数量。CCKS-IMLTC数据对应的无标注文本主要来自CCKS2021和CCKS2020中相关任务的纯文本数据。AAPD数据共55,840条标注样本，首先按标签分布等比例各拿出1,000条当作测试集与验证集，接着把剩下的标注样本一半当作训练集，另一半去除掉标注标签当作无标注数据。

5.2 评价指标

本文用三种指标全面评价模型性能。第一种指标为精确率(Precision, P)、召回率(Recall, R)和F1值(F1-measure, F1)，该指标常用于分类任务的模型性能的直观评估。后两种评价指标

[1]https://www.biendata.net/competition/ccks2022_eventext/data/

分别为退化率α_p和误导率β_p，计算如式(8-9)所示。该指标由负采样NER(Li et al., 2020)提出，用于评估分析缺失实体对模型的退化和误导两方面的影响。本文将该指标从NER任务中迁移到多标签文本分类中，以评估缺失标签对模型的影响，同时验证PST框架对两方面影响的削弱效果。

$$\alpha_p = \frac{f_0^a - f_p^a}{f_0^a} \tag{8}$$

$$\beta_p = \frac{f_p^a - f_p}{f_p^a} \tag{9}$$

其中f_p和f_p^a分别表示在缺失概率p数据上是否经过调整缺失标签损失训练后模型分类的F1值。具体地，我们通过将人工构建的合成数据中的缺失标签当作其它标签不计算其loss来调整缺失标签损失。而在真实数据集中，我们并不知道具体的缺失标签，因此实验部分我们只将该指标作用到合成数据AAPD上。f_0^a表示在原始完整数据上训练模型的分类F1值。从公式(8)可以发现，退化率α_p指标消除了缺失标签对loss的影响，主要是为了评价正例标签数量减少导致分类器训练不充分的退化影响。从公式(9)可以发现，误导率β_p指标计算相同缺失概率下是否消除缺失标签作为负例对loss的影响，从而进一步评价不完全标注数据对分类器的误导影响。

5.3 实验设置与对比模型

我们将PST框架运用到多种常见的多标签文本分类模型上进行实验，相关的模型如下：1)BERT-CLS(CLS)(Devlin et al., 2018):首先利用预训练语言模型将文档编码到向量空间，然后独立地输出每个标签的概率。2)BERT-TextCNN(TextCNN)(Kim, 2014):将BERT作为文本编码结合CNN卷积提取单词信息。3)LSAN(Xiao et al., 2019):借助注意力机制获得特定标签的文本表示。4)Focal loss(FL)(Lin et al., 2017):一种简单但广泛使用的加权损失分类策略。5)Rebalanced focal loss(R-FL)(Wu et al., 2020):重平衡加权与focal loss的组合。6)Class-balanced loss(CB)(Cui et al., 2019):以每个类的有效数量为指导的按类重新加权的损失函数。7)Distribution-balance loss(DB)(Wu et al., 2020):利用负容忍正则化缓解模型对负类的过度抑制而造成的分类边界偏移。8)HTTN(Xiao et al., 2021):利用元学习将头部标签与尾部标签联系起来，以缓解长尾分布中尾部标签数据量少的问题，是一种基于迁移学习的多标签分类器。9)LACO(Zhang et al., 2021):通过引入多任务学习方法来增强标签相关性反馈，并利用联合编码机制同时获得文本和标签的向量表示。10)FLEM(Zhao et al., 2022):集成标签增强与模型训练，从而有效地挖掘不同标签的隐含相对重要性信息。11)BERT-TextCNN+CBLoss(TextCNN-CB)，作为PST框架的基础教师模型，用于合成数据集上补充分析实验。

我们使用开源的预训练模型bert-base-cased[2]和bert-base-chinese[3]分别作为英文和中文的编码层，通过教师模型在验证集的性能，设定超参。具体地，我们设置最大句长为256，学习率为2e-5，batch-size为16，教师模型的训练轮次为20，self-training的轮次为10，随机种子为1227，dropout为0.5，正例阈值为0.6，负例阈值为0.4，线性层维度大小为300，FLEM模型中α和β均为0.01，HTTN模型中头部标签数量为84，TextCNN中滤波器为200、窗口大小为$[1, 3, 5, 7]$，LSTM隐层大小为256，其它模型超参均遵循其模型原论文中的设置。

5.4 实验结果与分析

5.4.1 真实数据集上的实验

表3展示了在CCKS-IMLTC数据集上不同系统的实验结果，其中Teacher表示采用标准有监督训练，Self-Training表示采用标准Self-Training框架，PST(Ours)表示采用本文所提PST框架。横向对比分类模型在不同框架下的结果发现，标准Self-Training框架对某些模型的性能有所提升，然而这种提升并不稳定，而本文提出的PST框架对所有教师模型都有较高的性能提升，充分证明了PST框架的有效性与通用性。对比不同框架下的性能发现，标准Self-Training框架通过优化模型的准确率(P值)提升性能，相反PST框架主要优化模型的召回率(R值)以提升整体性能。并且相较于Self-Training，PST框架对教师模型性能的提升更加稳定且有效。进一步分

[2]https://huggingface.co/bert-base-cased
[3]https://huggingface.co/bert-base-chinese

析实验结果发现，教师模型与传统的Self-Training框架训练中大都将缺失标签当作负例标签训练，因此其预测结果大都是训练中见过的正例标签，进而导致预测出的正例标签准确率较高即整体的P值较高。而PST框架通过双阈值策略和全局标签状态补充利用缺失的正例标签，进而使得可以预测出更多的正例标签，即使得模型整体的R值升高。

图2展示了不同模型在CCKS-IMLTC中标签缺失严重的Few标签集合上的实验结果，我们发现相较于表3整体标签上的实验结果，PST框架在缺失比例高的标签上对教师模型的提升更为明显。同时我们发现当教师模型性能过低时，标准Self-Training框架反而会给模型带来负优化，而PST框架对所有教师模型都有不同程度的提升，更加充分地证明了PST框架的通用性，可以兼容现有多种不同的多标签文本分类模型。进一步对比不同模型在整体标签和Few标签集合上的实验结果，我们发现相较于注重标签间关系的模型FLEM和LACO，针对标签分布设计的模型如CB，DB等随着标签缺失率升高性能下降的较为缓慢。

模型	Teacher			Self-Training			PST(Ours)				
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	Δ_T(F1)	Δ_{ST}(F1)
CLS	**79.91**	63.81	70.95	77.37	63.08	69.50	76.60	**67.25**	**71.62**	+0.67	+2.12
TextCNN	76.69	71.91	74.22	75.91	**73.87**	74.88	**78.06**	72.22	**75.03**	+0.81	+0.15
LSAN	75.51	59.89	66.80	**77.24**	60.89	68.10	75.34	**63.92**	**69.16**	+2.36	+1.06
FL	80.99	68.83	74.41	**82.14**	67.83	74.30	80.44	**70.38**	**75.07**	+0.66	+0.77
RFL	80.65	69.67	74.76	**81.97**	69.37	**75.15**	80.36	**70.58**	**75.15**	+0.39	+0.00
CB	81.16	70.28	75.33	**81.83**	69.39	75.10	80.51	**71.03**	**75.48**	+0.15	+0.38
DB	74.76	74.50	74.63	73.99	**76.66**	75.30	**77.32**	74.68	**75.97**	+1.34	+0.67
HTTN	81.46	67.81	74.01	**81.56**	68.38	74.39	80.79	**69.88**	**74.94**	+0.93	+0.55
LACO	78.19	**71.45**	74.65	78.85	70.46	74.42	**79.53**	70.68	**74.84**	+0.19	+0.42
FLEM	80.54	69.39	74.55	**83.91**	67.81	75.01	82.05	**69.57**	**75.30**	+0.75	+0.29
TextCNN-CB	76.56	74.35	75.44	77.00	73.53	75.22	**77.59**	**74.68**	**76.11**	+0.67	+0.89

Table 3: 不同模型在CCKS-IMLTC数据集上的实验结果

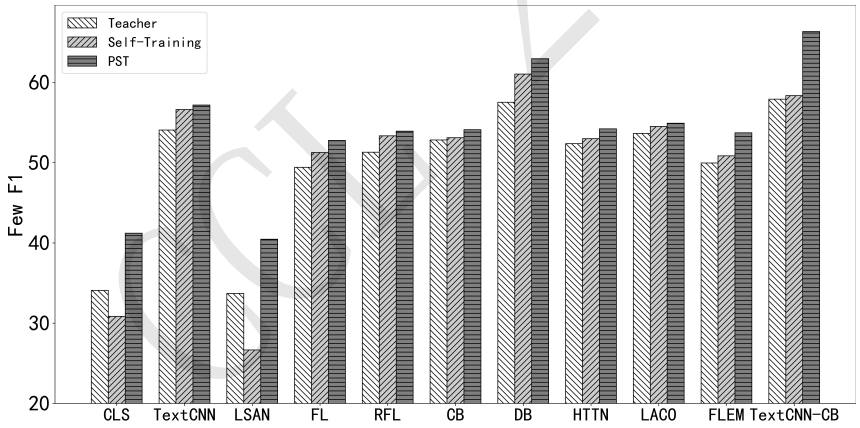
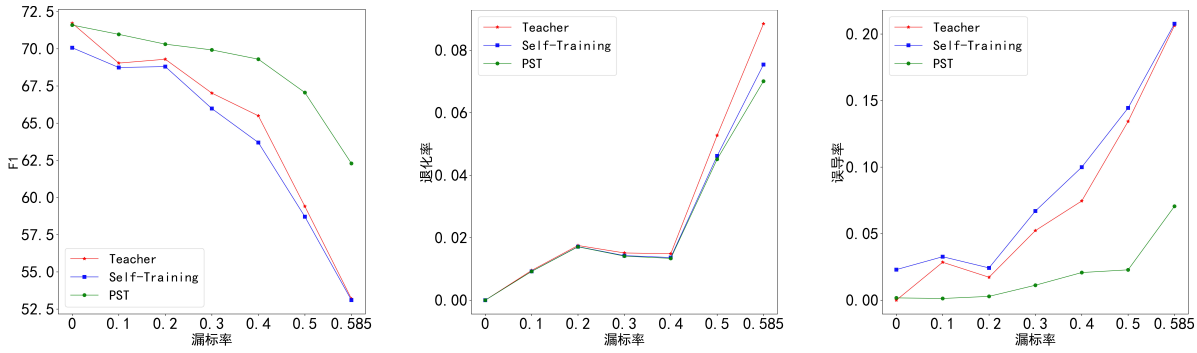


Figure 2: 不同模型在CCKS-IMLTC上Few标签的实验结果

5.4.2 合成数据集上的实验

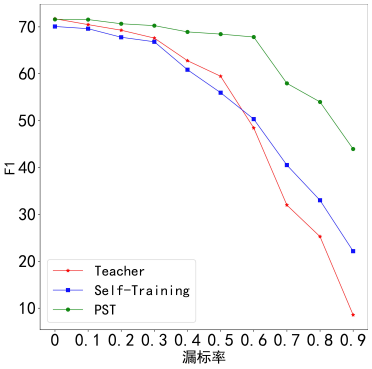
图3展示了两种不同标签缺失方案(详见节5.1.1)下TextCNN-CB在人工合成数据集AAPD上的实验结果。图3(a)与图3(d)为不同框架下F1值随标签缺失比例变化的折线图，可以发现随着缺失率的增加所有框架下模型的性能都会有所下降，但相较于教师模型和标准Self-Training框架，本文提出的PST框架下降得更加缓慢，即PST框架可以有效缓解缺失标签对模型的负面影响进而提升模型性能，并且标签缺失问题越严重提升效果越明显。同时我们发现传统的Self-Training框架在两种不同标签缺失方案下性能基本都不如教师模型，说明教师模型受到缺失标签的影响在对训练集打分时可能会导致标注错误标签进而影响模型下一轮训练，而我们提出的PST半监督学习框架可以通过双阈值策略和全局标签状态来缓解这种错误标签的出现，同时提升教师模型补充缺失标签的能力。



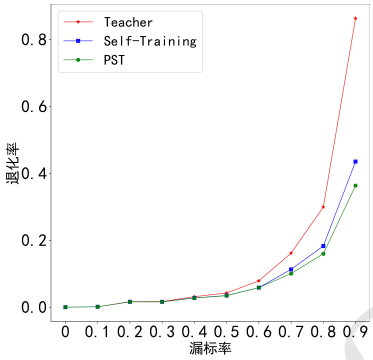
(a) 方案一F1曲线图

(b) 方案一退化率曲线图

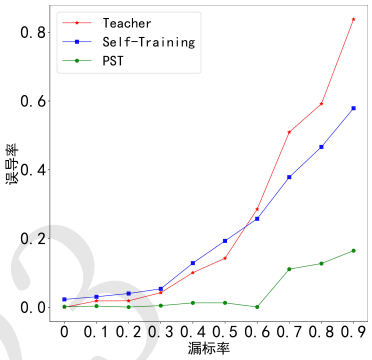
(c) 方案一误导率曲线图



(d) 方案二F1曲线图



(e) 方案二退化率曲线图



(f) 方案二误导率曲线图

Figure 3: 不同标签缺失方案下人工合成数据AAPD上的实验结果

图3(b)与图3(e)为不同框架下退化率随标签缺失比例变化的折线图，可以发现标签缺失比例不足50%时不同框架下缺失标签带来的退化影响都很小最高仅为5%。当缺失率达到70%时，因缺失标签带来的退化影响陡增，而PST框架可以明显地缓解，尤其是缺失率为90%时PST框架可以将退化率从90%降至40%。图3(c)与图3(f)为不同框架下误导率随标签缺失比例变化的折线图，可以发现随着标签缺失率的增加越来越多的缺失标签被当作负例学习误导模型，而PST框架将可能产生误导的缺失标签转为其它标签忽略其损失，进而有效的减弱模型被误导的概率。总的来说，实验结果表明PST框架可以缓解缺失标签对模型误导和退化两方面的影响。

5.4.3 不同正负例阈值选择对PST框架的影响

图4展示了PST框架中正负例阈值不同组合时TextCNN-CB在CCKS-IMLTC数据上的实验结果热力图，颜色越暗模型的F1值越高，相反颜色越淡模型的F1值越低，我们发现PST框架的正例阈值设为0.6，负例阈值设为0.4时模型性能最优。

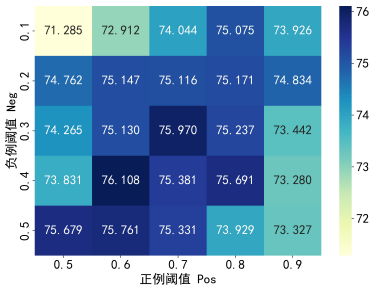


Figure 4: 不同正负例阈值组合在CCKS-IMLTC数据上的实验结果

5.4.4 消融实验

为了进一步分析PST框架中各个组件对整个框架的影响，我们在CCKS-IMLTC上进行了消

融实验，实验结果如表4所示，其中Few和All分别指的是标签缺失严重的Few标签集合上的结果与整体全部标签上的结果，右下角正负数表示较于教师模型性能的变化。首先我们发现去除外部无标注数据只在人工标注数据集上使用PST框架，教师模型的性能仍有明显提升，侧面说明PST框架的确可以对不完全标注数据补充部分缺失标签；接着我们去掉双阈值策略，只保留正例阈值整体框架退化为标准Self-Training，此时整体性能有所降低，但Few标签上的性能仍有微弱提升；最后保留双阈值去掉其它标签，此时Few标签性能提升1.36%但较完整PST框架仍有很大差距。总的来说，PST框架可以无需外部无标注数据辅助直接在不完全标注的数据上提升教师模型性能，同时PST框架中的双阈值策略，其他标签设置等各组件都是整体框架中不可或缺的部分。

	Few(F1%)	**All(F1%)**
TextCNN-CB	57.90	75.44
TextCNN-CB w/PST	$66.33_{+8.43}$	$76.11_{+0.67}$
-Unlabel data	$63.99_{+6.09}$	$75.82_{+0.38}$
-负例阈值T_{Neg}	$58.35_{+0.45}$	$75.22_{-0.22}$
-Partial label	$59.26_{+1.36}$	$75.37_{-0.07}$

Table 4: TextCNN-CB在CCKS-IMLTC上的消融实验

6 总结与展望

本文提出了一种全新的基于局部标注的自监督框架(PST)，以缓解不完全标注问题在多标签文本分类中的影响。该框架是一种模型无关的插件式框架，可以兼容多种不同的教师模型。充分利用外部无标注数据来优化教师模型的同时，对不完全标注数据的缺失标签补充利用，进而削弱了缺失标签给模型带来的影响。实验结果表明，我们提出的框架具有通用性，并且能一定程度缓解数据不完全标注问题带来的影响。

我们发现教师模型的选择也影响着我们框架的上限，效果更好的教师模型通过PST框架可以更好的补充缺失标签，从而更大程度地缓解不完全标注的问题。因此如何设计更优雅、高效的教师模型，提升该任务的整体性能，是我们今后的研究方向。

参考文献

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

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

John Dougrez-Lewis, Maria Liakata, Elena Kochkina, and Yulan He. 2021. Learning disentangled latent topics for twitter rumour veracity classification. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 3902–3908.

Cunxiao Du, Zhaozheng Chen, Fuli Feng, Lei Zhu, Tian Gan, and Liqiang Nie. 2019. Explicit interaction model towards text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6359–6366.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 935–944.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2840–2850.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

John Langton, Krishna Srihasam, and Junlin Jiang. 2020. Comparison of machine learning methods for multi-label classification of nursing education and licensure exam questions. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 85–93.

Xin Li, Haoran Xie, Yanghui Rao, Yanjia Chen, Xuebo Liu, Huan Huang, and Fu Lee Wang. 2016. Weighted multi-label classification model for sentiment analysis of online news. In *2016 International conference on big data and smart computing (bigcomp)*, pages 215–222. IEEE.

Yangming Li, Lemao Liu, and Shuming Shi. 2020. Empirical analysis of unlabeled entity problem in named entity recognition.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.

Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3855–3864.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. *Advances in neural information processing systems*, 30.

Nikolaos Pappas and James Henderson. 2019. Gile: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 629–640.

Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Isaac Triguero, Salvador García, and Francisco Herrera. 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42:245–284.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 466–475.

Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. 2021. Does head label help for long-tailed multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14103–14111.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.

Junjie Yu, Xing Wang, Jiangjiang Zhao, Chunjie Yang, and Wenliang Chen. 2022. Stad: Self-training with ambiguous data for low-resource relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2044–2054.

Ximing Zhang, Qian-Wen Zhang, Zhao Yan, Ruifang Liu, and Yunbo Cao. 2021. Enhancing label correlation feedback in multi-label text classification via multi-task learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1190–1200.

Xingyu Zhao, Yuexuan An, Ning Xu, and Xin Geng. 2022. Fusion label enhancement for multi-label learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*.

肖琳, 陈博理, 黄鑫, 刘华锋, 景丽萍, and 于剑. 2020. 基于标签语义注意力的多标签文本分类. 软件学报, 31(4):1079–1089.

附录.不同损失函数经过其它标签修正前后的计算公式

- **BCE loss**

$$L_{BCE} = \begin{cases} -\log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -\log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

$$L_{BCE}^{Partial} = \begin{cases} -\log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -\log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

- **Focal loss(FL)**: γ为超参

$$L_{FL} = \begin{cases} -(1 - p_i^k)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -(p_i^k)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

$$L_{FL}^{Partial} = \begin{cases} -(1 - p_i^k)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -(p_i^k)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

- **Rebalanced-Focal loss(RFL)**: $\widehat{r}_{DB}$详见(Wu et al., 2020)

$$L_{RFL} = \begin{cases} -\widehat{r}_{DB}(1 - p_i^k)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -\widehat{r}_{DB}(p_i^k)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

$$L_{RFL}^{Partial} = \begin{cases} -\widehat{r}_{DB}(1 - p_i^k)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -\widehat{r}_{DB}(p_i^k)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

- **Class-balanced focal loss(CB)**: ϵ, γ为超参

$$r_{CB} = \frac{1 - \epsilon}{1 - \epsilon^{freq}}$$

$$L_{CB} = \begin{cases} -r_{CB}(1 - p_i^k)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ -r_{CB}(p_i^k)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

$$L_{CB}^{Partial} = \begin{cases} -r_{CB}(1 - p_i^k)^\gamma \log\left(p_i^k\right) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -r_{CB}(p_i^k)^\gamma \log\left(1 - p_i^k\right) & \text{otherwise} \end{cases}$$

- **Distribution-balanced loss(DB)**: λ为超参 $\widehat{r}_{DB}, q_i^k$详见(Wu et al., 2020)

$$L_{DB} = \begin{cases} -\widehat{r}_{DB}(1-q_i^k)^{\gamma}\log\left(q_i^k\right) & \text{if } y_i^k = 1 \\ -\widehat{r}_{DB}\frac{1}{\lambda}(q_i^k)^{\gamma}\log\left(1-q_i^k\right) & \text{otherwise} \end{cases}$$

$$L_{DB}^{Partial} = \begin{cases} -\widehat{r}_{DB}(1-q_i^k)^{\gamma}\log\left(q_i^k\right) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -\widehat{r}_{DB}\frac{1}{\lambda}(q_i^k)^{\gamma}\log\left(1-q_i^k\right) & \text{otherwise} \end{cases}$$

附录.案例分析

文本：涉案的美国三大投行遭到重罚,花旗集团和摩根大通因涉嫌财务欺诈被判有罪，向安然公司的破产受害者分别支付了20亿、22亿和6900万美元的赔偿罚款。					
文本的相关标签：重大赔付，财务造假，破产清算 标注员标注标签：重大赔付，财务造假 部分不相关标签：债务违约，股东减持					
正例阈值T_{Pos}：0.6，负例阈值T_{Neg}：0.4 PST过程中样本标签状态变化：					
初始状态值：	重大赔付_2	财务造假_2	破产清算_0	债务违约_0	股东减持_0
epoch1打分：	/	/	0.53	0.47	0.23
epoch1状态：	重大赔付_2	财务造假_2	破产清算_0	债务违约_0	股东减持_-1
epoch2打分：	/	/	0.65	0.44	0.16
epoch2状态：	重大赔付_2	财务造假_2	破产清算_1	债务违约_0	股东减持_-2
epoch3打分：	/	/	0.71	0.46	/
epoch3状态：	重大赔付_2	财务造假_2	破产清算_2	债务违约_0	股东减持_-2
⋮					
epoch10打分：	/	/	/	/	/
epoch10状态：	重大赔付_2	财务造假_2	破产清算_2	债务违约_-2	股东减持_-2

Table 5: 一条不完全标注样例的标签状态在PST过程中的变化。标签数量过多这里只选取了两个不相关标签演示，其中绿色标签作为下一轮次训练的正例标签，红色标签作为下一轮次训练的负例标签，黑色标签作为其它标签不用来下一轮模型训练。

融合汉越关联关系的多语言事件观点对象识别方法

李格格[1,2]，**郭军军**[1,2]，**余正涛***[1,2]，**相艳**[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1303717217@qq.com, guojjgb@163.com
ztyu@hotmail.com, sharonxiang@126.com

摘要

越南语观点对象识别是越南语事件观点分析的重要研究内容。由于汉越两种语言的语法结构上存在差异，使得多语言事件关联复杂，观点对象表征困难。现有研究方法仅能实现汉越双语的表征，未能有效捕获并利用汉越双语事件中要素的关联关系。因此，本文提出一种融合汉越关联关系的多语言事件观点对象识别方法，利用中文和越南语事件间的要素共现和整体语义关联构建汉越多语言事件表征网络，使用多语言预训练语言模型获得要素节点的特征向量，利用图卷积网络对节点信息进行聚合，得到同一语义空间下汉越双语的公共表征，实现汉越事件观点对象的识别。实验结果表明本文模型能够更有效地构建多语言关联信息，其F1值较多个基线模型都有明显提高。

关键词： 观点对象识别；多语言事件关联；图卷积网络

A Multilingual Event Opinion Target Recognition Method Incorporating Chinese and Vietnamese Association Relations

Gege Li[1,2]，**Junjun Guo**[1,2]，**Zhengtao Yu***[1,2]，**Yan Xiang**[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
1303717217@qq.com, guojjgb@163.com
ztyu@hotmail.com, sharonxiang@126.com

Abstract

Vietnamese opinion target recognition is an important research component of Vietnamese event opinion analysis. Due to the differences between the grammatical structure of Chinese and Vietnamese, it makes the association of multilingual events complex and the representation of opinion target difficult. Existing research methods can only realize the bilingual representation of Chinese and Vietnamese, and fail to effectively capture and utilize the association relationship of elements in Chinese and Vietnamese events. Therefore, this paper proposes a multilingual event opinion target recognition method that integrates Chinese and Vietnamese association relations, using element co-occurrence and overall semantic association between Chinese and Vietnamese events to build a network of Chinese and Vietnamese multilingual event representation, using a multilingual pre-trained language model to obtain the feature vectors of element nodes, and using graph convolutional network to aggregate node information to obtain a common representation of Chinese and Vietnamese in the same semantic space. In

*余正涛（通讯作者）：ztyu@hotmail.com
基金项目：国家自然科学基金（U21B2027，61972186，62266027，62266028）；云南省科技重大专项（202302AD080003，202103AA080015）；云南省基础研究计划项目（202301AS070047，202301AT070444）

order to achieve the recognition of Chinese and Vietnamese opinion target recognition. The experimental results show that the model in this paper can construct multilingual association information more effectively, and its F1 values are significantly improved compared with several baseline models.

Keywords: Opinion target recognition , Multilingual event correlation , Graph convolutional network

1 引言

互联网的快速发展推动了中越两国交流，从社交媒体评论文本中挖掘两国用户的观点，掌握用户对事件的关注，对处理好与越南的国际关系、区域经济发展和文化交流有着重要的作用，同时为政府及企业正确把握汉越舆情动态并及时做出应对措施提供有效保障。越南语标注数据资源的稀缺，阻碍了其观点对象识别方法的研究，可通过多语言观点对象识别（Multilingual Opinion Target Recognition）的方法，利用具有丰富标记数据的中文通过知识迁移帮助标注资源稀缺的越南语实现多语言观点对象识别任务。

Figure 1: “新冠疫情”数据集上的汉越社交媒体评论样例

如图1所示的是汉越“新冠疫情”数据集中有关“新冠疫苗”和“防疫措施”两种不同观点对象的评论句。图中左半部分描述的是汉越两国用户针对“新冠疫苗”观点对象发出的评论，右侧评论则是针对“防疫措施”的讨论，通过观察上图可以发现中文和越南语评论在针对同一事件时讨论的内容较为接近，关注的重点也较为相似，利用这种关联特征可以较好地捕获汉越双语评论的全局特征（汉越评论之间的关联关系）和局部特征（评论中关键词所携带的语义信息）。通过对关联关系和语义信息进行建模，能够得到信息互补的特征表示学习模型，从而较好地完成迁移任务，解决越南语标注资源稀缺的问题。

目前，在多语言观点对象识别的研究中，主要通过基于传统机器学习的方法和基于深度学习的方法进行观点对象识别，根据每个领域的评论表征来学习特定的观点对象分类器。基于传统机器学习的方法通过制定相关规则并融入领域相关信息等外部知识利用算法提升识别性能，基于深度学习的方法通过使用神经网络提取数据特征进行观点对象的识别。这些模型的训练过程需要大规模且高质量的标注数据集，但是在面对不同的应用场景时，构建这样规模的训练数据集成本较高，同时利用传统的特征编码模式只能考虑到单语语料库中各评论文本的局部特征，不能很好的做到多语言间的知识迁移。

针对以上问题，本文提出一种融合汉越关联关系的多语言事件观点对象识别方法，该方法通过将汉越社交媒体评论文本和其中的关键词(高频词)作为节点构建异构图，结合评论文本节点的输入表征，通过图卷积网络准确地捕获汉越双语评论间观点对象的关联信息，提高观点对象表征学习和模型识别性能。本文的主要贡献如下:

（1）在中文和越南语评论文本上利用异构图进行关联关系构建，通过构建多种类型的节点和边关系，捕捉各节点之间丰富的关系结构，得到汉越评论文本数据在同一嵌入空间下的对应关系。

（2）使用多语言预训练语言模型获取评论文本的特征向量，并将其作为评论文本节点的输入表征，使用图卷积网络学习节点特征并基于图结构迭代更新评论文本表征，进行汉越观点对象的识别。

（3）在所构建的汉越评论数据集上进行了实验，相比已有的基准模型，所提模型的性能都有较大的提升。

2 相关工作

观点对象是由带有情感偏见的情感词所修饰的对象，通常是社交网络中的一个特定主题，或者是电子商务平台上的一个特定产品或产品评论的一部分。观点对象识别是从预定义的标签集合中为评论文本分配对应的标签，观点对象识别策略可以分为以下两大类方法。

2.1 基于传统机器学习的方法

基于传统机器学习的方法主要是通过人工制定规则来分析语料或者通过融入领域相关信息等外部知识利用机器学习的算法提升识别性能，主要分为基于规则和统计的学习方法以及基于机器学习的方法。

2.1.1 基于规则和统计的学习方法

基于规则和统计的学习方法主要对语料库进行分析，结合分析制定词性规则、词序列规则和句法规则。倪茂树and 林鸿飞 (2007)提出了一种利用关联规则和极性分析方法挖掘观点特征的算法，从而更好的识别出商品评论中观点对象的类别。Qiu et al. (2011)用情感词识别观点对象的修饰关系和整体的从属关系词，取得了良好的实验结果。这些方法非常依赖规则和具体语言，难以覆盖所有情况导致只适合小规模的数据，而且系统移植性不强，根据任务的不同需要设定新的规则，建设周期长并且代价比较高昂。

2.1.2 基于机器学习的方法

基于机器学习的方法通过融入领域相关信息等外部知识利用机器学习的算法提升识别性能。Titov and McDonald (2008)采用多粒度的主题模型分析并识别文本中的观点对象，并在分析结果的基础上，归类出相同的观点对象，对相似度较大的观点对象进行聚类。Moghaddam and Ester (2011)利用狄利克雷分布（LDA）提取观点对象和相应的评级产品在线评论。Li et al. (2012)提出了一种新的关系自适应引导(RAP) 算法，通过利用标记的源域数据来获得主题词和观点对象之间的关系。Li et al. (2018)将每一个时间戳对应的观点对象特征与原始抽取的观点对象特征进行融合，另外利用观点对象识别过程中的坐标结构和抽取到的观点对象特征进行交互，通过探索到的两种信息提升观点对象识别模型的性能。基于机器学习的方法比起基于规则和统计的学习方法有一定的改进，但还是需要人工对文本特征进行标记，人为的主观因素会影响模型的性能，同时机器学习需要依赖先验知识的质量和大量的标记数据，执行的速度会比较慢，难以适应如今信息量爆炸的时代。

2.2 基于深度学习的方法

基于深度学习的方法通过训练神经网络，使用CNN、RNN和LSTM等各种典型的神经网络对观点对象的类别进行识别，Ding et al. (2017)提出使用规则在循环神经网络上生成辅助监督，以学习每个单词领域不变的隐式特征表示。Nguyen and Le Nguyen (2018)在SenTube数据集上提出了卷积N-gram BiLSTM词嵌入，用于进行多语言观点对象的识别。

这些模型会优先考虑文本的局部信息和顺序信息，能够很好的捕获连续词序列中的语义和句法信息，但是它们忽略了多语言的全局词共现，而全局词共现中携带了不连续以及长距离的

语义信息。随着深度学习的不断发展，近些年来图卷积网和多语言预训练语言模型被广泛应用到观点对象识别的任务当中。

2.2.1 基于图卷积网络的方法

图卷积神经网络可以通过在节点之间传递信息建立图模型，Yao et al. (2019)提出文本图卷积神经网络TextGCN，基于词共现和文档词关系建立单语言图，该方法可以同时学习单词和文档的嵌入。在多语言的任务中，Li et al. (2020)将图神经网络应用在元学习方面。Wang et al. (2021)提出CLHG模型，使用了类似于TextGCN方法的构图方式，将图神经网络用在多语言文本分类中，解决了原有模型只注重语义信息而忽略句法信息这一缺点。图卷积网络虽然擅长将图中的全局信息卷积成一个文本，但不能同时捕捉上下文的相关性与关联信息，剥离了文本中词与词之间的关联性，难以获得高效的性能。

2.2.2 基于多语言预训练语言模型的方法

在单语预训练语言模型的不断发展下，部分学者将焦点聚集在多语言预训练语言模型的训练中，并且训练出来的模型在许多下游任务中表现出优异的跨语言迁移能力。Kenton and Toutanova (2019)通过在104种语言的维基百科语料库上进行训练推出了mBert，在多语言迁移方面取得了不错的效果。Lample and Conneau (2019)提出的XLM模型，通过构造编码器编码多种语言的句子到同一嵌入空间来增强多语言的共享词汇。Conneau et al. (2020)在之前的基础上推出了多语言预训练语言模型XLM-R，增加了模型的语言数量和训练示例的数量。尽管多语言预训练语言模型已经取得了许多最新的成果，这些模型没有明确考虑语言之间的句法差异，导致目标语言的泛化性能下降，同时任务特定的结构依赖问题为模型性能的进一步提高带来了许多限制。

3 模型介绍

本文提出一种融合汉越关联关系的多语言事件观点对象识别方法，不仅关注汉越双语评论之间语义差异的问题，同时也关注汉越双语评论中观点对象之间的对齐关系，模型总体架构如图2所示，它主要包含四个网络：节点特征表示、多语言异构图的构建、节点特征学习和观点对象类别预测。

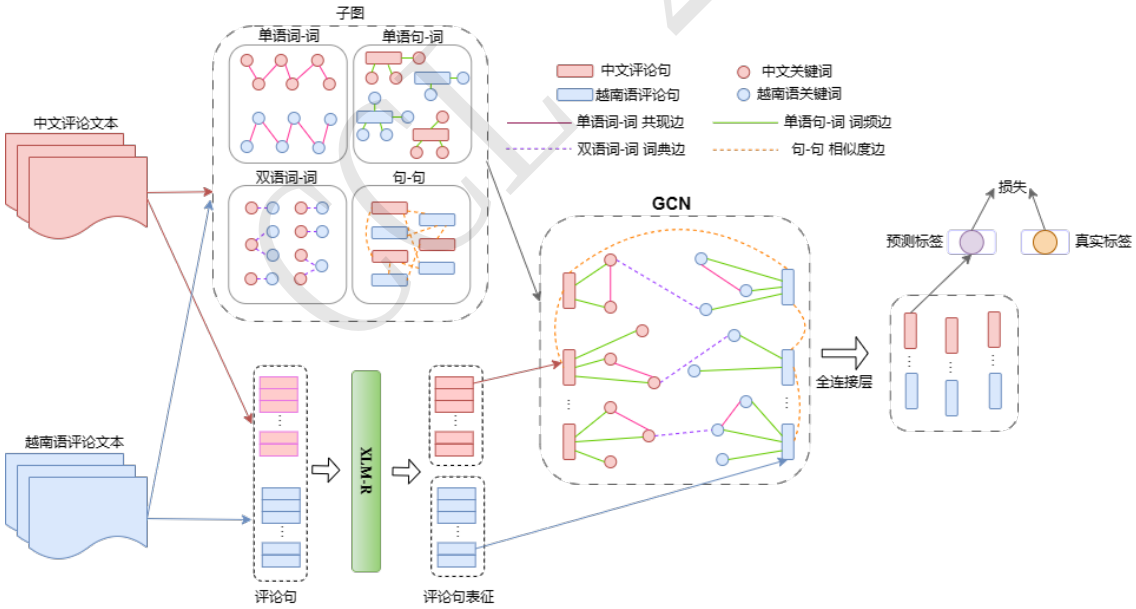


Figure 2: 融合汉越关联关系的多语言事件观点对象识别模型

使用多语言预训练语言模型获取汉越社交媒体评论文本的节点特征，将评论文本和其中的关键词作为异构图的节点，并基于评论文本中词共现、词对齐、词频信息和语义相似度的关系构边，利用图卷积网络对节点特征进行学习，并对节点进行线性转换输出评论文本节点的预测，并在训练期间与真实标签进行比较。

3.1 评论文本节点表征

类似于TextGCN(Yao et al., 2019)，论文中矩阵$\mathrm{X} = \mathrm{I}_{\mathrm{n_{doc}+n_{word}}}$被用作初始节点特征，其中$\mathrm{n_{doc}}$是文档节点的数量，$\mathrm{n_{word}}$是单词节点的数量（包括训练集和测试集），在本文中我们使用多语言预训练语言模型XLM-R来获得汉越双语评论文本的嵌入，并将它们作为异构图中评论文本节点的输入表示。

$$\mathrm{X} = \begin{pmatrix} X_{S_c} \\ X_{S_v} \\ 0 \\ 0 \end{pmatrix}_{(n_{S_c}+n_{S_v}+n_{K_c}+n_{K_v})\times \mathrm{d}} \tag{1}$$

其中，n_{S_c}、n_{S_v}、n_{K_c}、n_{K_v}分别表示中文评论文本数量、越南语评论文本数量、中文关键词数量和越南语关键词数量，中文评论文本节点和越南语评论文本节点嵌入由$\mathrm{X_{S_c}} \in \mathbb{R}^{n_{S_c}\times d}$和$\mathrm{X_{S_v}} \in \mathbb{R}^{n_{S_v}\times d}$表示，其中d是文本嵌入的维度。由于不考虑关键词节点的特征表示，因此将中文关键词和越南语关键词的嵌入置为0。

3.2 多语言异构图的构建

由于越南语标注数据资源稀缺，且以往所提出的表示学习模型仅学习到单语语料中的文本信息而忽略了同一事件下多语言观点对象之间的对齐关系。本文将汉越语料库中的各种实体和关系整合到一个异构图中，设计一种新的多语言表示学习模型，将语义信息和拓扑信息封装到一个低维联合嵌入的观点对象识别任务中，通过构建一个包含关键词节点和评论文本节点的异构图，节点数$n = (n_{S_c} + n_{S_v} + n_{K_c} + n_{K_v})$是中文和越南语评论文本数量和双语评论中关键词数量的总和，表1包含了异构图中节点和边的详细信息。

No.	节点	描述
1	S_c	中文评论句
2	S_v	越南语评论句
3	K_c	中文关键词
4	K_v	越南语关键词
No.	边	描述
1	$S_c \leftrightarrow K_c$	中文评论句中包含中文关键词
2	$S_v \leftrightarrow K_v$	越南语评论句中包含越南语关键词
3	$K_c \leftrightarrow K_v$	中文关键词与越南语关键词词典
4	$S_c \leftrightarrow S_v$	中文评论句与越南语评论句有较高的语义相似度

Table 1: 汉越多语言异构图的节点及边关系

使用汉越社交媒体评论文本数据集中的评论句和其中的关键词作为节点构建汉越多语言异构图，其中关键词之间、评论句和关键词以及评论句之间均有不同的关系种类，主要包括关键词之间的词共现和词对齐关系，评论句和关键词的词频关系，评论句之间的语义相似度关系。

关键词之间的词共现和词对齐关系：

为了更好的利用单语关键词的共现信息，通过基于词共现关系构建关键词节点之间的边，对语料库中所有的评论句使用一个固定大小的滑动窗口来收集词的共现信息，分别在中文和越南语评论文本上使用点互信息（PMI）计算两个关键词节点之间的权重，单语关键词对$\{i, j\}$的PMI值计算公式为：

$$PMI(i,j) = \log \frac{p(i,j)}{p(i)p(j)} \tag{2}$$

$$p(i,j) = \frac{\#W(i,j)}{\#W} \tag{3}$$

$$p(i) = \frac{\#W(i)}{\#W} \tag{4}$$

其中$\#W(i)$表示滑动窗口中包含关键词i的数量，$\#W(i,j)$是指滑动窗口中同时包含关键词i和j的数量，$\#W$是语料库中所有滑动窗口的数量。当PMI值为正时表示两个词之间的语

义相关性较高，而PMI值为负时表示两个词之间的语义相关性很少或没有，只在PMI值为正的关键词对之间添加边。

考虑挖掘汉越双语关键词之间的关系，基于双语词对齐构建关键词节点之间的边，对于汉越多语言观点对象识别的研究中，汉越双语关键词对相较于其他词对对模型预测性能产生的影响更大，利用汉越双语种子词典，匹配语义相似的双语关键词作为词节点并添加对齐的边关系，根据匹配出的双语关键词对进行多语言词级对齐和聚合，从而将两种语言的词级关系融入图结构中。

评论句和关键词的词频关系：

基于关键词在评论文本中出现的次数构建关键词与评论句之间的边，使用TF-IDF计算词频，其中TF是单词在评论句中出现的次数，IDF指的是由包含该单词的句子数量的对数缩放的逆分数，在评论句与关键词之间添加边并将计算的TF-IDF值作为边的权重。

评论句之间的语义相似度关系：

为了在评论句之间添加更直接的连接，使汉越两种语言的评论句可以更好的进行同一嵌入空间下的迁移学习，通过多语言预训练语言模型XLM-R得到汉越两种语言评论句的嵌入向量(A_i, B_j)，同时利用余弦相似度计算两个嵌入向量之间的相似性。

$$\cos\theta = \frac{A_i \cdot B_j}{|A_i| \times |B_j|} \tag{5}$$

其中$\mathrm{A_i} \in \mathrm{X_{S_c}}$表示第$i$条中文评论文本嵌入向量，$\mathrm{B_j} \in \mathrm{X_{S_v}}$表示第$j$条越南语评论文本嵌入向量。当余弦值越接近1表示两个向量的夹角越接近0度，也就是两个向量越相似，设置超参数Q作为阈值，找到余弦相似度最大的Q个评论文本添加边关系。

3.3 基于图卷积网络的节点特征学习

在构建多语言异构图后，将不同关系类别的子图进行融合，嵌入到一个简单的二层图卷积网络中。图卷积网络是一种多层神经网络，可以根据节点的领域属性引入节点的嵌入向量。GCN可以通过一层卷积来捕获关于近邻节点的信息，当堆叠多个GCN层时，图上更多的信息就会被整合起来。两层GCN可以允许信息在最多两步长的节点之间传递信息，对于一层GCN，新的s维节点特征矩阵$L^{(1)} \in \mathbb{R}^{n \times s}$为：

$$L^{(1)} = \rho\left(\tilde{A}XW_0\right) \tag{6}$$

其中$\tilde{A} = \widetilde{D}^{-1/2}A\widetilde{D}^{-1/2}$表示标准化对称邻接矩阵，$W_0 \in R^{d \times s}$表示权重矩阵。$\rho$是激活函数，本文使用的是RELU。通过叠加多个GCN层来学习合并更高阶的领域信息，学习更深层的节点特征。可以表示为：

$$L^{(j+1)} = \rho\left(\tilde{A}L^{(j)}W_j\right) \tag{7}$$

其中j表示层数，而$L^{(0)}$表示原始邻接矩阵。

3.4 评论文本观点对象类别预测

观点对象识别过程是判断当前节点属于哪一类别，属于分类过程。在图神经网络的第二层将评论文本嵌入维度映射成与类别标签相同的维度大小，然后送入到分类器中：

$$Z = \mathrm{softmax}\left(\tilde{A}\,\mathrm{Relu}\left(\tilde{A}XW_0\right)W_1\right) \tag{8}$$

其中$\mathrm{softmax} = \frac{1}{z}\exp(x_i)$，而$z = \sum_i \exp(x_i)$。

模型的损失函数使用交叉熵损失：

$$L = -\sum_{d \in y_D}\sum_{f=1}^{F} Y_{df} \ln Z_{df}。 \tag{9}$$

其中y_D是具有标签的评论索引集，F表示输出特征的维度，与类别数量相同，Y是标签矩阵。

4 实验设置

4.1 数据集

为了证明实验的有效性，本文参考Conneau et al. (2018)构建的XNLI多语言文本数据集格式，构建了汉越多语言观点对象识别数据集。利用网络爬虫技术在Twitter和新浪微博上爬取“新冠疫情”和“亚裔歧视”相关评论作为实验数据，通过语种识别方法清除非汉越数据，利用emoji数据包和正则表达式去除文本中的表情、符号以及超链接等，再进行数据筛查和整理完成数据清洗，对数据集按照6:2:2的比例划分训练语料、验证语料和测试语料，汉越观点对象数据集的具体划分信息和观点对象类别如表2，3所示:

种类	语种	训练语料	验证语料	测试语料
新冠疫情	中文	3000	1000	1000
	越南语	2000	600	600
亚裔歧视	中文	3000	1000	1000
	越南语	2000	600	600

Table 2: 汉越观点对象识别数据集（单位：条）

种类	新冠疫情	亚裔歧视
观点对象类别	新冠病毒	游行
	疫苗接种	亚裔
	疫情防控	种族歧视
	其它	其它

Table 3: 汉越评论文本观点对象类别

4.2 评价指标

与其他分类任务类似，本文实验使用测试数据集上准确度Acc（Accuracy）、精确度P（Precision）、召回率R（Recall）和F1值的结果作为评价指标，从而衡量模型的性能。公式如下:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2PR}{(P + R)} \tag{13}$$

其中TP表示正类被正确预测，FP表示负类被错误预测，FN表示正类被错误预测，TN表示负类被正确预测。

4.3 实验参数设置

利用Adam优化器对图卷积网络和分类器进行联合优化，实验使用多语言预训练模型mBert得到汉越双语评论文本的特征表示，向量维度为768，对每个评论句取相似度最高的Q个评论句，使用dropout防止过拟合。设置模型的最大训练批次为100个，设置early stopping在10个批次后当连续10次Epoch（或者更多次）没达到最佳精度时则模型训练终止，在验证集上选择最佳模型，所有实验都在单个GeForce RTX 3090 GPU上进行，具体信息如下表4所示。

参数	值
dropout	0.5
学习率	0.0005
最大轮次	100
滑动窗口大小	20
评论文本相似度阈值Q	3
GCN层数	2
GCN隐藏层维度	200

Table 4: 参数设置

4.4 基线模型

- MT+LM(Zhai and Lafferty, 2017)：将训练的越南语评论翻译为中文评论，利用预训练语言模型对评论句进行表征并训练观点对象分类器，最终在测试数据上实现观点对象识别。
- TF-IDF+LR(Yoo and Yang, 2015)：具有术语频率和反向文档频率加权的词袋模型加监督学习中经典的分类方法，以线性回归为理论支持，通过Sigmoid函数引入了非线性因素，解决分类任务。使用源语言中文训练的基线模型，并仅依靠双语词嵌入对目标进行分类。
- CNN(Kim, 2014)：采用TextCNN模型，使用源语言中文训练的基准模型，并仅依靠双语词嵌入对目标进行分类，设置卷积核大小为$\{3, 4, 5\}$。
- Node2vec(Grover and Leskovec, 2016)：Node2vec通过网络中的二阶随机游走来学习图的嵌入，通过在验证集上对$p, q \in \{0.25, 0.5, 1, 2, 4\}$进行网格搜索，为实验选择最佳的参数设置。
- MT+TextGCN(Yao et al., 2019)：将训练后的中文评论翻译为越南语评论，利用翻译后的文本进行异构图构建，并利用TextGCN对节点特征进行学习。
- CLHG(Wang et al., 2021)：使用基于异构图的图卷积网络，通过机器翻译对不同语言的文档进行翻译，文档和词之间存在的不同关系创建异构图结构。
- Bert+GCN(She et al., 2022)：使用图卷积网络获得评论文本的句法结构信息，多语言预训练语言模型获取文本的上下文信息，通过动态融合门对两个信息进行融合得到融合向量，对融合向量的文本进行识别分类。

5 实验结果分析

5.1 基线模型实验对比结果

表5列出了本文模型与基线模型在“新冠疫情”和“亚裔歧视”两个数据集上的实验对比结果。从实验结果可以看出，本文模型与其他基准模型相比有较大的优势，具体分析如下：

（1）将本文模型与MT+LM进行对比，以“新冠疫情”数据集为例，本文提出模型的F1值提升了25.51个百分点，分析原因在于翻译得到的标注语料含有大量噪声，同时只使用双语词嵌入的方法尚不具备捕获中文和越南语评论中观点对象关联信息的能力。

（2）对比本文模型与TF-IDF+LR基线时，本文模型的性能同样高于该基线的性能。推测原因是两种语言具有完全不同的词汇表，在多语言数据集上基于词袋模型捕获到的双语特征差距大，而MT+LM的Accuracy相比较TF-IDF+LR要高出1-6个百分点，说明机器翻译能够起到弥补语义鸿沟的问题。

（3）对比本文模型与CNN基线模型时，以”新冠疫情”数据集为例，本文模型的Accuracy和F1值分别高出23个和12个百分点。而CNN模型性能要比MT+LM模型性能好，说明利用CNN能够编码出更好的评论特征，同时也验证了仅使用嵌入向量无法完成观点对象关联信息捕获的问题。

（4）分析本文模型与Node2vec、MT+TextGCN的结果，Node2vec模型是基于同构网络设计，比直接对评论进行特征编码的性能有所提升，这一观察结果证实了异构信息能够提高模型

数据集	方法	Acc	P	R	F1
新冠疫情	TF-IDF+LR	0.5358	0.4455	0.4321	0.4796
	MT+LM	0.5900	0.42204	0.4241	0.4512
	CNN	0.7286	0.5429	0.5333	0.5810
	Node2vec	0.7600	0.6514	0.6243	0.6084
	MT+TextGCN	0.7700	0.6481	0.6067	0.6264
	CLHG	0.6920	0.6537	0.6537	0.6537
	Bert+GCN	0.9250	0.6743	0.6538	0.6639
	本文模型	**0.9625**	**0.7028**	**0.7098**	**0.7063**
亚裔歧视	TF-IDF+LR	0.5080	0.4167	0.4255	0.4426
	MT+LM	0.5150	0.4421	0.3608	0.4480
	CNN	0.6900	0.5725	0.5250	0.5204
	Node2vec	0.6975	0.6695	0.6284	0.6457
	MT+TextGCN	0.7100	0.6593	0.6546	0.6569
	CLHG	0.7220	0.7278	0.7278	0.7278
	Bert+GCN	0.9400	0.8446	0.8552	0.8499
	本文模型	**0.9625**	**0.9607**	**0.8787**	**0.9179**

Table 5: 汉越多语言事件观点对象识别方法性能对比

的表示学习能力。而MT+TextGCN模型的Accuracy和F1值相较于Node2vec普遍有1-3个百分点的提升，分析原因认为在进行节点特征学习的过程中，TextGCN将节点的新特征计算为节点自身及其二阶邻居节点的加权平均值，使得评论节点的标签信息能够进一步传递到相邻的其他评论节点和词节点中。

（5）对比本文模型与CLHG模型时，本文提出模型的Accuracy和F1值均高于CLHG模型，分析原因认为相比利用机器翻译缩小语言差异，利用词节点收集全面的评论标签信息，并且利用语义相似度计算捕获图中关联信息作为异构图中的关键路径，从而使标签信息可以传播到整个图中。

（6）分析本文模型与Bert+GCN的结果，以“新冠疫情”数据集为例，本文模型的Accuracy和F1值分别高出3.75个和4.24个百分点，多语言预训练语言模型可以学习到评论文本上下文的语义特征信息，有利于模型性能的提升，同时证实了图卷积网络能够学习邻居节点的特征信息，提高模型的表示学习能力。

5.2 不同多语言预训练语言模型的实验结果

为了验证不同的多语言预训练语言模型对本文模型方法的影响，本文分别使用mBert、XLM和XLM-R对数据集中的评论文本节点进行表征，词嵌入维度分别为786、1280和786，其它所有参数设置均相同，实验结果如表6所示。

数据集	多语言预训练语言模型	Acc	R	F1
新冠疫情	mBert	0.9318	0.6741	0.6638
	XLM	0.9475	0.6891	0.6893
	XLM-R（本文）	**0.9625**	**0.7098**	**0.7063**
亚裔歧视	mBert	0.9336	0.8306	0.8714
	XLM	0.9575	0.8505	0.8982
	XLM-R（本文）	**0.9625**	**0.8787**	**0.9179**

Table 6: 不同多语言预训练语言模型对实验结果的影响

观察表6可以发现，在使用不同的多语言预训练语言模型对节点进行表征，所有参数设置相同时，选择多语言预训练模型XLM-R做表征时模型效果最好。

5.3 图卷积层数设定对实验结果的影响

考虑到在图卷积学习的过程，图卷积层数的设定对聚合邻居节点信息程度有影响。本节针对图卷积层数在“新冠疫情”和“亚裔歧视”两个数据集上进行实验分析，实验结果如下图3所示：

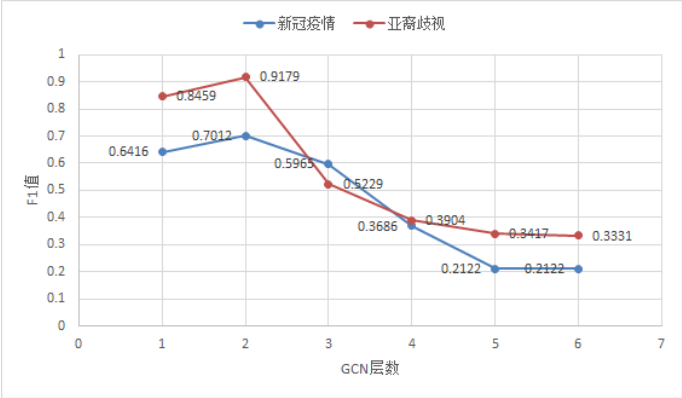


Figure 3: GCN层数设定对模型结果的影响

图3显示了在两个数据集上不同的卷积层数下模型F1值的结果，可以观察到模型的F1值首先随着卷积层数的增大而增加，当卷积层数为2时，模型的性能达到最佳，这表明卷积层数不足2时，卷积网络对信息聚合的能力不足，模型性能较低，当卷积层数超过2层后，随着层数的增加，模型性能有所下降并趋于稳定，因此本文提出的模型中将卷积层数的大小设定为2。

5.4 不同图结构对实验结果的影响

在本文中，我们通过使用汉越社交媒体评论文本数据集中的评论句和其中关键词作为节点构建汉越多语言异构图，构建了四种不同关系种类的子图，其中单语言图包括关键词之间的词共现关系以及评论文本和关键词的词频关系，多语言图包括词对齐关系和评论文本之间的语义相似度关系。本节根据不同构图方式在“新冠疫情”和“亚裔歧视”两个数据集上进行实验分析，实验性能的对比结果如下表7所示：

数据集	构图方法	Acc	R	F1
新冠疫情	单语言图	0.9250	0.6538	0.6639
	多语言图	0.9425	0.6993	0.6862
	本文	**0.9625**	**0.7098**	**0.7063**
亚裔歧视	单语言图	0.9400	0.8552	0.8499
	多语言图	0.9475	0.8501	0.8677
	本文	**0.9625**	**0.8787**	**0.9179**

Table 7: 不同图结构的性能对比

观察表7可以发现，在“新冠疫情”和“亚裔歧视”上构建多语言图比构建单语言图的F1值高，证明了进行跨语言的有效性，而本文方法是将单语言的两种子图和多语言的两种子图进行融合，所取得的F1值均高于单独构建子图的效果，实验结果表明添加不同类型边的关系会取得更好的性能。

6 结论

本文提出了一种融合汉越关联关系的多语言事件观点对象识别方法，利用关联事件下的汉越社交媒体评论文本数据作为模型训练语料，结合汉越评论文本之间以及关键词之间的关联关系构建多语言异构图，随后利用图卷积网络对该图进行建模，从而聚合邻居节点信息并捕获高阶领域信息，利用该方法能够识别出汉越双语评论文本中的观点对象，实验结果证明了本文所提方法的有效性。下一阶段的工作我们将重点研究如何融入观点对象信息对汉越评论文本进行情感倾向性的分析。

参考文献

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Ying Ding, Jianfei Yu, and Jing Jiang. 2017. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. corr abs/1408.5882. *arXiv preprint arXiv:1408.5882*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–419.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. *arXiv preprint arXiv:1805.00760*.

Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang. 2020. Learn to cross-lingual transfer with meta graph learning across heterogeneous languages. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2290–2301.

Samaneh Moghaddam and Martin Ester. 2011. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 665–674.

Huy Tien Nguyen and Minh Le Nguyen. 2018. Multilingual opinion mining on youtube–a convolutional n-gram bilstm word embedding. *Information Processing & Management*, 54(3):451–462.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Xiangrong She, Jianpeng Chen, and Gang Chen. 2022. Joint learning with bert-gcn and multi-attention for event text classification and event assignment. *IEEE Access*, 10:27031–27040.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120.

Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021. Cross-lingual text classification with heterogeneous graph neural network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 612–620.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

Jong-Yeol Yoo and Dongmin Yang. 2015. Classification scheme of unstructured text document using tf-idf and naive bayes classifier. *Advanced Science and Technology Letters*, 111(50):263–266.

Chengxiang Zhai and John Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM New York, NY, USA.

倪茂树and 林鸿飞. 2007. 基于关联规则和极性分析的商品评论挖掘. In 第三届全国信息检索与内容安全学术会议, volume 635, page 642.

基于网络词典的现代汉语词义消歧数据集构建

严福康、章岳、李正华*
苏州大学计算机科学与技术学院，江苏省，苏州市
20215227039@stu.suda.edu.cn; hillzhang1999@qq.com; zhli13@suda.edu.cn

摘要

词义消歧作为自然语言处理最经典的任务之一，旨在识别多义词在给定上下文中的正确词义。相比英文，中文的一词多义现象更普遍，然而当前公开发布的汉语词义消歧数据集很少。本文爬取并融合了两个公开的网络词典，并从中筛选1,083个词语和相关义项作为待标注对象。进而，从网络数据及专业语料中为抽取相关句子。最后，以多人标注、专家审核的方式进行了人工标注。数据集[1]包含将近2万个句子，即每个词平均对应约20个句子。本文将数据集划分为训练集、验证集和测试集，对多种模型进行实验对比。

关键词： 数据集；词义标注；词义消歧；网络资源

Construction of a Modern Chinese Word Sense Dataset Based on Online Dictionaries

Fukang Yan, Yue Zhang and Zhenghua Li
School of Computer Science and Technology, Soochow University, Suzhou, China
20215227039@stu.suda.edu.cn; hillzhang1999@qq.com; zhli13@suda.edu.cn

Abstract

The task of word sense disambiguation (WSD) is one of the most classic tasks in natural language processing, aiming to identify the accurate sense of polysemous words in a given context. In Chinese, the phenomenon of polysemy is more prevalent compared to English. However, there is a lack of publicly available Chinese word sense dataset. In this paper, we crawled and integrated two publicly accessible online dictionaries, from which we selected 1,083 words and their corresponding senses for annotation. Additionally, we extracted relevant sentences from web data and specialized corpora. Finally, a manual annotation process was conducted through multi-annotator labeling and expert review. The dataset comprises nearly 20,000 sentences, averaging around 20 sentences per word. We divided the dataset into training, validation, and testing sets, and conducted experimental comparisons with various models.

Keywords: dataset , word sense annotation , word sense disambiguation , network resource

* 通讯作者 Corresponding Author.
[1]https://github.com/SUDA-LA/Modern-Chinese-Word-Sense-Annotated-dataset

第二十二届中国计算语言学大会论文集，第43页-第53页，哈尔滨，中国，2023年8月3日至5日。

1 引言

词义消歧是自然语言处理中最经典的任务之一。它旨在识别多义词在给定上下文中的准确词义，以便更好地理解句子的含义(Weaver, 1952)。在汉语中多义词比例相对较低，但它们在自然语言中却被广泛使用。因此，词义消歧与语音识别、机器翻译、信息检索等领域密切相关，消歧的准确性直接影响这些领域的相关应用效率。

词义消歧数据集是词义消歧任务的基础语料，其质量关乎后续消歧任务的开展。鉴于其重要性，已有不少学者就如何构建高质量的词义消歧数据集展开了系统性的研究。一个完整的词义消歧数据集通常包含两种不同类型的数据集：1)词义数据集，包含待消歧词语的所有词义信息；2)标注语料数据集，能够将具有多种含义的词汇与其在语境中的正确词义联系起来，通常用于模型的训练和评估。

英文的词义消歧数据集建设已经趋于成熟，常用的词义数据集有WordNet(Miller et al., 1990)和BabelNet(Navigli and Ponzetto, 2012)。标注语料数据集则有许多不同的语料库可供使用，例如SemCor语料库(Miller et al., 1993)、Senseval-2语料库(Edmonds and Cotton, 2001)、Senseval-3语料库(Snyder and Palmer, 2004)，以及在语义测评比赛SemEval中使用的语料库：SemEval-2007 Task 17(Pradhan et al., 2007)、SemEval-2013 Task(Navigli et al., 2013)和SemEval-2015 Task 13(Moro and Navigli, 2015)，这些语料库都是基于WordNet中的词义构建而成。

相较于英文，汉语词义消歧研究起步较晚，数据资源相对匮乏。不过许多学者已开展相关工作并取得了不错的成果。例如北京大学的汉语词义标注语料库(STC)、汉语二语教学词义标注语料库、基于构词法的汉语词义消歧语料库(FiCLS)以及古汉语词义标注语料库。STC语料库(吴云芳and 俞士汶, 2006)使用《现代汉语语义词典》作为词义语料，对2000年1-3月和1998年1月的《人民日报》语料(共计642万字)进行多义词标注，标注了966个多义词义项。汉语二语教学词义标注语料库(王敬et al., 2017)和FiCLS语料库(Zheng et al., 2021)以《现代汉语词典》(CCD)为标注体系，前者对汉语二语教材文本(约350万字)中的1181个多义词进行标注；后者基于中文维基百科构建了有7064个多义词，囊括121655条标注数据的汉语词义消歧语料库。古汉语词义标注语料库(舒蕾et al., 2022)整合了多个词典资源，其语料库规模超过117.6万字，包含了3.87万条标注数据，极大的丰富了古代汉语领域的语言资源。

STC语料库构建时间较早，整体规模较小且缺乏时效性。汉语二语教学词义标注语料库和FiCLS语料库规模庞大，选用的都是近十年的数据作为标注语料，有较强的多样性和时效性，但二者都基于《现代汉语词典》进行语料标注，语料数据因词义数据集版权问题无法公开。因此当前汉语词义消歧任务面临高质量数据集获取困难的问题。同时本文注意到，鲜有数据集运用中文互联网资源，然而当前人们对一个词语的意思有疑惑时，会直接通过网络进行查询，从多个网络词典中找到最合适的解释。由此本文决定充分利用网络资源，基于网络数据构建一个高质量、公开的、由人工标注的且有一定规模的词义消歧数据集。

本文通过网络资源对多音节的多义词进行词义搜寻整合，筛选出1,083个词语作为待标注对象，从网络数据和专业语料中分别抽取句子开展词义标注，建成了超过85万字规模的现代汉语词义消歧数据集。以该库为基础，本文利用几种词义消歧模型进行词义消歧预测，其中模型F1值最高达到了77.74%。本文进一步分析了网络数据与专业语料的差别，证实了合理利用网络数据的可行性与可靠性。

2 网络词典爬取和融合

肖航和杨丽姣(2010)指出，构建词义消歧数据集存在两个难点。第一个难点是词典中的词义区分若不够清晰，可能导致标注结果不一致，第二个难点是词典提供的词义不够全面，会导致在标注时出现无法匹配的情况。同时，随着当前词义消歧技术的发展，义项作为词义的直观文字表示，其作用被逐渐挖掘。最近几年，向神经网络模型中添加义项信息来辅助消歧工作已成为主流，GlossBERT、BEM、ESCHER(Huang et al., 2019; Blevins and Zettlemoyer, 2020; Barba et al., 2021)这些充分利用义项信息的模型都取得了当时词义消歧任务的最佳水平。研究表明，当前神经网络模型能够有效的识别出不同的义项信息，同时义项信息对词义消歧研究有

项目资助：国家自然科学基金（62176173）、江苏高校优势学科建设工程资助项目

着极大的帮助。因此本文决定在构建词义数据集时，为同一词义添加多种文字描述，并尽可能提供全面的词义解释。

2.1 词义语料库选择

许多词典网站经过多年的更新和发展已经形成规模。本研究在经过深度分析及实际测试后发现，百度汉语[1]、在线新华字典[2]、汉典[3]这些在线词典词语丰富、词义质量较高且具备较高的专业性，但由于词义区分规则不同及构建人员文学认知水平存在差异，这些词典的义项大多不同，以“冲突”一词为例，百度汉语、在线新华字典、汉典分别给出了不同的义项，如表1所示。

网络词典	“冲突”义项
百度汉语	1.(动)矛盾激烈。
在线新华字典	1.有矛盾；争斗；争执。2.两种或几种动机同时存在又相互矛盾的心理状态。3.指文艺作品中人和人，人和环境，或人物内心的矛盾及其激化。
汉典	1.冲杀奔突。2.对立的、互不相容的力量或性质(如观念、利益、意志)的互相干扰。3.以争吵、摩擦和对立为特色的持久的不和。4.意见不合,发生争执。

Table 1: “冲突”在多种不同网络词典中的义项

汉典网站始建于2004年，是一个有着巨大容量的字、词、词组、成语等其它中文语言文字形式的免费在线辞典。其词典内的义项既参考了百度百科、维基百科等网络数据，又参考了王同亿编著的《高级汉语词典》，每一个词语义项都对有相应英文对应，英文解释是其进行义项区分的一个重要因素。汉典经过十多年的更新优化，其内部词义知识体系已趋渐完善，且义项充分，在使用者中获得了广泛好评。

不同于词义“粒度过细”的汉典，由无忧无虑中学语文网构建的在线新华字典词义更为凝练，针对同一个词义，在线新华字典与汉典的义项表述往往不同，正符合本文对词义数据集构建的需求。不过在线新华字典内部分词语存在义项冗杂、重复的问题，该词典在使用时需要提前进行专业的人工筛选。

在比对汉典与在线新华字典的数据特色后，本文拟将汉典与在线新华字典义项进行融合，充分利用不同词典的义项资源，在合并词典前后分别对两个词典数据进行人工清洗修改，以确保最终获得的词义数据集质量。

2.2 词语义项的获取及处理

2.2.1 网络词典的获取

为了获取汉典和在线新华字典的词典资源信息，本文根据两种网站的页面格式设计了不同的词语抽取策略。

在线汉语字典的网页源码采用HTML格式存储，整体风格较为简洁，包含了“基本解释”、“分解解释”、“相关词语”和“更多相关词语”等几个栏目。我们使用Beautiful Soup工具获取单个词语的网页源码，并通过正则表达式匹配获取“基本解释”中该词语的所有义项信息。在收集完当前词语的信息后，通过“相关词语”和“更多相关词语”中的链接依次爬取其他词语的网页源码，重复上述流程直到获得所有词语的信息。通过这种方式，本文共获得了368,024个词语解释，其中包含60,683个多义词。

汉典网站的词语解释页面功能较为丰富，单个词语的源码较大。为了提高爬取效率，我们采用了URL拼接的策略来获取特定词语的网页源码。例如，将“zdic.net/hans/”与“冲突”拼接即可组成汉典中“冲突”对应的URL：“zdic.net/hans/冲突”，使用Beautiful Soup工具便可获取“冲突”在汉典中的源码数据。采用这种方式对已经从在线汉语字典中获取的词语进行汉典源码爬取，取得源码中“词语解释”栏目中的相应词条作为对应词语的义项信息。基于这一策略，本文从汉典中共得到了302,273个词语解释，其中有62,620个词语是多义词。

[1]https://hanyu.baidu.com/
[2]http://xh.5156edu.com/
[3]https://www.zdic.net/

初步获得的在线新华字典和汉典数据存在诸多问题，包括网络词典本身存在的义项错误问题，以及由于爬取规则不全面引发的错误。为了使这些数据满足后续研究的需要，我们对各词典中多音节的多义词进行了系统整理。通过自动处理和专家人工审核相结合的方式对数据进行了清洗，确保各词典的词义满足以下要求：1)各个义项只包含词义的文本表示，不包含其他信息，如“英文解释”、“词性”、“音标”等。2)各义项表示的词义相互独立，不存在词义重叠的情况。3)如果义项中带有例句，则将各例句放在义项最后并用“|”分隔。最终得到的词典数据如表2所示。

网络词典	多义词数	词语词义数均值	单个义项长度均值
汉典	60,288	3.1	33.3
在线新华字典	58,995	2.4	36.2

Table 2: 处理后的网络词典数据

2.2.2 义项合并

在完成词典整理后，本文将对两个词典数据进行融合。一般的融合策略通常需要通过人工去重来保证融合后的语料中的每个词语义项都清晰可分。然而，在实际的融合过程中存在许多挑战。例如，不同词义的边界往往难以确定，同时两个词义可能存在交集，这给选择合适义项带来了困难。这些融合过程中的问题直接影响着后续的标注任务。因此，本文在融合词典数据时作出以下规定：1）允许在词义解释中存在表示同一词义的不同义项。2）不同词义可以存在交集。为了替代人工操作，本文采用了自动融合的方法，具体的融合流程如下所述：

步骤1：以在线新华字典数据为新词义数据集基础，将汉典中词语逐个添加至新词义数据集中，若原本在线新华字典中存在该词语词义解释，则跳转步骤2；若原本在线新华字典中不存在该词语词义解释，则跳转步骤3;

步骤2：将该词语在汉典里的义项逐个加入，加入时与在线新华字典中该词所有义项对比句子相似度，若句子相似度高于85%，说明二者表述极为相似，则不添加至新词义数据集;

步骤3：直接向新词义数据集添加该词语;

最终，本文得到了一个包含多种词义表述且有较广覆盖率的汉语多义词词义语料库，该语料库共有多义词59655个，每个词语平均有2.7个义项，每个义项平均有34.8个字，具体义项个数分布如图1所示。

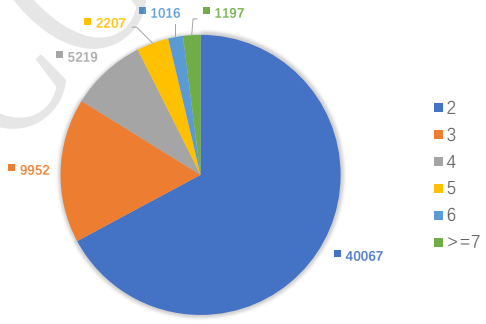


Figure 1: 不同义项数词语的分布情况

相比于人工融合，通过上述方法构建而成的词义语料库义项处理更为简单，极大的节约了词义语料库构建的人力物力成本。同时涵盖了更多更全面的义项，能够为当前深度学习模型的训练提供更多有价值的信息。

3 词义消歧数据集构建

3.1 待标注词语筛选

考虑到实际标注需要花费的人力成本，挑选1000个词语作为待标注词语，可以让每个词语都有足够多的标注语料与其对应，满足词义消歧模型的训练需要。同时为确保后续标注语料数据集的全面性和可靠性，本研究先对1998及2000年的《人民日报》中各词语进行词频统计，再将词语按照义项数分成六份(如图1中的划分方式)，依据词频抽取出这六份词语中各自出现最多的200个词语。

随后，本研究组织两名专家对这1,200个词语进行人工筛选，如果一个词语中只有一个词义是常用词义，其余词义在现代汉语中几乎不会被使用，则将该词语筛除。例如“等待”有两个义项：1.“不采取行动，直到期望或意料中的人、事务或情况出现；”2.“犹等到”。这里义项2出自《水浒传》第二回：“ 史进回到庄上，将陈达绑在庭心内柱上，等待一发拿了那两个贼首，一解官请赏。”，在现代汉语中“犹等到”这个义项几乎不再使用，若对含有“等待”的语句进行标注，标注结果单一，不符合词义消歧的需要，因此要将“等待”去除。最终，本文得到了1,083个包含多种常用词义的多义词作为后续标注语料数据集的主要标注词语。

3.2 标注语料采样及预处理

为确保标注数据的多样性和代表性，本研究拟从多种不同的语料库中抽取待标注语料。

《人民日报》和CoNLL2009中文语料是经过专业机构筛选和整理后的高质量语料库，有着规模大、内容丰富、文本质量高、语言风格统一的特点。而网络数据由本研究从各种网络资源爬取获得，是用户自由发布的文本内容，其包含大量的新闻报道、博客文章、舆论评价等信息，与另外两个语料相比，网络数据的语言风格更为多样化，其中词汇用法和上下文也更加复杂。采用多种语料库构建标注语料数据集，充分保证了数据的多样性。

虽然网络数据让标注语料更为多样，但其质量难以保证，存在错别字较多、句子过短、包含非文本相关的符号或标记等问题。针对这些问题本研究对获取的待标注句子进行人工预处理，确保即将被标注的句子满足以下原则：1)长度适中，待消歧词有较充足的上下文本。2)不含有错别字或逻辑错误。3)无与文本无关的内容。

基于上述原则，本研究从网络数据、1998年及2000年的《人民日报》语料和CoNLL2009中文语料中按照词语的五倍义项数抽取相应数量的句子，同时采用6：4的比例将网络数据和专业语料随机抽取的句子进行合并，得到24,455个句子。最后，对合并后的语料进行了人工处理，得到21,396个句子以供后续标注。

3.3 人工标注

本研究基于新构建的词义数据集和待标注语料，开展了语料标注工作。为确保标注质量，我们组织了共87名具有较强词汇理解能力的本科生和研究生参与人工标注。整个标注过程分为两个阶段。在第一阶段，我们使用了少量标注数据进行摸索尝试；依据第一阶段的标注经验，我们优化了标注流程和系统；第二阶段则展开了大规模的标注工作。

3.3.1 具体标注流程

参与标注的人员分为标注人员和审核人员，标注人员由82个本科生组成，审核人员由5个研究生组成。我们将对一个句子的标注称为”标注实例”，一个标注实例包含以下基础信息：1)含有待消歧词语的完整句子。2)待消歧词语的所有义项。每个标注实例会被分配给两个不同的标注人员，标注人员根据待消歧词语的上下文选择出适合此语境下的义项。如果两个标注人员选择的义项相同，该实例会被直接存储进数据集中。如果两个标注人员选择的义项不同，那该实例会进入审核流程，由审核人员进行判断并给出审核理由；经审核后的实例会再返回给标注人员，如果标注人员无异议，则进行学习；如果标注人员有异议，可以给出理由并进行投诉，被投诉的实例会交由另一个审核人员处理并给出最终答复。详细标注流程如图2所示。

3.3.2 标注规则

本研究构建的词义数据集包含重叠、相离和包含三种义项关系。重叠指的是两个义项的意思相同，但表述方式不同；相离指的是两个义项的意思不同；包含则指一个义项的意思被另一个义项所包含。以“接受”一词为例，其包含以下五个义项：1.收受。2.根据法令把机构、财产等拿过来。3.接纳。4.接纳；收受。5.依据法令收归己方所有。其中义项2和义项5之间存在重叠关

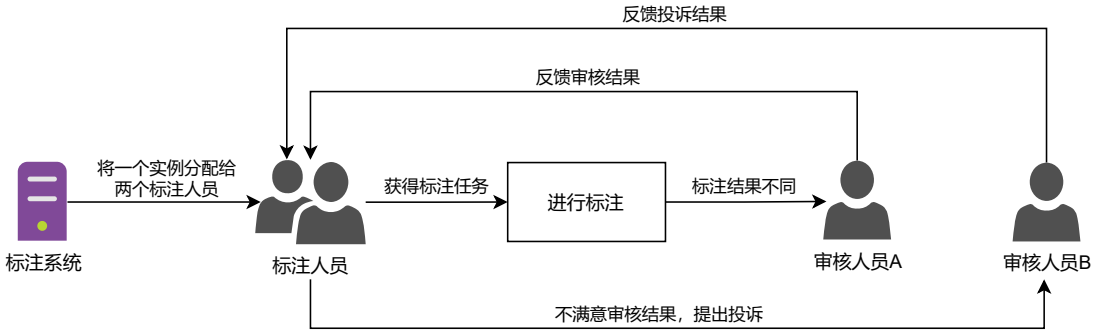


Figure 2: 标注流程

系，义项1和义项2之间是相离关系，义项1和义项4之间是包含关系。在标注过程中，标注者需要充分理解待标注句子，并根据义项的关系进行多选，具体多选规则如下：

规则一 不允许在一个标注实例中多选相离关系的义项。如果存在多个相离的义项都符合语境，则选择“待标注句子中词语有歧义”。

规则二 如果义项间存在重叠关系，则所有符合语境且重叠的义项都需要被选择。

规则三 如果义项间存在包含关系，则当被包含的义项被选择时，另一个义项也需要被选择。

规则四 如果只有一个义项符合语境，则单选该义项。

在实际标注中，有时候待标注句子可能存在一些本身就有错误或无法对应词义的情况。为了解决这些问题，我们特别添加了两个选项：“待标注句子本身有错误”和“没有合适的词义”供标注人员选择。同时，基于上述标注流程和多选规则，我们构建了一个轻量级的标注网站，使得人工标注变得更加方便和高效。图3展示了该标注网站的用户界面。

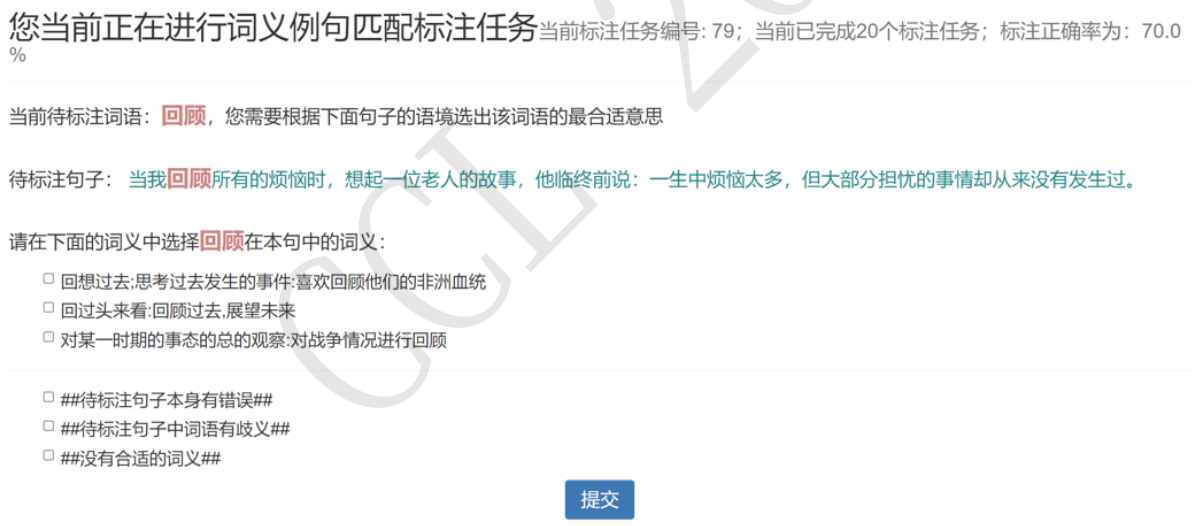


Figure 3: 标注系统界面

3.3.3 实际标注情况

在第一阶段的标注中，我们抽取了一小部分数据进行标注。每个标注人员需要逐一完成100个标注实例，这些实例是完全随机分配的。这一阶段的主要目的是筛选出准确率较高的标注人员，同时暴露出了以下问题：1)标注人员每次都需要学习新的词语义项，因此单个实例的学习成本过高。2)审核人员数量相对较少，标注人员完成标注后很难及时获得审核反馈。如果标注人员的词义理解出现错误，存在错误延续情况。

根据第一阶段的标注情况，我们挑选出了准确率较高的63名标注人员参加第二阶段的标

注。针对第一阶段中的问题，我们对标注系统进行了以下改进：1)将每个词语的所有实例分配给两个固定的标注人员，每个标注人员只有在完成一个词语的所有标注实例后才能开始下一个词语的标注。2)标注人员在获取一个词语的标注实例时，会先获得该词语30%的实例进行标注，只有在这些实例全部被审核完后，才会获得剩余的实例。在第二阶段，每个标注人员至少需要完成500个实例的标注工作。此外，在标注过程中，如果因为义项处理问题导致标注错误，这些错误也会被记录。在所有标注工作完成后，我们依据这些记录对词义数据集内的义项进行进一步优化处理。

所有标注完成后，我们对标注情况进行了统计。标注人员的平均标注准确率为74.6%，标注一致性达到了68.8%，进一步分析发现，一个词语前30%的实例的标注一致性仅为58.3%，标注者此时并未收到审核反馈，因此标注一致性较低，在学习审核结果后，标注一致性达到了76.2%。审核人员给出审核结果后，标注人员的投诉率为22.8%，其中有36%的投诉被采纳。从总体标注情况来看，标注人员对于有异议的审核结果能够积极反馈，且能从审核中充分学习到义项正确信息。

3.4 整体规模及义项分布

在本研究的标注过程中，有61名标注人员完成了词义消歧任务，共得到19,759个被正确标注的句子。其中，213个句子存在句子本身错误，160个句子中的词语具有歧义，304个句子无法找到合适的词义。最终，我们获得了19,082条高质量的标注语料，每条语料仅含有唯一的待消歧词，该词义消歧数据集涵盖了1,023个词语，总共包含4,831个义项，其中3,790个义项被正确标注的句子覆盖，义项标注覆盖率为78.45%。

数据集中各义项分布如图4所示，其中，有65.5%的义项在数据集中出现不足10次，且有18.4%的义项仅出现一次。对这些义项做进一步分析可发现：义项仅出现一次大多是因为该义项在近代几乎无人使用，但原词典包含了相应的例句，这些例句被收录在标注数据集中；一个词语的本义是最常被使用的词义，对应的义项在数据集中相较其它义项也更频繁。

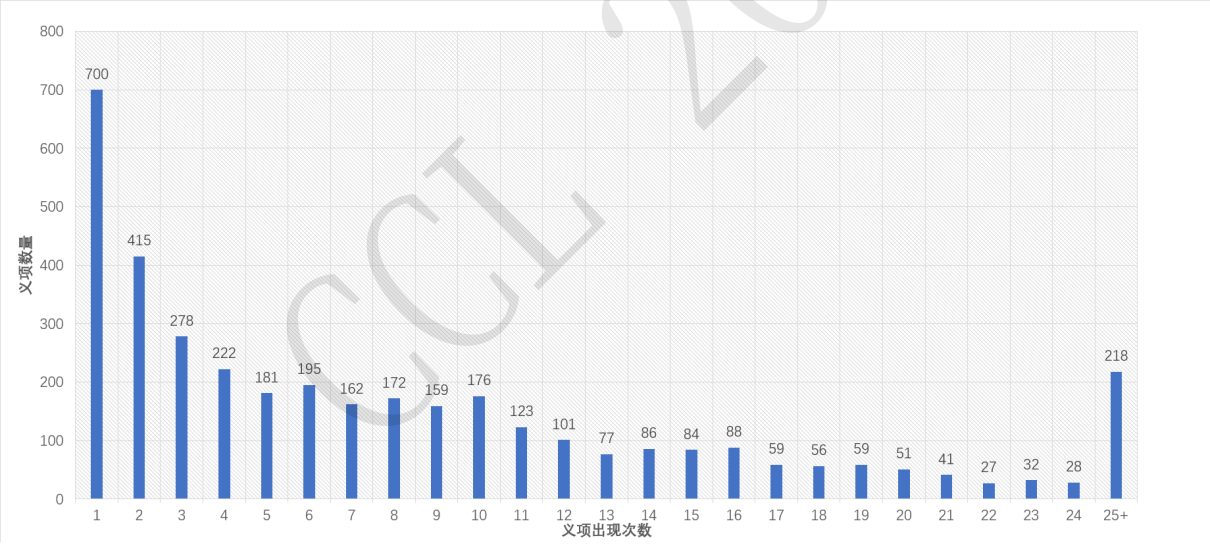


Figure 4: 数据集中各义项分布情况

3.5 网络数据与专业语料分析

本文构建的数据集共有19082条标注语料。其中，有9766条属于专业语料，10036条属于网络数据，这些语料的详细对比如表3所示。

可以看到，网络数据包含了更多的义项，极大的提高了整个数据集的词义覆盖率，同时，网络数据的格式与内容更为多样，弥补了专业语料格式单一的问题。在为数据集带来提升的同时，网络数据存在着以下问题：1)如图5所示，初始的网络数据存在很多噪音，需要花费较多人力进行清洗。2)网络数据的句长相对较短，句子提供的上下文信息有限。

语料资源	标注语句数	词语数	义项数	词义覆盖率	标注语句平均长度
专业语料	9,766	1,023	3,228	66.8	46.5
网络数据	10,036	1,021	3,689	76.3	41.3

Table 3: 网络数据与专业语料详细统计分析

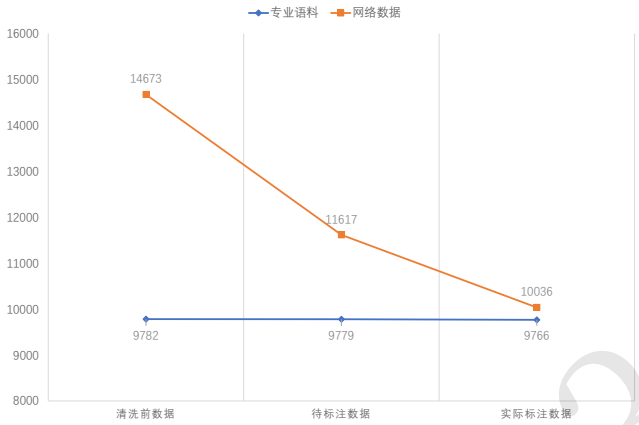


Figure 5: 标注数据规模变化

最终构建的标注数据集中，网络数据与专业语料占比近似1：1，这既保证了数据的专业性，又确保了语料的多样性和时效性。

4 词义消歧实验

4.1 模型介绍

GlossBERT 一种基于BERT(Devlin et al., 2018)的预训练模型，用于解决词义消歧问题。不同于传统的词义消歧模型，GlossBERT将义项信息也融入到神经网络中，并将词义消歧任务转化为句对分类问题。具体地，给定一个包含待消歧词w的文本c和其在词义数据集S中的对应义项$s_1, s_2, ..., s_n$，GlossBERT将c和S中各词义分别组合，并输入BERT进行句对分类。分类总共有两个标签：yes, no，分别表示s_n（是/不是）w的对应词义。该模型的创新之处在于首次将词语义项与BERT相融合，从而取得了较好的性能表现。

BEM 也是一种基于BERT的预训练模型。该模型采用双编码器的结构，将待消歧词w所在文本c和w对应的义项$s_1, s_2, ..., s_n$分别用BERT编码，并提取各自的[CLS]表示信息。相比于GlossBERT，BEM将所有义项都单独进行编码，有助于模型更好地理解不同义项表示的词义信息。实验结果表明，BEM在低频词上表现良好，进而提高了模型的整体性能表现。

ESCHER 是一种基于BART(Lewis et al., 2019)的预训练模型，它将词义消歧问题重新定义为一个跨度抽取问题。该模型把一个包含待消歧词w的文本c和w所对应的所有义项$s_1, s_2, ..., s_n$输入BART中，然后指定各个义项在模型输入中的位置信息，最终让模型生成与w最匹配的义项的位置。ESCHER首次将所有义项都放在同一个输入中让模型判断，相比于BEM，它在低频词上有着更好的效果。此外，它引入了高频噪音机制，通过向模型输入中随机添加高频词，从而变相降低高频词的选择率。在英文数据集上，ESCHER的F1值达到了80.7%。

4.2 数据集划分

本研究将之前构建的标注语料数据集按照7：1：2的比例划分为训练集，验证集，测试集。具体情况如表4所示。

数据集	词语数	义项数	标注语句数	标注语句平均长度	义项平均长度
训练集	918	3,259	15,244	44.2	11.0
验证集	604	1,453	2,893	43.1	11.7
测试集	678	1,778	3,645	44.6	11.6

Table 4: 数据集统计

为了评估模型在不同词义分布下的性能，我们采用了Zheng(2021)提出的测试集划分方法。将测试集按照词义在训练集中的出现频率分为四个子集：(1)最常用词义，即在训练集中对应词语出现最多的词义。(2)较常用词义，即在训练集中出现了超过5次，但不是对应词语出现最多的词义。(3)低频词义，即在训练集中出现过，但出现次数不多于5次的词义。(4)零样本词义，即该词义在训练集中从未出现过的词义。由于我们的测试集规模偏小，通过随机抽取的方式难以保证测试集中有足够数量的低频词义和零样本词义，因此本文在构建数据集时随机抽取了几个词语，将这些词语的所有语料放入测试集中，以此提高测试集中低频词义及零样本词义的比例。

值得注意的是，我们的标注数据集采用了多选策略，因此一个标注句子通常对应多个义项。同时，我们的数据集中待消歧词并不具有歧义，即每个标注句子对应的义项之间不存在相离的情况。这一特点保证了模型在训练过程中可以根据损失函数有效收敛。然而，这也需要我们在BEM 和ESCHER 模型加载数据时采用特殊的处理方式。

假定一个标注句子c的对应义项为$L = l_1, l_2, ..., l_k$。当使用BEM 或ESCHER 这类以最合适词义为输出的模型时，我们要对待消歧句子做单选衍变。一个标注句子进行单选衍变时会衍生为k个数据，这k个数据分别对应L中的一个义项，在进行词义预测时，对应词义数据会自动将L中其余义项从候选义项选择中去除。图6展示了这种单选衍变过程，图中原标注句子c有2个对应义项，4个候选义项。

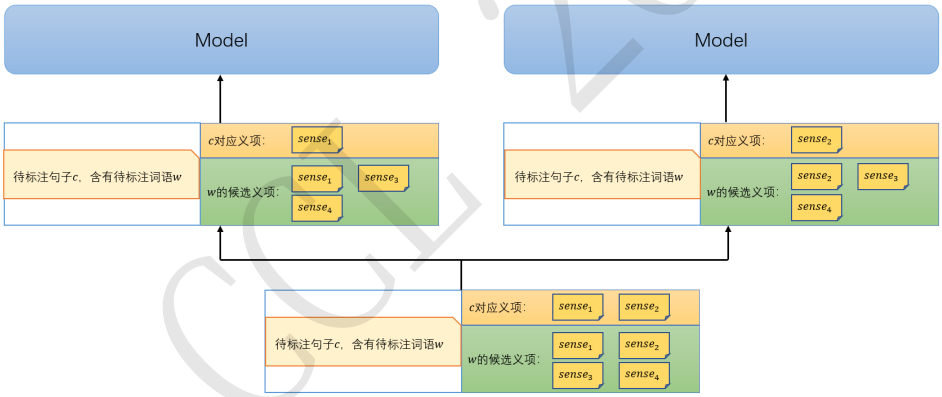


Figure 6: 数据单选衍变过程

4.3 实验结果分析

使用最常用词义(MFS)作为默认的基线参考，GlossBERT、BEM、ESCHER均采用原论文给定的参数设置，且都基于chinese-roberta-wwm-ext-large[1]进行训练微调。模型评估方面，我们以Macro-F1作为比较模型效果的依据。在本文构建的词义消歧数据集上，最终实验结果如表5所示。

相较于其它模型，GlossBERT的F1值偏低。在英文大规模数据集上GlossBERT有着不错的性能表现，但在我们的数据集上效果并不显著。从表5给出的结果不难发现，GlossBERT在低频词义及零样本词义方面效果较差，因此可以推断GlossBERT在小规模的数据集上难以学习到足够的词义知识。ESCHER是当前表现最好的模型，在小规模的数据集上，ESCHER已具备不

[1]https://huggingface.co/hfl/chinese-roberta-wwm-ext-large

模型	验证集	测试集	最常用词义	较常用词义	低频词义	零样本词义
MFS	26.78	41.85	×	×	×	×
GlossBERT	56.90	57.04	86.53	68.71	27.83	25.74
BEM	71.04	71.12	**92.32**	80.47	44.53	48.77
ESCHER	**76.81**	**77.74**	91.73	**84.38**	**52.57**	**54.53**

Table 5: 消歧模型结果

错的消歧能力，不论是BEM还是ESCHER，其模型性能的提升很大程度上源自于其在低资源词义数据上的改进。这种改进我们可以理解为是模型对于不同词义知识“理解”能力的提升。

考虑到标注语料的规模会对模型预测性能产生影响，我们从原1023个词语中分别抽取600、700、800、900个词语，依据4.2的划分规则对这些词语进行数据划分，并利用ESCHER模型分别训练、测试这些数据集，不同数据集规模的ESCHER表现如表6所示。

模型	总词语数	标注语料总数	测试集F1值
ESCHER_{600}	600	11,082	72.91
ESCHER_{700}	700	13,357	74.55
ESCHER_{800}	800	14,823	75.83
ESCHER_{900}	900	17,001	76.07
ESCHER_{all}	1,023	19,082	**76.81**

Table 6: 不同数据集规模下的ESCHER模型表现

当标注语料数不满15,000时，模型性能会随着语料的增加有较大提升，不过随着语料的不断增加，模型性能的提升幅度逐渐放缓，预估在标注语料数达到25,000时，模型性能受到语料规模的影响几乎可以忽略。不过值得注意的是模型性能在很大程度上也受限于数据集中的低频词义与零样本词义。因此，改善模型结构，提高模型对汉语词义的“理解”能力，让模型在更少量的数据上学习到更多的词义知识，是今后汉语词义消歧模型改进的一个重要方向。实践证明，模型能够从本文构建的现代汉语词义消歧数据集中学习到有效的词义知识，本文构建的数据集可以为后续的汉语词义消歧研究提供帮助。

5 结论

本文主要以现代汉语词义消歧数据集为研究对象，对两个公开的网络词典中多义词进行融合，筛选处理出1,083个词语作为待标注对象，以网络数据及专业语料作为源语料库进行标注语料抽取，并依据标注规则进行了人工标注。最终，本文构建的词义消歧数据集包含将近2万条标注数据，规模超过85万字。本文利用多种词义消歧模型对该数据集进行测试，既验证了数据集的质量，还探讨了汉语词义消歧模型发展的趋势。后续该数据集会无偿公开，方便更多学者进行汉语词义消歧研究。

不过本文构建的数据集规模依旧偏小，不太适用于对数据量有较大需求的模型，为了满足今后的科研需要，我们希望从以下几点继续改进现有的资源：1)对那些低频词义进行标注语料补充。2)用更多种类的语料扩充当前数据集，使语料覆盖更广。3)不局限于多音节词，对更多的多义词标注，努力实现全词标注。

参考文献

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. Esc: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

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

Philip Edmonds and Scott Cotton. 2001. Senseval-2: overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 288–297.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 87–92.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

Warren Weaver. 1952. Translation. In *Proceedings of the Conference on Mechanical Translation*.

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

吴云芳and 俞士汶. 2006. 信息处理用词语义项区分的原则和方法. 语言文字应用, (2):126–133.

王敬, 杨丽姣, 蒋宏飞, 苏靖杰, and 付静玲. 2017. 汉语二语教学领域词义标注语料库的研究及构建. 中文信息学报, 31(1):221–229.

肖航and 杨丽姣. 2010. 基于词典的语料库词义标注研究. 语言文字应用, (2):135–141.

舒蕾, 郭懿鸾, 王慧萍, 张学涛, and 胡韧奋. 2022. 古汉语词义标注语料库的构建及应用研究. 中文信息学报, 36(5):21–30.

基于多意图融合框架的联合意图识别和槽填充

尹商鉴, 黄沛杰*, 梁栋柱, 何卓棋, 黎倩尔, 徐禹洪
华南农业大学, 数学与信息学院, 广东广州, 510642
sjy8460@163.com, pjhuang@scau.edu.cn, liang_dz@stu.scau.edu.cn,
13428897035@163.com, li@stu.scau.edu.cn, xuyuhong@scau.edu.cn

摘要

近年来，多意图口语理解（SLU）已经成为自然语言处理领域的研究热点。当前先进的多意图SLU模型采用图-交互式框架进行联合多意图识别和槽位填充，能够有效地捕捉到词元级槽位填充任务的细粒度意图信息，取得了良好的性能。但是，它忽略了联合作用下的意图所包含的丰富信息，没有充分利用多意图信息对槽填充任务进行指引。为此，本文提出了一种基于多意图融合框架（MIFF）的联合多意图识别和槽填充框架，使得模型能够在准确地识别不同意图的同时，利用意图信息为槽填充任务提供更充分的指引。我们在MixATIS和MixSNIPS两个公共数据集上进行了实验，结果表明，我们的模型在性能和效率方面均超过了当前最先进的方法，同时能够有效从单领域数据集泛化到多领域数据集上。

关键词： 多意图口语理解；多意图融合框架；联合多意图识别和槽位填充

A Multi-Intent Fusion Framework for Joint Intent Detection and Slot Filling

**Shangjian yin, Peijie Huang*, Dongzhu Liang, Zhuoqi He,
Qianer Li, Yuhong Xu**
College of Mathematics and Informatics, South China Agricultural University, China
sjy8460@163.com, pjhuang@scau.edu.cn, liang_dz@stu.scau.edu.cn,
13428897035@163.com, li@stu.scau.edu.cn, xuyuhong@scau.edu.cn

Abstract

In recent years, multi-intent spoken language understanding (SLU) has become a research hotspot in the field of natural language processing. The current state-of-the-art multi-intent SLU model uses a graph-interaction framework for joint multi-intent detection and slot-filling, which can effectively capture fine-grained intent information for lexical element-level slot-filling tasks and achieves good performance. However, it ignores the rich information contained in the intent under joint action and does not fully utilize the multi-intent information to guide the slot-filling task. To this end, this paper proposes a joint multi-intent detection and slot-filling approach based on the multi-intent fusion framework (MIFF), which enables the model to accurately identify different intents while using the intent information to provide more adequate guidance for the slot-filling task. We conducted experiments on two public datasets, MixATIS and MixSNIPS, and the results show that our model outperforms current state-of-the-art methods in terms of performance and efficiency, while being able to effectively generalize from single-domain dataset to multi-domain dataset.

*通讯作者

第二十二届中国计算语言学大会论文集，第54页-第63页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

Keywords: Spoken language understanding , Multi-intent fusion framework , Joint intent detection and slot filling.

1 引言

口语理解（Spoken Language Understanding, SLU） (Tur et al., 2011; Young et al., 2013)是面向任务的对话的关键组成部分系统，其目的是创建一个语义框架，总结用户的请求。利用意图识别来识别用户意图，利用槽填充来提取相关语义成分，构建语义框架。由于意图识别和槽填充这两个子任务紧密相关，主流SLU系统采用联合模型来建模它们之间的相关性。在现实场景中，用户通常在话语中表达多个意图，如亚马逊内部数据集中52%的例子是多意图的(Gangadharaiah et al., 2019)。Figure 1展示了一个两意图示例，其中包含一个分类任务来对意图标签进行分类(即，预测意图为`atis_aircraft` 和`atis_city`) 和一个序列标记任务来预测槽标签序列(即，将语句标记为{`O`, `O`, `O`, `O`, `O`, `O`, `O`, `B-aircraft_code`, `O`, `O`, `O`, `O`, `O`, `O`, `B-city_name`, `O` })。然而，以往的工作大多只关注简单的单意图场景，无法有效处理原始网络的多意图。最近，多意图SLU逐渐受到关注，因为它在我们日常生活中具有重要的应用价值 (Gangadharaiah et al., 2019; Qin et al., 2000)，它可以处理包含多个意图的语句。为了满足现实生活的需要，Xu等人 (2013)和Kim等人 (2017)开始探索多意图SLU。然而，他们的模型只考虑了多重意图识别，而忽略了槽填充任务。最近，Gangadharaiah 和Narayanaswamy (2019) 首次尝试提出了一个多任务框架来联合建模多重意图识别和槽填充。Qin等 (2020)进一步提出了一种自适应交互框架（AGIF）来实现细粒度的多意图信息集成。基于图注意网络[7]和非自回归方式的思想，Qin等人 (2021)提出了一个全局局部图交互网络(global - local graph Interaction network, GL-GIN)，该网络建模了多个意图和话语中所有槽之间的槽依赖和交互，获得了先进的性能。虽然该模型取得了较好的效果，然而，它在模型浅层交互中只是将多意图和槽填充信息简单的拼接，没有充分挖掘意图中对槽的填充具有重要的指导意义的信息，从而让训练效率和模型性能大打折扣。

为了进一步增强多意图与槽填充的关联性以及协同优化的效果，我们提出了一个多意图融合框架，其核心是多意图融合层，该层包含了三种信息的融合过程，实现了在意图识别和槽填充任务之间建立更加稳定有效的连接。考虑到意图识别任务在联合效应中起着重要作用，我们提出了意图强化层，获得了更丰富的多意图信息表示，提高了整体框架的鲁棒性。在MixATIS (Hemphill et al., 1990)和MixSNIPS (Coucke et al., 2018)两个公共数据集上的实验结果表明，我们的框架获得了最先进的性能。此外，由于充分利用了意图信息，我们的模型更容易协调意图识别和槽填充任务，大大提高了训练效率。

我们的贡献总结如下：1）我们构建了一个意图强化层，更好地丰富了多意图信息和缓解信息遗忘问题。2）为了更好地利用意图信息，我们首次（在我们认知下）尝试构建了用于联合多意图识别和槽填充的多意图融合层。3）在两个公开数据集上的实验结果表明，我们的框架不仅达到了最先进的性能，而且更容易协调意图识别和槽填充，大大提高了训练效率，为对话系统的下游任务提供了一个创新的方法。

2 相关工作

尽管SLU已经有很长的历史，但对于多意图SLU的研究在近几年才出现作为一个新任务。多意图SLU是由Gangadharaiah等人 (2019) 首先提出的，并由Qin等人(2020; 2021)证明了其重要性。在这一节中，我们将介绍SLU历史中一些经典的工作。SLU任务通常由两个子任务组成，分别是槽填充和意图识别。槽填充任务可以被视为一个序列标注任务，而意图识别任务可以被视为一个分类任务。传统的方法包括用条件随机场（Conditional Random Fields，CRF） (Raymond et al., 2007)用于槽填充，以及用于意图识别的支持向量机（Support Vector Machines, SVM） (Haffner et al., 2003)和Adaboost (Schapire et al., 2000)，都取得了很好的效果。自从深度学习开始蓬勃发展以来，这两个子任务的性能也达到了一个更高的水平。循环神经网络（Recurrent Neural Network，RNN）首次被Yao等人引入SLU任务 (Yao et al., 2000)。随后，Yao等人(2014)在SLU任务中利用了长短期记忆网络（Long Short Term Memory，LSTM）的优势。他们都取得了显著的效果。从此，SLU任务步入了深度学习的时

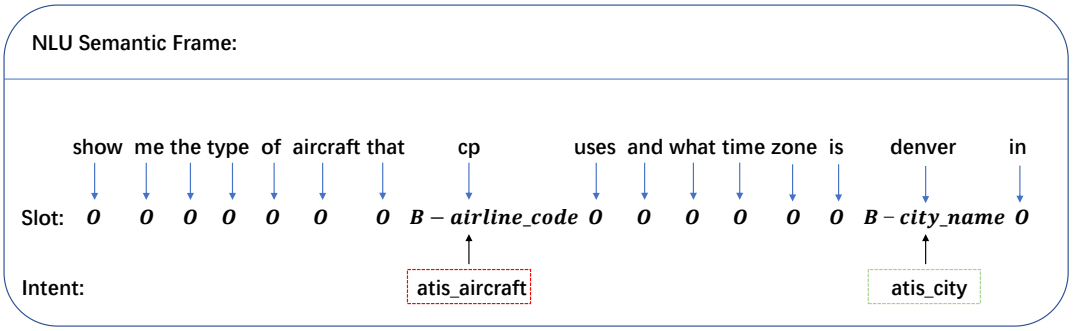


Figure 1: 多意图SLU框架

代。近年来，多任务训练，如联合学习考虑了意图与槽之间的相互关联，取得了很好的效果，推动了SLU的发展。Zhang等人(2016)通过引入一个共享的RNN编码器来建模意图和槽之间的关联，以此来实现槽与意图之间的联合，这可以被视为一个隐式的联合模型。Qin等人 (2019)提出了一个堆叠传播模型（Stack-Propagation），以更好地利用意图语义信息来指导槽填充。上述模型可以被视为单一信息流向的联合模型。E等人(2019)提出了一个新颖的SF-ID网络，同时考虑到槽到意图和意图到槽的双向影响，为槽填充和意图识别提供了双向影响的机制。最近，多意图SLU逐渐受到关注，因为它在我们日常生活中具有重要的应用价值 (Gangadharaiah et al., 2019; Qin et al., 2000)，它可以处理包含多个意图的语句。Gangadharaiah 等人(2019) 首次尝试提出了一个多任务框架，以联合建模多重意图识别和槽填充。Qin等人(2020)进一步提出了自适应交互框架（AGIF），以实现更细粒度的多意图信息集成。基于图注意网络和非自回归方式的思想，Qin等人(2021)提出了全局局部图交互网络（GL-GIN），该网络可以建模多个意图和话语中所有槽之间的槽依赖和交互，获得了先进的性能和效率。虽然该模型取得了较好的效果，然而，它在模型浅层交互中只是将多意图和槽填充信息简单的拼接，不能充分挖掘意图中对槽的填充具有重要的指导意义的信息，从而让训练效率和模型性能大打折扣，我们的模型主要基于GL-GIN上改进。此外，最新的研究还有Xing等人(2022)通过构建Co-guiding模型种实现两个任务之间的相互指导的新型模型，通过异构语义标签图实现多个意图识别和槽填充之间的相互指导。

为了进一步增强多意图与槽填充的关联性以及协同优化的效果，我们提出了一个多意图融合框架，其中核心结构是多意图融合层，该层包含了三种信息的融合过程，可以在意图识别和槽填充任务之间建立更稳定有效的连接。考虑到意图识别任务在联合效应中起着重要作用，我们还应用了意图强度层，以获得更丰富的意图在话语中的表示，从而提高整体框架的鲁棒性。我们在MixATIS(Hemphill et al., 1990)和MixSNIPS(Coucke et al., 2018)两个公共数据集上进行了实验，结果表明，我们的框架获得了最先进的性能。此外，由于我们充分利用了多意图信息，因此我们的模型更容易协调意图识别和槽填充任务，从而大大提高了训练效率。

3 问题定义

多意图识别. 给定输入序列$x = (x_1, \ldots, x_n)$，多意图识别可以被定义为一个多标签分类任务，其输出一个序列意图标签$o_I = (o_1^I, \ldots, o_m^I)$，其中$m$ 是给定话语中意图的数量，n 是话语的长度。

槽填充. 槽填充可以被视为一个序列标注任务，将输入话语x 映射到一个槽输出序列$o_S = (o_1^S, \ldots, o_n^S)$。

4 模型

在本节中，我们将详细介绍我们的MIFF模型。模型的体系结构如Figure 2所示。MIFF模型由一个共享编码器、两个解码器和两次融合过程组成。

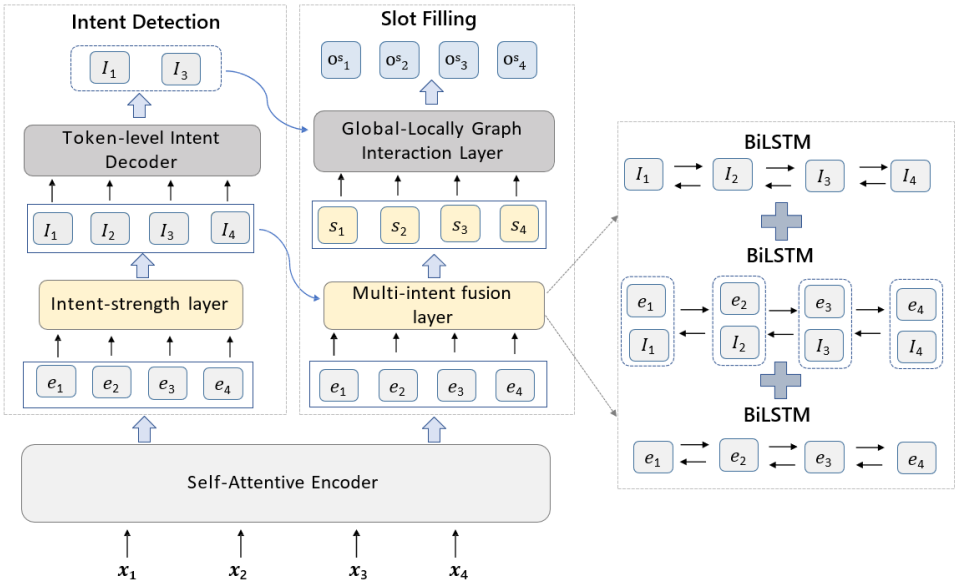


Figure 2: 模型架构和多意图融合过程

4.1 共享编码器

给定一个带有词元序列的句子 $\{t_1, t_2, \ldots, t_n\}$，输入嵌入层$\varphi^{emb}$ 将标记序列映射为嵌入序列 $X = x_1, x_2, \ldots, x_n \in \mathbb{R}^{n \times d}$（$d$ 表示嵌入维度）。随着Qin 等人 (2019) 提出的方案，具有双向LSTM（BiLSTM）的自注意力编码器被用来捕获在词元顺序和上下文信息中的特征。BiLSTM (Chen et al., 2017) 通过使用hi = BiLSTM(x_i) 生成上下文敏感的隐藏状态 $H = h_1, h_2, \ldots, h_n$。受Vaswani 等人 (Shaw et al., 2018) 的启发，词元矩阵表示上使用了自注意力机制。A = Self-Attention(H)。将H 和A 并联成一个矩阵，获取更丰富的语句编码信息：

$$\boldsymbol{E} = \boldsymbol{H} \,||\, \boldsymbol{A}. \tag{1}$$

4.2 意图强化层

我们构建了一个由带有残差连接 (He et al., 2016)的BiLSTM组成的意图强化层，利用更丰富的意图表示，最大限度地减少信息遗忘。我们将BiLSTM的输出和编码表示融合为意图表示：

$$\hat{h} = BiLSTM(\boldsymbol{E}), \tag{2}$$

$$h = BiLSTM(\alpha\ \hat{h}\ +\ \beta\ \boldsymbol{E}), \tag{3}$$

$$I = \sigma(\boldsymbol{W_I}(LeakyReLU(\boldsymbol{W_h}h + b_h)) + b_I), \tag{4}$$

其中$\hat{h}$和h用于增强特定任务的表征；$I = \{I_1 \ldots\ldots I_n\}$ 代表强化后的多意图隐藏层表示；α 和 β 是控制上下文意图信息强化超参数；σ表示sigmoid激活函数；W_h和W_I是可训练矩阵参数。

4.3 Token级别的意图识别解码器

我们使用token级别的多标签多意图识别 (Qin et al., 2021)来进一步提取有效的意图信息，其中通过对所有预测的tokens投票来获得句子结果。句子的意图结果o_k^I 可由下列公式获得：

$$o_k^I = \{o_k^I | \sum_{i=1}^{n} \mathbb{1}[I_{(i,k)} > 0.5] > n/2\}, \tag{5}$$

其中$I_{(i,k)}$表示i对o_k^I的分类结果。当标签在所有n标记中获得一半以上的正向预测时，我们就将其预测为语料意图，这时我们可以捕捉到更细化的意图表示。

4.4 层次递进的多意图融合框架

4.4.1 多意图融合层

我们提出的多意图融合层主要由三个部分组成：首先，应用一个意图感知的BiLSTM来建模更详细的意图连接来丰富句子中的意图表示，同时我们将上下文编码E提供给一个槽感知的BiLSTM来增强其对于特定任务的表示。此外，我们利用一个BiLSTM来产生槽-意图隐藏表示，来加强多意图与槽信息的交互。最后，我们将来自三个不同维度的信息通过按元素相加，以获得更丰富的意图-槽信息表示，给槽填充提供更有效的指导，提高联合效果。融合过程可以定义为:

$$S'_1 = BiLSTM(\boldsymbol{E}), \tag{6}$$

$$S'_2 = BiLSTM(\boldsymbol{I}), \tag{7}$$

$$S'_3 = BiLSTM(\boldsymbol{I} \mid\mid \boldsymbol{E}), \tag{8}$$

$$S = S'_1 + S'_2 + S'_3, \tag{9}$$

其中||代表了一个并联操作；$S = \{s_1.....s_n\}$代表了多意图融合层之后的最终槽意识隐藏表示。

4.4.2 全局-局部交互层

我们将全局-局部图交互层应用到我们的框架中，以提取更加细颗粒的意图-槽交互信息 (Qin et al., 2021)。局部槽位感知图交互层是为了建立槽位间的依赖关系，缓解槽位不协调的问题。全局槽位-意图图交互层是为了实现句子级的意图-槽位交互，所有预测的多个意图和序列槽位都被连接，实现平行输出槽位序列的目的。他们的工作基于图注意网络(GAT)，将初始节点特征$\widetilde{H} = \{\widetilde{h}_1, ..., \widetilde{h}_N\}$，旨在产生更抽象的表示$\widetilde{H}' = \{\widetilde{h}'_1, ..., \widetilde{h}'_N\}$ 作为其输出。一个典型的GAT的注意机制可以概括如下:

$$\tilde{\boldsymbol{h}}'_i = \Big\|_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \boldsymbol{W}_h^k \tilde{h}_j\right) \tag{10}$$

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[W_h \tilde{h}_i \| W_h \tilde{h}_j\right]\right)\right)}{\sum_{j' \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\boldsymbol{W}_h \tilde{h}_i \| \boldsymbol{W}_h \tilde{h}'_j\right]\right)\right)} \tag{11}$$

其中，$W_h \in \mathbb{R}^{F' \times F}$和$a \in \mathbb{R}^{2F'}$是可训练的权重矩阵；$N_i$表示节点$i$（包括$i$）的邻居；$\alpha_{ij}$是规范化的注意力系数，$\sigma$代表非线性激活函数；$K$是多头注意力数。

在通过线性层对每个图的邻居进行信息聚合后，我们能够获得更细颗粒的意图-槽交互信息表示。GAT中第l层的信息聚合过程可以表示为:

$$\boldsymbol{s}_i^{l+1} = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \boldsymbol{W}_l \boldsymbol{s}_j^l\Big), \tag{12}$$

其中, $\mathcal{N}_i$是一组顶点，表示连接的槽-槽或意图-槽。在堆叠了L层之后，我们得到了上下文意图-槽感知的隐藏特征$\boldsymbol{S}^{L+1} = \boldsymbol{s}_1^{L+1}, \ldots, \boldsymbol{s}_n^{L+1}$。

4.5 槽位预测

经过L层的传播，我们得到最终的槽位表示$\boldsymbol{s}_i^{l+1}$，用于槽位预测，它可以被表述为为:

$$y_t^S = softmax(W_s \boldsymbol{s}_i^{l+1}), \tag{13}$$

$$o_t^S = argmax(y_t^S), \tag{14}$$

其中W_s是一个可训练的参数，o_t^S是预测的槽位，t表示一个预测句子中的第t个词元，$\boldsymbol{s}_i^{l+1}$是经过L层的传播的聚合槽位信息。

4.6 协同训练

考虑到两个子任务之间的相关性，我们训练我们的模型，并联合更新参数，意图识别的目标可以表述为：

$$\mathcal{L}_1 = -\sum_{i=1}^{n}\sum_{j=1}^{n_I} \hat{y}_i^{(j,I)} \log\left(y_i^{(j,I)}\right), \tag{15}$$

其中n_I是意图的编号，$\hat{y}_i^{(j,I)}$是黄金意图标签。

同样地，槽填充的任务目标被表述为：

$$\mathcal{L}_2 = -\sum_{i=1}^{n}\sum_{j=1}^{n_S} \hat{y}_i^{(j,S)} \log\left(y_i^{(j,S)}\right), \tag{16}$$

其中n_S是槽的编号，$\hat{y}_i^{(j,S)}$是黄金槽标签。最终的联合目标是：

$$\mathcal{L} = \gamma\mathcal{L}_1 + (1-\gamma)\mathcal{L}_2, \tag{17}$$

其中γ是一个用与平衡意图识别和槽填充任务的超参数。

5 实验

5.1 数据集

我们在两个公开的多意图SLU数据集MixATIS和MixSNIPS上进行了实验。MixATIS数据集是一个多意图数据集，由单意图数据集ATIS构建而成，用于评估自然语言理解模型的性能，它来自航空公司的查询，包括13,162个用于训练的语料，756个用于验证的语料和828个用于测试的语料。MixSNIPS数据集包含来自餐厅、酒店、电影等领域的查询，是由单意图数据集SNIPS构建而成的多意图数据集，它包括39,776、2,198和2,199个用于训练、验证和测试的语料。

5.2 实验设置

我们设定自我注意编码器的隐藏单元为256，丢弃率为0.4，LSTM隐藏单元的维度为256，意图嵌入层的维度为128, 批量大小为16, 多头注意力数为6, 图注意力网络的层数为2, γ为0.75, α为0.8, β为0.2, 槽位预测解码器的维度为128。我们的模型以及复现的研究进展模型AGIF，GL-GIN 和Co-guiding都是使用在验证集上表现最好的模型，并在测试集上评估它的的表现。我们用F1分数来评估填槽填充的性能，用准确率来评估意图识别的性能，用总体准确率来评估语义解析的性能，它代表意图和槽位在语篇中全部被正确预测。我们所有的实验都是在RTX3090Ti上完成。

5.3 对比基线

我们将我们的模型与8个先进的基线模型进行比较，包括单意图SLU (Liu et al., 2016; Wang et ., 2018; E et al., 2019; Qin et al., 2019)和多意图SLU (Gangadharaiah et al., 2019; Qin et al., 2020; Qin et al., 2021; Xing et al., 2022)。

- Attention BiRNN (Liu et al., 2016): 提出了一种基于注意力的神经网络模型，用于联合意图识别和槽填充，对于许多语音理解和对话系统影响深远。
- Bi-Model (Wang et ., 2018)：一种用于语音理解的模型，它考虑了意图和槽填充之间的交叉影响。
- SF-ID (E et al., 2019)：一种用于语音理解的模型，它是一种新颖的双向关联模型，用于联合意图识别和槽填充。
- Stack-Propagation (Qin et al., 2019)：一种基于堆栈传播的框架，用于意图识别和槽填充任务。

Model	MixATIS		
	Slot(F1)	Intent(Acc)	Overall(Acc)
Attention BiRNN (Liu et al., 2016)	86.4	74.6	39.1
Bi-Model (Wang et ., 2018)	83.9	70.3	34.4
SF-ID (E et al., 2019)	87.4	66.2	34.9
Stack-Propagation[sig] (Qin et al., 2019)	87.4	71.9	41.0
Joint Mutiple ID-SF(Gangadharaiah et al., 2019)	84.6	73.4	36.1
AGIF (Qin et al., 2020)	86.9	72.2	39.2
GL-GIN (Qin et al., 2021)	87.2	76.0	42.5
Co-guiding (Xing et al., 2022)	86.53	74.03	43.35
MIFF	**87.7***	**77.2***	**45.0***

Table 1: 在MixATIS的实验结果。带*的数字表示我们的模型对所有基线的改进在t检验下具有统计学意义，p <0.05。

Model	MixSNIPS		
	Slot(F1)	Intent(Acc)	Overall(Acc)
Attention BiRNN (Liu et al., 2016)	89.4	95.4	59.5
Bi-Model (Wang et ., 2018)	90.7	95.6	63.4
SF-ID (E et al., 2019)	90.6	95.0	59.9
Stack-Propagation[sig] (Qin et al., 2019)	93.2	94.6	71.9
Joint Mutiple ID-SF(Gangadharaiah et al., 2019)	90.6	95.1	62.9
AGIF (Qin et al., 2020)	93.8	95.1	72.7
GL-GIN (Qin et al., 2021)	93.9	95.5	72.5
Co-guiding (Xing et al., 2022)	93.8	95.1	72.7
MIFF	**94.2***	**95.8***	**74.3***

Table 2: 在MixSNIPS中的实验结果。带*的数字表示我们的模型对所有基线的改进在t检验下具有统计学意义，p<0.05。

- Joint Mutiple ID-SF (Gangadharaiah et al., 2019)：一种多任务框架，可以联合学习槽填充（SF）和多个意图识别（ID）。
- AGIF (Qin et al., 2020)：一种自适应图交互框架，用于联合多个意图识别和槽填充。
- GL-GIN (Qin et al., 2021)：一种快速准确的非自回归模型，用于联合多个意图识别和槽填充。
- Co-guiding (Xing et al., 2022)：一种实现两个任务之间的相互指导的新型模型，通过异构语义标签图实现多个意图识别和槽填充之间的相互指导。

5.4 主要结果

Table 1和2展示了我们的模型在MixATIS和MixSINPS数据集上的实验结果,我们有以下观察和分析：

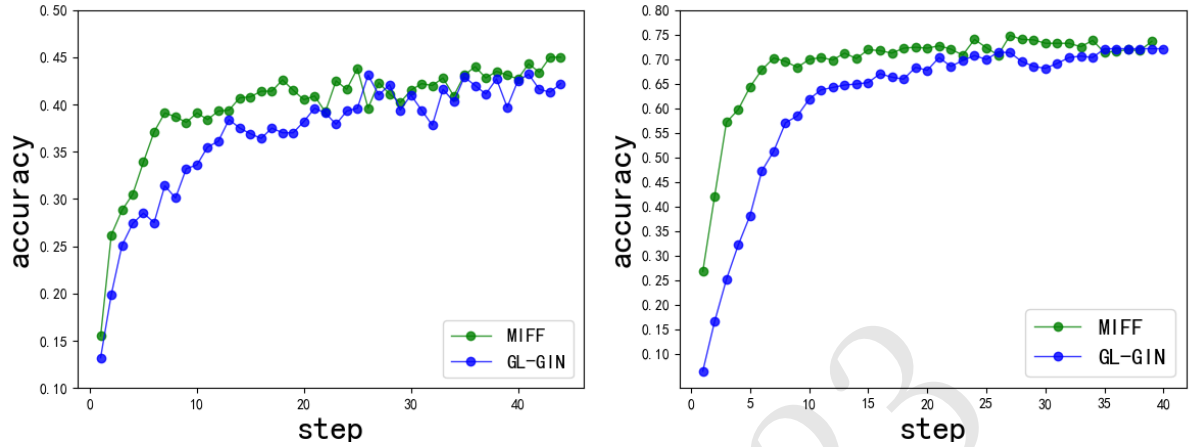
（1）在槽位填充任务上，我们的框架在两个数据集上的F1得分超过了所有的强基线，这表明我们的框架通过层次性多意图融合，有效利用丰富的意图信息来指导槽位填充。

（2）与GL-GIN相比，我们的框架在MixATIS和MixSNIPS的总体精度上分别实现了2.5%和1.8%的提高。我们认为，多意图融合框架能有效捕捉粗细颗粒的意图-槽信息表示，提高总体任务的联合性能。

（3）最重要的是，我们的框架在所有评价指标上都达到了最先进水平，展示了我们的框架在联合意图识别和槽填充任务的优越性，为今后的工作留下了进一步探索融合机制的空间。

Model	MixATIS		
	Slot(F1)	Intent(Acc)	Overall(Acc)
MIFF	87.7	77.2	45.0
- w/o 意图强化层	88.0	76.4	44.0
- w/o 多意图融合层	87.7	76.0	42.6

Table 3: 在MixATIS数据集上的消融实验.



(a) MixATIS (b) MixSNIPS

Figure 3: 训练效率的对比

5.5 进一步分析

在本节，我们主要在MixATIS数据集上研究分析了意图强化层和多意图融合层在我们整体模型框架的作用，结果如Table 3所示。

5.5.1 消去意图-强度层

为了检验意图强度层的有效性，我们将其替换为单一的BiLSTM层进行对照实验。实验结果显示，其在整体语义解析的准确性上降低了1.0%，在意图识别准确率上降低了0.8%。这表明我们提出的意图强化层对于丰富了多意图信息和缓解信息遗忘问题有重要作用，它可以通过集连多个维度的信息，获得更全面的意图表示信息，从而提高了多意图识别效能，同时提升整体语义解析的性能。

5.5.2 消去多意图融合层

为了进一步检验多意图融合层的有效性，我们将三种信息融合通道(意图-意图，槽-槽，槽-意图) 替换为单一的槽-槽信息通道。结果显示，它的语义解析精度降低了2.4%。这意味着多意图融合对于联合意图识别和槽位填充非常重要。它可以为槽位任务提供更多有用的意图信息，提高联合意图识别和槽位填充之间的性能。

5.5.3 对槽填充任务的影响

从Table 3可以看出，在意图强化层和多意图融合层联合作用下，槽填充F1性能没有得到显著地提高，然而总体的语义解析性能得到改善，我们认为，我们提出的多意图融合框架更适用于提升总体的联合多意图识别和槽填充性能。

5.6 效率评估

我们在16个批次下评估了我们的框架和目前效率最高的非自回归框架(GL-GIN)之间的训练效率，如图3所示，在这两个数据集上，我们的框架相比GL-GIN有更卓越的训练效率。我们认为，多意图融合层可以更加准确地提供细粒度的意图表示信息，以加快槽的填充任务。同时我

们发现，我们的模型框架在多领域数据集MixSNIPS上的表现更加突出。我们认为，多意图融合框架能有效泛化到多领域的语料任务，这为今后的工作留下了进一步探索融合机制的空间。

6 总结与未来工作

在本文中，我们提出了一个用于联合多意图识别和槽位填充的多意图融合框架，它可以提取更丰富的意图信息来指导槽位填充任务，在槽位填充和意图识别之间建立更强的相关性，同时提高训练效率。我们在MixATIS和MixSNIPS两个公共数据集上进行了实验，结果表明，我们的模型在性能和效率方面均超过了当前最先进的方法，同时能够有效从单领域的MixATIS数据集泛化到多领域的MixSNIPS数据集上。

致谢

本文受到广东省自然科学基金(2021A1515011864)、国家自然科学基金(71472068)、广州市智慧农业重点实验室(201902010081)、广东省普通高校特色创新项目(2020KTSCX016)、国家级大学生创新训练计划项目(202210564069)的资助。

参考文献

Gokhan Tur and Renato De Mori, "Spoken language understanding: Systems for extracting semantic information from speech" *Wiley, New York, 2011.*

Steve Young, Milica Ga˘si c, Blaise Thomson, and Jason D Williams, "Pomdp-based statistical spoken dialog systems: A review" Proceedings of the IEEE, vol. 101, no. 5, pp. 1160–1179, 2013.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy, "Joint multiple intent detection and slot labeling for goal-oriented dialog," *In Proceed- ings of the 2019 Conference of the North American Chapter of the Association for Computational Linguis- tics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 564–569*

Puyang Xu and Ruhi Sarikaya, "Convolutional neu- ral network based triangular CRF for joint intent detec- tion and slot filling," In 2013 IEEE Workshop on Au- tomatic Speech Recognition and Understanding, Olo- mouc, Czech Republic, December 8-12, 2013, pp. 78– 83

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na, "Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation" *In Pro- ceedings of the Second Conference on Machine Trans- lation, 2017, pp. 562–568.*

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu, "AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling" In Find- ings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, pp. 1807–1816

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li'o, and Yoshua Bengio, "Graph attention networks" *CoRR, vol. abs/1710.10903, 2017*

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanx- iang Che, and Ting Liu, "GL-GIN: fast and accurate non-autoregressive model for joint multiple intent detection and slot filling" In Proceedings of the 59th Annual Meeting of the Association for Computational Linguis- tics and the 11th International Joint Conference on Nat- ural Language Processing (ACL-IJCNLP 2021), (Vol- ume 1: Long Papers), Virtual Event, August 1-6, 2021, pp. 178–188

Charles T Hemphill, John J Godfrey, and George R Dod- dington, "The atis spoken language systems pilot corpus" *In Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990, 1990*

Alice Coucke, Alaa Saade, Adrien Ball, Th eodore Bluche, Alexandre Caulier, David Leroy, Cl ement Doumouro, Thibault Gisselbrecht, Francesco Calta- girone, Thibaut Lavril, Ma el Primet, and Joseph Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice inter- faces" CoRR, vol. abs/1805.10190, 2018.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang, "Im- proving sentiment analysis via sentence type classifica- tion using bilstm-crf and cnn" *Expert Systems with Ap- plications, vol. 72, pp. 221–230, 2017*

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations" CoRR, vol. abs/1803.02155, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition" *In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27- 30, 2016, pp. 770–778.*

Bing Liu and Ian Lane, "Attention-based recurrent neu- ral network models for joint intent detection and slot filling" In Proceedings of the 17th Annual Confer- ence of the International Speech Communication As- sociation (INTERSPEECH 2016), San Francisco, CA, USA, September 8-12, 2016, pp. 685–689

Yu Wang, Yilin Shen, and Hongxia Jin, "A bi-model based RNN semantic frame parsing model for intent de- tection and slot filling" *In Proceedings of the 2018 Con- ference of the 16th North American Chapter of the As- sociation for Computational Linguistics: Human Lan- guage Technologies (NAACL-HLT 2018), New Orleans, Louisiana, USA, June 1-6, Volume 2 (Short Papers), 2018, pp. 309–314.*

Haihong E, Peiqing Niu, and Zhongfu Chen, "A novel bi-directional interrelated model for joint intent detection and slot filling" In Proceedings of the 57th Conference of the Association for Computational Lin- guistics (ACL 2019), Florence, Italy, July 28- August 2, Volume 1: Long Papers, 2019, pp. 5467–5471.

Libo Qin, Wanxiang Che, and Yangming Li, "A stack-propagation framework with token-level intent detection for spoken language understanding" *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), Hong Kong, China, November 3-7, 2019, pp. 2078–2087.*

Bowen Xing and Ivor W. Tsang "Co-guiding Net: Achieving Mutual Guidances between Multiple Intent Detection and Slot Filling via Heterogeneous Semantics-Label Graphs" Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022

Christian Raymond and Giuseppe Riccardi, "Generative and discriminative algorithms for spoken language understanding" *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, pages= 1605–1608*

Patrick Haffner, Gokhan Tur, and Jerry H, "Optimizing SVMs for complex call classification" In proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, April 6-10, 2003, pp. 632–635.

Robert E. Schapire and Yoram Singer "BoosTexter: A Boosting-based System for Text Categorization" *Machine Learning, 2000, 39(2/3):135–168.*

Kaisheng Yao, Baolin Peng, and Geoffrey Zweig, et al. "Recurrent conditional random field for language understanding" In Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), Florence, Italy, May 4-9, 2014, pp. 4077–4081.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig and Yangyang Shi, "Spoken language understanding using long short-term memory neural networks" *in Proceedings of the 2014 IEEE Workshop on Spoken Language Technology (SLT 2014), South Lake Tahoe, NV, USA, December 7-10, 2014, pp. 189–194.*

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy, "Joint multiple intent detection and slot labeling for goal-oriented dialog" in Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, MN, USA, June 2-7, Volume 1 (Long and Short Papers), 2019, pp. 564–569.

Libo Qin, Xiao Xu, Wanxiang Che, et al. "Towards fine-grained transfer: An adaptive graph-interactive framework for joint multiple intent detection and slot filling" *in Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November, 2020, vol. EMNLP 2020 of Findings of ACL, pp. 1807– 1816.*

基于词频效应控制的神经机器翻译用词多样性增强方法

史学文[1], 鉴萍[2]*, 唐翼琨[2], 黄河燕[2]

[1]东北财经大学 数据科学与人工智能学院，中国 辽宁 大连，116025
[2]北京理工大学 计算机学院，中国 北京，100081
polarlion@qq.com {pjian, tangyk, hhy63}@bit.edu.cn

摘要

通过最大似然估计优化的神经机器翻译（NMT）容易出现不可最大化的标记或低频词精度差等问题，这会导致生成的翻译缺乏词级别的多样性。词频在训练数据上的不均衡分布是造成上述现象的原因之一。本文旨在通过限制词频对 NMT 解码时估计概率的影响来缓解上述问题。具体地，我们采用了基于因果推断理论的半同胞回归去噪框架，并结合本文提出的自适应去噪系数来控制词频对模型估计概率的影响，以获得更准确的模型估计概率，并丰富 NMT 译文用词的多样性。本文的实验在四个代表不同资源规模的翻译任务上进行，分别是维吾尔语-汉语、汉语-英语、英语-德语和英语-法语。实验结果表明，本文所提出的方法在提升 NMT 译文词级别多样性的同时，不会损害译文的质量。另外，本文提出的方法还具有模型无关、可解释性强等优点。

关键词： 神经机器翻译 ； 译文多样性 ； 因果推断

Improving Word-level Diversity in Neural Machine Translation by Controlling the Effects of Word Frequency

Xuewen Shi[1], Ping Jian[2]*, Yi-Kun Tang[2], Heyan Huang[2]

[1]School of Data Science and Artificial Intelligence, Dongbei University of Finance and Economics, Dalian, Liaoning, China, 116025
[2]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, 100081
polarlion@qq.com {pjian, tangyk, hhy63}@bit.edu.cn

Abstract

Neural machine translation (NMT) optimized by maximum likelihood estimation is prone to problems such as unargmaxable tokens or poor accuracy of low-frequency words, which leads to the lack of word-level diversity in the generated translations. The unbalanced distribution of word frequency on the training data is one of the reasons for the above phenomenon. This paper aims to alleviate the above problems by limiting the impact of word frequency on the estimated probability when decoding NMT . Specifically, we adopt a denoising framework of Half-Sibling Regression based on causal inference theory, combined with the adaptive denoising coefficient proposed in this paper to control the effect of word frequency on estimated probability, in order to obtain more accurate model estimated probability, and enrich the diversity of the words used in NMT translations. The experiments in this paper are carried out on four translation tasks representing different resource scales: Uyghur-Chinese, Chinese-English, English-German and English-French. In addition, the proposed method is model-agnostic and interpretable.

Keywords: Neural machine translation , Translation diversity , Causal inference

* 通讯作者：鉴萍；Corresponding author: Ping Jian

源语言	警方 现 追缉 涉案 的 另 一 名 男子 。
参考译文	Police are still hunting for the other man involved in the case .
	$\log f_{req}(\text{“hunting”}) = -4.85$
NMT译文	Police are looking for the other man in connection with the case .
	$\log f_{req}(\text{“looking”}) = -3.97$

图 1: NMT生成译文与训练数据中的原译文对比

1 引言

近年来，端到端的神经机器翻译（Neural Machine Translation，NMT）(Sutskever et al., 2014; Bahdanau et al., 2015) 在机器翻译领域取得了令人瞩目的成就，在某些特定的翻译任务上，机器译文已经接近人类译文的水准 (Wu et al., 2016; Vaswani et al., 2017; Hassan et al., 2018)。NMT 模型通常构建在编码器-解码器 (Cho et al., 2014) 架构上，其中，编码器的作用是将源语言序列 $\mathbf{x} = \{x_1, \ldots, x_{T_x}\}$ 转换成一组隐状态表示 $\mathbf{h} = \{h_1, \ldots, h_{T_x}\}$，解码器则被用来对如公式 (1) 所示的译文 $\mathbf{y} = \{y_1, \ldots, y_{T_\mathbf{y}}\}$ 的翻译概率建模：

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T_y} p(y_t|\mathbf{y}_{<t}, \mathbf{h}) 。 \tag{1}$$

经典的 NMT 方法 (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) 通常利用最大似然估计（maximum likelihood estimation，MLE）对 NMT 模型进行优化，训练的损失函数 $\mathcal{L}_{nmt}$ 通常采用负对数似然的形式：

$$\mathcal{L}_{nmt} = -\log p(y_t|\mathbf{y}_{<t}, \mathbf{x}, \theta), \tag{2}$$

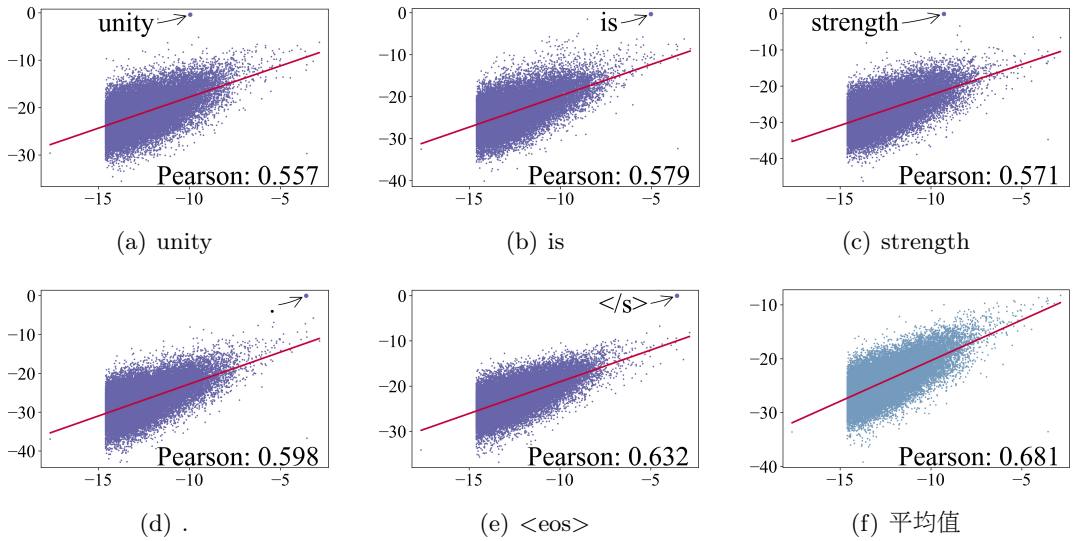
其中 θ 表示 NMT 模型的自由参数集合。

在机器翻译训练数据中，单词的词频分布是不平衡的，因此，经过 MLE 训练的 NMT 模型在解码阶段倾向于生成更高频的单词，而不是最适合的单词。例如，我们利用一个训练完成的汉语-英语 NMT 模型去重新翻译该模型所使用的训练集中的句子：“警方 现 追缉 涉案 的 另 一 名 男子 。”，图 1 给出了模型生成的译文与训练集中原始译文的对比。如图 1 所示，NMT 模型生成的译文与训练集的参考译文拥有相似的句法结构，但是对于源语言单词“追缉”，模型生成的译文使用了训练集中词频更高的单词“looking”而不是“hunting”。NMT 模型倾向于选择更高频单词的现象可能会引起“不可最大化的标记（unargmaxable tokens）”(Demeter et al., 2020; Grivas et al., 2022) 和“低频词准确率低”(Koehn and Knowles, 2017; Ott et al., 2018) 等问题。

针对上述问题，已有的工作主要分为以下两种方向：（1）在训练 NMT 时引入自适应的损失函数 (Lin et al., 2017; Gu et al., 2020; Xu et al., 2021; Zhang et al., 2022)；（2）尽可能消除词表示中与词频有关的部分信息 (Gong et al., 2018; Yang and Liu, 2020; Liu et al., 2020)。这两类工作的核心思想是尽量消除或缓解训练集中词分布不平衡对于训练损失信号或词向量分布的影响。上述方法必须作用于 NMT 的训练过程中，无法对已经存在的优化好的模型使用。

在本文，我们提出了一种基于词频效应控制的 NMT 译文用词多样性增强方法。该方法引入了半同胞回归 (Schölkopf et al., 2016) 去噪框架，结合本文提出的自适应降噪系数，通过调整目标语言单词的词频信息在 NMT 解码时的影响，以缓解 NMT 解码时倾向于选择高频词的问题，从而增强 NMT 译文的多样性。本文在 4 个不同语言对的翻译任务上进行了实验以验证提出的方法的有效性，分别是维吾尔语到汉语翻译（维-汉），汉语到英语翻译（汉-英），英语到德语翻译（英-德）以及英语到法语翻译（英-法）。上述四种翻译任务分别代表了四种任务类型：低资源翻译（维-汉），中等资源翻译（汉-英，英-德）以及丰富资源翻译（英-法）。实验结果表明，本文提出的方法在不损害译文质量的前提下，增强了 NMT 译文词级别多样性。

第二十二届中国计算语言学大会论文集，第64页-第77页，哈尔滨，中国，2023年8月3日至5日。



(a) unity (b) is (c) strength

(d) . (e) <eos> (f) 平均值

图 2: 词频分布与模型估计概率的关系

2 观察与讨论

2.1 模型估计概率与词频分布的相关性

观察发现，NMT 模型在解码时输出的概率分布（记作 O）通常与训练数据中目标语言词汇频率分布（记作 F）有很高的正相关性。例如，图 2(a) ∼ 图 2(e) 展示了汉-英翻译任务中，将源语言句子“团结 就 是 力量 。”翻译成目标语言句子“unity is strength .”时的每个解码步骤中，NMT 模型估计的目标语言词汇概率分布与目标语言词频分布的关系；而图 2(f) 展示了解码时模型输出概率的平均值与词频分布的关系。图中 x 轴表示目标语言单词频率的对数（即 $\log F$），y 轴表示对应单词的解码输出概率的对数（即 $\log O$）；图中直线表示以 x 轴为输入数据，y 轴为输出数据拟合的线性回归的曲线；“Pearson”表示 x 轴数据与 y 轴数据的 Pearson 相关系数，正值表示正相关，反之表示负相关。此外，图 2(a) ∼ 图 2(e) 标注了对应解码步骤模型估计概率排名最高的单词的位置。如图 2 所示，从线性回归的曲线图像和 Pearson 相关系数的数值（均大于 0.5）可以看出：在该翻译案例解码过程中，对于大多数目标语言单词，其在训练数据的词频与 NMT 模型解码时估计的概率值具有较高的正相关性。

2.2 相关性在不同概率区间的差异

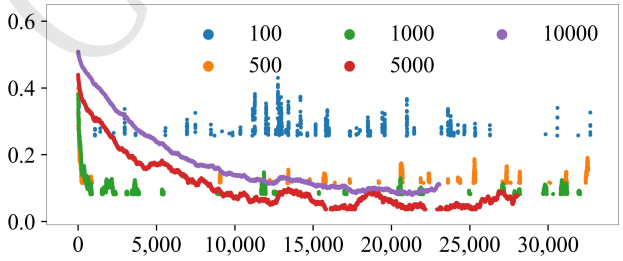


图 3: 不同概率区间下词频与模型生成概率的 Pearson 相关性对比

O 与 F 的相关性在不同的估计概率区间上的表现是有差异的。据观察，在每个解码步骤，排位较高的估计概率通常与对应的词频分布有较高的相关性。为了说明这一现象，我们选择图 2(c) 的解码步骤作为示例，在图 3 展示了在不同模型生成概率区间的 F 与 O 的 Pearson 相关系数。图 3 中横坐标表示窗口中单词的起始编号，其中单词编号越小表示对应单词的模型估计概率越大，编号为 1 的单词为训练数据中词频最高的单词；纵坐标表示区间内单词对应的 F 和 O 的 Pearson 相关系数。表中不同曲线表示不同的窗口大小，以窗口“500”对应

第二十二届中国计算语言学大会论文集，第64页-第77页，哈尔滨，中国，2023年8月3日至5日。

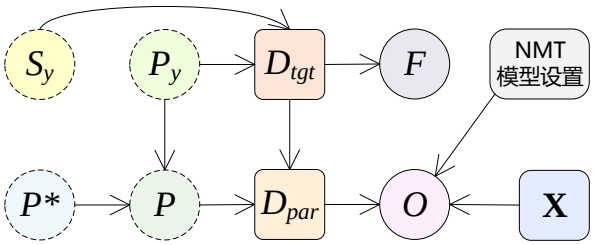


图 4: 词频与模型输出关系的因果有向图

的曲线为例，横坐标 0 对应的数据表示单词编号 1∼500 区间，而横坐标 10000 表示单词编号 10001 ∼ 10500 区间。在图 3 中，我们展示了 5 种不同的窗口大小对应的 Pearson 相关系数变化曲线，没有数据的部分表示该区间对应的 Pearson 相关系数未通过显著性检验（ p 值大于 0.01）。从图 3 可以看出，多数情况下，窗口起始位置越靠前，其对应的 Pearson 相关系数越大，即该区间内单词的 F 和 O 相关性越强。

2.3 讨论：模型估计概率与词频分布的因果关系

本小节将讨论章节 2.1 所述现象的可能原因。在解码步骤 t，给定源语言序列 $\mathbf{x}$ 和之前生成的译文序列 $\mathbf{y}_{<t}$，则目标语言单词 y_t 的条件概率分布 $P = p(y_t|\mathbf{y}_{<t}, \mathbf{x})$。$P$ 可以视为模型估计的概率分布（记作 O ）希望近似的理想值。根据贝叶斯法则，$p(y_t|\mathbf{y}_{<t}, \mathbf{x})$ 可以展开为：

$$P = p(y_t|\mathbf{y}_{<t}, \mathbf{x}) = \frac{p(y_t)p(\mathbf{y}_{<t}, \mathbf{x}|y_t)}{p(\mathbf{y}_{<t}, \mathbf{x})}。 \tag{3}$$

由于 $p(\mathbf{y}_{<t}, \mathbf{x})$ 对于任意 y_t 是确定值，由此可得到：

$$P = p(y_t|\mathbf{y}_{<t}, \mathbf{x}) \propto p(y_t)p(\mathbf{y}_{<t}, \mathbf{x}|y_t)。 \tag{4}$$

在实践中，先验概率分布 $p(y_t)$ 通常以训练数据中的词频分布 F 作为近似，这样，由公式 (3) 可以得出，条件概率分布 P 与 F 相关，该结论证明了章节 2.1 中所展示的现象的合理性。公式 (3) 表明 F 与 P 在理论上存在相关关系，尽管如此，我们仍认为在实践中 F 与 O 的相关系数要高出理论上的合理数值，也因此造成了如图 1 所呈现的问题。我们假设在训练过程中词频分布的偏差被转移到 NMT 模型中，在 NMT 的解码过程，该偏差会作为噪声干扰模型输出。

可观测性	变量名	说明
不可观测	D_{tgt}	NMT 模型训练数据中的目标语言数据
	S_y	构造 D_{tgt} 时使用的采样方法
	P_y	先验概率分布 $p(y_t)$
	P	理想的翻译概率（即 $p(y_t|\mathbf{y}_{<t}, \mathbf{x})$）
	P^*	影响 P 的其他因素，例如公式 (3) 中的 $p(\mathbf{y}_{<t}, \mathbf{x}|y_t)$
	D_{par}	平行数据，即 NMT 模型的训练数据，同时受到 D_{tgt} 和 P 的影响，这里不假设 D_{tgt} 是由源语言数据翻译而成的，例如先有目标语言数据，再经人工翻译成源语言数据，或源语言与目标语言数据均由第三语言经人工翻译而来
可观测	$\mathbf{x}$	源语言
	F	D_{tgt} 中的单词频率分布
	O	由 NMT 模型生成的估计概率，$\mathbf{x}$ 表示 NMT 模型的源语言输入
	模型设置	NMT 的模型架构、参数规模、训练算法等可能影响 NMT 性能的模型设置相关的因素

表 1: 因果有向图中各变量的物理意义

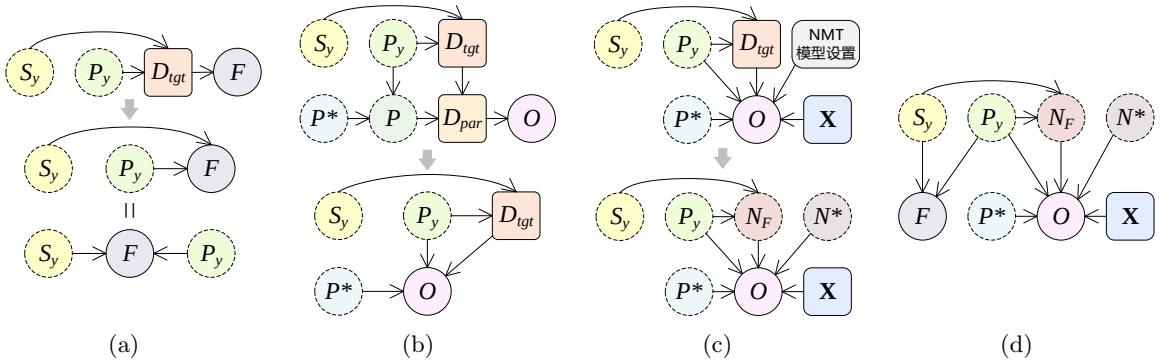


图 5: 对图 4 因果有向图的分解和简化过程

如图 4 所示，本文引入了一个因果有向无环图（Directed Acyclic Graph，DAG）说明上述假设，图 4 中各变量说明如表 1 所示。在图 4 中，S_y，P_y，P^*，P 均为不可观测的变量，无法获取其精确值。由于训练数据的源语言部分不在本文的研究范围，因此组成 D_{par} 的源语言数据、源语言分布等源语言相关因素均未在图中体现。在图 4 中，D_{par} 所示的双语平行数据，即 NMT 模型的训练数据，同时受到 D_{tgt} 和 P 影响。我们不假设 D_{tgt} 是由源语言数据翻译而成的，事实上，在语料构建过程中确实存在符合该情景的情况，例如：（1）先有目标语言数据，再经人工翻译成源语言数据，或（2）源语言与目标语言数据均由第三种语言翻译而来。

关于 F 与 O 的因果 DAG 的简化：由于本文主要关注词频分布 F 与模型输出 O 之间的关系，而图 4 所示关系涉及的变量过多，增加了研究的复杂度，因此我们在不破坏如图所示的因果关系的情况下，对图 4 进行进一步地简化。图 5(a) ∼ 图 5(c) 展示了具体的简化过程：

(a) 图 4 中 $P_y \to D_{tgt} \to F$ 和 $S_y \to D_{tgt}$ 两条路径上不存在其它因变量，因此可以合并为 $S_y \to F \leftarrow P_y$，如图 5(a) 所示；

(b) 同理，路径 $P^* \to P \to D_{par} \to O$ 和 $D_{tgt} \to D_{par} \to O$ 也可以合并，这样则可省略中间节点 P 和 D_{tgt}，如图 5(b) 所示；

(c) 根据观测到的现象，已知在 D_{tgt} 和 “模型设置” 的共同作用下得到的 O 并不准确，于是我们将 D_{tgt} 和 “模型设置” 对 O 起负面作用的因素拆分为两个不可观测变量 N_F 和 N^*，其中 N_F 表示与词频分布相关的噪声，N^* 表示其他噪声，而 D_{tgt} 和 “模型设置” 产生积极作用的部分则已经包含在路径 $P_y \to O$ 中，上述简化过程如图 5(c) 所示。

综合上述简化过程，图 4 经上述过程简化后的因果 DAG 如图 5(d) 所示，图中可观测的变量为 F、O 和 $\mathbf{x}$，其余均为不可观测的变量。由于 $\mathbf{x}$、N^* 与文本的研究内容无关，因此我们将二者合并为无关变量的集合，记作 U，这样得到最终的简化后的因果模型如图 6 所示。

3 方法

本文认为 NMT 模型解码过程中各步骤的估计概率 O 与训练数据的词频分布 F 之间的相关关系超出了合理的范围，且由此造成了 NMT 译文单词多样性下降的问题。我们假设存在与

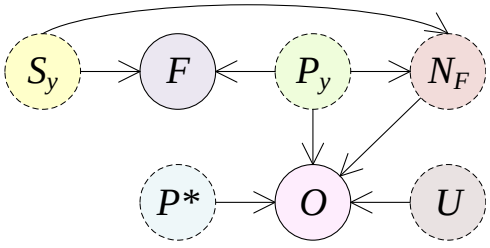


图 6: 简化后的词频与模型估计概率的因果关系图

算法 1 基于 HSR 的模型估计概率降噪方法

输入: 目标语言词汇表：V_y；NMT 模型估计的目标语言单词概率分布 O，其中 $|O| = |V_y|$；目标语言单词词频分布 F，其中 $|F| = |V_y|$；调节参数 $\alpha \in [0, 1]$。

1: 利用回归模型 $R(\cdot)$ 估计 $\mathrm{E}[O|F]$，以 F 为自变量数据，O 为因变量数据，根据最小均方误差找到回归函数参数 θ_R 的最优解：

$$\theta_R^* = \arg\min_{\theta_R} |R(F; \theta_R) - O|^2 。 \tag{7}$$

2: 从模型输出 O 中消除目标语言词频分布 F 的部分影响:

$$O' \leftarrow O - \alpha R^*(F; \theta_R^*), \tag{8}$$

其中 $R^*(\cdot)$ 表示最优化后的回归模型，θ_R^*为回归模型最优化参数。

输出: 降噪后的目标语言单词的概率分布 O'。

目标语言词频相关的噪声 N_F，该噪声造成了上述现象。由图 4 和图 6 所展示的因果关系可知，F 与 N_F 共享 S_y 和 P_y 两个原因变量，根据结构因果模型（structural causal model）中分叉结构的性质 (Pearl et al., 2016) 可知:

$$p(N_F|F, S_y, P_y) \neq p(N_F|S_y, P_y), \tag{5}$$

即 N_F 与 F 相互关联（$N_F \perp\!\!\!\perp F$）。基于 N_F 与 F 的关联性，本文试图通过控制 F 对 O 的作用，以减弱 N_F 对 O 造成的负面影响。

本文采用半同胞回归（Half-Sibling Regression，HSR）(Schölkopf et al., 2016) 的方法控制噪声 N_F 对 O 的影响。根据 HSR (Schölkopf et al., 2016)，假设存在不可观测变量 U_1 和 U_2 同时作用于可观测变量 O_1，若存在可观测变量 O_2 满足 $O_2 \not\perp\!\!\!\perp U_2$，则可通过控制 O_2 在 O_1 中产生的效应 $\mathrm{E}[O_1|O_2]$ 来控制 U_2 对O_1 的影响。在本文提出的场景中，如图 6 所示，由于 N 与 F 相关联（即 $F \not\perp\!\!\!\perp N$），通过控制 F 在 O 产生的效应，则可实现缓解噪声 N_F 对 O 的影响:

$$O' \leftarrow O - \alpha \mathrm{E}[O|F], \tag{6}$$

其中 O' 表示降噪后的模型估计概率，$\mathrm{E}[O|F]$ 表示 F 在 O 上产生的效应，在实践中，$\mathrm{E}[O|F]$ 通过回归模型估计。$\alpha \in [0, 1)$ 表示降噪系数。由图 6 可知，F 包含了 P_y 的信息，为防止执行公式 (6) 的操作时过度破坏 P_y 对 O 的影响，因此本文引入降噪系数 α 以控制降噪操作的力度。

上述方法通过对 NMT 模型解码时输出的估计概率进行降噪处理来缓解词频分布的偏置对于 NMT 模型预测精度的影响，无需更改模型设置或对 NMT 的训练进行干预，是完全模型无关（model agnostic）的方法。算法 1 展示了上述方法的操作过程。

3.1 分区间回归与自适应的降噪系数

在算法 1 中，线性回归模型在 F 与 O 的全集上优化得到（记作 SR，Set-based Regression），降噪系数 α 采用固定的常数（记作 CDR，Constant Denoising Ratio），上述 SR+CDR 的方法假设：对于不同频率的目标单词 y_i，其对应的模型估计概率 o_i 受到的与词频 f_i 相关的影响函数是一致的。然而，根据章节 2.2 中所讨论的现象，不同的目标语言词频分布 F 的区间与对应的估计概率 O 的区间的相关关系是有差异的，因此，利用 F 与 O 的全集优化得到的线性回归模型可能无法拟合不同词频区间上的真实情况，这样估计得到的 $\mathrm{E}[O|F]$ 误差较大。为缓解该问题，本文引入了分区间回归和自适应的降噪系数，以优化对降噪项（即 $\alpha\mathrm{E}[O|F]$ ）的估计，二者的具体操作方法如下：

（1）分区间回归（记作 PR，Partition-based Regression），将回归模型的训练数据 $< F, O >$ 切分成 N_R 组不同的区间 $\{< F_1, O_1 >, \ldots, < F_{N_R}, O_{N_R} >\}$，并以此分别计算得到 N_R 个回归模型 $\{R_1, \ldots, R_{N_R}\}$，则公式 (8) 将改写为：

$$O'_i \leftarrow O_i - \alpha R_i^*(F_i; \theta_{R_i}^*), \quad i \in \{1, \ldots, N_R\}。 \tag{9}$$

上述区间的划分由人工凭借经验完成，本文首先将词汇表按词频从大到小排序，采用固定的步长划分区间，其中步长等于 4,000，即编号 1 ∼ 4,000 的单词对应的词频和估计概率作为第一组数据 $< F_1, O_1 >$，号 4,001 ∼ 8,000 的单词对应的词频和估计概率作为第二组数据 $< F_2, O_2 >$，以此类推；

（2）自适应的降噪系数（记作 SADR，Self-Adaptive Denoising Ratio），即对不同的训练数据对 $< f_i, o_i > \in < F, O >$ 使用不同的降噪系数 α_i：

$$o_i' \leftarrow o_i - \alpha_i R^*(f_i; \theta_R^*), \quad i \in \{1, \ldots, V_y\}。 \tag{10}$$

本文 α 通过对 F 取自然对数的相反数再经过最大-最小值缩放（Max-Min Scaling）后得到：

$$\alpha_i = \frac{-\log f_i - \min(-\log F)}{\max(-\log F) - \min(-\log F)}, \quad i \in \{1, \ldots, V_y\}, \tag{11}$$

其中 max(·) 和 min(·) 分别表示取集合中最大值和最小值的函数。上述降噪系数的计算方法的物理意义为：对于词频较高的单词取较小的降噪系数，降噪力度较小，反之对于低频词则取较大的降噪系数，降噪力度较大。

本文的实验部分（章节 4 ）将对比两种回归模型的优化方式（ SR 和 PR ）和两种降噪系数设置方法（ CDR 和 SADR ）共 4 种组合方式：（1）SR+CDR，基于全集的回归模型结合固定的降噪系数；（2）PR+CDR，分区间回归模型结合固定的降噪系数；（3）SR+SADR，基于全集的回归模型结合自适应的降噪系数；（4）PR+SADR，分区间回归模型结合自适应的降噪系数。在实验中，SR 设置为 0.01。

4 实验

4.1 实验数据

为验证本文提出的方法，我们在如下 4 组翻译任务上进行实验：维吾尔语到汉语翻译（维-汉），汉语到英语翻译（汉-英），英语到德语翻译（英-德）以及英语到法语翻译（英-法）。上述语料的具体信息如下：

维-汉：训练集数据、测试集数据以及验证集数据均来自于 CCMT2019 机器翻译评测维吾尔语到汉语新闻翻译任务 (Yang et al., 2019)。维吾尔语端除了词例化 (Koehn et al., 2007) 和词拆分 (Kudo and Richardson, 2018) 外，并未使用其他分词工具。

汉-英：训练集数据提取自多组 LDC 语料[1]. 本文采用 NIST'02 测试数据作为汉-英翻译的验证集数据，NIST'03∼NIST'06 数据作为汉-英翻译的测试集数据.

英-德和英-法：英-德和英-法翻译任务的训练集数据来自于 WMT'14 (Bojar et al., 2014) 的公开数据，验证集和测试集数据分别是 newstest2013 和 newstest2014。

上述四个翻译任务分别代表低资源翻译（维-汉）、中等资源翻译（汉-英、英-德）和丰富资源翻译（英-法）等三类翻译任务。翻译任务的数据均经过 Moses (Koehn et al., 2007) 的 *tokenizer.perl* 脚本[2]进行词例化，之后由 sentence-piece (Kudo and Richardson, 2018) 工具进行词拆分以缩小词表。对于维-汉和汉-英翻译任务，汉语端文本在词例化之前先利用 LTP (Che et al., 2010) 中文分词工具进行分词。上述翻译任务数据经预处理后的统计信息如表 2 所示。

	维-汉	汉-英	英-德	英-法
平行句对的数目	0.17M	1.3M	4.5M	18M
源语言词表规模	32K	37K	37K	30K
目标语言词表规模	27K	33K	37K	30K

表 2: 各翻译任务训练数据的统计信息

[1]包括 LDC2005T10，LDC2003E14，LDC2004T08 以及LDC2002E18. 其中 LDC2003E14 是文档级别对齐语料，我们采用Champollion 句对齐工具 (Ma, 2006) 从中提取平行句对.

[2]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

4.2 模型设置与评价指标

基线模型设置：本文采用 Transformer（base） (Vaswani et al., 2017) 作为 NMT 模型的具体实现。参考 Vaswani et al. (2017)，训练完成后，我们选择在验证集上表现最佳的 5 个模型存储点并求输出层平均，得到最终的测试用模型。在测试时，NMT 的解码采用柱搜索（beam search）方法，柱搜索的宽度为 4。

机器翻译评价指标：机器翻译采用大小写不敏感的 BLEU (Papineni et al., 2002) 作为评价指标，本文实验采用 Moses (Koehn et al., 2007) 提供的 *multi-bleu.perl* 脚本[3]对译文进行打分。对于维-汉翻译，译文评价是在中文字符级别上进行的 BLEU 评分。

词级别译文多样性评价指标：本文从（1）生成译文中独立的 N 元文法（N-gram）的占比（记作 *Dist-N*）和（2）译文的词频分布（Mean 和 Median）两个角度评价机器翻译译文词级别多样性，其计算方法如下：

（1）*Dist-N*：生成译文中独立的 N 元文法在全部 N 元文法的总数目中的占比 (Li et al., 2016a)，其中 $N \in \{1, 2, 3, 4\}$，该评价指标的数值越高，表示译文多样性程度越高；

（2）Mean 和 Median：本文通过计算译文词汇的平均词频（记作 Mean）和词频的中位数（记作 Median）对译文的词频分布情况进行量化，该指标对应的数值越低，表示译文受到词频偏置的影响越小。本文所有的词频或概率值均以自然对数数值的形式进行展示。

4.3 机器翻译实验结果

本文提出的方法旨在不影响译文质量的情况下提升机器译文的词汇多样性。为验证该假设，本文在四个代表着不同资源丰富度的机器翻译任务上进行了实验，实验结果在表 3 和表 4 中展示。在表 4 中，“汉-英”列对应的数据表示将 NIST03~06 合并后再计算得到的 BLEU 值。从表 3 和表 4 可以看出，本文提出的 4 种方法生成的译文相对于基线模型在 BLEU 指标上均有所提升，说明本文提出的方法在不改变模型设置和训练方法的情况下，仅通过在解码时对模型的输出层降噪即可提升译文质量。

另一方面，从表 3 和表 4 可以看出，“+SADR”的方法通常能得到最佳的 BLEU 评分，证明了本文提出的自适应降噪系数的有效性，同时侧面反映了模型输出 O 和词频 F 之间的关联确实是与词频所在的区间相关的。

模型	NIST03	NIST04	NIST05	NIST06
Transformer	42.73	45.76	43.53	44.34
+SR+CDR	42.87	45.77	43.85	44.48
+PR+CDR	42.86	45.76	43.81	44.51
+SR+SADR	43.27	**45.79**	44.00	**44.52**
+PR+SADR	**43.37**	45.77	**44.02**	44.36

表 3: 汉-英翻译任务 4 组测试数据的译文 BLEU 对比

模型	汉-英	维-汉	英-德	英-法
Transformer	44.34	39.71	27.33	40.08
+SR+CDR	44.57	39.73	27.39	40.09
+PR+CDR	44.56	39.74	27.43	40.09
+SR+SADR	**44.73**	39.78	**27.61**	**40.17**
+PR+SADR	44.71	**39.83**	**27.61**	40.15

表 4: 不同方法在四个翻译任务上的生成的译文 BLEU 对比

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

模型	*Dist*-*N*				词频分布	
	Dist-1	*Dist*-2	*Dist*-3	*Dist*-4	平均数	中位数
维-汉测试集	.1572	.6710	.9039	.9601	-4.8390	-7.6416
Transformer	.1452	.6460	.9026	.9663	-4.8259	-7.5277
+SR+CDR	.1468	.6517	.9046	.9669	-4.8371	-7.6088
+PR+CDR	.1468	.6520	.9051	.9671	-4.8380	-7.6088
+SR+SADR	.1503	.6560	.9063	**.9674**	-4.8421	-7.6278
+PR+SADR	**.1516**	**.6578**	**.9066**	.9673	**-4.8472**	**-7.6406**
汉-英测试集	.0697	.4544	.7758	.9058	-4.6298	-7.4397
Transformer	.0593	.3969	.7217	.8806	-4.4997	-7.0097
+SR+CDR	.0601	.4016	.7258	.8827	-4.5093	-7.0893
+PR+CDR	.0601	.4015	.7255	.8824	-4.5088	-7.0893
+SR+SADR	.0616	.4078	.7314	.8856	-4.5169	-7.1433
+PR+SADR	**.0621**	**.4113**	**.7350**	**.8880**	**-4.5218**	**-7.2065**
英-德测试集	.1522	.6565	.9072	.9727	-4.9994	-7.9051
Transformer	.1459	.6252	.8901	.9684	-4.9290	-7.6590
+SR+CDR	.1467	.6295	.8921	.9691	-4.9346	-7.6939
+PR+CDR	.1467	.6295	.8921	.9690	-4.9347	-7.6939
+SR+SADR	.1491	**.6365**	**.8950**	**.9697**	-4.9426	**-7.8249**
+PR+SADR	**.1492**	**.6365**	**.8950**	**.9697**	**-4.9428**	**-7.8249**
英-法测试集	.1055	.4878	.7811	.9117	-4.5886	-6.8775
Transformer	.1020	.4733	.7697	.9057	-4.5824	-6.7374
+SR+CDR	.1026	.4782	.7745	.9089	-4.5921	-6.8319
+PR+CDR	.1034	**.4822**	**.7781**	**.9108**	**-4.5988**	**-6.9009**
+SR+SADR	**.1040**	.4816	.7773	.9106	-4.5960	-6.8775
+PR+SADR	.1039	.4817	.7773	.9106	-4.5959	-6.8775

表 5: 不同方法生成译文的词汇多样性量化指标对比

4.4 词级别译文多样性实验结果

表 5 给出了本文提出的方法在词级别多样性评价指标下的实验结果。其中 *Dist*-*N* 表示数据中的独立 *N*-gram 的占比，是直观展示词汇多样性的指标，该指标数值越高，则表示用词越丰富；而“词频分布”则可以看出数据中的词频组成情况，其中低频词越多，可以侧面反映出词汇多样性越高。

从表 5 可以看出，在全部四个翻译任务上，本文提出的方法相对于基线模型在“*Dist*-*N*”和“词频分布”两类指标上均有所提升。在维-汉、汉-英和英-德翻译任务上，“+PR+SADR”方法在多数评价指标上效果最佳。此外“+PR”和“+SADR”方法得到的实验结果均优于采用固定降噪系数的方法（“+CDR”），说明本文提出的分区间回归和自适应的降噪系数可以很好的应对章节 2.2 所展示的问题。

4.5 不同方法解码效率对比

为展示引入本文提出的降噪方法对于 NMT 模型解码速率的影响，本节以英-德翻译任务为例，展示了不同方法下 NMT 模型的解码速率。速率的测量单位是“词/秒”，即每秒钟机器翻译模型生成词数。实验采用的软件平台是 Centos 7.5.1804+Python 3.9.2+PyTorch 1.12.1+CUDA 12.0，硬件方面，处理器为英特尔 E5-2650 v4，显卡型号为 Nvidia GTX 1080 Ti。所有实验均在硬件空闲时完成，且实验结果取 3 次重复实验的平均值。

	Transformer	+SR+CDR	+PR+CDR	+SR+SADR	+PR+SADR
解码速率（词/秒）	1366.13	1251.67	1049.10	1211.57	1047.13
解码速率比	1.00	0.92	0.77	0.89	0.77

表 6: 不同方法解码效率对比

实验结果如表 6 所示，“速度比”表示各个方法的速率与基线模型 Transformer 的解码速率的比值，以便于更好的展示解码速率之间的差异。从表 6 中可以看出，本文提出的降噪方法对 NMT 的解码速率虽有负面影响，但影响较小，具体的解码速率损失在 10% ∼ 25% 之间，另外，从表中可以看出，“+SADR” 方法的解码效率较高，且 “+SADR”在译文 BLEU 和译文词汇多样性等指标上综合表现较好，因此，从实验中得到的多个指标综合考虑，推荐使用 “+SADR” 的方法。

4.6 译文案例

本节给出了本文提出方法与基线模型的译文案例对比，如表 7 所示。该案例来自于中-英翻译任务测试数据 NIST'06，其中“源语言”表示输入的源语言句子，“参考译文1”表示数据集中四个参考译文中的一个。表 7 在各机器译文的下方给出了句子级的 BLEU 评分，并标出了译文用词的主要差异，以及差异词在训练数据中的词频的对数（$\log f_{req}(\cdot)$）。

从表 7 可以看出，本文提出的方法生成的译文与基线系统生成的译文在句子结构上是相似的，唯一区别在于对源语言单词“发表”的翻译时选词：“made” 和 “delivered”。如表 7 所示，基线系统（“Transformer”）选择了更高频的单词 “made”（做，作出）作为对“发表”的翻译，而本文提出的方法则选择了相对低频但更符合语境的单词 “delivered”（发表），而该单词也与参考译文中选择的单词一致。

方法	译文
源语言	澳大利亚 总理 霍华德 还 在 追悼 仪式 中 发表 了 讲话 。
参考译文 1	Australian Prime Minister Howard also delivered a speech at the memorial service .
Transformer	Australian Prime Minister Howard also made a speech at the memorial ceremony .
	$\log f_{req}$(“made”) = −7.11, 句子级 BLEU：59.23
+SR+CDR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24
+PR+CDR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24
+SR+SADR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24
+PR+SADR	Australian Prime Minister Howard also delivered a speech at the memorial ceremony .
	$\log f_{req}$(“delivered”) = −9.84, 句子级 BLEU：84.24

表 7: 译文案例

5 相关工作

译文多样性受限是 NMT 的经典问题之一，早期对译文多样性研究主要关注句子级译文多样性，旨在最大化同一输入源文对应的众多候选译文之间的差异，并尽可能地保证译文质量 (Li et al., 2016b; Wu et al., 2020; Sun et al., 2020; Lin et al., 2022)。近年来， NMT 受训练数据偏见影响的问题逐渐受到越来越多的关注，其中代表性的问题包括 NMT 性别偏见 (Stanovsky et al., 2019; Costa-jussà et al., 2022) 和个性化翻译(Lin et al., 2021)等。关于数据偏见的重要表现形式之一是 NMT 训练数据的词频分布不平衡问题，这使得低频词翻译一直是 NMT 面临的挑战(Koehn and Knowles, 2017)。

针对上述问题，早期的研究工作主要关注引入更细的翻译单元或优化词汇表 (Sennrich et al., 2016; Lee et al., 2017; Kudo, 2018)，通过将单词拆分成成更小的单元从而减少低频词的个数，调整输入单元的频率分布。此外，通过消除 NMT 模型词向量中与词频相关的信息，也可以达到缓解词频分布影响的问题： Yang and Liu (2020) 采用 HSR (Schölkopf et al., 2016) 方法静态消除词向量中的词频信息，而 Gong et al. (2018)和 Liu et al. (2020) 则分别通过对抗训练和课程学习（curriculum learning）的方法在训练过程消除词向量中的词频信息。最近，一些方法根据目标语言词频 (Gu et al., 2020)和双语互信息 (Xu et al., 2021)通过自适应权重的损失函数来缓解这个问题。类似地 Zhang et al. (2022) 则提出了一种标记级对比学习方法，并引入了频率感知软权重来自适应地对比目标词的表示。

上述前人的工作需要在 NMT 的训练过程或训练之前介入，无法对已经存在的优化好的模型使用。而本文提出的方法无需修改 NMT 模型的架构和训练模式，并且是模型无关的，适用于目前已知的各类 NMT 模型。

6 结论

本文针对 NMT 训练时易受训练集词频分布偏置影响，从而导致模型输出译文用词多样性受限的问题，提出了一种基于半同胞回归的 NMT 模型估计概率 O 去噪方法，通过从 O 中消除与目标语言词频 F 相关的部分信息，从而实现缓解 O 受数据偏置影响的问题。上述方法是模型无关的，且从理论推导而来，具有完整的可解释性。在四种不同规模数据翻译任务的实验结果表明，本文提出的方法可以在不损坏译文质量的情况下，提升译文的用词多样性。

在未来的工作中，我们将继续继续分析模型估计概率与词汇分布的相关性与训练数据规模以及训练轮次之间的关系，即探索本文提出的方法在不同模型规模和训练数据规模下的适用性。另外，我们也将尝试为其他自然语言生成任务引入本文提出的方法。

致谢

感谢所有匿名审稿人的宝贵意见，由于会议论文的篇幅有限，无法将这些意见悉数吸纳到此版本的论文中，因此，未及时采纳的意见将呈现在未来版本的论文中。本文的工作得到了国家重点研发计划（批准号：2017YFB1002103）的资助。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese language technology platform. In *Coling 2010: Demonstrations*, pages 13–16, Beijing, China, August. Coling 2010 Organizing Committee.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October. Association for Computational Linguistics.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11855–11863. AAAI Press.

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2191–2197, Online, July.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

ChengYue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. FRAGE: frequency-agnostic word representation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1341–1352.

Andreas Grivas, Nikolay Bogoychev, and Adam Lopez. 2022. Low-rank softmax can have unargmaxable classes in theory but rarely in practice. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6738–6758, Dublin, Ireland, May.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, Online, November. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 10.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4008–4018, Online, August. Association for Computational Linguistics.

Huan Lin, Baosong Yang, Liang Yao, Dayiheng Liu, Haibo Zhang, Jun Xie, Min Zhang, and Jinsong Su. 2022. Bridging the gap between training and inference: Multi-candidate optimization for diverse neural machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2622–2632, Seattle, United States, July. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online, July.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley.

Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. 2016. Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. USA*, 113(27):7391–7398.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Zewei Sun, Shujian Huang, Hao-Ran Wei, Xinyu Dai, and Jiajun Chen. 2020. Generating diverse translation by manipulating multi-head attention. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8976–8983. AAAI Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Xuanfu Wu, Yang Feng, and Chenze Shao. 2020. Generating diverse translation from model distribution with dropout. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1088–1097, Online, November. Association for Computational Linguistics.

Yangyifan Xu, Yijin Liu, Fandong Meng, Jiajun Zhang, Jinan Xu, and Jie Zhou. 2021. Bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 511–516, Online, August.

Zekun Yang and Tianlin Liu. 2020. Causally denoise word embeddings using half-sibling regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9426–9433, Apr.

Muyun Yang, Xixin Hu, Hao Xiong, Jiayi Wang, Yiliyaer Jiaermuhamaiti, Zhongjun He, Weihua Luo, and Shujian Huang. 2019. Ccmt 2019 machine translation evaluation report. In Shujian Huang and Kevin Knight, editors, *Machine Translation*, pages 105–128, Singapore. Springer Singapore.

Tong Zhang, Wei Ye, Baosong Yang, Long Zhang, Xingzhang Ren, Dayiheng Liu, Jinan Sun, Shikun Zhang, Haibo Zhang, and Wen Zhao. 2022. Frequency-aware contrastive learning for neural machine translation. pages 11712–11720.

基于语音文本跨模态表征对齐的端到端语音翻译

周国江[1,2]，董凌[1,2]，余正涛*[1,2]，高盛祥[1,2]，王文君[1,2]，马候丽[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1845716340@qq.com,46761956@qq.com,ztyu@hotmail.com,
gaoshengxiang.yn@foxmail.com,175360805@qq.com,1341584939@qq.com

摘要

端到端语音翻译需要解决源语言语音到目标语言文本的跨语言和跨模态映射，有限标注数据条件下，建立语音文本表征间的统一映射，缓解跨模态差异是提升语音翻译性能的关键。本文提出语音文本跨模态表征对齐方法，对语音文本表征进行多粒度对齐并进行混合作为并行输入，基于多模态表征的一致性约束进行多任务融合训练。在MuST-C数据集上的实验表明，本文所提方法优于现有端到端语音翻译跨模态表征相关方法，有效提升了语音翻译模型跨模态映射能力和翻译性能。

关键词： 端到端语音翻译；跨模态；多任务；表征对齐

End-to-end Speech Translation Based on Cross-modal Representation Alignment of Speech and Text

Guojiang Zhou[1,2],Ling Dong[1,2],Zhengtao Yu[1,2],Shengxiang Gao[1,2], Wenjun Wang[1,2],Houli Ma[1,2]
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming 650500, China
1845716340@qq.com,46761956@qq.com,ztyu@hotmail.com,
gaoshengxiang.yn@foxmail.com,175360805@qq.com,1341584939@qq.com

Abstract

End-to-end speech translation aims to address cross-language and cross-modal mapping from source language speech to target language text. Under the limitation of labeled data, establishing an unified mapping between speech and text representation and alleviating cross-modal differencials become the keys to improve speech translation performance. In this paper, we propose a cross-modal representation alignment method for speech and text. The representation of speech and text are aligned at multiple granularities and mixed as parallel inputs to model, and multi-task training is performed based on the consistency constraint of multi-modal representation. Experiments on MuST-C dataset show that the proposed method outperforms existing related methods in end-to-end speech translation, and it effectively improves the cross-modal mapping capability and speech translation performance.

Keywords: end-to-end speech translation , cross-modal , multitasking , representation alignment

*余正涛（通信作者）：ztyu@hotmail.com
基金项目：国家自然科学基金（U21B2027, 61972186）；云南高新技术产业发展项目（201606）；云南省重大科技专项计划（202103AA080015）；云南省基础研究计划（202001AS070014）；云南省科技人才与平台计划（202105AC160018）

1 引言

端到端语音翻译任务将源语言语音直接翻译为目标语言文本，在多语言视频字幕、多语言会议同传等场景中具有广阔的应用前景。相较于先对源语言语音进行识别再翻译为目标语言文本的级联系统，端到端语音翻译系统具有更低的延迟和更少的参数量，避免了错误传播问题(Bérard et al., 2016)，因此备受研究者关注(Inaguma et al., 2020; Wang et al., 2020; Zhao et al., 2020)。

目前，面向端到端语音翻译任务的标注数据相对较少，有限标注数据条件下，输入语音和输出文本间的模态差异在较大程度上影响着语音翻译模型的性能(Liu et al., 2020)。这种模态差异主要表现在：语音长度远远大于其对应的文本长度，同时语音和文本的结构不同，语音是连续的时序信号，而文本是离散的符号序列，导致模型难以学习到语音和文本的对齐关系(Xu et al., 2021)。目前端到端语音翻译大多利用机器翻译、语音识别领域中较为丰富的数据通过预训练(Weiss et al., 2017; Alinejad and Sarkar, 2020; Stoian et al., 2020)，多任务训练(Tang et al., 2021; Ye et al., 2021)，知识蒸馏(Liu et al., 2019; Inaguma et al., 2021)等方式进行语音翻译辅助训练。然而机器翻译中的训练数据仅为文本模态，语音识别中的训练数据并不具备跨语言特性，故使用这类数据进行语音翻译辅助训练易导致编解码器跨模态映射能力不匹配(Cheng et al., 2022)，因此，如何有效缓解语音和文本之间的模态差异，提升语音翻译模型的跨模态映射能力是端到端语音翻译任务面临的一个重要问题。

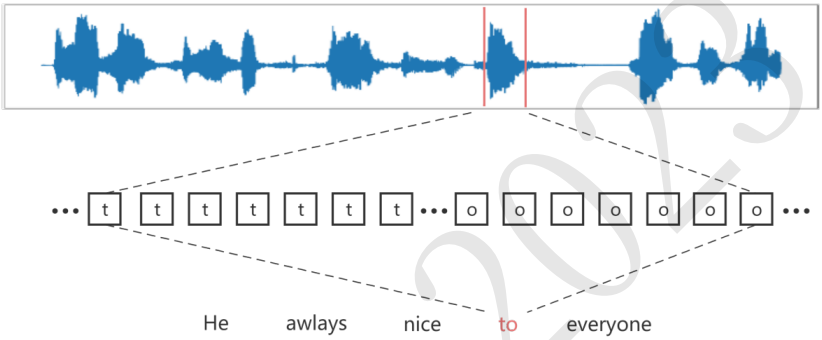


图 1. 语音与文本间的长度与结构差异

近期Liu et al. (2020)等人的工作表明，实现语音和文本的跨模态表征有助于减缓端到端语音翻译中文本与语音间的模态差异。Wei et al. (2022)和Yao et al. (2022)使用不成对的语音文本数据进行跨模态预训练以统一语音和文本的编码表示，这种多模态预训练模型相较于单模态预训练模型取得了更好的结果；Wang et al. (2022)提出了一种离散跨模态对齐方法，利用一个共享的离散词汇空间来容纳和匹配语音和文本模态，利用非平行数据对有效提升了端到端语音翻译的性能。Ye et al. (2022)提出利用对比学习对长度不一致的语音与文本表征进行约束，在此基础上Ouyang et al. (2022)通过对比学习来统一语音与文本两种模态的词级表征，证明了在不同粒度上进行跨模态一致性表征的可行性。针对语音与文本因长度造成的模态差异，Zeng et al. (2022)通过引入预测词边界任务使模型学习到语音与文本间词级的长度对齐信息，Han et al. (2021)将语音与文本表征映射为统一且固定大小的抽象表征，有效统一了语音与文本的表征空间，但固定长度的表征空间可能限制了模型的表达。Fang et al. (2022)通过混合语音与文本的浅层表征作为并行输入，增强数据的同时缓和了模态差异，但其混合表征与语音表征之间仍存在较大的长度差异。

本文提出词级和句子级的语音文本表征对齐方法，在对齐基础上对语音和文本表征进行交叉混合得到混合表征作为模型输入，将语音与文本映射到同一表征空间。针对语音和文本表征因长度造成的模态差异，使用长度归一化融合模块对混合表征与语音表征进行长度统一。针对不同模态的表征内容差异问题，在多任务联合训练框架对语音文本模态表征进行一致性约束。从而实现语音表征和对应文本表征的跨模态对齐，从而提升模型跨模态映射能力。

本文主要贡献如下：（1）本文提出的语音文本表征对齐方法和多粒度表征混合方法，验证了跨模态一致性表征对语音翻译的积极作用。（2）本文提出长度归一化融合方法，验证了其对缓解语音文本模态间长度差异的有效性。（3）在MuST-C数据集上的实验表明，在相同数据条

件下，本文所提方法优于现有端到端语音翻译跨模态表征相关方法。

2 方法

模型由声学编码器、语音文本多模态表征对齐模块、翻译解码器构成，训练阶段语音通过声学编码器得到语音表征，文本通过嵌入得到文本表征，语音表征与文本表征通过多粒度混合得到句级和词级的混合表征。语音文本多模态表征对齐模块由长度归一融合模块、共享语义编码层、门控融合模块组成，长度归一化模块用于统一各表征的长度以减轻混合表征与语音表征之间由长度导致的模态差异，共享语义编码层用于提取抽象语义表征，门控融合模块则将经过编码模块的两个混合表征再次进行融合以降低解码难度，经对齐的语音表征与混合表征并行输入翻译解码器进行解码，输出目标语言文本词序列，模型架构如图 2所示。

训练使用的数据集包含语音、转录文本以及翻译文本三元组数据，记为D(s，x，y)。模型训练分为预训练与多任务训练两阶段，在预训练阶段，使用源语言语音s和目标语言文本y对共享语义编码层和翻译解码器进行文本翻译预训练，在多任务训练阶段使用转录文本x作为辅助输入，在推理阶段以源语言语音s作为输入，不使用转录文本x。

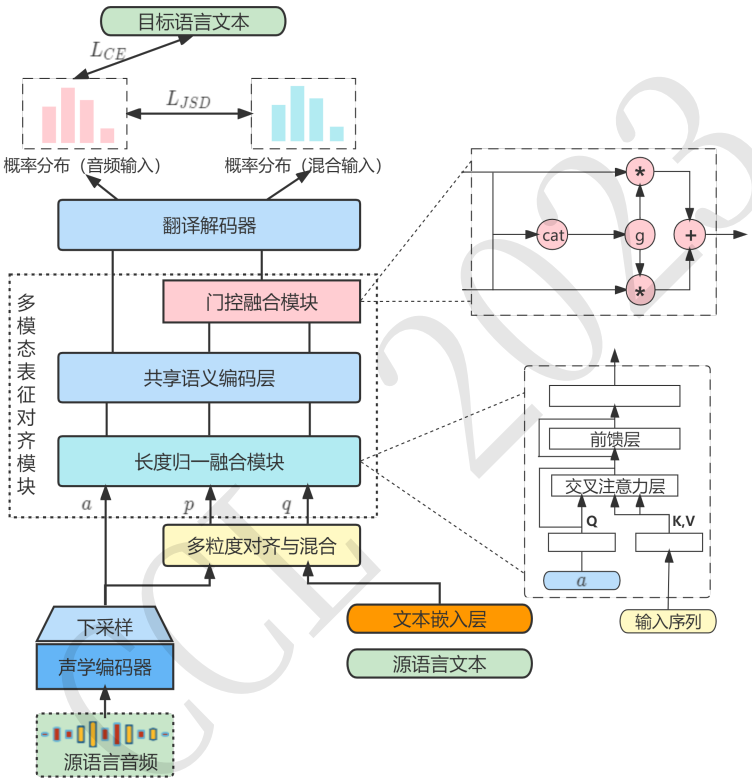


图 2. 基于语音文本跨模态表征对齐的端到端语音翻译

2.1 文本嵌入与声学编码

源语言语音s经声学编码器编码得到语音表征作为长度归一融合模块的输入。转录文本x经文本嵌入得到文本表征，与语音表征进行对齐并混合后作为模型并行输入。

声学编码器:主要用于提取语音信号的浅层表征。近期的一些研究(Gállego et al., 2021)表明使用经预训练的Hubert(Hsu et al., 2021)作为声学编码器可以提高语音翻译的性能，因此本文采用与其相同的声学编码方式$Hubert(\cdot)$。基于Ye et al. (2021)的研究，本文在Hubert的基础上加入两个卷积层对输出的语音表征进行下采样$DS(\cdot)$。编码过程如式 (1)所示，下采样倍数为4，编码后得到语音表征序列$a = [a_1, a_2, ..., a_{l_a}]$，其中$a \in \mathbb{R}^{d_a}$,$d_a$为语音表征维度，$l_a$为序列长度。

$$a = Ds(Hubert(\mathrm{s})) \tag{1}$$

文本嵌入:如式 (2)所示，对于训练时的文本输入，使用Unigram SentencePiece[1] 学习双语词表，并对文本输入x进行编码$Unigram(\cdot)$，编码后经文本嵌入$Emb(\cdot)$得到文本表征$e = [e_1, e_2, ..., e_{l_e}]$，其中$e \in \mathbb{R}^{d_e}$,$d_e$为文本嵌入表征维度，$l_e$为序列长度。

$$e = Emb(Unigram(\mathrm{x})) \tag{2}$$

2.2 语音文本表征的对齐与混合

针对语音长度远大于其对应的文本长度，导致对齐关系难以学习的问题，本节将语音与文本表征在句子级和词级上进行混合作为多模态表征对齐模块的并行输入，让模型学习到跨模态的词级和句子级对齐信息，混合过程如图 3所示。

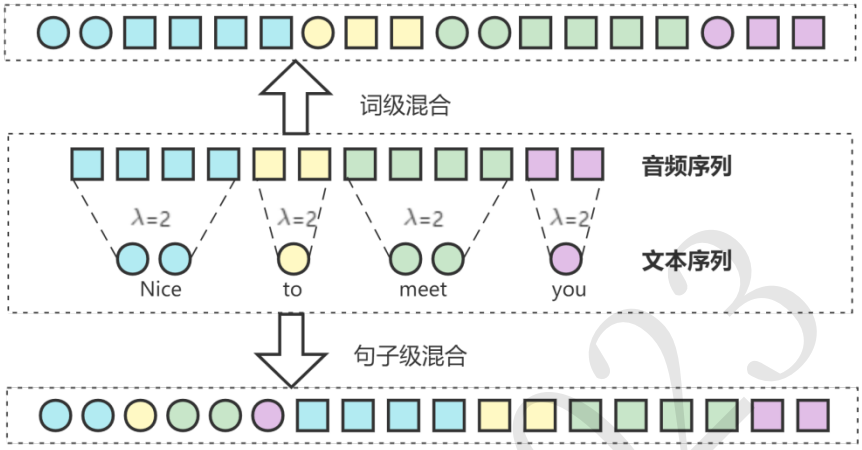


图 3. 语音文本表征的对齐与混合

词级语音文本表征混合：将编码后的语音表征序列a与文本表征序列e按一定比例切割对齐，如式 (3)，先计算出训练集中所有语音编码表征序列长度l_a与文本嵌入序列长度l_e的数学期望之比λ ，$E(\cdot)$用于计算数学期望，$Int(\cdot)$用于向下取整。

$$\lambda = Int(E(l_a)/E(l_e)) \tag{3}$$

对于文本表征序列e中的任意单词序列e_j，如式 (4)，使用λ进行对齐得到其对应语音表征序列的起止位置u_j和v_j，$|\cdot|$用于计算单词序列长度。

$$u_j = \sum_{\mathrm{i}=1}^{\mathrm{j}-1} \lambda \left|\mathrm{e_i}\right|, \quad v_j = u_j + \lambda \left|\mathrm{e_j}\right| = \sum_{\mathrm{i}=1}^{\mathrm{j}} \lambda \left|\mathrm{e_i}\right| \tag{4}$$

根据位置信息u_j和v_j对语音进行对齐得到m_j。对整个语音表征序列进行对齐后可表示为$m = [m_1, m_2, ..., m_{l_m}]$，其中$m \in \mathbb{R}^{d_a}$,$l_m$为序列长度。

$$m_j = \begin{cases} \left[a_{u_j} : a_{v_j}\right] & v_j \le l_a \text{ and } j \le l_e \\ \left[a_{v_{j-1}} :\right] & j > l_e \\ \left[a_{u_j} :\right] & v_j > l_a \end{cases} \tag{5}$$

如式 (5)所示，对齐后的语音表征序列m_j为原音频表征序列a中对应起始位置u_j与终止位置v_j之间的序列。当最后一个单词序列对应的终止位置小于l_a时，对于剩余还未对齐的音频序列，有$j > l_e$，取剩余音频序列作为m_j。当终止位置v_j大于语音表征序列长度l_a时，取起始位置u_j之后的所有序列作为m_j，若$u_j > l_a$，则a_{u_j}与m_j均为空序列。

将对齐后的序列混合后进行拼接$Concat(\cdot)$，得到词级混合表征p，过程如式 (6)所示。

[1]https://github.com/google/sentencepiece

$$p = Concat(\mathrm{m}_1, \mathrm{e}_1,\ \mathrm{m}_2, \mathrm{e}_2 \ldots, \mathrm{m}_{l_m}, \mathrm{e}_{l_e}) \tag{6}$$

句子级语音文本表征混合：句子级混合不需要对齐,混合过程如式 (7)所示，q表示句子级混合表征，词级与句子级混合表征序列长度相同。

$$q = Concat(\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_{l_a}, \mathrm{e}_1, \mathrm{e}_2 \ldots, \mathrm{e}_{l_e}) \tag{7}$$

2.3 多模态表征对齐模块

语音文本多模态表征对齐模块由长度归一融合模块、共享语义编码层、门控融合模块组成。训练阶段以语音表征a和两种混合表征p，q作为输入，通过长度归一化模块映射为具有相同维度的表征，经共享语义编码层得到抽象语义表征，词级与句级的混合语义表征通过门控融合模块融合为多粒度语义表征。

长度归一融合模块：本文使用长度归一的融合模块再次对语音表征与混合表征进行融合,融合编码方法为多头交叉注意力$CMHA(\cdot)$ 。如式 (8)所示，语音表征a、词级混合表征p、句子级混合表征q经长度归一融合后分别得到融合表征h_a，h_p，h_q，其中W_q，W_k，W_v均为随机初始化的参数矩阵,输入Q始终为语音表征a，K、V则为对应的输入表征。

$$h_{out} = CMHA(QW_q, KW_k, VW_v) \tag{8}$$

共享语义编码层：本文使用的编码层遵循Transformer编码层的结构，层数为6，每一层都包含一个自注意、残差、前馈和归一化模块。如式 (9)所示，共享语义编码层$encoder(\cdot)$的输入h_{input}分别为长度归一融合模块的输出h_a、h_p、h_q，经语义编码后得到对应的抽象语义表征h_a^A、h_p^A、h_q^A。

$$h_{out}^A = encoder(h_{input}) \tag{9}$$

门控融合模块：该模块将词级与句级的混合语义表征h_p^A、h_q^A进一步融合，综合两种混合表征的特点，同时降低翻译解码器解码压力。如式 (10)所示，$*$表示计算矩阵乘法，先将h_p^A、h_q^A在隐层维度拼接，使用可学习的W_g进行线性映射得到门控单元系数γ，γ的隐层维度为1，激活函数$\sigma(\cdot)$为sigmoid，最后使用γ对h_p^A、h_q^A进行融合得到多粒度融合表征h_g。

$$\gamma = \sigma(Concat(h_p^A, h_q^A) * W_g), h_g = \gamma * h_p^A + (1-\gamma) * h_q^A \tag{10}$$

2.4 翻译解码器

解码器由6层transfomer解码层构成。语音输入s经声学编码器得到语音表征，通过长度归一融合模块和共享语义编码层得到抽象语义表征，再通过翻译解码器生成目标语言的单词序列，简化表示为h(s)。同理，语音输入s与其转录文本输入y计算得到多粒度融合表征，经翻译解码器生成目标语言的单词序列，简化表示为h(s, x)。对于输入序列s和期望输出y，我们定义损失函数如式 (11)。

$$\mathrm{L_{CE}}(\mathrm{h(s)}, \mathrm{y}) = \sum_{\mathrm{i=1}}^{\mathrm{|y|}} \log\left(\mathrm{P}_\theta\left(\mathrm{y_i} \mid \mathrm{y_{<i}}, \mathrm{h(s)}\right)\right) \tag{11}$$

使用交叉熵损失作为语音输入得到结果与目标语言文本的损失，语音表征和融合表征之间使用Jensen-Shannon散度$JSD(\cdot)$计算得到一致性约束损失，下文简称为JSD损失，计算过程如式 (12)所示。本文没有引入混合表征作为输入时其结果与目标语言文本的约束损失，使模型更加关注以语音作为输入时的结果，同时降低模型拟合的难度。

$$\mathrm{L_{JSD}}(\mathrm{h(s)}, \mathrm{h(s,x)}, \mathrm{y}) = \sum_{i}^{\mathrm{|y|}} JSD\left(\mathrm{P}_\theta\left(\mathrm{y_i} \mid \mathrm{y_{<i}}, \mathrm{h(s)}\right), \mathrm{P}_\theta\left(\mathrm{y_i} \mid \mathrm{y_{<i}}, \mathrm{h(s,x)}\right)\right) \tag{12}$$

综上，微调阶段总损失L(s, x, y)如式 (13)所示，β为JSD损失权重系数。

$$L(s,x,y) = L_{CE}(h(s),y) + \beta * L_{JSD}(h(s),h(s,x),y) \quad (13)$$

3 实验设置与结果分析

3.1 实验设置

3.1.1 数据集

实验使用了MuST-C(Di Gangi et al., 2019)数据集，该数据集包含英语到多个语言对数据，本文使用了英语(En)到越南语(Vi)、德语(De)、意大利语(It)、俄罗斯语(Ru)、西班牙语(Es)、法语(Fr)、罗马尼亚语(Ro)、荷兰语(Nl)和葡萄牙语(Pt)9个语言对，具体参数如表 1所示。实验中使用MuST-C中的dev集作为验证集，tst-COMMON作为测试集。

语言	En-De	En-Es	En-Fr	En-Nl	En-It	En-Pt	En-Ro	En-Ru	En-Vi
时长（h）	408	504	492	442	465	385	432	489	441
句数（k）	234	270	280	253	258	211	241	270	230

表 1. MuST-C数据集

3.1.2 数据预处理

为保证实验公平性，文本采用Fairseq(Ott et al., 2019)中对MuST-C数据集的预处理方式，并使用Unigram SentencesPiece学习得到每个语言对的源、目标语言双语共享词表，词表大小为10000。语音输入为16bit，16kHz的单通道原始语音，使用开源Hubert[2]作为声学编码器，提取语音的768维语音表征，并用两层卷积神经网络对其进行下采样，卷积核大小为5，步长为2，隐藏层维度为1024。根据语音长度与声学编码器中的维度变化计算语音经表征序列长度，与文本嵌入表征长度信息结合，计算得到λ，λ的值为3，根据λ进行对齐得到位置信息u，v。

3.1.3 模型配置与评价指标

为保证实验结果可比性，本文所做实验均基于Fairseq框架。共享语义编码层与解码层的层数均为6，多头注意力头数为8，隐层变量维度为512，前馈层网络维度为2048，dropout为0.1。

在文本翻译预训练阶段，使用源语言-目标语言对来对翻译编解码器进行训练，设置学习率为7e-4，每批可使用的序列长度最多为4k。

在多任务训练阶段，每批使用最多2M的源语音帧，学习率为1e-4，JSD权重系数β为4，设置每8个批次进行一次梯度更新以模拟使用8张显卡进行计算。为避免过拟合，设置最大训练周期为30，若验证集上的损失在十个周期内没有减少，提前停止训练。上述两个训练阶段所使用优化器均为Adam(Kingma and Ba, 2014)，设置β_1 为0.9，β_2 为0.98，交叉熵损失标签平滑率为0.1。学习率预热步长为4000，4000步后学习率将与步数的平方成反比下降。

在推理阶段，对最后十个周期得到的模型参数进行平均以用于评估。解码使用大小为5的束搜索算法，使用区分大小写的SacreBLEU[3] (Post, 2018) 作为模型性能的评价指标，所有训练过程均在1张Tesla V100 GPU上进行。

3.2 实验结果

3.2.1 与其它方法的对比实验

为验证所提方法的有效性，在MuST-C数据集上进行对比实验，实验参数与所提方法一致。选择了以下几个端到端语音翻译方法：Fairseq-ST (Wang et al., 2020) 、Chimera(Han et al., 2021)、STEMM(Fang et al., 2022)、ConST(Ye et al., 2022)、W2V2-ST，具体基线模型介绍如下：

（1）Fairseq-ST：使用语音识别任务进行预训练，在端到端语音翻译任务上微调。

（2）Chimera：引入了一个共享的语义空间映射层，将语音和文本映射成固定维度的语义表示并用于翻译。

[2]https://github.com/facebookresearch/Fairseq/tree/main/examples/hubert
[3]https://github.com/mjpost/sacrebleu

（3）STEMM：将语音和文本的浅层表征进行混合作为翻译任务的并行输入，使用自学习训练框架约束训练结果。

（4）ConST：在多任务交替训练中引入对比损失约束语音和文本表征的一致性，让语义相似的语音和文本具有相似的表示。

（5）W2V2-ST：使用Wav2vec2.0(Baevski et al., 2020)模型提取语音表征，使用transformer编解码器进行训练，结果取自(Cheng et al., 2022)。

模型	额外数据			BLEU								
	Speech	MT	MFA	Vi	De	Es	Fr	It	Nl	Pt	Ro	Ru	Avg.
Fairseq-ST	×	×	×	—	22.7	27.2	32.9	22.7	27.3	28.1	21.9	15.3	24.8
Chimera	√	√	×	—	**27.1**	30.6	35.6	25.0	29.2	30.2	24.0	17.4	27.4
STEMM	√	×	√	—	25.6	30.3	36.1	25.6	30.1	31.0	24.3	17.1	27.5
ConST	√	×	×	—	25.7	30.4	36.8	26.3	30.6	32.0	24.8	17.3	28.0
W2V2-ST	√	×	×	23.2	24.3	29.6	35.2	25.1	29.1	30.3	23.4	16.5	26.7
本文所提方法	√	×	×	**24.7***	26.6*	**31.0***	**37.2***	**26.4***	**30.8***	**32.4***	**25.1***	**17.8***	**28.4***

表 2. 在MuST-C数据集多语言对上与其它方法的比较

如表 2，*表示本文所提方法结果强于W2V2-ST基线模型结果，加粗部分表示达到了最佳翻译效果，speech表示使用外部语音数据，MT(Machine Translation)表示使用外部翻译数据，MFA[4](McAuliffe et al., 2017)表示使用外部强制对齐模型。因其它方法在英语-越南语语言对上实验数据的缺乏，本文对除越南语外的8个语言对的BLEU求均值得到Avg.结果。

本文方法与W2V-ST基线模型相比，BLEU值平均提高了1.8，验证了本文所提语音文本跨态表征对齐方法的有效性。与使用了额外文本翻译数据的Chimera相比，在可比较语言对上BLEU值平均提升了1.0，本文方法同样将语音和文本映射到了同一长度但没有将输出限制到固定大小，验证了基于语音长度归一化融合的有效性。

STEMM方法在进行混合时使用了外部强制对齐模型，本文则根据数据集本身语音与文本表征长度进行混合。相较于STEMM，本文方法进行了多粒度混合且得到的混合序列更长，但经长度一致性融合与门控融合后仅增加了少量训练时间，并在可比较语言对上取得了平均0.9 BLEU值的提升，证明了本文所提多粒度对齐与长度归一融合方法的有效性。

在相同数据条件下，本文方法较ConST方法在可比较语言对上平均BLEU值提升了0.4，ConST为使用对比损失将语音和文本进行了平均池化，一定程度上忽略了语音和文本表征间的局部差异，本文方法在使用JSD损失进行约束时并未造成局部信息损失，表明了在JSD损失约束下的多任务训练框架有效性。

3.2.2 预训练对语音翻译结果的影响

遵循Fairseq中基于端到端语音到文本的模型设置，分别基于Fbank特征和Hubert特征进行英语到越南语翻译实验作为对比，基于Fbank特征的Fairseq-ST实验参数与Wang et al. (2020)等人一致，基于Hubert特征的实验参数与上节中提到的一致。为进一步验证文本翻译预训练对结果的影响，使用CCMatrix(Schwenk et al., 2019)中英语-越南语平行语料作为额外数据进行文本翻译预训练。

模型	预训练	训练时间	推理时间	BLEU
Fairseq-ST	语音识别	1.00x	1.00x	20.8
Hubert-Transformer	Hubert	4.23x	1.83x	22.6
Hubert-Transformer	Hubert+文本翻译	4.23x	1.83x	23.4
本文方法	Hubert+文本翻译	5.52x	1.83x	24.7
本文方法+额外翻译数据	Hubert+文本翻译	5.52x	1.83x	25.4

表 3. MuST-C英语-越南语不同预训练方法BLEU值对比

[4]https://mfa-models.readthedocs.io/en/latest/index.html

在tst-COMMON测试集上的结果如表 3所示，训练时间与推理时间分别为训练与推理阶段花费的时长。在没有经过文本预训练的情况下，使用经预训练的Hubert作为声学编码器进行训练BLEU值达到了22.6，比在语音识别预训练下以Fbank作为特征输入进行训练得到的BLEU结果高出1.8，验证了Hubert作为声学编码器的有效性。在使用文本翻译进行预训练后，结果再次提高了0.8 BLEU，表明进行跨模态的预训练对端到端语音翻译是有效的。在此基础上，本文所提跨模态表征对齐方法训练得到的BLEU再次提升了1.3，验证了本文所提方法的有效性。引入额外翻译数据进行文本翻译预训练后BLEU结果达到了25.4，进一步验证了使用文本翻译进行跨模态预训练对语音翻译的积极作用。以Fairseq-ST结果1.00x为基准，使用Hubert作为声学编码器使训练时间增长为4.23x，推理时间增长为1.83x。本文方法引入了词级与句子级混合表征进行多任务训练，训练时间达到了5.52x,推理阶段只有音频作为输入，故推理时间与Hubert-Transformer方法相同。

3.2.3 不同并行输入特征的对比实验

为探究混合表征对语音翻译结果的影响，分别设置词级、句级混合表征以不同组合方式作为训练时的并行输入进行实验。

为了验证模型缓和跨模态差异的能力，使用tst-COMMON测试集中的英语语音-文本对作为输入，分别得到其经过共享语义编码层后的文本语义表征和语音语义表征，计算语音-文本语义表征对的平均余弦相似度作为特征相似度指标，它反应了模型在语义层面将不同模态表征进行共同映射的能力。

有无混合	并行输入	BLEU	特征相似度（%）
无混合	无	23.4	97.88
	文本表征	24.2	97.26
有混合	词级混合表征	24.5	98.14
	句子级混合表征	24.4	98.00
	句子+词级混合表征	24.7	98.28

表 4. 混合表征对英语-越南语实验结果影响

如表 4所示，在没有进行混合的情况下，使用文本表征作为并行任务的输入BLEU提高了0.8，表明在多任务训练框架下使用额外的翻译数据是有益的。在多任务训练框架下，特征相似度与BLEU结果成正相关，表明缓和模态差异对端到端多任务语音翻译的积极影响。利用词级或句级混合表征都能对语音翻译产生正向效果，因为混合表征与文本表征相比具有额外的语音信息和不同模态信息间的相对位置信息。相较于句子级的混合表征，使用词级的混合表征进行训练特征相似度提高了0.14%，BLEU提高了0.1，表明用词级的混合表征作为并行输入使模型能够捕捉到不同模态表征长度期望之间的联系，从而更好地缓解语音和文本在长度上造成的模态差异，提升翻译性能。相较于仅利用单一混合表征，使用多粒度的混合表征能够进一步提升语音翻译性能和特征相似度，表明同时使用词级混合表征与句子级混合表征使模型学习到更多不同模态间的位置信息，二者具有互补性。

3.2.4 长度归一融合模块不同输入下的对比实验

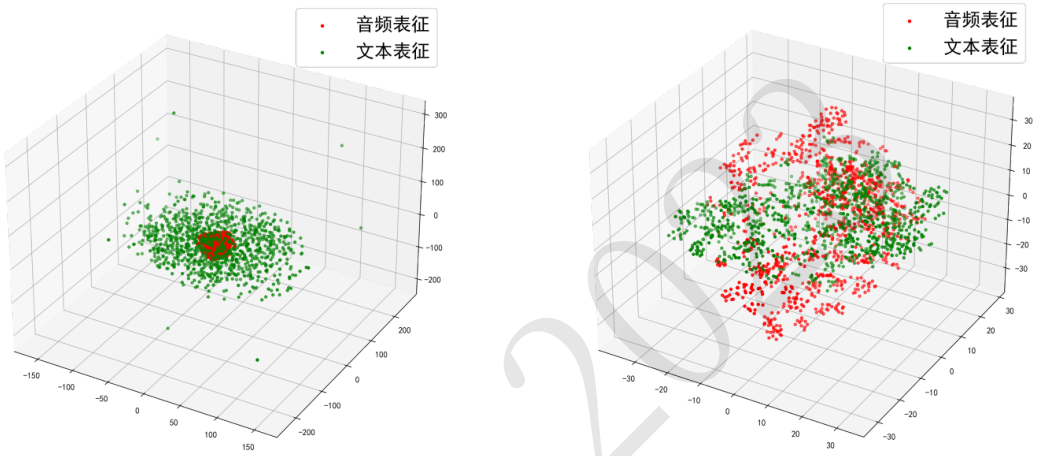
为探究融合特征与语音表征长度之间的关系对语音翻译性能的影响，在不改变主任务语音表征输入的情况下，将并行任务中长度归一融合模块的语音输入改为文本表征与语音表征分别进行了实验。为避免词级与句级混合表征共同作为输入时，两种混合表征间可能造成的影响，实验以表 4中仅使用词级混合表征为基础进行。

相较于语音表征，融合表征中还包含完整文本表征，以所含信息量评估，使用其作为输入BLEU值应更高。但如表 5所示，将语音输入更改为融合表征输入后BLEU值下降了0.1，特征相似度下降了0.73%，表明在融合表征作为辅助任务输入时，将其映射为较长的表征向量导致其与语音翻译主任务中语音表征产生较大长度差异，增加了模型跨模态映射难度和训练复杂度。将语音输入替换为文本表征之后BLEU与特征相似度下降情况更加明显，这表明将融合表征映射为与语音统一长度对缓解因长度造成的模态差异问题是有效的。

Q	K，V	BLEU	特征相似度（%）
词级混合表征	词级混合表征	24.4	97.41
文本表征	词级混合表征	23.4	80.03
语音表征	词级混合表征	24.5	98.14

表 5. 长度归一融合模块输入特征类型对英语-越南语翻译效果影响

使用T-SNE(Van der Maaten and Hinton, 2008)将各表征的隐层维度由512维简化为3维，对经过长度归一融合模块进行融合前后的文本和语音表征进行三元核密度估计可视化。如图 4所示，在进行长度归一化融合之前语音表征分布更为集中，导致模型难以学习到跨模态的对齐关系，表现了语音与文本不同模态表征之间存在的分布差异，而融合后两种表征的分布更为均匀，表明长度归一化融合模块可以有效将两种模态的表征映射到同一表征空间，缓解了不同模态间的分布差异，提升了翻译性能。



(a) 融合前表征分布　　(b) 融合后表征分布

图 4. 长度归一融合模块对表征分布的影响

3.2.5 门控融合模块对翻译性能影响

为验证门控融合模块在多任务训练框架下的作用，使用双JSD损失约束代替门控融合模块进行实验，经共享语义编码层的词级和句级融合表征都作为翻译解码器的并行输入，分别得到词级混合表征和句子级混合表征对应的目标文本词序列，过程分别简化表示为$h_1(x,y)$，$h_2(x,y)$，使用两个JSD损失对其进行约束，训练损失$L_1(s,x,y)$如式 (14)所示。

$$L_1(s,x,y) = L_{CE}(h(s),y) + \beta * L_{JSD}(h(s),h_1(s,x),y) + \beta * L_{JSD}(h(s),h_2(s,x),y) \quad (14)$$

为验证门控融合模块的有效性，对两个经共享语义编码层的混合表征求均值作为解码层输入，进行对比实验。

融合	JSD约束对象	BLEU	特征相似度（%）
无	句子与词级表征	24.1	97.09
均值融合	多粒度融合表征	24.5	98.11
门控融合	多粒度融合表征	24.7	98.28

表 6. 不同融合方式对英语-越南语翻译结果的影响

结果如表 6所示，在不对词级与句子级混合表征进行融合的情况下BLEU值为24.1，而使用取均值进行融合使BLEU提高了0.4，特征相似度提高了1.02%，表明过多的并行输入和损失约束增加了解码器的解码压力，导致翻译性能的下降。与均值融合相比，使用门控融合方法提高了0.2BLEU值，表明门控融合模块能够关注到词级与句子级混合表征间不同的模态与位置差异，实现对两种混合表征的有效整合。

3.2.6 多任务训练框架下不同损失约束的对比实验

为验证训练过程中损失函数的作用，如式 (15)所示，新增多粒度融合表征的交叉熵损失，用以评估模型在不同损失下的翻译效果。

$$\mathrm{L}_{\mathrm{CE}_m}(\mathrm{h}(\mathrm{s},\mathrm{x}),\mathrm{y}) = \sum_{\mathrm{i}=1}^{|\mathrm{y}|} \log\left(\mathrm{P}_\theta\left(\mathrm{y_i} \mid \mathrm{y_{<i}}, \mathrm{h}(\mathrm{s},\mathrm{x})\right)\right) \tag{15}$$

L_{CE_m}	L_{JSD}	BLEU	特征相似度（%）
×	×	23.4	97.88
√	×	23.6	93.00
√	√	24.2	97.81
×	√	24.7	98.28

表 7. 多任务训练中损失函数对英语-越南语翻译结果影响

如表 7所示，当使用交叉熵对融合表征作为输入的辅助任务进行约束时其BLEU值提升了0.2，但使模型更加关注于翻译效果更好的文本输入，导致特征相似度的下降。在此基础上加入JSD损失规范两个输出的预测，BLEU值进一步提升了0.6，特征相似度提升了4.81%，表明使用JSD损失进行一致性约束在多任务框架中对缓和模态差异的积极作用。当仅使用JSD损失进行一致性约束时，特征相似度与BLEU值得到了进一步提高，表明仅使用JSD损失与针对语音的交叉熵损失使模型更加关注语音翻译任务，同时降低了模型拟合的难度。

JSD损失是约束输出一致性的重要因素，为选定最优JSD损失比例，我们将JSD权重系数β分别置为0、1、2、3、4、5进行实验。由图 5可知，在β为4时得到最佳翻译效果。

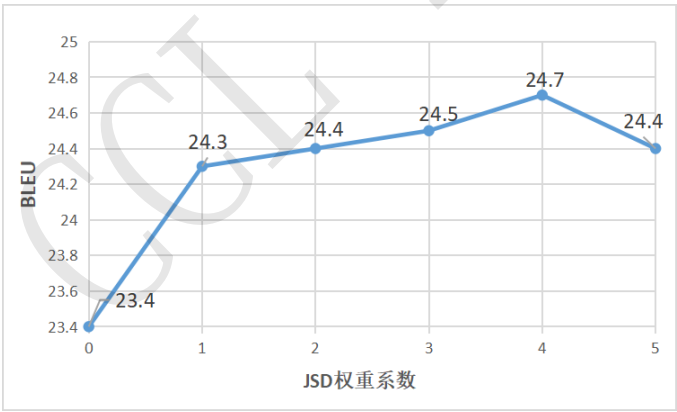


图 5. JSD权重系数对英语-越南语翻译结果影响

4 结论

针对端到端语音翻译过程中存在的跨模态问题，本文根据音频与文本之间长度的关系，提出语音文本跨模态表征对齐方法，使用多粒度混合特征作为模型并行输入，对不同模态的表征进行归一化融合与对齐，使用改进了损失约束的多任务训练框架约束混合表征与音频表征的一致性，使模型将语音与文本输入映射到同一表征空间。实验和分析表明了本文提出方法在不同层面对缓解跨模态表征差异的有效性，提高了端到端语音翻译的性能。未来的工作将在现有跨模态一致性表征工作的基础上对跨模态数据的利用进行研究，探索在使用更多外部数据的情况下，对端到端语音翻译进行增强的同时保持其跨模态表征的一致性。

参考文献

Ashkan Alinejad and Anoop Sarkar. 2020. Effectively pretraining a speech translation decoder with machine translation data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8014–8020.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2022. M3st: Mix at three levels for speech translation. *arXiv preprint arXiv:2212.03657*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*.

Gerard I Gállego, Ioannis Tsiamas, Carlos Escolano, José AR Fonollosa, and Marta R Costa-jussà. 2021. End-to-end speech translation with pre-trained models and adapters: Upc at iwslt 2021. *arXiv preprint arXiv:2105.04512*.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. *arXiv preprint arXiv:2105.03095*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Enrique Yalta Soplin, Tomoki Hayashi, and Shinji Watanabe. 2020. Espnet-st: All-in-one speech translation toolkit. *arXiv preprint arXiv:2004.10234*.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. *arXiv preprint arXiv:2104.06457*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*.

Yuchen Liu, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2020. Bridging the modality gap for speech-to-text translation. *arXiv preprint arXiv:2010.14920*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Siqi Ouyang, Rong Ye, and Lei Li. 2022. Waco: Word-aligned contrastive learning for speech translation. *arXiv preprint arXiv:2212.09359*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213. IEEE.

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

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Chen Wang, Yuchen Liu, Boxing Chen, Jiajun Zhang, Wei Luo, Zhongqiang Huang, and Chengqing Zong. 2022. Discrete cross-modal alignment enables zero-shot speech translation. *arXiv preprint arXiv:2210.09556*.

Kun Wei, Long Zhou, Ziqiang Zhang, Liping Chen, Shujie Liu, Lei He, Jinyu Li, and Furu Wei. 2022. Joint pre-training with speech and bilingual text for direct speech to speech translation. *arXiv preprint arXiv:2210.17027*.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Qi Ju, Tong Xiao, Jingbo Zhu, et al. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. *arXiv preprint arXiv:2105.05752*.

Zhuoyuan Yao, Shuo Ren, Sanyuan Chen, Ziyang Ma, Pengcheng Guo, and Lei Xie. 2022. Tessp: Text-enhanced self-supervised speech pre-training. *arXiv preprint arXiv:2211.13443*.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. *arXiv preprint arXiv:2205.02444*.

Xingshan Zeng, Liangyou Li, and Qun Liu. 2022. Adatrans: Adapting with boundary-based shrinking for end-to-end speech translation. *arXiv preprint arXiv:2212.08911*.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2020. Neurst: Neural speech translation toolkit. *arXiv preprint arXiv:2012.10018*.

基于离散化自监督表征增强的老挝语非自回归语音合成方法

冯子健[1,2]，王琳钦[1,2]，高盛祥*[1,2]，余正涛[1,2]，董凌[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1456644199@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.com,
ztyu@hotmail.com,46761956@qq.com,

摘要

老挝语的语音合成对中老两国合作与交流意义重大，但老挝语语音发音复杂，存在声调、音节及音素等发音特性，现有语音合成方法在老挝语上效果不尽人意。基于注意力机制建模的自回归模型难以拟合复杂的老挝语语音，模型泛化能力差，容易出现漏字、跳字等灾难性错误，合成音频缺乏自然性和流畅性。本文提出基于离散化自监督表征增强的老挝语非自回归语音合成方法，结合老挝语的语言语音特点，使用老挝语音素粒度的标注时长信息构建非自回归架构声学模型，通过自监督学习的预训练语音模型来提取语音内容和声调信息的离散化表征，融入到声学模型中增强模型的语音生成能力，增强合成音频的流畅性和自然性。实验证明，本文方法合成音频达到了4.03的MOS评分，基于离散化自监督表征增强的非自回归建模方法，能更好的在声调、音素时长、音高等细粒度层面刻画老挝语的语音特性。

关键词： 语音合成；老挝语；非自回归；预训练语音模型

A Discretized Self-Supervised Representation Enhancement based Non-Autoregressive Speech Synthesis Method for Lao Language

Zijian Feng[1,2],Linqin Wang[1,2],Shengxiang Gao*[1,2],Zhengtao Yu[1,2], Ling Dong[1,2]
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming 650500, China
1456644199@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.com,
ztyu@hotmail.com,46761956@qq.com,

Abstract

*高盛祥（通信作者）：gaoshengxiang.yn@foxmail.com
基金项目：国家自然科学基金（61972186,U21B2027）；云南高新技术产业发展项目（201606）；云南省重大科技专项计划（202103AA080015）；云南省基础研究计划（202001AS070014）；云南省科技人才与平台计划（202105AC160018）

Speech synthesis of Lao language is significant to the cooperation and communication between China and Lao, but Lao speech pronunciation is complex, and there are pronunciation characteristics such as tones, syllables and phonemes, and the existing speech synthesis methods do not work well on Lao language. The autoregressive model based on attention mechanism modeling is difficult to fit complex Lao speech, and the model has poor generalization ability and is prone to catastrophic errors such as word omission and word skipping, and the synthesized audio lacks naturalness and fluency. In this paper, we propose a non-autoregressive speech synthesis method for Lao based on discrete self-supervised representation enhancement, combining the linguistic phonetic features of Lao, using the annotated temporal information of Lao phoneme granularity to construct a non-autoregressive architectural acoustic model, extracting discrete representations of speech content and intonation information through a pre-trained speech model with self-supervised learning, and incorporating them into the acoustic model to enhance the model's speech generation capability of the model and the fluency and naturalness of the synthesized audio.The experiments demonstrate that the synthesized audio of this paper achieves a MOS score of 4.03, and the non-autoregressive modeling approach based on discrete self-supervised representation enhancement can better portray the speech characteristics of Lao language at the fine-grained level of intonation, phoneme duration, pitch, etc.

Keywords: TTS , Laotion , Non-Autoregressive , Pre-trained Speech Model

1 引言

老挝语是东南亚地区重要的语言之一，是老挝人民的官方语言，也在泰国、柬埔寨、越南等国家被广泛使用，研究老挝语语音合成对中老两国合作与交流意义重大。同时，研究如何融入语音特征提高老挝语语音合成性能，对后续研究非通用语种语音合成具有重要帮助。

老挝语与中英语种区别较大的地方是老挝语是声调语言，音调会直接添加到字符的上方，音调的改变会改变词语本身的意思，使得老挝语的语音合成需要在音节及音调上准确建模，由于其独特的发音方式，通用的语音合成方法在老挝语上难以保持语音合成效果，因此研究工作还存在很大困难。现有老挝语的语音合成缺乏自然度和流畅度，并且合成速度较慢。最近，(Anh et al., 2022)首次实现了基于神经网络的老挝语语音合成，然而该方法只是在基准模型上复现了老挝语语音合成。针对老挝语的语音合成任务还值得探索。

在语音合成任务中，主要目标有两个方面：(1)高质量：为了提高合成语音的自然度，模型应该捕捉到自然语言中的细节部分。(2)快速：在实际应用场景下，高速生成语音是至关重要的。一个完整的语音合成模型包括文本前端(Lai et al., 2021)，声学模型和声码器(Oord et al., 2016)(Kong et al., 2020)。本文从声学模型出发，结合老挝语语言语音特性研究如何构建更加高质量且快速的声学模型。声学模型从文本前端提供的信息中生成梅尔频谱图，然后使用单独训练的声码器根据梅尔频谱图来

合成语音，基于神经网络的语音合成系统显著的提高了合成音频的质量和自然度，并出现了很多面向特定领域的实际应用系统，Shen等人提出的Tacotron2模型架构(Shen et al., 2018)，相比于之前基于统计参数和级联系统语音合成方法，极大程度上改进了合成音频的质量。Ren等人提出基于Transformer(Vaswani et al., 2017)的非自回归架构FastSpeech(Ren et al., 2019)，移除了传统注意力机制对齐文本-语音的方法，选用了基于预测时长对齐的方式，使模型在解码的时候可以并行计算，极大地的提高了解码速度，同时解决了以往语音合成模型漏字和跳字等鲁棒性的问题。在此之后，Ren等人继续提出FastSpeech2(Ren et al., 2020)，训练和推理速度比自回归架构声学模型在速度上有极大提升。也有工作是在Tacotron2的基础上发展为非自回归模型(Elias et al., 2021)。这些工作在中、英等大语种上的语音合成取得了较好的成果，而针对非通用语种例如老挝语的语音合成工作还不足。为了解决老挝语语音合成音频缺乏自然度等问题，本文根据老挝语特性构建数据集，并利用微调的预训练语音模型提取老挝语语音特征，在传统FastSpeech2的语音合成声学模型架构体系上融合语音的离散化自监督表征来提升老挝语语音合成模型的性能。本文的贡献如下：

（1）实现了非自回归老挝语语音合成任务，解决了老挝语语音合成任务中，计算效率低，生成速度慢，声学模型泛化能力差的问题。

（2）提出了预训练语音模型融合机制策略，微调预训练语音模型，实现了将语音特征融入到声学模型中，改进了通用语音合成方法在老挝语上表现差的问题。

（3）在1小时左右的音频文本训练数据上的老挝语语音合成任务达到了4.03的MOS值。

2 相关工作

(1)自回归语音合成(Autoregressive Speech Synthesis)是一种基于序列模型的语音合成方法。在自回归语音合成中，语音信号被视为一个序列，每个样本都依赖于前面的样本。在传统的自回归语音合成方法中，通常采用的是循环神经网络（Recurrent Neural Network，RNN)(Grossberg, 2013)或者卷积神经网络(Convolutional Neural Network，CNN)(Gu et al., 2018)来建模语音信号的序列关系。这些模型能够学习到语音信号的时序特征，从而实现从文本到语音的转换。自回归语音合成方法具有一些优势，已被证明能生成连贯的语音信号，同时可以生成高质量的音频样本。

然而，由于其自回归的特性，有着昂贵的计算成本，自回归语音合成方法在生成速度方面相对较慢，因此在一些实时场景中不太适用，且在老挝语语音合成数据集较少的情况下，自回归语音合成模型可能会出现跳字、漏字等现象。

(2)非自回归语音合成（Non-autoregressive Speech Synthesis）是一种与自回归语音合成相对的语音合成方法。与自回归语音合成不同的是，非自回归语音合成模型不需要依赖于前面的样本来生成当前的样本，因此具有更高的生成速度和更低的延迟。Ren等人提出的FastSpeech2在FastSpeech的基础上添加了音调(Pitch)、音高（Energy）、时长（Duration）的外部语音信息，提高了合成语音的流畅度与自然度。另外，FastPitch(Łańcucki, 2021)基于FastSpeech对频率轮廓进行调节，提高了合成语音的整体质量。非自回归模型能以令人满意的速度生成语音音频，适用于实时语音合成等场景。

老挝语语音合成发音规律复杂，通用非自回归语音合成方法可能无法学习到足够的语音变化，从而导致过拟合或泛化能力不足，导致难于取得较好的效果。

第二十二届中国计算语言学大会论文集，第90页-第101页，哈尔滨，中国，2023年8月3日至5日。

3 方法

本文基于FastSpeech2的模型结构，根据老挝语语言语音特点构建数据集，在训练时融合语音特征，受Fang等人的启发(Fang et al., 2022)，本文将Wav2vec2.0提取的语音特征编码为隐状态，该隐状态序列与适应层的输出进行混合训练，对解码器输出的分布添加额外的JSD（Jensen-Shannon Divergence）损失来增强语音特征对生成语音的影响并提升训练效率。

3.1 数据构建

由于老挝语极其缺少语音文本数据对，很难利用MFA等工具(McAuliffe et al., 2017)对老挝语数据集进行切割，因此，为了做到音素对齐，本文在构建数据集阶段，对采集到的语音数据进行预处理，以使其适合进一步的分析和建模，包括对语音信号进行滤波、预加重、分割、去除噪声等预处理操作。其次使用Praat等语音分析工具(Styler, 2013)，从预处理后的语音信号中提取音高、音量、语速、音素持续时间等语音特征，对采集到的语音数据进行标注，对语音进行音素级别的标注和语音类型的分类，按照发音单元进行分割后，将预处理后的语音数据、提取的语音特征、标注信息和分割信息等整合在一起，构建数据集。老挝语音节的发音方式与其书写系统密切相关，老挝语以音节为最小发音单位，以音素为最小标注单位。并且老挝语没有官方的拉丁音译系统。本文首先老挝语辅音进行了详细的音素标注。如图1所示：

字符	ມ	ນ	ຍ	ງ	ບ	ດ	ປ	ຕ	ກ	ຜ, ພ
发音	[m]	[n]	[ɲ]	[ŋ]	[b]	[d]	[p]	[t]	[k]	[p^h]
字符	ຖ,ທ	ຂ, ຄ	ຝ, ຟ	ສ, ຊ	ຫ, ຮ	ຈ	ວ	ຢ	ຣ, ລ	ອ
发音	[t^h]	[k^h]	[f]	[s]	[h]	[tɕ]	[w]	[j]	[l]	[ʔ]

Figure 1: 老挝语辅音表

在老挝语中，音节是最小的发音单位，要组成音节则需要辅音与元音进行组合，图2中列出了具体的标注方式。

字符	ກະ	ກັ	ກິ	ກຶ	ກຸ	ເກະ	ເກັ	ເແກະ	ເແກັ	ໂກະ	ກໍ
发音	[ka]	[ka]	[ki]	[kɯ]	[ku]	[ke]	[ke]	[kɛ]	[kɛ]	[ko]	[ko]
字符	ເກາະ	ກັອ	ເກິ	ກາ	ກີ	ກຶ	ກຸ	ເກ	ເແກ	ໂກ	ກໍ
发音	[kɔ]	[kɔ]	[kɤ]	[ka:]	[ki:]	[kɯ:]	[ku:]	[ke:]	[kɛ:]	[ko:]	[kɔ:]
字符	ກອ	ເກີ	ເກັຍະ	ກັງ	ເກືອ	ກົວະ	ກົອ	ໄກ	ໃກ	ກັຍ	ເກຍ
发音	[kɔ:]	[kɤ:]	[kiə]	[kiə]	[kɯə]	[kuə]	[kuə]	[kai]	[kai]	[kai]	[ki:ə]
字符	ກງ	ເກືອ	ກົວ	ກອ	ກາຍ	ກ໋າ					
发音	[ki:ə]	[kɯ:ə]	[ku:ə]	[ku:ə]	[ka:i]	[kam]					

Figure 2: 老挝语音节组合实例

3.2 模型结构

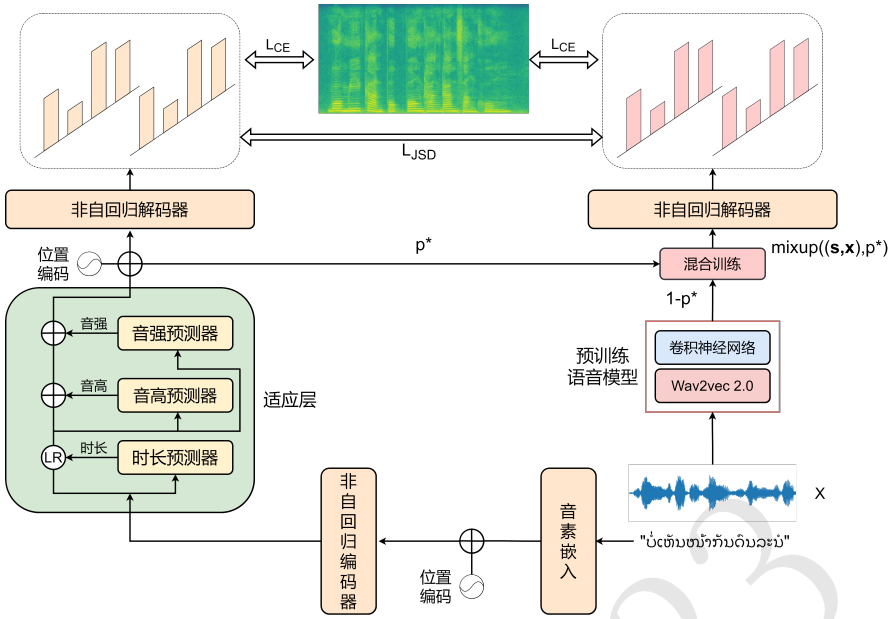


Figure 3: 基于离散化自监督表征增强的老挝语非自回归语音合成方法

模型的整体结构如图3所示，模型的输入为平行的文本-音频数据对，图中s为输入文本序列所对应的音素序列，x为文本序列对应的音频序列。采用非回归形式的编码器+解码器的架构，其中编码器、解码器分别由N个transformer层组成(N=4)，在编码层与解码层之间引入变换适应层(Variance Adaptor)用来作音素之间停顿的预测以及音调、音强的预测，使模型更好地建模音频特征。适应层预测的输出与提取的语音特征做混合训练，由于文本与语音结构不同，模态差异较大，语音是连续的时序信号，而文本是离散的符号序列，通过mixup方法，模型可以在不同模态之间建立联系和相互影响，混合文本-音频的数据可以帮助模型学习到文本到语音的相关性和一致性信息，从而提高模型对输入的理解和表达能力(Fang et al., 2022)。对解码器输出的分布添加额外的JSD损失，引入JSD损失用于度量生成的样本分布与真实样本分布之间的差异，它可以帮助模型学习生成更逼真的样本，来增强语音特征对生成语音的影响并提升训练效率(Gulrajani et al., 2017)。该损失函数为：

$$\mathcal{L}_{\text{JSD}}(\mathbf{s},\mathbf{x},\mathbf{y},p^*) = \sum_{i=1}^{|\mathbf{y}|} \text{JSD}\{p_\theta(\mathbf{y}_i \mid \mathbf{y}_{<i}, \mathbf{h1}(\mathbf{s})) \parallel p_\theta(\mathbf{y}_i \mid \mathbf{y}_{<i}, \mathbf{h2}(\text{mixup}((\mathbf{s},\mathbf{x}),p^*)))\} \quad (1)$$

其中h1(s)为文本通过音素编码器和适应层输出的上下文表示，h2(mixup((s,x),p^*))为通过预训练语音模型提取的语音特征向量经过声学编码器输出的向量与适应层输出的混合表征。

加上两次交叉熵损失，最终的损失函数如下：

$$\mathcal{L} = \lambda\mathcal{L}_{\text{JSD}}(\mathbf{s},\mathbf{x},\mathbf{y},p^*) + \mathcal{L}_{\text{CE}}(\mathbf{s},\mathbf{y}) + \mathcal{L}_{\text{CE}}(\text{mixup}((\mathbf{s},\mathbf{x}),p^*),\mathbf{y}) \quad (2)$$

其中 λ 是控制JSD损失的权重系数。

3.3 微调预训练语音模型

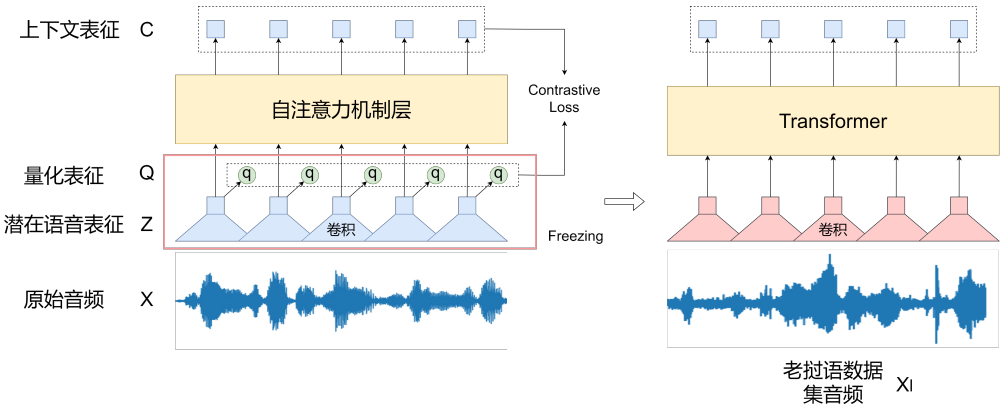


Figure 4: 微调wav2vec2.0

老挝语与中英等大语种差异较大，在通用方法下只对文本到音频进行建模使得模型难以充分训练，本文提出通过融入语音特征来提高模型的性能。

受wav2vec2.0启发，本文微调了预训练的多语言语音模型，利用无监督的语音预训练模型，迁移到语音合成任务中。具体而言，本文用预训练的wav2vec2.0模型结构中添加了一个自注意力机制层来增加语音量化表征上下文表征向量的相似度，根据数据集调整了学习率以及迭代次数。使用测试集来评估微调模型的性能。用评估结果来进一步调整模型架构和训练超参数。实验结果表明本文微调过的预训练语音模型所提取出来的语音特征可以提高语音合成的质量与流畅度。

在微调过程中，由于本文的数据集具有精确的音素持续时长标注，本文对特征编码器的输出采用了与SpecCutout类似的屏蔽策略(Kriman et al., 2020)：随机选择一些起始时间步长，对这些时间步长的数个后续时间步长的语音信号，将语音信号的频谱图进行分割，得到一些小块的频谱子图。在这些频谱子图中随机选择一些子图，然后将这些子图内的所有频谱值全部屏蔽（即用0来替换这些子图内的所有频谱值）。将所有被屏蔽的频谱子图拼接起来，得到一张被局部屏蔽的频谱图来代替原本的频谱图。本文使用与预训练时相同的屏蔽时间步长嵌入。图4中的对比损失为：

$$L = -\log \frac{\exp\left(\mathrm{sim}\left(c_t, q_t\right)/\kappa\right)}{\sum_{\hat{q}\sim Q_t} \exp\left(\mathrm{sim}\left(c_t, \hat{q}\right)/\kappa\right)} \tag{3}$$

其中

$$\mathrm{sim}(a,b) = a^T b/\|a\|\|b\| \tag{4}$$

4 实验

4.1 实验设置

本文在内部时长约一小时的老挝语数据集上进行了实验，音频采样率为22.05kHz，中间输出为特征维度大小为80的梅尔频谱图。音频由母语为老挝语的单人在专业录音

室录制，其训练集和验证集的大小比为4:1。训练时batchsize为32，使用$\beta_1 = 0.9,\beta_2 = 0.999,\varepsilon = 10^{-6}$的Adam优化器，学习率为$10^{-3}$。

4.2 评价指标

主观评价部分采用两种评价指标对模型综合能力进行评价。指标一：采用平均意见分（Mean Opinion Score Score，MOS）来评价合成语音的自然度和流畅度，听者根据自身感受对合成音频进行打分，MOS值评分共分为1-5五个等级，1分差，2分一般，3分正常，4分良好，5分最优，最后根据所有听者给出的意见分计算平均意见分。指标二：采用ABtest方式选择出听者感受更好的音频，ABtest方案会设置两组不同的语音合成系统的合成音频，听者盲听两组的音频并选择出较优的一方，最后计算听者的选择占比。以下评价模型优劣均为老挝语为母语的听众完成。

客观评价部分采用三种评价指标，指标一：实时因子RTF(Real Time Factor)，该值是评估非自回归模型推理速度的客观指标。指标二：MCD(Mel cepstral distortion)值，它表示的是转换后语音的MFCC 特征与标准输出语音的MFCC 特征的差距，用来验证本文的合成语音在语音特征上的保留程度。指标三：SSIM (Structural Similarity),它表示的是两个图像之间的相似程度，本文用来计算合成音频与真实音频之间的梅尔谱图相似度。

4.3 实验结果与分析

4.3.1 主观评价及客观评价

为了证明本文方法能够在保证快速生成语音的同时提高语音的自然度与流畅度，本文设置了与4个基准模型的对比实验，其中Tacotron2(Elias et al., 2021)是自回归语音合成模型，Tacotron2+GA是在Tacotron2模型的基础上加入了guide attention(Tachibana et al., 2018)，占比权重$\alpha = 1$。FastSpeech2(Ren et al., 2020)是本文的基准模型。FastPitch(Łańcucki, 2021)是另一个基于FastSpeech的非自回归模型。所有基准模型以及本文提出的模型都使用预先训练的HiFi-GAN(Kong et al., 2020)作为声码器，各基准模型的实验设置与提出该基准模型的原论文一致。各评价指标得分如表1所示。

Table 1: 客观评价指标与主观评价指标MOS评分。

方法	MCD	SSIM	RTF	MOS
Ground Truth	-	-	-	4.52±0.07
Tacotron2	7.76	0.45	2.31×10^{-1}	3.71±0.08
Tacotron2+GA	7.82	0.48	2.55×10^{-1}	3.86±0.07
FastSpeech2	7.78	0.43	$\mathbf{1.98} \times 10^{-2}$	3.85±0.08
FastPitch	8.38	0.46	2.16×10^{-2}	3.82±0.08
Our Model	**7.61**	**0.53**	2.07×10^{-2}	**4.03±0.07**

通过对表1的MOS评分进行分析，本文提出的模型能在只有一小时的老挝语音频文本训练数据下训练出完整的语音合成模型，并且使FastSpeech2在主观听觉上的表现超过了自回归的Tacotron2，证明本文通融合老挝语语音特征训练模型能够在一定程度上提升老挝语语音合成模型在韵律上的表现，并且取得了相比于基准模型更高的评分，相较于基准模型FastSpeech2,增加了0.18的MOS评分。

第二十二届中国计算语言学大会论文集，第90页-第101页，哈尔滨，中国，2023年8月3日至5日。

本节同时对实验结果进行客观评估，包括MCD(Kominek et al., 2008)值以及SSIM值(Wang et al., 2004)以及RTF值。其中MCD值的范围在0-10之间，数字越小说明两个音频之间的差距越小。SSIM的范围在0-1之间，数字越大说明两个图片的相似度越高。RTF值的计算方式为处理音频的时长/音频时长，在统一的设备配置下，RTF值越小说明实时率越高。通过表1进行分析，可以看出本文提出的模型在MCD指标上相比基准模型FastSpeech2降低了0.17。且SSIM值提高了0.1。

4.3.2 推理时长表现

为了验证本文提出模型是否具有快速的语音生成速度，本文对各基准模型在不同长度文本上进行了实验，推理时长实验结果如图5所示。

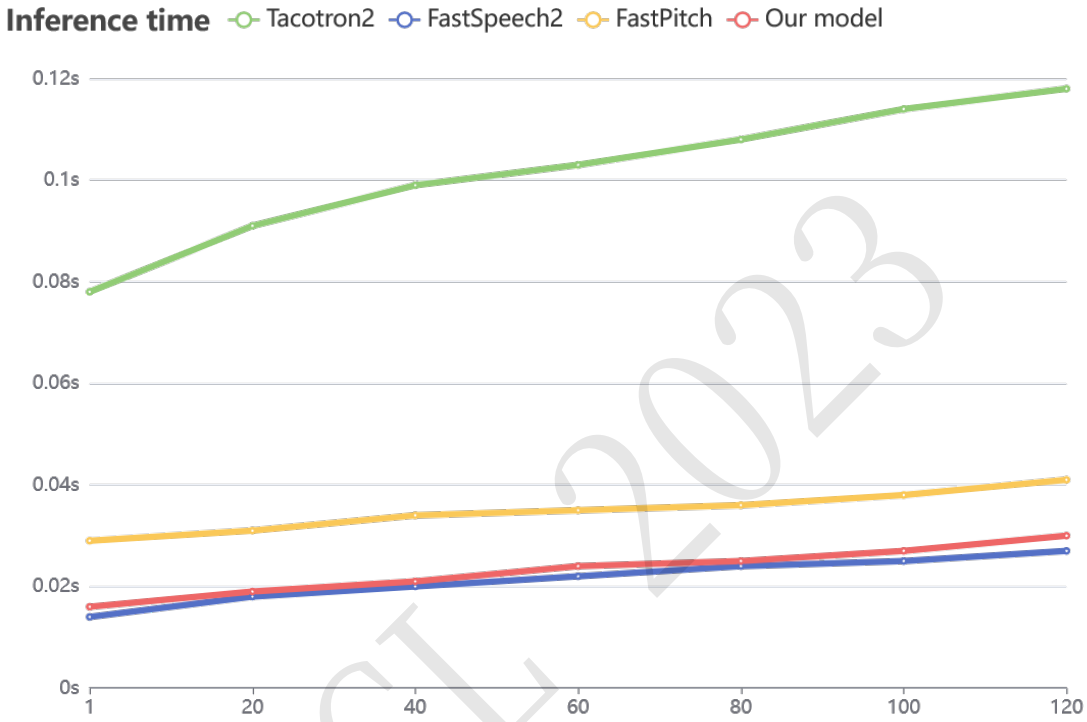


Figure 5: 模型推理时长表现

通过图5可以看出。由于自回归架构中每个样本都依赖于前面的样本，Tacotron2的推理时长相较于其他基准模型是花销更大的。而能够并行计算的FastPitch、FastSpeech2以及本文方法都具有较快的推理速度。本文方法在推理速度上的表现远远超出Tacotron2，相较于FastSpeech2也几乎没有增加开销。

4.3.3 消融实验

为了探索融合语音特征的最佳方式以及探索微调后的语音特征是否对模型性能提升具有决定性作用，本文进行了消融实验。分别尝试了对Tacotron2+GA模型直接融合提取的语音特征，对FastSpeech2将适应层的输出与特征向量直接进行拼接、加入注意力机制(Cross Attention)以及最终的模型。实验结果如表2所示。

Table 2: 消融实验：不同融合语音特征方式对声学模型性能的影响

方法	MCD	SSIM	RTF	MOS
Tacotron2+GA without finetuned wav2vec2.0	7.82	0.48	2.55×10^{-1}	3.86±0.07
Tacotron2+GA with finetuned wav2vec2.0	7.88	0.47	2.71×10^{-1}	3.87±0.06
FastSpeech2 without finetuned wav2vec2.0	7.78	0.43	$\mathbf{1.98} \times 10^{-2}$	3.85±0.08
FastSpeech2 with finetuned wav2vec2.0	7.98	0.46	2.01×10^{-2}	3.88±0.08
通过Cross Attention方式融合语音特征	7.64	0.51	2.51×10^{-2}	3.92±0.07
Our Model	**7.61**	**0.53**	2.07×10^{-2}	**4.03±0.08**

通过对表2进行分析，本文提出的融合语音特征方法在MOS评分上取得了最好的效果。FastSpeech2不融合语音特征的基线模型MOS评分为3.85，在直接融合了语音特征之后提升了0.03的MOS评分，达到了3.88，在一定程度上提高了语音的自然度和流畅度。而Tacotron2+GA基线模型的MOS评分达到了3.86，在直接融合语音特征之后提升了0.01的MOS评分，原因是Tacotron2模型受限于自回归模型本身泛化能力差，容易出现漏字、跳字等灾难性错误的问题。可以得出结论，本文主实验中的性能提升主要来源于本文提出的方法上的改进，而非在老挝语数据集对wav2vec2.0上进行微调。本文提出的对FastSpeech2融合语音特征方法，在不同融合方式上都提高了模型所生成语音在听觉上的表现，并且在MCD值和SSIM值上的表现均有提高。虽然FastSpeech2模型在RTF上有比最终模型有更好的性能，但由于本文提出的方法在生成语音的自然度和流畅度上有更好的表现，且RTF值相比基线模型的损失并不大，本文还是以获得了最高MOS评分、MCD值以及SSIM值的方法作为最终模型。

4.3.4 梅尔频谱图分析

为了分析本文提出的方法是否能借助预训练的语音模型在细节部分进行更精细地建模，本文对具有相同内容的语音进行了梅尔频谱图的分析。如图6所示。

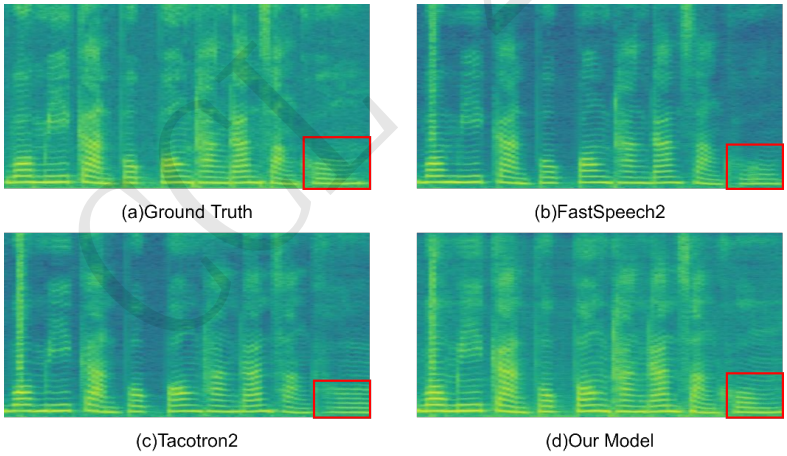


Figure 6: 梅尔频谱图分析

对图6中的红色部分进行梅尔频谱图分析，对于同样内容的语音，它的ground truth转换的梅尔频谱图如图中(a)所示，用FastSpeech2生成的梅尔频谱图如图中(b)所

示，可以看出在所选区域的梅尔频谱图有明显失真。而用Tacotron2模型生成的梅尔频谱图的对应位置可以发现实际上虽然Tacotron2合成该音频保留了发音，但是在结构上与原始音频相差很大，如图中(c)所示。而本文提出的模型可以完整保留该部分发音，在梅尔频谱图结构细节上与原始谱图保持了高度一致，如图中(d)所示。

4.3.5 ABtest实验

为了更直接地对比不同方法在主观上的优劣，本文进行了ABtest实验，该实验是让听者盲听两个不同模型对同一文本合成出来的音频，并选择出音频较优一方。参与该测试的一共有30人。测试结果如图7所示。

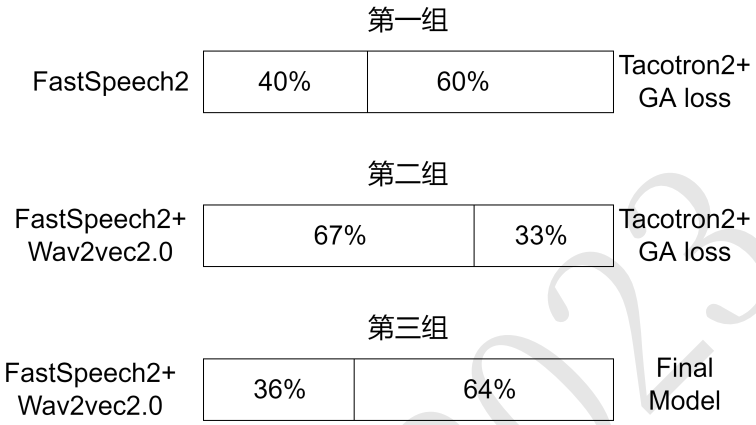


Figure 7: ABtest实验

测试显示，在第一组ABtest实验中，有超过一半的听众选择了自回归架构的Tacotron2模型合成的语音，说明在该数据集上仅仅基于FastSpeech2所合成的模型生成的语音在主观上不如Tacotron2。而通过第二组对比表明，基于FastSpeech2融合语音特征的模型所合成的语音的效果就超过了Tacotron2，这说明对于非自回归模型来说，融合语音特征的效果是有利于模型生成更自然、流畅的语音的。通过第三组实验可以对比出不同融合方式下，听众对语音优劣的主观选择，从而帮助本文定位出最适合融合语音特征的方法。

5 结论

针对通用语音合成方法在老挝语上表现较差的问题,本文提出基于离散化自监督表征增强的老挝语非自回归语音合成方法，结合老挝语的语言语音特点，在老挝语音素粒度上标注时长信息，使用非自回归架构建模声学模型提高老挝语语音合成速度，通过自监督学习的预训练语音模型来提取语音内容和声调信息的离散化表征，融入到声学模型中增强模型的语音生成能力，增强合成音频的流畅性和自然性，并且保证推理时长几乎不增加。实验证明，本文方法合成音频达到了4.03的MOS评分。

第二十二届中国计算语言学大会论文集，第90页-第101页，哈尔滨，中国，2023年8月3日至5日。

参考文献

Nguyen Thi Ngoc Anh, Nguyen Tien Thanh, et al. 2022. Development of a high quality text to speech system for lao. In *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–5. IEEE.

Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu. 2021. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*.

Stephen Grossberg. 2013. Recurrent neural networks. *Scholarpedia*, 8(2):1888.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. 2018. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans.

John Kominek, Tanja Schultz, and Alan W Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE.

Tuan Manh Lai, Yang Zhang, Evelina Bakhturina, Boris Ginsburg, and Heng Ji. 2021. A unified transformer-based framework for duplex text normalization. *arXiv preprint arXiv:2108.09889*.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Will Styler. 2013. Using praat for linguistic research. *University of Colorado at Boulder Phonetics Lab*.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4784–4788. IEEE.

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

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

面向机器翻译的汉英小句复合体转换生成能力调查

邢富坤
浙江外国语学院/ 浙江杭州
xingfukun@126.com

徐佳宁
青岛大学/ 山东青岛
461096572@qq.com

摘要

小句复合体由小句组合而成，不同语言在小句的组合模式上存在差异，该差异对机器翻译有何影响尚不清楚。本文以汉英机器翻译为例，选取多语体的汉语小句复合体及专家译文，从话头共享关系和共享类型两方面对主流机器翻译系统以及ChatGPT开展调查。结果显示，与专家译文相比，机器翻译的小句复合体转换生成能力存在较大不足，表现为机器翻译在话头补足、转换、提炼等方面的能力较弱，小句组合模式单一且带有明显的汉语原文痕迹，译文的地道性受到较大影响。

关键词： 小句复合体；转换生成能力；机器翻译；ChatGPT

Investigation of the Clause Complexes Transfer and Generation Capability from Chinese to English for Machine Translation

Xing Fukun
Zhejiang International Studies University
xingfukun@126.com

Xu Jianing
Qingdao University
461096572@qq.com

Abstract

Clause complexes are combined of clauses and the combination patterns differ across languages, and these differences play a crucial role in machine translation quality. However, our understanding of the exact impact is limited. To shed light on this issue, this paper focuses on Chinese-English translation. It examines Chinese clause complexes and expert translations in various genres, and evaluates mainstream machine translation systems and ChatGPT in terms of naming sharing relationships and sharing types. The findings reveal a significant disparity between the ability of machine translation systems and expert translators to transfer and generate clause complexes. Machine translation systems exhibit weaker performance in completing, converting, and refining naming information. They also demonstrate a limited range of clause combination patterns and retain many traces of the original Chinese text, which greatly compromises the quality of the translation.

Keywords: clause complexes , transfer and generation capability , machine translation , ChatGPT

1 引言

小句复合体由具有紧密逻辑语义关系的小句序列构成，是小句之上，篇章之下的语言单位。近年来小句复合体研究取得重要进展，宋柔（2022）对小句复合体的语法特征和性质进行了系统描写，为揭示小句复合体层面的语言规律提供了重要支持。小句复合体具有语种普遍性，其内部的话头共享机制及共享类型普遍存在于汉语、英语等语言之中，但话头共享类型的使用分布在不同语种间存在显著差异（宋柔，2022），这种差异给小句复合体结构分析与转换生成带来很大困难。下面是一个专家译文与机器译文在小句复合体层面的对比。

例1

原文：（选自2022年国务院政府工作报告）

各方面要围绕贯彻这些重大政策和要求，

细化实化具体举措，

形成推动发展的合力。

专家译文：

All of us involved

must adopt detailed and effective measures

‖ to meet these major policy requirements

and create synergy for driving growth.

翻译引擎A：

All parties should focus on implementing these major policies and requirements,

refine and implement specific measures,

and form a joint force to promote development.

上面的汉语原文、专家译文和机器译文都通过换行缩进形式标注出小句间的话头共享关系和共享类型。汉语原文是一个由3个标点句组合而成的小句复合体，每个标点句一行，其中第2和第3行都缩进到了第1行的“各方面要”之后，表示这两行与第1行的“各方面要”之间存在话头共享关系，第2行和第3行分别与共享的话头“各方面要”组合，可生成完整的小句，分别是：①各方面要围绕贯彻这些重大政策和要求。②各方面要细化实化具体举措。③各方面要形成推动发展的合力。由于共享的话头位于第一个小句的起始位置，因此第2、3小句的话头共享类型都属于分支模式。

与汉语原文相仿，使用换行缩进对专家译文和机器译文进行话头共享的标注。在专家译文中，其话头共享关系和共享类型与汉语原文存在差异，表现为：话头共享关系存在差异，译文第2行中的measures作为话头被后句共享，但其对应的原文“举措”在汉语原文中并非是话头，不存在共享关系；话头共享类型也存在差异，汉语原文的话头共享类型只有分支模式一种，而专家译文中则是第2个小句共享“All of us”，第4个小句共享“All of us must”，这两个小句的共享类型都是分支模式，而第3个小句则共享了第2个小句的末尾成分“measures”，这种共享前面小句的非开始成分做话头的类型属于新支模式，因此专家译文包含分支模式和新支模式两种话头共享类型。机器译文在话头共享关系和共享类型层面与汉语原文完全相同，第2、3小句都共享了第1句的“All parties should”，且话头共享类型都是分支模式。

观察机器译文的每一个小句，其质量基本不存在问题，但如果观察译文小句组合而成的小句复合体，则机器译文质量较专家译文有较大差距，具体表现为机器译文的同一个主语，后面连续跟了三个具有并列结构的限定性动词短语，这种多个动词短语并列的表达形式在英语中较少使用，带有明显的中式英语痕迹，机器在小句组合模式层面原样照搬了汉语的小句组合模式，未作出适合英语表达习惯的小句组合模式调整，从而影响了机器译文的整体质量。

现有机器翻译研究缺少对小句复合体转换生成能力的关注，小句组合模式的差异对机器译文质量存在哪些影响以及影响程度如何尚不清楚。本文以汉英机器翻译为例，选取多语体的汉语小句复合体及专家译文，从话头共享关系和共享类型两方面对主流机器翻译系统以及ChatGPT的小句复合体转换生成能力开展调查。调查的机器翻译系统包括百度翻译、小牛翻译和有道翻译三个主流翻译系统，并将ChatGPT纳入调查范围，目的是调查ChatGPT这类生成式大语言模型在小句复合体层面的生成能力，从而能够对专家、机器翻译以及生成式大语言模型在小句复合体层面的转换生成能力进行更全面的对比分析。

2 小句复合体理论和机器翻译现状

2.1 小句复合体理论

语言学界对于小句复合体结构本质的共识就是逻辑语义关系，Halliday（1985）认为小句复合体是通过衔接手段将语义上相互依赖的各小句连成一体，各小句间有一定的逻辑关系。宋柔（2018）则从小句复合体的形式规律层面开展研究，对小句复合体的语法结构进行了系统描写，提出了话头（naming）和话身（telling）等核心概念，认为话头是话语字面上的出发点，话身是对于话头的陈述，并基于话头共享结构界定了汉语的小句复合体，认为小句复合体不仅是小句间具有紧密逻辑语义结构的标点句序列，而且有特定的语法结构，即小句间遵循一定的模式来共享话头和话身。小句复合体的话头话身共享大体分为四种类型，即分支模式、新支模式、后置模式和汇流模式。其中分支模式的特点是一个话头被多个右置的话身共享，是汉语语篇中出现最多的话头共享模式。

宋柔和葛诗利（2015）基于小句复合体理论，面向篇章机器翻译对英汉翻译单位和翻译模型进行探究，以NT小句为基本单位，设计了篇章机器翻译的拆分、翻译、装配三步模型（PTA模型），并以PTA模型的翻译过程为标注内容，建设英汉篇章NT小句对齐语料库。宋柔与葛诗利团队（2020）基于宾州英语树库的华尔街日报建立了一个包含5000个英语句子的英语语料库，并对其进行NT小句划分，结果显示小句复合体理论的的话头话身关系类型可以对语料库中99%以上的句子进行描写，进一步证实了该理论的语种通用性。英语中同样存在小句复合体，小句复合体中的小句之间可以通过成分共享进行分析，共享类型与汉语一致，包括分支、新支、后置、汇流四种模式，但模式的使用分布与汉语有较大差异，集中表现在英语中多用新支模式，而汉语多用分支模式，具体统计结果可参考宋柔（2022）。

小句复合体理论从话头共享角度描写小句间关系，不仅揭示出小句间组合的形式规律，同时也为不同语言间的对比研究提供了重要视角，从已有研究结果看（宋柔, 2022；张学贞, 2022），小句复合体理论一方面可以全覆盖地对汉语、英语和日语进行描写，另一方面三种语言间在话头共享模式层面又各具特点，各不相同。

2.2 机器翻译现状

机器翻译的概念起源于20世纪30年代，历经基于规则的机器翻译、基于统计的机器翻译、基于实例的机器翻译和基于不同方法应用的机器翻译（胡开宝，2016）。2016 年，谷歌推出基于神经网络的机器翻译，机器翻译效果有了极大提升，并取代统计机器翻译成为谷歌、百度、微软等商用在线机器翻译系统的核心技术。

随着机器翻译的不断发展，人们对机器翻译的需求也不仅局限于原先的句子级别，而是到了句子之上的篇章级别，神经网络机器翻译的发展也带动了篇章机器翻译的发展，目前有关机器篇章翻译的热点主要集中在神经机器翻译领域。虽然篇章神经翻译的对象是篇章，但由于受到计算资源限制，大部分模型仍先依次翻译篇章中每个句子，再拼接得到最终篇章译文。因为需要关注上下语义的连贯，目前篇章神经机器翻译大多注重译文中篇章的共指性和连贯性，从机器的上下文建模、模型分析和模型训练几个方面进行研究，篇章层面的翻译仍存在一定的问题（苏劲松，2020）。

以句子为单位机器翻译系统在处理篇章信息时会忽略上下文信息（李晗佶，2020），同时也欠缺相关的技术对小句复合体内部小句间逻辑语义的分析。目前机器翻译较少关注汉英转换在小句复合体层面的生成能力，研究方面也较为欠缺对汉英机器翻译在小句复合体层面生成能力的调查。

美国OpenAI公司于2022年11月底推出ChatGPT模型，虽然该模型并非是专门针对翻译任务而训练的模型，但由于其依托于大语言模型（Large Language Model, LLM），具有语言间的翻译能力，在翻译领域有一定的应用（朱光辉，2023）。目前已经有研究者（Wang&Lyu, 2023; Jiao et al.,2023）对ChatGPT等大语言模型的翻译能力开展调查和评价工作，但评价方法与标准存在一定局限性，表现为现有译文评价基本沿袭已有翻译数据集并基于BLEU值进行译文评价。由于主流翻译模型基本是在该评价框架下进行训练优化，因此利用该框架评价ChatGPT与主流机器翻译系统具有一定的不公平性，难以反映出ChatGPT这类大语言模型独有的翻译能力。更重要的是，BLEU评价方法得出的结果反映的是机器译文与人工译文在Ngram层面的相似度，但由于小句间的话头共享关系具有跨距特点，且有时会跨越有多个小句，因此很难被Ngram捕获，因此现有的译文自动评价方法难以反映出不同译文在小句复合体

层面的差异，进而无法对不同模型的小句复合体翻译能力做出评价。本文针对小句复合体层面的转换生成能力开展人工调查与评价，目的是调查不同模型以及专家在小句复合体层面的翻译情况，进而给出不同翻译主体在小句复合体层级的翻译特点与能力，为优化机器翻译提供支持。

3 调查方法

3.1 翻译引擎选取

本研究选取百度翻译、小牛翻译和有道翻译三个主流翻译引擎和大语言模型ChatGPT作为调查对象，前三个翻译引擎以翻译引擎A、翻译引擎B和翻译引擎C来代表。前三个翻译引擎均使用主流的神经机器翻译理论和技术，通过深度神经网络使计算机先对语料库进行自动学习，再自动生成机器译文。ChatGPT是新兴的生成式大语言模型，通过海量无标注语料的预训练以及高质量标注语料进行有针对性的微调而成，具有较为广泛的能力，本文只对其翻译能力进行调查。

这四类翻译引擎应用广泛，市场上有很大的客户群体，其翻译能力在机器翻译行业具有一定的代表性。本研究对这四类翻译引擎的汉译英机器翻译进行标注、分析和总结，反映出目前整体机器翻译引擎的汉译英在小句复合体层面的机器翻译能力。

3.2 调查语料

本文选择政府工作报告、白皮书、法律以及小说四类文本作为调查语料，其中政府工作报告为2022年全国人大政府工作报告，白皮书为2019年白皮书，包括《新时代的中国国防白皮书》《中国新型政党制度白皮书》《中国应对气候变化的政策与行动白皮书》，法律为中国法律节选，包括《法官法》《个人信息保护法》《国防法》《行政处罚法》《反食品浪费法》。政府工作报告、白皮书、法律这三类信息类文本作为政府官方文件，汉语特点鲜明且语言严谨。小说选取《围城》和《蛙》的节选片段。补充文学类文本为的是增加样本中汉语语言的多样性和代表性。其中2022年政府工作报告的译文为新华网发布的国务院官方权威英译文，白皮书译文为国务院新闻办公室发布的英文版，中国法律节选的英文版出自全国人民代表大会发布的官方权威英译文，《围城》的译文为珍妮·凯莉和茅国权的英译版，《蛙》的译文为中国文学翻译家葛浩文所译的英译版。选取的译文均为该领域的专家译文，译文质量和权威性较高，可作为调查的基准与机器译文进行对比，探究机器的译文生成能力。

3.3 调查设计与分析方法

本文选取了汉语中使用最广泛的分支模式作为调查对象，从调查语料中共选出共44个具有分支模式的小句复合体，共154个小句，其中政府工作报告小句70句，白皮书、法律、小说这三类语体的小句各28句。目前的调查样本的绝对数量不大，但并不会对调查目的造成本质影响，一方面是考虑了语体因素，在多种语体中采样，保证调查具有较大的覆盖范围，另一方面是选用了多个机器翻译系统进行调查，同一个原文提交给不同的机器翻译系统翻译，然后对机器译文进行比较分析，调查不同机器翻译系统间是否存在某些共性特点。未来将进一步扩大调查的语料规模，提高调查结果的可靠性和代表性。调查的基本单位是小句复合体，内容包括汉语原文的小句复合体、原文对应的专家译文小句复合体、机器译文小句复合体以及ChatGPT翻译得到的译文小句复合体。

针对调查目标，对原始语料进行标注，基本内容包括：汉语原文的标点句切分及话头共享关系标注、专家译文的小句切分和话头共享关系标注以及机器翻译引擎A、B、C和ChatGPT的译文做小句切分和共享关系标注，同时在小句复合体层面将汉语原文与专家译文、机器译文进行对齐。

调查的基本方式是，以专家译文为基准，将专家译文以及四类机器译文的标注结果分别与汉语原文的标注结果进行对比，标注出译文与原文在话头共享关系和话头共享类型层面的异同。

“话头共享关系”是指汉语原文小句中的话头与对应译文小句中的话头之间的对应关系，如果原文和译文的话头一致，则标注为“话头一致”，反之标注为“不一致”。对于话头不一致的情况，进一步调查不一致的类型，分别是话头补足、话头变形、话头转换和话头提炼。“话头共享类型”是指汉语原文小句的话头共享类型与译文小句的话头共享类型间的对应关系，也分为一致

和不一致两类，如果不一致，则进一步标注不一致的小句话头共享类型的具体情况，主要包括三类，独立小句、新支句和后置句。调查的内容体系具体如下图所示。

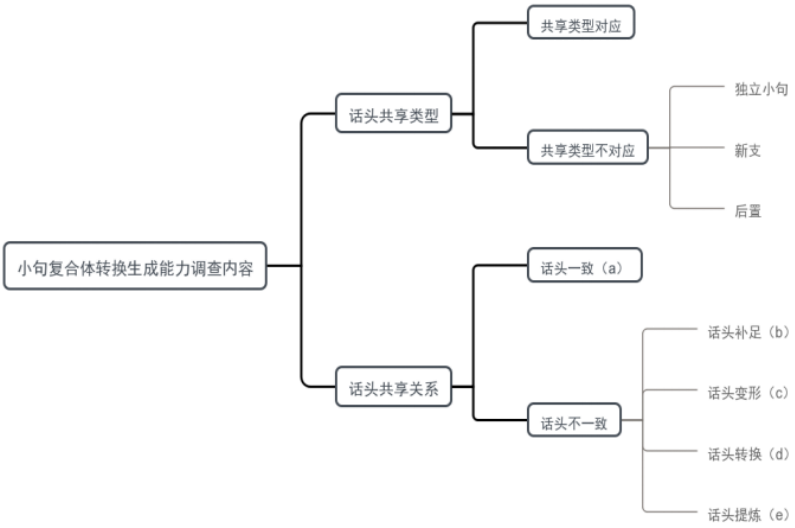


图 1: 小句复合体转换生成能力的调查体系

本文根据以上调查内容，设计了具体的标注规范，下面是具体标注实例。例2给出的是原文与译文在话头共享类型上具有一致性的情况，这类情况，在译文对应小句后以双斜线分隔，标注为“一致”，表示译文与原文的话头共享类型一致。

例2

原文：（选自2022年国务院政府工作报告）

国内生产总值达到114 万亿元，

增长8.1

翻译引擎A：

The gross domestic product reached 114 trillion yuan, //一致

up 8.1%. //一致

例3则是话头共享类型不一致的实例，对于不一致的实例，本文使用其话头共享类型名称进行标注，如例3译文中第1个小句的话头共享类型为“后置”类型，而其对应的原文第2行则是共享第1行起始位置的成分，共享类型是“分支”，原文与译文不一致，因此在译文第1个小句后标注为“后置”。

例3

原文：（选自2019年新时代的中国国防白皮书）

中国国防费按用途划分，

主要由人员生活费、训练维持费和装备费构成。

翻译引擎A：

D‖According to purposes, //后置

China’s defense expenditure is mainly composed of personnel living expenses, training and maintenance expenses and equipment expenses.

下面给出了“话头共享关系”的标注实例，例4是话头一致情况，即原文话头与译文话头具有一致关系，该类标注为“a”。

例4

原文：（选自2022年国务院政府工作报告）

{我们}加强大宗商品保供稳价，

着力解决煤炭电力供应紧张问题。

翻译引擎B：

{We will}Strengthen the supply and price stability of bulk commodities, //a

and strive to solve the problem of tight supply of coal and electricity. //a

原文第1行中的话头没有显式出现，但从上下文可以推断出隐含话头是“我们”，因此在标注中以大括号的形式补充出话头，译文的第1行与汉语一致，也没有出现话头，标注中同样以大括号的形式补充出话头“We will”，原文与译文均未出现隐含的话头，且补充后的话头具有对译关系（忽略助动词，下同），这类情况被归为话头一致。

（2）“话头不一致”的情况主要包含四类：①“话头补足”，即在汉语原文没有出现话头，但话头可在上下文中找到，译文将原文隐含的话头显式补足出来的情况，类标注为“b”；②“话头变形”，即英语小句中以其他的形式如代词等表现出汉语隐含的话头的情况，类标注为“c”；③“话头转换”，即英语小句中的话头由汉语原文中非话头成分转换而来，如汉语原文中的宾语成分在英语译文中转换为话头的情况，类标注为“d”；④“话头提炼”，即英语小句的话头是基于汉语原文提炼总结得到的，类标注为“e”。例5给出了话头提炼的情况，其他情况不再一一给出实例。

例5

原文：（选自2022年国务院政府工作报告）

{我们}加强大宗商品保供稳价，

着力解决煤炭电力供应紧张问题。

ChatGPT：

Efforts must be made to stabilize the supply and prices of bulk commodities, //*e*

and to tackle the problem of coal and electricity shortages. //*e*

例5中汉语原文的话头为“我们”，但译文的话头“Efforts（举措）”则在原文中没有出现，是译者基于原文提炼得到，原文讲到的“加强大宗商品保供稳价”和“着力解决煤炭电力供应紧张问题”都可被认为是“Efforts（举措）”的具体表现，因此译者提炼出一个在原文中并不存在但符合原文表达意思的话头，这类情况称之为“话头提炼”。

可以看出，无论话头共享关系，还是话头共享类型，都是发生在小句与小句之间的组合上，与小句中具体词或短语翻译无关，因此是小句复合体层面的语言现象，通过调查机器翻译在这两类语言现象上的表现，可以反映出机器在小句复合体转换生成上的基本能力。

4 调查结果与分析

4.1 调查结果

由于专家译文与机器翻译引擎的译文在小句总数量上不完全一样，为了对不同来源的译文进行统一对比，本文在原始数据基础上，进行归一化处理，下表给出不同情况的绝对数量和其占各自译文小句总数的百分比。其中四类不同的翻译引擎A、B、C、ChatGPT与专家译文话头共享类型对应情况调查结果如表1所示。

	翻译引擎A	翻译引擎B	翻译引擎C	翻译引擎总	ChatGPT	专家译文
一致	87.01%	88.89%	86.58%	87.50%	87.84%	62.94%
独立小句	10.39%	9.15%	10.07%	9.87%	6.08%	11.68%
新支	0.65%	1.31%	2.01%	1.32%	2.70%	19.29%
后置	1.95%	0.65%	1.34%	1.32%	3.38%	6.09%

表 1: 翻译引擎、ChatGPT与专家译文话头共享类型对应情况调查结果

表1给出了四类翻译引擎和专家译文在话头共享类型层面的对应情况，从中可以看出，四类翻译引擎在话头共享类型的对应关系中，“一致”情况具有很大的相似性，其占比均为超过85%，位于85% 89%之间，而专家译文的“一致”情况只有62.94%，低于机器翻译引擎约20个百分点，反映出专家译文在话头共享类型上，与原文有较大差异，而机器译文与原文的差异程度远小于专家译文的差异程度。三种翻译引擎“一致”情况总占比与ChatGPT都是87%多一些，这两类翻译引擎在话头共享类型上都与原文保持了较高的相似程度，而与专家译文则有较大差异。这初步说明机器译文在话头共享类型上更多地受到源语的影响。

在“不一致”的三种情况中，机器译文与专家译文也表现出较大的差异性。首先，译文是“独立小句”的情况中，占比最高的是专家译文，达到11.68%，而最低的是ChatGPT，为6.08%，二者相差5个百分点，而其他三个专门的翻译引擎则较为相似，基本在10%左右。该数据说明，

相较机器译文，专家译文更多使用独立小句，即译文小句中包含完整的话头和话身，对于汉语原文中使用话头共享关系的小句，在翻译中会将其话头补全，并明确地翻译出来。ChatGPT的译文则更少地使用独立小句，相应地ChatGPT会使用更多的组合小句来进行翻译，这表现出ChatGPT的独特之处，其原因和译文效果需要进行专门评估。

在“新支小句”中，则专家译文的使用比例达到19.29%，远远超过机器译文使用新支小句的比例，翻译引擎A、B、C的比例基本为2%或更少，ChatGPT也只有2.7%。由于在本文选取的汉语原文语料中，话头共享类型都是分支类型，没有新支类型，因此，如果译文出现新支类型，说明在翻译转换中在一定程度上摆脱了汉语原文的影响。从这个角度看，专家译文高比例地使用新支类型，表现出专家译文相对于机器译文的特别优势。同时，由于ChatGPT使用新支类型的比例在一定程度上高于其他三个机器翻译引擎，因此也可以初步判断ChatGPT在小句复合体层面的译文转换生成能力强于已有的机器翻译引擎。

在“后置小句”中，专家译文同样表现出远高于机器译文的水平，与新支小句同理，由于汉语原文中没有后置类型，因此后置类型的使用也能表现出译文脱离原文影响的程度。专家译文受到原文影响的程度最小，ChatGPT居中，另外三个机器翻译引擎则较大地受到了原文话头共享类型的影响。

下面是一个机器译文、ChatGPT译文和专家译文在话头共享类型层面的对比情况。

例6

原文：（选自中华人民共和国国防法）

现役部队是国家的常备军，

主要担负防卫作战任务，

按照规定执行非战争军事行动任务。

翻译引擎A：

The active force is the standing army of the country,

which is mainly responsible for defense operations

and performs non war military operations in accordance with regulations.

翻译引擎B：

The troops in active service are the standing army of the country,

which are mainly responsible for defensive operations

and perform non-war military operationsin accordance with regulations.

翻译引擎C：

The active forces are the state’s standing forces,

mainly responsible for defensive operations,

and carry out non-war military operations in accordance with regulations.

ChatGPT：

The active-duty troops are the country’s regular army

and mainly hold the responsibility of defensive combat missions,

and execute non-war military operations tasks as required.

专家译文：

D||As the standing forces of the State,

active-duty forces shall be mainly tasked with defense operations

and perform non-war military operations in accordance with regulations.

例6中汉语原文是一个分支类型的话头共享结构，翻译引擎A、B、C及ChatGPT与原文的话头共享类型一致，呈现出较强的原文影响程度；而专家译文话头共享类型与原文不完全一致，使用了后置类型，表现出专家译文相比于机器译文在翻译转换上的优势。

通过表2可以较为直观地看出四类翻译引擎和专家译文在话头共享关系层面的对应情况，其中“话头一致”情况，机器翻译引擎之间仍有较大的相似性，且占比较高，位于78% 89%之间；而专家译文的“话头一致”情况只有68.02%，低于机器翻译引擎约10-20个百分点，反映出专家译文在话头共享关系上与原文差异较大，而机器译文与原文的差异程度仍小于专家译文的差异程度。将机器翻译引擎分为两大类，在“话头一致”方面三种专门翻译引擎与ChatGPT均为80%以上，这两类翻译引擎在话头共享关系上与原文相似程度也较高，而与专家译文相比有一定的差

第二十二届中国计算语言学大会论文集，第102页-第112页，哈尔滨，中国，2023年8月3日至5日。

	翻译引擎A	翻译引擎B	翻译引擎C	翻译引擎总	ChatGPT	专家译文
话头一致(a)	81.17%	88.31%	78.67%	82.75%	80.54%	68.02%
话头补足(b)	9.74%	1.95%	14.00%	8.52%	1.34%	5.58%
话头变形(c)	6.49%	7.79%	6.00%	6.77%	5.37%	5.58%
话头转换(d)	2.60%	1.95%	1.33%	1.97%	6.04%	19.80%
话头提炼(e)	0.00%	0.00%	0.00%	0.00%	6.71%	1.02%

表 2: 翻译引擎、ChatGPT与专家译文话头共享关系对应情况调查结果

异，机器译文不仅在话头共享类型上更多地受到源语的影响，在话头共享关系方面也呈现出相似的情况。

“话头不一致”的四种情况，专家译文与机器译文也存在一定的差异性。本文认为这四种情况，即“话头补足”、“话头变形”、“话头转换”和“话头提炼”的话头生成的难度逐级递增，话头生成能力也越来越强。首先，译文为“话头补足”的情况中，三个专门的翻译引擎占比最高，达到8.52%，而专家译文次之，为5.58%，ChatGPT占比最低，仅为1.34%，与三个专门翻译引擎相比，相差7个百分点。该数据说明，机器译文更多地使用话头补足，即在英语译文中对汉语上下文中出现但在小句复合体中省略的主语进行的形式上的补足，专家译文相对较少使用这种方法，而ChatGPT使用占比更少。在“话头变形”的情况中，机器译文与专家译文在占比方面均无较大差异，均在5%-7%之间，表明机器译文和专家译文在翻译译文时均会在一定程度上使用话头变形，即将呈现为名词等情况的原话头，在对应译文以及补充缺失话头时转换为同义名词或代词等情况的翻译方法。在“话头转换”的情况中，专家译文呈现出明显的差异，占比高达19.80%，远超于ChatGPT的6.04%和翻译引擎A、B、C的1.97%，展现出专家译文相比于机器译文，脱离原文句式对话头进行转换的能力。而译文为“话头提炼”的情况中，ChatGPT显示出独特的优势，其占比为6.71%，不仅高于翻译引擎A、B、C的0%，也高于专家译文的1.02%。在其他翻译引擎在话头提炼能力展现出欠缺的情况下，ChatGPT可以做到对语义的理解总结提炼出原文没有明确表达出的内容作为话头，表现出其独特的优势和较高的话头生成能力。

例7给出了话头补足的情况，原文第1行没有出现话头，但可以从上下文中推断出隐含的话头是“我们”，译文中则补充出话头“We will”。由于本数据集大多为信息类文本，缺失主语大多为“We”，补足难度较低，根据数据发现，机器翻译引擎ABC话头补足能力相对不错，为8.52%，高于专家译文的5.58%和ChatGPT的1.34%。虽然在数据上超越了专家译文和ChatGPT，但专家译文使用频率低的原因是话头补足在话头生成方面相对简单，较多使用了较为复杂的话头转换，而ChatGPT则是更多的使用了更为复杂的话头提炼。

例7

原文：（选自2022年国务院政府工作报告）

{我们}加强大宗商品保供稳价，

着力解决煤炭电力供应紧张问题。

翻译引擎A:

We will strengthen supply and price stability for bulk commodities, //*b*

and work hard to solve the problem of tight supply of coal and electricity. //*a*

话头变形是指原文话头与译文话头的所指对象一致，但表达形式上有变化，例如原文话头为名词，在译文中转换为另一个同义名词或代词的情况，这类情况翻译引擎、ChatGPT和专家译文差别不大，不详细列举。

话头转换是将原文非话头成分，例如宾语等成分转换为译文的话头，涉及句式的转换，是一类更为复杂的话头生成能力。如例8所示，原文第2行中的“多场专题新闻发布会”不是话头成分，而是宾语成分，但其对应的译文为“A number of special press conferences”充当了译文的话头，这种情况属于话头转换。此外，译文第2句的“major issues”是新支句话头，对应原文的“重大事项”，同样也发生了话头的转换。专家译文中话头转换的占比为19.80%，明显高于翻译引擎的的1.97%和ChatGPT的6.04%。专家译文多使用话头转换，可以实现句式的灵活性和多样性，而机器翻译引擎的译文较多地受原文影响，话头转换能力较弱，ChatGPT则位于二者之间，强于机器翻译，而弱于专家译者。

例8

原文：（选自2019年新时代的中国国防白皮书）

{国防部}围绕深化国防和军队改革、裁减军队员额等重大事项，
召开多场专题新闻发布会。
组织近百家中外媒体多次赴部队、军事院校参观采访。

翻译引擎A：

A number of special press conferences were held around major issues //*d* ‖ such as deepening the reform of national defense and the military, and reducing military posts. //*d*

Nearly 100 Chinese and foreign media have been organized to visit the troops and military academies for many times. //*d*

话头提炼是需要通过对语义的理解总结提炼出原文没有明确表达出的内容，并将其作为话头。由于需要语义理解，因此这一类的话头生成较为复杂和困难。例9中，原文的话头是"我们"，但译文话头"Efforts（举措）"是基于原文提炼得到，原文讲到的"加强大宗商品保供稳价"和"着力解决煤炭电力供应紧张问题"都是"Efforts（举措）"的具体表现，因此译者提炼出一个在原文中并不存在但符合原文表达意思的话头。机器翻译引擎的话头提炼方面存在明显不足，在调查的样例中没有出现此类话头；ChatGPT的话头提炼占比则高达6.71%，显著高于专家译文的1.02%，表现出ChatGPT这类生成式大语言模型在话头转换生成方面的独特优势，其能力甚至超过专家译者。深层原因值得进一步调查分析，仅从原文对译文的影响角度看，机器译文受到原文形式的影响最大，没有在原文形式层面出现的内容，很难被翻译生成出来，导致翻译引擎生成提炼式话头的能力最弱；专家译者基于对原文的理解进行翻译，能够在一定程度上摆脱原文形式影响，生成一些未在原文中出现，但符合原文语义的话头，但这种能力有着较大个体差异，优秀译者此方面能力较强，而一般译者则较弱；ChatGPT这类大语言模型具有较强的语义表示能力，基于深层的语义表示进行转换，从而较好地摆脱原文形式束缚，而能够较为灵活地生成提炼式话头。

例9

原文：（选自2022年国务院政府工作报告）

{我们}加强大宗商品保供稳价，
着力解决煤炭电力供应紧张问题。

ChatGPT：

Efforts must be made to stabilize the supply and prices of bulk commodities, //*e* and to tackle the problem of coal and electricity shortages. //*e*

4.2 翻译引擎与专家译文话头共享类型和话头共享关系的对比分析

为进一步探究不同翻译主体的译文在话头共享类型和共享关系方面的整体差别，本文引入交叉熵进行差异度量。交叉熵是来度量两个概率分布之间差异程度的概念，交叉熵越大，则两个概率分布相差越大，交叉熵越小，则两个概率分布的差异越小（李郝林，2014）。例如两个样本集分别为P={}，Q={}，交叉熵的定义为：

$$\mathrm{H}(P,\ Q)=\sum_{i=1}^{n}(p_i)log(1/q_i)$$

式中，P为后验概率，Q为先验概率，H(P, Q)则为概率P和概率Q的交叉熵，本文将专家译文相应的概率作为后验概率，不同翻译引擎的相应概率作为先验概率，分别计算出专家译文和四类不同翻译引擎的话头共享类型和话头共享关系分布的交叉熵，如图1和图**??**所示，其中数据为0的对数为0，并最终对结果保留小数点后四位数。

	翻译引擎A	翻译引擎B	翻译引擎C	ChatGPT
话头共享类型	0.6789	0.6499	0.5972	0.5697
话头共享关系	0.5390	0.5734	0.5990	0.4926

表 3: 翻译引擎与专家译文话头共享类型及话头共享关系分布的对比情况

交叉熵的越大，翻译引擎和专家译文在共享类型的使用上的差异越大，根据表3可以看出翻译引擎A、B、C和ChatGPT的话头共享类型与专家译文的相似程度，其中依托于大型语言模

型的ChatGPT与专家译文的差异最小，相似程度更高；而神经网络机器翻译引擎A、B、C则与专家译文差异较大。在话头共享关系方面也呈现相似的结果，ChatGPT相较于其他三类翻译引擎，与专家译文的差异较小，相似度更高，话头的生成能力与专家译者更为接近。

4.3 话头共享模式对应和话头共享关系对应情况的调查结果

根据数据调查发现，翻译引擎ABC的译文的话头共享模式转换类型大多为与原文分支模式的类型相同的对应转换，转换为独立小句、新支模式以及后置模式的能力相对较弱，本文推测基于神经网络的翻译引擎受原文句式的影响较大，故模式较为单一，英语的地道性相对较弱。

而专家译者在话头共享模式上相对灵活，与原文共享模式相同的类型占比相较于基于神经网络的机器翻译引擎和ChatGPT都较少，转换为独立小句、新支模式和后置模式占比相对较高，表现出共享模式和句式的多样化，可读性相对较强。

ChatGPT在话头共享模式的转换能力上弱于专家译者，但相比于基于神经网络的翻译引擎，其共享模式的转换能力的能力更强，受原文影响较小，与专家译者更为接近。

而通过话头对应的类型和数量来观察基于神经网络的机器翻译和话头生成能力发现，在与原文话头完全对应占比最小的仍为专家译文，基于神经网络的翻译引擎的译文大多为与原文话头对应，呈现出大量译文与汉语原文话头完全对应的情况，在汉语主语缺失时译文也对应翻译为主语缺失的英语译文，汉语痕迹十分明显，译文的地道性、可读性上都较为欠缺，表现出话头生成能力的弱势。

基于神经网络的翻译引擎拥有一定的话头补足能力，但较为复杂的话头转换和话头变形等话头生成能力较弱，更为复杂的话头提炼较为欠缺。ChatGPT前两项能力与专家译文较为接近，话头转换能力优于基于神经网络的翻译引擎但弱于专家译文，但其话头提炼较为优越，弥补了之前的机器翻译在此项能力上的欠缺，甚至在数据上高于专家译文，本文推测ChatGPT在话头生成方面受原文影响和约束程度较小，表现出大语言模型在翻译时话头生成能力的独特优势。

5 总结

本研究对目前机器翻译在小句复合体层面的转换生成能力进行调查，通过对四类机器翻译引擎和专家译文进行话头共享类型、话头共享关系两方面的标注，并将基于神经网络的翻译引擎ABC的译文、基于大型语言模型的ChatGPT译文与专家译文进行对比和分析，不仅可以较为客观地表现出目前机器翻译在小句复合体层面的翻译生成能力，还可以总结出目前机器翻译在小句复合体层面的部分不足。翻译引擎、ChatGPT与专家译文相比，在话头共享模式和话头对应方面的翻译能力仍有提升的空间，而其中ChatGPT作为基于大型语言模型的新兴机器翻译，与前三类神经网络机器翻译相比，在话头共享类型和话头共享关系两个方面与专家译文的相似度更高，甚至在话头共享关系最复杂的“话头提炼”上，不仅弥补了机器翻译引擎在这一方面能力的欠缺，甚至使用数量上高于专家译文，话头提炼需要对文章内容进行总结归纳，需要较高的语义理解能力，而ChatGPT相比于专家译文拥有更强的话头提炼能力，展现出较强的话头生成能力，本文猜测可能是原文对于基于大语言模型的ChatGPT的约束程度要小于原文对于专家译文的约束程度，反映出ChatGPT生成式模型在翻译方面的独特优势，后续可以基于汉语无主语小句的数据集进一步开展ChatGPT话头总结能力的探究。

本文选择了44个小句复合体作为调查样本，在绝对数量上存在一定的不足，未来需进一步扩大调查规模。为弥补样本数量的不足，本文选择了多个语体文本作为调查样本，同时选择了多个翻译引擎进行调查。调查结果显示，专家译者、神经网络机器翻译系统以及ChatGPT在小句复合体层面的转换生成能力存在较大差异，机器翻译系统的复杂话头生成能力较人类专家和ChatGPT都有较大差距，未来需要进一步探究优化方法，提升机器翻译系统在小句复合体层面的译文质量。同时，本文的调查结果也显示出人类专家在小句复合体层面的转换能力具有一定的局限性，容易受到原文的约束，而降低译文的地道性，针对此类问题，需要将小句复合体层面的翻译能力作为专门对象，加以研究并开展针对性训练，为提升人的翻译能力提供支持。

致谢

本文得到宋柔教授的宝贵意见与建议，在此深表谢忱，文中所有错误缺漏全由本文作者自负。

本文是国家社科基金项目“汉英小句级对齐语料库研制与应用研究”（19BYY081）的阶段成果，同时也得到浙江省高等教育“十四五”教学改革项目（jg20220440）的支持。

参考文献

Halliday M A K. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London.

Wang L, Lyu C, Ji T, et al. 2023. Document-Level Machine Translation with Large Language Models. *arXiv preprint arXiv*:2304.02210.

Jiao W, Wang W, tse Huang J, et al. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine.

葛诗利, 宋柔. 2020. 基于成分共享的英汉小句对齐语料库标注体系研究. 中文信息学报,34(06):27–35.

胡开宝,李翼. 2016. 机器翻译特征及其与专家翻译关系的研究. 中国翻译,37(05):10–14

李晗佶,陈海庆. 2020. 机器翻译技术困境的哲学反思. 大连理工大学学报*(*社会科学版*)*,41(06):122-128.

刘满芸. 2016. 翻译技术时代翻译模式的裂变与重构. 中国科技翻译,29(04):17-20.

宋柔, 葛诗利. 2015. 面向篇章机器翻译的英汉翻译单位和翻译模型研究. 中文信息学报,29(05):125-135.

宋柔. 2023. 汉语小句复合体和话头结构. 中国社会科学出版社, 北京.

宋柔. 1985. 小句复合体的语法结构. 商务印书馆, 北京.

朱光辉, 王喜文. 2023. *ChatGPT*的运行模式、关键技术及未来图景. 新疆师范大学学报*(*哲学社会科学版*)*,44(04):113-122.

张学贞. 2022. 基于汉英日三语可比语料库的*NT*小句复合体对比研究. 青岛大学.

基于端到端预训练模型的藏文生成式文本摘要

黄硕 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 h17852656271@163.com

闫晓东* 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 yanxd3244@sina.com

欧阳新鹏& 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 2855973230@qq.com

杨金朋 中央民族大学 信息工程学院 国家语言资源监测与研究少数民族语言中心 国家安全研究院 语言信息安全研究中心 15713862215@163.com

摘要

近年来，预训练语言模型受到了广泛的关注，这些模型极大地促进了自然语言处理在不同下游任务中的应用。文本摘要作为自然语言处理中的一个重要分支，可以有效的减少冗余信息，从而提高浏览文本速度。藏文作为低资源语言，缺乏用于大规模的训练语料，藏文生成式文本摘要研究还处于起步阶段，为了解决藏文生成式文本摘要的问题，本文首次提出将端到端的预训练语言模型CMPT（Chinese Minority Pre-Trained Language Model）用于藏文生成式文本摘要研究，CMPT模型通过对其他不同低资源语言文本进行去噪和对比学习，同时为了提高编码器的理解能力，在编码器的输出层增加一个单层掩码语言模型(MLM)解码器，进行Seq2Seq的生成和理解的联合预训练。通过进一步微调可以有效地提高在藏文文本摘要任务上的性能。为了验证模型的性能，我们在自己构建的5w条藏文文本摘要数据集和公开数据集Ti-SUM上进行实验，在两个数据集上的实验表明，我们提出的方法在藏文生成式文本摘要的评测指标上有显著提升。同时，该方法不仅可以应用于藏文文本摘要任务，也可以拓展到其他语言的文本摘要任务中，具有较好的推广价值。

关键词： 预训练语言模型；藏文；文本摘要；CMPT；Seq2Seq

Abstractive Summarization of Tibetan Based on end-to-end Pre-trained Model

Shuo Huang, Xiaodong Yan*, Xinpeng Ouyang&, Jinpeng Yang
Minzu University of China
School of Information Engineering
National Language Resources Monitoring and Research Center on Minority Languages
Language Information Security Research Center
Institute of National Security MUC
h17852656271@163.com, yanxd3244@sina.com, 2855973230@qq.com, 15713862215@163.com

Abstract

In recent years, pre-trained language models have received widespread attention, and these models have greatly facilitated the application of natural language processing in different downstream tasks. Text summarization, as an important branch of natural language processing, can effectively reduce redundant information and improve the speed of text browsing. As a low-resource language, Tibetan lacks large-scale training corpus, and research on Tibetan generative text summarization is still in its infancy. In order to solve the problem of Tibetan generative text summarization, this paper

国家语委项目，多语言网络谣言检测研究，ZDI145-61
基金项目：国家自然科学基金项目（61972436）
* 通讯作者：yanxd3244@sina.com
& 同等贡献

proposes for the first time to use the end-to-end pre-trained language model CMPT (Chinese Minority Pre-Trained Language Model) for Tibetan generative text summarization research. The CMPT model denoises and compares other different low-resource language texts. At the same time, in order to improve the comprehension ability of the encoder, the encoder A single-layer masked language model (MLM) decoder is added to the output layer for joint pre-training of Seq2Seq generation and understanding. The performance on the Tibetan text summarization task can be effectively improved by further fine-tuning. In order to verify the performance of the model, we conducted experiments on the 50,000-entry Tibetan text summary dataset constructed by ourselves and the public dataset Ti-SUM. There is a significant improvement in the evaluation metrics of summaries. At the same time, this method can not only be applied to Tibetan text summarization tasks, but also can be extended to text summarization tasks in other languages, which has good promotion value.

Keywords: Pre-trained language model , Tibetan , Text summarization , CMPT , Seq2Seq

1 引言

随着互联网技术的高速发展和信息化时代的到来，人们面临的信息量呈爆炸式增长，想要高效、快速、准确地获取有价值的信息成为一大难题，文本摘要技术的出现有效缓解了这个问题。作为自然语言处理的一个重要分支，文本摘要技术可以帮助人们从文本中快速、准确获得重要信息,从而提高了人们浏览文本的效率。

根据实现技术的不同，文本摘要一般可以分为两大类：抽取式摘要和生成式摘要。抽取式摘要是指从原始文本中选择最相关的，包含最多信息的，以及最能代表整篇文章的多个句子作为文章的摘要。生成式摘要则是通过理解文本的含义和语法规则，从原始文本中生成全新的概括性文本。由于生成式摘要需要理解和创造语言表达方式，其可靠性和准确性通常比抽取式摘要要低，因此大多数研究都集中在抽取式摘要上(Gambhir and Gupta, 2017)。近年来，随着深度学习技术的不断发展，生成式摘要逐渐成为研究热点。

藏文是作为藏族人民的书面交际工具，是世界公认的成熟的文字之一，在藏语信息化处理的过程中，同样也涉及到自然语言处理的各种任务。因此藏文的文本摘要也是一个值得关注的问题。相对于中英文来说，藏文的文本摘要还面临着许多问题和困难，首先，藏语有丰富的语法和语义特征，如主谓宾语的排列、合成词和分词等，这增加了文本摘要的难度。对于机器学习算法来说，这些语言学特征需要进行更加细致和复杂的处理，以确保正确的理解和提取。其次，目前藏语文本摘要缺乏大规模的标注数据集，这使得使用监督学习方法进行文本摘要变得困难，同时缺乏标注数据也使得评估文本摘要算法的性能变得更加困难，最后对基于语义表示的方法存在一定的局限性，现有的模型无法捕捉文本中的全部语义信息，这会影响文本摘要的质量。

本文提出了使用预训练少数民族语言模型(CMPT)(Li et al., 2022)来完成藏文生成式摘要任务，首先，本文使用的CMPT模型采用中英和多种少数民族语言进行预训练，中国少数民族语言具有文化传播的相似性和邻接性特征，模型通过对不同低资源语言文本进行去噪和对比学习的预训练，提高了语言理解能力。为了提高编码器模型的理解能力，该模型参考CPT(Shao et al., 2021)的设置，在编码器输出层增加一个单层掩码语言模型(MLM)(Taylor, 1953)解码器，进行生成和理解的联合训练，在一定程度上克服了语义表示方法存在的局限性。与其他模型相比，该模型能够有效地生成藏文摘要。

本文的主要贡献如下：

1）首次将预训练语言模型用于藏文生成式文本摘要研究，取得了较好的效果，为后续的藏文生成式摘要研究提供了参考；

2）把标题作为文章摘要，并人工对数据集进行校对，构建了5万条藏文文本摘要数据集，解决藏语文本摘要缺乏大规模的标注数据集的问题。

3）训练多个模型完成藏文生成式文本摘要任务，进行结果对比分析。

2 相关工作

预训练语言模型在自然语言处理相关的下游任务上取得巨大进步，本文先梳理预训练语言模型在生成式摘要的研究进展，再介绍藏文生成式文本摘要的发展状况。

受预训练Transformer句子编码器的工作的启发，Zhang等人(Zhang et al., 2019b)提出了HIBERT进行文档编码，以及一种使用未标记数据对其进行预训练，然后将预训练的HIBERT 应用于摘要模型。同年由OpenAI开发的一种基于Transformer架构的预训练语言模型GPT-2(Radford et al., 2019)，其在大规模语料上进行了无监督的预训练，并可以通过微调适应不同的自然语言处理任务。研究人员通过对GPT-2进行微调，将其应用于生成式文本摘要任务取得优异效果。Zhang等人(Zhang et al., 2019a) 提出了一种新颖的基于预训练的编码器-解码器框架，在编码器端使用BERT 将输入序列编码为上下文表示。对于解码器端有两个阶段，在第一阶段，使用基于Transformer 的解码器来生成草稿输出序列。在第二阶段，屏蔽草稿序列的每个单词并将其提供给BERT，然后通过组合输入序列和BERT 生成的草稿表示，然后使用基于Transformer 的解码器来预测每个掩码位置的改进单词，并将该方法应用于文本摘要任务。Song等人(Song et al., 2020)期望通过改进通用单文档摘要的框架来实现生成不同文本重用比例的摘要，提出一个基于Transformer但仅包含解码器的模型来控制生成摘要的复制率，在训练和解码阶段采取了多种策略生成从完全抽取到高度生成度的不同摘要。Google Research开发的一种基于Transformer架构的序列到序列的预训练语言模型PEGASUS(Zhang et al., 2020)，以间歇句生成为预处理目标，为生成式文本摘要定制。Facebook AI在2020年提出了一个新的预训练序列到序列模型的去噪自动编码器BART(Lewis et al., 2019)，通过用任意噪声函数破坏文本来训练的，以及学习一个模型来重构原始文本，被许多研究者应用在文本摘要任务上取得了优异的效果。

由于缺乏大规模的训练语料，目前针对藏语的文本摘要研究多数停留在抽取式方法。安见才让提出基于句子抽取的文本摘要算法，将每个句子的权重分解为特征词权重和句子结构权重，根据权重挑选候选句子，然后进行平滑处理，抽取出一定质量的摘要(安见才让, 2010)。南奎娘若在此基础上又基于不同特征加权，然后根据权重进行度量来实验基于敏感信息的藏文摘要抽取(南奎娘若and 安见才让, 2016)。李维提出了两种藏文文本摘要方法，一种改进TextRank的藏文抽取式摘要生成方法。该方法将外部语料库的信息以词向量的形式融入到TextRank算法，通过TextRank与词向量的结合，把句子中每个词语映射到高维词库形成句向量，进行迭代为句子打分，并选取分值最高的句子重新排序作为文本的摘要(李维et al., 2020)；另一种是一种将抽取式摘要和生成式摘要相结合的藏文摘要生成统一模型，使用双向Bi-GRU神经网络从藏文新闻中提取句子。其次，将指针网络融入到基于注意力的seq2seq模型中生成摘要(Yan et al., 2020)，此方法为藏文生成式摘要任务提供了一个可以参考的基线。李亮通过预训练一个藏文的ALBERT模型完成藏文抽取式文本摘要任务(李亮, 2020)，主要思想是把藏文抽取式任务转化为句子分类任务，验证了预训练语言模型在藏文文本摘要任务上的有效性。

3 模型架构

3.1 模型描述

对于低资源语言，多语言预训练可以比单一语言预训练表现的更好，但是多语言模型对方言和少数民族语言建模是个很困难的问题，尽管如此，哈工大讯飞联合实验室为中国少数民族语言开发了第一个预训练语言模型CINO，该模型提供了藏语、蒙语（回鹘体）、维吾尔语、哈萨克语（阿拉伯体）、朝鲜语、壮语、粤语等少数民族语言与方言的理解能力(Yang et al., 2022)，但是在下游生成任务上表现不尽人意。受CPT工作的启发，将理解和生成任务结合到CMPT模型中，CMPT是一个基于Transformer(Vaswani et al., 2017)，在BART的基础上，加入DeepNorm预训练的超深层生成模型,支持多种语言。它有256个隐藏状态、8个注意力头、128个编码器层和128个解码器层。如图1所示，为了更好的适应理解和生成任务，对Transformer结构做了以下四部分修改:

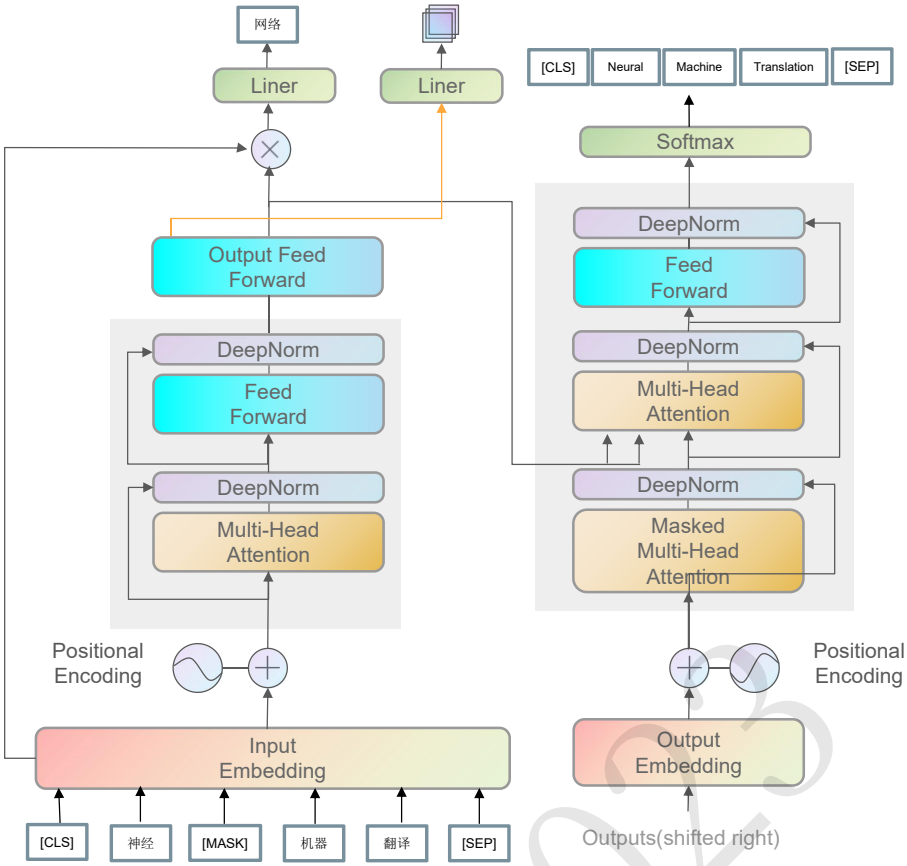


图1.CMPT(Chinese Minority Pre-Trained)语言模型结构图

1）双向编码器:使用双向自注意力编码器，它可以利用语义表示和文本含义。

2）掩码解码器：对双向编码器的输出采用单个线性层，其中输入嵌入(input embedding)乘以输出，被称为MLM头来支持MLM预训练任务的训练。

3）自回归解码器:采用交叉注意实现自回归解码。

4）相似解码器：将编码器的CLS 向量输入到单层相似度解码器中以提取语义向量。

3.2 模型预训练

本文主要研究藏文生成式摘要方法，所以对模型预训练主要介绍和生成任务相关的方法。CMPT为了更好地利用低资源语料库来学习语言知识，采取了四个具有两阶段策略的预训练任务。

1) 掩码语言模型(MLM)任务。以15%的概率随机屏蔽输入文本。并且要求Mask 解码器分别预测掩码标记，以便模型可以学习更深层次的语义信息。输入嵌入（input embedding）与编码器的输出一起用于此MLM 任务。

2) 去噪自动编码(DAE)任务。使用噪声函数来随机破坏输入文本，然后使用掩码填充相应的位置。动机是自回归解码器可以学习重建原始噪声输入。 3) 文本翻译(TT)任务。在第二阶段，将DAE任务更改为监督训练，将多语言翻译对输入到预训练的语言模型中，而MLM任务保持其原始设置。

4) 跨语言对比学习(CCL)任务。在第二阶段，添加相似度解码器来比较和学习相互翻译对的CLS输出，从而缩短具有相同语义的文本之间的向量空间距离。

在预训练阶段，首先使用了Xavier Norm(Glorot and Bengio, 2010)来初始化模型参数，其中E是编码器的层数，D是解码器的层数。

$$\alpha^{Encoder} = 0.81(E^4 \cdot D)^{\frac{1}{16}} \tag{1}$$

$$\alpha^{Decoder} = (3D)^{\frac{1}{4}} \tag{2}$$

$$\beta^{Encoder} = 0.87(E^4 \cdot D)^{-\frac{1}{16}} \tag{3}$$

$$\beta^{Decoder} = (12D)^{-\frac{1}{4}} \tag{4}$$

参考DeepNet(Wang et al., 2022)设置，模型为标准参数归一化设置α和β值:

$$std_{Encoder} = \beta^{Encoder} \times \sqrt{\frac{2}{fan\_in + fan\_out}} \tag{5}$$

$$std_{Decoder} = \beta^{Decoder} \times \sqrt{\frac{2}{fan\_in + fan\_out}} \tag{6}$$

$$W_{Encoder} \sim N(0, std_{Encoder}) \tag{7}$$

$$W_{Decoder} \sim N(0, std_{Decoder}) \tag{8}$$

其中fan_in是输入网络连接的数量，fan_out是该层输出网络连接的数量。

模型为每一层在LayerNorm中添加了残差结构。

$$Layer^{Output}_{Encoder} = LyerNorm(x \times \alpha^{Encoder} + f(x)) \tag{9}$$

$$Layer^{Output}_{Decoder} = LyerNorm(x \times \alpha^{Decoder} + f(x)) \tag{10}$$

首先使用编码器将句子编码为特征矩阵H，$H \in R^{x \times d \times t}$，然后将其输入到三个不同的解码器层。生成任务主要是应用自回归解码器，模型采用交叉注意力机制将特征矩阵H融合到自回归编码器中，注意力函数可以描述为查询(Q) 和一组键值对(K-V)映射到输出，其中Q、K、V和输出都是向量。输出可以通过值的加权和而计算得出。这些Q、K、V 之间的计算如下所示:

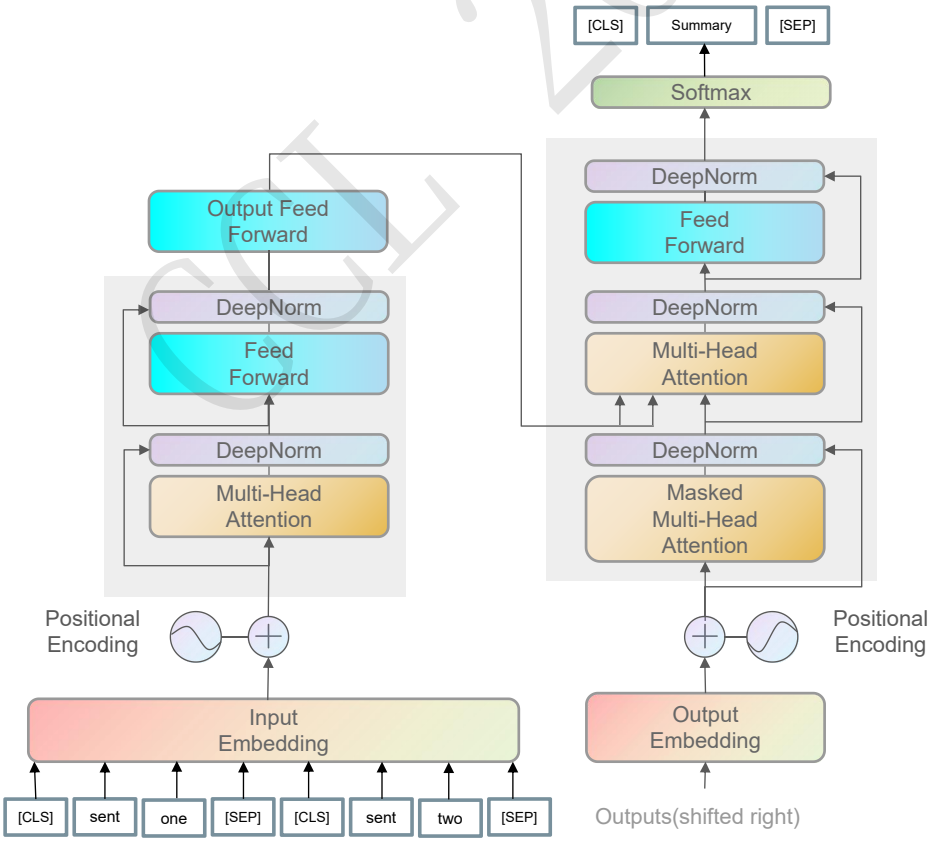


图2.用于生成式文本摘要任务的CMPT模型图结构图

$$\dot{H}_D^t = MultiHead\_Self\ Att(H_D^t\ ) \tag{11}$$

$$\ddot{H}_D^t = MultiHeadAt(\dot{H}_D^t, H, H) \tag{12}$$

$$H_D^{t+1} = LyerNorm(H_D^t \times \alpha^{Decoder} + \ddot{H}_D^t)) \tag{13}$$

其中t表示当前时间，并且整个计算被实现为进一步自回归的递归过程。CMPT模型在超过10G的汉英维藏蒙语料中进行受限预训练，最终的模型大小是390M，具有较强的理解与生成性能。

3.3 摘要任务微调

CMPT模型在翻译任务上的表现，说明预训练的CMPT模型有较好的语义理解和生成性能，我们把该模型用在藏文文本生成式摘要任务上，如图2所示，并在下游摘要任务微调阶段，使用CMPT模型的权重进行初始化，选用自回归解码器作为CMPT模型的解码器，把数据集按照8:1:1的比例划分为训练集，验证集以及测试集，为了更好的适应藏文，我们依旧选用CINO的词表，使用交叉熵作为损失函数进行实验，相比于其他方法，我们的实验结果有显著的提升。

4 实验

4.1 数据集

本文使用Python爬虫工具从香格里拉藏文网、新华网、人民网藏文版等多家新闻媒体网站上爬取58642篇藏语新闻文本作为文本摘要的原始语料，我们采用新闻标题作为参考摘要，剔除过长或过短的新闻原文以及过短的新闻标题，对于新闻标题，我们又进行了人工校验，剔除了标题较为抽象的新闻篇幅，对于藏文原始文本进行了剔除HTML标签以及标点符号过滤等数据清洗操作，最终保留了51221条语料作为实验的数据集。

4.2 超参数设置

表1展示了实验阶段的超参数设置情况。

Parameter	Value
batch_size	8
epochs	10
learning_rate	1e-04
warmup_steps	500
weight_decay	0.001
max_input_length	1024
max_targe_length	128
vocab_size	135259

表1.超参数设置

4.3 评测方法

文本摘要的评价方法分为两种：人工评价方法和自动评价方法。人工评价就是由专家对生成的摘要进行评价，但是评价成本高不利于大规模语料评测，另外人工评价带有主观性容易受外界因素干扰。自动评价是比较模型生成的摘要和参考摘要的相似度。目前，Lin等人参考机器翻译自动评测方法Bleu(Papineni et al., 2002)，提出了ROUGE(Recall-Oriented Understudy for Gisting Evaluation)评测方法(Lin, 2004)，其基本思想是通过将由一系列算法或技术自动生成的摘要或翻译与一组通常由人工生成的理想摘要或翻译进行比对，通过对两者之间的重叠单元（n元语法，单词序列和单词对）进行计数，从而得出分值，以衡量自动生成的摘要或翻译与

参考文本之间的相似性，来评价算法有效性。ROUGE系列评价指标包括ROUGE-N、ROUGE-L、ROUGE-S、ROUGE-W。最常见的评价指标是ROUGE-N,它基于n-gram共现统计,n的范围是从1到4。计算如公式(14)所示：

$$ROUGE-N=\frac{\sum_{S\in\{Refsummaries\}}\sum_{n-grams\in S}Count_{match}(n-gram)}{\sum_{S\in\{Refsummaries\}}\sum_{n-grams\in S}Count(n-gram)} \tag{14}$$

其中$Refsummaries$表示引用摘要，$Count(n-gram)$表示引用摘要中的个数，$Count_{match}(n-gram)$表示生成的摘要和引用摘要中的公用个数。

ROUGE-L是基于最长公共子串的统计，ROUGE-W所做的工作就是给连续的匹配给到更多的权重，让连续匹配的比非连续匹配的有更高的分数。ROUGE-S是ROUGE-N的一种扩展，N-gram是连续的，Skip-bigram是允许跳过中间的某些词，同时结合了ROUGE-L的计算方式。不同的方法对不同类型的总结评价有不同的影响。

4.4 实验结果和分析

本节使用CMPT预训练模型在下游藏文生成式文本摘要任务上微调，分别在我们构建的数据集和公开的Ti-SUM数据集(闫晓东et al., 2022)上进行了实验。因为目前除CMPT模型以外，还没有其他的预训练语言模型可以做藏语的生成式任务，所以我们选择了同样使用标题作为摘要且数据量为5W条的基于统一模型的藏文新闻摘要方法作为基线进行了实验结果对比，对比结果如表2所示：

	ROUGE-1	**ROUGE-2**	**ROUGE-L**
统一模型	19.81	13.27	16.90
CMPT模型（本文数据集）	49.16	33.43	48.66
CMPT模型（Ti-SUM数据集）	39.53	26.42	38.02

表2.不同模型和数据集的实验结果

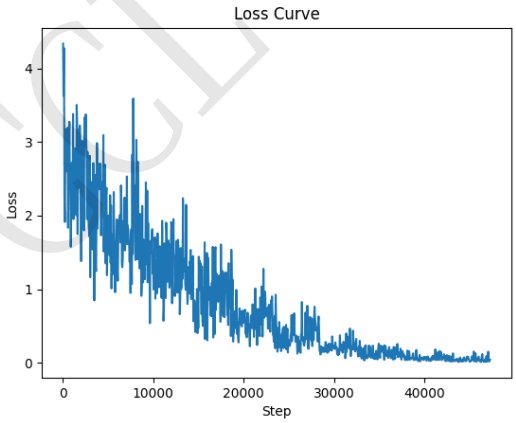


图3.CMPT模型在微调训练过程中的Loss曲线

通过实验结果分析，我们得出如下结论：

1）使用CMPT模型的方法比统一模型方法取得了更好的性能，ROUGE评分分别提高了29.35、20.16、31.76，我们认为CMPT能够取得如此优异的表现主要取决于模型对不同低资源语言文本进行去噪和对比学习的预训练，并在编码器输出层增加一个单层掩码语言模型解码器，进行生成和理解的联合训练，使得CMPT模型具有强大的语义理解和文本生成能力。

2）在Ti-SUM数据集CMPT模型的评测结果略有下降，相较于本文自己构建的数据集ROUGE评分分别下降了24.21%、27.48%、25.53%。对于这样的结果，我们分析认为可能

是两方面原因导致的，首先，Ti-SUM数据集的特点是摘要通常为两到三句，我们参数在设置的时候生成最大长度为128，这样的设置会影响到ROUGE评分的结果。其次，Ti-SUM数据集的数据量只有1000条，微调数据量太小可能会导致模型过拟合Ti-SUM数据集，而忽略了预训练过程中所学到的更广泛的知识，我们将会在4.5节具体展开分析不同数据量对模型的影响情况。

3）我们发现，在使用本文自己构建的数据集训练过程中，CMPT模型的Loss曲线前期震荡幅度较大，我们分析可能有以下原因导致，首先是学习率调整的问题，微调大型模型时前期学习率大，模型参数更新幅度大，使得loss的值波动比较大。其次当从预训练模型切换到微调阶段时，输入数据的分布通常会发生变化，这可能包括数据集的不同领域、任务的不同类别等，模型需要适应新的数据分布，因此可能会导致前期loss的不稳定性。最后模型参数的初始值也会对训练过程和loss 曲线的动荡性产生影响。

4.5 分析数据量对模型的影响

为了分析不同数量的数据对模型的影响，我们把原始的数据集随机抽取成6份（1000/5000/10000/20000/30000/40000），然后进行实验，为了避免随机抽样的影响，我们用不同的样本重复了每个实验5次，并报告了它们的平均结果，如图4-5所示。 根据这组数据，

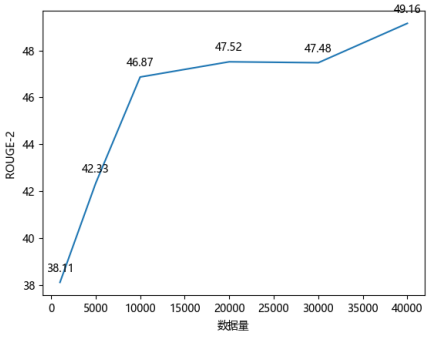


图4.不同数据量的ROUGE-1得分

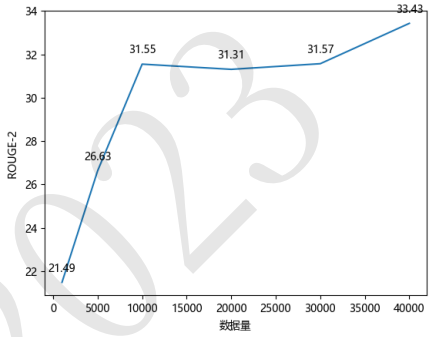


图5.不同数据量的ROUGE-2得分

可以发现随着数据量的增加，ROUGE得分有所提高，这可能是因为模型在更大的数据集上进行训练，获得更多的语言特征，从而提高了文本摘要的准确性。但是在1000-10000数据量时，ROUGE值大幅度增加，这可能是因为更大的数据集提供了更多的数据多样性，使得模型能够更好地捕捉到藏文生成式文本摘要任务中存在的一些规律和特征，从而提高了模型的性能，在这之后，随着训练数据量的增加，ROUGE得分逐渐趋于平缓，这说明此时模型的训练数据的规模已经不再是限制模型的重要因素，想要继续提高模型的性能，可以从其他因素继续研究。

4.6 分析数据长度对模型的影响

此外，本文进一步探究了文本长度对模型生成摘要效果的影响，我们挑选了文本长度在500-800、1000-1300、1500-1800和2000-2500的数据各2000条，重新对模型进行训练，同样，为了避免随机抽样的影响，我们用不同的样本重复了每个实验3次,实验结果取平均，实验结果如表3所示。 实验结果表明，文本的长度也是影响模型生成效果的一个重要因素。

	ROUGE-1	ROUGE-2	ROUGE-L
500-800	41.98	25.47	40.68
1000-1300	34.73	16.23	32.65
1500-1800	31.19	16.63	31.12
2000-2500	32.20	16.59	30.79

表3.不同数据长度的的实验结果

从表中可以看出，当文本长度小于1024时，ROUGE-1值稳定在42左右，而当文本长度大于1024时，ROUGE得分开始降低，模型生成摘要的效果明显变差。我们认为文本长度影响模型生成效果的主要原因是模型最大能编码的序列长度为1024，在对长文本进行生成时，模型无法编码完整的文本数据，会有一定的重要信息丢失，模型不能捕获到重要信息，导致生成效果不佳。在文本长度大于1024时，随着文本长度的继续增加，ROUGE评分波动幅度不大，这可能是因为我们用来构建数据集的文本主要为新闻文本，然而新闻文本的重要信息主要集中在文章的开头部分，为了验证我们的猜想，首先我们选取了5000条藏文新闻文本数据，并对文章分句编号处理，然后使用贪婪的方法把文章的每个句子和标题做ROUGE计算，我们把ROUGE得分作为评价文中句子重要程度的指标，并把ROUGE得分排序，返回句子的编号，结果如图6所示。接下来我们会继续增加数据集文本的类别，提高数据集的质量，推动藏文生成式文本摘要的研究。

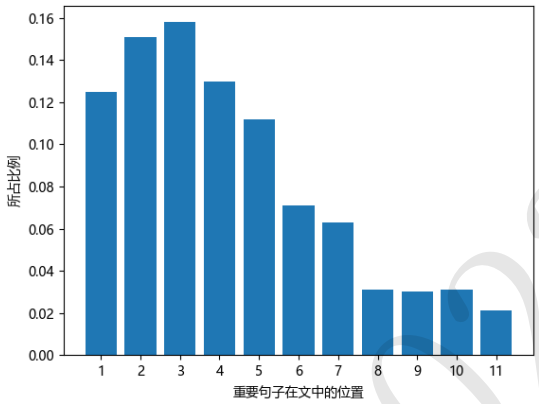


图6.重要句子在文中的位置所占比例

原文：

ཞིན་ཧྭ་གསར་འགྱུར་ཁང་གི་གློག་འཕྲིན་ལྟར།ཚེས7ཉིན་གྱི་ཕྱི་དྲོའི་ཆུ་ཚོད3པར།རྒྱལ་ཡོངས་དམངས་ཆེན་སྐབས་བཅུ་བཞི་པའི་གྲོས་ཚོགས་ཐེངས་དང་པོས་མི་དམངས་ཚོགས་ཁང་ཆེན་མོ་ནས་ཚང་འཛོམས་གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་ཤིང་།རྒྱལ་ཡོངས་དམངས་ཆེན་རྒྱུན་ཨུ་ལྷན་ཚོགས་ཀྱི་ཨུ་ཡོན་ཀྲང་ལི་ཀྲན་ཧྲུའུ་ཡིས་ཕུལ་བའི་རྒྱལ་ཡོངས་མི་དམངས་འདུས་མི་ཚོགས་ཆེན་རྒྱུན་ལས་ཨུ་ཡོན་ལྷན་ཁང་གི་བྱ་བའི་སྐོར་གྱི་སྙན་ཞུ་ལ་མཉན་པ་དང་།ཆེས་མཐོ་མི་དམངས་ཁྲིམས་ཁང་གི་ཡོན་ཀྲང་ཀྲོའུ་ཆང་གི་ཆེས་མཐོ་མི་དམངས་ཁྲིམས་ཁང་གི་བྱ་བའི་སྐོར་གྱི་སྙན་ཞུ་ལ་མཉན་ལ།ཆེས་མཐོ་མི་དམངས་ཞིབ་དཔྱོད་ཁང་གི་ཞིབ་དཔྱོད་དཔོན་ཀྲང་ཅུན་གྱི་ཆེས་མཐོ་མི་དམངས་ཞིབ་དཔྱོད་ཁང་གི་བྱ་བའི་སྐོར་གྱི་སྙན་ཞུ་ལ་མཉན།རྒྱལ་སྲིད་ཨུ་ཡོན་ནམ་རྒྱལ་སྲིད་སྤྱི་ཁྱབ་ཁང་གི་དྲུང་ཡིག་ཆེན་མོ་ཞའོ་ཅེ་ཡི་རྒྱལ་སྲིད་སྤྱི་ཁྱབ་ཁང་གི་སྒྲིག་གཞི་བཅོས་བསྒྱུར་ཇུས་གཞི་སྐོར་གྱི་གསལ་བཤད་ལ་མཉན།

据新华社电。7日下午3时。***十四届全国人大一次会议在人民大会堂举行第二次全体会议***。听取了全国人大常委会委员长栗战书关于全国人大常委会工作的报告。听取最高人民法院院长周强关于最高人民法院工作的汇报。听取最高人民检察院检察长张军关于最高人民检察院工作的报告。听取国务委员、国务院秘书长肖捷关于国务院机构改革方案的说明。

参考摘要：

རྒྱལ་ཡོངས་དམངས་ཆེན་སྐབས་བཅུ་བཞི་པའི་གྲོས་ཚོགས་ཐེངས་དང་པོའི་ཚང་འཛོམས་གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་པ། གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་པ།

十四届全国人大一次会议第二次全体会议举行第二次会议的召开

生成摘要：

རྒྱལ་ཚོགས་ཆེན་སྐབས་བཅུ་བཞི་པའི་གྲོས་ཚོགས་ཐེངས་དང་པོས་ཚང་འཛོམས་གྲོས་ཚོགས་ཐེངས་གཉིས་པ་འཚོགས་པ།

十四届***国会***一次会议举行第二次全体会议

图7.生成摘要实例

4.7 摘要生成实例分析

从图7生成的摘要结果分析，我们的方法可以理解文本语义信息，能够捕捉到文本重要信息的所在位置，生成可读性和准确性较强的摘要，这主要归功于模型进行了生成和理解的联合预训练，又通过足够多的语料进行了监督训练，使得模型更好地学习数据的特征，从而提高模型

的泛化能力。另外生成的摘要里存在原文没有出现的词，这可能是因为模型在生成摘要时出现了一些语法或语义的歧义，导致生成的摘要略有瑕疵，接下来我们会针对这个问题对模型的解码端继续改进。

5 总结

文本摘要是自然语言处理的重要分支，在藏语信息化进程中，也需要跟上深度学习发展的步伐。本文提出使用预训练CMPT模型解决藏文生成式文本摘要问题，通过对不同低资源语言文本进行去噪和对比学习的预训练，并在生成式文本摘要任务上进行微调，通过在不同的数据集上的效果表明，生成的摘要具有较强的原文相关性和可读性。但是仍有许多问题亟待解决，首先，使用标题做参考摘要存在部分标题不足以总结全文的问题。其次，ROGUE评测主要基于匹配单词、短语等方式计算，与语法和语义的理解存在局限性。接下来，我们会继续完善藏文文本摘要的数据集，以及改进该模型并将其拓展到其他低资源语言的文本摘要任务中，同时，我们还要预训练一个支持蒙、藏、维三种少数民族语言的T5模型，致力推进中国少数民族语言文本摘要的发展。

参考文献

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022. A multi-tasking and multi-stage chinese minority pre-trained language model. In *Machine Translation: 18th China Conference, CCMT 2022, Lhasa, China, August 6–10, 2022, Revised Selected Papers*, pages 93–105. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8902–8909.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

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

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. Deepnet: Scaling transformers to 1,000 layers.

Xiaodong Yan, Xiaoqing Xie, Yu Zou, and Wei Li. 2020. 基于统一模型的藏文新闻摘要(abstractive summarization of tibetan news based on hybrid model). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 479–490.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. *arXiv preprint arXiv:2202.13558*.

Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019a. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

南奎娘若and 安见才让. 2016. 基于敏感信息的藏文文本摘要提取的研究. 网络安全技术与应用, (4):58–59, 1.

安见才让. 2010. 藏文搜索引擎系统中网页自动摘要的研究. 微处理机, 31(5):77–80, 1.

李亮. 2020. 基于albert 的藏文预训练模型及其应用.

李维, 闫晓东, and 解晓庆. 2020. 基于改进textrank 的藏文抽取式摘要生成. 中文信息学报, 34(9):36–43.

闫晓东, 王羿钦, 黄硕, 杨金朋, and 赵小兵. 2022. 藏文文本摘要数据集. 中国科学数据*(*中英文网络版*)*.

融合多粒度特征的缅甸语文本图像识别方法

何恩宇[1,2]，陈蕊[1,2]，毛存礼[*1,2]，黄于欣[1,2]，高盛祥[1,2]，余正涛[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
2329863182@qq.com, 1226211036@qq.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, ztyu@hotmail.com

摘要

缅甸语属于东南亚低资源语言，缅甸语文本图像识别对开展缅甸语机器翻译等任务具有重要意义。由于缅甸语属于典型的字符组合型语言，一个感受野内存在多个字符嵌套，现有缅甸语识别方法主要是从字符粒度进行识别，在解码时会出现某些字符未能正确识别而导致局部乱码。考虑到缅甸语存在特殊的字符组合规则，本文提出了一种融合多粒度特征的缅甸语文本图像识别方法，将较细粒度的字符粒度和较粗粒度的字符簇粒度进行序列建模，然后将两种粒度特征序列进行融合后利用解码器进行解码。实验结果表明，该方法能够有效缓解识别结果乱码的现象，并且在人工构建的数据集上相比“VGG16+BiLSTM+Transformer”的基线模型识别准确率提高2.4%，达到97.35%。

关键词： 缅甸语文本图像识别；多粒度识别；字符簇

Burmese Language Recognition Method Fused with Multi-Granularity Features

Enyu He[1,2], Rui Chen[1,2], Cunli Mao[*1,2], Yuxin Huang[1,2], Shengxiang Gao[1,2], Zhengtao Yu[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
2329863182@qq.com, 1226211036@qq.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, ztyu@hotmail.com

Abstract

Burmese is a low-resource language in Southeast Asia, and Burmese text image recognition is of great significance for carrying out tasks such as Burmese machine translation. Since Burmese language is a typical character-combined language, there are multiple character nests in one receptive field. The existing Burmese language recognition methods mainly recognize character granularity, and some characters may not be recognized correctly during decoding, resulting in partial Garbled characters. Considering that there are special character combination rules in Burmese, this paper proposes a Burmese text image recognition method that integrates multi-granularity features. The two granular feature sequences are fused and then decoded by a decoder. The experimental results show that this method can effectively alleviate the phenomenon of garbled recognition results, and the recognition accuracy rate of the baseline model of "VGG16+BiLSTM+Transformer" is increased by 2.4% to 97.35% on the artificially constructed dataset.

*毛存礼（通讯作者）：maocunli@163.com
国家自然科学基金（62166023，U21B2027，61972186），云南省科技重大专项（202103AA080015，202203AA080004），云南省基础研究计划项目（202201AT070858）

Keywords: Burmese text image recognition , Multi-granularity recognition , character clusters

1 引言

缅甸语是一种东南亚低资源语言，其文字具有独特的形态和结构。随着数字化技术的迅速发展，缅甸语文本图像识别逐渐成为了一个重要的研究领域。缅甸语文本图像识别可以帮助我们将印刷或手写的缅甸语文本转换为可编辑的数字形式，这对于数字化文献、信息检索和自然语言处理等领域都具有重要的应用价值。然而现有的缅甸语识别模型在识别缅甸语时，由于缅甸语中存在着大量组合字符，导致识别过程中易发生漏识、错识某些关键字符，容易出现乱码的现象。

(a) 含上下标图像

(b) 含左拼/右拼字符图像

Figure 1: 缅甸语字符组合规则示例

缅甸语字符编码顺序以及组合规则和中英文字符存在较大的差异。中英文字符编码遵循从左到右的顺序，同时中英文不存在组合字符的现象，然而由于缅甸语字符包括辅音字符、左拼元音字符、上拼元音字符、右拼元音字符、下拼音调字符/后拼音调字符、后拼其他字符。由于缅甸语中大量存在组合字符，如Figure1中的(a)的“ငှဲ့”是由辅音字符“င”、下拼音调字符“့”和上拼元音字符“ှ”组成的。我们将纵向堆叠的字符定义为一个字符簇，切分后的结果如表1中示例所示。同时缅甸语编码顺序和视觉上呈现的顺序一致。如Figure1中的(b)的字符簇“ဖတ့်”是由辅音字符“တ”、左拼元音字符“ဖ”、上拼元音字符“ိ”和下拼音调字符“့”组成的。虽然我们看到该字符簇的第一个字符为左拼元音字符“ဖ”，但是在文本编码端“ဖ”其实是该字符簇的第二个字符。在传统CRNN(Shi et al., 2016)网络中，一个感受野只会对应一个特征，但在缅甸语文本图像中一个感受野往往包含多个缅甸语字符，按照单个字符解码的方式会导致一些关键字符的丢失，从而出现局部乱码的现象。

缅甸语文本	字符级	字符簇级
တထေရာတည်း	တ ေ ထ ရ ာ တ ည ် း	တ ေ ထ ရာ တ ည် း
နေဝင်မှနေထွက်ချိန်။	ေ န ဝ င ် မ ့ ေ န ထ ွ က ် ခ ျ ိ န ် ။	ေ န ဝ င် မ့ ေ န ထွ က် ချ ိ န် ။
အောက်ခြေ	ေ အ ာ က ် ေ ခ ြ	ေ အာ က် ေ ခြ

Table 1: 缅甸语文本切分实例

此外，研究人员针对缅甸语文本图像的识别任务做了一些尝试，毛存礼et al. (2022)提出了利用知识蒸馏的方式，将教师网络学习到的单个字符的特征传递给学生网络，以提升学生网络对于缅甸语字符的特征提取能力，从而在一定程度上解决了缅甸语文本图像识别过程中的某些字符丢失的情况，但是该方法忽略了深度卷积神经网络的底层语义信息及相关特征。Liu et al. (2021)提出一种基于融合多层语义特征的缅甸语文本图像识别方法，将特征提取层提取到的多层缅甸语特征进行融合，达到提高主干网络对缅甸语文本图像的特征提取的能力的目的。但是该主干网络在提取缅甸语文本图像的边缘特征上表现的并不是很好，同时模型的编解码效率较低。Wang et al. (2022)提出一种融合通道注意力和空间注意力的缅甸语文本图像识别方法，在提取图像特征的同时构建空间注意力和通道注意力，最后利用多头注意力机制对融合结果进行注意力计算，但是该方法忽略了缅甸字符的组合规则，在真实场景应用中依然会出现由于识别结果中某些缅甸语字符丢失导致的乱码现象。

为了解决上述问题，本文针对某些关键字符识别丢失或者出现识别错误导致的局部乱码现象，受到缅甸语分词林颂凯et al. (2018)的启发，本文将纵向堆叠字符作为一个字符簇，提出一

种融合多粒度特征的缅甸语文本图像识别方法，在编码端获得字符粒度和字符簇粒度的两种特征序列，然后将两种粒度的特征序列进行融合，然后将利用字符组合规则切分得到的字符簇作为解码字典，最后通过解码器得到识别结果。

本文工作主要有以下贡献：

(1)我们提出了一种融合多粒度特征的缅甸语文本图像识别方法，使用了字符簇粒度的编解码字典，解决了现有缅甸语识别模型未能正确识别某些字符导致的乱码现象。

(2)我们在图像特征提取层做了改进，分别提取到字符粒度和字符簇粒度的图像特征并进行序列建模，将两种粒度特征序列进行融合后利用解码器进行解码，提高了缅甸语识别模型的精度。

(3)在人工构建的的缅甸语文本图像数据集上，实验结果表明所提方法的缅甸语文本图像识别的精度达到97.35%，优于多个对比模型。

2 相关工作

(1)基于联结主义时间分类的文本图像识别方法

基于联结主义时间分类（Connectionist Temporal Classification, CTC）的文本识别方法，使用CTCLoss作为目标优化函数。该算法的核心思想是定义如何将预测结果转化为真实标签，并使用动态规划算法从输出概率分布中获取多条状态转移路径，将路径概率之和或最大值作为目标优化函数。因此，CTC算法可以进行端到端的训练，只需要输入文字级标签，而不需要字符级标签。这种方法使得文本识别的训练更加高效，同时减少了标注数据的成本。Shi et al. (2016)利用卷积神经网络（Convolutional Neural Networks, CNN）提取文本图像中的文本特征，利用循环神经网络（Recurrent Neural Networks, RNN）对特征进行编码，提出了一种将CNN和RNN相结合的识别模型（CRNN）。通过CRNN将文本图像转化为特征序列，然后通过长短时记忆（Long Short-Term Memory, LSTM）增强上下文的语义建模，最后将输出的特征序列输入到CTC模块进行解码得到最后的识别结果。Chandio et al. (2022)使用视觉几何组网络（Visual Geometry Group Network, VGGNet）来提取图像特征，使用基于RNN 的结构将特征序列解码为概率分布，最后，将CTC 函数应用于RNN序列之上，以将每帧预测转换为标签的目标序列。Bhatt et al. (2023)提出了一种混合模型，将新颖的属性签名表示（表征单词中基本视觉形状和字符的出现和位置）与CTC 框架中的LSTM 结合在一起。然而CTC算法假设每个时间片都是相互独立的，但在OCR中，相邻几个时间片中往往包含着高度相关的语义信息，它们并非相互独立的，这种特性使得CTC算法存在一定的缺陷。

(2)基于注意力机制的文本图像识别方法

基于注意力机制的文本识别方法首先使用编码器将文本图片转化为中间语义特征。接着，基于注意力模型的解码器可以将这些中间语义特征转化为识别结果。这种方法可以学习任意长度序列之间的对齐关系，从而减轻了序列对齐的问题。Wojna et al. (2017)使用CNN特征提取器处理图像，然后通过注意力机制进行加权，然后将加权后的数据传递给RNN进行解码。Liao et al. (2019)将文本图片编码为二维特征，用一个结合注意力机制的全卷积网络做像素级的分类，再用一个后处理模块输出字符序列以实现任意形状文本的识别。Zhong et al. (2022)首先利用语义生成对抗网络（Generative Adversarial Network, GAN）生成简单的语义特征，然后利用平衡注意力模块对场景文本进行识别。但是此类模型的计算量较大、对图像噪声和畸变的鲁棒性不高、对长文本的识别效果不佳以及对于语义的理解能力有限。

(3)基于Transformer的文本图像识别方法

如今Vision Transformer(Dosovitskiy et al., 2020)在计算机视觉取得了广泛的应用。在文本图像识别方面，CNN在长依赖建模上存在局限性。而Transformer因为可以在提取特征的同时关注到全局的信息，解决了这一问题。Zhao and Gao (2022)等人使用DenseNet作为图像的文本特征提取网络，将提取到的特征通过Transformer进行编解码，结合注意力精炼模块（Attention Refinement Module, ARM）得到最终输出。Xie et al. (2022)等人分别利用角点检测器和传统图像特征提取网络得到角点特征和图像特征，然后将图像的特征将通过多头自注意力机制进一步建模全局特征，同时角点图的特征将通过多头交叉注意力机制与图像全局特征融合，编码器的输出和字符序列Embedding输入Transformer解码器获得特征序列，得到最终输出。Li et al. (2021)首先将输入文本图像调整为384×384，然后将图像分割成16x16patch的序列，送入Transformer中进行编解码并最终得到输出。

以上方法均为本文解决缅甸语文本图像识别任务提供了较好的思路，本文的主要方法与现有方法的主要区别是提出了一种融合多粒度的编解码方法，并在识别的过程中融合了缅甸语的字符组合规则，进而缓解了在识别缅甸语文本图像时由于某些关键字符丢失导致的乱码现象。

3 模型架构

本文提出的识别模型架构如Figure2所示，整个模型分为基于视觉几何组网络VGGNet(Sengupta et al., 2019)的图像特征提取模块、多粒度图像特征编码模块、多粒度特征融合模块、缅甸语文本解码模块。图像特征提取模块将文本图像进行特征提取，多粒度图像特征编码模块将多粒度图像特征进行序列建模，多粒度特征融合模块将字符粒度和字符簇粒度的序列进行融合、缅甸语文本解码模块将融合特征序列解码得到缅甸语文本输出。

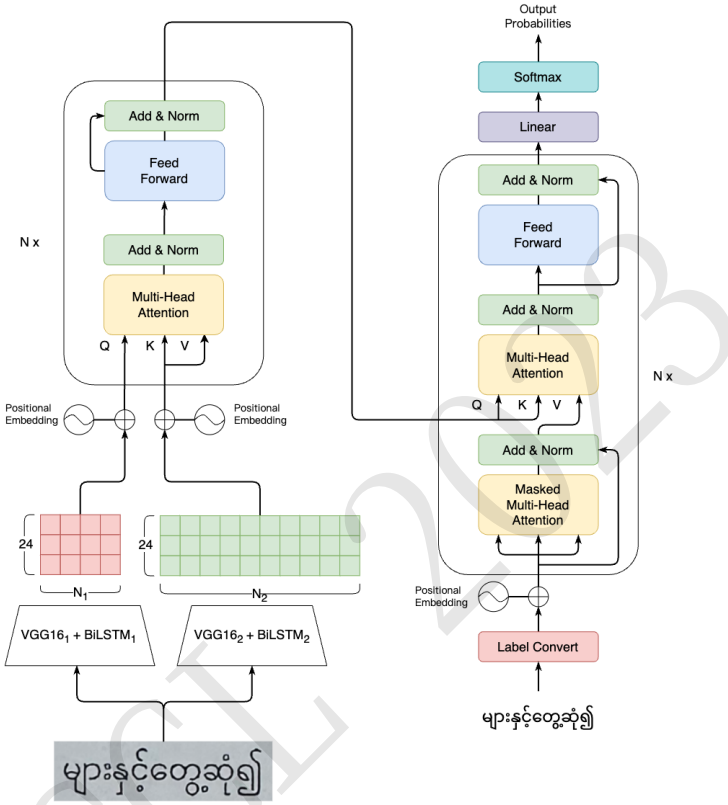


Figure 2: 融合多粒度特征的缅甸语识别模型框架

3.1 缅甸语字符簇字典构建

我们基于缅甸语拼写规则构建了字符簇单独编码字典。以字符簇为单位的编解码字典的构建：我们研究缅甸语拼写规则后发现，缅甸语文本组成的基本单位是缅甸语音节。其中每个缅甸语音节是由辅音字符为主题，然后加上左拼、上拼、右拼元音字符或其他上拼、右拼字符组成。参考现有中英文OCR识别方法不难发现，其编码顺序均是将文字序列从左往右依次编码，然而缅甸语中的编码规则却和中英文的编码方式有着较大的差异，一般来说，缅甸语的编码规则为：辅音字符、左拼元音字符、上拼元音字符、上拼其他字符、右拼元音字符/右拼其他字符。同时我们使用一些算法规则并以字符簇为单位，将缅甸语文文本从左到右切分为纵向字符簇。切分字符簇伪算法如下所示。

3.2 缅甸语文本图像特征提取模块

为了提取到字符粒度的图像特征和字符簇粒度的图像特征，我们在VGGNet的基础上分别构建了适应于提取缅甸语字符粒度和字符簇粒度的特征提取网络。通过特征提取网络分别得到512维的字符粒度特征$X_1 \in R^{C\times H\times W}$和字符簇粒度特征$X_2 \in R^{C\times H\times W}$,其中$C$，$H$，$W$ 分别为通道数、高度和宽度。

Algorithm 1 缅甸语切分字符簇伪算法

Input: တထေရာတည်း

Output: တ ေ ထ ရ တ ည် ်း

```
lst = list( တထေရာတည်း )
clusterList = [" "]
for index in range(len(lst)) do
    if (clusterList[-1] + lst[index]) is not a cluster then
        clusterList.add(lst[idx])
    else
        clusterList[-1] += lst[index]
    end if
end for
return တ ေ ထ ရ တ ည် ်း
```

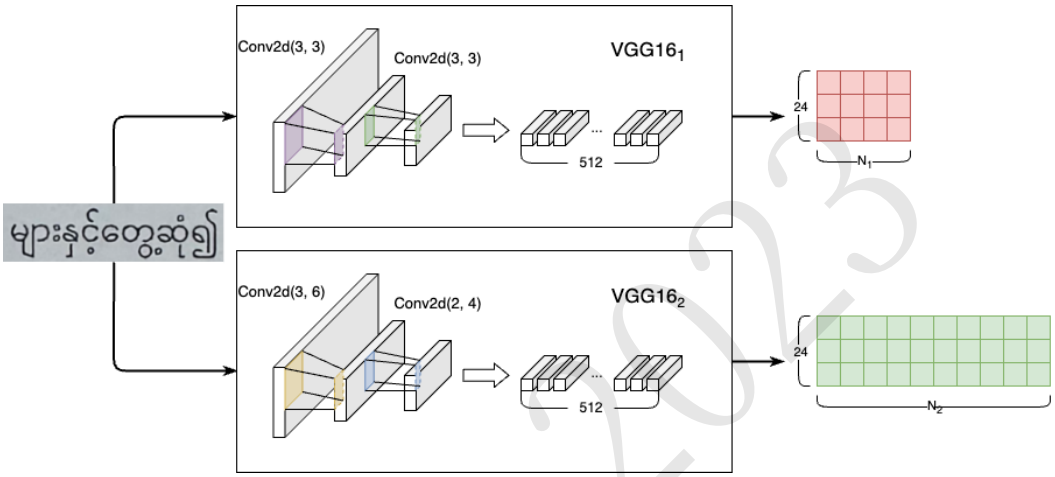


Figure 3: 缅甸语文本图像特征提取模块

考虑到字符簇粒度在纵向上的分布相比于横向分布的占比更大，考虑使用横纵比为1：2的卷积核来提取字符簇粒度的图像特征。

如Figure3中所示，$VGG16_1$为提取字符粒度特征的特征提取网络，$VGG16_2$为提取字符簇粒度特征的特征提取网络。$VGG16_1$将输入图片分别经过(3×3)、(2×2)的卷积核以提取缅甸语文本图像的字符粒度的图像特征。$VGG16_2$将输入图片分别经过(3×6)、(2×4)、(3×3)、(2×2)的卷积核以提取缅甸语文本图像的字符簇粒度的图像特征。

3.3 缅甸语文本图像特征编码模块

为了更好地对文本图像的文本特征进行建模，排除图像中噪声、畸变等条件的干扰，从而获取质量更好的图像特征，我们使用BiLSTM(Bi-directional Long Short-Term Memory, BiLSTM)对通过特征提取网络获取到512维的缅甸语文本图像特征图进行建模，以提高模型对图像噪声和畸变的鲁棒性以及对缅甸语文本图像的表征能力。

$$F_1 = BiLSTM_1(X_1) \tag{1}$$

$$F_2 = BiLSTM_2(X_2) \tag{2}$$

$F_1 \in R^{C \times B \times N_1}$，$F_2 \in R^{C \times B \times N_2}$。其中$C$，$B$，$N_1$，$N_2$分别为通道数、最大字符预测长度、字符级编解码字典长度和字符簇级编解码字典长度。

3.4 多粒度特征融合模块

为了使用字符粒度的特征来优化字符簇粒度的特征，我们使用基于Transformer的多粒度特征融合模块来将$F_1 = \{v_1, v_2, ..., v_{N_1}\}$、$F_2 = \{V_1, V_2, ..., V_{N_2}\}$进行融合，其中$v_i \in R^{B \times N_1}$，$V_i \in R^{B \times N_2}$。

我们的模型采用了Multi-Attention来对视觉特征向量进行编码。由于视觉特征向量缺乏位置信息，因此我们使用Transformer中的位置编码方法来对其进行编码。在进行位置编码之前，我们先将视觉特征向量按维度大小进行压缩，得到两个视觉特征向量F_1和F_2，它们的大小分别为(C, W_1)和(C, W_2)。为了让注意力机制更加有效，同时让F_1和F_2在水平方向上具有位移不变性，我们采用了一种基于正弦和余弦函数的位置编码方式，该方式已在Vaswani et al. (2017)的研究中得到了证明。

$$TE_1(pos_1, 2i) = sin\left(\frac{pos_1}{10000^{2i/f}}\right) \tag{3}$$

$$TE_1(pos_1, 2i+1) = cos\left(\frac{pos_1}{10000^{2i/f}}\right) \tag{4}$$

$$TE_2(pos_2, 2i) = sin\left(\frac{pos_2}{10000^{2i/f}}\right) \tag{5}$$

$$TE_2(pos_2, 2i+1) = cos\left(\frac{pos_2}{10000^{2i/f}}\right) \tag{6}$$

其中，$pos_1 \in \{0, 1, 2, ..., N_1-1\}$，$pos_2 \in \{0, 1, 2, ..., N_2-1\}$，$i \in \{0, 1, 2, ..., c-1\}$。将$F_1$、$F_2$和位置编码融合得到$F_1'$、$F_2'$，为了使用$F_1'$优化$F_2'$，我们使用交叉注意力对$F_1'$和$F_2'$进行融合。该注意力模块输入为$Q$，$K$，$V$。这里我们将$F_1'$ 作为Q，F_2'作为K，V。

$$F_{fusion} = Softmax\left(\frac{Q^i \times K^T}{\sqrt{c}}\right)V \tag{7}$$

3.5 缅甸语文本解码模块

文本解码模块将视觉特征F_{fusion}转化为字符，关注视觉特征、从文本特征中学到特定的语言知识以及隐式的学习到相关的字符组合规则。文本解码模块由4个Transformer解码器组成。使用Transformer解码器而不使用RNN作为解码器的原因是：RNN在解码的时候是串行解码的，并且RNN在某一时刻的输出依赖上一时刻的输出。我们通过将视觉特征F_{fusion}通过Transformer模块得到最终的预测序列F_{pred}。最后将F_{pred}输入解码器Convert得到对应的缅甸语文本text。

$$F_{pred} = Transformer\left(F_{fusion}\right) \tag{8}$$

$$text = Convert(F_{pred}) \tag{9}$$

模型训练时我们使用交叉熵损失作为整个模型的损失函数。

$$Loss_{attn} = -\sum \ln P(y_t|M, \theta) \tag{10}$$

其中，M为输入的缅甸语文本图像，θ为当前识别网络的模型参数，$y_t|M$为缅甸语文本图像的第t个特征序列对应的标签。

4 实验结果与分析

为了验证融合多粒度的缅甸语文本图像识别方法的有效性，我们在缅甸语文本图像数据集上进行实验分析。

4.1 数据集

本文中使用了自构的缅甸语文本图像数据集进行实验，因为缅甸语属于南亚东南亚低资源语言，目前没有公开的缅甸语数据集。该数据集包括800万张缅甸语文本图像，其中大约3万张是人工标注的，其余数据是通过算法合成的。合成数据考虑了不同的背景、噪声和倾斜度等因素，以最大程度模拟真实场景中的数据情况。为了验证模型的有效性，我们随机抽取了20万张图像作为测试数据集和验证数据集。为了提高模型的训练速度，我们使用“.mdb”文件来存储数据集。具体规模如表2所示。

实验采用评价指标为序列率精确率（Sequence Accuracy, SA），如公式11所示：

$$SA = \frac{SL}{LN} \times 100\% \tag{11}$$

其中，SA、SL、LN分别表示缅甸语文本图像识别的字符串正确率、正确识别文本图像中的字符长度、正确文本长度。

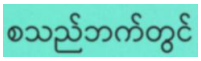
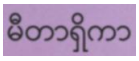
数据集	数量	样例	标签
训练集	800万		အနားယူပါ။
测试集	20万		စသည်ဘက်တွင်
验证集	20万		မီတာရှိကာ

Table 2: 缅甸语文本图像数据集样例及实例

4.2 实验结果分析

为了评估融合多粒度特征的缅甸语文本图像识别方法的有效性，我们在自构的缅甸语文本图像数据集上进行了实验。所有模型在相同的实验环境下进行训练和测试。我们使用Adam优化器，初始学习率设置为1，并使用CosineAnnealing策略逐渐减小学习率。批处理大小设置为128，训练步长为1,000,000。实验结果以测试集的字符错误率（CER）和词错误率（WER）表示，同时选择训练过程中精度最高的模型作为最终结果。实验结果如表3所示。

实验一：主要实验结果及分析

本文在自构的缅甸语文本图像数据集上进行了实验，同时与当前主流识别模型的实验结果做了对比。

CNN+BiLSTM+CTC(Shi et al., 2016)：该文本图像识别方法首先通过卷积神经网络提取文本图像的特征，接着采用BiLSTM网络融合特征向量来捕获字符序列中的语义信息。然后，对每一列特征进行分类，得到其概率分布。最后，采用CTC方法对概率分布进行预测，以得到最终的文本序列。

CNN+BiLSTM+Attention(Baek et al., 2019)：解码部分采用基于注意力机制的解码器对序列进行解码。

CNN+Transformer：使用CNN从文本图像中提取出高层次的特征表示，利用Transformer进行序列建模，学习特征之间的依赖关系，得到更加准确的文本序列预测结果。

毛等人(毛存礼et al., 2022)：构建了教师网络和学生网络，并利用卷积神经网络和循环神经网络的框架进行特征提取和序列建模，教师网络和学生网络共同进行训练，以提高模型的泛化能力和减少过拟合风险。

刘等人(Liu et al., 2021)：利用深度卷积网络获取多层语义特征图并进行融合，以缓解上下标字符特征丢失的问题；使用MIX UP的训练策略进行模型的训练，以提高模型的鲁棒性和泛化能力。

王等人(Wang et al., 2022)：提出了融合通道注意力和空间注意力的缅甸语文本图像识别方法，在提取图像特征的同时构建空间注意力和通道注意力，最后利用多头注意力机制对融合结果进行注意力计算，得到文本图像的预测。

具体方法	SA(%)
VGG16+LSTM+CTC	84.5
VGG16+BiLSTM+CTC	90.4
VGG16+BiLSTM+Attention	90.6
VGG16+Transformer	93.3
毛等人	93.5
刘等人	94.2
王等人	95.3
Resnet50+LSTM+CTC	85.3
Resnet50+BiLSTM+CTC	91.5
Resnet50+BiLSTM+Attention	92.1
Resnet50+Transformer	94.8
Ours	**97.3**

Table 3: 主要对比实验结果

如表3所示，本文所提的识别方法在缅甸语文本图像识别任务上准确率达到97.3%，到达了当前最高水平。相比基于“VGG16+BiLSTM+CTC”和“ResNet50+BiLSTM+CTC”的识别方法，分别提升了6.9%、5.8%，说明本文方法使用新的编解码字典，极大程度上避免了识别结果中的乱码现象；相比基于“VGG16+BiLSTM+Attention”和“ResNet50+BiLSTM+Attention”识别方法，分别提升了6.7%、5.2%，说明本文方法在识别含有大量噪声的文本图像时，也能有较好的表现；相比“VGG16+Transformer”和“ResNet50+Transformer”的识别方法，提升了4.0%、2.5%，说明本文方法使用字符粒度的特征优化了字符簇粒度的特征，提高了字符簇识别的准确率；相比于现有的缅甸语识别方法，提升了2.0%，说明本文方法在关注到上下标的同时也关注到了整个字符簇整体，减少了缅甸语文本图像识别过程中某些关键辅音字符丢失导致的乱码现象。

实验二：多粒度特征和字符簇解码字典消融实验结果对比

为了验证多粒度特征以及字符簇解码字典的有效性，我们分别在基线模型上对其做了消融实验。实验结果如表4所示（“✓”表示融合，“×”表示未融合）。

方法类别	字符粒度	字符簇粒度	SA(%)
VGG16+BiLSTM+Transformer	✓	×	94.9
	×	✓	96.5
	✓	✓	**97.3**

Table 4: 关于多粒度特征的消融实验结果

如表4所示，其中字符粒度、字符簇粒度分别表示模型使用字符粒度和字符簇的特征进行推理。从实验结果可以看出，模型使用字符簇粒度的特征可以更好的提取缅甸语中纵向堆叠字符的特征。当模型使用字符粒度和字符簇粒度的融合特征后，模型既可以关注到缅甸语文本图像的边缘特征，又可以关注到图像中的纵向堆叠字符的特征，提高了基线模型的精度，证明了所提方法的有效性。

实验三：融合多粒度的编解码端对不同识别模型识别效果的影响

为了验证多粒度特征融合策略的有效性，我们分别在多个主流识别模型上做了实验。实验结果如表5所示（“✓”表示融合，“×”表示未融合）。

方法类别	字符粒度	字符簇粒度	SA(%)
VGG16+BiLSTM+CTC	✓	×	90.4
	×	✓	91.2
	✓	✓	**92.1**
VGG16+BiLSTM+Attention	✓	×	90.6
	×	✓	91.4
	✓	✓	**92.6**

Table 5: 融合多粒度特征编解码端对其他模型的影响

如表5所示，其中字符粒度表示模型提取字符粒度的特征，字符簇粒度表示模型提取字符簇粒度的特征。从实验结果可以看出，只使用字符簇粒度特征的情况下，当前主流识别网络均有一定的提升，证明字符簇粒度特征确实使模型关注到了整个纵向堆叠的字符簇；当融合字符粒度和字符簇粒度特征后，这些模型均有明显的提升，证明了融合两种粒度的缅甸语识别方法，既可以使模型关注到字符的边缘特征，又可以关注到整个纵向字符簇的特征。

实验四：真实场景缅甸语文本图片数据测试

为了保证模型在真实场景中的应用，本文使用人工标注的1000张真实场景缅甸语文本图像作为测试集。本文在该真实场景缅甸语文本图像测试集上进行实验，实验结果如表6所示。

本文的方法在对1000张真实场景中的缅甸语文本图像的识别中仍然保持着最高的精度，相比基于CTC的识别模型有着5.6个百分点的提高，证明了本文使用的字符簇粒度的编解码字典确实可以规避大量的乱码现象；比基于注意力机制的模型有着5.8个百分点的提高，证明了本文方

第二十二届中国计算语言学大会论文集，第124页-第134页，哈尔滨，中国，2023年8月3日至5日。

方法类别	SA(%)
VGG16+LSTM+CTC	82.5
VGG16+BiLSTM+CTC	89.7
VGG16+BiLSTM+Attention	89.5
Ours	**95.3**

Table 6: 真实场景缅甸语文本图片测试结果

法可以对图像特征进行更好的建模以及在面对图像中的噪声以及文本图像畸变时拥有着更高的鲁棒性。

实验五：多种模型训练速度与推理速度对比

为了验证我们所提模型的训练速度与推理性能，本文对多个主流识别模型进行了训练速度与推理速度的对比实验。

具体方法	训练时间(s)	推理时间(s)
VGG16+LSTM+CTC	*	5.7
VGG16+BiLSTM+CTC	1250	7.3
VGG16+BiLSTM+Attention	16897	12.4
VGG16+Transformer	1590	6.6
Resnet50+LSTM+CTC	*	6.1
Resnet50+BiLSTM+CTC	1750	7.6
Resnet50+BiLSTM+Attention	23631	13.2
Resnet50+Transformer	2206	6.9
毛等人	*	*
刘等人	11560	11.2
王等人	1632	7.4
Ours	**1664**	7.8

Table 7: 模型训练速度与推理速度对比

如表7所示，我们统计了多个主流模型在缅甸语文本图像识别任务上的训练时间以及推理时间。为了验证我们的模型在训练以及推理速度上的性能，我们在相同的数据集上做了多种主流模型的性能测试，测试过程中我们取模型训练2000步的时长作为训练速度，并使用训练的模型对同样的缅甸语文本图像进行推理预测得到模型的推理速度。从实验结果可以看出，我们的模型相比于“VGG16+Transformer”和“ResNet50+Transformer”模型的训练速度和推理速度变化不大，说明本文所提方法在几乎没有增加多余的时间开销的同时提高了模型识别的精度。此外，虽然我们的模型在推理速度上不及“VGG16+BiLSTM+CTC”等模型，但是我们的模型在识别精度上相较于这些模型均有提升，依然可以说明本文所提方法的有效性。

4.3 测试样例

表8给出了缅甸语文本图像识别的实例。在针对组合字符的识别上，基于“VGG16+BiLSTM+CTC”的识别模型会存在某些关键辅音字符错识、漏识，进而导致识别结果出现乱码。而本文方法使用字符簇粒度的编解码字典，从而避免了识别结果因为错识、漏识导致的乱码。在针低质图片或有畸变存在的文本图像上的识别中，基于“CTC+BiLSTM+Attention”的识别模型由于模型鲁棒性较差，导致识别结果较差。而本文方法使用鲁棒性更好的Transformer框架对图像特征序列进行建模，从而在一定程度上解决了低质图像识别结果查的问题。在针对复杂场景中或含有大量组合字符的文本图像的识别中，基于“VGG16+BiLSTM+Transformer”的识别模型由于未能较好的关注到字符的边缘特征和纵向字符簇的整体特征导致未能准确的识别出图像中的文本。而本文方法将字符粒度特征和字符簇粒度的特征进行融合，使模型可以同时关注到文本的边缘特征和整个纵向字符簇的特征，进而提升识别模型的精度。

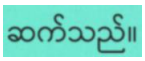
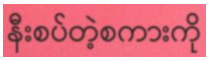
测试样例	CTC	Attention	Transformer	ours
	ကျွ်ိုးကျင်မှ	ကျွင်းကျင်ျမှု	ကျွမ်းကျင်မွ	ကျွမ်းကျင်မှ
	ဆက်သင်္။	ဆတ်သည်။	ဆက်သညီ။	ဆက်သည်။
	ိုးစပ်တဲ့စကားကို	နိုးစပ်ပီတဲ့စကားကို	နိုးစပ်တဲ့စကားကဲ့	နိုးစပ်တဲ့စကားကို

Table 8: 测试样例

5 结束语

针对现有缅甸语识别模型识别过程中容易出现局部乱码的问题，提出了一种融合多粒度特征的缅甸语文本图像识别方法，使用特征提取网络提取到的字符粒度和字符簇粒度图像特征并进行序列建模，将两种粒度特征序列进行融合后利用解码器进行解码，大幅度减少缅甸语文本图像识别中的乱码情况，提高了缅甸语识别的精度。并在自构的数据集上进行了实验，准确率达到97.35%，验证了所提方法的可行性。本文工作不仅解决了缅甸语识别中字符丢失导致的乱码现象，还为类似缅甸语的南亚东南亚小语种基于字符簇的识别方法提供了解决思路。在下一步的工作中，针对南亚东南亚小语种等具有组合字符语言的文本图像识别的研究中，我们将进一步探索将两种粒度的识别结果进行更好的融合。

参考文献

Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4715–4723.

Ravi Bhatt, Anuj Rai, Sukalpa Chanda, and Narayanan C Krishnan. 2023. Pho (sc)-ctc—a hybrid approach towards zero-shot word image recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(1):51–63.

Asghar Ali Chandio, MD Asikuzzaman, Mark R Pickering, and Mehwish Leghari. 2022. Cursive text recognition in natural scene images using deep convolutional recurrent neural network. *IEEE Access*, 10:10062–10078.

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

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2019. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8714–8721.

Fuhao Liu, Cunli Mao, Zhengtao Yu, Chengxiang Gao, Linqin Wang, and Xuyang Xie. 2021. 融合多层语义特征图的缅甸语图像文本识别方法(burmese image text recognition method fused with multi-layer semantic feature maps). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 176–185.

Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. 2019. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

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

Fengxiao Wang, Cunli Mao, Zhengtao Yu, Shengxiang Gao, Huang Yuxin, and Fuhao Liu. 2022. 融合双重注意力机制的缅甸语图像文本识别方法(burmese image text recognition method with dual attention mechanism). In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 355–365.

Zbigniew Wojna, Alexander N Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. 2017. Attention-based extraction of structured information from street view imagery. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 844–850. IEEE.

Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. 2022. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 303–321. Springer.

Wenqi Zhao and Liangcai Gao. 2022. Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 392–408. Springer.

Dajian Zhong, Shujing Lyu, Palaiahnakote Shivakumara, Bing Yin, Jiajia Wu, Umapada Pal, and Yue Lu. 2022. Sgbanet: Semantic gan and balanced attention network for arbitrarily oriented scene text recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 464–480. Springer.

林颂凯, 毛存礼, 余正涛, 郭剑毅, 王红斌, and 张家富. 2018. 基于卷积神经网络的缅甸语分词方法. 中文信息学报, 32(6):62–70.

毛存礼, 谢旭阳, 余正涛, 高盛祥, 王振晗, and 刘福浩. 2022. 基于知识蒸馏的缅甸语光学字符识别方法. 数据采集与处理, 37(1):10.

TiKEM: 基于知识增强的藏文预训练语言模型

邓俊杰[1,3] 陈龙[1,3] 张廷[1,2,3] 孙媛[1,3,4,*] 赵小兵[1,3,*]
[1]中央民族大学 信息工程学院，北京 100081
[2]中央民族大学 中国少数民族语言文学学院
[3]国家语言资源监测与研究少数民族语言中心
[4]民族语言智能分析与安全治理教育部重点实验室
*通讯作者：孙媛，赵小兵
tracy.yuan.sun@gmail.com

摘要

预训练语言模型在中英文领域有着优异的表现，而低资源语言数据获取难度大，预训练语言模型在低资源语言如藏文上的研究刚取得初步进展。现有的藏文预训练语言模型，使用大规模无结构的文本语料库进行自监督学习，缺少外部知识指导，知识记忆能力和知识推理能力受限。为了解决以上问题，本文构建含有50万个三元组知识的藏文知识增强预训练数据集，联合结构化的知识表示和无结构化的文本表示，训练基于知识增强的藏文预训练语言模型TiKEM，以提高模型的知识记忆和推理能力。最后，本文在文本分类、实体关系分类和机器阅读理解三个下游任务中验证了模型的有效性。

关键词： 藏文；知识增强；预训练语言模型；文本分类；实体关系分类；机器阅读理解

TiKEM: Knowledge Enhanced Tibetan Pre-trained Language Model

Junjie Deng [1,3] Long Chen[1,3] Ting Zhang[1,2,3] Yuan Sun [1,3,4,*] Xiaobing Zhao[1,3,*]
[1] School of information engineering, Minzu University of China, Beijing 100081
[2] School of Chinese Ethnic Minority Languages and Literature, Minzu University of China
[3] National Language Resources Monitoring and Research Center for Minority Languages
[4] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
*Corresponding author: Yuan Sun and Xiaobing Zhao
tracy.yuan.sun@gmail.com

Abstract

The pre-trained language model has excellent performance in the Chinese and English fields. But the research of pre-trained language models in low-resource languages such as Tibetan have just made initial progress. The main reason is that it's difficult to obtain low-resource language data. The existing Tibetan pre-trained language model uses a large-scale unstructured text corpus for self-supervised learning, which lacking external knowledge guidance. So their knowledge memory and knowledge reasoning abilities are limited. To solve the above problems, this paper build a Tibetan knowledge enhancement pre-trained dataset containing 500,000 triples of knowledge. Then, this paper combine structured knowledge representation and unstructured text representation to train the knowledge enhanced Tibetan pre-trained language model TiKEM. This method can improve the knowledge memory and reasoning abilities of the model. Finally, this paper verifies the effectiveness of the model in Text Classification, Tibetan relationship classification and Tibetan machine reading comprehension.

Keywords: Tibetan , Knowledge Enhancement , Pre-trained Language Model , Text Classification , Entity Relationship Classification , Machine Reading Comprehension

第二十二届中国计算语言学大会论文集，第135页-第144页，哈尔滨，中国，2023年8月3日至5日。

1 引言

预训练语言模型可以从大规模无标签的数据中学习丰富的上下文表征，通过迁移学习在多个下游任务上取得了优秀的性能，这对于低资源语言上的自然语言处理研究存在重要意义。目前，预训练语言模型在英文等语言上取得了很好的发展，一系列预训练语言模型相继出现，如GPT(Radford et al., 2018)、BERT(Kenton and Toutanova, 2019)和RoBERTa(Liu et al., 2019)等。为了进一步优化预训练语言模型，研究人员通过增加模型参数量，提高预训练语言模型在下游任务中的性能，如GPT-3(Brown et al., 2020)、T5(Raffel et al., 2020)、盘古α(Zeng et al., 2021)等。模型参数量的增加，虽然显著提升了模型性能，但是在知识获取能力方面依然存在不足。对于该问题，研究人员将知识图谱中的事实知识融合到预训练语言模型中，如ERNIE(Zhang et al., 2019)、KEPLER(Wang et al., 2021)、ERNIE3.0(Sun et al., 2021a)等，显著提高了模型的认知能力。

以上研究均是在英文等资源比较丰富的语言上的研究，为了解决在低资源语言上数据稀疏等问题，人们提出多语言预训练模型mBERT(Pires et al., 2019)、XLM-R(Conneau et al., 2020)等，但上述多语言预训练模型在藏文上的效果并不理想，例如mBERT在藏文分类数据集TNCC(Qun et al., 2017)上的F1为5.5%，XLM-R-base的F1为21.1%(Yang et al., 2022)。为此，Liu等人(Liu et al., 2022)提出藏文预训练语言模型TiBERT，Yang等人(Yang et al., 2022)提出了少数民族多语言预训练模型CINO，Deng等人(Deng et al., 2023)提出了少数民族多语言预训练语言模型MiLMo。以上模型推动了藏文自然语言处理的研究，但目前的藏文预训练语言模型都是使用大规模无标注数据进行自监督学习，缺少外部知识指导，知识记忆能力和知识推理能力存在不足。

针对上述存在的不足，本文提出基于知识增强的藏文预训练语言模型TiKEM，联合结构化知识和无结构化文本，并在下游任务中评估模型性能，主要贡献如下：

（1）为了显示表示知识，本文构建了一个含有50万个三元组的藏文知识库，并将其与藏文语料库结合，构建藏文知识增强预训练数据集；

（2）为了提高藏文预训练语言模型的知识记忆和知识推理能力，本文提出基于知识增强的藏文预训练语言模型TiKEM，统一建模结构化知识表示和无结构文本表示，对知识进行掩码。同时，为了增强模型的句子表达能力，本文将下一个句子预测任务扩展为句子重排序任务和句子间的距离关系任务；

（3）为了评估模型的性能，本文在文本分类、实体关系分类、阅读理解三个下游任务中进行了对比实验，实验结果表明本文提出的藏文预训练语言模型的性能有着显著的提升。

2 相关工作

预训练语言模型已经在自然语言处理的多项下游任务中取得了优秀的性能，包括OpenAI GPT(Radford et al., 2018)、BERT(Kenton and Toutanova, 2019)、XLNet(Yang et al., 2019)等，可以有效获取句法和语义信息来进行文本表示。尽管预训练语言模型的上下文表示已经包含了句法、语义等知识，但挖掘上下文表示所蕴含的知识的研究较少，它对于文本理解非常重要。Zhou等人(Zhou et al., 2020)在不同具有挑战性的测试中检验GPT、BERT、XLNet和RoBERTa的常识获取能力，发现模型在需要更多深入推理的任务上表现不佳，这也表明常识获取依然是一个巨大挑战。

知识图谱存储着丰富的知识，利用知识图谱让模型显式学习人类对世界的认知，是融合知识的预训练模型采用的重要方法(王海峰et al., 2022)。该方法可以提高预训练模型的知识获取和知识推理等能力。ERNIE(Zhang et al., 2019)首先对文本中提到的命名实体进行识别提取，然后将实体与知识图中对应的实体对齐，利用文本语义作为知识图的实体嵌入，再使用TransE方法学习图的结构。然后利用掩码机制，将知识图中的实体遮蔽，使模型聚合上下文和知识图共同预测遮掩的令牌和实体，使得预训练模型不仅可以将图谱三元组中的事实知识更好地融合到模型中，而且还可以通过丰富的实体描述，有效地学习实体和关系的知识表示。其提供了整合异构数据的一种示范方法。Sun等人(Sun et al., 2020)认为，通过多种不同的知识表示学习获得的实体嵌入，并在预训练阶段进行融合的方法，不能够充分学习到相应知识，并且当知识图谱发生变化时需要重新训练实体嵌入表示模型。因此在CoLAKE模型中提出词-知识图的

概念，将文本序列看作是全连接的词图，以构成一个同时包含词语、实体和关系的词语-知识图。CoLAKE利用遮蔽注意力来控制信息流，将掩码策略分为词节点掩码、实体节点掩码、关系节点掩码，从而能够同时融合训练语料中的语言知识和图谱中的知识。然而，CoLAKE更加关注于实体在知识图谱中的建模，却忽略了实体在训练语料中的表述，在一定程度上削弱了语言模型的泛化能力。为此，ERNIE3.0(Sun et al., 2021a)提出知识图谱与文本平行预训练的方法，使用文本表述知识。其将知识图谱中的三元组与对齐文本统一编码，作为预训练语言模型的输入，同时利用掩码策略，掩盖三元组中的关系和文本中的实体，促使模型融合三元组知识和文本信息。

在藏语方面，哈工大讯飞联合实验室提出了包含藏语、蒙语（回鹘体）、维吾尔语、哈萨克语（阿拉伯体）、朝鲜语、壮语、粤语七种少数民族语言的多语言预训练模型CINO(Yang et al., 2022)，该模型基于多语言预训练模型XLM-R(Conneau et al., 2020)开发，在多种少数民族语料上进行了二次预训练，该模型在藏文分类数据集TNCC(Qun et al., 2017)上相比其它基线模型获得了显著的性能提升。Liu等人(Liu et al., 2022)提出了藏文预训练语言模型TiBERT，其构建了覆盖语料库99.95%的词汇表。该模型在文本分类和问题生成任务中取得较好效果。Deng等人(Deng et al., 2023)提出了包含蒙古语、藏语、维吾尔语、哈萨克语和韩语五种少数民族语言的多语言预训练模型MiLMo，该模型在构建的多语言文本分类数据集上证明了模型的有效性，推动了少数民族语言信息化的建设。安波等人(安波and 龙从军, 2022)提出了藏文预训练语言模型BERT-base-Tibetan，并将该模型应用于藏文文本分类，实验表明预训练语言模型能显著提升藏文文本分类的性能。目前的藏文预训练语言模型取得了不错的进展，但大都是使用大规模无标注数据进行自监督学习，缺少外部知识指导，知识记忆能力和知识推理能力存在不足。因此如何使用知识库来增强藏文预训练模型的表示能力是藏文预训练模型研究和应用的难点之一。

3 TiKEM模型

本文构建的藏文知识增强预训练语言模型总体结构如图1所示。首先将知识库中的三元组与语料库中的文本进行拼接作为训练数据。本文使用[SEP]分隔三元组与文本并将其作为预训练语言模型的输入，不使用额外的模型对知识库中的三元组进行知识表示，从而统一建模结构化知识表示和无结构化文本表示。然后分别根据实体掩码预测任务、知识掩码预测任务、句子重排序任务和句子间的距离关系任务的要求，对输入文本进行处理。如知识掩码预测任务中，随机掩码三元组中的关系或文本中的实体。TiKEM与CharBERT(Ma et al., 2020)结构类似，融合子词与字符表示，使用多层双向Transformer作为通用知识文本表示。在此基础上，使用多任务学习框架，分别对四个任务进行建模，最后以四个任务的加权损失函数值作为模型总体损失值。

3.1 预训练任务

基于掩码语言模型和下一个句子预测任务，本文使用实体掩码预测、知识掩码预测、句子重排序和句子间的距离关系共四个任务作为预训练任务，具体如下。

3.1.1 实体掩码预测

BERT的掩码预测任务是预训练语言模型最重要的预训练任务，它可以帮助模型更好地理解语言中的上下文和语义信息。在BERT的掩码预测任务中，模型需要预测输入文本中被随机遮掩的子词，但是随机掩码子词会影响全词表示，不利于模型对实体、单词、短语的理解。因此本文随机选取输入文本中15%的令牌进行掩码，其中实体占80%，即在文本中代表具体事物、人物、地点等的词语，而剩余20%的令牌则是非实体的单词或短语。在这个过程中，为了提高模型鲁棒性，80%的令牌使用特殊标记[MASK]进行替换，10%的令牌随机替换为其它令牌，剩余10%的令牌不进行处理。掩码语言模型的损失值$loss_{mlm}$即为其损失值，如公式（1）所示。

$$loss_{mlm} = -\sum_{i=1}^{n} p_m \log p'_m \tag{1}$$

其中，p'_m为预测的token。p_m为真实token。

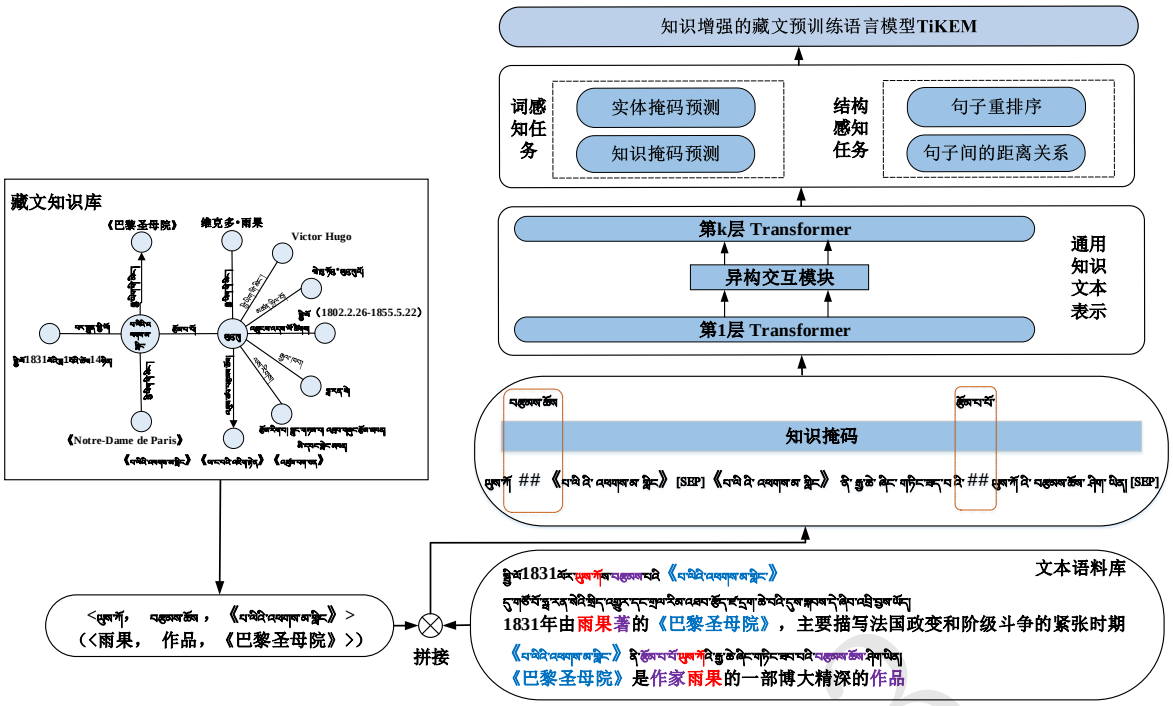


图 1: 基于知识增强的藏文预训练语言模型TiKEM

3.1.2 知识掩码预测

知识掩码预测任务是指将知识库中的三元组与给定文本作为预训练模型的序列输入，以特殊标记[SEP]隔开，并随机掩码三元组中的关系或文本中对应的实体，让模型结合三元组知识与文本知识，预测三元组的关系或文本中的实体,与实体掩码预测相同，掩码语言模型的损失值$loss_{mlm}$同样为其损失值。

在传统的语言模型中，模型需要学习词汇之间的关系和语言结构，而在知识掩码预测任务中，模型需要学习知识库中的三元组关系，并将其与文本融合，从而增强模型对知识的理解和应用能力。因此本文将藏文知识库中的三元组与藏文语料库中的文本使用“[SEP]”进行拼接，作为序列输入。其中三元组以主体-关系-客体的形式呈现，如图1所示，（ཡུས་ཀོ （雨果），བརྩམས་ཚོས （创作），པ་ལིའི་འཐབས་མ་སྒྲིང （巴黎圣母院））。本文在序列中随机掩码三元组中的关系或文本中对应的实体，例如在图1中，三元组中的“ བརྩམས་ཚོས （创作）”以及语料中的“ རྩོམ་པ་པོ་ （作家）”被遮掩，这促使模型结合三元组知识与文本蕴含知识，预测三元组的关系或文本中的实体，从而学习三元组中主客体之间的关系和知识。其本质类似于实体关系抽取中的远程监督算法(Mintz et al., 2009)。远程监督算法假设如果两个实体参与了一个关系，那么任何包含这两个实体的句子都可以表达这种关系。

与传统语言模型中的掩码预测任务相比，知识掩码预测任务使模型能够显示学习人类对世界的认知，从而更好地应对自然语言处理中的任务和问题。

3.1.3 句子重排序

句子重排序任务是将给定的输入序列分成多段文本，并随机打乱，让模型对文本重新排序。其目的是让模型学习文本中句子间的结构关系，以便更好的理解和生成连贯的自然语言文本。例如，在机器翻译、摘要生成等生成式任务中，模型需要生成具有流畅语境的自然语言文本，因此对句子的正确排序非常重要。

本文将训练样本中的每个文本语料分成多段文本，然后随机抽取50%的训练样本，将其中的多段文本随机打乱，让模型预测其顺序。模型首先需要理解每段文本中所包含的信息，包括每段文本的主题、语义和上下文信息。然后根据语境的先后顺序以及句子间的关联关系，给出

多段文本的顺序，将多段文本重新组合成具有流畅语境的自然文段。其损失值计算如公式（2）所示。

$$loss_{sort} = -\sum_{i=1}^{m} p_s \log p'_s \tag{2}$$

其中，p'_s为段落的顺序预测。p_s为真实答案。

3.1.4 句子间的距离关系

句子间的距离关系预测任务是预训练语言模型中的下一个句子预测任务的扩展，其目的是让模型能够学习文档级别的信息。文档级别的信息往往需要考虑到句子之间的关系，例如同一篇文章中的句子往往具有相关性，句子间的距离关系也能够体现文章的结构特点。因此该任务可以为其他自然语言处理任务提供有益的信息，例如文本分类、信息检索等任务。

该任务要求模型从一个文本序列中预测出句子间的距离关系，包括同一篇文章中相邻的句子、同一篇文章中不相邻的句子和不同文章中的句子。为了让模型学习到这些距离关系，本文采用了一种比较直观的方法：将给定的训练文本分成多段，随机选取一段文本，以25%的概率替换为同文档的其它句子，25%的概率替换为不同文档的句子，剩余50%的概率不替换，这种方法可以让模型在学习时考虑到不同文本之间的语境差异。

为了更好地训练模型，本文将句子间的距离关系任务视为三分类任务，其损失值计算如公式（3）所示。

$$loss_{relation} = -\sum_{i=1}^{3} p_r \log p'_r \tag{3}$$

其中，p'_r为句子间的距离关系预测。p_r为句子间的真实关系。

3.1.5 模型总体损失值

根据多任务学习框架，本文将上述各任务损失值的加权和作为模型总体损失值，如公式（4）所示。

$$loss_{all} = loss_{mlm} + \alpha loss_{sort} + \beta loss_{relation} \tag{4}$$

其中α和β是可调整的权重参数。

3.2 数据集

由于目前没有公开的大规模藏文知识语料库，为了增强模型的知识表示能力，本文构建了一个知识增强预训练数据集。本文通过爬取21个藏文网站如云藏网、西藏新闻网、西藏人民网等，收集大量的结构化知识和无结构化文本。然而，由于网络上的语料不够规范和完整，会存在大量的错误和噪声数据。因此，为了提高数据集的质量，在构建数据集的过程中，本文进行了以下数据清洗和预处理操作:

（1）本文将数据中的图片、链接、特殊字符等无意义的内容剔除。同时，过短的文本包含的文本信息不足，因此本文将词数低于100的文本去除，只保留了词数超过100的文本数据。

（2）爬取到的表格数据中偶尔会有不完整的三元组，如缺失实体或者关系，因此本文将信息不完整的三元组剔除。并且由于不同的文本语料中可能会包含相同的三元组，因此本文需要对三元组进行去重处理，确保知识库中每个三元组只出现一次。

通过以上的数据清洗和预处理，本文构建了一个包含50万个三元组、大小为4GB，藏文知识增强预训练数据集。该数据集共2.45亿个token。其中50万个三元组构成藏文知识库，无结构化文本作为藏文语料库。该数据集包含多个领域知识，如：经济、社会、科技、法律、体育等。最终用于增强预训练模型的知识表示能力。

第二十二届中国计算语言学大会论文集，第135页-第144页，哈尔滨，中国，2023年8月3日至5日。

3.3 词表构建

藏文是一种拼音文字，其单词的最小单位是一个音节，包含一个或最多七个字符，音节间以“.”来分割，但基于音节的分词并不能很好的表达语义结构。藏语的文字由一个或多个音节组成，同样以“.”分割。藏文包含七种结构：基字、上加字、下加字、前加字、后加字、后后加字和元音符号，共有155个字符。在预训练模型中，通常使用子词切分文本，而子单词表示可能不包含细粒度字符信息和全词的表示，本文同时编码藏文词表示和对应的藏文字符序列表示，使模型能够捕获不同粒度之间的语言知识，提升模型语言表达能力。

（1）字符表构建

除藏文的155个字符外，本文对训练语料中包含的其它字符进行了统计，频次在400以上的字符有941个，取其中前845个字符与藏文字符构成大小为1000的字符表，以做字符嵌入。

（2）子词表构建

如果以藏语文字进行分词，为了减少未被录入词表的单词数即未登录词（out-of-vocabulary，OOV），我们需要构建一个非常大的词表，这加大了机器的运算量，并且需要花费大量的时间和计算资源。针对OOV问题，本文使用sentencepiece(Kudo and Richardson, 2018)训练一个藏文分词模型，构建了一个大小为30,005，覆盖语料库99.99%字符的词表，并使用该分词模型对训练数据进行分词。

4 实验评估

本文使用文本分类、实体关系分类、机器阅读理解三个下游任务评估TiKEM模型性能。

4.1 藏文文本分类

本文藏语新闻数据集TNCC(Qun et al., 2017)，评估TiKEM模型对文本的分类能力。该数据集包含9,203条样本，涉及政治、经济、教育、旅游、环境、艺术、文学、宗教等12个领域。因原始数据集没有切分，本文按8:1:1的比例将其划分为训练集、验证集、测试集，评价指标为Accuracy和Macro-F1。

本文将知识增强预训练模型TiKEM与基于藏文音节分词的CNN分类模型、Transformer(Vaswani et al., 2017)、TextCNN(Guo et al., 2019)、DPCNN(Johnson and Zhang, 2017)等经典分类模型进行比较，同时也与少数民族多语言预训练模型CINO-base和藏文预训练模型TiBERT、BERT-base-Tibetan进行比较。实验结果如表1所示。

模型	Accuracy(%)	Macro-F1(%)
Transformer(Vaswani et al., 2017)	28.63	28.79
CNN(syllable)	61.51	57.34
TextCNN(Guo et al., 2019)	61.71	61.53
DPCNN(Johnson and Zhang, 2017)	62.91	61.17
TextRCNN(Lai et al., 2015)	63.67	62.81
BERT-base-Tibetan(安波and 龙从军, 2022)	-	51
TiBERT(Liu et al., 2022)	71.04	70.94
CINO-base(Yang et al., 2022)	73.1	70.0
TiKEM	**74.46**	**72.61**

表 1: 藏文文本分类结果

由表1可以看到，Transformer在藏文文本分类上的表现不如经典分类模型以及预训练语言模型，而TiKEM模型在藏文文本分类上的准确率超过了经典分类模型如TextCNN、DPCNN等，比TextCNN高了12.75%，DPCNN与TiKEM模型相差11.55%。同时TiKEM模型比TiBERT高了3.42%，并略微高于CINO-base模型。在Macro-F1值方面，TiKEM模型的表现同样超过了众多基线模型。与经典分类模型比较，CNN(syllable)与TiKEM模型之间相差15.27%，而TiKEM模型采用的同样是transformer结构，但是Macro-F1值却远远高于Transformer，这表明使用大规模藏文语料训练模型的方法提高了模型对藏文的理解能力。在预训练模型方面，TiKEM模型比BERT-base-Tibetan高

了21.61%，比CINO-base高了2.61%，比TiBERT高了1.67%，相对于只使用了大规模无结构化藏文语料训练的藏文预训练语言模型，融合藏文知识库的TiKEM模型在文本分类上有着明显的性能提升。

4.2 实体关系分类

为了验证TiKEM模型对知识的记忆及融合运用能力，本文构建了6,433条三元组-文本对齐数据集，三元组中共有11种关系。该任务要求在给定两个实体和包含该实体的对应文本后，给出两个实体之间的关系类别。本文按8:1:1的比例将其划分为训练集、验证集、测试集。本文使用FastText(Joulin et al., 2016)、DPCNN等作为基线模型，并与藏文预训练语言模型TiBERT和多语言预训练模型MiLMo、CINO-base进行比较。评价指标为Accuracy(%)、Macro-P(%)、Macro-R(%)和Macro-F1(%)，实验结果如表2所示。

模型	Accuracy(%)	Macro-P(%)	Macro-R(%)	Macro-F1(%)
FastText(Joulin et al., 2016)	55.80	34.05	32.98	31.61
DPCNN	70.94	54.21	49.23	48.65
TextCNN	72.38	71.03	59.11	56.76
TiBERT(Liu et al., 2022)	84.70	76.66	68.82	67.94
CINO-base(Yang et al., 2022)	85.31	75.48	69.12	66.73
MiLMo(Deng et al., 2023)	85.76	77.13	68.97	68.57
TiKEM	**90.12**	**91.73**	**75.61**	**76.34**

表 2: 藏文实体关系分类结果

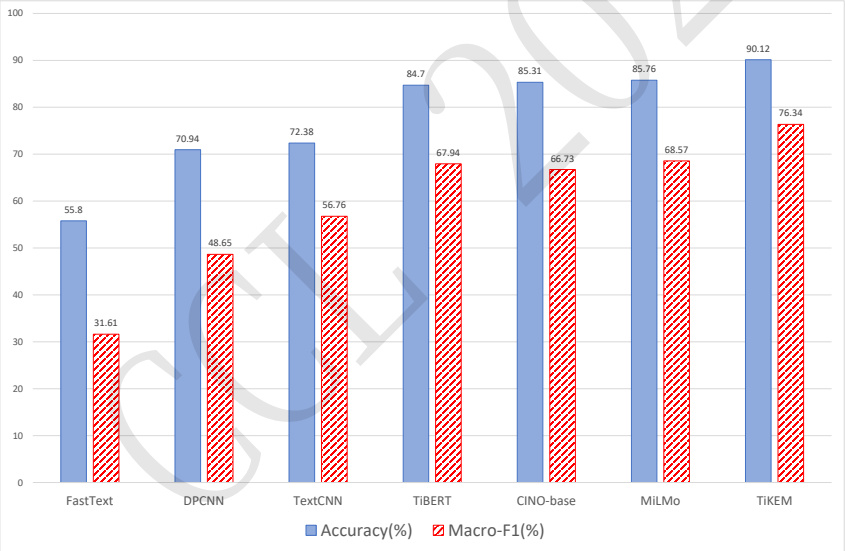


图 2: 模型在藏文实体关系分类中的Accuracy与Macro-F1值对比

由表2可以看到，FastText相对其它模型表现较差，TiKEM模型的准确率比FastText高了34.32%，比TextCNN高了17.74%。从总体上看，预训练模型在实体关系分类中比基线模型表现更好。同时融合了知识的藏文预训练模型TiKEM在该任务中，比TiBERT的准确率高了5.42%，比多语言预训练模型CINO-base和MiLMo的准确率分别高了4.81%和4.36%。

为了更清晰地观察各模型性能之间的差异，本文将各模型的准确率和Macro-F1值进行对比，绘制成柱状图，如图2所示。在以Macro-F1作为评价指标中，各模型的性能差异总体与准确率作为评价指标时的趋势一致。不同的是，在准确率中TiBERT比CINO-base低了0.61%。而在Macro-F1中，TiBERT比CINO-base高了1.21%。Macro-F1是各类别F1值的平均值，一定程度上反应了模型在不同类别中实体关系分类性能的偏差。因此可以看出CINO-base在一些类别中，实体关系分类性能比TiBERT更好，而总体上TiBERT的实体关系分类性能比CINO-base更稳定。此外，我们也可以看到，TiKEM模型在实体关系分类中的Macro-F1值高于其余模型。

其中，比预训练模型CINO-base、TiBERT和MiLMo分别高了9.61%、8.4%和7.77%。这表明，融合了知识的藏文预训练模型TiKEM有着更加丰富的知识，并且对给定的知识更加擅于去融合及运用。

4.3 藏文机器阅读理解

机器阅读理解任务是给定一段文本和一个问题，让模型回答对应问题。这需要模型理解问题和上下文语义，然后进行推理、判断等，给出具体答案。本文使用藏文机器阅读理解数据集TibetanQA(Sun et al., 2021c)对模型的阅读理解能力进行评估，该数据集包含了1,513篇文章和20,000个问答对。为了评估模型性能，本文使用EM值（精确匹配）和F1值作为评价指标。

本文以8:2的比例将数据划分为训练集和测试集，并使用机器阅读理解的经典模型R-Net(Wang et al., 2017)、BiDAF(Seo et al., 2017)、QANet(Yu et al., 2018)作为基线模型，这些模型在英文数据集上有着出色的表现，同时本文还将TiKEM与藏文预训练语言模型TiBERT和藏文抽取式机器阅读理解模型Ti-Reader(Sun et al., 2021b)进行比较。此外为了验证模型的知识推理能力，本文在TibetanQA数据集基础上，增加了1,823条包含三元组的藏文问答数据样本，并将数据同样以8:2的比例划分为训练集和测试集。实验结果如表3所示。

模型	TibetanQA		TibetanQA（含三元组）	
	EM(%)	F1(%)	EM(%)	F1(%)
R-Net(Wang et al., 2017)	55.8	63.4	-	-
BiDAF(Seo et al., 2017)	58.6	67.8	-	-
QANet(Yu et al., 2018)	57.1	66.9	-	-
TiBERT(Liu et al., 2022)	53.2	73.4	54.1	73.9
Ti-Reader(Sun et al., 2021b)	67.9	77.4	-	-
TiKEM	**69.4**	**80.1**	**72.6**	**81.3**

表 3: 预训练模型在藏文阅读理解上的应用

由表3可以看到，TiBERT在TibetanQA上F1值超过了R-Net等基线模型，但是EM值却低于基线模型。EM衡量模型预测与标准答案完全一致的占比，F1值评估模型预测与标准答案的重叠程度。F1值高而EM值低，说明TiBERT可以很好地确定答案范围，但是对于答案边界上的判断能力明显不足。而融合了知识之后的藏文知识增强预训练模型TiKEM在机器阅读理解上的性能相较于TiBERT有了极大的提升，并且EM值的提升幅度大于F1值的提升幅度。一方面实体掩码预测提高了模型对事物、人物、地点等实体的边界预测准确性，同时知识的融入提高了模型推理能力。另一方面句子关系预测等任务提高了模型对上下文结构和语境的理解能力。基于此，我们也可以看到TiKEM模型超过了藏文抽取式机器阅读理解模型Ti-Reader在TibetanQA上的表现。

在加入包含三元组的藏文问答数据样本后，TiBERT和TiKEM模型性能有了明显提升，TiBERT的EM值和F1值分别提升了0.9%和0.5%，TiKEM模型的EM值和F1值分别提升了3.2%和1.2%。显然，TiKEM模型提升幅度相较于TiBERT的提升幅度更大，这表明TiKEM模型比TiBERT更加擅长理解和运用知识，并进行知识推理。

5 总结

本文构建了一个包含50万个三元组的知识增强预训练数据集，在此基础上训练了一个基于知识增强的藏文预训练语言模型TiKEM，将结构化的藏文知识库和无结构化的文本统一表征。同时，针对知识的融合，将掩码预测任务扩展为实体掩码预测任务和知识掩码预测任务。为了学习句子间关系和文档级信息，本文将下一个句子预测任务扩展为句子重排序任务和句子间的距离关系任务。最后本文在在文本分类、关系分类、机器阅读理解三个下游任务上进行了实验。TiKEM模型性能均超过对比模型，证明了TiKEM模型在知识记忆、运用和推理能力等方面的有效提升。

第二十二届中国计算语言学大会论文集，第135页-第144页，哈尔滨，中国，2023年8月3日至5日。

致谢

本论文得到了国家自然科学基金项目（61972436）和国家社会科学基金项目（22&ZD035）的资助。

参考文献

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. Milmo:minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

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

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Charbert: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Nuo Qun, Xing Li, Xipeng Qiu, and Xuanjing Huang. 2017. End-to-end neural text classification for tibetan. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 5*, pages 472–480. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuan-Jing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021a. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Yuan Sun, Chaofan Chen, Sisi Liu, and Xiaobing Zhao. 2021b. Ti-reader: 基于注意力机制的藏文机器阅读理解端到端网络模型(ti-reader: An end-to-end network model based on attention mechanisms for tibetan machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 219–228.

Yuan Sun, Sisi Liu, Chaofan Chen, Zhengcuo Dan, and Xiaobing Zhao. 2021c. 面向机器阅读理解的高质量藏语数据集构建(construction of high-quality tibetan dataset for machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

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

W Wang, N Yang, F Wei, B Chang, and M Zhou. 2017. R-net: Machine reading comprehension with self-matching networks. *Microsoft Research Asia, Beijing, China, Tech. Rep*, 5.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9733–9740.

安波and 龙从军. 2022. 基于预训练语言模型的藏文文本分类. 中文信息学报, 36(12):85–93.

王海峰, 孙宇, and 吴华. 2022. 知识增强预训练模型. 中兴通讯技术, (16-24).

TiKG-30K：基于表示学习的藏语知识图谱数据集

庄文浩[1,3,&] 高歌[2,3,&] 孙媛[1,3,4,*]
[1]中央民族大学 信息工程学院，北京 100081
[2]中央民族大学 中国少数民族语言文学学院，北京 100081
[3]国家语言资源监测与研究少数民族语言中心
[4]民族语言智能分析与安全治理教育部重点实验室
[&]共同第一作者：庄文浩，高歌
*通讯作者：孙媛
tracy.yuan.sun@gmail.com

摘要

知识图谱的表示学习旨在通过将实体和关系映射到低维向量空间中来学习知识图谱数据之间的复杂语义关联，为信息检索、智能问答、知识推理等研究提供了支撑。目前知识图谱的表示学习研究主要集中在英、汉等语言，公开高质量数据集（如FB15k-237，WN18RR）对其研究起到非常重要的作用。但是，对于低资源语言（如藏语），由于缺少公开的知识图谱数据集，相关研究任务还处于起步阶段。基于此，本文提出一个公开的藏语知识图谱数据集TiKG-30K，包含了146,679个三元组，30,986个实体和641种关系，可应用于知识图谱的表示学习及下游任务。针对现有藏语知识图谱数据量少、数据稀疏的问题，本文利用藏文三元组中实体的同指关系，借助其他语言丰富的知识库和非文本介质对知识库进行扩充，通过跨语言近义词检索、合并同义实体和关系、修正错误三元组等技术对知识图谱进行多层优化，最终构建了藏语知识图谱数据集TiKG-30K。最后，本文采用多种经典表示学习模型在TiKG-30K进行了实验，并与英文数据集FB15k-237，WN18RR以及藏文数据集TD50K进行了对比，结果表明，TiKG-30K可以与FB15k-237、WN18RR数据集相媲美。本文将TiKG-30K数据集公开，https://tikg-30k.cmli-nlp.com/。

关键词： 藏语知识图谱；表示学习；知识图谱嵌入；链接预测

TiKG-30K: A Tibetan Knowledge Graph Dataset Based on Representation Learning

Wenhao Zhuang[1,3,&] Ge Gao[2,3,&] Yuan Sun[1,3,4,*]
[1] School of Information Engineering, Minzu University of China, Beijing 100081
[2] School of Chinese Ethnic Minority Languages and Literature, Minzu University of China
[3] National Language Resources Monitoring and Research Center for Minority Languages
[4] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
[&]Co-first authors: Wenhao Zhuang and Ge Gao
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com

Abstract

Representation learning of knowledge graphs aims to learn the complex semantic relationships between entities and relations in knowledge graph data by mapping them into a low-dimensional vector space, providing support for information retrieval, intelligent question answering, knowledge reasoning, and other research areas. Currently, research on representation learning of knowledge graphs mainly focuses on languages such as English and Chinese, and high-quality public datasets (such as FB15k-237, WN18RR) have played an important role in their research. However, for low-resource languages such as Tibetan, relevant research is still in the initial stages due to the lack of public knowledge graph datasets. In this paper, we propose a publicly available Tibetan knowledge graph dataset TiKG-30K, which contains 146,679 triples, 30,986

第二十二届中国计算语言学大会论文集，第145页-第154页，哈尔滨，中国，2023年8月3日至5日。

entities, and 641 relations, and can be applied to representation learning of knowledge graphs and downstream tasks. To address the problem of the small and sparse Tibetan knowledge graph dataset, we use the same-as relations between entities in Tibetan triples, and leverage other language-rich knowledge bases and non-text media to expand the knowledge base. We optimize the knowledge graph through multiple layers of techniques such as cross-lingual synonym retrieval, merging synonymous entities and relations, and correcting incorrect triples. Finally, we conduct experiments on TiKG-30K using classical representation learning models, and compare it with English datasets FB15k-237, WN18RR, and Tibetan dataset TD50K. The results show that TiKG-30K is comparable to FB15k-237 and WN18RR datasets. We make the TiKG-30K dataset public at https://tikg-30k.cmli-nlp.com/.

Keywords: Tibetan knowledge graph , Representation learning , Knowledge graph embedding , Link prediction

1 引言

藏语是中国少数民族语言之一，具有丰富的文化历史和独特的语言结构特点。近年来，随着人工智能领域的快速发展，基于藏语的知识图谱构建和知识表示学习成为研究热点之一。知识图谱是一种用于描述实体之间关系的结构化数据，它可以帮助我们更好地理解和利用丰富的实体知识。知识表示学习旨在将自然语言等符号化的知识转化为计算机可处理的向量表示(刘知远et al., 2016)，以便于机器学习算法的应用，表示学习的研究需要知识图谱作为数据支撑。

现有的知识图谱数据集大多针对中英文设计，如英文大规模通用知识图谱Freebase(Bollacker et al., 2008)，包含了超过8,000万个实体，三元组数量达到12亿条，从中提取知识事实构建的英文知识图谱FB15k(Bordes et al., 2013)，它包含14,951个实体、1,345个关系以及592,213个三元组。FB15k-237是FB15k的子集，在FB15k的测试集中，很多三元组可以通过训练集中简单的反转关系来获得，因此专家学者对出现的反向关系进行了去除，构建了更为有效的FB15k-237。中文大规模通用知识图谱CN-DBpedia(Xu et al., 2017)由复旦大学构建，涵盖超过2,200万的实体和2亿条三元组。相比之下，藏语等低资源语言的知识图谱比较匮乏，目前已有的藏语知识图谱如TD50K(Sun et al., 2021)，三元组数量为53,797，关系数量为3,285，三元组数据量较少，数据稀疏。针对现有藏语知识图谱数据量少、数据稀疏的问题，本文构建了一个藏语知识图谱数据集TiKG-30K，该数据集包含了30,986个实体以及641种关系类型，三元组数量为146,679，可以用于藏语表示学习的研究和链接预测、关系预测等相关任务。本文的主要贡献如下：

（1）针对现有藏语知识图谱数据量少、数据稀疏的问题，本文利用藏文三元组中实体的同指关系，借助其他语言丰富的知识库和非文本介质对知识库进行了扩充。

（2）在扩充三元组时，中英文专业词汇有时难以找到对应的藏语专业术语，导致产生歧义或者混淆语义。例如，中文学名“紫苏梗”、“紫草”、“紫花地丁”、“紫花针茅”对应着不同植物，但对应的藏语是相同的“ཟི་མིག་” （紫胶）。因此，本文采用三元组修正技术，合并同义实体和关系、删除不必要的实体和关系、修正错误的三元组等方式，进行了四个版本的优化更新，进一步构建了一个关系稠密、规模适中且适合用于表示学习任务的藏语知识图谱数据集TiKG-30K。

（3）采用TransE(Bordes et al., 2013)、DistMult(Yang et al., 2014)、ComplEx(Toutanova and Chen, 2015)、RotatE(Sun et al., 2019)、pRotatE(Sun et al., 2019)、HAKE(Zhang et al., 2020)多种经典表示学习模型在TiKG-30K进行了实验，并与英文数据集FB15k-237，WN18RR以及藏文数据集TD50K进行了对比，为藏文知识图谱表示学习提供了可开放测试的基线数据。

2 相关工作

知识图谱是谷歌在2012年提出的概念，其本质上是一种结构化的知识库，通过三元组（头实体，关系，尾实体）的形式来表示单条知识，以实体为节点，关系为边，大量的三元组可构建图谱结构，用来发掘不同实体间更为复杂的关系，现有的国外大规模通用知识图谱有DBpedia(Auer et al., 2007)、Yago(Suchanek et al., 2007)、Freebase(Bollacker et al., 2008)、Wikidata(Vrandečić and Krötzsch, 2014)等。国内大型知识图谱项目有CN-DBpedia(Xu et al., 2017)、XLORE(Wang et al., 2013)、openKG等，涵盖百科及细分领域的大量知识图谱。对于藏语知识图谱的构建，由于缺乏大规模的公开知识库，且三元组抽取技术的限制，现有的藏语知识图谱集中在特定领域，如汉藏双语旅游领域知识图谱(冯小兰and 赵小兵, 2019)，藏语数据集TD50K(Sun et al., 2021)等。此外，藏文知识图谱的数量与质量也远不能媲美中英文知识图谱。比如，在FB15k-237(Toutanova and Chen, 2015)英文数据集中，97.8%的实体出现两次及以上，每个实体平均拥有20个三元组，而在藏语数据集TD50K(Sun et al., 2021)中，只有48%的实体出现两次及以上，且一个实体平均仅有2个三元组，同时三元组数量也只有FB15k-237的17.3%。

知识图谱表示学习模型是一类广泛应用于知识图谱数据的机器学习模型。这些模型旨在通过将实体和关系映射到低维向量空间中来学习知识图谱数据之间的复杂语义关联，从而为许多基于知识图谱的任务提供更好的性能。2013年Bordes等人提出基于向量空间的表示学习模型TransE，它将实体和关系都映射到向量空间中，通过计算三元组中头实体与关系向量之和与尾实体向量的距离，来判断三元组是否成立(Bordes et al., 2013)。在TransE模型被提出后，TransH(Wang et al., 2014)、TransR(Lin et al., 2015)、TransD(Ji et al., 2015)等一系列基于TransE的改进模型被提出。2015年Yang等人提出的DistMult是一种基于张量分解的表示学习模型，它将实体和关系都表示为向量，并使用张量乘积来计算三元组的分数(Yang et al., 2014)。Toutanova等人提出一种基于复数向量的表示学习模型ComplEx，能够捕捉实体和关系之间的更复杂的交互(Toutanova and Chen, 2015)。RotatE是Sun等人在2019年提出的一种基于旋转操作的表示学习模型，它将关系表示为一个旋转矩阵，并通过对头实体和关系向量进行旋转，来预测尾实体，与前面几种模型相比，RotatE能够更好地处理对称性和反对称性关系(Sun et al., 2019)。2020年，在RotatE的工作基础上，Zhang等人对语义层次结构进行建模，将实体映射到极坐标系中，同心圆可以自然地反映等级，据此提出的HAKE模型在多个基准数据集上进行链接预测时取得了较好的结果(Zhang et al., 2020)，本文所用模型的简要信息如表1所示。

模型	得分函数	参数
TransE(Bordes et al., 2013)	$-\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_{1/2}$	h, r, t $\in R^k$
DistMult(Yang et al., 2014)	$\mathrm{h}^\top diag(\mathbf{r})\mathbf{t}$	h, r, t $\in R^k$
ComplEx(Toutanova and Chen, 2015)	$\mathrm{Re}\left(\mathbf{h}^\top diag(\mathbf{r})\bar{\mathbf{t}}\right)$	h, r, t $\in C^k$
RotatE(Sun et al., 2019)	$-\|\mathbf{h}\circ\mathbf{r}-\mathbf{t}\|_2$	h, r, t $\in C^k, \|r_i\|=1$
HAKE(Zhang et al., 2020)	$-\ \|\mathbf{h}_m\circ\mathbf{r}_m-\mathbf{t}_m\|_2-\lambda$	$\mathrm{h}_m, \mathbf{t}_m \in R^k, \mathbf{r}_m \in R^k_+,$
	$\|\sin\left(\left(\mathbf{h}_p+\mathbf{r}_p-\mathbf{t}_p\right)/2\right)\|_1$	$\mathrm{h}_p, \mathbf{r}_p, \mathbf{t}_p \in [0,2\pi)^k, \lambda \in R$

表 1: 表示学习模型的得分函数与参数对比

3 TiKG-30K数据集的构建及优化

3.1 藏语知识图谱扩充

本文在前期爬取了中国藏族网通、宗喀巴网等许多藏文网站大量的原始藏文文章，并依据主题划分成了常识、旅游、法律、地理等不同的类型，使用词性标注系统对所有文章的字词进行了标注。参考高定国等人对藏语单句句型的研究(高定国and 扎西加, 2014)，本文首先对藏文文章中符合三元组提取的句型进行筛选，接下来，根据藏语单句中的词性标注，采用基于词性

第二十二届中国计算语言学大会论文集，第145页-第154页，哈尔滨，中国，2023年8月3日至5日。

组合的规则对单句中的结构进行拆分(扎西吉, 2018)，得到了主语、谓语、宾语等不同句子成分，组合成原始的三元组30余万条。

藏语知识库中三元组比较稀疏，因此如何借助其他语言丰富的知识库和非文本介质对知识库进行扩充，是本文的重要研究内容。我们通过藏文三元组中实体的同指关系（比如藏文三元组<སྐུ་འབུམ་དགོན།，རྒྱ་ཡིག་གི་མིང་།（中文名），塔尔寺>），借助于“塔尔寺”在百度百科、维基百科已有信息盒的“实体—属性—值”三元组关系，利用构建的18万汉藏对照词典，以及15,387条的藏汉语命名实体库，扩充藏文实体关系及属性值。另外，借助于地图兴趣点（POI）等非文本介质信息，获得对应中文实体的其他关系属性，传递给藏文实体，扩充藏文实体关系及属性值，解决藏语知识库数据量少、数据稀疏的问题。

对于低资源语言来说，单纯的获取文本语料并抽取实体和关系的规模比较有限，比如在藏语知识库中，存在<སྐུ་འབུམ་དགོན།，རྒྱ་ཡིག་གི་མིང་།（中文名），塔尔寺>三元组，而སྐུ་འབུམ་དགོན།（塔尔寺）这一实体只存在一个关系、属性三元组，这种数据稀疏的现象在藏语知识库中非常普遍。因此，本文借助其他语言丰富的知识库和非文本介质对知识库进行扩充，如图1所示。

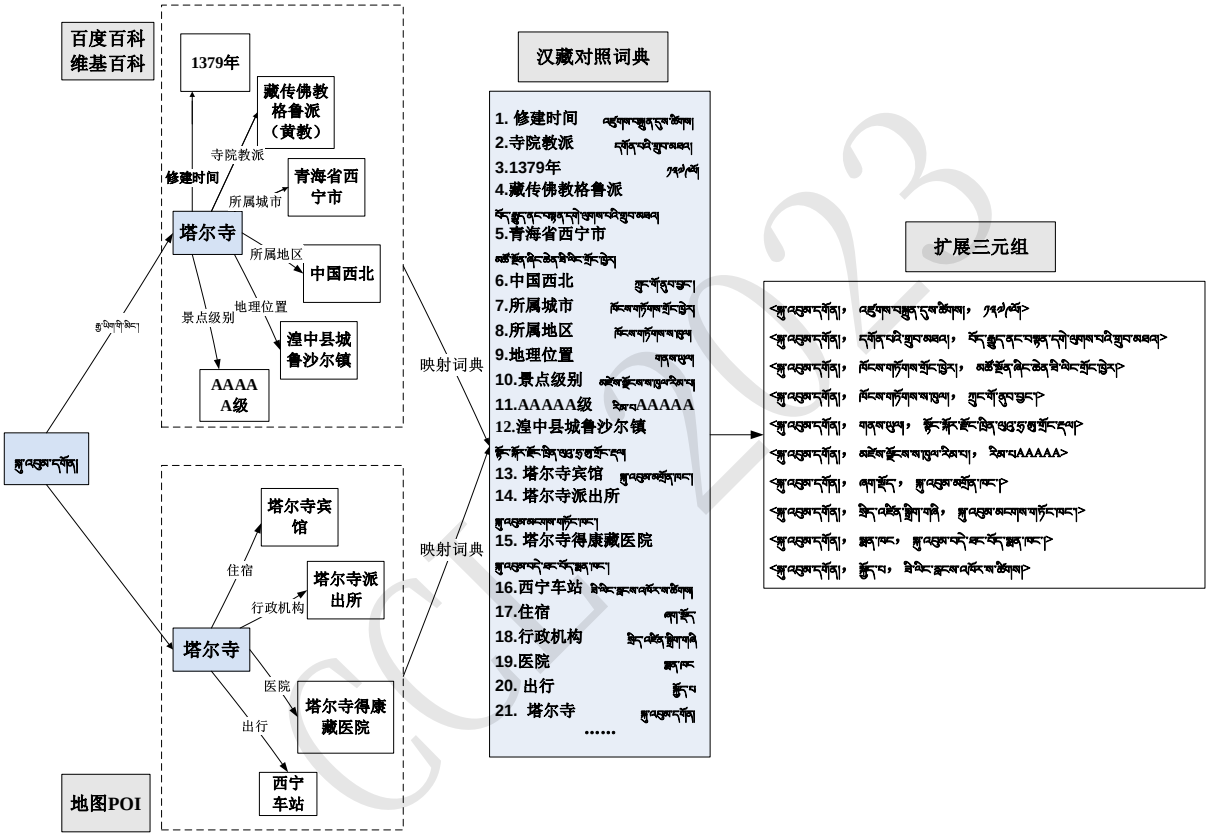


图 1: 藏语知识库扩充方法

我们借助百度百科、维基百科查找到“塔尔寺”这一实体，可以获得已有信息盒的“实体—属性—值”三元组关系，同时利用地图POI信息，扩充“塔尔寺”周边信息，最后利用汉藏对照词典，实现藏文实体关系属性的扩充，解决藏文知识库数据稀疏问题，如表2所示。结合百科与地图POI信息，我们将三元组数量扩充，去重后得到完整的三元组498,258条，此时实体数量达到348,596条，关系类型16,765种，以它作为初始的藏文知识图谱TiKG-V0。

3.2 藏语知识图谱优化

3.2.1 筛选实体和关系类型

TiKG-V1：在TiKG-V0中，由于实体和关系类型的数量相对较大，即使我们对藏语知识库进行了扩充，但整体依然很稀疏，不便于表示学习模型学习到很好的特征。因此，我们采用出现次数作为筛选标准，对所有实体和关系类型进行了统计。考虑到所需知识图谱的规模，我们

扩充前的藏文三元组	百度百科、维基百科、地图POI扩充	扩充后的藏文三元组
<སྐུ་འབུམ་དགོན།，ཀུ་ཡིག་གི་ཨིང་།，塔尔寺>	<塔尔寺，修建时间，1379年>	<སྐུ་འབུམ་དགོན།，འཛུགས་བསྒྲུབ་དུས་ཚོགས།，1379/ལོ།>
	<塔尔寺，寺院教派，藏传佛教格鲁派（黄教）>	<སྐུ་འབུམ་དགོན།，དགོན་པའི་གྲྭ་མཚན།，བོད་རྒྱུད་ནང་བསྒྲུབ་དགེ་ལུགས་པའི་གྲྭ་མཚན།>
	<塔尔寺，所属城市，青海省西宁市>	<སྐུ་འབུམ་དགོན།，ཞོང་གཏོགས་གོང་ཁྲེས།，མཚོ་སྨོན་ཞིང་ཆེན་ཞི་ཡིང་གོང་ཁྲེས།>
	<塔尔寺，所属地区，中国西北>	<སྐུ་འབུམ་དགོན།，ཞོང་གཏོགས་ས་ཁུལ།，ཀྲུང་གོ་རྒྱལ་བྱང་།>
	<塔尔寺，地理位置，湟中县城鲁沙尔镇>	<སྐུ་འབུམ་དགོན།，གནས་ཡུལ།，སྐོང་སྐོར་རྩོང་ཁྲིན་ལུའུ་ཏུ་ཚུ་ཏོང་དཀྱི།>
	<塔尔寺，景点级别，AAAAA级>	<སྐུ་འབུམ་དགོན།，མཐོ་སྡེད་ས་ཁུལ་རིམ་པ།，རིམ་པAAAAA>
	<塔尔寺，住宿，塔尔寺宾馆>	<སྐུ་འབུམ་དགོན།，ཞག་རྩོད་，སྐུ་འབུམ་མཇོན་ཁང་།>
	<塔尔寺，行政机构，塔尔寺派出所>	<སྐུ་འབུམ་དགོན།，ཕྱི་འཛིན་ལྷག་གཞི，སྐུ་འབུམ་མངགས་གཏོང་ཁང་།>
	<塔尔寺，医院，塔尔寺得康藏医院>	<སྐུ་འབུམ་དགོན།，སྨན་ཁང，སྐུ་འབུམ་བདེ་ཞང་བོད་སྨན་ཁང་།>
	<塔尔寺，出行，西宁车站>	<སྐུ་འབུམ་དགོན།，སྐྱོད་པ，ཟི་ཡིང་སྒངས་འཕོར་ཚོགས།>

表 2: 藏语知识库扩充

经过多次尝试，最终选择了筛选出现次数不小于8次的实体和不小于15次的关系类型。经过筛选，共得到166,086条三元组。

FB15k、WN18RR等常用于表示学习的数据集，往往需要有训练集（train）、验证集（valid）、测试集（test）三部分，且验证集和测试集中出现的实体和关系必须在训练集中出现过，这是因为这些数据集的目标是评估表示学习模型对于已知实体和关系的推理能力，而不是处理未知实体和关系。据此，我们再对166,086条三元组进行进一步的划分，首先选出127,664条三元组作为训练集，将剩下的三元组进行判定，若一条三元组的实体和关系都在训练集中出现过则保留，得到验证集和测试集各15,000条。

3.2.2 合并同义实体和关系类型

TiKG-V2：经过筛选后，TiKG-V1的稠密程度相较TiKG-V0有了很大提升，但是我们发现在TiKG-V1中依然存在一些问题，典型的就是语义重复的实体占据了一定数量。比如，“བེ་ཇིང་།”（北京）、“བེ་ཇིན་གོང་ཁྲེ་”（北京市）、“ཀྲུང་ཧྭ་མི་དམངས་སྤྱི་མཐུན་རྒྱལ་ཁབ་ཀྱི་རྒྱལ་སྡ་”（中华人民共和国首都）、“ཀྲུང་གོའི་བེ་ཇིན།”（中国北京），这些实际意义相同的实体在数据集中以独立实体出现，倘若能将这些实体合并，会降低数据冗余，构建出关系更丰富的图谱，无论对于知识推理还是提高查询准确率都有一定帮助。

关于如何寻找藏语近义词的研究较少，基于规则的寻找方法也不够准确。考虑到在跨语言对比中，有些近义词往往会有相同的映射，如果将同一映射中的内容粗略地视作近义词，会合并很多近义实体和关系。采用这种方式，本文对符合条件的近义词进行合并，合并之后会出现少许重复，再次去重得到TiKG-V2。

TiKG-V3：在TiKG-V2中，利用跨语言对比的方式虽然能合并一些近义实体，但仍然会漏掉一些，比如上述的“བེ་ཇིང་།”（北京）和“བེ་ཇིན་གོང་ཁྲེ་”（北京市）在跨语言对比中容易得到相同的映射，但“ཀྲུང་ཧྭ་མི་དམངས་སྤྱི་མཐུན་རྒྱལ་ཁབ་ཀྱི་རྒྱལ་སྡ་”（中华人民共和国首都）仅仅通过跨语言对比很难与前两个实体进行合并。

通过更细致的观察，我们发现在扩充的三元组中，很多头实体并不相同，但是对应的关系和尾实体相同，这是因为在扩充时，很多别名或者代称在百科中会有相同的搜索结果，例如，“ཤར་ཧོག་ཀྱི་རྒྱ་ཆེ་ཁ།”（东方之珠）、“དངོས་ཟོག་ཚོང་ཁྱབ་བྱེད་ས།”（购物天堂）、“གྲུང་གིའི་ཞང་ཀང་།”（中国香港）这三个关键词在百科中都会对应到香港特别行政区。对此，我们在TiKG-V1的基础上不进行TiKG-V2的合并，直接对有着相同关系类型及尾实体的头实体进行合并，并且把TiKG-V2中有相同映射的联系进行合并，得到TiKG-V3。

TiKG-30K：在以上几个版本的优化中，仅靠跨语言对比、寻找等方式并不能充分、正确地对近义词进行合并，因此在最后，需要加入人工来审查及合并。在TiKG-V3的基础上，我们检查了关系类型，调整不准确的近义关系，合并了新的关系类型。检查已经合并的实体，修正少量错误，同时参考TiKG-V2中具有相同映射的近义实体，对语义相同但格式有区别的实体（如“གྲུང་གོ་ཅོ་ནན་ནན་དཔང་།”（中国·河南·南阳）、“ཅོ་ནན་ཞིང་ཆེན་ནན་དཔང་ཚོང་ཁྲོང་།”（河南省南阳市））、尾实体中与头实体语义相似的实体进行合并。

在完成上述人工审查合并工作后，我们得到了最终的知识图谱TiKG-30K。它的实体和关系类型较为常见，内容与中国地理具有较强相关性，实体和关系可靠准确，数据公开，开放性和可扩展性强，能够为针对藏语知识图谱的表示学习模型、藏语问答系统等领域提供支持。TiKG知识图谱各版本信息如表3所示。

数据集	实体	关系类型	训练集	验证集	测试集
TiKG-V0	348,596	16,765	498,258	-	-
TiKG-V1	34,836	698	127,664	15,000	15,000
TiKG-V2	33,521	659	125,995	14,960	14,958
TiKG-V3	32,164	659	117,633	14,834	14,818
TiKG-30K	**30,986**	**641**	**117,051**	**14,820**	**14,808**

表 3: TiKG知识图谱各版本的对比

4 实验结果与分析

在一般情况下，数据集关系类型的数量远少于实体数量，所以实体链接预测难度相较关系预测难度也更大(曾平, 2018)，因此本文在测评任务中选择难度较大的实体链接预测。

4.1 评估指标

（1）Mean Reciprocal Rank：平均倒数排名，简称MRR，用于衡量知识图谱中实体链接任务性能的常用指标之一。给定一个查询q_i，模型需要在知识图谱中为其找到相应的实体。由于知识图谱中可能存在多个与查询实体相似的实体，因此需要对这些实体被选中的概率值由高到低排序，假设正确的尾实体e_i的排名为rank(e_i)。对所有的链接预测任务计算正确排名的平均值，可以得到Mean Rank（平均排名）指标，简称MR，但是该指标受待预测实体数量的影响，并不能真实地反映模型的性能。

$$MR = \frac{\sum_{q_i \in Test} rank(e_i)}{\mid Test \mid}$$

而MRR将排名的倒数进行求和取平均，将评测指标范围限制在0 ∼ 1之间，数值越大，客观上说明模型的性能越好，因此本文采用MRR作为评测指标之一。

$$MRR = \frac{\sum_{q_i \in Test} \frac{1}{rank(e_i)}}{|Test|}$$

（2）Hits@k：k命中率，正确排名rank(e_i)排在top-k所占的比例，取值范围0 ∼ 1，取值越大，模型性能越好，在本文测评中，k值分别取1，3，10。另外需要注意的是，给定头实体和关系进行预测时，在有多个正确的尾实体时，如果测试样例给出尾实体的排名为n，排在前边n-1个的实体中如果出现正确的尾实体也会被模型认为是错误的，这会影响正确的实验结果。因此在进行评测时，需要先将其他正确的尾实体过滤掉，经过过滤处理的记为Filtered，未经

处理的记为Raw。因过滤处理后的评测结果更为合理，故本文的评测指标全部基于过滤后的数据。

$$Hits@k = \frac{\sum_{q_i \in Test} I\left(\frac{1}{rank(e_i)} \leq k\right)}{|Test|}$$

4.2 TiKG-30K与基准数据集对比实验结果

由于目前公开可用的藏文数据集相对较少，我们使用TD50K(Sun et al., 2021)作为藏文基准数据集。TiKG-30K与基准数据集的对比如表4所示。

数据集	实体	关系类型	训练集	验证集	测试集
WN18RR	40,943	11	86,835	3,034	3,134
FB15k-237	14,541	237	272,115	17,535	20,466
TD50K	12,573	3,285	27,754	9,253	9,251
TiKG-30K	30,986	641	117,051	14,820	14,808

表 4: TiKG-30K与基准数据集的对比

与已有的藏文知识图谱TD50K相比，TiKG-30K提供更多、更全面、更准确的实体关系信息，模型能够学到更好的表示，根据表5实验结果，TiKG-30K更适合用于表示学习的研究。

	TransE				DistMult			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TD50K	-	-	-	.25	-	-	-	.31
TiKG-30K	.496	.419	.548	**.625**	.399	.367	.416	**.457**

表 5: TiKG-30K与TD50K的实验对比结果

使用链接预测常用基准数据集FB15k-237、WN18RR与TiKG-30K进行对比实验。WN18RR是WN18(Toutanova and Chen, 2015)数据集的子集，由于WN18中存在测试集泄露的问题，因此改进的WN18RR内容更加合理。对比实验结果如表6所示。

	TransE				DistMult			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
WN18RR	.226	-	-	.501	**.43**	**.39**	**.44**	**.49**
FB15k-237	.294	-	-	.465	.241	.155	.263	.419
TiKG-30K	**.496**	.419	.548	**.625**	.399	.367	.416	.457
	ComplEx				RotatE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
WN18RR	.44	.41	.46	.51	.476	.428	.492	.571
FB15k-237	.247	.158	.275	.428	.338	.241	.375	.533
TiKG-30K	**.479**	**.437**	**.502**	**.554**	**.529**	**.483**	**.553**	**.612**
	pRotatE				HAKE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
WN18RR	.462	.417	.479	.552	.497	.452	.516	.582
FB15k-237	.328	.230	.365	.524	.346	.250	.381	.542
TiKG-30K	**.526**	**.468**	**.557**	**.630**	**.534**	**.483**	**.561**	**.629**

表 6: TiKG-30K与英文基准数据集的实验对比结果

将TiKG-30K与WN18RR、FB15k-237的Hits@10指标进行直观对比，如图2所示。结合表5、6进行分析，本文提出的TiKG-30K在实验中的各项指标相较基准数据集均有所提高，说明TiKG-30K在知识图谱链接预测任务上具有更好的性能表现。

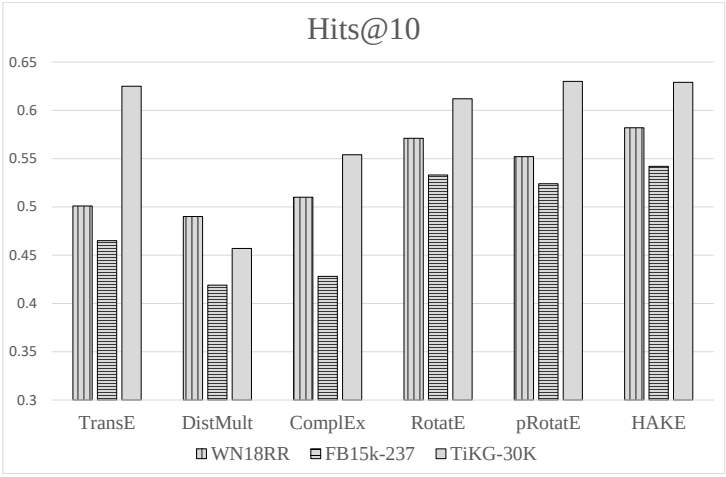


图 2: TiKG-30K与WN18RR、FB15k-237在不同模型上Hits@10的实验结果

4.3 消融实验结果

本文构建TiKG-30K时，通过跨语言近义词检索、合并同义实体和关系、修正错误三元组等技术对知识图谱进行多层优化，为了验证优化方式有效，需要对TiKG-V1、TiKG-V2、TiKG-V3、TiKG-30K四个不断优化的知识图谱进行消融实验。对于每个进行链接预测的模型，在相同的参数设置下，分别记录MRR、Hits@1、Hits@3、Hits@10指标的实验结果如表7所示，将实验结果中Hits@10指标进行对比如图3所示。

	TransE				DistMult			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TiKG-V1	.440	.346	.505	.594	.360	.322	.382	.428
TiKG-V2	.446	.351	.514	.599	.358	.320	.381	.426
TiKG-V3	**.496**	**.423**	.542	.621	**.403**	**.371**	**.422**	**.460**
TiKG-30K	**.496**	.419	**.548**	**.625**	.399	.367	.416	.457
	ComplEx				RotatE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TiKG-V1	.432	.383	.462	.521	.484	.430	.512	.578
TiKG-V2	.436	.388	.465	.521	.492	.441	.519	.584
TiKG-V3	.476	.435	.499	.551	.528	**.485**	.549	.609
TiKG-30K	**.479**	**.437**	**.502**	**.554**	**.529**	.483	**.553**	**.612**
	pRotatE				HAKE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TiKG-V1	.480	.410	.523	.604	.490	.427	.526	.601
TiKG-V2	.490	.423	.529	.608	.498	.437	.533	.604
TiKG-V3	.520	.464	.550	.623	.528	.477	.554	.622
TiKG-30K	**.526**	**.468**	**.557**	**.630**	**.534**	**.483**	**.561**	**.629**

表 7: TiKG各版本数据集的消融实验对比结果

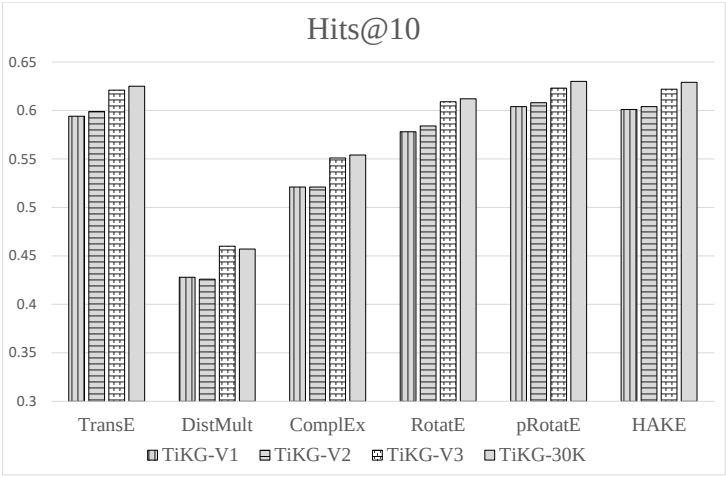


图 3: 不同模型上Hits@10的消融实验结果

综合上述消融实验结果，可以得出以下结论：

（1）在ComplEx、pRotatE、HAKE模型上进行链接预测任务时，TiKG-30K的各项指标均领先于前三个版本数据集的同类指标；

（2）在TransE、RotatE模型中，TiKG-30K仅有Hits@1指标落后于TiKG-V3，平均落后0.3%；仅在DistMult模型中，TiKG-30K各项指标小幅度落后于TiKG-V3；

（3）总体上TiKG-V1、TiKG-V2、TiKG-V3、TiKG-30K在不同模型中的各项指标均随着数据集的优化而提升，这证实本文的优化方式是有效的。综上可以认为TiKG-30K是这四个版本中结构更加合理、内容更加准确的知识图谱数据集，更适合应用于藏语表示学习领域。

5 总结与展望

本文构建了藏语知识图谱TiKG-30K用于藏语表示学习的研究，借助百科与地图POI信息对知识图谱进行扩充，有效缓解了藏文知识图谱三元组数量少、数据稀疏的问题，通过多种方式合并近义实体和关系，减少数据冗余和语义重复，有效解决了知识图谱中关系稀疏的问题。在TransE、DistMult、ComplEx、RotatE、pRotatE、HAKE表示学习模型上进行链接预测任务评估，实验结果表明，对比基准数据集与TiKG各版本数据集，TiKG-30K有着更好的表现，证实本文数据优化方式的有效性。由于TiKG-30K比已有的藏语知识图谱有着更丰富合理的实体关系表示，所以它能够对藏语表示学习的研究、知识推理以及藏语问答系统等多个领域提供高质量的基础数据。在未来我们将探索更多数据源，对现有的藏语知识图谱进行扩充，对TiKG-30K进行更多方向的优化，以能够胜任更多藏语自然语言处理的任务，推动自然语言处理在藏语等低资源语言领域的发展。

致谢

本论文得到了国家自然科学基金项目（61972436）和国家社会科学基金项目（22ZD035）的资助。

参考文献

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 722–735. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Yuan Sun, Andong Chen, Chaofan Chen, Tianci Xia, and Xiaobing Zhao. 2021. A joint model for representation learning of tibetan knowledge graph based on encyclopedia. *Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–17.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 57–66.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. 2013. Xlore: A large-scale english-chinese bilingual knowledge graph. In *ISWC (Posters & Demos)*, pages 121–124.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part II*, pages 428–438. Springer.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3065–3072.

冯小兰and 赵小兵. 2019. 汉藏双语旅游领域知识图谱系统构建. 中文信息学报, 33(11):64–72.

刘知远, 孙茂松, 林衍凯, and 谢若冰. 2016. 知识表示学习研究进展. 计算机研究与发展, 53(2):247–261.

扎西吉. 2018. 基于pcfg的藏语句法分析. Master's thesis, 青海师范大学.

曾平. 2018. 基于文本特征学习的知识图谱构建技术研究. Ph.D. thesis, 国防科技大学.

高定国and 扎西加. 2014. 藏语单句的基本句型研究. 中国藏学, (4):127–133.

噪声鲁棒的蒙古语语音数据增广模型结构

马志强[1,2*], 孙佳琦[1], 李晋益[1], 王嘉泰[1]
[1] 内蒙古工业大学数据科学与应用学院, 呼和浩特, 010080
[2] 内蒙古自治区基于大数据的软件服务工程技术研究中心, 呼和浩特, 010080
mzq_bim@imut.edu.cn

摘要

蒙古语语料库中语音多样性匮乏，虽然花费人力和经费收集数据在一定程度上能够增加语音的数量，但整个过程需要耗费大量的时间。数据增广能够解决这种数据匮乏问题，但数据增广模型的训练数据包含的环境噪声无法控制，导致增广语音中存在背景噪声。本文提出一种TTS和语音增强相结合的语音数据增广方法，以语音的频谱图为基础，从频域和时域两个维度进行语音增强。通过多组实验证明，蒙古语增广语音的合格率达到70%，增广语音的CBAK和COVL分别下降了0.66和0.81，WER和SER下降了2.75%和2.05%。

关键词： 语音增强；数据增广；噪声鲁棒性；蒙古语

Noise robust Mongolian speech data augmentation model structure

Ma Zhiqiang[1,2*], Sun Jiaqi[1], Li jinyi[1], Wang Jiatai[1]
[1] College of Data Science and Application Inner Mongolia University of Technology, Huhhot, 010000
[2] Inner Mongolia Autonomous Region Software Service Engineering Technology Research Center Based on Big Data, Huhhot, 010000
mzq_bim@imut.edu.cn

Abstract

There is a lack of phonetic diversity in Mongolian corpus. Although manpower and funds spent on data collection can increase the number of phonetic sounds to some extent, the whole process needs a lot of time. Data augmentation can solve the problem of data scarcity, but the environmental noise contained in the training data of the data augmentation model cannot be controlled, resulting in background noise in the augmentation speech. In this paper, a speech data augmentation method combining TTS and speech enhancement is proposed. Based on the speech spectrum graph, speech enhancement is carried out from two dimensions: frequency domain and time domain. Multiple experiments show that the qualified rate of Mongolian augmented speech reaches 70%, the CBAK and COVL of augmented speech decrease by 0.66 and 0.81, and WER and SER decrease by 2.75% and 2.05%, respectively.

Keywords: Speech enhancement , Data augmentation , Noise robustness , Mongolian

第二十二届中国计算语言学大会论文集，第155页-第163页，哈尔滨，中国，2023年8月3日至5日。

1 引言

蒙古语作为一种低资源语言，在深度学习任务中其训练数据十分有限。为了扩充蒙古语音数据集的规模并提高训练数据的数量，可以采用语音数据增广。语音数据增广可以通过原始语音数据生成新的语音样本，以有效地解决语音数据匮乏和多样性不足的问题。TTS是一种常用的语音数据增广方法，可以提高蒙古语语料库中语音说话人的多样性。然而，与在受控条件下录制蒙古语语音语料库不同，使用TTS语音数据增广方法常需要依赖于域外语音，以引入域外说话人的特征。这可能会导致增广语音受到无法控制的混响和环境噪声的影响。此外，在蒙古语发音中，存在较为复杂的发音现象，比如气音、元音的颤音、非单元音辅音和元音拼接等。这些现象容易导致语音数据增广模型过度拟合域外语音中的噪声，从而导致增广后的语音失真。因此，本文旨在解决语音数据增广过程中的噪音鲁棒性问题，将SE方法应用于基于语音合成(Text to Speech，TTS)的增广框架中，以提高增广语音数据的质量。

语音增强[1](Speech Enhancement，SE)旨在通过减少背景噪声来提高语音质量，是语音数字信号处理中的一个重要环节，其主要目标是提高语音的清晰度，进而提高下游任务对噪声的鲁棒性。语音增强主要分为信号处理和深度神经网络两种方法，标准的信号处理方法以频谱减法和维纳滤波为主，频谱减法[2]通过减去非语音活动期间计算的频谱噪声偏差来抑制语音中的静音噪声，然后衰弱减法后的残余噪声。维纳滤波[3]使用均方误差去估计纯净信号谱，通过估计线性滤波器对不同频段的噪声进行不同程度的抑制。但基于信号处理的SE方法的应用范围有限，如噪声具有一定的平稳性，干净语音和噪声不相关等，这限制了算法的性能和应用场景。基于深度神经网络的语音增强具有更好的性能和更加丰富的应用场景，该方法主要用于直接重建干净语音[4]或从噪声信号中估计掩码[5]，采用监督的方式进行去噪，将倒谱或频谱表示转换为波形表示，相比提取梅尔频率倒谱系数能够保留更加完整的语音信息。受深度神经网络的启发，将SE与语音数据增广模型相结合提高语音增广模型的噪声鲁棒性，消除域外语音中引入的噪音，提高增广语音的质量。

目前，将语音增广作为下游任务的SE研究并不多，本文将SE方法加入到基于TTS的蒙古语语音增广模型中，该方法结合了蒙古语的发音特点和噪声对增广语音的影响，以提高增广模型的噪音鲁棒性。

2 相关工作

基于深度学习的SE方法主要包括DNN、CNN和LSTM等网络。Karjol等人[6]提出基于多个DNN的增强方法，利用一个门控网络提供权重来组合多个输出，该方法有效的提高了语音质量感知评估(Perceptual Evaluation of Speech quality，PESQ)。Y Zhao等人[7]通过在损失函数中加入语音可理解度的度量进行DNN的扩展，在多种信噪比和噪声类型下提高语音清晰度。Bagchi等人[8]也进行DNN模型的扩展，将模拟损失与传统损失结合起来训练语音增强模型。但在基于DNN的SE网络需要大量参数，此外在低信噪比条件下还会导致增强语音恶化。

为了有效的学习语音时间信息，SE方法开始由DNN向循环神经网络(Recurrent Neural Network，RNN)和卷积神经网络(Convolutional Neural Network，CNN)转变。Maas等人[9]使用RNN增强带噪声语音的MFCC特征，利用立体噪声和干净音频特征进行训练，预测噪声语音的干净特征。Gao等人[10]提出了一种LSTM的渐进式学习框架，将输入和中间目标的估计进行拼接后学习下一个目标，这种方法可以充分利用多个学习目标的信息，缓解信息丢失，提高低信噪比环境下的SE性能。此外，采用RNN还能够实现非平稳加性噪声[11]、混响[12]和多通道噪声语音[13]的语音去噪。基于CNN的语音增强能够处理语音的局部时间信息，可以有效地分离噪声信号中的语音和噪声信息，在频谱域和波形域都能完成语音增强。Kinoshita等人[14]受语音分离的时域卷积网络[15]启发，采用CNN在时域上进行掩蔽估计去噪，完成语音增强。P Plantinga等人[16]使用残差网络去除残余噪声重构输入信号，借助ResNet频谱映射架构来提高语音增强性能，将该方法与下游模型联合训练，实现下游任务中模型的噪声鲁棒性。

通过RNN和CNN神经网络进行语音增强可以捕获语音信号的时序特征，通过多次迭代学习到更多的特征以此来提高语音增强的准确性，但RNN和CNN网络模型存在某些参数不具备可解释性，因此网络结构上的优化比较困难，导致模型训练时可控性较差。将SE方法加入到基于TTS的蒙古语语音增广模型中可以在模型训练时对所需的性能进行调整和优化，从而更好地控制优化模型得到输出语音质量更好的语音数据。

第二十二届中国计算语言学大会论文集，第155页-第163页，哈尔滨，中国，2023年8月3日至5日。

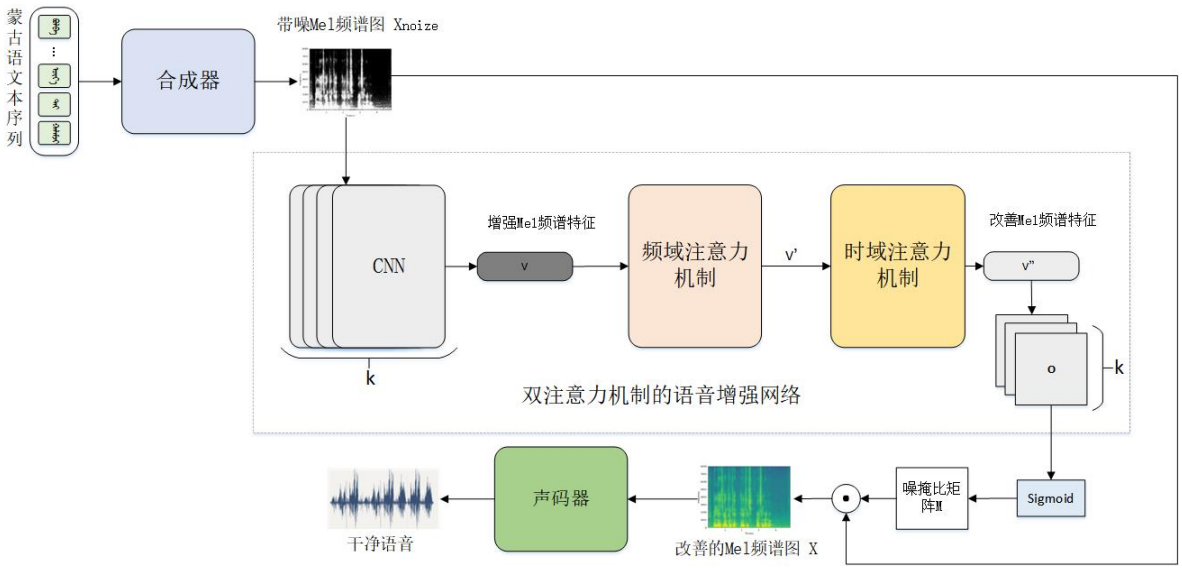


Figure 1: 噪声鲁棒的语音生成器架构图

3 方法

3.1 噪声鲁棒的语音生成器

为了提高基于TTS增广模型对噪声的鲁棒性，在语音生成器中加入SE网络，噪声鲁棒的语音生成器架构如图1所示，合成器将蒙古语文本转换为Mel频谱图，因为Mel频谱图能够同时表示语音时域和频域的信息，SE网络利用这一特性使用双注意力机制进行增强，图1中虚线部分为双注意力机制的语音增强网络(Double Attention Speech Enhancement，D-Attention-SE)。$X_{noize} = R^{T*F}$表示域外语音的Mel频谱向量，T表示时间维度，F表示频率维度。使用k层CNN对X_{noise}增强，输出增强后的Mel频谱特征V，然后利用频域注意力机制从频域维度增强得到特征V'；将V'输入到时域注意力机制完成时域增强，得到改善后的增强频谱特征V"。双注意力机制语音增强网络将改善后的特征V"细化为$O = R^{T_k F_k}$,其中k表示k层CNN改善的频谱特征。最后，细化特征O通过Sigmoid计算输出噪掩比矩阵M，将该矩阵与原始频谱图进行点积，过滤掉被噪声破坏的Mel频谱图，输出最终改善的Mel频谱图X，声码器利用X将Mel频谱图转换为语音，实现频域和时域维度的语音增强。

3.2 双注意力机制网络

双注意力机制包括频域注意力机制和时域注意力机制，其网络结构如图2所示，图中上半部分为频域注意力机制，下半部分为时域注意力机制。该网络结构的计算流程为$V' = a_F \odot V$，$V'' = a_T \odot V'$，其中a_F和a_T分别表示频率注意和时域注意的权重，增强Mel频谱特征V与a_F点积，得到频域增强特征V'；V'与a_T点乘，得到时域增强特征V"，实现两个维度的增强改善。在频域注意力机制中，首先从输入特征的信道维度进行最大池化和平均池化，两者的池化结果进行拼接后，通过时间池化进行最大池化和平均池化。最后通过卷积运算和Sigmoid激活函数计算得到频域注意力权重a_F，将a_F与V进行点积，得到频率增强特征V'。时域注意力机制与频域注意力机制类似，对频域注意力机制增强的特征V'进行信道池化和频域池化，利用卷积层和Sigmoid激活函数计算得到时域注意力权重a_t，通过广播后扩展维度，与V'点积得到时域增强特征，实现频域和时域两个维度的Mel频谱增强。

3.3 模型训练

两阶段的语音增强网络使用均方差(Mean Square Error，MSE)作为的损失函数，其损失函数如公式(1)所示。

$$Loss = \frac{\sum \|X_{noise} \odot M - X_{clean}\|^2}{n} \quad (1)$$

其中,X_{noise}表示带噪Mel频谱图，X_{clean}表去噪的Mel频谱图，n表示Mel的频率转换尺度，M表示Mel频谱的噪掩比矩阵。双注意力机制的SE网络在训练过程主要更新频域注意力

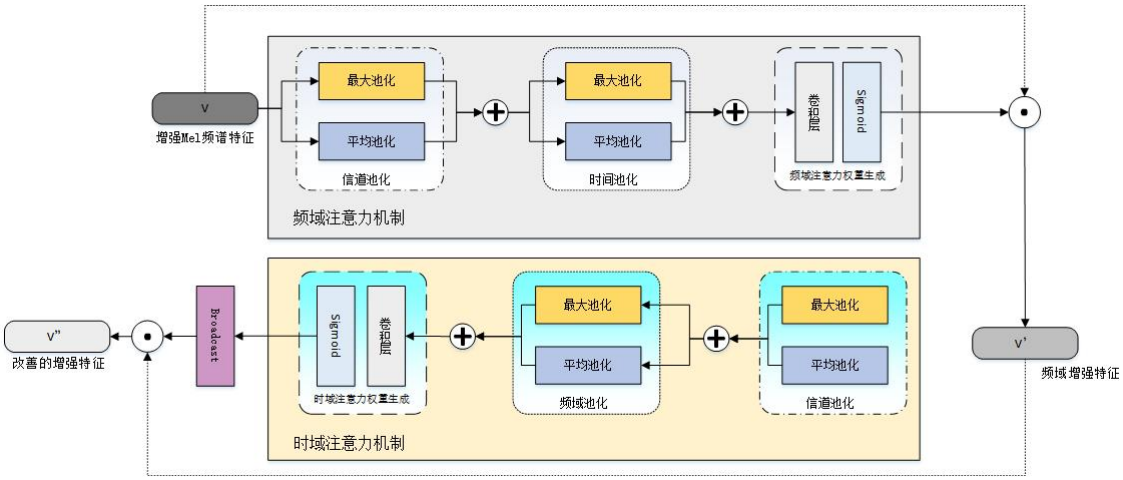


Figure 2: 两阶段的语音增强网络结构图

机制和时域注意力机制的权重，通过输入增强Mel频谱特征，计算出频域注意力权重a_F和时域注意力权重a_T，利用a_F和a_T求出改善后的Mel频谱特征V"，该训练算法如表1所列。

Table 1: 模型训练算法

输入： 带噪语音Mel频谱图X_{noize}；干净语音Mel频谱图X_{clean}；训练轮数$epoch$；
输出： 噪掩比M；改善的Mel频谱图X；

For $i = 0$; $epoch$ **do**
  $V = Train_{cnn}\left(X_{noise}, X_{clean}\right)$ //增强网络CNN训练
  $a_F = F\_Attention(V)$ // 为频域注意力权重
  $V' = V \bigodot a_F$
  $a_T = T\_Attention\left(V'\right)$ // 为时域注意力权重
  $V'' = V' \odot a_T$
  $\mathrm{H} = Sigmoid\left(V''\right)$
  If $X_{noise} \odot M \neq X_{clean}$ **:**
    $a_F = backpropogation(Loss)$
    $a_T = backpropogation(Loss)$
  Else:
    Return X

4 实验

4.1 实验设置

在语音增强网络的训练过程中，将Audio Set[17]数据集中的噪声加入到蒙古语语音数据集IMUT-MC1中，得到干净语音所对应不同环境噪音下的语音，用于训练语音增强网络。Audio Set中包含五类真实世界中的噪声环境，包括人类噪声、动物噪声、自然噪声、音乐噪声和事物噪声，训练数据如表2所列。

Table 2: 语音增强网络训练数据

噪声类别	文本句子	词数	人数	性别	平均词个数	平均时长	总时长
5类	1255条	2237	1	女	6	12秒	4.1h

在两阶段的语音增强网络的训练过程中，语音的采样率为16kHz，将带噪语音和干净语音归一化到[-1,1]，从归一化的结果中提取帧。使用Adam优化器对随机梯度下降进行优化，批量

处理大小为4条语音，为了匹配最长语音的大小，将批处理中较短的语音用0进行填充。此外，训练轮数为500轮，学习率为2e-4。

4.2 评价指标

在噪声鲁棒的语音增广模型实验中，从增广语音的质量、自然度和消除噪声的效果三个角度进行模型评价。使用平均意见得分(Mean Opinion Score，MOS)和生成合格率(Generated Pass Rate，GPR)对语音质量进行评价。该方法的指标包括缺字句数(Missing Words，MW)、多字句数(Insert Words，IW)、字序错误(Word Sequence Error，WSE)、发音错误(Pronunciation Error，PE)、噪音句数(Noise Sentence，NS)和无域外说话人句数(Without Foreign Speakers，WFS)。其中，MW表示增广语音存在字符缺失的句数；IW表示增广语音插入了非对应文本词的句数；WSE表示增广语音发音顺序错误的句数；PE表示增广语音发音错误的句数；NS表示增广语音存在噪音的句数；WFS表示生成语音中没有加入域外说话人的语音。生成合格率(Generated Pass Rate，GPR)：GPR指人工评价中没有错误的句子数占总句数(Total Sentences，TS)的百分比，见式(2)。

$$GPR = \left(1 - \frac{MW+IW+WSE+PE+NS+WFS}{TS}\right) \times 100\% \quad (2)$$

使用梅尔频谱失真(Mel Cepstral Distortion，MCD)对语音的自然度进行评价。使用背景干扰综合测度(The Composite Measure for Background Interferences，CBAK)和总体综合测度[18](The Overall Composite Measure，COVL)评价语音中噪声的抑制程度和效果，其取值范围为1到5。此外，使用增广说话人的语音训练语音识别模型，通过语音识别的词错误率（Word Error Rate，WER）和句错误率(Sentence Error Rate, SER)来评价提出方法的有效性。

4.3 实验结果与分析

4.3.1 有效性实验

为了验证语音增广模型的噪声鲁棒性，使用五种不同类型的噪声作为域外语音进行增广，从每个噪声类型中选择10个不同说话人，每个说话人生成相同的200条语音，增广数据集详情如表3所列。IMUT-MC1-N1、IMUT-MC1-N2、IMUT-MC1-N3、IMUT-MC1-N4和IMUT-MC1-N5表示域外语音中噪声类型数不同的增广数据集，其中IMUT-MC1-N1 ⊆ IMUT-MC1-N2 ⊆ IMUT-MC1-N3 ⊆ IMUT-MC1-N4 ⊆ IMUT-MC1-N5，IMUT-MC1-N5中包含了人类噪声、动物噪声、自然噪声、音乐噪声和事物噪声五类噪声，不同噪声语音增广所对应的文本标签相同，没有相同说话人。

Table 3: 地区数不同的增广数据集

名称	噪声类别数	说话人数	总句数（万）	总时长（h）
IMUT-MC1-N1	1	10	0.2	2.4
IMUT-MC1-N2	2	20	0.4	4.8
IMUT-MC1-N3	3	30	0.6	5.2
IMUT-MC1-N4	4	40	0.8	7.6
IMUT-MC1-N5	5	50	1.0	10.0

使用带有不同噪声的域外语音进行语音增广，对IMUT-MC1-N5中五类噪声背景下的增广语音进行评价，每类噪声背景选择相同文本对应的200条增广语音，从语音的质量、自然度和噪声抑制程度来验证提出方法的有效性，增广语音的评价结果如表4所列。分析表3可知，五种噪声环境下增广语音的GPR都达到65%，增广语音的整体合格率达到70%；MOS均达到4.0以上，与静音环境下增广语音的MOS值接近，这表明在增广模型中引入语音增强，有效的提高了噪声环境下增广语音的质量。MCD的评价结果均小于19.20，Mel频谱失真低于与静音场景下增广语音的19.61，表明语音增强网络有效的提高了增广语音的自然度。此外，CBAK和COVL的值都达到2.6以上，相比域外语音中的1.5，CBAK和COVL有了明显的提示，表明语音增强网络显著的抑制了不同类型的背景噪声。

为了验证Double-Attention-SE针对噪声数据增广的鲁棒性，使用包含10位说话人的语音作为测试集，分别对不同噪声环境下增广数据集训练的声学模型进行测试，测试集从五类噪声中挑选文本相同的200条语音。使用不同噪声环境下的增广数据集构建的声学模型测试结果如

Table 4: 噪声鲁棒的增广语音评价

噪声类型	质量评价		自然度评价	噪声抑制度评价	
	GPR	MOS	MCD	CBAK	COVL
人类噪声	65%	4.10	19.05	2.77	2.71
动物噪声	75%	4.20	18.67	2.82	2.79
自然噪声	70%	4.13	18.23	2.69	2.64
音乐噪声	65%	4.08	19.11	2.66	2.61
事物噪声	75%	4.22	18.58	2.88	2.85

表5所列，在训练数据中加入不同噪声环境下的增广语音，声学模型的I、D和S均呈现下降趋势，而且下降幅度逐渐增大，其中I、D和S对比原始训练数据分别下降2.36%、3.51%和1.33%，这表明加入Double-Attention-SE的增广语音数据能够有效降低声学模型的插入错误和删除错误。此外，WER和SER分别下降7.20%和6.88%，当训练集中包含五类不同噪声环境的增广语音时，IMUT-MC1-N5数据集训练的声学模型在测试数据上的WER和SER最低，这表明加入不同噪声环境下的增广语音作为训练数据没有降低语音识别的准确率，反而因为域外说话人的增加提高了语音识别的识别准确率。

Table 5: 语音识别实验结果

噪声类型	I	D	S	WER	SER
IMUT-MC1	7.29%	12.26%	6.99%	26.54%	31.04%
IMUT-MC1-N1	7.17%	12.14%	6.56%	25.87%	29.91%
IMUT-MC1-N2	6.93%	11.89%	6.42%	25.24%	29.72%
IMUT-MC1-N3	6.55%	11.01%	6.10%	23.66%	28.15%
IMUT-MC1-N4	5.32%	10.16%	5.91%	21.39%	25.88%
IMUT-MC1-N5	4.93%	8.75%	5.66%	19.34%	24.16

4.3.2 消融实验

在基于TTS语音增广模型中加入语音增强网络消除域外语音引入的噪声，为了避免增强过程中破坏域外说话人特征，语音增强网络从频域和时域进行增强，准确消除噪声的同时不破坏说话人特征。增强网络以单个频域注意力机制Single-Attention-SE为基础，加入频域注意力机制，完成两阶段语音增强网络模型Double-Attention-SE的构建。消融实验从增广语音的质量、自然度、噪声抑制程度和语音识别四个方面进行评估，验证两阶段语音增强的去噪效果。该实验选取不同方法增广的语音进行评价，挑选与训练集中相同说话人的干净语音作为测试集进行语音识别测试，消融实验结果如表6所列。分析表6可知，Single-Attention-SE方法与None-SE相比，GPR和MOS分别提高了2%和0.05；MCD降低了0.13；CBAK和COVL分别下降了0.42和0.25，各项指标的改善较小，表明Single-Attention-SE的有效性不足。虽然加入语音增强网络提高了增广语音的质量，但语音识别评价的WER和SER升高了1.78%和1.77%，分析表明语音增强网络破坏了域外语音中的说话人特征，导致语音识别准确率下降。Double-Attention-SE方法与None-SE相比，GPR和MOS分别提高了5%和0.10；MCD降低了1.95；CBAK和COVL分别下降了0.66和0.81。此外，WER和SER下降了2.75%和2.05%，这表明Double-Attention-SE语音增强网络在提高增广语音质量的过程同时没有破坏域外说话人特征。

Table 6: 消融实验结果

方法	GPR	MOS	MCD	CBAK	COVL	WER	SER
None-SE	65%	4.01%	20.68%	3.54%	3.66%	22.09%	26.21%
Single-Attention-SE	67%	4.06%	20.55%	3.12%	3.41%	23.87%	27.98%
Double-Attention-SE	70%	4.11%	18.73%	2.88%	2.85%	19.34%	24.16

4.3.3 对比实验

对比实验将提出的Double-Attention-SE增强方法与基于深度学习的语音增强方法进行对比，包括基于DNN、RNN、LSTM和ResNet的语音增强方法，同时与时单通道域语音增强方法Time-Domain-SE进行对比，将上述方法加入到基于TTS的语音增广模型中，采用相同的文本和说话人相同的噪声语音进行增广，对增广语音进行评价验证所提方法的先进性，对比结果如表7所列。分析表7可知，所有方法对应增广语音的GPR均达到60%以上，其中Double-Attention-SE最高为70%，表明提高方法有效的提高了增广语音的合格率。此外，Double-Attention-SE取得了该组对比实验最低的MOS、MCD、CBAK和COVL，均优于其它方法。通过对比发现，加入Double-Attention-SE 方法的增广语音CBAK和COVL降低幅度较小，均保持在1.00之内，但WER和SER的有较为明显的下降，与Multiple-DNN-SE方法的差距最大，分别为10.98%和9.86%。表明不同的增强方法都可以有效的抑制增广语音的噪声，但在一定程度上会破坏说话人特征，导致语音识别的准确率降低，Double-Attention-SE在增强过程中对说话人特征的破坏最低，相比其它方法取得最低WER和SER。

Table 7: 语音增强对比实验结果

方法	GPR	MOS	MCD	CBAK	COVL	WER	SER
Single-Attention-SE	60%	3.98%	21.12%	3.72%	3.81%	30.32%	34.02%
Double-Attention-SE	62%	4.01%	21.05%	3.46%	3.59%	25.65%	29.88%
Single-Attention-SE	65%	4.05%	20.88%	3.41%	3.53%	24.87%	28.35%
Double-Attention-SE	68%	4.07%	20.19%	3.32%	3.45%	23.92%	28.09%
Single-Attention-SE	68%	4.09%	19.78%	3.11%	3.24%	21.66%	25.79%
Double-Attention-SE	70%	4.11%	18.73%	2.88%	2.85%	19.34%	24.16%

4.3.4 分析实验

为了验证Double-Attention-SE增强网络的适应性，分析实验分别使用Aishell-1中文数据集，Librispeech-clean英文数据集，IMUT-MC1蒙古语数据集，JSUT日语数据集和Ruslan俄语数据集五类不同语言进行增广，将五类噪声分别加入到三种数据集中，从各类增广语音中选取200条语音数据进行评价。使用GPR和MOS评价不同语言增广语言的质量，使用MCD评价增广语音的自然度，CBAK和COVL验证增广模型的噪声鲁棒性。汉语、英语、蒙古语、日语和俄语的分析实验结果如表7所列。分析表7可知，在语音增广模型中加入Double-Attention-SE进行汉语、英语、蒙古语、日语和俄语增广，增广语音的合格率GPR均达到65%以上；MOS均达到4.0以上，这表明该模型针对不同语言进行增广时，语音的质量均能达标。此外，分析表8可知汉语、英语、日语、俄语和蒙古语增广语音的MCD、CBAK和COVL没有明显差距，蒙古语与汉语增广语音的差距最大，为0.29、0.13和0.27，这表明使用Double-Attention-SE进行噪声鲁棒的语音增广方法同样适用于汉语、英语、日语和俄语，从而证明该方法对不同语言具有适应性。

Table 8: 多语言实验结果

评价指标	汉语 Aishell-1	英语 Librispeech-clean	蒙古语 IMUT-MC1	日语 JSUT	俄语 Ruslan
GPR	67%	68%	70%	66%	69%
MOS	4.01	4.09	4.11	4.02	4.07
MCD	19.02	18.98	18.73	18.68	18.94
CBAK	3.01	2.91	2.88	2.93	2.91
COVL	3.12	2.98	2.85	2.74	2.88

5 结论

本文针对基于TTS蒙古语语音数据增广模型的噪声鲁棒性展开研究，提出基于频谱特征的语音增强单元，该单元从时域和频域两个维度进行去噪，消除域外语音中引入的噪声，降低语

音增强对说话人特征的影响。实验结果表明加入语音增强单元后，蒙古语增广语音的合格率达到70%，增广语音的CBAK和COVL分别下降了0.66和0.81，WER和SER下降了2.75%和2.%，这表明基于频谱特征的语音增强网络消除了域外语音中的噪声，提高了蒙古语语音数据增广模型的噪声鲁棒性，提升了增广语音的合格率。

References

[1] Asri Rizki Yuliani et al. "Speech Enhancement Using Deep Learning Methods: A Review". In: *Jurnal Elektronika dan Telekomunikasi* 21.1 (2021), pp. 19–26.

[2] Steven Boll. "Suppression of acoustic noise in speech using spectral subtraction". In: *IEEE Transactions on acoustics, speech, and signal processing* 27.2 (1979), pp. 113–120.

[3] Tim Van den Bogaert et al. "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids". In: *The Journal of the Acoustical Society of America* 125.1 (2009), pp. 360–371.

[4] Yong Xu et al. "An experimental study on speech enhancement based on deep neural networks". In: *IEEE Signal processing letters* 21.1 (2013), pp. 65–68.

[5] Arun Narayanan and DeLiang Wang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 7092–7096.

[6] Pavan Karjol, M Ajay Kumar, and Prasanta Kumar Ghosh. "Speech enhancement using multiple deep neural networks". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5049–5052.

[7] Yan Zhao et al. "Perceptually guided speech enhancement using deep neural networks". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5074–5078.

[8] Deblin Bagchi et al. "Spectral feature mapping with mimic loss for robust speech recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5609–5613.

[9] Andrew Maas et al. "Recurrent neural networks for noise reduction in robust ASR". In: (2012).

[10] Tian Gao et al. "Densely connected progressive learning for lstm-based speech enhancement". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5054–5058.

[11] Martin Wöllmer et al. "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 6822–6826.

[12] Felix Weninger et al. "The Munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks". In: *Proceedings of the 2nd CHiME workshop on machine listening in multisource environments*. 2013, pp. 86–90.

[13] Xiaofei Li and Radu Horaud. "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory". In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2019, pp. 298–302.

[14] Keisuke Kinoshita et al. "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7009–7013.

[15] Yi Luo and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8 (2019), pp. 1256–1266.

[16] Peter Plantinga, Deblin Bagchi, and Eric Fosler-Lussier. “An exploration of mimic architectures for residual network based spectral mapping”. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2018, pp. 550–557.

[17] Jort F Gemmeke et al. “Audio set: An ontology and human-labeled dataset for audio events”. In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[18] Neil Shah, Hemant A Patil, and Meet H Soni. “Time-frequency mask-based speech enhancement using convolutional generative adversarial network”. In: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2018, pp. 1246–1251.

基于数据增强的藏文机器阅读有难度问题的生成

旦正错[1,3] 陈龙[1,3] 邓俊杰[1,3] 庞仙[2,3] 孙媛[1,3,4,*]
[1]中央民族大学 信息工程学院，北京 100081
[2]中央民族大学 中国少数民族语言文学学院
[3]国家语言资源监测与研究少数民族语言中心
[4]民族语言智能分析与安全治理教育部重点实验室
*通讯作者: 孙媛
tracy.yuan.sun@gmail.com

摘要

问题生成是机器阅读理解数据集构建的子任务，指让计算机根据给定有（无）答案的上下文，生成流利通顺的问题集。在中英文领域，以端到端为基础的问题生成模型已经得到了很好的发展，并且构建了大批高质量的问答对。但是在低资源语言（藏文）领域，以机器阅读理解、智能问答系统为代表的数据驱动型任务中仍然普遍存在数据量较少和问答对过于简单的问题。因此，本文提出了三种面向藏文机器阅读的有难度问题的生成方法:（1）基于藏文预训练语言模型进行掩码、替换关键词生成不可回答问题。（2）根据相似段落的问题交叉生成不可回答的问题。（3）根据三元组生成具有知识推理的问题。最后，本文在构建的数据集上进行了实验，结果表明，包含不可回答、知识推理等类型的机器阅读理解数据集对模型的理解能力提出了更高的要求。另外，对构建的不可回答问题，从数据集的可读性、关联性和可回答性三个层面验证了数据集的质量。

关键词: 藏文；不可回答；有难度；数据集；机器阅读理解

Difficult Question Generation of Tibetan Machine Reading Based on Data Enhancement

Zhengcuo Dan[1,3] Long Chen [1,3] Junjie Deng [1,3] Xian Pang [2,3] Yuan Sun [1,3,4,*]
[1] School of information engineering, Minzu University of China, Beijing 100081
[2] School of Chinese Ethnic Minority Languages and Literature, Minzu University of China
[3] National Language Resources Monitoring and Research Center for Minority Languages
[4]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com

Abstract

Question generation is a sub task of constructing machine reading comprehension datasets, aimed at enabling computers to generate fluent question sets based on the context of given (no) answers. In the field of Chinese and English, generative models based on end-to-end have been well developed, and a large number of high-quality question and answer pairs have been constructed. However, in the field of low resource language (Tibetan), there are problems of less data and too simple Q&A pairs in data-driven tasks such as machine reading comprehension and intelligent question answering. Therefore, this paper proposes three methods to generate difficult questions for Tibetan machine reading. (1) Mask and replace keywords based on a Tibetan pre-trained language model to generate unanswerable questions. (2) Generate unanswerable questions based on question exchange in similar paragraphs. (3) Generate questions with knowledge reasoning based on triples. Finally, this paper conducts experiments on the constructed dataset, and the results show that machine reading comprehension datasets containing unanswerable, knowledge reasoning, and other types

第二十二届中国计算语言学大会论文集，第164页-第173页，哈尔滨，中国，2023年8月3日至5日。

put forward higher requirements for the model's understanding ability. In addition, for the constructed unanswerable questions, the quality of the dataset was verified from three aspects: readability, relevance, and answerability.

Keywords: Tibetan , Unanswerable , Difficult , Data set , Machine reading comprehension

1 引言

机器阅读理解（MRC）指机器根据给定的上下文回答相关问题，测试机器对自然语言的理解程度。早期的机器阅读理解依赖于人工制定的规则或基于统计学习模型，存在可移植性差、人工成本高、产生周期长的问题，很难在实际中广泛应用。近年来，大规模、高质量的数据集极大地推动了机器阅读理解的发展。(Hirschman et al., 1999)等人第一次构建了面向机器阅读理解的数据集，包括3-6年级的阅读材料和简单的5W问题。随后出现了面向多项选择式机器阅读理解的数据集MCTest(Richardson et al., 2013)、RACE(Lai et al., 2017)；面向完形填空式机器阅读理解的数据集Children's Book Test(CBT)(Hill et al., 2016)、CNN&Daily Mail(Hermann et al., 2015)；面向区间答案式机器阅读理解的数据集SQuAD(Rajpurkar et al., 2016)和面向自由问答式机器阅读理解的数据集MARCO(Nguyen et al., 2016)、DuReader(He et al., 2018)。随着这些大规模数据集的创建与应用，如R-Net(Wang et al., 2017)、BiDAF(Seo et al., 2016)等模型相继被提出并在机器阅读理解任务上取得不错的效果，目前在SQuAD数据集上最好的模型成绩达到了95.71[0]，超过人类的表现。

根据上述数据集训练出来的模型容易受到对抗样本的攻击，如在原先的数据集中加入根据文本内容所产生的干扰片段(Jia and Liang, 2017)、删除问题或片段的重要部分以把当前问题变得不可回答(Mudrakarta et al., 2018)之后模型仍然能根据文章给出合理但不正确的答案。针对此类问题，(Rajpurkar et al., 2018)等人提出了包含"不可回答"问题的数据集SQuADRUN，在原来的SQuAD十万个问题——答案对的基础上，新增了超过五万个由人类众包者设计的无法回答的问题。此数据集让模型先判断当前问题是否存在答案，然后确定答案。这类任务也可以让模型更加符合人类做阅读理解的习惯和思维方式，目前，SQuADRUN上表现最好的模型成绩为93.21[1]，超过人类89.45的表现。

在低资源语言（藏文）领域，TibetanQA(Sun et al., 2021)是面向藏文抽取式、可回答的机器阅读理解数据集，该数据集规模可观、涵盖内容比较全面，但是无法满足不可回答问题、基于多个信息进行多步推理、加入外部知识等藏文机器阅读理解任务的需求。生成大规模有难度机器阅读理解数据集的研究还处于初步阶段，主要原因在于：（1）通过众包的形式从海量无结构化文本数据中构建相应数据集存在人工成本过高、构建周期过长、质量难以掌控的问题。（2）藏文属于黏着语，藏文中的虚词往往与实词组合的形式出现，如"ཡི་"（的），导致词与词之间没有明确的划分，因此，目前中英文领域的模型无法直接套用在藏文机器阅读理解任务上。针对以上问题，本文提出了三种面向藏文机器阅读理解的有难度问题的自动生成方法，并在相关的实验上验证了数据集的质量。本文的主要贡献如下：

（1）本文通过三元组的隐式实体关系链，提出基于三元组的知识推理问题生成方法，并将其与基于语法规则生成的简单问题进行了比较。

（2）本文通过众包的形式构建了面向藏文机器阅读理解不可回答问题数据集，包含2,200对问答对，并使用藏文预训练语言模型，提出了基于掩码、替换关键词的藏文机器阅读理解不可回答问题的生成方法。

（3）本文通过计算藏文机器阅读理解数据集TibetanQA的段落相似度，设置上下限阈值实现藏文机器阅读理解不可回答数据集的增广。

2 相关研究

问题生成的方法分为基于规则的问题生成和基于神经网络的问题生成。基于规则的问题生

[0]https://paperswithcode.com/sota/question-answering-on-squad11

[1]https://paperswithcode.com/sota/question-answering-on-squad20

成主要利用句法分析和知识库的辅助制定相关的规则，将陈述句改为疑问句来生成问题。最新研究表明，在非常成熟的第三方语义资源和强大的句法分析技术的支撑下，基于规则的问题生成效果优于神经网络模型(Dhole and Manning, 2020)，但由于不同语言、同语言不同领域之间的差异性，规则移植性差，难以扩展，并且人为制定的规则限制了生成问题的多样性。

为了减少人力和缩减构建周期，(Du et al., 2017)等人和(Zhou et al., 2018)等人首次提出基于神经网络模型的问题生成研究。(Du et al., 2017)等人提出了一种基于全局注意力机制的序列学习模型来生成问题。之后，众多学者从不同角度和侧重点对端到端的问题生成展开研究。为了解决问题生成中疑问词与答案类型不匹配、复制机制提取的片段与答案词不相关的问题，提出了将词法、词汇特征(Zhou et al., 2018)、答案信息(Wang et al., 2020)、答案的位置信息(Sun et al., 2018)等各种特征作为输入来提升模型性能。但是，将已知答案信息作为输入特征会使模型自动生成的问题中可能包含目标答案，因此，(Kim et al., 2019)等人使用[mask]标记文本中的答案词，在分开的答案词信息中捕获关键信息，最后采用检索式词生成器(Ma et al., 2018)生成完整的、不包含答案词的问题。比起句子级问题生成，段落级问题生成(Zhao et al., 2018)由于融入了更多的语义信息，其生成的问题质量往往更好。

自预训练语言模型（例如BERT(Kenton and Toutanova, 2019)）及其变体被提出，在机器阅读理解等自然语言处理的下游任务中，其表现远超序列到序列的神经网络模型，但对于不可回答、多跳、加入外部知识等机器阅读理解的复杂任务，模型还是无法做出强有力的判断。因此，(Zhu et al., 2019)等人把SQuADRUN作为不可回答问题生成模型的训练数据，pair-sequence（Pair2Seq）作为问题生成模型，自动生成面向机器阅读理解不可回答问题集。除此之外，在知识推理数据集方面，出现了给每个文档附带多个相关文档的TriviaQA(Joshi et al., 2017)、利用知识图谱构建的QAngaroo(Welbl et al., 2018)、基于多个文档且问题不局限于任何已有的知识库或知识模式的HotpotQA(Yang et al., 2018) 等多跳数据集。

目前，在中英文领域，已经出现了很多机器阅读理解不可回答、知识推理、加入外部知识等的复杂数据集及相关研究。但是在低资源语言（藏文）中，其相关研究还处于初步阶段，因此，本文提出了三种低资源语言（藏文）的不可回答和知识推理的机器阅读理解数据集增广方法，以促进低资源语言（藏文）机器阅读理解的发展。

3 模型架构

3.1 知识推理有难度问题的生成

基于深度学习的机器阅读理解模型已经取得了很大的进步，但是与人类相比，其理解能力有四个方面的不足，主要表现在推理能力弱、可解释性差、缺少外部知识、答案可塑性差。本文通过提取文本中包含的三元组实体的隐式关系，构建了基于三元组的知识推理数据集，并与基于规则生成的简单数据集进行了比较。

3.1.1 基于规则的简单问答对生成

为了检验知识推理问题集的质量，本文根据三元组、三元组显式关系的同义词以及相关的藏文格助词添接语法生成了基于规则的藏文机器阅读理解简单问答对，其构建方法分为三元组提取及匹配，三元组关系的同义词统计，基于规则的问题生成。

1、三元组提取及匹配

王丽客等人构建了103,509条藏文知识库(王丽客et al., 2021)，本文使用TibetanQA的文本与其对齐，没有对齐到的文本，提取并标注其适合的三元组，得到<实体1，关系，实体2，文章>格式的数据，如<ལུའུ་ཨུན།，མཚན་དངོས་，ཀྲིག་རྒྱུའུ་རེན་，ལུའུ་ཨུན། ཁོང་གི་མཚན་དངོས་ལ་ཀྲིག་རྒྱུའུ་རེན་ཞུ་བ་རེད། གསར་བཟེ་པ་ཡིན།>(鲁迅，原名，周树人，鲁迅，他的原名叫周树人，革命家)。

2、三元组关系统计及规则制定

对于对齐得到的三元组，即实体和关系，为了减少因关系出现次数太少而产生的噪音，筛选出现次数最高的前4个关系作为制定规则的依据，包含国家、出版物、出生日期、死亡日期，共有1,846条数据，最高出现次数为892，最低出现次数为24。另外，统计三元组关系的同义词，4种关系共11个不同的同义词。最后利用实体、实体关系的同义词得到藏文简单问答对，数据示例如表1所示。

表1中，根据实体1、关系和语法规则生成简单的问题集，而实体2作为生成问题的答案。另外，对于同一组三元组，实体关系的同义词不同，其生成的问题也不同。文中的关

实体1	关系	实体2	生成的问句
ལུའུ་ཞུན། (鲁迅)	Mother	ལུའུ་རུའི། (鲁瑞)	ལུའུ་ཞུན་གྱི་ཡུམ་ཚེན་ནི་སུ་ཡིན།, ལུའུ་ཞུན་གྱི་མ་ཡུམ་ནི་སུ་ཡིན། (鲁迅的母亲是谁？)
ལུའུ་ཞུན། (鲁迅)	Birthday	1881.9.25	ལུའུ་ཞུན་ནི་དུས་ནམ་ཞིག་ལ་སྐྱེ་འཕྲུང་བ་ཡིན། (鲁迅是什么时候出生的？)

表 1. 基于规则生成的简单问题

系Mother有“ཡུམ།”,“ཡུམ་ཚེན་”,“ མ་མ་ ”等同义词，可以生成ལུའུ་ཞུན་གྱི་ཡུམ་ཚེན་ནི་སུ་ཡིན།或者ལུའུ་ཞུན་གྱི་མ་ཡུམ་ནི་སུ་ཡིན། 等不同的问题。

3.1.2 基于三元组生成的多跳阅读理解数据集

三元组是表示文本结构最常用的一种方法。本文从原始段落中抽取三元组，使用图的节点表示三元组的实体，连线表示实体间的关系，若为实线则表示该实体间的关系是显式可控，若为虚线，则表示实体间的关系为隐式不可控。根据实体间的可推理路径，构建了基于三元组多跳的知识推理数据集，阅读此类数据集需要根据当前段落提供的线索，进行2-4的跳级检索才能得到问题的答案，其数据示例如图1所示。

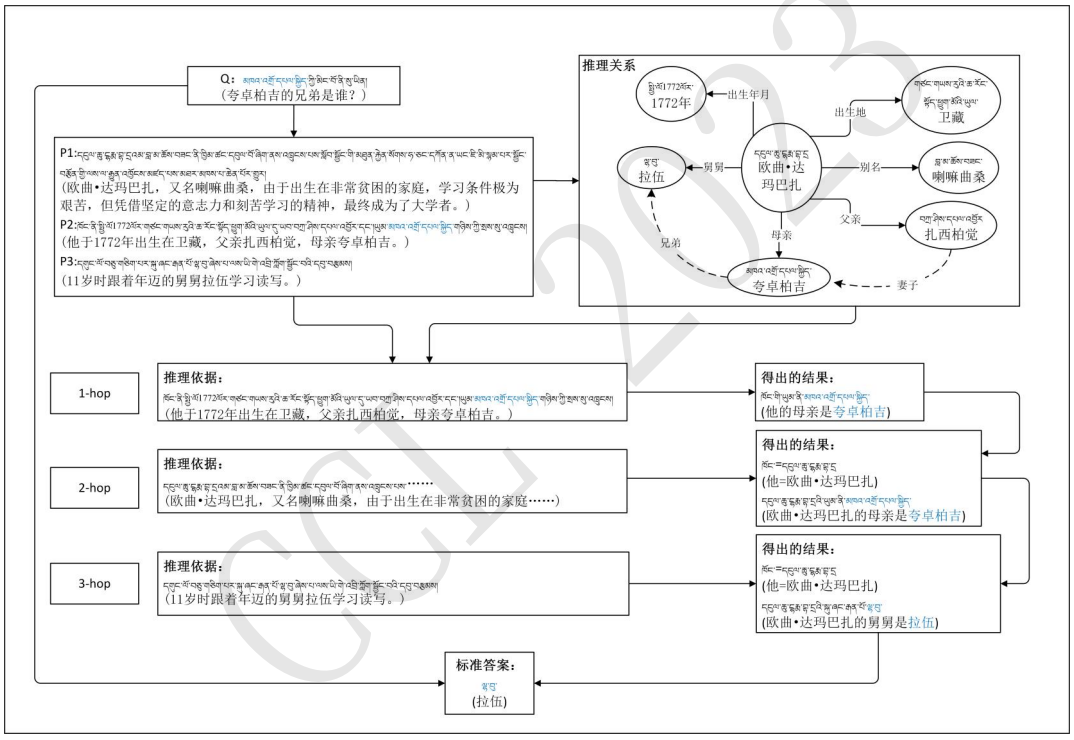


图 1. 基于三元组的知识推理问题生成方法

图1中，问题མཚན་འགོ་དཀར་སྐྱིད་ཀྱི་མིང་ཙོ་ནི་སུ་ཡིན། （夸卓柏吉的兄弟是谁？），在文中没有直接的答案。根据文章，在第一、二段找出实体དཏལ་ཆུ་དཔའ་རྒྱ་དོ། （欧曲·达玛巴扎）与མཚན་འགོ་དཀར་སྐྱིད། （夸卓柏吉）的显式关系为母亲，第三段中找出实体དཏལ་ཆུ་དཔའ་རྒྱ་དོ། （欧曲·达玛巴扎）与ལ་ཕུ（拉伍）的显式关系为舅舅，从而推断出དཏལ་ཆུ་དཔའ་རྒྱ་དོ། （欧曲·达玛巴扎）的舅舅ལ་ཕུ （拉伍）是དཏལ་ཆུ་དཔའ་རྒྱ་དོ། (欧曲·达玛巴扎)的母亲མཚན་འགོ་དཀར་སྐྱིད་(夸卓柏吉)的兄弟。

3.2 不可回答问题的生成

通常，机器在做阅读理解时，需要根据文档判断出当前问题是否可答后才进行下一步的答案提取工作。不可回答数据集是实现这一任务的基础。本文提出了两种藏文机器阅读理解不可回答数据集的构建方法。

3.2.1 基于相似度计算的不可回答问题的生成

一词多义问题是机器阅读理解、机器翻译等自然语言处理下游任务中的难点。藏文中，在不同的语境下，ཉི་མ་(太阳)，མེ་དོག་ (花朵)表示人名，ནམ་མཁའི་ཆོར་ཐུ་ (空中珍宝)、པདྨའི་གཉེན་ (莲花之友)表示太阳。这种情况下传统的编辑距离计算不出这些词语的远近关系，但是这也给机器阅读理解不可回答数据集的构建提供了新的思路。

本文计算了TibetanQA的段落相似度，相似度在0.6-0.9之间的段落在书写层面被认为具有一定的相关性，因此，互相替换该段落的问题集，得到基于相似度计算的藏文机器阅读理解不可回答数据集，数据实例如图2所示。

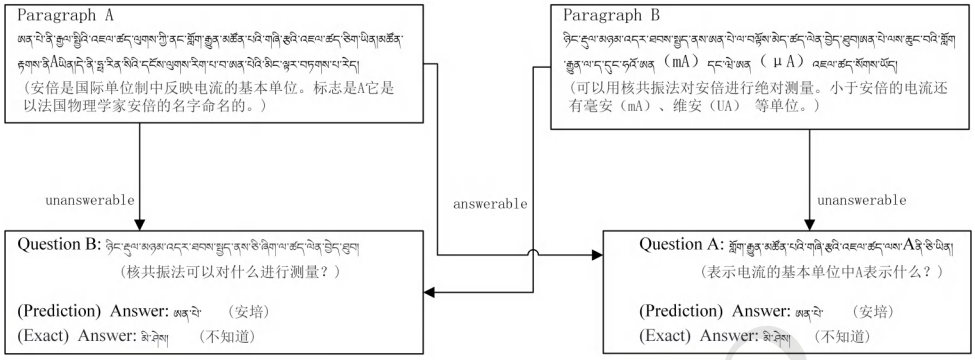


图 2. 基于相似度计算的不可回答问题生成方法

图2中，Question A（B）是Paragraph A（B）的可回答问题，通过计算两个段落的相似度，具有相关性的段落对应的问题集交换并进行人工校对，得到了Paragraph A（B）的不可回答问题Question B（A）。

3.2.2 基于藏文预训练语言模型的不可回答问题的生成

前期，通过众包的形式构建了2,200对面向藏文机器阅读理解不可回答数据集。为了避免下游任务的模型通过简单的启发式搜索或单词匹配的方式在文中找到当前问题的答案，在问题构建过程我们遵循了以下三个原则：（1）对于每篇文章，问题的最佳数量定为1到10；（2）问题集没有相关的答案，但是有看似合理的答案；（3）问题与原上下文内容高度相关。另外，验收时，我们删除了问题数量小于100的创建者提供的问答对，有效避免了没有全面理解此类任务而创建问题的人为噪音。其数据示例如图3所示，其中包含段落、根据该段落提出的不可回答问题和取自该段落的看似合理的答案。

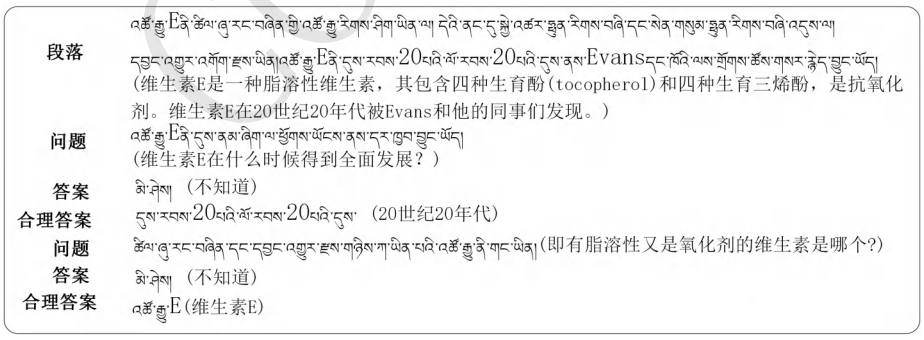


图 3. 通过众包构建的不可回答数据集样例

为了在短周期内获得更具有挑战性的数据集，本文利用藏文预训练模型TiBERT(Liu et al., 2022)对可回答问题进行掩码，替换关键词的方式自动生成不可回答问题，其过程分为确定可回答问题集的关键词、对应文本中的关键词进行掩码和预测，用预测出的关键词替换问题集的关键词生成不可回答问题，数据示例和构建过程如图4所示。

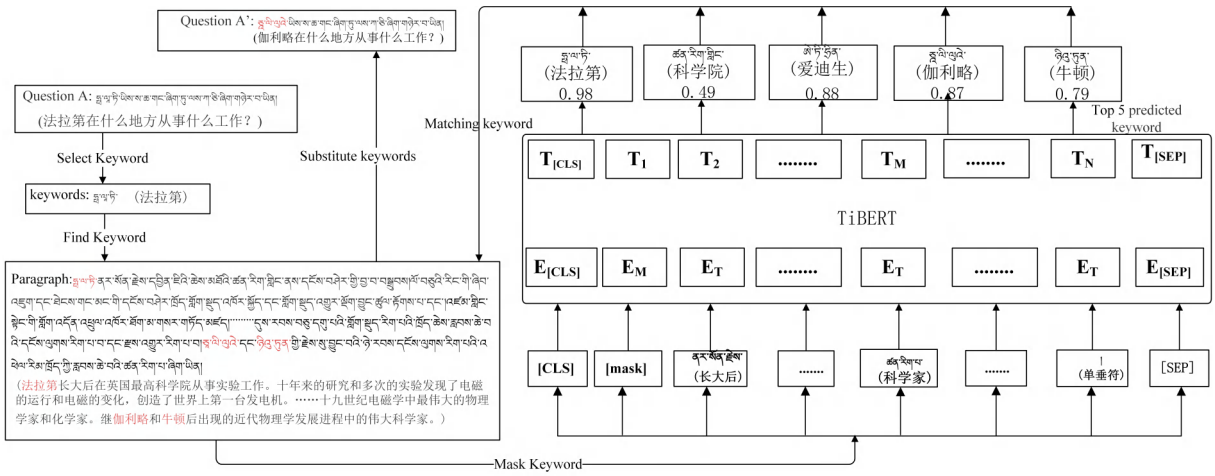


图 4. 基于TiBERT的不可回答问题生成方法

1、确定可回答问题集的关键词

本文使用了TibetanQA中的问题集。注意到问题集中包含许多人名、地名、组织机构名等专有名词，先用sentence piece(Kudo and Richardson, 2018)对语料库进行一体化分词并随机掩码，结果并不理想，主要原因是文本内容涵盖广泛的主题，每个主题的句子结构有比较鲜明的对比。为了正确提取当前问题中的关键词信息，按照文章主题将数据集分为更细粒度的子集，包括人物、时间、科学等类别。本文筛选TibetanQA问题集及相关段落中包含人物名字的数据，提取问题集中的人物名字作为关键词。如图4的Question A:ཧྥ་ལཱ་ཏི་ཡིས་ས་ཆ་གང་ཞིག་ཏུ་ལས་ཀ་ཅི་ཞིག་གཉེར་བ་ཡིན།（法拉第在什么地方从事什么工作？)中将“ཧྥ་ལཱ་ཏི་”(法拉第)标为该问题的关键词信息。

2、关键词掩码，预测并生成不可回答问题

确定了问题集中的关键词，接着使用[mask]对文本中的关键词进行掩码，最后使用藏文预训练语言模型TiBERT将其预测，输出排名前五的预测结果。

为了保证得到的问题集是不可回答且具有一定的难度，一方面，TiBERT预测出的关键词不得等同于掩码之前的关键词以及文中表明的该关键词的同义词，另一方面，TiBERT预测出的关键词需跟当前段落有一定的关联，避免模型根据单词重叠等简单的方式就能判断出问题的可回答性。为此，本文将模型预测出的五个关键词返回到问题及相关段落中进行匹配，过滤掉与关键词相同、不存在于当前段落中的预测值。最后，根据藏文格助词语法的添接规则，使用新预测到的关键词替换可回答问题集中的原关键词，产生新的不可回答问题。

如图4所示，[mask]原文中的“ཧྥ་ལཱ་ཏི་”（法拉第）时按照概率输出的预测结果分别为ཧྥ་ལཱ་ཏི་（法拉第）、ཨེ་ཏི་སིན་（爱迪生）、ཀྲ་ལི་ལུའེ་（伽利略）、ཉིའུ་ཊུན་（牛顿）、ཚན་རིག་གླིང་（科学院）。其中ཧྥ་ལཱ་ཏི་（法拉第）、ཨེ་ཏི་སིན་（爱迪生）是预测值最高的两个输出，但是ཧྥ་ལཱ་ཏི་（法拉第）是原问题中的关键词，而ཨེ་ཏི་སིན་（爱迪生）不存在于当前文本中，用前者替换问题中的关键词没有意义，用后者关键词替换的新问题过于简单，模型根据单词重叠就能得出问题的准确答案。去除两者干扰项，最终得到的预测值为ཀྲ་ལི་ལུའེ་（伽利略），替换可回答问题Question A中的关键词ཧྥ་ལཱ་ཏི་（法拉第），产生新的不可回答问题Question A’:ཀྲ་ལི་ལུའེ་ཡིས་ས་ཆ་གང་ཞིག་ཏུ་ལས་ཀ་ཅི་ཞིག་གཉེར་བ་ཡིན།（伽利略在什么地方从事什么工作？）。

4 实验结果与分析

4.1 实验数据集

本文在TibetanQA和构建的各类数据上进行了实验。将数据按照8:2的比例分为训练集和测试集，分布如表2所示。

TibetanQA(Sun et al., 2021)：采用众包的形式构建的数据集，包含20,000对藏文机器阅读理解可回答问答对和1,513篇文章，文本数据选自云藏百科。

数据集		问答对		
		训练集	测试集	总
推理问题	TibetanQA	16,000	4,000	20,000
	TibetanQA+MultiHop	16,740	4,183	20,923
	TibetanQA+Trip	17,477	4,369	21,846
不可回答问题	TibetanQA+Unanswerable	17,760	4,440	22,200
	TibetanQA+Mask	16,800	4,200	21,000
	TibetanQA+Sim	16,800	4,200	21,000

表 2. 实验所用的数据集

TibetanQA+MultiHop：在TibetanQA中加入了根据三元组多跳的形式构建的知识推理数据集。

TibetanQA+Trip：在TibetanQA加入了三元组及规则生成的简单问答对。

Unanswerable：本文采用众包的形式构建的不可回答数据集，包含2,200对问答对。

TibetanQA+Mask：在TibetanQA加入根据藏文预训练语言模型TiBERT进行掩码，替换关键词的方式生成的藏文机器阅读理解不可回答数据集。

TibetanQA+Sim：在TibetanQA加入了根据可回答数据集相似段落的问题交叉产生的藏文机器阅读理解不可回答数据集。

4.2 实验结果

本文使用(Liu et al., 2022)等人提出的TiBERT在TibetanQA及构建的各类数据集上进行了实验，使用EM和F1值对实验结果进行了评价。

4.2.1 推理问题集的实验结果分析

TibetanQA+Trip、TibetanQA+MultiHop以及TibetanQA数据集在TiBERT上的实验结果如表3所示。

数据集	EM	F1
TibetanQA	53.2	73.4
TibetanQA+MultiHop	47.6	69.9
TibetanQA+Trip	52.9	73.6

表 3. TiBERT在藏文MRC可回答数据集上的实验结果

表3中，TiBERT在TibetanQA数据集上的EM和F1值分别为53.2%和73.4%，根据规则生成的简单问答对上的EM和F1值分别为52.9%，73.6%，其EM值比前者下降了0.3%，而F1值却提高了0.2%，总体对模型的影响较小。其原因如下：该数据集包含的关系较少、生成的问题种类不够丰富、关系词及其同义词具有比较鲜明的特点，如表示母亲的关系词及其同义词“ཨ་མ་”，“ཨ་ཡུམ་”都包含表示女性的词“མ”，这使得模型很容易识别当前问题，从而精准找到文中的答案区间。

采用三元组多跳形式生成的知识推理型数据集在TiBERT上的EM值和F1值分别为47.6%和69.9%，比TibetanQA数据集上的表现下降了5.6%和3.5%。其主要原因是根据多跳三元组产生的知识推理型数据集对模型的理解能力提出了更高的要求，回答此类数据集的问题，模型需要有一定的知识推理能力，无法根据单纯的启发式搜索或者加入以同义词为主的外部知识来获取答案。

4.2.2 不可回答的实验结果分析

本文使用三种方法构建的不可回答数据集TibetanQA+Unanswerable、TibetanQA+Mask、TibetanQA+Sim以及TibetanQA数据集在TiBERT上的实验结果如表4所示。

由表4得知，加入不可回答、知识推理等具有难度的数据集在TiBERT上的结果都呈下降趋势，表明机器阅读理解数据集的不同类型和难度会给模型带来不同程度的影响，有难度的数据集对模型的鲁棒性提出了更高的要求。

数据集	EM	F1
TibetanQA	53.2	73.4
TibetanQA+Unanswerable	50.1	72.6
TibetanQA+Mask	50.1	72.1
TibetanQA+Sim	51.6	73.5

表 4. TiBERT在不可回答数据集上的实验结果

TiBERT在人工构建的不可回答数据集TibetanQA+Unanswerable上的EM值和F1分别为50.1%，72.6%，比TibetanQA上的表现分别下降了3.1%和0.8%，对模型性能产生较大的影响。这也验证了人工方式构建的数据集在答案的不可回答性、答案与文章的相关性和合理答案的选择上具有明显的优势。

根据相似段落的问题交叉产生的不可回答数据集的EM值为51.6%，比TibetanQA数据集下降了1.6%，而其F1值为73.5%，比TibetanQA数据集高出0.1%，对模型的影响较小。表明除了数据量的可控因素之外，不可回答数据集中文章与问题的关联性是影响模型的因素之一。

根据TiBERT进行掩码,替换关键词方法生成的不可回答数据集在TiBERT上的EM值和F1值分别为50.1%，72.1%，比TibetanQA数据集下降了3.1%和1.3%，对模型性能的影响最大，表明将文章根据主题分为更细粒度的子集并进行关键词掩码和替换时产生的效果更好。

另外，本文在每类不可回答数据集中随机选择10%的样本，邀请三组藏族同学根据可读性、关联性和不可回答性三个维度对其进行打分，累计最高为3分，最差为0分。取三种指标的平均值作为最终结果，如表5所示。

可读性：数据集中语法的添接规则、疑问词的使用等书写内容，正确标为1，否则标为0，目的是为了保证数据集书写的规范；关联性：当前问题类型与文本类型是否匹配，即若文本内容是人物介绍类，而问题内容是景物或者其他与人物介绍完全不相关的标为0，否则标为1，其目的是为了避免模型以单词匹配等简单的方式识破不可回答问题；可回答性：当前问题根据给出的文本是否可答，如果不可回答标为1，否则标为0，其目的是检测生成问题的不可回答性。

指标 / 类型	可读性(%)	关联性(%)	不可回答性(%)	平均值(%)
Unanswerable	0.97	0.96	0.85	0.93
Mask	0.75	0.63	0.74	0.71
Sim	0.99	0.44	0.61	0.68

表 5. 不可回答数据集的人工评价结果

表5中，通过众包构建的unanswerable数据集上三种指标的平均值达到了93%，而根据预训练语言模型生成的不可回答数据集Mask和根据相似段落的问题交叉生成的数据集Sim在三种评价指标上的平均值只有71%和68%，比unanswerable数据集分别下降22%和25%。数据表明，机器自动生成的数据集质量还有进一步的发展空间。其主要原因如下：预训练语言模型在生成问题时，将原文中[mask]的人名部分预测成一个代词或者关联性不大的另一个人名，使得生成的问题没有明确的主语。藏文属于黏着语，因此，预测出的新词往往伴随着不同的格助词，这对于根据格助词的添接法则生成问题的规则非常不友好，使得生成的新问题出现语法错误和重复的问题，相似段落交替而生成的问题的可读性指标达到99%，因为相似段落的计算是在TibetanQA的基础上完成，而TibetanQA问题集的语法和疑问词的使用等书写较为规范。但是该类数据集的问题和段落相关性不大，导致其生成的问题与文章内容不相关，机器根据单词重叠的启发式搜索变能分辨当前问题不可答。

5 总结

本文提出了三种面向藏文机器阅读理解的数据增广方法，并且构建了对应的数据集。为了检验数据集的质量，利用藏文预训练语言模型在构建的不同类型的数据集上进行实验，并对藏文机器阅读理解不可回答数据集的可读性、关联性、可回答性进行了人工评价。实验结果表

明，机器阅读理解数据集的质量是影响模型性能的关键因素之一，同一个模型在不同类型数据集的表现大不相同。对于加入不可回答、知识推理等内容的复杂型数据，目前的藏文机器阅读理解模型并不能取得很好的成绩，此类数据对模型的鲁棒性和理解能力提出了更高的要求。在未来的工作中，我们将继续扩充我们的数据集，并针对藏文更具挑战性的机器阅读理解任务，开展进一步的研究和学习。

致谢

本论文得到了国家自然科学基金项目（61972436）和国家社会科学基金项目（22&ZD035）的资助。

参考文献

Kaustubh Dhole and Christopher D Manning. 2020. Syn-qg: Syntactic and shallow semantic rules for question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 752–765.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016*.

Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 325–332.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6602–6609.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961. IEEE.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 196–206.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *In Workshop on Cognitive Computing at NIPS*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3930–3939.

Y Sun, S S Liu, C F Chen, Z C Dan, and X B Zhao. 2021. Construction of high-quality tibetan dataset for machine reading comprehension. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.

Liuyin Wang, Zihan Xu, Zibo Lin, Haitao Zheng, and Ying Shen. 2020. Answer-driven deep question generation based on reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5159–5170.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3901–3910.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 662–671. Springer.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4238–4248.

王丽客, 孙媛, and 刘思思. 2021. 基于多级注意力融合机制的藏文实体关系抽取. 智能科学与技术学报, 3(466-473).

融合预训练模型的端到端语音命名实体识别

兰天伟
北京理工大学，计算机学院
北京，100081
lantianwei0818@qq.com

郭宇航
北京理工大学，计算机学院
北京，100081
guoyuhang@bit.edu.cn

摘要

语音命名实体识别(Speech Named Entity Recognition, SNER)旨在从音频中识别出语音中命名实体的边界、种类和内容，是口语理解中的重要任务之一。直接从语音中识别出命名实体，即端到端方法是SNER目前的主流方法。但是语音命名实体识别的训练语料较少，端到端模型存在以下问题:(1)在跨领域识别的情况下模型的识别效果会有大幅度的下降。(2)模型在识别过程中会因同音词等现象对命名实体漏标、错标，进一步影响命名实体识别的准确性。针对问题(1)，本文提出使用预训练实体识别模型构建语音实体识别的训练语料。针对问题(2)，本文提出采用预训练语言模型对语音命名实体识别的N-BEST列表重打分，利用预训练模型中的外部知识帮助端到端模型挑选出最好的结果。为了验证模型的领域迁移能力，本文标注了少样本口语型数据集MAGICDATA-NER，在此数据上的实验表明，本文提出的方法相对于传统方法在F1值上有43.29%的提高。

关键词： 语音命名实体识别 ； 融合预训练模型 ； 外部知识 ； 跨领域识别 ； 少样本训练

End-to-End Speech Named Entity Recognition with Pretrained Models

Tianwei Lan
Beijing Institute of Technology
School of Computer Science
Beijing 100081, China
lantianwei0818@qq.com

Yuhang Guo
Beijing Institute of Technology
School of Computer Science
Beijing 100081, China
guoyuhang@bit.edu.cn

Abstract

Speech Named Entity Recognition (SNER) aims to recognize the boundary, type and content of named entities in speech from audio, which is one of the important tasks in spoken language understanding. Recognizing named entities directly from speech, that is, the end-to-end method is the current mainstream method of SNER. However, the training corpus for speech named entity recognition is less, and the end-to-end model has the following problems: (1) The recognition effect of the model will be greatly reduced in the case of cross-domain recognition. (2) During the recognition process, the model may miss or mislabel named entities due to phenomena such as homophones, which further affects the accuracy of named entity recognition. Aiming at problem (1), this paper proposes to use a pre-trained entity recognition model to

第二十二届中国计算语言学大会论文集，第174页-第185页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

construct a training corpus for speech entity recognition. For problem (2), this paper proposes to use the pre-trained language model to re-score the N-BEST list of speech named entity recognition, and use the external knowledge in the pre-trained model to help the end-to-end model select the best result. In order to verify the domain migration ability of the model, we labeled the MAGICDATA-NER data set with few samples. The experiment on this data shows that the method proposed in this paper has an improvement of 43.29% in F1 value compared with the traditional method.

Keywords: Speech Named Entity Recognition , Fusion of Pre-trained Models , External Knowledge , Cross-domain Recognition , Few-shot Training

1 引言

命名实体识别(Named Entity Recogination, NER)是自然语言处理中的一项重要任务，传统的命名实体识别旨在将文本中的命名实体信息进行抽取，用特定的类别进行分类，如人名（PER）、机构名（ORG）、地名（LOC）等。任务要求同时确定命名实体的边界以及类别信息。基于文本的命名实体识别研究目前已经有了较为丰富的研究成果(Zhang and Yang, 2018; Huang et al., 2015; Gui et al., 2019a; Gui et al., 2019b)，和比较多的进展(Chiu and Nichols, 2016; Lample et al., 2016; Nadeau and Sekine, 2007)。近年来，语音命名实体识别(Speech Named Entity Recognition, SNER)作为一项口语理解任务得到了越来越多的关注(Caubrière et al., 2020)，在包括屏蔽音频医疗记录中患者姓名(Cohn et al., 2019)等隐私保护领域以及线上直播、视频会议等场景中都有着较大的应用价值。

传统的语音命名实体识别采用级联模型完成(Hatmi et al., 2013)，即首先使用语音识别(Automatic Speech Recogination, ASR)模型对音频识别出对应的文本(Li et al., 2020)，然后使用针对文本的模型在此基础上进行命名实体识别。这类级联模型的识别过程如图1所示。

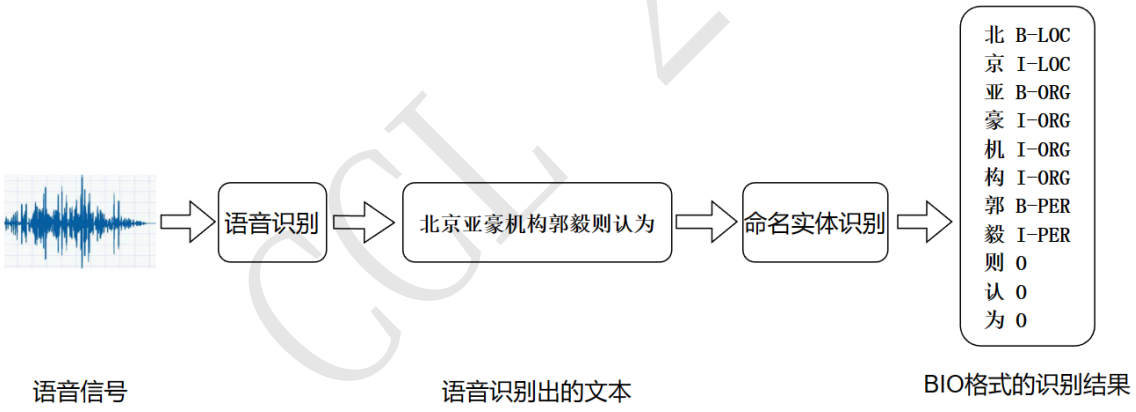


图1.级联模型进行语音命名实体识别的过程

这种方法有一定的缺点。首先，存在误差的级联传播问题(Jannet et al., 2015)。ASR过程中产生的错别字，语义不通顺等问题会增加NER模型识别命名实体的难度，使得这种情况下的准确率相比于在没有错误的文本上进行识别的准确率有所降低。其次，对ASR系统的评价指标一般是字/词错误率，而针对NER系统的评价指标一般是准确率、召回率以及F1值，评价指标的不同使得模型的两个部分无法进行统一的调优，同时也使得模型的训练变得愈发复杂。

已有的一些工作(Ghannay et al., 2018; Yadav et al., 2020; Chen et al., 2022)研究了针对语音命名实体识别的端到端模型,提出了实体感知的语音识别的概念。实体感知的语音识别指的是这些工作普遍将用于识别命名实体的标签直接加入到与音频相对应的文本中来体现对命名实体的标记，例如用方括号[]表示人名，圆括号()表示地点，尖括号⟨⟩表示机构名。在识别模型的训练过程中，用于训练的文本就是已经打好标签的文本，所以模型可以学习音频与打出的标签之间的对齐，直接针对输入的音频输出带有标签的文本，从而完成端到端的语音命名实体识

别。目前这类方法已经被成功应用到了法语、英语以及中文的数据集上(Ghannay et al., 2018; Yadav et al., 2020; Chen et al., 2022)。端到端模型的识别过程如图2所示。

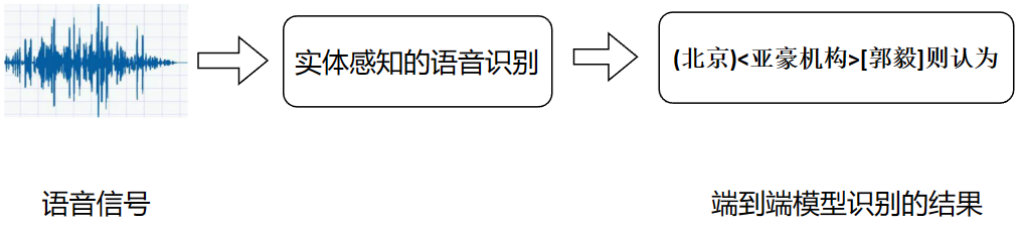


图2.端到端模型进行语音命名实体识别的过程

然而端到端语音命名实体识别仍然面临一些问题。

一方面，语音命名实体识别领域面临标注语料较少的问题。在单一数据集上训练出的语音命名实体识别的模型，往往不能直接用于其他领域数据的识别，在此情况下准确率会有大幅度的下降，模型的跨领域性能较差，而重新标注数据集有着较高的人工成本。在这种情况下，模型的泛化能力是一项重要的评价指标。如果模型在融合实体识别预训练模型中的知识之后或者在一个领域的大规模语料上训练之后，仅在另一领域的少量数据上进行训练就可以产生质量较高的识别结果，从而成为一个通用的语音命名实体识别模型，便可以在很大程度上减少标注成本和训练成本。

另一方面，在端到端场景下，不仅要求正确识别出命名实体的边界以及类别信息，还要正确识别命名实体的内容，所以识别系统输出的错别字以及不准确识别的实体也会导致错误。

表1列举了几种因为错别字以及实体识别原因而导致的错误。在第一个例子中虽然模型同时识别正确了命名实体的边界和类型，但是因为把“荆州”错误的识别成了“金州”，所以依然没有识别正确。第二个例子中错别字直接导致识别出的文本与正确答案有很大的语义差别，没有识别出人名实体。第三个例子中，端到端模型没有识别出和后面的地名实体并列的前一个实体。综合观察这几个错误，本文发现，受限于训练语料规模的不足，使用本地语料训练的端到端模型，在进行实体识别的时候对于一些特定的地理名词以及人名会产生错别字问题，这导致即使在正确识别实体边界和类型的情况下也不能得出正确的答案。

表1.端到端模型的错误识别示例

模型输出结果	正确识别结果
(湖北省)<**金州市安监局**>召开<安良百货>电梯事故情报通报会	<**荆州市安监局**>
出清就是<中国女排>的核心	[**朱婷**]
(崇礼县)发展较成熟的**万龙滑雪场**和(云顶滑雪场)	(**万龙滑雪场**)

(Chen et al., 2022)提出将预训练模型融入到级联模型的识别过程中并取得了结果的提高，但是目前融合了预训练模型的端到端系统并不常见。本文提出一种将预训练模型融入端到端语音命名实体识别模型来提高识别效果的方法。

针对语音命名实体识别训练语料较少、跨领域识别困难的问题，本文提出使用命名实体识别预训练模型帮助构建语音实体识别的训练语料，利用较易获得的语音识别平行语料解决了模型的跨领域问题同时保持了端到端模型的简洁性。针对同音词等现象导致的对命名实体的漏标、错标问题，采用预训练语言模型对语音命名实体识别的N-BEST列表进行重打分，利用预训练模型中的外部知识帮助端到端模型挑选出最好的结果。

为了验证模型的跨领域识别能力，本文标注了少样本数据集MAGICDATA-NER并设计了零样本实验来验证预训练模型对整体模型的泛化能力的提升效果。在AISHELL-NER为训练语料，MAGICDATA-NER为测试语料的场景下，相比于传统方法，本方法的F1值可以提升43.29%。

在AISHELL-NER上的中文实验结果表明，当融合了预训练语言模型BERT和GPT2重排序之后，F1值从74.31%分别提升至了77.27%和79.26%。在DATA2上的英文实验中，F1值从87.0%分别提升至了88.9%和88.7%。

以上实验均证明了本文所提方法的有效性。

本文的贡献总结如下：

(1)用预训练命名实体识别模型构建语音命名实体识别训练语料，解决了语音命名实体识别训练语料较少的问题，利用较易获得的语音识别平行语料解决了模型的跨领域问题。

(2)通过重打分的方式将预训练语言模型融合到语音命名实体识别模型当中，在不同语种的多个语料中取得了比之前更好的结果，并通过各项实验结合实例分析了取得进步的原因。

(3)提出了少样本数据集MAGICDATA-NER，证明了结合预训练模型能够提升模型整体的泛化性能，为今后降低语音命名实体识别模型训练的人工标注成本以及计算成本提供了参考。

2 融合外部知识的语音命名实体识别

2.1 融合声学模型和预训练实体识别模型

在单一数据集上训练出的语音命名实体识别的模型，往往不能直接用于其他领域数据的识别，在此情况下准确率会有大幅度的下降，模型的跨领域性能较差。因此本文提出使用预训练的命名实体识别模型来对较为容易获得的语音识别平行语料的转录文本部分进行实体标记，使之成为可以用于训练语音命名实体识别的伪语料，最终用这份语料来训练语音命名实体识别模型。

在实际训练阶段，先用经过人工标记的语料训练模型，在将模型应用于某个具体的领域时，使用以上方法利用该领域内的语音识别平行语料将预训练实体识别模型中的知识蒸馏到语音命名实体识别模型中，完成领域迁移的工作。如此训练出的模型在应用阶段可以在不借助语言模型的情况下取得较好的识别效果，保持了端到端模型的简洁性同时完成了领域迁移，相关实验结果在将第四部分展示。

2.2 融合声学模型和预训练语言模型

本文采用的方法如图3和图4所示。在实体识别阶段，本文依然采用端到端方法中实体感知语音识别的思想，在输出时直接在命名实体的左右打好标签。所不同的是，这一次模型输出的不是贪婪搜索得到的最好的结果，而是采用束搜索算法得到的N-BEST列表，这个列表是实体识别模型产生的一个候选者列表，通过设定束搜索宽度，可以控制得到的列表长度，即候选句子的个数，同时也可以得到相应的声学模型的打分S_{AM}。接着，本文将获取的预训练语言模型在带括号的文本语料上进行微调，以便让模型学习打标签的相关知识。最后用包含外部知识的语言模型对N-BEST列表中的句子进行重打分，得到语言模型的分数S_{LM}，将两个得分相结合，最终挑选出得分最高的结果输出。通过以上过程，既避免了传统级联模型中误差累积的问题，也改善了端到端模型中因训练语料不足以及无法融合预训练模型而导致的错别字问题。在接下来的部分中，本文详细介绍了模型各阶段所使用的方法。

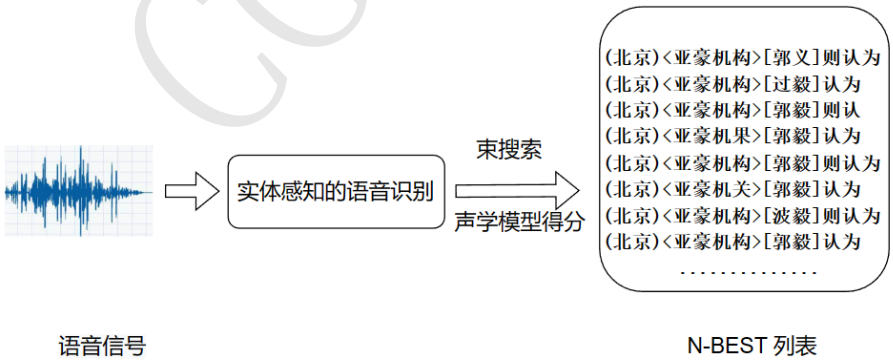


图3.使用束搜索算法得到N-BEST列表

2.3 实体感知的语音识别过程

在语音识别的过程中，本篇文章使用端到端语音命名实体识别所采用的方法，用添加过标签的文本数据对实体识别模型进行训练，另外向词表中添加相应的标签，来让实体识别模型在解码时能够直接得出已经打好标签的文本结果。

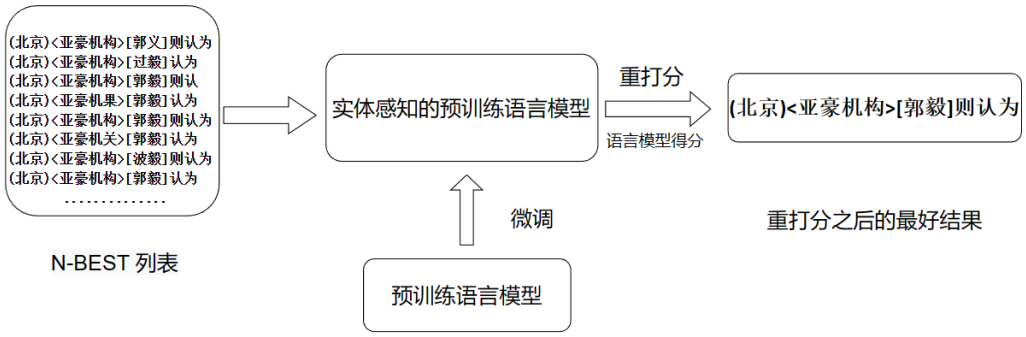


图4.使用预训练模型重打分

同时，为了用预训练的语言模型给语音识别模块输出的N-BEST列表进行重打分，以融合外部知识，语音识别的过程共分为两步。第一步，经过特征提取的音频信息输入到语音识别模型进行编码，每段经过采样的音频会产生一个长度为词表大小的向量表示，这些向量拼接在一起形成了将要进行束搜索的矩阵。可以将这个矩阵看作一个全连接的Lattice图，通过束搜索算法计算两个元素之间的转移概率，就可以按照概率大小解码出进行重打分需要的N-BEST列表。第二步，应用束搜索算法在该矩阵上进行搜索，通过设置束搜索的宽度可以确定最终获得的N-BEST列表的长度，即在重打分的过程中候选者的数目。同时在进行束搜索时，由转移概率可以计算出相应的得分，这个得分就是声学模型部分的打分S_{AM}。

2.4 预训练语言模型重打分

2.4.1 GPT2

本篇文章所使用的预训练模型GPT2由多层Transformer decoder堆叠而成，是一个单项模型。采用GPT2对语音识别过程所获得的N-BEST列表中的每个句子进行打分，以获得相应的语言模型分数S_{LM}。打分时GPT2基于给定的单词来计算下一个单词出现的对数似然概率，将一个句子中每个单词的对数似然概率相加，就得到了一个给定句子的得分，如公式(1)所示。其中W表示给定的句子，W_t表示句子中的第t个字，$W_{<t}$表示句子的前t-1个字，Θ表示用来打分的语言模型的参数。

$$Score_{LM}(\mathbf{W}) = \sum_{t=1}^{|W|} \log P_{LM}(\mathbf{w}_t|\mathbf{W}_{<t};\Theta) \tag{1}$$

2.4.2 BERT

本篇文章所使用的预训练模型BERT由多层Transformer encoder堆叠而成，与GPT2不同的是，BERT是一个双向模型。在给句子进行打分的时候，BERT会同时考虑到一个字左边和右边的字，结合之前和之后的情况进行打分，如公式(2)所示。其中W表示给定的句子，W_t表示句子中的第t个字，$W_{\backslash t}$表示句子中除第t个字以外的其他字，Θ表示用来打分的语言模型的参数。

$$Score_{\mathrm{LM}}(\mathbf{W}) = \sum_{t=1}^{|\mathbf{W}|} \log P_{\mathrm{LM}}(\mathbf{w}_t|\mathbf{W}_{\backslash t};\Theta) \tag{2}$$

2.4.3 重打分

公式(1)和(2)的计算方式会使得长度较长的句子获得较低的得分，因为每一个对数似然概率都是一个负数，句子长度较长意味着有更多的负数相加并获得一个较低的得分。而在语音识别的过程中有可能更长的句子中包含了更多的正确信息，所以在实际应用过程中，本文首先获取语言模型对整句的打分(Salazar et al., 2019)，然后用整句的分数除以每个句子的长度以得到每个单词的平均对数似然概率，来当作语言模型部分句子的得分，如公式(3)所示。其中T表示给定的句子的长度。

$$S_{LM}(\mathbf{W}) = \frac{1}{T} Score_{LM}(\mathbf{W}) \tag{3}$$

同时，由于预训练模型本身的训练文本中不包含用括号标记处的命名实体，未经过微调的预训练模型在进行打分的时候可能会在计算括号标签的转移概率的时候输出一个过低的数值。所以为了取得更好的打分效果，还需要在域内数据上对预训练模型进行微调。

2.5 融合声学得分和语言模型得分

N-BEST列表中每一个句子最终得分的计算方法如公式(4)所示。

$$\mathrm{Score}(\mathbf{W}) = (1-\lambda) \cdot \mathrm{S_{AM}}(\mathbf{W}) + \lambda \cdot \mathrm{S_{LM}}(\mathbf{W}) \tag{4}$$

采用线性插值的方法将两个部分的得分融合在一起，λ 是本文定义的一个超参数，用来调节两个模型之间的权重比例。最终本文挑选N-BEST列表中得分最高的句子作为整个语音命名实体识别系统的输出。

如此，因为在解码阶段同时考虑了声学模型部分和语言模型部分的打分，本文在原有模型的基础上融合了预训练模型中的外部知识，这些外部知识可以帮助模型在解码时减少因为错别字而造成的错误，用以获得更加合理的输出结果，这样的效果在中文命名实体识别中应当较为明显，因为中文中有很多同音字。本文预计，通过结合预训练模型，可以使输出句子中的错别字更少，更加通畅，进而提高结果的精确率（Precision）。而通过对预训练模型进行微调，可以使结果中做出预测的数量变得更多，进而提高结果的召回率（Recall），在第四部分的实验中，本文的猜想也得到了进一步的证实。

3 MAGICDATA-NER数据集

为了检验本文的方法在零样本以及少样本情况下的表现，即预训练语言模型对模型整体的跨领域识别能力的提升效果，本文标注了数据集MAGICDATA-NER。MAGICDATA-NER基于开源的语音识别数据集MAGIADATA，该数据集包含755个小时的语音数据，由1080名来自中国大陆的说话人用普通话朗读。与AISHELL-NER中包含财经、科技、体育、娱乐和新闻的较为正式的语音数据不同，该数据集的语料取自日常生活对话。本文在MAGICDATA中抽取部分语音以及对应的文本用以标注本文的跨领域少样本数据集MAGICDATA-NER。MAGICDATA-NER的语音以及原始的转录文本部分抽取自MAGICDATA的验证集，共包含11793条，约14个小时的语音数据，这部分数据将会在论文录用后公开。通过比较在零样本以及少样本训练情况下模型融合大规模预训练语言模型前后的命名实体识别效果，来检验预训练语言模型对整体的跨领域识别能力是否有提升效果。

3.1 标注过程

在标注过程中本文采用了和以往工作(Chen et al., 2022)类似的标注方法，标注了三个种类的命名实体，即人名(PER)、地名(LOC)、机构名(ORG)。首先本文获取了一个用于在中文文本上进行命名实体识别的预训练模型，该模型取自Hugging Face(Wolf et al., 2020)。然后把该模型在MSRA数据集上进行微调，并在测试集上取得了91.59%的F1值。最后本文用微调过后的模型标注从MAGICDATA中抽取的转录文本，以此作为人工标注的起点。

在人工标注阶段，将数据集中的语句分为2000条一组，针对每一组数据，两个学习了命名实体标注规则的研究人员会分别对机器标注的数据进行检查并对机器标注错误的实体进行修改以及标记。之后对比两组经过人工查验的数据，当两组数据的一致性达到95%以上时才接受其为语料数据，并对其中不一致的地方进行进一步探讨。当一致性低于95%时，更换研究人员进行重新标记。最终形成本文标注的跨领域少样本数据集MAGICDATA-NER。

3.2 语料数据

表2中展示了MAGICDATA-NER中各个集合中命名实体的情况。在将数据集分割成训练集、开发集、测试集时，为在进行验证的时候避免因集合过小而造成的偶然性因素，本文将验证集和测试集的大小控制在2000条语料以上。同时作为少样本语料，本数据集的训练集相对较小，用以检验模型在少样本以及零样本情况下的命名实体识别能力。各集合的实体类型分布基本一致，保证了数据分布的一致性。

表2.MAGICDATA-NER各个集合中命名实体的情况

数据集	句子数量	人名实体数量	地名实体数量	机构实体数量	实体总数量
MAGICDATA-NER	11793	2036	770	242	3048
训练集	6538	1135	391	146	1672
验证集	2310	365	155	55	575
测试集	2945	536	224	41	801

4 实验

本文分别进行了在AISHELL-NER上的中文实验，在DATA2上的英文实验，和在MAGICDATA-NER上的跨领域实验，以下是对实验过程以及结果的详细说明。

4.1 模型架构

中文实验的实体识别模块使用的是ESPNET中的Conformer模型，模型的基本设置与原文(Chen et al., 2022)相同,在AISHELL-NER数据集上进行了实验。在MAGICDATA-NER数据集上的跨领域以及少样本实验同样使用此模型，只在训练过程中需要在相应的数据集上进行后续的训练和验证。为了验证本篇文章所提出的方法对于不同数据集以及不同语种的有效性，本文用这篇文章(Yadav et al., 2020)使用的DeepSpeech2框架，在其提出的英文数据集DATA2上同样做出了验证实验。

实验中所用到的预训练模型均获取自Huggingface Transformers(Wolf et al., 2020)。中文实验使用的GPT2语言模型是uer/gpt2-chinese-cluecorpussmall(Radford et al., 2019; Zhao et al., 2019)，使用的BERT语言模型是bert-base-chinese(Devlin et al., 2018)。少样本实验使用的预训练模型与中文实验相同，只在微调语料上不同。英文实验使用的GPT2语言模型是gpt2(Radford et al., 2019),使用的BERT语言模型是bert-base-uncased(Devlin et al., 2018)。各模型均在对应的训练语料上进行5个epoch的微调，然后用于重打分。预训练实体识别模型使用的是ckiplab/bert-base-chinese-ner(Ckiplab, 2020)。

4.2 数据集

中文实验使用的AISHELL-NER(Chen et al., 2022)数据集包含超过170小时的普通话语音数据，语料库涵盖五个领域:“财经”、“科技”、“体育”、“娱乐”和“新闻”。英文实验使用的DATA2(Yadav et al., 2020)数据集包含约150个小时的英文语音数据，共39769条语料。这些语料是在英语数据集Librispeech(Panayotov et al., 2015)、CommonVoice(WikipediaContributors, 2020)、Tedlium(Rousseau et al., 2012)和Voxforge(WikipediaContributors, 2019)的子集基础上进行标注的。少样本、零样本以及跨领域实验使用的是本篇文章所提出的数据集MAGICDATA-NER，共包含11793条，约14个小时的语音数据，内容为日常生活对话。

4.3 实验结果以及分析

4.3.1 预训练语言模型对于识别结果的影响

本实验采用conformer结构的中文语音命名实体识别作为基线模型，并对比了使用不同语言模型进行重打分对于识别结果的影响。其中transformer模型从头在域内数据上进行了15个epoch的训练，而GPT2和BERT则在域内数据上进行了5个epoch的微调。在进行实验时，为了确定语言模型的最佳权重λ，首先在验证集上进行实验，观察在不同权重条件下验证集上的F1值，如图5所示。

在获得结果后将在验证集上取得最好结果的权重值应用于测试集以获得最终的实验结果，如表3所示。Conformer+BERT-PT表示在先前工作(Chen et al., 2022)中级联模型取得的最好结果。

根据实验结果本文可以看出，通过结合Transformer模型，命名实体识别的F1值增长至75.32%，使用BERT让结果进一步提升至了77.27%，而结合微调过后的GPT2模型之后实验结果有了大幅度的增长来到了79.26%，并且在精确率以及召回率上都有较大幅度的提升。由此本文可以得出，结合语言模型对N-BEST列表进行重打分的策略对提升命名实体识别的准确

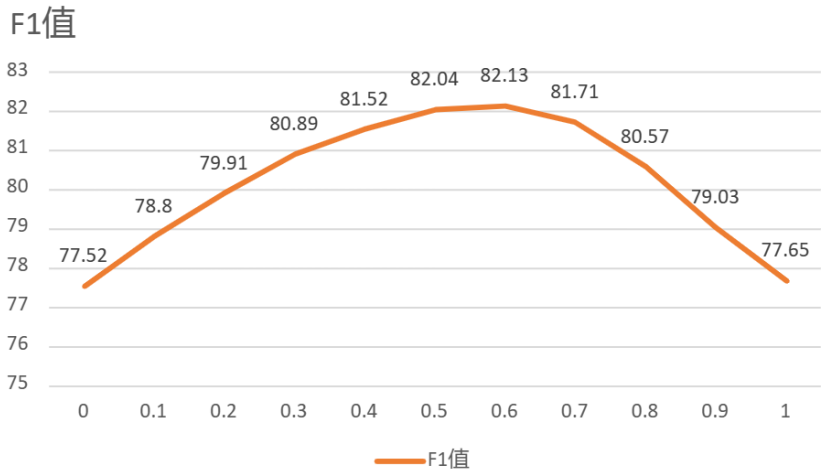


图5.验证集的F1值随λ的变化情况(以Conformer + GPT2_lm为例)

表3.增加预训练语言模型对于识别结果的影响

模型	精确率	召回率	F1值
Conformer(baseline)	77.56	71.32	74.31
Conformer+BERT-PT(Chen et al., 2022)	75.54	74.27	74.90
Conformer + Transformer_lm	78.37	72.50	75.32
Conformer + BERT_lm	80.21	74.54	77.27
Conformer + GPT2_lm	**82.55**	**76.13**	**79.26**

性有正向效果，而不同语言模型对于F1值的提升效果有所不同。Transformer模型只在域内的数据上进行了训练，碍于数据规模的大小，模型的能力要落后于先预训练后进行微调的预训练模型，这体现出了在预训练阶段模型中融合的大规模外部知识对于整个系统识别的指导作用，也验证了本篇文章所提出的方法的有效性。

4.3.2 选取不同候选列表长度对于识别结果的影响

在使用束搜索算法以及语言模型进行解码的过程中，可以通过设定束搜索宽度来决定候选名单的长度。候选的句子是根据声学模型打分从高到低来排列的，设定的候选列表长度越大，进行重打分的时候可以参考的句子数量就越多，下面的实验验证了在融合GPT2模型的情况下候选列表长度与最终结果之间的关系。实验结果如表4所示。

表4.选取不同候选列表长度对于识别结果的影响

候选列表长度	精确率	召回率	F1值
1	77.56	71.32	74.31
5	80.57	74.54	77.43
10	81.35	75.48	78.31
20	81.96	75.89	78.79
30	82.45	75.89	79.04
40	82.64	76.04	79.20
50	**82.55**	**76.13**	**79.26**

从实验结果可以看出，当候选列表长度越大，也就是候选的句子数目越多的时候，结果的得分就会越高，这是因为考虑的范围越广泛，挑选出最佳句子的机率就越大。但同时也要注意到的是，由于候选的句子是根据声学模型打分从高到低来排列的，越往后的句子的参考价值也会越小。从实验结果可以看出，增加候选句子的数量对于实验结果的增益效果是越来越小的，并在候选列表长度为50的时候基本收敛，此时模型的识别效果达到最优。

4.3.3 识别结果实例分析

表5.识别结果实例分析

模型	识别结果
Conformer	(湖北省)<**金州市安监局**>召开<安良百货>电梯事故情报通报会 **出清**就是<中国女排>的核心 (崇礼县)发展较成熟的**万龙滑雪场**和(云顶滑雪场)
Conformer + GPT2_lm	(湖北省)<**荆州市安监局**>召开<安良百货>电梯事故情报通报会 [**朱婷**]就是<中国女排>的核心 (崇礼县)发展较成熟的**(万龙滑雪场)**和(云顶滑雪场)

在本部分，本文挑选了几个比较典型的例子来说明融合GPT2语言模型之后，外部知识对于命名实体识别的提升作用，如表5所示。第一个例子体现的是预训练模型中的外部知识对于同音字、近音字的区别作用。因为“金”和“荆”的发音相近，所以语音识别的过程中可能会因错别字而导致实体识别错误，而预训练模型因为有地名相关知识，即荆州才是湖北省下辖的一个市从而做出正确的选择。第二个例子中，因为在语音识别的结果中有噪声所以基线模型未能正确识别出人名“朱婷”。而“朱婷”是一个较为知名的公众人物的名字，预训练模型的大规模训练语料中会包括相关内容，所以在融合预训练模型之后模型能够识别出“朱婷”是一个人名。人名、地名的相关知识在语音识别中是一个比较难以解决的问题，单纯依靠训练本地模型因为受到数据规模大小的限制，难以完善的结合相关知识，而前两个例子证明通过结合预训练模型可以较好地改善这一问题。第三个例子中，基线模型漏掉的地名实体“（万龙滑雪场）”可以在融合GPT2之后正确的识别出，由此可以看出预训练模型提高了系统对于上下文的感知能力，意识到了万龙滑雪场是和云顶滑雪场并列的一个地名实体。

4.3.4 英文数据集实验

为了验证本篇文章所提出的方法对于不同数据集以及不同语种的有效性，使用DeepSpeech2框架，本文在英文数据集DATA2(Yadav et al., 2020)上同样做出了验证实验，实验结果如表6所示。其中DeepSpeech2 + NER tagger表示先使用DeepSpeech2进行语音识别，再进行命名实体识别的级联模型的结果，而其余三项实验中使用DeepSpeech2做端到端的语音命名实体识别。

表6.英文数据集DATA2实验结果

模型	精确率	召回率	F1值
DeepSpeech2+NER tagger(Yadav et al., 2020)	80.0	59.0	63.0
DeepSpeech2(Yadav et al., 2020)	96.0	80.0	87.0
DeepSpeech2 + GPT2_lm	**99.5**	80.0	88.7
DeepSpeech2 + BERT_lm	99.1	**80.6**	**88.9**

从实验结果可以看出，通过融合英文预训练模型，识别的准确性也获得了提高，这印证了在不同语种的不同数据集上，本文的方法都对端到端语音命名实体识别有提升效果。

4.3.5 零样本及少样本实验

在零样本实验中，模型在AISHELL-NER数据集上进行训练然后直接用于识别少样本数据集MAGICDATA-NER测试集中的语音。在少样本实验中，模型首先在AISHELL-NER数据集上进行训练而后在MAGICDATA-NER的训练集上进行小规模后续训练，最后用于识别少样本数据集MAGICDATA-NER测试集中的语音。应用于两个实验的语言模型均在MAGICDATA-NER训练集的文本上进行了微调。以下为实验结果。

从实验结果可以看出，在两种情况下融合预训练模型后识别结果均有明显提升。零样本实验中，识别结果的F1值从24.79%提升至了29.96%，少样本实验中，识别结果的F1值从51.34%提升至了60.98%。这说明在缺少训练样本乃至没有对应领域训练样本的情况下，预训练模型很好的帮助了实体识别过程，提升了模型的泛化能力。相比于中英文实验，预训练模型在零样本以及少样本实验中对于实验结果的提升更为明显，这体现了预训练模型在训练样本较

表7.零样本实验结果

模型	精确率	召回率	F1值
Conformer(baseline)	36.67	18.73	24.79
Conformer + GPT2_lm	45.06	22.22	29.77
Conformer + BERT_lm	**43.96**	**22.72**	**29.96**

表8.少样本实验结果

模型	精确率	召回率	F1值
Conformer(baseline)	53.84	49.06	51.34
Conformer + GPT2_lm	62.25	53.93	57.79
Conformer + BERT_lm	**62.75**	**59.30**	**60.98**

少情景中对于提升模型表现的显著作用。这一结果进一步说明使用本文提出的方法融合预训练模型，通过先大规模语料训练后少样本语料微调的方法，可以大幅提升模型在新领域的识别表现，进而帮助减少标注成本和训练成本，降低语音命名实体识别模型在实际应用中的开发难度。

4.3.6 融合预训练实体识别模型的跨领域实验

在本实验中，模型首先在AISHELL-NER上训练至收敛，然后对MAGICDATA的测试集进行识别，取得未经过跨领域训练的实验结果。而后使用经过预训练实体识别模型标记的MAGICDATA的训练集进行下一步训练，同样对MAGICDATA的测试集进行识别，取得跨领域训练后的实验结果，用以检验预训练模型对跨领域的实体识别是否有帮助，实验结果如下。其中Conformer(baseline)表示只在AISHELL-NER上训练未融合预训练实体识别的模型，GPT2_LM和BERT_LM表示使用了语言模型重打分的方法，fusion表示使用了融合预训练实体识别模型的方法。

表9.融合预训练实体识别模型的跨领域实验

模型	精确率	召回率	F1值
Conformer(baseline)	41.74	22.80	29.49
Conformer+GPT2_LM	49.97	27.03	35.09
Conformer+BERT_LM	48.85	27.79	35.42
Conformer+fusion	75.91	62.76	68.71
Conformer+fusion+GPT2_LM	79.12	64.63	71.14
Conformer+fusion+BERT_LM	**78.84**	**67.59**	**72.78**

本文采用消融实验的方式验证了语言模型重打分和构造训练语料两种方法对于提升识别效果的作用，从实验结果可以看出，通过重打分，F1值从29.49%提升至了35.42%，通过融合预训练实体识别模型，在跨领域的情况下语音实体识别的准确率有了大幅的提升，F1值提升至了68.71%，并基本接近了在AISHELL-NER上使用人工标注的数据进行训练的识别准确率。在重打分的同时结合预训练语言模型的情况下，模型识别的F1值还可以进一步提升至72.78%。这说明在跨领域的情况下，两种方法对于提升识别效果都有正向作用，为尽可能提高识别的准确率可以使模型同时结合预训练语言模型和预训练实体识别模型，以达到最好的识别效果。

5 结论

针对端到端语音命名实体识别模型训练语料较少，在跨领域识别情况下效果大幅下降以及识别过程中的错标、漏标问题，本篇文章提出了通过融合预训练实体识别模型构建训练语料并用预训练语言模型重打分的方法，提升了端到端模型的识别效果，弥补了以往端到端方法未融合预训练模型的不足。为了验证模型的跨领域识别能力，本文标注了少样本数据集MAGICDATA-NER并设计实验来验证预训练模型对整体模型的泛化能力的提升效果。实验结果表明，使用本文的方法融合预训练模型，在跨领域的情况下本文的方法对提升命名实体识别效果有显著作用，同时在中英文数据集上语音命名实体识别的F1值都获得了提升。本篇文章针对实验结果进行了分析，并且验证了所提出方法在不同语种、不同数据集上的有效性。本文提出的方法为今后语音命名实体识别模型在跨领域情况下的训练以及在训练过程中降低人工标注以及计算成本提供了参考，同时新的数据集也可以为其他口语理解以及语音识别任务的跨领域实验提供帮助。在未来的工作中，可以通过在更多的数据集上进行实验以验证本方法对于效果的提升作用，同时也可以探索在融合预训练模型之后提高模型计算效率的方法。

致谢

本研究受科技创新2030-“新一代人工智能”重大项目(2020AAA0106600)资助。

参考文献

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4514–4520, Marseille, France, May. European Language Resources Association.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. Aishellner: Named entity recognition from chinese speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8352–8356.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Ckiplab. 2020. Ckip bert base chinese. Website. https://github.com/ckiplab/ckip-transformers.

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio de-identification: A new entity recognition task. *arXiv preprint arXiv:1903.07037*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent, and E. Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4982–4988. International Joint Conferences on Artificial Intelligence Organization, 7.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuanjing Huang. 2019b. A lexicon-based graph neural network for Chinese NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1040–1050, Hong Kong, China, November. Association for Computational Linguistics.

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, and Sylvain Meignier. 2013. Named entity recognition in speech transcripts following an extended taxonomy. In *SLAM@INTERSPEECH*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv e-prints*, page arXiv:1508.01991, August.

Mohamed Ameur Ben Jannet, Olivier Galibert, Martine Adda-Decker, and Sophie Rosset. 2015. How to evaluate asr output for named entity recognition? In *Interspeech*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online, July. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

WikipediaContributors. 2019. Voxforge wikipedia, the free encyclo- pedia. Website. https://en.wikipedia.org/w/index.php?title=VoxForge&oldid=913093799.

WikipediaContributors. 2020. Common voice wikipedia, the free encyclopedia. Website. https://en.wikipedia.org/w/index.php?title=CommonVoice&oldid=939008593.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *arXiv preprint arXiv:2005.11184*.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia, July. Association for Computational Linguistics.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

基于词向量的自适应领域术语抽取方法

唐溪[1]，**蒋东辰**[1*]，**蒋翱远**[2]
1.北京林业大学信息学院，北京，100083
2.中原银行股份有限公司，郑州，450046
{tangxi,jiangdongchen}@bjfu.edu.cn jiangaoyuan@zybank.com.cn

摘要

术语分布呈现长尾特性。为了有效提取低频术语，本文提出了一种基于词向量的自适应术语抽取方法。该方法使用基于假设检验的统计方法，自适应地确定筛选阈值，通过逐步合并文本的强关联性字符串获得候选术语，避免了因固定阈值导致的低频术语遗漏问题；其后，本文基于掩码语言模型获得未登录候选术语的词向量，并通过融合词典知识的密度聚类算法获得候选术语归属的领域簇，将归属于目标领域簇的候选术语认定为领域术语。实验结果表明，我们的方法不仅在F值上优于对比方法，而且在不同体裁的文本中表现更为稳定。该方法能够全面有效地抽取出低频术语，实现领域术语的高质量提取。

关键词： 术语抽取；自适应；假设检验；词向量

An Adaptive Domain-Specific Terminology Extraction Approach Based on Word Embedding

Xi Tang[1],**Dongchen Jiang**[1*],**Aoyuan Jiang**[2]
1. School of Information Science and Technology,
Beijing Forestry University, Beijing, 100083, China
2.Zhong Yuan Bank Co.,Ltd, Zhengzhou, 450046, China
{tangxi,jiangdongchen}@bjfu.edu.cn jiangaoyuan@zybank.com.cn

Abstract

Terminology distribution shows a long-tail pattern. This study presents an adaptive term extraction method based on word embedding to effectively extract low-frequency terms. Using a hypothesis-testing statistical approach, the method adaptively sets filtering thresholds and acquires candidate terms by incrementally merging strongly related text strings, avoiding omission of low-frequency terms due to fixed thresholds. Word embeddings for out-of-vocabulary candidates are obtained through a masked language model, and a dictionary-integrated density clustering algorithm identifies domain clusters for these terms. Candidates within target domain clusters are recognized as domain-specific terms. Experimentally, our method outperforms competitors in F-score and maintains stability across diverse text genres. This approach effectively extracts low-frequency terms, ensuring high-quality domain-specific term extraction.

Keywords: Terminology extraction , Self-adaptation , Hypothesis testing , Word embedding

* 通讯作者

第二十二届中国计算语言学大会论文集，第186页-第195页，哈尔滨，中国，2023年8月3日至5日。

1 介绍

术语是特定领域内用以传达专业概念的约定俗成的语言符号。术语抽取作为自然语言处理领域的一项核心技术，能够从大规模语料库中自动提取领域术语，从而降低手工发现术语的人力成本。此外，术语抽取可应用于诸如文本分类、信息抽取、机器翻译等下游任务中，为其提供更为丰富的领域概念知识。例如，在机器翻译任务中，将领域术语作为先验知识引入，能够有效提升翻译质量(Michon et al., 2020; 游新冬et al., 2021)。因此，术语抽取在自然语言处理领域具有重要的应用价值。

研究发现，领域术语的分布遵循长尾分布(Williams et al., 2015)，低频术语在全部领域术语中所占比例极高。这意味着如果不能有效实现低频术语抽取，将导致大量术语无法被识别。在实际应用场景中，新兴术语和非规范术语作为术语标准化工作的重点(刘书剑and 彭道黎, 2011)，它们大都以低频形式出现。准确识别低频术语是及时发现新兴术语和规范术语使用的基础，能够有效辅助术语标准化工作。因此，关注低频术语识别具有重要意义，低频术语抽取质量直接影响术语抽取的全面性。

现有研究在抽取领域术语方面已取得一定成果，然而针对低频术语的抽取仍存在较大挑战。低频术语在语料库中出现次数较少，使得诸如基于统计、机器学习或深度学习的方法在处理这类术语时很难获得理想的效果。针对低频术语抽取困难这一问题，部分学者关注低频术语的构词规律，通过语法规则特征匹配来抽取低频术语(俞琰and 赵乃, 2018; 李思良et al., 2018)。尽管这类方法在低频术语抽取方面具有一定的有效性，但方法所依赖的构词规律却呈现出显著的领域依赖性，使得领域迁移性相对较弱。

为了实现各领域低频术语的有效识别，本文提出一种基于词向量的自适应领域术语抽取方法。该方法首先采用假设检验的统计方法自适应地获取目标文本中候选术语的频率阈值，对低频候选术语筛选具有显著效果。在术语确认阶段，本文采用词向量技术，通过融合词典知识的密度聚类评估候选术语的领域相关性，并将具有领域语义的候选术语确定为最终领域术语。本方法能够有效抽取低频术语，实现准确度较优的术语抽取效果。

本文的组织结构如下：第二节介绍了研究的相关工作。第三节详细阐述了我们的术语抽取框架的细节。第四节展示了我们的实验设置和结果。第五节给出了研究结论。

2 相关工作

目前，研究者主要采用语法规则、统计计算、传统机器学习以及深度学习四类方法抽取术语。基于语言规则的方法从术语的构词规律出发，通过语法模板匹配得到术语。对于英语等拉丁语系语言，词性分析准确率相对较高，因此可通过匹配词性序列来抽取文本中的术语，甚至是术语变体(Kafando et al., 2021)。但中文的情况更为复杂，在词性分析前，通常需要先对语料分词，这一额外步骤可能会带来累积误差，从而增加了词性分析的难度。因此，基于语言规则的中文术语抽取方法更多使用领域常用词根和词缀辅助分析(孙水华et al., 2016; 李思良et al., 2018)。这种方法简单且有效，但其适应不于同领域和文本风格的能力有限；同时，规则匹配的形式较为僵化，这也限制了识别新兴术语的能力。

术语识别是一项普遍应用于各领域的任务，而统计方法为此提供了一种不针对特定领域的通用解决方案。基于统计的术语识别方法通常包括两个步骤。首先，从文本中找出能够表达独立概念的语义单元，作为候选术语。这一步中常借助互信息和邻接熵(刘伟童et al., 2019; 李贞贞et al., 2022)、对数似然比(王大亮et al., 2008)等统计指标，评估语义单元内部结合强度以及整体独立性，一旦统计指标超过预定的阈值，即可确认该语义单元为候选术语。在第二步中，则是要综合领域相关度对候选术语的置信度排名，确定前若干名为术语。这一步往往根据候选术语在目标领域与普遍场景的词频分布差异，利用C-Value(Frantzi et al., 2000)、TF-IDF(董洋溢et al., 2017)及其改进版本(俞琰et al., 2020; Kosa et al., 2020)等统计量，评估候选术语的领域相关度。然而，这类方法存在一些问题：由于统计量大多以词频作为核心计算因素，低频的术语的统计指标可能会偏低，甚至接近于非术语；同时，为了有效地排除噪声，这类方法的统

基金项目：北京市优秀人才培养资助项目(2018000020124G063)；

计量往往会主观设定一个较高的阈值，这可能导致低频术语被遗漏(蒋婷, 2021)。因此，现有的基于统计的术语抽取方法，在识别低频术语上仍面临较大的挑战。

传统机器学习和深度学习方法均将术语抽取视作序列标注任务。在用于序列标注的机器学习模型中，条件随机场最为经典，已被用于多个领域的术语抽取任务(木合亚提·尼亚孜别克et al., 2016; 黄菡et al., 2019)。不同于最大熵模型或隐马尔科夫模型，条件随机场直接对标签序列的联合分布建模，从而允许引入各种复杂的特征，这在前两者中是较难实现的。然而，机器学习模型的性能通常受限于人工特征选择的准确性。近年来，深度学习的发展为这一问题提供了解决方案。神经网络模型，如门控循环单元(Kucza et al., 2018)、双向长短期记忆网络(吴俊et al., 2020)、图卷积网络(任秋彤et al., 2021)等，可以自动提取上下文特征；将这些特征应用于条件随机场，会进一步提升术语抽取的效果。但无论是特征工程方法还是深度学习方法，都需要一定规模的标注语料协助模型训练。这就需要专业人员参与数据标注，标注成本相对较高。另外，由于低频术语在训练语料中的出现次数较少，模型可能无法从有限的样本中学习到低频术语的特征知识，从而导致模型对低频术语的识别能力不足

由此可见，虽然现有的各种术语抽取方法从不同角度深入研究了术语抽取任务，但对于低频术语的抽取仍然存在不同的局限。因此，我们提出了一种术语抽取方法，该方法融合了基于假设检验的统计方法和词向量技术，能够有效地抽取低频术语。

3 方法

术语是指在特定领域或学科内，用于表示专业概念的一组约定性的语言符号。这些符号可以是单个词语、短语或缩略词，其目的是为了在学术研究、技术开发和行业应用等场景下，确保概念的精确表达与交流。本文参照术语的使用场景，认为领域术语应具有以下特点：

(1) 单元性：构成一个术语的字符串在文本中作为一个独立单元使用，单元内部呈现出强关联性，以术语“钢筋混凝土”为例，“钢筋”与“混凝土”在此共同参与构成了一个完整的术语，单独识别“钢筋”或“混凝土”是不符合单元性要求的；

(2) 领域性：术语是特定领域内的专门用语，术语所表达的语义内容应归属于其所属的领域；

(3) 专业认可性：术语在特定领域内具有权威性，术语的使用需要得到专业人士的认可，从而确保领域内沟通的准确、高效。

在上述标准中，专业认可性必须由领域专业人士确定，目前难以用机器替代；但单元性和领域性确是在一定程度上可由计算机完成。因此，针对术语的低频特性，我们基于术语的单元性和领域性设计了一种无监督领域术语抽取方法。整体流程如图1所示。该方法分别针对单元性、领域性要求，设计候选术语单元性识别、候选术语领域性筛选两个模块。候选术语单元性识别模块依靠文本自身信息，识别文本中语义完整、使用独立的字符串作为候选术语；候选术语领域性筛选模块评估候选术语与所属领域的领域相关性，将归属于特定领域的候选术语认定为领域术语。

术语自动抽取能够显著减少专家在术语整理和归纳过程中的工作量。通过将术语抽取技术融入专家工作，能够进一步优化术语标准化流程的效率与准确性。

3.1 候选术语单元性识别

参照单元性定义，本文将“钢筋混凝土”这样具有单元性的字符串称为一个语义单元，一个语义单元可能由一个词或多个词构成。本步骤将文本划分为若干语义单元，作为候选术语。

候选术语识别与中文新词识别任务有共同之处：术语是具有领域性的语义单元，新词是未登录的语义单元，两者都需要满足单元性要求。在新词识别中，识别语义单元的方法包括互信息、信息熵和假设检验等。前两者需人为设定阈值，达到阈值的候选项视为语义单元。然而，阈值设定通常具有主观性，其高低影响筛选效果。过低的阈值无法筛除干扰项，而过高的阈值可能导致低频词汇被误筛。基于假设检验的方法根据给定文本和语料库的频率信息自适应地为不同频率字符串设置不同的阈值，消除了主观设定阈值的局限性。Jiang等人(2022)提供了一个基于假设检验的语义单元识别方法，对语义单元尤其是低频语义单元具有良好的识别效果。

Jiang等人指出语义单元内汉字或词语所具备的两个关键特征：非偶然相邻性和强关联性，并设计了相应的假设检验方法用于判别两个特征。具体来说，针对语义单元内部任意相邻字符

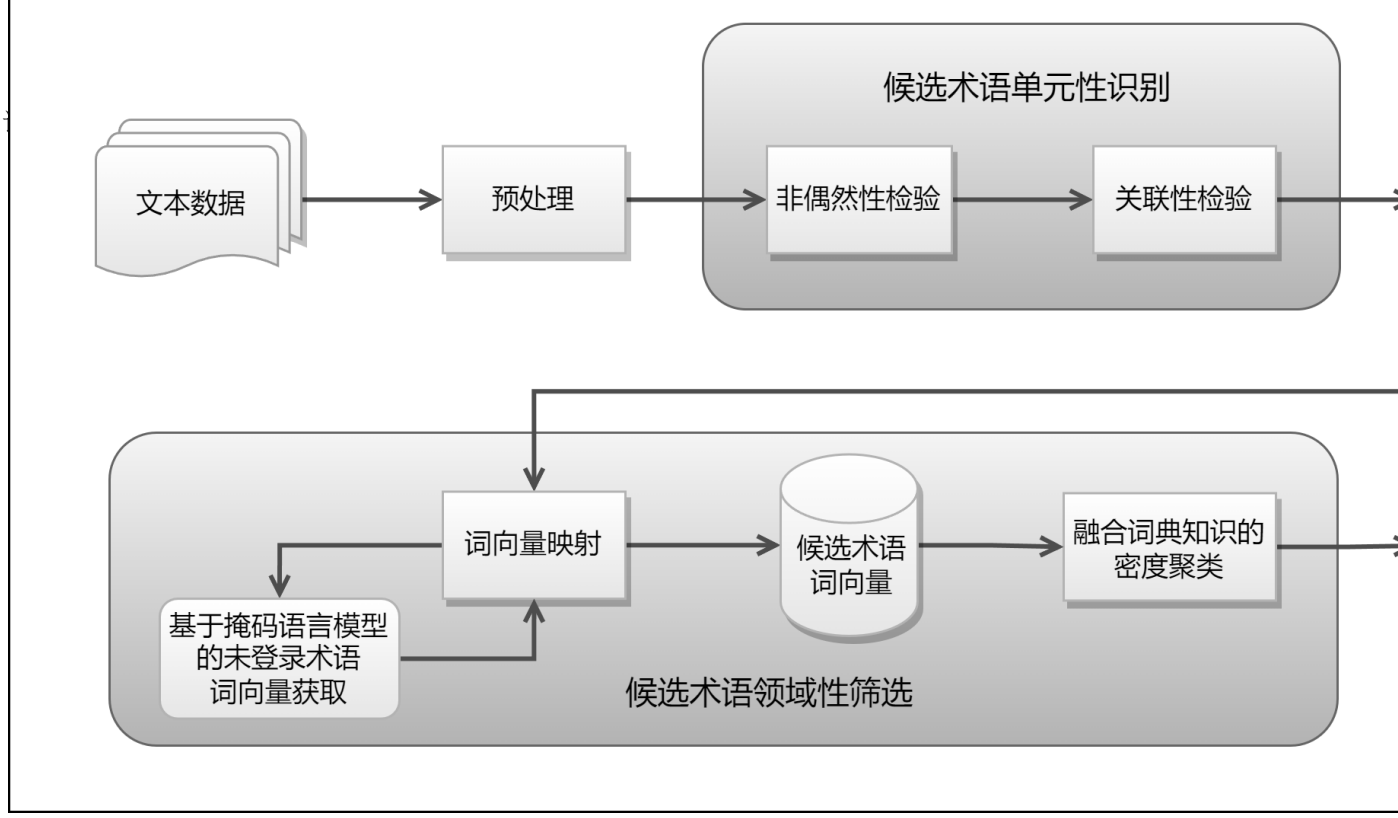


图 1: 领域术语抽取流程

之间的非偶然相邻特征，可以采用非偶然性检验评估相邻字符共同出现的显著性；考虑到构成语义单元的字符串作为一个整体在文本中表现出的强关联性时，可以使用关联性检验来判断相邻语义单元结合的紧密程度。

然而，这种自适应获得阈值的形式会为高频常见词设置较高的阈值，导致一些高频常见词不能通过非偶然性检验，进而影响了对包含高频常见词的语义单元的识别。不同于新词，许多术语含有高频常见词，如计算机领域术语“支持向量机”中的“支持”。为适应术语特点，本文将分词信息整合至Jiang等人的方法中，从而提高对包含高频常见词的语义单元的识别效果。具体而言，本文首先使用分词工具分割文本。现有的分词工具能够高效准确地识别高频常见词，因此我们无需对文本中全部相邻汉字对进行非偶然检验；相反，我们只需对相邻分词结果的邻接汉字对进行非偶然性检验，既提高了判断效率，又消除了非偶然性检验识别高频常见词的局限性。这样的融合策略能够更有效地识别高频常见词，进一步优化语义单元识别。

具体的，本文首先使用PKUSEG(Luo et al., 2019)对文本分词，分词结果存储为一个词语序列。对序列中相邻词语$[P,Q]$, 使用$< l_P, f_Q >$ 表示$[P,Q]$ 之间的相邻汉字对, 其中, l_P 表示词P 中的最后一个汉字, f_Q 表示词Q 中的第一个汉字。对$< l_P, f_Q >$ 进行非偶然性测试。

非偶然性检验的原假设为: 文本中任何相邻汉字对的频率都与它的一般频率特征一致。如果相邻汉字对的实际频率明显高于它的一般频率, 就可以判断该相邻汉字对为非偶然性相邻。

假设文本中任意两个汉字相邻出现的概率服从泊松分布, 则可以根据相邻汉字对的频率估计参数λ :

$$\lambda = N_T \times p = N_T \times \frac{n_{i,j}}{N} \tag{1}$$

式中$n_{i,j}$ 为相邻汉字对$< c_i, c_j >$ 在语料库中出现的频次，N 表示所有相邻汉字对出现的总频次，N_T 表示文本中相邻汉字对个数。可以通过以下公式计算$< c_i, c_j >$ 出现n 次的累计概率:

$$F_c\left(c_i, c_j, n\right) = \frac{\sum_{k=1}^{n} \frac{e^{-\lambda}\lambda^k}{k!}}{1 - e^{-\lambda}} \tag{2}$$

给定显著性水平α_p, N_a 表示$< c_i, c_j >$ 在文本中的实际出现次数, 若$F_c\left(c_i, c_j, N_a\right) > 1 - \alpha_p$, 则可以推断$< c_i, c_j >$ 是因非偶然性因素相邻的汉字对。

当$< l_P, f_Q >$ 满足非偶然性检验时, 才对相邻词语$[P,Q]$ 进行关联性检验。关联性检验的原假设为: 对于任意相邻词语$[A,B]$, A 、B之间的关联性不足以构成语义单元。

给定文本T, a, b, c, d 分别表示相邻语言单元$[A,B], [\bar{A},B], [A,\bar{B}], [\bar{A},\bar{B}]$ 出现在文本T 中的频

次, 用$[\bar{A}, B]$ 表示词B 的前一位置不是词A 的情况。构建统计量$Q^2_{A,B}$：

$$Q^2_{A,B} = \frac{(a+b+c+d) \times (ad-bc)^2}{(a+b) \times (c+d) \times (a+c) \times (b+d)} \tag{3}$$

给定显著性水平, 如果统计量$Q^2_{A,B}$ 的值在拒绝域中, 应拒绝原假设。这说明A 、B呈现强关联, 通过关联性检验。

若$[P, Q]$ 通过非偶然性检验、关联性检验，则意味着P 、Q可以共同构成语义单元或语义单元的一部分，应当合并。通过多轮迭代，合并所有能组成语义单元的词语，至文本中无可合并的词语为止。这时将当前所有获得的语义单元结果作为候选术语。

3.2 候选术语领域性筛选

术语的领域性是其另一个基本特征。由于候选术语是基于统计方法识别的，尚未判断其语义，导致筛选结果可能包含一些与领域无关的词汇。为解决这一问题，本文将进一步评估候选术语语义与所属领域语义的相关性，从而筛除领域外词汇，仅保留与领域相关的候选术语。为实现候选术语的领域性评估，本文采用基于词向量的聚类方法，这种方法相较于传统的TF-IDF方法具有优势，无需领域专用语料库即可执行。

词语的语义内涵可以通过词向量表示。词向量是基于大量语料训练得到的，常见的词向量模型如Word2Vec(Mikolov et al., 2013)能够有效表示训练语料中高频词汇。然而，针对领域术语的低频特性，Word2Vec模型存在一定的局限性。对于那些在训练语料中出现频次极低或根本未出现的词汇，Word2Vec无法生成相应的词向量，这类词汇很可能是新兴术语、非规范术语等低频术语，这一局限性使得Word2Vec在处理这类低频术语时面临诸多挑战。鉴于此，我们根据词向量的可用性，将候选术语分为已登录和未登录两类，使用基于Word2Vec模型的词向量数据集(Song et al., 2018)获取已登录候选术语的词向量。对于未登录候选术语，本文提出一种基于掩码语言模型的方法构建其向量表示，以克服低频特性对词向量表示的负面影响。

3.2.1 基于掩码语言模型的未登录候选术语词向量获取

掩码语言模型最初作为预训练任务在2018年提出，广泛应用于Bert(Devlin et al., 2019)、RoBERTa(Liu et al., 2019)等预训练模型。其思想方法为：按一定策略遮盖文本中的单词，让预训练模型根据遮盖位置的上下文信息，预测最适合填入遮盖位置的单词。具体而言，给定一个包含N个词的文本$\{x_1, x_2, x_3, \ldots, x_N\}$, 使用特殊标记[MASK] 遮盖其中第$j$ 个单词x_j, 掩码语言模型建模:

$$p\left(x_j \mid x_1, \ldots, x_{j-1}, [MASK], x_{j+1}, \ldots, x_N\right) \tag{4}$$

掩码语言模型在各类完形填空式任务中展现出良好的适应性，可直接应用于句法分析(Wu et al., 2021)、实体类型推断(Dai et al., 2021)等任务。本文利用掩码语言模型获取与未登录候选术语语义相近的替代词，并使用这些替代词的词向量构建未登录候选术语的词向量。具体来说，本文采用WoBERT预训练模型(Su, 2020)预测替代词。WoBERT是一种以词为训练单位的预训练模型，其训练任务基于掩码语言模型，因此更适合处理中文文本。

被遮盖候选术语	模型输入	模型输出结果
南荣油茶	攸县油茶、小果油茶、[MASK]、尾叶山茶等4个物种，鲜出籽率最高，达50%～60%以上。	大红袍, 油茶, 山茶
双苗砧嫁接	[MASK]效果好，当年嫁接苗高可达1.59cm，叶片数可达10片以上，它可以应用于快速培育嫁接大苗方面	嫁接, 移栽, 扦插
连续清查报告	根据《第9次全国森林资源[MASK]》，目前我国的国有林面积8274万公顷、集体林面积3874万公顷、个人所有林9673万公顷。	普查, 规划, 清查

表 1: WoBert生成替代词示例

为获得未登录候选术语的词向量，需要在WoBERT后添加softmax层，以便获取WoBERT对每一个替代词的预测概率。假设未登录候选术语w 在文本中出现M 次, 取出其所有出现位置的段落集合$S=\{S_1,S_2,\ldots,S_M\}$, 遮盖未登录候选术语后送入WoBERT，将所有位置的预测结果按预测概率由高至低排序，保留前top-K个替代词$\{w_i \mid i=1,2,3,\ldots,k\}$ 及其预测概率$\{p_i \mid i=1,2,3,\ldots,k\}$ 。由于替代词结果可能来自不同位置预测结果, 因此对预测概率归一化:

$$p'_i=\frac{e^{p_i}}{\sum_{j=1}^{K}e^{p_j}} \tag{5}$$

未登录候选术语的静态词向量$vecW$ 按如下公式计算得到:

$$vecW=\sum_{i=1}^{K}VEC\left(w_i\right)*p'_i \tag{6}$$

式中VEC (w_i)为词w_i 在词向量集合中映射得到的词向量，若VEC (w_i) 不存在, 则置为零。

3.2.2 融合词典知识的密度聚类

词向量聚类方法可以将语义相近的同领域术语映射到空间距离上的聚集，使得语义相近的同领域术语聚集在同一簇中，从而实现领域性筛选。由于术语抽取可能面临一个或多个学科分支主题的混合文本，这使得术语簇的形状难以确定。此外，候选术语中的与领域无关的词汇可能产生一定的噪声。因此，本方法采用具有抗噪声能力并能发现任意形状簇的密度聚类算法划分候选术语的语义类别。

为了在文本中的术语分布稀疏且数量较少的情况下，确保由候选术语映射的词向量在空间中能够规模化地聚集，本文从领域词典中收集已有的领域术语。将这些术语映射为词向量后，将它们与候选术语一起参与密度聚类。

本文将余弦相似度作为密度聚类中的距离指标，余弦相似度通过计算两个向量夹角的余弦来衡量两个词向量的语义相似度。词向量a,b 的余弦相似度记作$\mathrm{sim}(a,b)$ 。在此基础上，给出密度聚类中邻近集合的定义：给定向量v, 它的邻近集合$N_{\mathrm{eps}}\left(v\right)$ 定义为以v 为中心, 与v 的余弦相似度大于等于相似度阈值Eps 的向量集合, 即:

$$N_{eps}(v)=\{q \mid \mathrm{sim}(v,q)\geq \mathrm{Eps}\} \tag{7}$$

对于词向量v，给定一个密度阈值$minPts$，词向量v的邻近集合如果满足:

$$\left|N_{\mathrm{eps}}\left(v\right)\right|\geq \mathrm{minPts} \tag{8}$$

则称v 为在相似度阈值Eps，密度阈值$minPts$条件下的高密度点；自身不是高密度点但属于某个高密度点邻域集合的对象称为边界点；其余为噪声点。根据密度聚类的原理，从空间中任意一个高密度点Q开始，通过将Q附近的高密度点和边界点加入到Q所属的簇中，从而扩大簇的规模。通过不断地连接邻近集合中的高密度点，簇的方向得以拓展，最终实现对空间内词向量的聚类。在获取候选术语的聚类结果后，计算各个聚类簇与已有领域术语的交集。交集规模最大的簇被认定为领域簇，而属于领域簇的候选术语则被认定为领域术语。

4 实验

4.1 实验数据

在本节中，我们将测试所提方法的术语抽取能力，尤其是低频术语抽取能力，并与现有方法对比抽取效果。

由于尚无公开的中文术语抽取数据集，本文以林业领域为例手工构建实验文本。具体的，我们从国家林业和草原局政府网收集2022年4月至7月发布的林业碳中和主题新闻25篇，并选取湖南省常德市林业科学研究所编撰的《油茶优质高产栽培技术》一书的第四章“油茶良种繁育技术”，第八章“油茶科研应用及发展趋势”作为实验文本。实验文本主题涵盖了林业碳中和、良种繁育、林业科研成果三类不同林业主题，体裁包括林业新闻、技术类书籍，能够考察方法在不同场景的适用性。

本文采用词典与人工结合的方式标注实验文本中的林业术语，作为实验标准集。具体而言，如果一个词包含在《林学名词》、《中国林业辞典》、《中国树木志》《草业大辞典》林业辞典中，本文将其标注为林业术语。在此基础上，我们另外邀请五位林学专业研究人员以人工的方式标注出未在辞典内的林业术语，若一个词得到过半人数认可，则认定为林业术语。

预处理工作包括在去除实验数据中的图表内容与特殊字符。预处理后的文本数据如下：

主题文本	字数	术语数
林业碳中和	17,172	174
良种繁育	16,212	263
林业科研成果	15,943	208

表 2: 实验文本数据

4.2 实验结果与分析

实验采用Python3.7 编程语言和PyTorch 1.12 深度学习框架、scikit-learn0.24机器学习框架。候选术语单元性识别的显著性水平α_p 按建议固定在1×10^{-8} 。林业碳中和、良种繁育、林业科研成果三个实验数据集对应的密度聚类参数[Eps, minPts]分别设置为[0.75, 7]、[0.75, 5]、[0.75, 7]。

在实验评价标准上，本文选用准确率、召回率和综合评价指标F值评估实验结果。准确率反映抽取结果的正确性，召回率反映抽取结果的全面性，F值基于两者给出抽取质量的综合性评价。

为了验证本文提出的方法在领域术语抽取上的有效性，我们选取了中文开源术语抽取工具Termolato作为对比方法。Termolator 是一种结合了语法规则和统计的领域术语识别方法。它采用分词和词性匹配识别候选术语，并使用一般性语料库和领域语料库计算候选术语的相关统计指标，以统计指标评估候选术语的领域相关程度。在对比实验中，文本采用推荐配置运行Termolator，领域语料库选用《油茶优质高产栽培技术》除第四、第八章外的内容，一般性语料库则选用中文维基百科语料库*。

值得注意的是，可以通过设置词频阈值来调整Termolator的抽取效果。由于本文方法能够抽取最低词频为1的术语，为了在同等条件下对比，我们将Termolator的词频阈值也设置为1。在林业碳汇、良种繁育、林业科研成果等三个数据集上实验，结果见表3。

数据集	方法	准确率	召回率	F值
林业碳中和	本文方法	73.30	74.14	73.71
	Termolator	8.81	37.36	14.27
良种繁育	本文方法	76.30	78.33	77.30
	Termolator	16.90	56.27	25.99
林业科研成果	本文方法	73.10	78.40	75.63
	Termolator	12.81	59.61	21.11

表 3: 林业数据集上的实验结果

实验结果显示，本文方法在这三个数据集上的F值分别为73.71%、77.30%、75.63%；相较于Termolator方法，本文方法在准确率、召回率和F值上均实现了显著提升。同时，本文方法在不同体裁数据集上的F值表现稳定，方法有效性受体裁影响小。Termolator会误将部分与领域有间接关联、但并非术语的词也标记为术语，如“塑料薄膜”、“塑料棚”等。经分析，我们认为：此类词在领域语料库和通用语料库中的分布差异显著，在统计上表现出和术语相似的特征，这导致基于统计的Termolator方法产生了混淆。本文方法除使用统计特性，还会利用词向量从语义角度判断候选术语的类别归属，这能有效避免此类错误。另一方面，Termolator采用的词性规则更倾向于识别名词词组作为候选术语，这使得该方法对“皮下枝接”和“炼苗”等带有动词特性的术语产生了遗漏；而本文方法不受词性限制，具有更高的召回率。

*https://dumps.wikimedia.org/zhwiki/

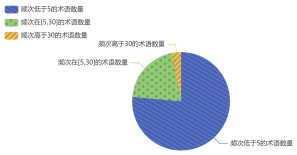


图 2: 良种繁育数据集术语词频分布

为了比较两种方法在低频术语上的抽取效果，本文首先统计良种繁育数据集的词频分布，结果见图2。统计结果显示，数据集中包含263个林业术语，其中出现频次低于5的低频术语有201个，占全部术语的76.43%。同时，所有术语的出现总频次为1246次，而低频术语出现频次仅为360次，占比28.89%。由此可见，良种繁育数据集中的术语分布具有明显的长尾特征，因此低频术语的抽取效果将对领域术语的整体抽取至关重要。

我们将本文方法与词频阈值分别设置为1、3的Termolator方法分别在良种繁育数据集上测试，并采用召回率来评价两种方法对低频术语抽取能力。从表4中可以看出，当词频阈值为1时，Termolator的低频术语召回率虽达到53.55%，但其抽取结果中包含大量干扰项，这导致其总体术语抽取的F值仅为22.13%。当将阈值提高至3，Termolator的F值提升了12.87%，但过于严格的阈值在筛除垃圾项的同时会误筛低频术语，导致低频术语召回率仅为26.38%。然而，本文方法在抽取术语最低词频为1的情况下，低频术语召回率达到74.88%，高出Termolator最佳结果21.33%；同时，总体术语F值达到77.30%。这表明本方法更适应于领域术语的长尾特征，能够在有效抽取低频术语的同时实现高质量的术语抽取。

方法	总体术语F值	低频术语召回率
本文方法	77.30	74.88
Termolator-1	25.99	53.55
Termolator-3	35.00	26.38

表 4: 在良种繁育数据集的抽取效果*

表5展示了本文提出的方法在两种体裁的语料上抽取到的林业术语示例。从展示样例可以看到，本文方法成功地抽取了诸如“木质化”、“树冠”和“间伐”等常用林业术语。同时，本方法还具有一定的林业新兴术语识别能力，例如“林业碳票”和“北京绿林认证”等，这些术语都是近两年应时代需求出现的林业新兴术语，在一定程度上反映了当前林业领域的热点问题。此外，本问方法还能成功抽取像“浙江红花油茶”这样的品种名称。

技术类书籍识别术语	**林业新闻识别术语**
木质化、韧皮部、腹接、嵌合枝接、常绿树、容器育苗、苗期、树冠、催芽、徒长枝、浙江红花油茶	林业碳票、北京绿林认证、林草碳汇、森林碳库、森林蓄积量、碳泄漏、间伐、中幼龄林、混交林

表 5: 识别术语样例

在实验结果中，我们还发现了一些语义相同但使用混乱的术语，如“无人机飞防”、“无人机

*Termolator-1,Termolator-3分别表示词频阈值设置为1和设置为3的Termolator方法。

防治”和“飞机防治”。这三个术语都表示“使用无人机喷洒农药以防治森林病虫害”，但由于缺乏统一标准，导致在具体使用中出现了用词混乱。这些词语识别将有助于反映了林业术语标准化需要关注和改进的方向。

5 结论

本文针对低频术语抽取所面临的挑战，提出了一种基于词向量的自适应术语抽取方法。这种方法不局限于特定领域，仅利用词典信息即可完成术语的无监督抽取。该方法分为候选术语单元性识别与候选术语领域性筛选两步。与现有研究成果相比，基于假设检验识别候选术语单元识别无需大规模语料库训练，能够根据目标文本中字符的统计特性自适应地设置参数，消除了由于主观设置的阈值而导致的低频词被筛除问题，从而克服了现有统计方法在低频术语识别方面的局限性。在判断候选术语领域性时，本文方法采用基于词向量的密度聚类方法筛选候选术语的领域相关性，该方法能够有效判别候选术语的领域性；针对未在词向量集内的未登录候选术语，本文采用基于掩码语言模型的方法获取其词向量，有效保证了新兴术语、非规范术语等未登录术语的筛选。

本方法在林业领域的三个数据集上开展实验，分别取得了73.71%、77.30%、75.63%的F值，并在低频术语抽取实验中达到74.88%的召回率。实验结果表明，本方法在不同体裁和主题的文本上表现稳定；相对于已有方法，本方法能够更为有效抽取低频术语，达到全面识别各频率术语的抽取效果。

参考文献

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-Fine Entity Typing with Weak Supervision from a Masked Language Model. pages 1790–1799, August.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, August.

Dongchen Jiang, Aoyuan Jiang, and Shuai Tang. 2022. An adaptive method for Chinese new word detection based on hypothesis testing. *Pattern Analysis and Applications*, 25(4):993–999, November.

Rodrique Kafando, Rémy Decoupes, Sarah Valentin, Lucile Sautot, Maguelonne Teisseire, and Mathieu Roche. 2021. ITEXT-BIO: Intelligent Term EXTraction for BIOmedical analysis. *Health Information Science and Systems*, 9(1):29, December.

Victoria Kosa, David Chaves-Fraga, Gennadiy Dobrovolskiy, and Vadim Ermolayev. 2020. Optimized Term Extraction Method Based on Computing Merged Partial C-Values. pages 24–49. January.

M. Kucza, J. Niehues, T. Zenkel, A. Waibel, and S. Stüker. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. *19th Annual Conference of the International Speech Communication, INTERSPEECH 2018; Hyderabad International Convention Centre (HICC)Hyderabad; India; 2 September 2018 through 6 September 2018*, page 2072.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *undefined*.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation, June.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating Domain Terminology into Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *undefined*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana, June. Association for Computational Linguistics.

Jianlin Su. 2020. WoBERT: Word-based Chinese BERT model - ZhuiyiAI. Technical report.

Jake Ryland Williams, Paul R Lessard, Suma Desu, Eric M Clark, James P Bagrow, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific reports*, 5:12209, August.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2021. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT, May.

任秋彤, 王昊, 熊欣, and 范涛. 2021. 融合GCN远距离约束的非遗戏剧术语抽取模型构建及其应用研究. 数据分析与知识发现, 5(12):123–136.

俞琰and 赵乃. 2018. 基于通用词与术语部件的专利术语抽取. 情报学报, 37(7):742–752.

俞琰, 陈磊, 姜金德, and 赵乃. 2020. 融合论文关键词知识的专利术语抽取方法. 图书情报工作, 64(14):104–111.

刘书剑and 彭道黎. 2011. 林业信息术语标准化研究. 林业调查规划, 36(01):104–107+116.

刘伟童, 刘培玉, 刘文锋, and 李娜娜. 2019. 基于互信息和邻接熵的新词发现算法. 计算机应用研究, 36(05):1293–1296.

吴俊, 程, 郝瀚, 艾力亚尔·艾则孜, 刘菲雪, and 苏亦坡. 2020. 基于BERT嵌入BiLSTM-CRF模型的中文专业术语抽取研究. 情报学报, 39(4):409–418.

孙水华, 黄德根, and 牛萍. 2016. 中医针灸领域术语自动抽取研究. 中文信息学报, 30(3):118–124.

木合亚提·尼亚孜别克, 古力沙吾利·塔里甫, and 达吾勒·阿布都哈依尔. 2016. 采用CRF模型的哈萨克语信息技术术语自动抽取技术研究. 西北师范大学学报*(*自然科学版*)*, 52(01):53–56.

李思良, 许斌, and 杨玉基. 2018. DRTE:面向基础教育的术语抽取方法. 中文信息学报, 32(3):101–109.

李贞贞, 钟永恒, 王辉, 刘佳, and 孙源. 2022. 基于深度学习与统计信息的领域术语抽取方法研究. 数据与计算发展前沿, 4(2):87–98.

游新冬, 杨海翔, 陈海涛, 孙甜, and 吕学强. 2021. 融合术语信息的新能源专利机器翻译研究. 中文信息学报, 35(12):76–83+93.

王大亮, 蒋宏潮, 涂序彦, 郑雪峰, and 佟子健. 2008. 基于选择倾向性的词汇获取方法. 计算机工程, (12):169–171.

董洋溢, 李伟华, and 于会. 2017. 文本特征和复合统计量的领域术语抽取方法. 西北工业大学学报, 35(4):729–735.

蒋婷. 2021. 学术文献术语抽取方案比较研究. 信息资源管理学报, 11(1):112–122.

黄菡, 王宏宇, and 王晓光. 2019. 结合主动学习的条件随机场模型用于法律术语的自动识别. 数据分析与知识发现, 3(6):66–74.

基于句法特征的事件要素抽取方法

余子健　朱桐　陈文亮
苏州大学计算机科学与技术学院, 江苏苏州2150062
{zjyu, tzhu7}@stu.suda.edu.cn
wlchen@suda.edu.cn

摘要

事件要素抽取（Event Argument Extraction, EAE）旨在从非结构化文本中提取事件参与要素。编码器—解码器（Encoder-Decoder）框架是处理该任务的一种常见策略，此前的研究大多只向编码器端输入文本的字词信息，导致模型泛化和远程依赖处理能力较弱。为此，本文提出一种融入句法信息的事件要素抽取模型。首先对文本分析得到成分句法解析树，将词性标签和各节点的句法成分标签编码，增强模型的文本表征能力。然后，本文提出了一种基于树结构的注意力机制（Tree-Attention）辅助模型更好地感知结构化语义信息，提高模型处理远距离依赖的能力。实验结果表明，本文所提方法相较于基线系统F1值提升2.02%，证明该方法的有效性。

关键词： 事件要素抽取 ； 成分句法分析 ； 事件抽取

Syntax-aware Event Argument Extraction

Zijian Yu　Tong Zhu　Wenliang Chen
School of Computer Science and Technology,
Soochow University, Suzhou, Jiangsu, 215006, China
{zjyu, tzhu7}@stu.suda.edu.cn
wlchen@suda.edu.cn

Abstract

Event Argument Extraction (EAE) aims to extract event arguments from unstructured text. The Encoder-Decoder framework is a common solution for this task. Most of the encoders in previous studies only receive plain text information, resulting in weak generalization and remote dependency processing capabilities of the models. To this end, we propose a new event argument extraction model incorporating syntactic information. Firstly, after the text is analyzed to obtain a constituent parsing tree, the part-of-speech tags of words and the syntactic component labels of nodes are encoded to enhance the context representation ability. Then, we propose a tree-structure-based attention mechanism (Tree-Attention) to assist the model to better perceiving structured semantic information and improve the model's ability to handle long-distance dependencies. The experimental results show that the method yields better results with the F1 value by 2.02% compared to the baseline system, which proves the effectiveness of our method.

Keywords: Event Argument Extraction , Constituency Parsing , Event Extraction

* 基金项目：2020-2024自然科学基金重点联合项目：自然语言对话交互的基础理论和方法(61936010)

1 引言

事件抽取是一项基础自然语言处理任务，旨在从非结构化文本抽取出结构化的事件信息。该任务主要包括事件识别和事件要素抽取两部分。事件识别主要是确定事件属于哪种类型，而事件要素抽取则是根据事件类型预先定义好的事件框架，抽取相应要素类型的论元。

对于事件要素抽取任务的研究，通常是围绕词粒度[0]的语言模型展开。例如Hsi et al. (2016)使用了Word2Vec作为词的向量表示，该方法虽然能够一定程度上表达词的含义，但是由于训练窗口大小的限制，其所学习到的语义信息较为有限。Kenton and Toutanova (2019)提出的BERT语言模型很好地解决了Word2Vec的窗口限制问题，相较之下能够更好地学习到全局信息，这也是目前应用最为广泛的语言模型之一。但同类词[1]之间被看作是相互独立的，缺少信息共享，导致泛化能力较弱。并且一些相关联的词随着距离的增大，模型对它们的处理能力也逐渐减弱。例如对中文句子中一个词语的表示，上述模型中并不能很好地体现其之间的依赖关系。传统的事件要素抽取方法通常是基于卷积神经网络(Li et al., 2019; Chen et al., 2015)和递归神经网络(Patchigolla et al., 2017; Zhao et al., 2018)来实现的，这些方法虽然能够学习一些底层的特征，但是同样存在泛化能力较弱和远距离依赖处理能力不强的问题。

为解决上述问题，增强模型的表达能力，本文提出融入成分句法树的模型框架。成分句法解析(Cross and Huang, 2016)（constituecy parsing）能根据给定的句子，分析出句子的短语结构句法树以及词性信息，能够为下游任务提供更丰富的句法语义信息。本文通过Tree-LSTM（Tree Structured Long Short-Term Memory）将成分句法树进行编码融入Encoder端，并且提出了树注意力机制（Tree-Attention）来更好地学习树的结构特征。

总体来说本文包含以下三点贡献:

- 将成分句法解析的结果作为额外特征，借助词性标签和树节点句法标签不同粒度的表示单元对文本进行编码，从而得到更好的文本表示，增强模型泛化能力;
- 利用Tree-LSTM，并提出了树的注意力机制，辅助模型感知句法解析树的结构和特征，并且通过加权的方式将注意力网络进行融合，缓解远距离依赖的问题。进而提高了事件要素抽取的效果;
- 该方案在CCKS2022 Task10数据上的实验结果表明，融入句法信息的事件要素抽取的F1值相较于基线系统提高2.02%，证明了本文所提方案的有效性。

2 相关工作

近些年基于深度学习的文本表示研究主要有以下几点。Zaremba et al. (2014)提出的循环神经网络(Recurrent Neural Network, RNN)，按照时序对文本进行计算，相较于CNN(LeCun et al., 2015)拥有更好的短期记忆Hochreiter and Schmidhuber (1997)提出的长短期记忆网络（long short-term memory, LSTM）是RNN 的一种变体，通过“输入门”、“输出门”和“遗忘门”的结构，让信息能够按序列传递下去。双向长短期记忆网络（Bi-LSTM）是LSTM的优化版本，通过正反两个方向进行单向的信息传递，最终通过拼接或加权的方式进行融合，该方法广泛应用于命名识别任务(成于思and 施云涛, 2020; 镇宇et al., 2019)。Kenton and Toutanova (2019)提出的基于预训练模型BERT（Bidirectional Encoder Rep-resentations from Transformers）的框架，不仅能够很好地保留局部的关键信息，还能更好地挖掘全局的信息，因此常作为抽取任务的Encoder(蔺志et al., 2022; Zhang et al., 2020)。

现有的事件要素抽取方法，很大程度上忽略了要素之间的关系和相互作用，而这一层信息已被证实能辅助模型更好地进行要素抽取(Li et al., 2021)。不仅要素之间的关系影响模型性能，事件要素类型和事件要素之间也具有强相关性。Xu et al. (2023)显式地利用了要素类型和要素的关系，通过对比学习和循环训练策略，提高了模型的性能。

句法信息能够将句子中不同词联系起来，对要素抽取任务有一定的辅助效果。Ding et al. (2022)针对各要素类型之间的联系，提出了显式角色交互网络，它允许动态捕获事件中不同参数角色之间的相关性，以提高事件要素抽取的效果,但是仅使用了词性标签，并未充分利用文本的结构化信息;Sun et al. (2022)将句法结构信息与命名实体识别任务（Named Entity Recornigtion, NER）结合起来，提升了模型的性能，证明了句法信息有助于模型理解能力的提升;Veyseh et al. (2020)基于图转换网络提出了SemSynGTN模型，充分利用句子的句法和语义信

[0]在中文文本处理时通常以字为粒度
[1]例如“可乐”中的“可”与“乐”

息，更好地学习文本的表示和结构，提升模型的要素抽取能力。本文将该工作作为对比模型进行比较。

Tai et al. (2015)等人提出了Tree-LSTM，与标准的LSTM相同，Tree-LSTM的每个单元都包含了三个门，但不同的是Tree-LSTM单元中门向量和细胞状态的更新依赖于所有与之相关的子单元的状态。另外，Tree LSTM拥有多个遗忘门，分别对应当前单元的每个子单元，这使得Tree-LSTM可以选择性地从子节点中获取信息。这种模型结构能够很好地将成分句法树融入文本语义表达中。

3 任务定义与数据

在本文中，抽取任务的输入为一段中文文本$S = \{w_1, w_2, w_3, \ldots, w_n\}$以及一段对相应事件的描述$M = \{m_1, m_2, m_3, \ldots, m_t\}$，描述$M$是对抽取事件主体的概括。输出为根据预先定义好的槽位抽取的事件论元。其中一段文本S所包含的事件描述M可以有一个或多个，每个槽位抽取的要素可以为空，也可以有多个。文本S与描述M均不为空，并且抽取的要素字段均出现在文本S中。

事件要素槽位包括以下三个：1)事件主体，表示在该事件描述下的施事方，通常为一个人、组织或机构；2)事件动作，表示事件主体的一些行为，大多为动词或包含动词的短语；3)事件客体，表示该事件的受事方，根据事件的类型不同，可能是一个人、机构、政策或动物等受施事方影响的客体。

如表1给出的样例所示，包含一段文本以及一个事件描述，事件要素包括：1) 事件主体为“美联储”；2) 事件动作为“采取”；3) 事件客体为“缩紧政策”。

文本	美元指数盘尾涨0.13%，报99.97，上周五曾触及100.19，2020年5月以来最高；由于预期美联储将采取更激进的紧缩政策，美债收益率急升，提振美元，指标10年期美债收益率周一达到2.793%的2019年1月以来最高。
事件描述	预期美联储将采取更激进的紧缩政策
事件主体	美联储
事件动作	采取
事件客体	缩紧政策

Table 1: 数据样例

我们使用HanLP(He and Choi, 2021)工具对文本解析，从而获得文本的成分句法结构。如图1所示，系统输入为一段文本，输出为该文本的成分句法树以及分词之后的词性结果。该工具的句法标签遵循PTB3.0(Xue et al., 2000)标注规范，其中，“IP”表示简单子句或句子，通常不带补语（如“的”、“吗”等）；“NP”表示名词短语，中心词通常为名词；“VP”表示动词短语，中心词通常为动词。

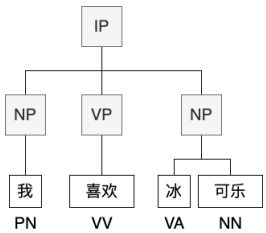


Figure 1: 成分句法解析样例

对于词性标签，假设原始文本为$S = \{w_1, w_2, w_3, \ldots, w_n\}$，分词之后得到$k$个词，结果为$D = \{d_1, d_2, d_3, \ldots, d_k\}$，其中$d_i$表示第$i$个词。并且$D$中每个元素$d_i$都对应一个词性标签。图1的实例中，“我”、“喜欢”、“冰”、“可乐”四个词的词性标签分别对应“PN（代词）”、“VV（动词）”、“VA（表语形容词）”、“NN（普通名词）”。词性标签的标注准则遵循PTB3.0(Xia, 2000)标注规范。

在该任务中，输入的句子可能包含多个简单句。解析器会自动识别并切分成多个简单句，每个简单句（IP）连接至根节点（Root），构成该句子的完整成分句法树。为了利用句子的句法成分信息，我们将“IP”标签的孩子节点标签作为其叶子节点的句法标签。具体的，在图1的例子中，“IP”标签下包含了“NP”、“VP”和“NP”三个孩子节点标签，则这三个句法标签分别对应了“我”、“喜欢”和“冰可乐”。若出现简单句的嵌套，即“IP”标签的孩子节点中存在另外一个“IP”标签（下称为子“IP”节点）。除子“IP”节点外的孩子节点与上述方法一致，作为其叶子节点的句法标签，而子“IP”节点则继续向下确定句法标签，以此类推。若数据中出现一个词语作为一个子句的情况，则失去了句法树的结构信息。

本文采用CCKS2022 Task10[2]面向金融领域的因果事件要素抽取学术评测提供的数据作为实验数据。该数据集主要来自公开的新闻和报告，共有4,000条标注规范的样本，其中包含9232个事件，平均一条文本包含2个事件，平均句子长度为191个字。在实验中，对该数据使用随机采样方法，将数据集打乱，按8:1:1的比例划分为训练集、开发集及测试集。

我们对句法分析结果进行统计分析，发现事件要素与词性标签以及句法标签有着密不可分的关联。其中，事件主体81.42%的句法标签为名词短语，词性标签包含63.13%的名词以及22.44%的形容词与名词的组合；事件动作88.74%的句法标签为动词短语，79.02%的词性标签为动词；事件客体94.12%的句法标签为名词短语，85.75%的词性标签为名词。

4 模型框架

本文考虑如何提高模型泛化能力和远距离表示能力，提出一种融合句法特征的事件要素抽取模型。我们的模型利用词性标签和句法标签建立不同词之间的联系增强模型泛化能力，基于句法树结构建立远距离依赖。

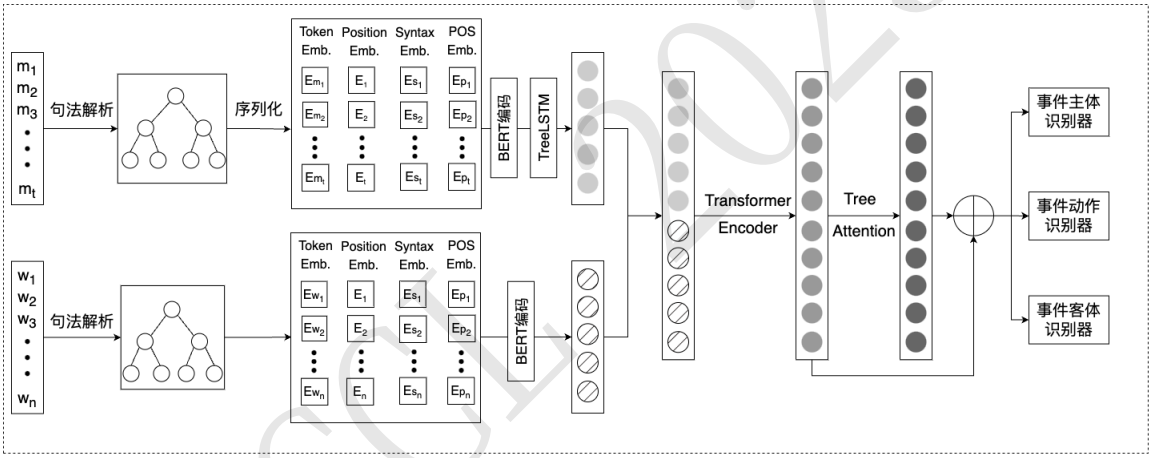


Figure 2: 总体模型框架图

图2是总体模型结构图，主要包括四个步骤：1）利用成分句法解析器对输入文本进行分析，得到成分句法树、词性和各叶子节点的句法标签。分别将句法标签和词性标签通过BMES的标注模式打上标签并编码。2）将事件描述M的成分句法树通过Tree-LSTM进行编码，再与第一步的编码进行拼接，作为Transformer Encoder的输入，并利用Transformer Encoder进行信息的交互。3）将第二步的结果作为Tree-Attention的输入，利用注意力机制，增强成分句法树在隐藏向量中的表示，并将增强后的向量与第二步的结果通过加权的方式进行融合，得到最终的语义表示。4）使用多层二元标志解码器对句子表示进行解码，识别出各事件要素的首尾位置，最终得到各槽位的论元。本节详细介绍我们的要素抽取模型，主要分为三个模块：编码、训练以及解码。

4.1 编码

4.1.1 词性及句法标签的嵌入

我们采用BMES标注模式对词性和句法标签进行表示，其中B表示某一标签内容的开始部分；M表示该字符为某标签内容的中间部分；E表示某标签的最后一个字符；S表示该字符单独

[2]https://www.biendata.net/competition/ccks-nec-2022/

作为一个标签的内容。具体而言，图1给出的例子中，“我”的词性标签为“S-PN”，“冰可乐”三个字符的句法标签分别对应“B-NP”，“M-NP”，“E-NP”。

将字符的词性及句法标签分别进行编号，并且进行随机初始化，之后连同文本一起作为BERT的输入，经过计算得到的句子表示为$\boldsymbol{H}=\{\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n\},\boldsymbol{H}\in\mathbb{R}^{L\times i}$，计算过程如式1所示。

$$\{\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n\}=\text{BERT}\left(x_t,x_c,x_s\right) \tag{1}$$

其中，x_t为句子的字符，x_c与x_s分别为词性标签与句法标签，与字符一一对应。每个字符的向量表示$\boldsymbol{h}_i\in\mathbb{R}^i$，$i$表示向量的维度，$L$为最大句长。

4.1.2 基于Tree-LSTM的句法树编码

本节介绍基于Tree-LSTM的事件描述M的成分句法树的编码方案。基于LSTM的树编码方式主要有N-ary Tree-LSTM和Child-Sum Tree-LSTM。其区别在于，前者所编码的树中除了叶子节点外的每一个节点的孩子数相同，而后者对于树的结构没有特别的要求。由于文本的多样性，文本句法解析树的结构往往无法满足N-ary LSTM的要求，因此本文采用Child-Sum Tree-LSTM的方法对成分句法树进行编码。

与传统的LSTM相同，Tree-LSTM包含输入门（i_j）、遗忘门（f_{jk}）、输出门（o_j）、记忆单元（c_j）以及隐藏状态h_j，但不同的是Tree-LSTM循环单元需要处理一个或多个子节点的隐藏状态和单元状态的信息。为了让关键信息在传递的过程中尽可能保留下来，Tree-LSTM的循环单元对每个子节点单独设立遗忘门。因此Tree-LSTM循环单元包含一个输入门、一个输出门、一个单元状态和多个遗忘门。

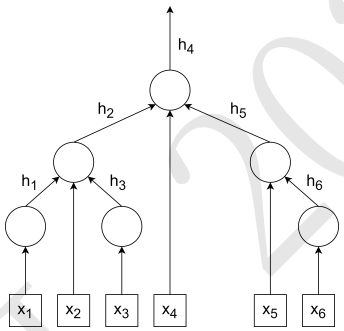


Figure 3: Tree-LSTM模型结构图

Tree-LSTM的结构如图3所示，不同于传统LSTM的线性循环单元，Tree-LSTM在树形结构上将循环单元自底向上递归展开，下面将具体计算过程及原理。在计算每个节点时，令该节点为j，其所有子节点的集合为$C(j)$，细胞状态为c_j，隐含状态输出为h_j。节点j的输入为当前节点j的输入x_j以及所有子节点的隐含状态$\mathbf{h}_i$（$i\in C(j)$）。对于每个子节点$i\in C(j)$都设置了一个遗忘门，以确定并保留子节点的哪一部分信息。每个子节点$k\in C(j)$的遗忘门输出取决于当前节点输入x_j和子节点的隐含状态输出$\boldsymbol{h}_k$，并且经过sigmoid函数激活得到权重值f_{jk}，计算过程如下式所示：

$$f_{jk}=\sigma\left(W^{(f)}\boldsymbol{x}_j+U^{(f)}\boldsymbol{h}_k+b^{(f)}\right) \tag{2}$$

取得所有子节点隐含状态之后，输入门首先对其求和得到$\widetilde{\boldsymbol{h}}_j$，根据$\widetilde{\boldsymbol{h}}_j$和当前节点输入$\boldsymbol{x}_j$决定输入门权重值$\boldsymbol{i}_j$及备选细胞状态$\boldsymbol{u}_j$，之后根据遗忘权重和输入权重更新当前节点的状态，具体计算公式如下：

$$\begin{cases}\tilde{\boldsymbol{h}}_j=\sum\limits_{k\in(j)}\boldsymbol{h}_k\\ \boldsymbol{i}_j=\sigma(W^{(i)}\boldsymbol{x}_j+U^{(i)}\tilde{\boldsymbol{h}}_j+b^{(i)})\\ \boldsymbol{u}_j=\tanh(W^{(u)}\boldsymbol{x}_j+U^{(u)}\tilde{\boldsymbol{h}}_j+b^{(u)})\\ \boldsymbol{c}_j=\boldsymbol{i}_j\odot\boldsymbol{u}_j+\sum\limits_{k\in C(j)}f_{jk}\odot\boldsymbol{c}_k\end{cases} \tag{3}$$

输出门根据当前节点的输入$\boldsymbol{x}_j$和子节点隐含状态之和$\widetilde{\boldsymbol{h}}_j$，计算得到输出门的权重$\boldsymbol{o}_j$，再通过当前细胞状态得到当前节点的输出$\boldsymbol{h}_j$，计算公式如下：

$$\begin{cases} \boldsymbol{o}_j = \sigma\left(W^{(o)}\boldsymbol{x}_j + U^{(o)}\widetilde{\boldsymbol{h}}_j + b^{(i)}\right) \\ \boldsymbol{h}_j = \boldsymbol{o}_j \odot \tanh\left(\boldsymbol{c}_j\right) \end{cases} \tag{4}$$

其中，σ表示sigmoid函数，通过sigmoid函数将向量的数值控制在0-1之间，$\odot$为点积表示矩阵对应元素相乘。W和U为可学习的权重矩阵，b为各个门的偏置，tanh为激活函数。

该部分对经过Tree-LSTM编码后的成分句法树和经过词性及句法增强的文本表示进行信息交互，以增强模型对成分句法树的感知，并且能够区分同一文本不同事件的隐藏表示。具体计算如下式所示：

$$\boldsymbol{h} = \text{BERT}\left(\text{concat}\left(\boldsymbol{h}_m, \boldsymbol{h}_t\right)\right) \tag{5}$$

其中，concat表示拼接操作，$\boldsymbol{h}_m$为成分句法树的语义表示，$\boldsymbol{h}_t$为词性和句法增强的文本语义表示。

4.1.3 融入树结构的注意力机制

为了进一步增强树的语义表达，我们提出了树的注意力机制（Tree-LSTM），通过依次计算父子两个词之间的相似度，将句法树中除叶子节点外的两个词联系起来。其计算过程如下。

首先在注意力机制中，输入向量与三个不同的变换矩阵相乘的到$\boldsymbol{q}$，$\boldsymbol{k}$，$\boldsymbol{v}$,分别为查询向量，关键词向量和值向量。每个关键词向量依次与查询向量进行点积运算计算相似度，再利用softmax进行归一化，将相似度转化为和为1的得分矩阵，最后与值矩阵相乘的到Attention的计算结果，具体计算如下所示：

$$\text{Attention}\left(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}\right) = \text{softmax}\left(\frac{\boldsymbol{q} \cdot \boldsymbol{k}^T}{\sqrt{d_w}}\right) \cdot \boldsymbol{v} \tag{6}$$

其中d_w为输入向量的维度。

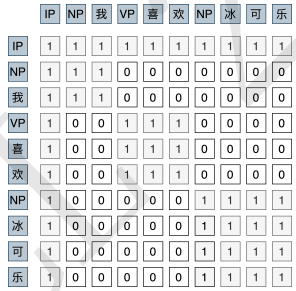


Figure 4: Mask矩阵示意图

Tree-Attention是在注意力机制中，添加一个二维的掩码（Mask）矩阵，矩阵元素1表示参与注意力的计算，0则表示不参与。Mask矩阵根据成分句法树的父子关系来确定，除叶子节点外，其余节点均会参与其所有子孙节点的注意力计算。具体地，如图4所示为“我喜欢冰可乐”的Mask矩阵。Tree-Attention部分的输入为经过语义交互层Transformer Encoder的向量$\boldsymbol{X}$，输出为经过Tree-Attention计算后的向量$\boldsymbol{Y}$，最终的文本表示$\boldsymbol{h}$通过公式7对$\boldsymbol{X}$和$\boldsymbol{Y}$加权融合所得，其中λ为0到1(含)之间的数。

$$\boldsymbol{h} = \lambda \cdot \boldsymbol{X} + (1 - \lambda) \cdot \boldsymbol{Y} \tag{7}$$

4.2 解码及训练

本文采用多层二元解码器对语义表示进行解码，以得到各槽位的事件论元。对于每个槽位，我们利用两个全连接层，将语义的隐层表示映射到二维的向量中，分别来预测其论元的起始位置和结束位置。计算过程如下所示：

$$p_i^{(s,c)} = \sigma\left(W_s^c \cdot \boldsymbol{h}_i + b_s^c\right) \tag{8}$$

$$p_i^{(e,c)} = \sigma\left(W_e^c \cdot \boldsymbol{h}_i + b_e^c\right) \tag{9}$$

其中，$\boldsymbol{h}_i$为第i个隐藏表示，s、e表示开始和结束位置。$p_i^{(s,c)}$和$p_i^{(e,c)}$分别表示第i个位置为要素类型c的开始和结束位置的概率。W为可学习参数，b为偏置，σ为激活函数。该部分模型的训练目标函数如下式所示：

$$loss = -\sum_{i=1}^{n} r_i^{(s,c)} \log\left(p_i^{(s,c)}\right) - \sum_{i=1}^{n} r_i^{(e,c)} \log\left(p_i^{(e,c)}\right) \tag{10}$$

其中n为文本序列的长度，$r_i^{(s,c)}$和$r_i^{(e,c)}$分别为标注数据的正确分类标签。之后通过以下规则确定最终抽取的事件要素：1)若在预测时没有预测出起始位置或者结束位置，则判定为该文本中不存在相应槽位的事件要素；2)若起始位置之后没有结束位置或者结束位置之前没有起始位置，则判定该结果为非法结果，不算作最终答案。

5 实验与分析

5.1 评价指标

在实验中，本文采用微平均的精确率(Precision,P)、召回率(Recall,R)、F1值(F1-measure,F1)来评价事件要素抽取的效果，即所有的事件要素类型一起计算P、R及F1值，具体计算方法如下式所示：

$$P_{micro} = \frac{\sum_{i=1}^{n} TP_i}{N_{predict}} \tag{11}$$

$$R_{micro} = \frac{\sum_{i=1}^{n} TP_i}{N_{gold}} \tag{12}$$

$$F1_{micro} = 2 \cdot \frac{P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}} \tag{13}$$

其中TP_i表示第i类事件要素预测正确的数量，$N_{predict}$为所有预测出的要素数量，N_{gold}表示所有金标答案的数量。除此之外，在预测时仅当预测的要素与标准答案的要素完全一致才算正确。

5.2 实验设置

本实验使用开源的预训练语言模型roberta-base-wwm[3]作为我们编码层的一部分。我们根据模型在开发集的性能设置超参数。最大句长为512，向量维度采用默认值768，mini-batch设置为16，迭代次数为20，采用Adam优化器以2e-5的学习率优化模型。为防止过拟合的情况，最终的隐藏表示通过dropout层随机丢失一些节点，丢失的比例为0.5。λ用于调节Tree-Attention过程中的向量加权融合的占比，为0.5。

5.3 对比模型

为了对比前人的事件要素抽取方法与本文方法的差异，我们选择了以下五种模型进行对比：

- BERT+Bi-LSTM+CRF(Jiang et al., 2019)：该模型将事件要素抽取作为序列化标注的问题，依靠其独特的门控机制，能够很好地学习并保留文本的局部和全局信息。该方法使用Roberta-base作为第一层Encoder，提取句子特征，紧接着利用Bi-LSTM的双向信息传播的特点对文本的信息进行交互并编码，最后通过CRF（Conditional Random Fields）进行解码，得到最终的抽取结果。
- BERT+GlobalPointer(Tao et al., 2022)：该方法利用全局归一化的思想，并且引入旋转式位置编码（RoPE），融入了要素在文中的开始和结束位置信息，最后通过要素首尾对的打分，得到抽取的结果，相较于CRF的解码方式更加合理，且泛化性更好。
- BERT+Biaffine(Yu et al., 2020)：该模型基于预训练语言模型对文本进行特征提取，利用Biaffine模型在多层神经网络输出句子中所有可能的要素的分数，然后对这些进行排序，返回各个槽位分数最高的要素作为答案。

[3]https://huggingface.co/hfl/chinese-roberta-wwm-ext

- MRC+SPAN(Li et al., 2020)：该模型将事件的描述作为针对文本提出的问题，并且通过特殊符号拼接在文本前面作为BERT的输入。最后通过三个二元分类器得到最终抽取的要素结果。
- SemSynGTN(Veyseh et al., 2020)：该模型利用了依存句法信息，将句法和语义信息转换为一个加权有向图，其中节点表示句子中的词，边表示依存句法和语义角色关系，权重表示边的重要程度，并且引入了图转换网络，每个图转换层通过聚合邻居节点和边的信息来更新当前节点的表示。最终通过全连接层抽取相应的要素。

5.4 实验结果

5.4.1 性能对比分析

方法	P/%	R/%	F1/%
BERT+Bi-LSTM+CRF	54.815	55.224	55.019
BERT+GlobalPointer	57.423	57.124	57.312
BERT+Biaffine	58.102	57.761	57.931
MRC+SPAN	58.682	58.146	58.413
SemSynGTN	60.024	59.113	59.565
本文方法	**61.457**	**59.613**	**60.521**

Table 2: 不同抽取方法的实验结果

表2是本文系统与对比模型的实验结果。从表中结果，我们发现如下现象：

BERT+Bi-LSTM+CRF无法针对不同的事件提取不同的语义特征，因此在本文的数据集上效果不佳。BERT+GlobalPointer和BERT+Biaffine可以利用特殊符号，将不同事件的信息与文本进行交互，提高了语义的表征能力，F1值相较于Bi-LSTM+CRF有明显提升。

MRC+SPAN利用了阅读理解的机制，将事件的Mention作为问题，并利用[CLS]和[SEP]特殊符号进行拼接作为输入并通过transformer进行信息交互，学习并保存了不同事件描述但相同文本的不同语义表示，再通过对要素起始位置和结束位置的预测，得到最终抽取的结果。我们将该模型作为基础模型框架（基线模型）搭建本文系统。

本文方法和SemSynGTN的F1值均高于其他系统，说明句法信息的确对事件要素抽取有帮助。而本文方法相较于SemSynGTN有0.956%的优势，说明本文方法能更好的利用成分句法信息来帮助本文的抽取任务,原因可能在于树结构相较于图结构更能体现句法树的依赖关系。

从表2的结果来看，本文的方法相较于BERT+Bi-LSTM+CRF提升了5.50%。对比BERT+GlobalPointer和BERT+Biaffine也有3.209%和2.59%的优势。从而证实了该方案的有效性。和他们方法最大的差异在于，本文方法将句法特征融入模型中，提升模型对语义强关联词的信息捕捉能力，一定程度上缓解了远距离依赖的问题。

总体来看，本文的方法相较于前人方法在性能上有着一定优势。相对于基线模型MRC+SPAN也有2.02%的提升。说明本文方法能够较好地融合成分句法树的信息，辅助模型挖掘更深层的信息，提高了事件要素抽取的精确率和召回率。

5.4.2 λ对结果的影响分析

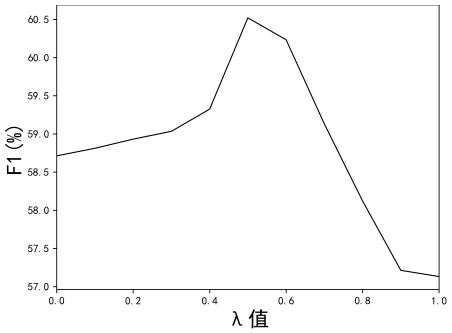


Figure 5: 模型的F1值随λ趋势变化图

通过Tree-LSTM和Transformer Encoder的文本嵌入与经过Tree-Attention后向量表示的融合是本文方法的核心。为探究Tree-Attention是否对模型有帮助以及其所占权重为多少的效果最佳，我们通过观察F1值随λ值的变化趋势，选择F1值最高的λ值作为融合的权重。

在开发集上，F1值随λ变化的趋势如图5所示，整体趋势随λ先上升后下降，在λ取0.5时达到峰值。从曲线在λ=0.5前后变化可以发现，句法信息和文字信息之间的重要程度接近。当λ=1.0时，模型仅对句法结构进行编码，此时出现的较快下降意味着简化的句法结构文本中含有的信息量较少。因此需要将文字信息和句法结构信息进行融合，同时增加模型的文本表征和长句依赖捕捉能力。

5.4.3 消融实验

本文模型在基础模型的基础上额外增加了4个重要组件，分别是词性嵌入、句法嵌入、句法树嵌入以及Tree-Attention。其中，上一节的实验表明，以0.5的权重融入Tree-Attention相比未添加Tree-Attention组件F1值提升了1.81%，充分证明其有效性，因此本节不再对该部分进行论证。在本节中，我们对其余3个组件进行消融实验，分别在基础模型上单独添加其中1个组件。

	P/%	R/%	F1/%
基础模型	58.682	58.146	58.413
＋词性嵌入	62.026	56.320	59.035
＋句法标签嵌入	61.153	57.374	59.203
＋句法树嵌入	61.042	57.721	59.335
＋以上三种组件	61.230	58.202	59.678

Table 3: 消融实验结果

消融实验结果如表3所示，词性嵌入、句法嵌入以及句法树嵌入三者相较于基础模型F1值分别有0.622%、0.79%以及0.922%的提升，说明三个组件均有助于模型性能提高。

词性标签来自于对分词后每个词的词性预测，词性嵌入将词性标签作为一种特征融入至模型之中。由于融入词性标签的分词以及词性信息，使得模型相较于基础模型能够更好地对文本进行表示，进而能够更好地识别要素在原文中的起始与结束位置，缓解了要素抽取的边界错误问题。

句法标签能够清晰表示文本中各个子句的组成成分，并且各字句的句法标签与抽取的要素有着密不可分的关系。例如事件动作的要素类型，其在子句中的句法标签通常为“VP”（动词短语）。依靠句法标签的句法嵌入能够辅助模型对要素的定位，缓解要素抽取错误以及非空槽位抽不出要素的问题。

句法树嵌入充分利用了文本的句法解析结果，不仅包含了句法标签的信息，还将成分句法树的结构信息融入模型，辅助模型对上下文的语义理解。句法树嵌入也是该三个组件中提升效果最明显的。

5.4.4 远距离依赖的效果分析

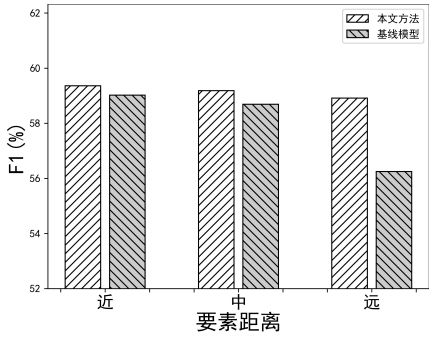


Figure 6: 模型F1值随主客体距离变化图

为了探究本文方法是否缓解远距离依赖的问题。根据事件主体与事件客体之间的距离，我

们将开发集划分成三等份，分别为远、中、近，并统计基线模型和本文模型在三个集合中事件主体与事件客体的F1值。

结果如图6所示，在事件主体与事件客体距离较近时，基线模型和本文方法效果接近。但随着其距离的增大，基线模型的F1值下降较为明显，本文方法在主客体之间距离较远时仍然有较好的表现。充分证明了本文方法对缓解远距离依赖问题的有效性。

5.4.5 实例分析

为更加直观地体现本文模型相较于基线模型的提升，本小节将展示具有代表性的3个文本以及抽取结果的实例。如表4所示，其中文本部分由于长度限制，仅展示含有要素的局部文本，省略上下文信息。

NO.	实例			
1	文本	...上海和吉林省是中国顶尖的汽车制造中心。最近，两地都实施了严格的疫情防控措施...		
	要素类型	事件主体	事件动作	事件客体
	金标答案	上海和吉林省	实施	严格的疫情防控措施
	基线模型	上海和吉林省	实施	严格的疫情
	本文模型	上海和吉林省	实施	严格的疫情防控措施
2	文本	...每一台进入冬奥会的服务机器人都经过严格的筛选评测，正是有了这些机器人的加入，才使得参会嘉宾有了别样新体验...		
	要素类型	事件主体	事件动作	事件客体
	金标答案	这些机器人	加入	冬奥会
	基线模型	这些机器人	加入	None
	本文模型	这些机器人	加入	冬奥会
3	文本	...还有部分区域受疫情影响，春耕用肥运输受阻，化肥难以送到田间地头。...		
	要素类型	事件主体	事件动作	事件客体
	金标答案	疫情	影响	部分区域
	基线模型	部分区域	受疫情影响	None
	本文模型	疫情	影响	部分区域

Table 4: 金标及预测的实例

实例1与实例2中，基线模型与本文模型均能正确识别事件主体和事件动作，但在事件客体的抽取中，实例1的基线模型由于边界问题未能完全正确地抽取正确答案，而本文模型融入句法特征，能更精准地识别各要素，减少边界问题带来的错误；实例2的基线模型未能识别出事件客体，原因可能是识别的要素为诸如“冬奥会”的专有名词时，基线模型并不能很好地理解，进而导致识别错误或者识别不出答案，相较之下，本文模型经过句法语义增强，即使要素为专有名词，也能在句法特征的加持下正确识别。

实例3是该任务的一个难题，原因主要是被动语态导致事件主体与事件客体的混淆，并且被动语态在数据集中的占比较少无法充分学习该特征。而本文模型将成分句法树进行编码，即使是被动语态也能够很好地识别各要素的成分。

6 总结与展望

为了提高模型泛化能力和远距离依赖处理能力，本文提出了一种基于句法特征的事件要素抽取模型。在所提模型中，我们利用词性和句法标签来提升模型泛化能力，通过对成分句法树建模来改进远距离依赖的处理能力。实验结果表明，本文提出的抽取模型能够充分利用成分句法信息，提高了系统性能。

同时，我们的工作也存在一些需要改进的地方。例如，本文使用的预训练语言模型是在大量语料上进行无监督训练，只包含文本的浅层语义信息。若在预训练阶段就融入句法信息等的深层特征，抽取性能可能会有进一步的提升。这也将是我们未来研究的一个重点。

第二十二届中国计算语言学大会论文集，第196页-第207页，哈尔滨，中国，2023年8月3日至5日。

参考文献

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

James Cross and Liang Huang. 2016. Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1–11.

Nan Ding, Chunming Hu, Kai Sun, Samuel Mensah, and Richong Zhang. 2022. Explicit role interaction network for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3475–3485.

Han He and Jinho D Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Andrew Hsi, Yiming Yang, Jaime G Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1201–1210.

Shaohua Jiang, Shan Zhao, Kai Hou, Yang Liu, Li Zhang, et al. 2019. A bert-bilstm-crf model for chinese electronic medical records named entity recognition. In *2019 12th international conference on intelligent computation technology and automation (ICICTA)*, pages 166–169. IEEE.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Junyi Li, Xiaobing Zhou, Yuhang Wu, and Bin Wang. 2019. Ynu-junyi in bionlp-ost 2019: Using cnn-lstm model with embeddings for seedev binary event extraction. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 110–114.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified mrc framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859.

Qian Li, Hao Peng, Jianxin Li, Jia Wu, Yuanxing Ning, Lihong Wang, S Yu Philip, and Zheng Wang. 2021. Reinforcement learning-based dialogue guided event extraction to exploit argument relations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:520–533.

Rahul VSS Patchigolla, Sunil Sahu, and Ashish Anand. 2017. Biomedical event trigger identification using bidirectional recurrent neural network based models. In *BioNLP 2017*, pages 316–321.

Mengtao Sun, Qiang Yang, Hao Wang, Mark Pasquine, and Ibrahim A Hameed. 2022. Learning the morphological and syntactic grammars for named entity recognition. *Information*, 13(2):49.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Shuheng Tao, Yi Chen, and Jiping Wang. 2022. A global pointer based entity relation extraction method for chinese pulmonary nodule medical records. In *2022 IEEE 16th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–8. IEEE.

Amir Pouran Ben Veyseh, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Graph transformer networks with syntactic and semantic structures for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3651–3661.

Fei Xia. 2000. The part-of-speech tagging guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series*, page 38.

Jing Xu, Dandan Song, Siu Cheung Hui, Fei Li, and Hao Wang. 2023. Multi-view entity type overdependency reduction for event argument extraction. *Knowledge-Based Systems*, 265:110375.

Nianwen Xue, Fei Xia, Shizhe Huang, and Anthony Kroch. 2000. The bracketing guidelines for the penn chinese treebank (3.0). *IRCS Technical Reports Series*, page 39.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR*, abs/1409.2329.

Hongkuan Zhang, Hui Song, Shuyi Wang, and Bo Xu. 2020. 基于bert 的端到端中文篇章事件抽取(a bert-based end-to-end model for chinese document-level event extraction). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 390–401.

Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 414–419.

成于思and 施云涛. 2020. 融合词典特征的bi-lstm-wcrf 中文人名识别. 中文信息学报, 34(4):69–76.

镇宇, 蒋盛益, 张礼明, and 包睿. 2019. 基于多特征bi-lstm-crf 的影评人名识别研究. 中文信息学报, 33(3):94–101.

蔺志, 李原, and 王庆林. 2022. 基于bert 改进的文化活动事件论元抽取研究. 中文信息学报, 36(12):115–122.

相似音节增强的越汉跨语言实体消歧方法

李裕娟[1,2]，宋燃[1,2]，毛存礼[*1,2]，黄于欣[1,2]，高盛祥[1,2]，陆杉[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1064406374@qq.com, song_ran@163.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, lushan88d@163.com

摘要

跨语言实体消歧是在源语言句子中找到目标语言相对应的实体，对跨语言自然语言处理任务有重要支撑。现有跨语言实体消歧方法在资源丰富的语言上能得到较好的效果，但在资源稀缺的语种上效果不佳，其中越南语-汉语就是一对典型的低资源语言；另一方面，汉语和越南语是非同源语言存在较大差异，跨语言表征困难；因此现有的方法很难适用于越南语-汉语的实体消歧。事实上，汉语和越南语具有相似的音节特点，能够增强越-汉跨语言的实体表示。为更好的融合音节特征，我们提出相似音节增强的越汉跨语言实体消歧方法，缓解了越南语-汉语数据稀缺和语言差异导致性能不佳。实验表明，所提出方法优于现有的实体消歧方法，在R@1指标下提升了5.63%。

关键词： 实体消歧；音节相似性；越汉跨语言

Similar syllable enhanced cross-lingual entity disambiguation for Vietnamese-Chinese

Yujuan Li[1,2], Ran Song[1,2], Cunli Mao[*1,2], Yuxin Huang[1,2], Shengxiang Gao[1,2], Shan Lu[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
1064406374@qq.com, song_ran@163.com, maocunli@163.com
huangyuxin2004@163.com, gaoshengxiang.yn@foxmail.com, lushan88d@163.com

Abstract

Cross-lingual entity disambiguation is to find the entity corresponding to the target language in the source language sentence, which is an important support for cross-language natural language processing tasks. Existing cross-lingual entity disambiguation methods can achieve good results in languages with abundant resources, but poor results in languages with scarce resources. Among them, Vietnamese-Chinese is a typical pair of low-resource languages; on the other hand, Chinese and Vietnamese are non-cognate languages with large differences, and cross-lingual representation is difficult; therefore, existing methods are difficult to apply to Vietnamese-Chinese entity disambiguation. In fact, Chinese and Vietnamese share similar syllable features that can enhance entity representation across Vietnamese-Chinese cross-languages. In order to better integrate syllable features, we propose a similar syllable-enhanced Vietnamese-Chinese cross-language entity disambiguation method, which alleviates the poor performance caused by the scarcity of Vietnamese-Chinese data and language differences. Experiments show that the proposed method is superior to existing entity disambiguation methods, and improves by 5.63% under the R@1 index.

*毛存礼(通讯作者):maocunli@163.com

第二十二届中国计算语言学大会论文集，第208页-第219页，哈尔滨，中国，2023年8月3日至5日。

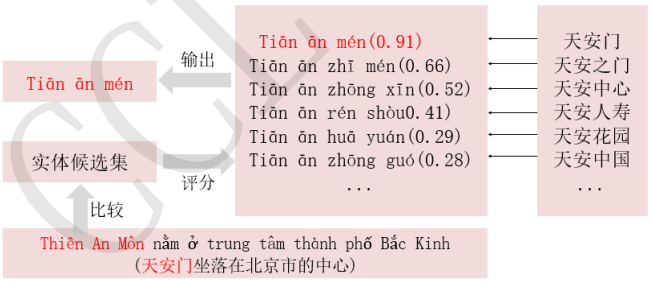
Keywords: Entity Disambiguation , Syllable similarity , Cross-lingual for vietnamese-chinese

1 引言

实体消歧（Entity Disambiguation，ED）的目的是将非结构化文本中提到的实体链接到相应的知识库实体。ED的挑战在于待处理实体的歧义性,例如在文本中提到的“世界杯”和知识库中的实体(如“FIFA世界杯”和“橄榄球世界杯”)存在歧义。ED模型通过建模提及实体的局部信息和文本的全局语义信息，以确定对应的目标实体。提高ED效果的关键在于有效地结合提及信息和上下文信息(Ganea and Hofmann, 2017; Guo and Barbosa, 2018)。

许多跨语言任务依靠多语言知识库提升其性能，如问答(Yang et al., 2017; Yang et al., 2018)、推荐(Cao et al., 2019)和信息抽取(Kumar, 2017)。跨语言实体消歧是够将原文本中提及的实体，在另一语言的知识库中进行匹配，能够为跨语言任务提供支持。目前，跨语言实体消歧主要依赖于多语言预训练模型，多语言预训练模型能够把源语言和目标语言的表示映射到同一语义空间下，以解决跨语言表示的问题并提升跨语言实体消歧效果。得益于预训练语言模型强大的表示能力，富资源语言上跨语言实体消歧已经达到很好的效果，如图1(a)所示，直接使用未经过预训练的跨语言模型计算提及实体和备选实体的相似度就能得到很好的效果。例如，对于提及实体“Washington”，可以在汉语中找到正确的对应实体“乔治·华盛顿”，并且具有高置信度，即目标实体与其他实体的得分差异度较大。

(a) 英中实体消歧　　(b) 越汉实体消歧



(c) 越汉实体消歧

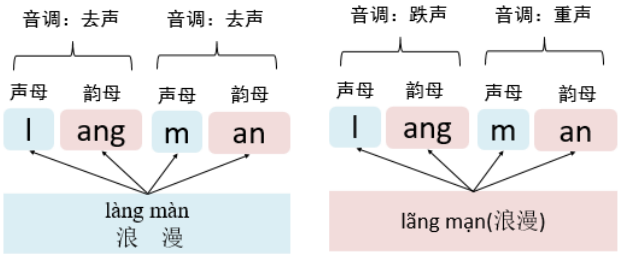
Figure 1: 实体消歧效果对比示意图

然而，大多数多语言预训练模型在越南语上表现不佳。这是因为在M-BERT(Devlin et al., 2018)的训练数据中，英语的数据量是越南语的116倍，导致越南语的表示效果远不如富资源语言，如英语、等。如图1(b)所示，在当利用模型计算越南语实体(Thiên An Môn)和汉语备选实体之间的相似度时，目标实体未被正确匹配，并且不同实体之间的相似度差距较小，置信度较低。因此，依赖于多语言预训练模型进行越汉跨语言实体消歧无法取得良好的效果。分析越南语和汉语之间的关系，并找到越南语-汉语之间的有效特征，能够改善越汉实体消歧的效果。

国家自然科学基金(U21B2027,61972186,62166023,62266028);云南省科技重大专项(202103AA080015,202202AD080003,202202AD080004);云南省基础研究计划项目(202301AS070047,202301AT070471)

事实上，越南语自古受到汉字文化的深远影响(李靖, 2008)，所以越南语的书写、句法以及读音等很多方面仍然保留着许多汉语特色。越南语和汉语类似，构词绝大多数以单音节(或称字)为单位。和多数汉语言以及壮泰语言一样，越南语的音节可由声母、韵母、声调三部分构成(陈雪, 2011)。汉语和越南语的声母中都有爆破音(b、p、t、g、k)和摩擦音(f、s、h)，在韵母中越南语和汉语都有单元音(a、e、i、o、u)和复元音(a、iê、i等)。

如图2，比如“浪漫”这个词，在汉语中拼音为“làng màn”，其声母为“l”(清辅音)，韵母为“ang”(开合中元音+鼻音)，声调为第四声。在越南语中为“lãng mạn”，其声母为“l”(清辅音)，韵母为“ang”(中元音+鼻音)，声调为重声。



(a) 汉语　　(b) 越南语

Figure 2: 汉语拼音和越南语音节分析图

基于以上观察，可以发现越南语和汉语拼音具有相似的音节，但并非所有单词都具有完全相似的结构。例如，“广告”一词在汉语中的拼音为“guǎng gào”，在越南语中的拼音为“quảng cáo”，两者的字符串存在差异。因此，需要将字符串分割成更细粒度字符片段进行比较，以便逐一比较各个字符片段。通过更细粒度的分割，汉语中的“uǎng”字符串和越南语中的“uảng”字符串将显示出更高的相似度，从而提高越南语和汉语拼音之间的整体相似度。因此，将跨语言实体消歧中的汉语和越南语提及的实体的音节结构纳入考虑，可以提高实体消歧的效果。

为了更好融合音节特征并提升汉语实体消歧的性能，本文提出一种相似音节增强的越汉跨语言实体消歧方法(Similar syllable enhanced cross-lingual entity disambiguation for Vietnamese-Chinese，VCED)。如图1(c)所示，首先利用越南语提及实体与汉语实体拼音之间的结构相似性增强实体表示。由于越南语与拼音并不是严格对齐的，因此我们利用N-gram编码对越南语和拼音进行不同字符粒度的切分，并计算他们之间的相似度。然后，使用M-BERT预训练语言模型计算越南语文本与汉语实体描述的语义相似度。最后，综合字符和语义的相似度得到目标实体。实验结果表明，本文提出的方法在越南语-汉语的跨语言实体消歧任务上准确率显著提高。

简而言之，本文工作贡献如下：

(1) 本文分析了越南语与中文之间的语言特点，并找到了音节之间的相似性，并构建了越南语-汉语的跨语言实体消歧数据集，以支持后续实体消歧的研究。

(2) 提出了面向越南语-汉语的音节字符及文本语义表示方法，增强了实体的表示，从而提升越南语-汉语跨语言实体消歧模型的性能。

(3) 我们进行了广泛的实验，结果表明所提出方法可以提升了越南语-汉语实体消歧的性能，效果优于现有的实体消歧模型。

2 相关工作

实体消歧方法分为两类。一种是基于实体特征的实体消歧方法，根据实体和关系的语义特征、实体和关系的上下文特征以及实体出现的频率特征来消除实体歧义。另一种是基于神经网络的实体消歧方法，该方法是利用知识图中的图结构特征，使用神经网络模型进行端到端的实体消歧。

基于实体特征的实体消歧方法提出了一种实体相似度模型来度量模糊实体之间的差异。命名实体消歧系统(Daiber et al., 2013)主要依靠实体上下文相似性度量来消歧。Adjali et al. (2020)使用实体语义相似性、上下文相似性和提及概率来消除实体歧义。MCKR(Hu et al.,

2021)采用多层感知器提取缺失数据与观测数据的交互特征。

Barrena et al. (2015)通过融合来自不同数据源的实体、名称、文本和维基百科信息的概率模型，发现这些特征在实体消歧中具有明显的互补作用。Tsai and Roth (2016)通过用相应的实体标记替换每个实体提及来联合训练单词和实体的单语嵌入，再基于维基百科的语言间链接，使它们学习从多种语言到英语词嵌入空间的投影函数，最后，把源语言实体文本的上下文嵌入同样也投影到英语空间，并与英语实体嵌入进行相似度计算完成实体消歧。Zwicklbauer et al. (2016a)为实体消歧提供了一种实体语义嵌入表示模型，他们使用Word2Vec(Mikolov et al., 2013)方法嵌入实体，并在RDF图上使用随机游走方法构建实体序列。Bouarroudj et al. (2022)提出了一种短文本的命名实体识别，他们使用WordNet进行实体上下文扩展，再基于语义和句法度量对候选词进行排名从而进行实体消歧。Tam et al. (2022)提出了使用知识图谱正则化的条件掩码实体模型（CMEM-KG），其中上下文中的多个提及可以在一次前向传递中消歧。Liu et al. (2022)提出使用KG嵌入进行语义表解释和实体消歧，旨在在先前识别的实体注释之后添加语义消歧步骤，考虑整个列作为上下文，并使用图嵌入来捕获实体之间的潜在关系，以提高它们的消歧性。无论是基于聚类还是基于实体链接，实体与实体、实体与文本、文本与文本之间的相似度计算是实体消歧的核心问题。这些计算方法主要利用自然语言处理技术来提取实体的特征。这些方法虽然取得了较好的性能，但越南语特征可扩展性差，表示能力不足，在实体消歧中容易造成错误传播。

基于神经网络的实体消歧方法采用端到端的机制来提高实体消歧的准确性。除了实体和关系的特征之外，研究人员还利用知识图的图结构特征进一步提高实体消歧的效果。RSJoint(Geng et al., 2021)集成了卷积神经网络和递归神经网络来消除实体歧义和提取关系。Guo and Barbosa (2014)通过估计候选实体的Topic-sensitive PageRank值(Haveliwala, 2002)，结合知识图上的随机游走方法进行实体消歧。Alhelbawy and Gaizauskas (2014)使用基于图的方法进行联合实体消歧，该方法将文本中的所有实体表示为图中的节点，然后根据节点的PageRank值对其进行排序，并根据值的大小进行实体消歧。Singh et al. (2011)利用图形模型来消除文档之间的实体歧义。Zwicklbauer et al. (2016b)设计了一种利用上述实体知识图上的个性化PageRank值的集体消歧方法，该方法依赖于集体链接算法进行实体消歧。

研究人员还尝试使用深度学习方法来消除歧义，并取得了良好的效果。Ganea and Hofmann (2017)采用知识图中的实体嵌入，采用基于注意的方法获得嵌入向量，并考虑实体之间的相干性进行联合消歧。与依赖监督或启发式方法预测实体关系不同，Le and Titov (2018)将实体关系作为神经实体链接模型中的隐变量，以端到端的机制实现实体消歧。DeepType(Raiman and Raiman, 2018)通过将一个符号特征和一个典型特征结合到神经网络推理中，解决了实体消歧问题。研究者设想了一种类型模型，并利用它来限制网络的输出以适应结构特征。他们提出了一种两阶段的实体消歧算法，首先创建一个类型系统，然后用它来训练神经网络。Hu et al. (2020)和Grover and Leskovec (2016)都是使用图神经网络模型解决实体消歧的问题。Phan et al. (2017)提出了一种深度神经网络方法NeuPL来计算实体之间的语义相似度。NeuPL是第一个使用长短期记忆网络消除实体歧义的模型。但基于神经网络的方法的局限性在于缺乏很好的解释性，且越南语的知识图较为稀疏，使用此方法并不能是越汉实体消歧达到很好的效果。

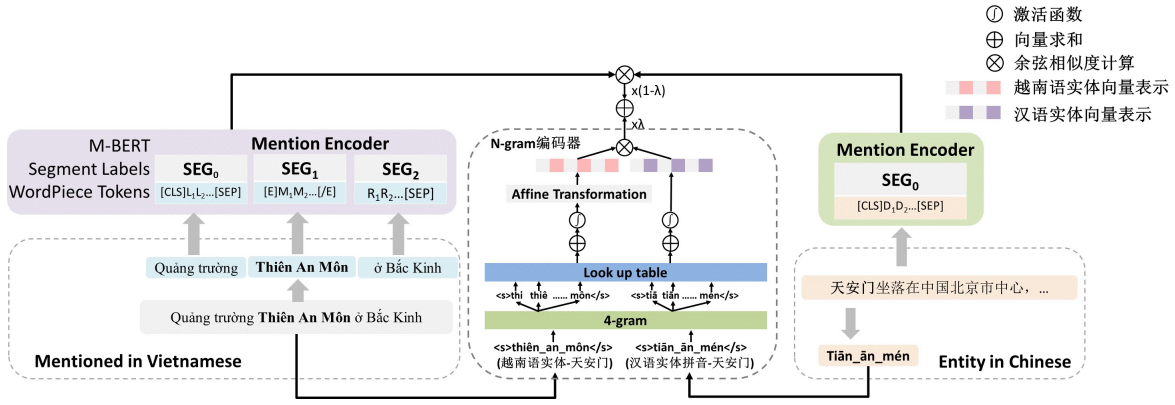


Figure 3: 相似音节增强的越汉跨语言实体消歧模型图

3 方法

我们方法的关键思想是融合越南语提及实体与汉语拼音之间的相似性和文本的相似性来计算实体自身之间的相似度。具体的**模型框架如图3所示**，主要由2个模块组成:(1)文本语义编码，主要是计算越南语实体提及文本和汉语实体描述的相似度；(2)音节字符编码，主要是用于计算越南语实体本身以及汉语实体本身的相似度。

3.1 文本语义编码

3.1.1 越南语文本编码

越南语文本编码是将越南语一段文本中字词序列和分句标签序列融合编码。给定模型输入的越南语文本表示T_{vi}，如图3中左下越南语"Quảng trường Thiên An Môn ở Bắc Kinh."和其中的越南语实体提及表示M_{vi}，如图3中"thiên_an_môn"。模型首先使用WordPiece分别切分越南语实体提及M_{vi}本身以及其所在文本T_{vi}的上下文文本得到对应的子词序列。再使用特殊的分隔符"[E]"和"[\E]"对实体提及和上下文文本进行标记，得到最终的子词序列，具体如图3所示。在送入多语言预训练模型M-BERT前，使用分句标签来对M_{vi}以及上下文文本进行标记，其中上文子词序列对应标签为"0"，实体提及子词序列对应标签为"1"，下文子词序列对应标签为"2"，得到最终的分句标签序列。最后把所得子词序列以及分句标签送入M-BERT模型得到最终的向量表示为公式1:

$$M_{vi}^{c} = M\text{-}BERT(subs_{vi} + seg_{vi}) \tag{1}$$

其中$subs_{vi}$为最终的子词序列，seg_{vi}为最终的分句标签序列。M_{vi}^{c}是M_{vi}的上下文向量表示。

3.1.2 汉语文本编码

汉语文本编码是将一段汉语文本中字词序列和分句标签序列融合编码。给定模型输入的汉语文本表示T_{zh}，该文本为实体在维基百科页面下介绍的第一句话，如图3中右下汉语"天安门坐落在北中国北京市中心..."和其中的汉语实体拼音字符表示E_{zh}。首先使用WordPiece切分该文本得到子词序列$subs_{zh}$。再使用分句标签"0"标记该子词序列得到分句标签序列seg_{zh}后，通过M-BERT得到汉语实体E_{zh} 的上下文向量表示E_{vi}^{c}，具体公式如2所示:

$$E_{zh}^{c} = M\text{-}BERT(subs_{zh} + seg_{zh}) \tag{2}$$

3.2 音节字符编码

N-gram能有效获取上下文的信息，因此本文通过使用N-gram模型丰富汉语以及越南语实体的编码向量表示。首先，将字符长度为m的字符串表示为字符序列$X = [x_1, x_2, ..., x_m]$,其中字符包括空格字符（空格以"_"代替）以及特殊的起始符和结束符，例如图3中越南语实体"thiên_an_môn"表示为字符序列["<s>"，"t"，"h",...,"n"，"</s>"]。为了获取字符串的多粒度表示，首先对越南语实体提及和汉语实体拼音表示进行多粒度4元切分，例如越南语实体"thiên_an_môn"，切分为"<s>thi"、"thiê"、"hiên"、"iên_"、"ên_a"、"n_an"、"_an_"、"an_m"、"n_mô"、"_môn"以及"môn</s>"共计11个4元组。

给定x_i^j来表示从位置i到位置j的子序列，例如$x_i^j = [x_i, x_{i+1}, ..., x_j]$。字符串$X$的4元嵌入表示$V_n$具体公示如3所示:

$$V_n = \left(\sum_{i=1}^{m} \mathrm{II}(x_i^j \in V) W_{x_i^j} \right) (j = i + n - 1, j \leq m + 2) \tag{3}$$

其中n是预定义的n-gram窗口大小，这里n=4。$m+2$表示字符串长度加上起始符和结束符的长度。$W \in R^{|V| \times d}$是嵌入矩阵，$W_{x_i^j} \in R^d$是x_i^j的向量嵌入表示。V是训练数据集中字符的所有4元组合，II()是指示函数，就是如果一个4元字符组合不在V中则丢弃。

给定汉语实体拼音表示E_{zh}^{py}以及越南语实体提及表示M_{vi}^{n}，通过上述过程得到最终的汉语实体以及越南语实体提及向量表示E_{zh}^{ngram}和E_{vi}^{ngram}，具体公式如4和5所示:

$$E_{zh}^{ngram} = N\text{-}gram(E_{zh}^{py}) \tag{4}$$

$$E_{vi}^{ngram} = N\text{-}gram(M_{vi}^{n}) \tag{5}$$

3.3 相似度计算

最后，将越南语文本提及上下文和汉语实体描述进行相似度计算以及越南语提及实体和汉语实体拼音进行相似度计算。通过以上过程获取得到越南语实体提及的上下文向量表示M_{vi}和自身向量表示E_{vi}^{ngram}，汉语实体描述文本向量表示E_{zh}^{c}和自身向量表示E_{zh}^{ngram}。为了综合考虑上下文全局信息以及自身局部信息，分别对全局信息向量M_{vi}^{c}和E_{zh}^{c}以及局部信息向量E_{vi}^{ngram}和E_{zh}^{ngram}进行相似度计算，具体公式如6和7所示：

$$sim_c = \frac{(M_{vi}^{c} E_{zh}^{c})^T}{||M_{vi}^{c}|| \times ||E_{zh}^{c}||} \tag{6}$$

$$sim_n = \frac{(E_{vi}^{ngram} E_{zh}^{ngram})^T}{||E_{vi}^{ngram}|| \times ||E_{zh}^{ngram}||} \tag{7}$$

在获得全局信息相似度得分sim_c以及局部信息相似度得分sim_n后，把两组相似度得分通过超参数λ进行线性组合，得到包含了上下文信息以及实体自身信息的综合相似度sim_{comb}，具体计算公式如8所示：

$$sim_{comb} = \lambda sim_c + (1-\lambda) sim_n \tag{8}$$

3.4 损失函数

本文选用HingeEmbeddingLoss损失来做为训练所用的目标损失函数。模型训练Loss的计算公式如9所示：

$$Loss = \begin{cases} sim_{comb}^{i} & if y_i = 1 \\ max(0, margin - sim_{comb}^{i}) & if y_i = 0 \end{cases} \tag{9}$$

其中，y_i表示第i对数据的真实标签，若越南语实体提及与中文实体是对齐实体则为1，否则为0。sim_{comb}^{i}表示第i对数据之间的距离，$margin$为两元素之间间隔距离允许的边界值，模型中设置为1。

4 实验

为了验证所提出方法的有效性，首先介绍自构数据集(4.1)和实验结果评估方法(4.2)，然后在常规的实验设置(4.3)下与其他实体消歧方法进行比较，还进行了消融实验(4.4)，以评估不同模块对实体消歧效果的影响和不同预训练语言模型对文本语义模块的影响。最后通过样例分析(4.5)进一步说明了本文提出的方法优于其他模型。

4.1 数据集

WikiANN(Pan et al., 2017)[0]是基于维基百科文章的跨语言命名实体识别和实体链接数据集，其中越南语数据共包含110,535条。我们基于WiKiANN以及维基百科[1]中的多语言链接来构建越汉跨语言实体消歧越南语部分的数据集，之后，再通过维基百科获取越南语相对应的汉语实体以及对应的描述。

数据集	总数量	正例数量	越南语实体数量	未在训练集出现过数量
训练集	60,000	10,000	6,303	—
验证集	10,000	2,000	1,515	755
测试集	11,825	2,365	1,737	801

Table 1: 越汉跨语言实体消歧数据集

数据集共计正例14,365条，划分为训练集、验证集和测试集分别为10,000条、2,000条以及2,365条。为了保证模型能学习到更好的知识，防止模型过拟合，增强模型鲁棒性，故采用交

[0]https://elisa-ie.github.io/wikiann/
[1]https://zh.wikipedia.org/wiki/

错匹配的方式构造出负例数据，其中正负比例为1:4，最终训练集、验证集和测试集的数据数量如表1所示，训练集、验证集和测试集中分别存在6,303、1,515和1,737个不相同的越南语实体提及。而验证集和测试集中分别有755、801个越南语实体提及没有出现在训练集中，分别占比为49.83%和46.11%。

4.2 评价指标

本文的实验选用召回率R@1来对实验结果进行评价，因为对于最终的系统比较，标准的做法是使用top检索实体的准确性(R@1)，R@1的具体计算公式如公式10所示：

$$R@1 = \frac{\sum_{i}^{PT} ||C(M_i^{vi})||}{PT} \tag{10}$$

PT为测试集中正例的数量，$C(M_i^{vi})$为该方法在测试集中越南语实体提及M_i^{vi}的正例和负例的结果，若全部正确则取1，否则取0。

Parameter	Value
BERT embedding size	768
BERT model	M-BERT
n-gram embedding size	100
dropout	0.3
Optimizer	Adam
Learning rate	5e-5
n-gram size	4
layers	9-12

Table 2: 模型主要超参数

4.3 参数设置

本文实验使用Pytorch1.7.1版本进行，选用Adam作为优化器，Batch大小设置为32，Epoch设置为20。学习率调整采用等间隔调整策略，每训练10轮调整学习率为当前学习率的百分之十。为了防止实验过拟合，在部分地方使用Dropout技术.所有实验均在一张RTX3090 Ti上训练，实验的主要超参数指标如表2所示。

4.4 实验结果及分析

4.4.1 对比实验

我们将$VCED$与3个基线模型进行了比较，3个基线都是基于实体特征的实体消歧方法，我们采用召回率R@1作为评价指标，R@1越高表示性能越好。

$WikiME$(Tsai and Roth, 2016)在2016年提出的通过联合训练单词和维基百科标题的多语言嵌入来将非英语文档中的实体提及内容与英语维基百科条目联系起来的模型。

$XELMS_{joint}$(Upadhyay et al., 2018)在2018年提出的一个结合多种语言监督的实体消歧方法。

$ModelF$(Botha et al., 2020)提出训练的双编码器模型，在先前工作的基础上改进了特征表示、负挖掘和辅助实体配对，以提高在低资源条件下实体消歧模型的性能。

通过实验结果可以发现，本文的方法可以有效提升越汉跨语言实体消歧模型的性能，效果优于其他几种对比方法。表3显示出$VCED$和$ModelF$均优于对比方法中的$WikiME$和$XELMS_{joint}$模型，主要是$VCED$和$ModelF$都使用了基于大规模语料训练的预训练多语言模型M-BERT，因此模型初始时就富含汉语和越南语的丰富语言知识，针对

模型	R@1(%)	未见实体R@1(%)
$WikiME$	35.05	16.12
$XELMS_{joint}$	38.69	17.67
$ModelF$	42.22	19.34
VCED	**47.85**	**21.96**

Table 3: 不同模型实验结果

低资源语言的模型训练来说预训练语言模型能提高模型性能。我们的模型在R@1指标下高于$ModelF$模型5.63%，虽然$ModelF$也使用M-BERT预训练语言模型充分地关注上下文语义信息，得到了更好的上下文特征表示向量，但在实体消歧任务中，除了上下文信息以外，实体自身的表示也同样重要，本文的方法加入了实体的音节字符信息，将汉语和越南语实体自身各粒度n元组进行编码从而丰富了实体自身的局部信息，提高了模型效果。

我们还在未见实体上做了模型性能测试，实验结果表明，VCED模型在未见实体上的性能也高于其他几个基线模型，这是因为预训练语言模型在没有见过实体的上下文表示比较弱，而相比较于Model F模型VCED模型加入了实体自身的表示，提高了模型的效果。

4.4.2 消融实验

现在预训练语言模型多种多样，我们对比了召回率与不同预训练语言模型之间的关系，在M-BERT，LaBSE(Feng et al., 2020)，SBERT(Reimers and Gurevych, 2019)预训练语言模型上做了实验。然后为了测试我们方法在两个模块上对实验结果影响，我们将音节字符编码模块和文本语义编码模块分别进行实验。

从表4可以看出，我们的模型使用的M-BERT预训练语言模型效果高于其他预训练语言模型，这是因为M-BERT富含了丰富的语义信息，所以使用含有丰富的语义信息的预训练语言模型在一定程度上可以提升实体消歧模型的性能。从单独用音节字符编码模块和单独用文本语义编码模块的实验结果来看，单独考虑文本语义信息或单独考虑音节字符信息去提升实体消歧模型性能是远远不够的，需要同时考虑文本语义信息和音节字符信息才能有效提升实体消歧的效果。

预训练语言模型	R@1(%)
$VCED_{onlyM-BERT}$	15.35
$VCED_{onlyN-gram}$	46.13
$VCED_{LaBSE}$	46.19
$VCED_{SBERT}$	45.53
VCED	**47.85**

Table 4: 消融实验

4.4.3 不同λ值实验

为了研究我们方法的召回率与超参数λ的大小之间关系，实验在超参数λ设置为0.1，0.3，0.5，0.7以及0.9时的准确率。实验结果如表5所示:

从表5可以看出超参数λ的大小对实验结果有着显著的影响。通过上述实验可以得知，超参数λ从0.1增长到0.5时，实验结果的召回率随着其增长也增长，λ到0.5时实验召回率达到最高的47.85%，但当λ从0.5继续增长到0.9时，实验结果的召回率随着其增长开始逐渐降低，此组实验证明了实体本身的信息和实体上下文信息在实体消歧任务中同样重要。λ从0.1到0.5，召回率提升了13.52%，而λ从0.5到0.9，召回率只下降了5.98%，从此结果可以看出上下文全局信息的

λ	R@1(%)
0.1	34.33
0.3	40.58
0.7	43.46
0.9	41.87
0.5	**47.85**

Table 5: 不同 λ 值实验结果

重要程度要高于实体自身，只有合理考虑上下文信息和实体自身信息才能时模型性能达到最好。

4.4.4 音节相似比例实验

为了验证测试集中音节相似比例对模型效果的影响，我们将测试集中数据的音节相似分为4种比例([0-0.25)，[0.25-0.5)，[0.5-0.75)，[0.75-1])，分别测试音节相似比例对模型准确率的影响，其中，值越小表示两个音节之间字符串的相似度越接近。由图4可以看出，随着音节相似度的降低，模型的准确率在不断的提高，这是因为在训练集中音节相似的数据占比较少，训练时模型对于音节相似的实体没有得到足够的学习所以模型的准确率较低，而音节不相似时模型的准确率较高是因为我们提出了4-gram编码器，对于较长的实体音节有更多组合的可能,丰富了音节的表示，从而提高了模型的准确率。

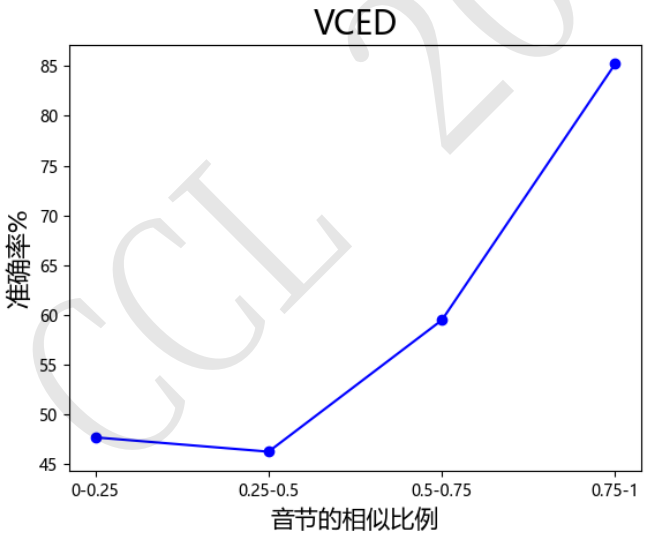


Figure 4: 不同比例的相似音节对模型准确率的影响

4.4.5 数据正负比例实验

为了验证我们方法的有效性并研究在构造数据集时正例与负例比例对实验结果的影响，我们分别在数据集中正负比例分别在1:2，1:3，1:4以及1:5时的结果进行对比，记录每组实验的R@1值，实验结果如表6所示。

从表6中可以看出，数据集中正负比例为1:2时，实验结果的召回率最低只有42.41%。随着负例数量上升到1:4时，实验的召回率随着负例的增加而增加，达到最高的47.85%，而当正负比例到1:5时，实验效果开始下降，可知，在原始数据集只有正例且正例数量固定的情况下，适当增加负例能帮助模型学习到更多的知识，能增加模型的鲁棒性，防止模型过拟合，但当负例达到某个数量时，再靠增加负例来使数据集数量变大提升模型效果是不可行的，因为负利数量过

正负比例	R@1(%)
1:2	42.41
1:3	45.88
1:5	46.13
1:4	**47.85**

Table 6: 不同正负比例实验结果

大会影响模型对正例的敏感度降低模型整体性能，还会无谓的增加训练时间。

4.5 样例分析

我们给出了越南语-汉语实体消歧的实例。基于实体特征的实体消歧模型在越汉实体消歧时，会出现目标实体未被找到的现象。如表7，越南语文本中的实体为“Khonkaen”，由于越南语是低资源语言这一事实，没有较多的文本语义信息，会导致实体消歧模型未找到汉语中相对应的实体“孔敬市”，而利用文本语义信息和音节字符信息结合，有效的提高了实体消歧的准确率。此外，仅仅基于实体的文本语义的方法($ModelF$)无法较好的识别语义信息较弱的越南语实体，而本文提出的方法结合了文本语义信息和音节字符信息，缓解了越南语-汉语数据稀缺导致性能不佳的问题，提高了实体消歧的准确率。第三行是VCED模型top1召回失败样例，原因是越南文本中出现了将军(Tướng)这个词，且提及实体和对应实体“阮知方(ruan zhi fang)”的音节至多只有一个字符是相似的，因此错误的将“阮知方军团”判断成消歧正确的实体。

越南语文本	提及实体	$ModelF$	$VCED$
Sau đó Ratchasuphawadi chuyển tới **Khonkaen**	Khonkaen	孔敬,孔敬大学, 孔敬总领事馆, **孔敬市**,孔敬机场	孔敬市, 孔敬总领事馆,孔敬, 孔敬大学,孔敬机场
Bà chết cùng ngày với **Lê Cung Hoàng**	Lê Cung Hoàng	刘恭煌,黎椿,**黎恭皇**	黎恭皇,黎椿,刘恭煌
Tướng **Nguyễn Tri Phương** bị trọng thương	Nguyễn Tri Phương	**阮知方军团**,阮知方, 阮知方路,阮知方街道, 阮知方军团酒店	**阮知方军团**,阮知方, 阮知方路,阮知方街道, 阮知方军团酒店

Table 7: 越汉跨语言实体消歧样例分析

5 结束语

针对现有实体消歧模型因越汉数据稀缺导致性能不佳的问题，本文提出了一个新的越汉实体消歧模型(VCED)，利用越南语和汉语文本语义信息和音节字符信息增强实体的表示，从而拉近两种语言实体在向量空间上的距离。实验结果表明本文方法的有效性和优越性，在相同的数据集下比其他实体消歧模型提升了5.63%。在之后的工作中，我们考虑融入知识图的图结构特征以提高实体消歧的效果。

参考文献

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *Advances in Information Retrieval: 42nd European Conference on IR*

Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I, pages 463–478. Springer.

Ayman Alhelbawy and Robert Gaizauskas. 2014. Graph ranking for collective named entity disambiguation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 75–80.

Ander Barrena, Aitor Soroa, and Eneko Agirre. 2015. Combining mention context and hyperlinks from wikipedia for named entity disambiguation. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 101–105.

Jan A Botha, Zifei Shan, and Daniel Gillick. 2020. Entity linking in 100 languages. *arXiv preprint arXiv:2011.02690*.

Wissem Bouarroudj, Zizette Boufaida, and Ladjel Bellatreche. 2022. Named entity disambiguation in short texts over knowledge graphs. *Knowledge and Information Systems*, 64(2):325–351.

Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *The world wide web conference*, pages 151–161.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th international conference on semantic systems*, pages 121–124.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*.

Zhiqiang Geng, Yanhui Zhang, and Yongming Han. 2021. Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, 429:132–140.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Zhaochen Guo and Denilson Barbosa. 2014. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 499–508.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Taher H Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526.

Linmei Hu, Jiayu Ding, Chuan Shi, Chao Shao, and Shaohua Li. 2020. Graph neural entity disambiguation. *Knowledge-Based Systems*, 195:105620.

Xuan Hu, Yongming Han, and Zhiqiang Geng. 2021. A novel matrix completion model based on the multi-layer perceptron integrating kernel regularization. *IEEE Access*, 9:67042–67050.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645*.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. *arXiv preprint arXiv:1804.10637*.

Jixiong Liu, Viet-Phi Huynh, Yoan Chabot, and Raphael Troncy. 2022. Radar station: Using kg embeddings for semantic table interpretation and entity disambiguation. In *The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings*, pages 498–515. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Minh C Phan, Aixin Sun, Yi Tay, Jialong Han, and Chenliang Li. 2017. Neupl: Attention-based semantic matching and pair-linking for entity disambiguation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1667–1676.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models.

Zhi-Rui Tam, Yi-Lun Wu, and Hong-Han Shuai. 2022. Improving entity disambiguation using knowledge graph regularization. In *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part I*, pages 341–353. Springer.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. *arXiv preprint arXiv:1809.07657*.

Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. 2017. Efficiently answering technical questions—a knowledge graph approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016a. Doser-a knowledge-base-agnostic framework for entity disambiguation using semantic embeddings. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Proceedings 13*, pages 182–198. Springer.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016b. Robust and collective entity disambiguation through semantic embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 425–434.

李靖. 2008. 汉语对越南语的影响. 时代人物, (7):120–121.

陈雪. 2011. 汉语拼音与越南语音节结构之对比研究. 新课程学习*(*下*)*.

英汉动物词的认知属性计量研究

华玲,李斌,冯敏萱,匡海波
南京师范大学文学院/ 南京210023
南京师范大学语言大数据与计算人文研究中心/ 南京210023
hualing0411@163.com,libin.njnu@gmail.com,
Fennel_2006@163.com，kuanghaibo.nlp@gmail.com

摘要

动物词承载了大量人类社会认知映射，不同民族对于同一个词的认知有所异同。通过隐喻研究动物词认知差异是近年来十分流行的趋势，反映人们对词语认知印象的认知属性就是一个简捷的切入口。本文选择《中华传统文化名词认知属性库》中的54种动物，借助中英文认知属性数据库，对比分析英汉语言中的认知属性差异。文章发现动物词的英汉认知属性之间有明显差异，且差异更多表现在主观属性上，并发现了中英文中动物词认知属性的整体异同。

关键词： 认知属性；英语；隐喻；动物词

Quantitative studies of cognitive attributes of English and Chinese animal words

Ling Hua,Bin Li,Minxuan Feng,Haibo Kuang
School of Literature, Nanjing Normal University / Nanjing 210023
Center of Language Big Data and Computational Humanities,
Nanjing Normal University / Nanjing 210023
hualing0411@163.com,libin.njnu@gmail.com,
Fennel_2006@163.com，kuanghaibo.nlp@gmail.com

Abstract

Animal words carry a large number of cognitive mappings of human society, and different ethnic groups have similarities and differences in cognition of the same word. The study of cognitive differences in animal words through metaphor is a very popular trend in recent years, and reflecting the cognitive attributes of people's cognitive impressions of words is a simple entry point. In this paper, 54 animals in the "Cognitive Attribute Database of Traditional Chinese Cultural Terms" are selected and the cognitive attribute differences in English and Chinese languages are compared and analyzed with the help of Chinese and English cognitive attribute databases. This paper finds that there are obvious differences between the cognitive attributes of animal words in English and Chinese, and the differences are more manifested in subjective attributes, and the overall similarities and differences between the cognitive attributes of animal words in Chinese and English are found.

Keywords: Cognitive Attributes , English , Animal Words , Metaphor

1 引言

基金项目:江苏省社科基金项目（20JYB004）

不同民族由于宗教、社会、历史、文化等因素，对于同一个概念的认知感受不尽相同。例如，对于“蛇”这个词，冰冷、粘腻、修长、冷漠等是各民族共有的认知印象，汉语中由于神话传说中的蛇精形象，蛇有妩媚、柔美、妖媚等意义，而英语中的蛇则有高贵、神秘的意义，两种语言中的认知感受有所差异。这种差异在跨文化交际，尤其在翻译和对外汉语教学中，会造成理解上的困难和障碍。而动物词在文化中承载着格外浓厚的联想隐喻意义，作为人类出生以后接触时间最久、认识最深刻、交互最多的活体种群，很自然地成为了人类情感和认知投射的对象。研究动物词在不同语言中表现出的认知异同，能够更高效、更直观、更丰富地展现出两种语言文化的对比碰撞，为跨语言交际提供补充，例如西方读者或许很难理解《白蛇传》中的蛇精角色，因其缺乏汉语语境下对蛇“柔美、娇媚”的认知。分析这种差异能够为理解双方文化、厘清跨语言交际中文化附加义带来的障碍提供帮助。

本文以认知语言学中的隐喻理论为依据，以动物词作为对象，利用英汉认知属性库中相关数据，对动物词在不同语言中的认知属性对比分析。

2 研究综述

2.1 隐喻研究

最早对隐喻作出系统性论述的是亚里士多德（Aristotle），他认为隐喻是“把一个本来是描述其他事物的词转移到另一个词上，包括从属到种、从种类到属类、从一个种类到另一种类或者通过类推。”之后的Quintilian（1927）、Richards（1981）在此前基础上深入讨论隐喻本质，但对隐喻本质的看法依旧囿于符号手段中，直到Lakoff和Johnson（1980）在Metaphors We Live By中提出全新的隐喻观，将隐喻看作概念系统中的跨域映射，是从始源域到目标域的映射：“隐喻在日常生活中无处不在，它不仅存在于我们的语言中，而且还存在于我们思维和行动中。”[0]Lakoff认为隐喻是在两个处于不同认知域的概念之间的映射，因为人们总是习惯借用熟悉的、具体的、明确的概念来认知不熟悉的、抽象的、模糊的概念。这一观点颠覆了以往将隐喻看作“用法”的论调，使隐喻研究从属于认知科学。

2.2 动物词的隐喻研究

我国对动物词的语际差异研究始于上世纪九十年代，许高渝（1991）、喻云根（1992）等人对中英、中俄等语言中动物词比较分析，以应用于解决文学翻译中的问题。此后研究汉语中动物词与其他语言中动物词语义差异逐渐成为热点问题，从英汉，到汉语与法语、维吾尔族语、韩语、俄语、日语等多种语言的对比分析，动物词语义的跨语言差异研究越来越成熟深入。早期的动物词对比分析多以质性分析为主。研究者举出某些代表性动物词作具体的语义分析，将详细的中英文语义对比得出差异结论，并附以对差异成因的猜测。

约2007年开始，关于动物词的隐喻研究在国内开始风靡，研究者引入认知语言学相关理论，从隐喻角度分析动物词负载的与文化、认知、社会相关的语义成分，动物词的跨语言对比有了一个全新的理论角度，如李子鹤和苏立昌（2007）等。

2.3 认知属性

前文提到的动物词研究以收集语料-分析语料的定性研究为主，定量研究较少见。语义中与文化、社会、社团认知、宗教历史等因素有关的成分，在认知语言学派中没有界限分明的划分，而作为隐喻研究框架下的“隐喻特征”（metaphor features）存在（Giora，1997；Veale，2007）。隐喻句中，联系本体和喻体的特征，通常反映了喻体词义中的隐喻知识，这种知识在传统语义学中被称作词的附加义，被Giora和Veale称为“隐喻属性”[1]，被陈小荷（2005）称为“显著特征”[2]，认为“显著”是指认知上显著，“特征”表示事物的独特之处，表现为属性和属性值的特异性。李斌（2012）在认为，与附加义相对的理性义、在隐喻义之外的一般文化义等都应包括在隐喻句获取的知识中，为了更好地指称这些成分，李斌提出一个新概念“认知属性”来指代隐喻属性，将其解释为“在特定的语言中，语言使用者对词语代表的概念或实体的认知体验凝结到词义中的各种属性”，并自建语料数据库进行量化研究。

[0]Lakoff, Johnson:《Metaphors We Live By》, London: The university of Chicago press,2003,第8-23页

[1]Tony Veale，Yanfen Hao：《Learning to Understand Figurative Language: From Similes to Metaphors to Irony》，Proceedings of CogSci 2007［C］，Nashville，USA，2007，第216页

[2]李斌，陈家俊，陈小禾：《基于互联网的汉语认知属性获取及分析》，语言文字应用，2012（3）：第135-143页

目前，国内外对汉语隐喻方面的研究重心在隐喻句的机器识别和分类。在获取词语，特别是名词的隐喻属性上，传统的手工搭建电子词典已经被抛弃，目前学界多采用半自动方式构建数据库，方法较为多样。主要途径分为两类：基于规则的获取与基于统计的获取。基于统计获取，是提供足量人工标注语料供机器学习后，通过概率计算识别，如早年国外学者采用潜在语义分析技术获取认知属性（Kinstch，2000），杨芸（2008）利用词语的相关度计算来捕捉认知属性等。基于规则获取，指指定一定逻辑上的规则，命令机器按照指令获取固定数据，如利用英语的明喻句式"A is as B as C"[3]从互联网上抓取认知属性（Veale and Hao，2007），搭建了大型英语名词隐喻属性知识库Sardonicus。这一方法后被引入国内，利用汉语明喻句式获取汉语名词的显著特征（贾玉祥，2009），李斌（2012）在这一技术基础上改进方法，构建了大规模的汉语词语认知属性库Cognitive Bank，词库中包含了各属性的频次信息。同样使用此类方法建库的还有Google语料相关的英文名词认知属性库（陈琛，2013）。汪梦翔、饶琪（2017）等人建库时则同时采用了句型获取和词典惯用语两种方式，并且对认知属性的类别信息如外观、颜色、特质等也作了详细的采集。汪、饶团队分设隐喻属性和隐喻特征，从文本中既提取隐喻属性也提取隐喻特征，之后再将隐喻特征归类到各个属性下，具有较高的学术应用价值。从认知属性方面来看，国内对于认知属性的研究起步较晚，目前研究成果中有不少规模较大的自建语料数据库，算法技术方面突破很大，与之配套的应用型研究却不多；从动物词相关研究来看，近年动物词的跨文化对比分析已逐渐依赖于隐喻理论分析，但与数据库结合的量化研究依然寥寥。本文借助李斌（2012）研发的中文认知属性库Chinese CogBank和Veale与Hao（2007）研发的英文名词隐喻属性库Sardonicus中部分数据，辅以同样手段获取的近期网络语料数据，对中英动物词的隐喻义差异进行对比研究，填补该领域内的空白。

3 数据来源

从汉语常用词的角度出发，本文以"中华传统文化名词认知属性库"为依托，从中选取"动物"义类下的23个名词，去除了其中重复的词语、偏僻的词语（如华虫、宗彝等），同时补全了中华民族传统文化中尤为重要的动物图腾十二生肖，构建出合适的动物词研究对象共29个词形。鉴于引用英文数据库中只收录词的认知属性，本文词表仅限于动物词，不包括动物概念的词组。

在确定中文词的英文形式时，发现部分中英文中概念到形式不能形成一对一的映射，只选择一种译文形式会造成语料数据的遗漏。为提高语料的覆盖率，本文对同一个中文概念赋予多个英文对应词形，如"鸡——chicken（小鸡）、rooster（公鸡）、hen（母鸡）"。本文在确定英文形式时，以常用、广义为选择标准，参考维基百科与百度百科两种语言引擎的搜索情况来确定中英文词形。只有对于没有唯一确切的动物概念，本文才将多种准确、必要的英文词形纳入词表。补充多个英文词形后，势必要相对应地补充不同的中文翻译，因此在词表中，部分动物概念下延伸出了多个中英文词形。

考虑到英文中不同词对应的文化附加义偏侧不同，如swine更偏向肮脏、丑陋等，sow偏向懒惰、贪婪等，对词表设置"属"与"种"二层，对英文中有多种词形的概念分设多个下辖小类进行属性分析。一条词表记录为"中文属——中文种——英文种"，如"猪——阉猪——hog"为一个记录，本文词表中共计包括了29个属，54个小类。

需要声明的是，本研究中的词表以中文常用的动物词为源头，按照常用度的基准映射到相对应的不同英文词形，再从英文词形映射到多个中文译名。两种语言互为补充，但本质上以中文概念为源头，有可能导致最终结论包含更多中文语境信息，更具有中文特色，而遗漏部分英文中常用的概念。

4 数据处理与分析

本文中的认知属性数据均来自相关数据库，其中中文认知属性来自李斌（2012）研发的Chinese CogBank Version1.0，英文认知属性来自Veale&Hao（2007）研发的Sardonicus库。部分数据缺失严重的，按照Chinese CogBank（李斌，2012）中的相同技术步骤抓取了互联网语料作为补充，总计获取中文认知属性1095条，英文认知属性630条。

[3]Tony Veale，Yanfen Hao：《Learning to Understand Figurative Language: From Similes to Metaphors to Irony》，Proceedings of CogSci 2007［C］，Nashville，USA，2007，第52页

4.1 前期处理

获得相关数据后，首先将同概念在双语中的认知属性对比观察，辨别出语义相近与不同的属性，并加以标注。根据匡海波、李斌在跨语言属性对比中作的工作，两种语言的认知属性之间可以被划分为“共现、相近、相异”[4]三种关系。示例如下表1。借用这种对比框架，对比中英文动物词的认知属性时，本文规定了三种关系，划定了1393对属性关系。

名词	英文属性	中文	属性关系
厨师chef	fastidious，skilled，expert	创新，专业，出色	共现
屠宰场abattoir	bloody	性感，无奈，恶心	相异
算盘abacus	primitive	死板，坚硬	相近

Table 1: 跨语言认知属性对比关系

在标注属性关系过程中，通过观察相异关系的属性可以发现，相异的属性多数表现出具有隐喻义、联想义、主观色彩义等特征，因此本文针对属性的描写角度作进一步的主观量划分。汪梦翔、饶琪等人在名词隐喻知识库研究中，自主提取了隐喻属性和隐喻特征，其中隐喻特征即本文中研究的认知属性，汪文中的隐喻属性则是对隐喻特征的概括或分类，如外观、颜色、特质等。参考汪文的隐喻属性情况，本文将描写名词外观、行为、状态等物理意义方面的特征属性认为是主观量较低的描写，对名词有联想义、隐喻义、附带主观色彩的描述认为是主观量较高的描写，对提取到的中英文属性进行标注。标注时，以能否被机器量化识别为依据判定主观量的高低：如能够通过机械传感器手段被量化识别，如快、明亮、重、白等，则视为主观量低；如需要人工标注、人工感知才能得到的属性，则视为主观量高，例如美、丑、可爱、讨厌等。标注时，相同关系标记为1，相近关系标记为2，相异关系标记为3，一致关系的数值越高，说明中英属性之间的差异越大。

4.2 整体情况

4.2.1 中文属性多于英文属性

分别统计中英双语中动物词下的属性分布情况，可以发现在部分动物词上属性数量差异明显，如猪、兔子、牛等；部分词的属性数量相近，如黄鹂、喜鹊、杜鹃等。在已知中文属性均值大于英文属性情况下，对动物词中英文属性的数量作单尾检验如表2，p =0.003, p值¡ 0.05，可知中文属性数量显著多于英文属性。

语种	平均值	标准差
英文属性计数	25.08	32.49
中文属性计数	43.8	49.75

Table 2: 中英文认知属性数量分析

一个概念的认知属性多少可以反应该词汇在语言中的使用频率、受重视程度、与当地文化的融合情况等。下表3展示了中英认知属性数量差值排名前十的动物词，这些词或许是差异研究的重点关注内容。

此外，还有一些词没有找到英文认知属性，在英文网络引擎上也搜索不到普罗大众的认知看法，将其整理发现，这些词都是来源于中国传统文化故事或发源于中国本土的动物，有些甚至没有准确的外文译名如饕餮、麒麟、鲲鹏等。这种民族文化专有的动物词在跨语言交际时造成的障碍是最困难的，需要重点关注。

4.2.2 中英文认知属性有所差异

分析1393对属性关系，发现中英共现属性总计517对，占比37.11%；中英相近属性总计291条，占比20.89%；中英相异属性总计585，占比42%。计算全部动物的“不一致”关系占

[4]匡海波,李斌,王嘉灵等.：《汉英词汇隐喻属性的对比分析与互增益技术》，中国中文信息学会.中国计算语言学研究前沿进展（2009-2011）.清华大学出版社,2011:第220-225页。

名词	中文属性计数	英文属性计数	中英差值
猪	25.08	32.49	
兔子	43.8	49.75	
牛	113	57	56
狗	135	100	35
老鼠	67	37	30
龙	35	8	27
猴子	47	22	25
马	44	20	24
蟾蜍	3	14	-11
羊	50	79	-29

Table 3: 中英文认知属性部分分布情况

关系总数比例的95%置信区间，得到动物词的中英属性不一致比例大概率在42.02%到61.87%之间，可以认为动物词的中英属性确实存在差异。

在项成东的研究中，他对36种动物进行隐喻义相同、相近、相异的判断，最后得出结论，“英汉动物隐喻之间异多同少”[5]。本文中相异关系占比低于50%的动物词仅有11个，无疑是符合这个结论的。对数据排序，可以发现差异较大的动物有杜鹃、蟾蜍、鹤、龙、马、鸡等，差异较小的动物有凤凰、老虎、兔子、蛇等。

平均值	标准差	置信区间
0.5195	0.22	[0.4202,0.6187]

Table 4: 中英属性“不一致”关系置信区间

4.2.3 中英差异受主观量影响

考虑到主客观属性对中英文属性关系的影响，分列各关系下的主客观属性分布情况，对高主观量与低主观量的属性词对应的一致关系统计分析，对二者的一致关系进行单尾检验如表5，单侧p值= 0.004，p值¡ 0.05, 高主观量与低主观量的属性词的一致关系差异显著。根据高主观量均值大于低主观量均值，可知高主观量属性词的中英差异更大，即动物词的中英认知属性差异更多地表现为高主观量属性词。

主观量	平均值	标准差
高主观量	1.83	0.84
低主观量	1.72	0.83

Table 5: 不同主观量的属性关系分析

分列中英文不同主观量地认知属性词的一致关系如表6，可以看到两种语言中相近关系的属性词中的主观量分布有所差异，另外两种关系中高、低主观量的比例接近。这种现象的产生是因为本文中对相近关系的判定较为宽泛，两个处在相近关系中的属性词既有语义关联，又在某些层面表现出相当的差异。与均值相比，中文中处于相近关系的以主观属性偏多，英文中处于相近关系的以客观属性偏多。列出具体词例发现，这种现象的产生与中英文之间翻译的不对等有关。许多英文中描述客观情景的词，在中文中被附加了许多主观色彩，成为各种各样的外延词。例如英文“fast”、“swift”、“quick”等表达速度快的词，在中文中的共现属性是“迅速”，但还有“迅猛”、“迅捷”、“利落”、“麻利”等多个中文相近属性词，这些词中附加了“威猛”、“敏捷身手好”、“做事干脆”等主观色彩，这种不对等造成在相近关系中的中文主观属性偏多，英文客观属性偏多。

[5]项成东,王茂: 《英汉动物隐喻的跨文化研究》，现代外语,2009,32(03):第239-247+328页。

名词类别	共现	相近	相异
中文主观	347（67.12%）	240（83.92%）	271（73.84%）
中文客观	170（32.88%）	46（16.08%）	96（26.26%）
英文主观	381（74.47%）	197（68.88%）	180（79.30%）
英文客观	136（25.53%）	89（31.32%）	47（20.70%）

Table 6: 三种双语属性关系下的主客观属性分布

4.3 整体属性差异

总结对比中英双语的全部认知属性，二者之间还有较为共通的一点相同和四点差异。

4.3.1 相同：怜爱幼年期动物

同一种动物有年龄上的差异，如猪有成年猪和小猪仔之分，兔子有成年兔和兔宝宝之分。人们对于不同年龄的同一物种有着不同的认知，这一现象在中英双语中都得到了印证。普遍来说，人们对幼年期的动物情感态度更积极、更怜爱，使用贬义形容词的概率更小，并习惯于将幼年动物与弱者、需要被庇护者相联系。

在本文研究中，同一动物概念衍生出不同年龄段称谓的有羊、狗、鸡、兔子、猪，五种动物的中英认知属性中均表现出了这种倾向。描述狗（dog）的认知属性中，许多贬义词汇如卑贱、丑陋、谄媚、愚蠢、恶心等，出现频率高，变形形式多，可以看出语言使用者对狗有浓烈的鄙视、厌恶情感；但在描述小狗（puppy）的认知属性中，几乎没有贬义词汇，多是可爱、快乐、听话、乖巧等褒义词汇，用幼儿、宠物的角度看待小狗。此外还有小鸡相较于公鸡、母鸡丢失了自满、愚蠢、疯狂等贬义词，兔宝宝相较于兔子、野兔丢失了胆怯、疯狂、顽皮、轻狂等贬义词，小猪只留下了快乐、可爱、胖嘟嘟等褒义属性词。

几种动物都证明，人们在描述幼年期动物时，会增加主观上的喜爱、怜惜色彩，下意识地忽略动物身上原本存在的不良特质，对于外表并不柔弱瘦小的幼年动物也会附加弱小、需要保护、乖巧的主观认知，这一点在中英双语中是共通的。

4.3.2 差异一：自由与苦难观

中文中多个动物下都出现了自由、奔放、随风飘荡、腾云驾雾、直上青云等表达对自由生活追求的属性，仅就“自由”一词，先后就有鹤、老鼠、龙、马、蛇、燕子、山羊、猪等动物作为喻体出现。这在英文属性中几乎找不到相似的概念。凡是对于行动灵活的动物如龙、凤凰、鹤、燕子、马等，中国人总习惯幻想自己如同动物一样追求自由、浪迹天涯，《庄子·逍遥游》中很早就表现出了这种向往：“乘天地之正，而御六气之辩，以游无穷者。”出现这种现象的根源，或许在于中国几千年的奴隶制与封建制统治，以及中国文化中乘奔御风、羽化登仙的精神追求。

中文中多个动物都出现了勤劳、深沉、刻苦、坚忍、耐劳等属性，来描述承受生活折磨痛苦的能力，常见于牛、马、狗、老鼠等；与之相对的是形容理想生活状态的属性，如不愁吃喝、清闲、无忧无虑、幸福、快乐、悠闲、能吃能睡等，常见于猪、鹤、山羊、兔子、狗等。这种对生活状态的思考与追求在英文中不常见，英文中与这种隐忍刻苦相类似的属性只出现了“tough”，只在马和老鼠的属性中出现过一次，也没有如中文对猪、小狗等幸福生活的大幅描写和追求。

汉语社团把“吃苦耐劳”看作一项值得称颂的能力、积极追求的品质，“刻苦”、“耐劳”、“坚强”、“坚毅”等用词无疑证明中国人赞同这种生活态度。《说文解字》中写道：“牛，大牲也。牛，件也；件，事理也。”“‘理’就是指牛的内在属性‘从顺、逊顺、驯顺’的诠释。‘事、理’也就是人们在对牛的长期驯养中获得的对牛‘任劳任怨’品质的认识。”（许家星，2006）但是面对违反这种品质的动物，尤以猪为代表，汉语社团的态度却又不是预想中的完全厌弃鄙视，像猪一样悠闲、无忧无虑的生活反而受到人们的追求与赞赏。这种矛盾体现了中国人在几千年历史文化生活中对自身苦难与处境的挣扎与思考，如《诗经·国风·魏风·伐檀》：“不稼不穑，胡取禾三百亿兮？不狩不猎，胡瞻尔庭有县特兮？彼君子兮，不素食兮！”长久的被奴役、被统治生活迫使人们培养应对苦难、“坚忍”“耐劳”的能力，但精神深处却并不赞同这种被迫、畸形的受苦生活，中国人依然向心宽体胖的动物们寄托着对幸福生活的渴望与歌颂。同时，受到先秦时期道

家宗教文化影响，汉族文化中始终对闲云野鹤、归隐山林的隐居生活心生向往，对鹤、鲲鹏、山羊等悠闲安静的动物赋予特殊的清闲意义。唐代李群玉《奉和张舍人送秦炼师归岑公山》中说：“闲云不系东西影，野鹤宁知去住心。”在汉语动物词所映射出的世界里，中国人身体在“吃苦受累”、“做牛做马”，心却向往“清闲快乐”的“猪狗生活”。这种投射在动物身上的矛盾挣扎在本文的英文研究中没有痕迹。

4.3.3 差异二：性别歧视

和汉语中增加一个辩义语素不同，英文中，同一种动物的称谓有雌雄之分，如公牛是bull，母牛是cow。对于同一动物的不同性别称谓，英文中的认知属性也有差异，这种差异在中文中则弱化得多。分析同一动物雌雄称谓之间认知属性的异同，可以发现，英文中的雌性动物词形象较雄性动物词更为消极负面，这在本文涉及的中文动物词研究中是少见的。

从动物词的隐喻角度出发研究性别偏见不是一个新出现的课题，但国内外关于它的系统性研究较少。罗芳春（2012）、周晓辉（2010）等人对动物隐喻的研究都表现出，动物词中反映了一定人类社会中的性别偏见现象，主要体现在对女性外貌、性格柔弱、社会地位低的歧视上。

本文中雌雄动物称谓有异的动物分别有鸡、牛、猪、狗。以鸡为例，鸡的总称“chicken”认知属性有褒有贬，诸如活泼、毛茸茸、繁殖力强、勇敢等，chicken的整体形象偏向于一个青壮年。公鸡（cock、rooster）的属性包含了大量映射人类社会中男性形象的词，如骄傲、好斗、雄健、大胆、有冒险精神等，整体形象具有攻击性。而母鸡（hen）的属性中多为负面消极的形容词，如焦虑、自满、愚蠢、刻薄等，且基本都为主观喻人义。

这个现象十分值得思考：公鸡与母鸡除了在鸡冠、羽毛等外貌，以及在下蛋、打鸣等习性上有所不同，其他外表品行上并无太大差异，为何英文使用者会对公鸡和母鸡两个名词产生截然不同的认知？这种现象不局限于鸡一种动物，在狗、牛、猪等动物身上都有表现，考虑到动物词的喻人义作用，人类社会中性别歧视映射或许是一个较为合适的解释。母鸡下蛋、孵蛋的行为让人能够联想到生育后的妇女，因此语言使用者将对已婚妇女常见的刻板印象刻薄、愚蠢、自满等贬义词汇加诸母鸡身上，而相对应的公鸡则继承了男人的雄健、威风、高傲等属性，这种刻板印象的背后正反映了人类社会中对女性不自觉的轻视、贬低。

如公狗、母狗中侮辱女性的意味就更重了，母狗的英文bitch、slut等词已演变为侮辱女人的专用詈语词汇，从网络语料中抓取到的认知属性早已不指向母狗这种犬科动物，绝大部分与bitch、slut相关的属性都是用来描述荡妇、妓女等人物形象。英文中的公狗没有特定的称谓，一般用male dog来特别指明是公狗，同时hound、dog一般默认指公狗。中文中的母狗没有明显的负面认知属性，都属于正常的动物属性词，如凶猛、警觉、美丽等。作为语义转移、褒贬色彩完全改变的词语，母狗一词是英文中表现性别歧视程度最深的动物词证明。

4.3.4 差异三：刻板印象

Gilbert和Fiske（1998）对刻板印象进行了明确阐述，“刻板印象是由人们对于某些社会群组的知识、观念和期望所构成的认知结构。作为一种特定的社会认知图式，刻板印象是有关某一群体成员的特征及其原因的比较固定的观念或想法。”[6]动物词的认知属性中，也存在人对动物的刻板认知印象，某个或某类动物的属性趋于一致，只是中英双语中的刻板印象程度有差异。

中英双语中都存在对动物认知的刻板印象，这是毋庸置疑的。张天然（2020）在其关于中国人对动物的刻板印象研究中，提出动物与人类社会中的部分群体有所关联，例如：“牛与工人关联，马与快递员关联，鼠与腐败分子关联”[7]。在本文研究中，某个动物族群的属性总有某个大致的靠拢方向，形成一个或强或弱的刻板印象。如对兔宝宝bunny的认知属性中，可爱、惊慌、胆小、活泼、乖巧、柔弱等属性频繁出现，各种同义词、近义词加重了对兔宝宝的刻板印象，塑造了一个胆小可爱的弱势幼童形象，其认知属性中基本没有与这一刻板印象相悖的属性如强壮、暴躁等出现。与之类似的还有矫健、可靠、劳累的马，愚蠢、笨重、胆小、温顺的奶牛，冷漠、精明、恶毒的蛇等等。

[6]Gilbert D T, Fiske S T, Lindzey G (Eds.)：《Handbook of social psychology (4th ed.,Vol. 2).》，Boston: McGraw-Hill, 1998.，第357 411页。

[7]张天然.：《中国人的动物刻板印象内容及对人印象评价的内隐效应》，华中师范大学硕士学位论文，2019，第358页。

但对比中英文属性的具体分布，英文动物词的刻板印象较中文程度轻，英文中的动物普遍拥有比中文中动物词更丰富多面的认知属性。例如英文中兔子的认知属性包含可爱、活泼、矫捷、温顺等大方向的刻板印象词，也包括了轻浮、阳刚、悲伤、有威胁力等角度丰富的属性，总计25条英文属性中，与中英文中主流刻板印象相类的词有20个，还有5个角度特殊的属性词：flighty（轻浮）、virile（阳刚）、sad（悲伤）、threatening（有威胁）、lucky（幸运）。这种丰富性在许多动物的英文属性中都有体现，即动物在大部分的刻板印象词之外，还拥有角度多样、评价丰富，甚至与主流刻板印象相悖的部分属性词。类似的还有狗词例下的轻率、文雅、擅长社交、专心等。英文中，表现出在刻板印象之外还有丰富属性的动物词有18个，中文中仅有8个，并且英文中非主流属性词一般存在3到5个，中文中的非主流属性词多数仅有1到2个。因此认为，英文中动物认知属性的刻板印象比中文要轻一些。

4.3.5 差异四：上位者视角

在狗的中文认知属性中，出现了大量表述卑微、低贱、谄媚等的属性词，狗在中文语境中与奴隶的形象挂钩，同样的还有被奴役的马、牛等，汉语社团从上位者的视角来描述这些动物的行为外貌。英文中与之类似的只有羊羔词例下的“obsequious（低三下四）”属性，这一属性与中文中谄媚、贱骨头、没尊严等属性的贬低意味有较大差距。此外，中文中格外强调动物对主人的忠心、服从、恭敬，分别在狗、小狗、猎狗、马、牛五种动物下出现了22个相近属性。英文中仅在狗下出现了submissive、loyal、obedient，羊下出现了submissive、compliant五个属性。

从语言使用者的角度来解释，中国人对待被奴役、被豢养的动物时，习惯于从上位者的角度贬低、猜测动物的行为，用描述奴隶的认知语义来解释动物的日常习性，例如对狗的摇尾巴动作赋予“乞怜”、“讨好”的主观意味。杨朝宝（2011）对狗的隐喻形象作了文化上的成因探讨，也指出了中国人对狗的忠诚的格外重视。[8]这或许与中华民族几千年的封建统治生活有关，长期与权力的交涉使人惯于代入统治者与被统治者的地位进行认知。《孟子·离娄下》中很早就将狗和马与臣子的角色相联系：“君之视臣如手足,则臣视君如腹心；君之视臣如犬马,则臣视君如国人；君之视臣如土芥,则臣视君如寇雠。”[9]可见汉语中上位者的描写视角由来已久。

5 结论与未来工作

本文针对动物词的英汉认知属性分析，主要从以下两个方面展开。

从数据层面来说，主要对中英双语中动物词的认知属性数量及双语间差异进行了比较。研究发现，从数量上来看，大部分动物词的中文认知属性要显著多于英文认知属性，中文中有一些属性十分丰富、承载重要认知意义的动物如猪、狗等，还有一部分动物词是汉语专有、缺少英文认知的，如鲲鹏、麒麟等。从双语差异来看，动物词的中文和英文认知属性相异的比例平均达到42%，有明显差异。对属性进行主客观划分后分析发现，中英双语认知属性上的差异更多地表现在了主观属性上。

从中英文认知属性的全局来看，中英文的总体认知属性表现出了一点共同和四点不同。共同之处在于，两种语言都习惯用怜悯、喜爱、赞颂的态度认知幼年期动物，忽略动物幼崽身上存在的和成年期动物相同的某些不好的方面，如兔子与兔宝宝、狗与小狗的差异。

不同之处在于：（1）英文中不同性别的同一动物有不同称谓，对两个称谓的不同认知属性反映出语言使用者下意识的性别歧视，这在鸡、牛、猪、狗等多种动物身上得到验证，在中文中却没有类似现象。（2）中文中马、牛等动物拥有高频的与刻苦、勤劳相关的属性，而猪、羊等动物拥有具有艳羡意的悠闲、自由等属性，两种截然相反的属性反映了汉语使用者对自由和苦难的思考观点，这种矛盾挣扎的迹象在英文中很少见到。（3）动物词英文属性表现出的刻板印象比中文程度轻，相较于中文语境，动物词拥有更多角度、更丰富的英文认知属性。（4）汉语使用者对于狗、马等畜养动物表现出了强烈的上位者视角，频繁使用具有贬低、压迫意味的属性词如卑贱、谄媚、没尊严等词描述动物，这在英文中也较为少见。

英汉动物词的认知研究对认知语义学是极有价值的理论空白补充，也是翻译与对外汉语教学界重要的参考研究，为了解、化解非汉语社团成员在接触汉语动物的认知印象时遇到的障碍提供帮助。跨语言交际者需要先了解认识目的语中截然不同的精神文化，才能更好地掌握这门

[8]杨朝宝：《汉英动物词汇对比研究》，云南师范大学硕士学位论文2008，第312页。
[9]王清：《动物隐喻的认知和应用探究》，上海交通大学硕士学位论文,2008，第67页

语言。本文依靠已有的数据库语料，聚焦动物词的认知属性，得到了详细明确的数据结果。同时借助文学典籍，分析了两个民族之间的文化差异，并提供了有效的解释说明。

由于研究深度和规模有限，本文研究未来还需开展进一步的工作深入。首先，本文研究中所使用的语料库成库较早，未来可仿照数据库手段获取新数据进行分析。其次，本文研究词量较少，未来可扩大研究范围，深入分析动物词认知属性。最后，本文研究结论为隐喻计算提供了新的研究视角，未来可结合隐喻计算，开展相应的实证性研究。

参考文献

Bergmann. 1979. *Metaphor and format semantic theory.* Poetics8.

Chandler. 1991. *Metaphor Comprehension: A connectionist approach to implications for the mental lexicon.* Metaphor and Symbolic Activity，6(4):227-258.

Cuddy A J C, Fiske S T, Glick P. 2004. *When professionals become mothers, warmth doesn't cut the ice.* Journal of Social issues,60(4): 701-718.

Fiske S T. 2004. *Social Beings: A core motives approach to social psychology.* John Wiley & Sons,398-400.

陈小荷. 2003. 属性分析说略. 语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集:216-223.

何燕红. 2012. 英汉“牛”文化的隐喻认知. 太原城市职业技术学院学报, 2012(01): 186-188.

贾玉祥,俞士汶. 2009. 基于实例的隐喻理解与生成. 计算机科学, 36(03):138-141.

匡海波,李斌,王嘉灵等. 2011. 汉英词汇隐喻属性的对比分析与互增益技术. 中国计算语言学研究前沿进展（2009-2011）:220-225.

李斌. 2017. 词语认知属性的知识库构建和应用. 世界图书出版有限公司北京分公司.

李斌,陈家俊,陈小荷. 2012. 基于互联网的汉语认知属性获取及分析. 语言文字应用，2012（3）：135-143.

李子鹤,苏立昌. 2007. 英汉动物词喻人义位统计分析及差异成因初探. 南开语言学刊, No.10(02):116-123+157.

汪梦翔,饶琪,顾澄等. 2017. 汉语名词的隐喻知识表示及获取研究. 中文信息学报, 31(06):1-9.

王清. 2008. 动物隐喻的认知和应用探究. 上海交通大学硕士学位论文.

项成东,王茂. 2009. 英汉动物隐喻的跨文化研究. 现代外语,32(03):239-247+328.

杨朝宝. 2008. 汉英动物词汇对比研究. 云南师范大学硕士学位论文.

杨芸. 2008. 汉语隐喻识别与解释计算模型研究. 厦门大学博士学位论文.

张天然. 2019. 中国人的动物刻板印象内容及对人印象评价的内隐效应. 华中师范大学硕士学位论文.

周晓辉. 2012. 指称人的动物隐喻中性别歧视的汉英对比研究. 荆楚理工学院学报，25 (04): 57 - 62.

融合词典信息的古籍命名实体识别研究

康文军　　左家莉*　　揭安全　　罗文兵　　王明文
江西师范大学　计算机信息工程学院　江西　南昌　330022
Email: {kwj, zjl, lwb, mwwang }@jxnu.edu.cn, jieanquan@163.com

摘要

古籍命名实体识别对于古籍实体知识库与语料库的建设具有显著的现实意义。目前古籍命名实体识别的研究较少，主要原因是缺乏足够的训练语料。本文从《资治通鉴》入手，人工构建了一份古籍命名实体识别数据集，以此展开对古籍命名实体识别任务的研究。针对古籍文本多以单字表意且存在大量省略的语言特点，本文采用预训练词向量作为词典信息，充分利用其中蕴涵的词汇信息。实验表明，这种方法可以有效处理古籍文本中人名实体识别的问题。

关键词： 古籍命名实体识别；词典信息；《资治通鉴》实体数据集构建

A Study on the Recognition of Named Entities of Ancient Books Using Lexical Information

Wenjun Kang　Jiali Zuo*　Anquan Jie　Wenbing Luo　Mingwen Wang
School of Computer and Information Engineering, Jiangxi Normal University,
Nanchang, Jiangxi 330022,China
Email: {kwj, zjl, lwb, mwwang }@jxnu.edu.cn, jieanquan@163.com

Abstract

Named entity recognition of ancient texts is of significant practical importance for the construction of a knowledge base and corpus of ancient entities. Currently, there are few studies on named entity recognition of ancient texts, mainly due to the lack of sufficient training corpus. In this paper, a dataset for the recognition of named entity in ancient books is manually constructed, starting from the History as a Mirror, as a way of starting to study the task of named entity recognition in ancient books. The paper uses pre-trained word vectors as lexical information in order to make full use of the lexical information contained in the text, which is characterised by a large number of omissions and single word meanings. Experiments show that this approach can effectively deal with the problem of recognising named entities in ancient texts.

Keywords: named entity recognition of ancient books , lexical information , construction of the History as a Mirror entities dataset

基金项目： 国家自然科学基金(61866018,62266023); 江西省教育厅研究生创新基金项目(YC2022-s329)
通讯作者： 左家莉

1 引言

1991年，Rau等人 (1991)首次提出了命名实体识别(named entity recognition，NER)任务，随即在1995年11月的第六届MUC会议上，命名实体识别任务正式被确认作为一个明确的概念和重要的研究领域 (刘浏和王东波, 2018)。作为信息抽取的一项子任务，命名实体识别主要是从一些非结构化文本中抽取出具有特定意义的命名实体并且准确归纳到预先定义好的类别，它在关系抽取、机器翻译、知识库构建等自然语言处理下游任务中发挥着重要作用。

先前的NER任务大多集中在英文数据集上，从SIGHAN Bakeoff-2006评测会议开始，中文NER逐渐走进人们的视野。与英文NER相比，由于汉字之间没有明显的分词边界，中文NER需要预先进行中文分词，但是这会造成分词错误传播。因此，之前的中文NER研究大部分基于字级别的方式 (Liu et al., 2010; Gui et al., 2019; Sui et al., 2019)，但是字级别的NER模型无法充分利用词汇以及词汇边界的信息。于是，针对这个问题，早期的一些研究主要通过对LSTM (Hochreiter and Schmidhuber, 1997)、CNN (LeCun and Bengio, 1995)、Transformer (Vaswani et al., 2017)等一些基础神经网络模型进行改进以此来引入外部词典信息，它有效地提升了模型在一些中文基准数据集上的效果 (Zhang and Yang, 2018; Ma et al., 2020; Li et al., 2020; Wu et al., 2022)。自预训练语言模型BERT (Devlin et al., 2019)被提出以来，一些研究开始利用预训练语言模型能够捕捉输入文本间隐式的语义和语法知识的能力，将外部词典信息结合到预训练语言模型中来提升NER模型的性能 (Chang et al., 2021; Nie et al., 2020; Liu et al., 2021)。

近年来，中文NER的研究主要集中在一些有限的领域和实体类型上，但很少对古文等特定领域文本中命名实体进行研究 (陈曙东和欧阳小叶, 2020)。这是由于古籍文本所使用的语言和现代文不同，其文字晦涩难懂且没有一套标准的标注规范，对研究人员的专业能力有一定的要求，而且已经标注并且公布出来的高质量古籍命名实体识别语料很少，因此很多需要靠人工对古籍语料进行预处理和标注。此外，基于深度学习的中文NER模型已经在现代文的数据集上取得了不错的成绩，而专门对古籍命名实体识别模型的研究较少。考虑到引入词典信息可以为模型提供词汇级别的信息，再结合古文多以单字表意且存在大量缩写的语言特殊性，我们认为融合词典信息的模型在古籍NER任务上也是可行的。基于此，我们展开了古籍命名实体识别任务的研究:

（1）本文根据施丁等人 (1994)编纂的《资治通鉴大辞典》，结合开源的《资治通鉴》的现代文翻译版本，研究了古籍命名实体识别的标注原则，并构建了《资治通鉴》古籍命名实体识别数据集，之后还从实体覆盖率和峰度系数两方面对标注数据进行了分析。

（2）本文使用了融合词典信息的不同命名实体识别模型，来对古籍命名实体识别任务进行研究，并对比了四种不同的预训练语言模型作为特征在各类别实体上的表现，最后用测试集中未登录词(out of vocabulary，OOV)的召回率反映模型在预测古籍文本中未登录词上的性能。

2 相关工作

以往的NER研究大多采用基于规则的方式和基于统计机器学习的方式，但是它们要求研究人员制定合适的特征工程以及具备相关的领域知识 (Zhou and Su, 2002; Takeuchi and Collier, 2002)。随着深度学习的发展，基于深度神经网络的NER模型几乎避免了那些方法的局限性，并在多个基准NER数据集上都取得了显著的性能提升 (Collobert et al., 2011; Lample et al., 2016)。Hammerton (2003)首次尝试使用神经网络模型来研究NER领域的任务。Huang等人 (2015)将双向LSTM-CRF模型用在中文命名实体识别等序列标注任务中。Peng等人 (2016)提出了一种结合中文分词模型的多任务联合训练学习的方式，它帮助模型更好地识别出了在中文社交媒体文本中的实体。尽管基于神经网络模型的方式直接有效地提升了NER模型的性能，但是融入外部词典等更多有用的额外知识仍能够在NER任务上取得较大增益。

本文对古籍NER任务的研究路线，与现有主流的利用外部词典信息来提升字级别NER模型性能的思想一致。Zhang and Yang (2018)首次提出了一种Lattice-LSTM网络结构，它能动态地将句子内所有与词典相匹配的词引入到字级别的中文NER模型中，但是它无法充分利用GPU的并行能力。Ma等人 (2020)通过构建一个SoftLexicon特征，将词典信息和边界信息结合到模型的输入表示层，它有效地避免了设计复杂的结构来融入词典信息，提升了模型的推理速度。Li等人 (2020)提出了一种FLAT-Lattice网络结构，它重新为每个字以及由词典

匹配到的词设计两个头尾位置索引，以此将词汇信息引入到Transformer结构，并采用相对位置编码 (Yan et al., 2019)去捕获每个字词之间的距离和方向信息。这种网络结构能够充分利用Transformer模型构建文本间长距离依赖的优势以及具有优越的并行能力。Wu等人 (2022)提出了一个新的InterFormer模块，它可以同时对原始字输入和词汇这两个不同的序列进行建模，从而获得融合了词边界和语义信息的字表征，再利用改进的Transformer模块对获得的字表示进行编码，这种方式有效地降低了FLAT模型的计算消耗，以及可以利用更大的词典进行匹配。

随着大规模预训练语言模型在多项自然语言处理基础任务上刷新了之前最好的结果，NER的研究重心也逐渐转向预训练语言模型。Chang等人 (2021)通过实验表明利用BERT去提取文本的特征向量要优于使用静态词向量的方式。Souza等人 (2019)简单地在BERT顶部添加了一个线性条件随机场层，并经过微调，就在其所用到的NER数据集上获得了出色的性能表现。Beltagy等人 (2019)在科技领域标注数据匮乏的情况下，向大规模预训练语言模型BERT中加入未标注的科技领域数据集进行继续预训练，之后在各种使用科技领域数据集的一些下游任务上取得不错的效果。

一些研究者开始考虑综合上述两种方式的优势，将外部词典信息融入到大规模的预训练语言模型中。Nie等人 (2020)提出了一个语义扩充模块对词典信息进行编码，以及对由词典匹配到的每个词赋予不同的权重，然后利用一个门控模块对BERT作为编码端时输出的原始字信息和扩充语义信息进行控制，它有效地缓解了社交媒体文本中数据稀疏的问题。Diao等人 (2020)通过在预训练语言模型BERT外部额外地添加了一个用来处理N-Gram词典的编码器，以此来显式地融入词语层级的信息。Tian等人 (2020)提出一种键值记忆神经网络的方式，将N-Gram词典的信息以及每个字在匹配到的N-Gram内的位置信息，融入到BERT编码的字信息中，它能够有效地利用特定领域未标注文本的信息来提升模型对未登录词的识别。Liu等人 (2021)将词典信息融入到BERT底层，利用BERT的语言表征能力学习到词典更深层的知识，最后该模型在命名实体识别等序列标注任务中呈现出卓越的性能。

现有古籍NER的研究工作也逐渐转向深度神经网络模型并且结合各种特征进行学习，以此来提升模型在古籍文本实体识别上的效果。徐晨飞等人 (2020)采用BiLSTM-CRF和BERT等四种基础的神经网络模型探究在《方志物产》云南卷语料库上四种实体识别的效果。包振山等人 (2022)提出了一种半监督学习的方法，并结合古籍语言学的特点以及词性等特征，在自建的中医古籍语料上达到了83.28%的效果。张朦等人 (2023)利用双向LSTM模型抽取出《花间集全译》语料的部首、声韵和格律等多个特征并和字向量特征融合，其模型效果达到了85.63%。受Gururangan等人 (2020)提出的基于领域自适应训练思想的启发，王东波等人 (2021)使用精校后的《四库全书》全文作为训练集，在预训练语言模型BERT和RoBERTa框架的基础上使用掩码语言模型任务进行预训练，最终获得了面向古文领域的SikuBERT和SikuRoBERTa预训练语言模型。Wang and Ren (2022)在包含道部、佛部等数十部古籍在内的大规模语料库上基于BERT预训练语言模型进行学习得到了词表更大的古文预训练模型bert-ancient-chinese。

3 古籍命名实体识别数据集

3.1 古籍数据集的选取

《资治通鉴》作为我国第一部编年体的通史，在史学和文学领域有着相当高的研究价值。它由北宋史学家司马光历时19年完成，涵盖了从周朝到后周16朝1362年的历史。随着古籍数字化研究工作日趋成熟，一些平台陆续公布了高质量《资治通鉴》文本，这对后续的古籍研究起到了推动作用。本文选取古诗文网[0]中收录的数字化《资治通鉴》为语料来源，经过分句和人工审校文白对照翻译后，在此语料上开展和完成专有名词的标注工作。

3.2 《资治通鉴》数据集标注

3.2.1 实体标注原则

目前，开源且受到广泛研究的中文命名实体识别基准数据集主要有Ontonotes 4.0(Weischedel et al., 2011)、MSRA(Levow, 2006)、Weibo(Peng and Dredze, 2016)、Resume (Zhang and Yang, 2018)以及Cluener(Xu et al., 2020)等，然而古籍语料从词汇到语法以及实体的构造规律都与现代文语料有所不同，现代文本的标注方法很难直接用于古籍实体标注。于是本文通过参考Ji等人 (2021)构建的“二十四史”命名实体识别数据集以及刘浏 (2018)针对古籍中

[0]https://www.gushiwen.cn/

人名、地名、时间三类实体类别和成分划分的探究，并结合《资治通鉴》中待标注的实体为例，制定了一套较为规范的五种类别实体的标注原则。

1、人名(PER)

人名实体是古籍文献中最重要且被研究最多的命名实体之一，它主要指古籍中出现人物的姓名以及一些可以指代人物的词语。与现代人名相比，古代人名种类相对要复杂得多，主要体现在古人除了有姓和名之外，还有字、谥号、爵位、排行、职官、尊称、庙号等，如表1所示。

构成成分	示例
姓+名	**王守澄**恶官者**田全操**
名	独**充国**留屯田
谥号	孝武皇帝、孝惠皇帝
尊称、爵位	沛公、郑伯、恒侯
字	昔鲍叔之于管仲，**子皮**之于**子产**
职官	庶人勇既废，**秦王**已薨

Table 1: 人名主要构成成分和示例

2、地名(LOC)

地名实体是另一种被研究较多的实体，主要分为地名、山川河流名、关隘名等。它多以单字或双字的形式出现，相对人名实体，实体数量较少且不存在大量缩写。此外，它在古籍中出现的位置具有一定的规律，例如一般加在人名前面便于区分标识和避免重名，出现在“于”、“至”、“居”、“迁”、“屯”、“破”和“攻”等一些指示词后面以及伴随在一些“东南西北”方位词附近等，而山水名和关隘名常和“氵”、“阝”等汉字部首相关，如表2所示。

构成成分	示例
人名前面	**新丰**王孝杰从刘审礼击吐蕃
“攻”等指示词	别将陈贞等攻**武陵**
汉字部首	庾亮还**芜湖**、**阳关**三百馀里
方位词	南破**零**、**桂**，东掠**武昌**

Table 2: 地点名主要构成成分和示例

3、官职名(JOB)

官职名实体是指在国家管理和行政工作中承担不同职位、具有不同职权范围和地位等级的一类人群的统称,大体上可以分为中央官职和地方官职两大类，对古籍文本中官职名实体的识别有助于研究当时的官职制度。它同地名实体一样，在古籍中出现的位置具有规律，例如出现在“拜”、“除”、“擢”、“出”、“封”和“迁”等一些表明官职任免升降的词语后面以及出现在人名实体的前面表明其身份等。

4、组织名(ORG)

《资治通鉴》记载了从周威烈王到后周世宗等16朝的历史，因此包含了很多像国家、诸侯国、少数民族部落等政治方面的地点。于是本文对地名实体进行了细分，将它们划分到组织实体。此外，它还包括家族名和官署名等。

5、时间名(TIME)

《资治通鉴》按照时间先后顺序记叙史事，识别出其中的时间实体信息，有助于研究人员梳理历史人物事件发展的脉络。时间实体主要分为月份、季节、年份三种成分，其中季节和月份同现代一样，分为四个季节和十二个月份，年份主要有太岁纪年、干支纪年、采用天子谥号或者尊号的方式纪年以及汉武帝之后的年号纪年。

对于同一个实体名称可以指代不同类别造成的歧义问题，本文结合上下文语境对实体类型进行判断和标注。例如“与刁协帝尽诛王氏”和“以昭仪王氏为德妃”两句中“王氏”在不同语境中分别作为人名和组织名。又如“韩王成又无功”和“臣为韩王送沛公”两句中的“韩王”是一个爵位，都指代的是韩成，但在第一句中“韩王”更倾向于官职名，后一句更倾向于人名。

3.3 数据集分析

Liang等人 (2021)从实体覆盖率和峰度系数两个角度对中文NER基准数据集进行分析，发现它们中存在着两种可能会影响模型泛化性能的数据偏差，于是本文也将从这两个方面对标注的数据集进行分析。

首先，分别计算验证集和测试集中在训练集出现过的实体比例，比例越高表明会影响模型在预测未见实体上的性能。如表3所示，《资治通鉴》数据集的验证集和测试集中各有43.6%和42.5%的实体在训练集出现过。与Liang等人 (2021)在中文NER基准数据集上计算的实体覆盖率相比，本文划分的数据集实体覆盖率不高。

数据集	验证集实体覆盖率	测试集实体覆盖率
OntoNotes 4.0	50.5%	51.4%
MSRA	55.4%	70.9%
Weibo	49.8%	42.9%
Resume	54.0%	54.4%
Cluener	61.5%	-
资治通鉴	43.6%	42.5%

Table 3: 在验证集和测试集中的实体覆盖率

其次，本文使用峰度系数 (Balanda and MacGillivray, 1998)去度量标注数据集中的fat-head实体，即出现频率较高的实体，高峰度系数意味着它比正态分布具有更多的异常数据，低峰度系数意味着它具有较少的异常值。

$$Kurtosis = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{X_i - \mu}{\sigma}\right)^4\right] \tag{1}$$

其中，X_i是指每个实体类别中各实体出现频次组成的数组，μ是指均值，σ是指标准差。

如表4所示，官职、人名和组织实体间存在着较多的fat-head实体，这是由于官职实体和组织实体的类型相对有限，而人名实体则是由于一些主要人物出现的频次较高。因此，未来我们将尝试增加数据集或使用实体替换算法来缓解这一问题。

实体类别	训练集	验证集	测试集
PER	113.6	21.0	91.8
LOC	58.4	15.2	8.5
JOB	183.9	42.2	59.3
ORG	87.3	12.9	15.0
TIME	14.0	7.6	4.5

Table 4: 不同实体类别的峰度系数

4 模型

本文主要研究古籍中扁平化命名实体 (Yan et al., 2021)的识别，在神经网络模型中，它通常作为一个序列标注任务来处理，也就是对序列中的每一个字分配一个标签。基于深度学习方式的NER模型的一般架构分为三部分：嵌入层、编码层和解码层。首先给定一个输入文本序列$T = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，其中$x_i$，i=1，2，...，n对应字级别古籍文本，$y_i$，i=1，2，...，n对应输出序列标签。在嵌入层获得每个原始字的特征向量后，再输入到编码端对上下文信息进行建模。

$$u_1, ..., u_n = Embedding(x_1, ..., x_n) \tag{2}$$

$$h_1, ..., h_n = Encoder(u_1, ..., u_n) \tag{3}$$

其中，Embedding()既可以使用静态的词向量，也可以使用基于上下文的词向量，例如BERT等。Encoder()可以是任何一种基础的神经网络模型架构。

在解码端，为了考虑连续标签之间的依赖性，通常采用线性条件随机场(CRF)来对标签序列进行约束和预测。首先将编码层最后的隐藏层状态向量通过一个线性变换层再作为CRF层的输入，然后计算输出标签序列的概率，如式(4)所示。在训练过程中，通过最小化负对数极大似然函数来不断更新模型的参数进行学习，如式(5)所示。在解码阶段，使用维特比算法找到得分最高的标签序列。

$$p(y \mid s) = \frac{\exp\left(\sum_i \left(O_{i,y_i} + T_{y_{i-1},y_i}\right)\right)}{\sum_{\tilde{y}} \exp\left(\sum_i \left(O_{i,\tilde{y}_i} + T_{\tilde{y}_{i-1},\tilde{y}_i}\right)\right)} \tag{4}$$

$$L = -\sum_j \log(p(y \mid s)) \tag{5}$$

其中$O_{i,\tilde{y}_i}$是经过一层线性层后的分数，T是转移分数矩阵，$\tilde{y}$表示所有可能的标签序列。

融合词典信息的NER模型也是基于序列标注的框架，一些研究通过改进不同的神经网络模型架构，将外部词典匹配得到的词汇信息在嵌入层或者编码层和输入文本信息结合，最后在线性CRF层进行解码。如图1所示，LEBERT模型通过词典适配器模块，将词典匹配到的词汇信息整合到预训练语言模型BERT的不同Transformer层中,来充分学习到词汇更深层次的特征。

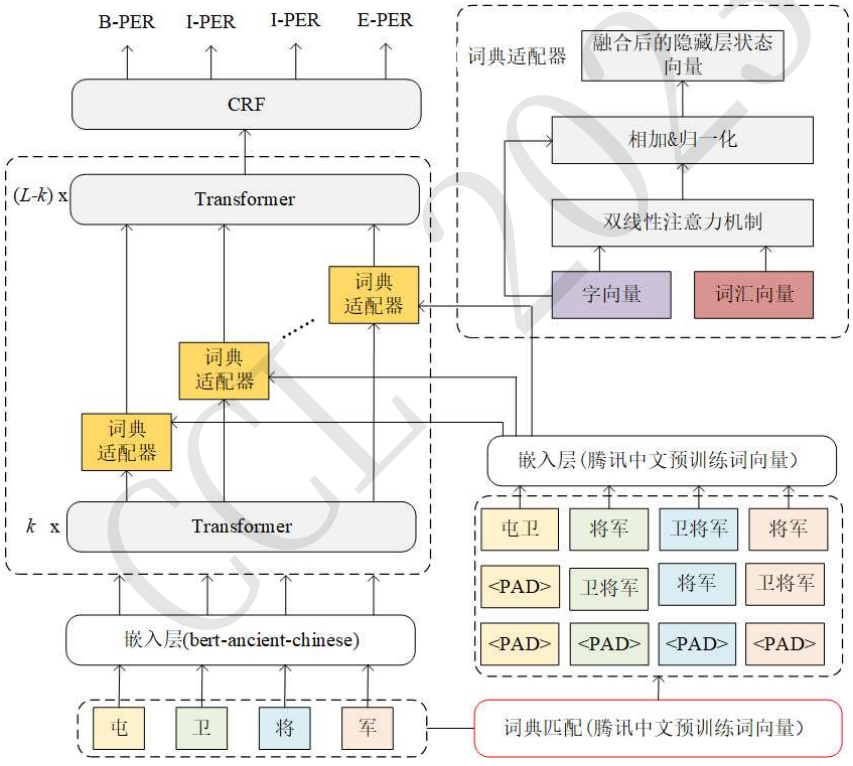


Figure 1: LEBERT(bert-ancient-chinese)模型整体架构

5 实验

5.1 数据集实体统计

本文借助Label Studio[1]进行实体的标注，并参照Resume(Zhang and Yang, 2018)等中文NER基准数据集的格式，将标注好的数据采用BIOES标注模式进行转换。然后，我们将构建的训练语料按照8:1:1的比例划分为训练集、验证集和测试集，再统计划分后实体的数量，

[1]https://labelstud.io/

如表5所示。在实验中，我们选取了本文构建的《资治通鉴》NER数据集和Ji等人(2021)公布出来的“二十四史”NER数据集作为研究对象，其中“二十四史”实体数量如表6所示。

实体类别	训练集	验证集	测试集
人名	10915	1271	1377
地名	3250	437	423
官职名	4802	429	403
组织名	1421	296	276
时间名	1743	74	96
总计	22131	2507	2575

Table 5: 《资治通鉴》命名实体数量详情

实体类别	训练集	验证集	测试集
人名	11532	756	859
地名	3625	220	236
官职名	2252	448	349
组织名	2041	4	45
总计	19450	1428	1489

Table 6: “二十四史”实体数量详情

5.2 实验参数设置与评估指标

本文在实验中采用了腾讯中文预训练词向量 (Song et al., 2018)作为外部词典信息，它提供了超过1200万个中文词汇和短语的200维词向量表示，这些词向量表示都是基于大规模语料库进行预先训练得到的。此外，输入文本的最大长度设置为200，batch size大小设置为4，epoch大小设置为50，学习率设置为1×10^{-5}。在评估指标的选择上，本文使用精确率(Precision)、召回率(Recall)以及F1值(F1-Score)来综合考虑这五个实体类别在各个模型上的效果，最后使用未登录实体词的召回率来计算模型在处理未见过实体词上的表现，如式(7)所示。

$$\mathrm{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \tag{6}$$

$$\mathrm{R}_{OOV} = \frac{\text{模型正确识别的未登录词}}{\text{测试集中未登录词总数}} \times 100\% \tag{7}$$

5.3 实验与分析

5.3.1 在“二十四史”NER数据集上的考察

本文选取了BERT-CRF模型以及融入词典信息的LEBERT模型 (Liu et al., 2021)，来对Ji等人(2021)公开的“二十四史”NER数据集进行考察。此外，我们选取了在中文语料上进行预训练的bert-base-chinese (Devlin et al., 2019)以及在古文语料上进行预训练的sikubert (王东波等人, 2021)分别作为模型的输入特征。

实验结果如表7所示，在使用bert-base-chinese作为输入特征时，LEBERT模型的F1值比BERT-CRF模型提升了1.32%。当换成sikubert预训练语言模型后，两个模型的F1值都进一步提高了，说明这两者的信息有助于对古籍命名实体识别任务的研究。然而，我们通过对“二十四史”NER语料中实体以及文本来源的分析，发现它数据来源于多部不同的史书并且实体标注稀疏，因此我们人工构建了一份《资治通鉴》NER数据集来研究古籍命名实体识别任务。

模型	PER	LOC	JOB	ORG	F1值
BERT-CRF(bert-base-chinese)	67.61%	55.39%	67.44%	48.00%	64.81%
BERT-CRF(sikubert)	67.46%	58.14%	68.22%	50.36%	65.48%
LEBERT(bert-base-chinese)	67.75%	59.65%	69.60%	49.37%	66.13%
LEBERT(sikubert)	70.20%	60.19%	70.26%	57.81%	68.22%

Table 7: 在“二十四史”NER数据集上的实验结果

5.3.2 在《资治通鉴》NER数据集上的研究

本文做了三组实验来对比目前加入词典信息的不同模型在古籍命名实体识别任务上的性能，第一组选取了不加入任何外部信息的基准模型，BiLSTM-CRF、BERT-

CRF和TENER (Yan et al., 2019)；第二组选取了传统的融入词典信息的模型，Lattice-LSTM (Zhang and Yang, 2018)、FLAT (Li et al., 2020)、NFLAT (Wu et al., 2022)和Lexicon-AugmentedNER (Ma et al., 2020)；第三组选取了将词典信息融入预训练语言模型中的模型，WMSEG (Tian et al., 2020)和LEBERT (Liu et al., 2021)。此外，本文使用谷歌发布的中文预训练语言模型bert-base-chinese (Devlin et al., 2019)抽取输入文本的特征。

实验结果如表8所示，相比于基准模型，加入词典信息有助于提升基础神经网络模型在古籍实体识别上的效果。在不使用预训练语言模型的情况下，LexiconAugmentedNER模型效果最优，F1值为73.21%。在使用预训练语言模型后，古籍实体识别的效果达到了77.26%，由于预训练模型BERT在处理中文文本时通常聚焦于字级别的输入，当加入词级别的信息后，模型在古籍文本实体识别上的效果进一步提升了，F1值高达81.24%。

模型	精确率	召回率	F1值
BiLSTM-CRF	68.82%	62.20%	65.34%
TENER	71.22%	65.36%	68.17%
BERT-CRF(bert-base-chinese)	77.59%	76.93%	77.26%
Lattice-LSTM	72.25%	65.51%	68.72%
FLAT	75.62%	68.65%	71.96%
NFLAT	75.24%	69.75%	72.39%
LexiconAugmentedNER	77.86%	69.09%	73.21%
WMSEG(bert-base-chinese)	79.97%	80.16%	80.06%
LEBERT(bert-base-chinese)	81.24%	81.24%	81.24%

Table 8: 在标注的《资治通鉴》NER数据集的实验结果

5.3.3 在不同预训练语言模型上古籍各类别实体识别的效果

为了研究使用不同的预训练语言模型能否进一步提升模型在识别古籍文本实体上的效果，我们选取了在《资治通鉴》NER数据集上F1值最高的LBERT模型作为基准模型，然后分别加入chinese-bert-wwm (Cui et al., 2021)、sikubert (王东波等人, 2021)和bert-ancient-chinese (Wang and Ren, 2022)这三个不同的预训练语言模型作为特征进行对比实验。

实验结果如表9所示，从模型的F1值来看，换成bert-ancient-chinese预训练语言模型后模型的性能最优，F1值上升了2.79%，其次是sikubert预训练语言模型F1值上升了1.02%，chinese-bert-wwm预训练语言模型F1值上升了0.46%，这表明特定领域上的预训练语言模型能够进一步帮助模型提高对古籍实体的识别，尤其是利用了更大规模的古籍语料文本进行预训练的语言模型。从各类别实体识别的效果来看，模型对人名实体和地点实体识别的F1值提升最多，分别上升了3.63%和4.48%，这说明相对于其它实体类型而言，它们由于实体构成多样且词汇用法较现代多有不同，需要来自更多特定领域的信息来帮助模型准确地识别出来。

模型	PER	LOC	JOB	ORG	TIME	F1值
LEBERT(bert-base-chinese)	85.18%	80.77%	79.72%	67.51%	74.85%	81.24%
LEBERT(chinese-bert-wwm)	86.22%	81.26%	77.73%	67.60%	74.44%	81.70%
LEBERT(sikubert)	86.57%	83.30%	79.06%	66.06%	77.27%	82.26%
LEBERT(bert-ancient-chinese)	88.81%	85.25%	79.64%	68.20%	77.53%	84.03%

Table 9: LEBERT模型在不同预训练语言模型上的实验结果

5.3.4 在预测未见过实体词上的效果

未登录词的识别是命名实体识别模型中的一个挑战，因此本文通过计算测试集中未登录词的召回率，来研究融入词典信息的NER模型在缓解古籍文本未登录词上的效果。从表10可以看出，加入词典信息和古文领域预训练语言模型都能够提升模型在识别OOV词上的性能。

模型	R_{OOV}
测试集未登录实体数量	1480
BERT-CRF(bert-base-chinese)	78.04%
WMSEG(bert-base-chinese)	82.84%
LEBERT(bert-base-chinese)	82.77%
LEBERT(sikubert)	84.86%
LEBERT(bert-ancient-chinese)	86.96%

Table 10: 在测试集未见过实体词上的召回率

6 案例分析

相对其它实体类型，人名实体构成灵活且常常被省略姓氏，因此模型在识别某些人名实体上难度较高，为此我们挑选出模型对人名实体识别的样例进行分析。如表11示例一所示，在某些情况下，LEBERT模型在识别一些被省略姓氏的人名实体上表现较好。

人名实体存在一定的歧义性，如表11示例二所示，“定兴”在不同的语境下可以是人名实体，也可以是地名实体。尽管我们结合了上下文语境对实体进行标注，但融入外部词典的模型仍无法正确识别出某些实体在当下语境的实体类型，这是由于“定兴”在词典中同时是人名或地名，它无法提供给模型对应的上下文信息。当把bert-ancient-chinese古文预训练语言模型作为LEBERT模型的输入特征后，它能有效地缓解这种歧义问题，这可能是它在大规模古文领域数据集上进行预训练时，学习到了这种上下文的信息。

示例一： 引入词典信息的NER模型实体抽取结果	
古籍文本片段	上欲遣淮南太守**戴僧静**将兵讨子响，**僧静**面启曰...
实际标注标签	B-PER I-PER E-PER B-PER E-PER
Lattice-LSTM	B-PER I-PER E-PER O S-PER
FLAT	B-PER I-PER E-PER O O
LEBERT(bert-base-chinese)	B-PER I-PER E-PER B-PER E-PER
古籍文本片段	**子如**曰：“**消难**亦通**子如**妾，此事正可掩覆。...”
实际标注标签	B-PER E-PER B-PER E-PER B-PER E-PER
Lattice-LSTM	O S-PER O O O O
FLAT	O S-PER O O O O
LEBERT(bert-base-chinese)	B-PER E-PER B-PER E-PER B-PER E-PER
示例二： 加入古文领域预训练语言模型实体抽取结果	
古籍文本片段	... 应募隶屯卫将军**云定兴**，说**定兴**多赍旗鼓为疑兵，...
实际标注标签	B-PER I-PER E-PER B-PER E-PER
LEBERT(bert-base-chinese)	O B-LOC E-LOC B-LOC E-LOC
LEBERT(sikubert)	O B-LOC E-LOC B-LOC E-LOC
LEBERT(bert-ancient-chinese)	B-PER I-PER E-PER B-PER E-PER
古籍文本片段	...今若杀**山阳**，与雍州举事，...则霸业成矣！**山阳**持疑不进...
实际标注标签	B-PER E-PER B-PER E-PER
LEBERT(bert-base-chinese)	B-LOC E-LOC B-LOC E-LOC
LEBERT(sikubert)	B-LOC E-LOC B-LOC E-LOC
LEBERT(bert-ancient-chinese)	B-PER E-PER B-PER E-PER

Table 11: 测试集中各模型预测示例

7 总结与未来工作

本文首先考察了“二十四史”NER数据集，发现它语句来源分散且标注稀疏，因此我们选取《资治通鉴》作为研究语料，并对数据集中五个类别的实体进行标注，人工构建了一份《资治

通鉴》命名实体识别数据集。此外，本文采用不同方式融入词典信息的模型，以及加入特定领域的预训练语言模型作为特征，来研究古籍命名实体识别任务。实验表明，加入词典信息能够帮助模型识别出被省略姓氏的人名实体，并提升模型在预测未见过实体上的性能。然而，融合词典信息的模型不具备消歧能力，无法解决一词多义的现象。

我们通过对标注数据集的分析，发现某些实体类别存在一些fat-head实体，会对模型的泛化能力造成影响，因此我们将增加语料的规模和语料的来源，并标注更多的实体类型。在古籍标注的过程中，我们观察到现代文翻译很好地对古文进行了补充，因此这些翻译中也包含了许多的信息。在未来，我们将尝试利用这些翻译的信息，来对古籍命名实体识别任务进行研究。

参考文献

K. Balanda and H. MacGillivray. 1988. Kurtosis: a critical review. *The American Statistician*, 42:111–119.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research 12(Aug)*:2493–2537.

Yuan Chang, Lei Kong, Kejia Jia, Qinglei Meng. 2021. Chinese named entity recognition method based on bert. *2021 IEEE International Conference on Data Science and Computer Application*, pages 294-299.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 3504-3514.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4982–4988.

Zhiheng Huang, Wei Xu, and Kai Yu 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 172–175.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zijing Ji, Yuxin Shen, Yining Sun, Tian Yu, and Xin Wang. 2021. C-CLUE: a benchmark of classical chinese based on a crowdsourcing system for knowledge graph construction. *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: chinese ner using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Associa-tion for Computational Linguistics*, pages 6836–6842.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced chinese sequence labeling using bert adapter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5847–5858.

Zhangxun Liu, Conghui Zhu, and Tiejun Zhao. 2010. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? In *Advanced intelligent computing theories and applications. With aspects of artificial intelligence, Springer*, pages 634–640.

Guanqing Liang and Cane Wing-Ki Leung. 2021. Improving model generalization: a chinese named entity recognition case study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 992–997.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,pages 260–270.

Yann LeCun and Yoshua Bengio. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*,3361(10):1995.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: ord segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2020. Simplify the usage of lexicon in chinese ner. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5951–5960.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1383–1391.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 149–155.

L. F. Rau. 1991. Extracting company names from text. *The Seventh IEEE Conference on Artificial Intelligence Application*, pages 29-32.

Yan Song, Tong Zhang, Yonggang Wang and Kai-Fu Lee. 2021. ZEN 2.0: continue training and adaption for n-gram enhanced text encoders. *arXiv preprint arXiv:2105.01279*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–180.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3830–3840.

Fábio Souza, Rodrigo Nogueira, Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.

Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002*.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020. Improving chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Shuang Wu, Xiaoning Song, Zhenhua Feng, and Xiaojun Wu. 2022. NFLAT: non-flat-lattice transformer for chinese named entity recognition. *arXiv preprint arXiv:2205.05832*.

Pengyu Wang and Zhichen Ren. 2022. The uncertainty-based retrieval framework for ancient chinese cws and pos. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, et al. 2011. OntoNotes release 4.0. *Web Download. Philadelphia: Linguistic Data Consortium*.

Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. CLUENER2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5808–5822.

Hang Yan and Bocao Deng and Xiaonan Li and Xipeng Qiu. 2019. TENER: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Associa-tion for Computational Linguistics*, pages 1554–1564.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480.

包振山,宋秉彦,张文博,孙超. 2022. 基于半监督学习和规则相结合的中医古籍命名实体识别研究. 中文信息学报, 36(6): 90-100.

陈曙东,欧阳小叶. 2020. 命名实体识别技术综述. 无线电通信技术,46(03):251-260.

刘浏. 2018. 古汉语典籍中的实体知识挖掘研究. 南京大学,DOI:10.27235/d.cnki.gnjiu.2018.001041.

刘浏,王东波. 2018. 命名实体识别研究综述. 情报学报,37(03):329-340.

施丁, 沈志华, 陈东林, 和龚. 1994. 资治通鉴大辞典. 吉林人民出版社.

苏棋, 胡韧奋, 诸雨辰, 严承希, 王军. 2021. 古籍数字化关键技术评述. 数字人文研究1(3): 83-88.

王东波，刘畅，朱子赫，刘江峰，胡昊天，沈思，李斌. 2021. SikuBERT与SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究.

徐晨飞, 叶海影, 包平. 2018. 基于深度学习的方志物产资料实体自动识别模型构建研究. 数据分析与知识发现, 4(8): 86-97.

张朦, 刘忠宝. 2023. 数字人文环境下融入多特征的词命名实体识别. 计算机系统应用,32(3):300-308.

结合全局对应矩阵和相对位置信息的古汉语实体关系联合抽取

胡益裕[2]　左家莉[1]　曾雪强[1]　万中英[1]　王明文[1,2]
[1]江西师范大学　计算机信息工程学院　江西　南昌　330022
[2]江西师范大学　数字产业学院　江西　上饶　334000
Email: 329272494@qq.com, {zjl, xqzeng, libby, mwwang}@jxnu.edu.cn

摘要

实体关系抽取是信息抽取领域中一项重要任务，目前实体关系抽取任务主要聚焦于英文和现代汉语领域，关于古汉语领域的数据集构建和方法的研究目前却较少。针对这一问题，本文在研究了开源的《资治通鉴》语料后，人工构建了一个古汉语实体关系数据集，并设计了一种结合全局对应矩阵和相对位置信息的实体关系联合抽取方法。最后通过在本文构建的数据集上进行实验，证明了该方法在古汉语实体关系抽取任务上的有效性。

关键词： 古汉语数据集构建；实体关系联合抽取；全局对应矩阵；相对位置信息

Joint Extraction of Ancient Chinese Entity Relations by Combining Global Correspondence Matrix and Relative Position Information

Yiyu Hu[2]　**Jiali Zuo**[1]　**Xueqiang Zeng**[1]　**Zhongying Wan**[1]　**Mingwen Wang**[1,2]
[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China
[2]School of Digital Industry, Jiangxi Normal University, Shangrao, Jiangxi 334000, China
Email: 329272494@qq.com, {zjl, xqzeng,libby, mwwang}@jxnu.edu.cn

Abstract

Entity relation extraction is an important task in the field of information extraction. Currently, entity relation extraction is mainly focused on the fields of English and modern Chinese, but there are few researches on the construction and methods of data sets in the field of ancient Chinese. To solve this problem, this paper constructs an ancient Chinese entity relation dataset by hand after studying the open-source corpus of "Comprehensive Mirror for Aid Government", and designs a joint entity relation extraction method combining global correspondence matrix and relative location information. Finally, experiments are carried out on the dataset constructed in this paper to prove the effectiveness of the proposed method for entity relation extraction in ancient Chinese.

Keywords: Ancient Chinese datasets construct , Joint extraction of entity relationships , global corresponding matrix , relative position information

通讯作者: 左家莉
基金项目: 国家自然科学基金(61866018, 62266023, 62266021); 江西省教育厅科学技术研究项目(GJJ2200330)

1 引言

实体关系抽取(Entity Relation Extraction)任务旨在识别出非结构化文本中的实体和实体与实体之间的语义关系，是信息抽取(Information Extraction, IE)领域中一项重要任务。

实体关系抽取任务的研究工作最早开始于上世纪90年代(Brin, 1998)，随着研究的深入，该任务对于大规模标注数据的需求也在不断上升。为此，近年来，学术界和工业界构建了各种基于英文领域(Riedelet al., 2010; Gardent et al., 2017)和现代汉语领域(Xu et al., 2017)的实体关系数据集，这些工作极大的推动了实体关系抽取任务在上述两个领域中的研究。相较而言，实体关系抽取在古汉语领域中的研究工作则相对较少，其主要原因在于：(1)目前古汉语实体关系标注数据集较少；(2)相较于现代汉语实体关系标注工作而言，古汉语实体关系标注工作在标注原则设定、标注类型选定等方面难度更大，要求标注人员具有扎实的古汉语专业知识。最近，王鑫等人(2021)基于对“二十四史”语料的研究，构建了一份“二十四史”实体关系数据集。然而，通过对该数据集进行分析(表1)，我们发现该数据集具有如下特点：(1)标注规模较小；(2)在数据标注上存在标注稀疏问题，每条标注文本仅标注了少量的实体和关系。上述特点使得实体关系抽取模型较难从“二十四史”数据集中学习到足够的信息，最终导致其在该数据集上的性能较差。

类型	训练集	验证集	测试集
句子数	3113	882	418
每条标注数据标注的关系数	1	1	1
每条标注数据标注的实体数	2	2	2
总实体数	6226	1764	836
总关系三元组数	3113	882	418

Table 1: “二十四史”数据集部分特点描述

而在实体关系抽取方法的研究上，早期的实体关系抽取方法大多是基于特征的方法(Ren et al., 2017; Li and Ji, 2014)。随着深度学习的发展，后来的研究者提出了各种基于深度学习的实体关系抽取方法，这些方法大致可分为基于特殊标签识别的方法(Zheng et al., 2017; Wei et al., 2020; Wang et al., 2020; Zheng et al., 2021; Shang et al., 2022)，以及基于序列生成的方法(Zeng et al., 2018; Sui et al., 2021)。其中，Zheng等人(2021)和Shang等人(2022)采用了一种基于全局对应矩阵(Zheng et al., 2021)的实体关系联合抽取方法，该方法主要通过使用矩阵建模主客体之间(Zheng et al., 2021)、主客体和关系之间(Shang et al., 2022)的关联信息，最终通过该关联信息完成对实体、关系的抽取。目前，该方法在各种英文领域数据集上均取得了SOTA(state-of-the-art)性能。然而表2显示，该方法在目前的古汉语实体关系数据集上容易生成长度较长的异常实体，其主要由于全局对应矩阵是通过实体头尾词来表示完整实体，模型在学习实体信息时主要关注实体的头尾信息，较少关注实体的边界信息，从而使得模型更难精确的识别实体的边界，最终导致模型容易生成表2所示的长度异常实体。

古汉语文本	真实值	预测值
河内公独孤信，南阳公赵贵。	（**赵贵**，任职，南阳公）	（**独孤信，南阳公赵贵**，任职，南阳公）
内牙上都监使章德安数与之争，右都监使李文庆不附于，乙巳，贬德安于处州。	（**章德安**，名，德安） （李文庆，任职，右都监）	（**章德安数与之争，右都监使李文庆**，名，德安）

Table 2: 古汉语数据集上的部分测试结果。其中，“预测值”中粗体字为长度异常实体

鉴于目前的古汉语实体关系数据集存在标注规模较小、标注稀疏等问题，本文研究了开源的《资治通鉴》语料，结合上下文语义和关系触发词，人工重新构建了一份实体和关系更加丰富的古汉语实体关系数据集，并设计了一种基于全局对应矩阵的实体关系联合抽取方法。对于基于矩阵的方法容易生成长度较长的异常实体问题，本文尝试了在全局对应矩阵上引入字与字之间相对位置信息的方法，最终通过大量实验证明了该方法的有效性。

第二十二届中国计算语言学大会论文集，第241页-第252页，哈尔滨，中国，2023年8月3日至5日。

2 相关工作

2.1 实体关系抽取数据集构建

近年来，为了开展实体关系抽取任务而构建的数据集主要有NYT(Riedelet al., 2010)、WebNLG(Gardent et al., 2017)、SemEval[0]等英文数据集，以及Chinese Literature Text(Xu et al., 2017)、DuIE2.0[1](Li et al., 2019)等现代汉语领域数据集。相较而言，目前在古汉语实体关系抽取数据集上的相关研究则相对较少。

最近，王鑫等人(2021)基于对“二十四史”语料的研究，构建了一份“二十四史”实体关系数据集[2]，为古汉语实体关系抽取任务的研究提供了一份数据标注基准。而王一钒等人(2021)则提出了一套由“关系配价标注”、“命名逻辑标注”以及“单一关系存在”原则构成的数据标注原则(Wang et al., 2021)，填补了古汉语实体关系数据集标注工作在标注规范上存在的空白。

2.2 实体关系联合抽取

早期的实体关系联合抽取方法主要是基于特征的方法，如：(Ren et al., 2017; Li and Ji, 2014)，该方法主要是利用设置的特征函数获得数据特征信息，然后通过该特征信息联合识别实体和关系。然而由于该方法在建立特征工程上严重依赖NLP工具和大量人工操作(Zheng et al., 2021)，使得其难以处理数据规模较大的情况。

之后，深度学习技术不断发展，结合深度学习的实体关系联合抽取方法研究受到了广泛关注。其中，Sun等人(2017)选择将实体关系抽取任务建模为序列标注任务，利用一种包含实体、关系信息的标记方案联合建模实体和关系，然而该方案由于只为每个token分配了单个标签，无法处理存在单个token对应多个标签现象的三元组重叠问题。为了解决上述问题，Zeng等人(2018)基于LSTM(Hochreiter and Schmidhuber , 1997)，提出了一种结合实体复制机制的序列到序列模型。该复制机制由于可以对同一token进行多次复制，使得每个token可以参与不同三元组的构建，提升了模型解决三元组重叠问题的能力。

近年来，随着Bert(Devlin et al., 2019)等基于大规模语料的预训练模型的提出，结合预训练模型的实体关系联合抽取方法受到广泛研究。其中，Wei等人(2020)以Bert为编码器，设计了一种“通过主体和关系识别客体”的实体关系联合抽取方法，并取得了新的SOTA性能。然而，由于该模型在实体头尾词匹配时采用“领近匹配”原则(Wei et al., 2020)，导致其无法处理嵌套实体问题。同时还因为该方法采用先识别主体后识别客体的多阶段方式，这使得模型存在错误传播问题。为了解决上述错误传播问题和实体嵌套问题，Wang等人(2020)设计了一种利用大小为n^2(n为输入文本长度)的矩阵(Zheng et al., 2021)提取关系三元组的实体关系联合抽取方法。而Sui等人(2021)则是以Bert为编码器，将实体关系识别任务重新考虑为三元组序列生成任务，通过非自回归的解码方式生成三元组序列，该方式解决了自回归方式(Zeng et al., 2018)生成三元组序列时仍需要按照三元组序列顺序解码的弊端，提升了模型的解码效率。上述两种方法在解决重叠三元组问题上都取得了较好的结果，然而由于两者均是通过不同模块识别实体和关系，不同模块间缺乏信息的交互，从而使得实体和关系之间的相互约束不足，最终导致实体和关系在匹配时出现信息冗余问题(Shang et al., 2022)。

最近，在解决重叠三元组和实体嵌套等复杂问题上。Zheng等人(2021)和Shang等人(2022)均采用全局对应矩阵(Zheng et al., 2021)联合建模实体和关系。其中，Zheng等人(2021)选择将实体关系任务划分为实体提取、主客体对齐和关系判断三个子任务，采用先进行潜在关系预测后完成主、客体识别的方式完成实体关系联合抽取。在主、客体对齐任务上，该方法主要利用一个全局对应矩阵学习主体和客体的关联性。虽然该方法提升了模型解决三元组重叠和嵌套实体问题的能力，然而由于该方法是采用先预测潜在关系，后通过潜在关系完成主、客体对齐任务的方式，这使得潜在关系预测阶段出现的错误会传递到主、客体对齐任务上，即模型存在错误传播问题。为了解决上述所说的信息冗余问题和错误传播问题，Shang等人(2022)设计了一种结合特殊标签和全局对应矩阵联合建模实体和关系的方法，该方法解决了上述所说的错误传播问题和信息冗余问题，并在实体关系抽取任务上取得了新的SOTA。

[0]https://github.com/thunlp/OpenNRE/blob/master/benchmark/download_semeval.sh
[1]https://github.com/PaddlePaddle/PaddleNLP/tree/develop/examples/information_extraction/DuIE
[2]https://github.com/jizijing/C-CLUE

然而，本文通过在古汉语数据集上进行了大量实验后发现：全局对应矩阵容易生成长度异常的实体。为了解决该问题，在研究了(Li et al., 2019)引入实体与字符间相对位置信息的方法后，本文设计了一种引入字与字之间相对位置信息的方法，最终通过大量实验证明：引入相对位置信息对于缓解“全局对应矩阵容易生成长度异常的实体”问题的有效性。

3 数据集构建

3.1 数据集的来源

《资治通鉴》是由北宋史学家司马光主编的一部编年体史书，该书记录了从周威烈王二十三年(公元前403年)到五代后周世宗显德六年(公元959年)期间共计1362年的历史。

本文构建的数据集语料来源于古诗文网[3]公开的资治通鉴语料，经过对该语料进行分句、筛选处理后，最终挑选出其中约10000条语句进行标注。目前完成标注的字符总数为76025，句子的平均长度为37.92。

3.2 数据标注准则

本文在实体关系标注类型上，参考了王鑫等人(2021)公布的“二十四史”实体关系数据集上定义的实体关系类型，而在实体和关系的标注原则上，则采用了张欢(2020)针对实体标注提出的简单性原则、易操作性原则、一致性原则。

首先是简单性原则。在实体标注上，本研究将古汉语实体类型简要分为人名(PER)、官职名(JOB)、组织名(ORG)、地名(LOC)四种实体类型，类型数目适中。针对成分复杂的类型，如：人名，本研究并未对其进行细分，对于个别在语义上产生交叉的实体，本文在标注时进行类别统一。在关系标注上，为了降低标注难度，本文选定的均是具有关系触发词或者表义明显的关系类型。所以，本文的实体标注和关系标注符合简单性原则。

其次是易操作原则。本研究为了提升标注工作的易操作性。分别对定义的实体、关系类型和实体、关系标注过程进行了详细的说明（具体见下文），符合易操作性原则。

最后是一致性原则。实体定义类型和关系类型定义是实体关系数据标注的第一步，对于容易混淆的实体类型和关系类型，本研究对其进行合并统一，遵循了一致性原则。

3.3 实体、关系标注说明

本文定义了人名(PER)、官职名(JOB)、组织名(ORG)、地名(LOC)四种古汉语实体类型和“任职”、“隶属于”和“去往”等24种关系类型，具体标注说明如下文所述。

3.3.1 实体标注说明

首先是人名(PER)，在古汉语中，“人名(PER)”成分种类繁多，包括名、字、氏、姓、爵位、排行、谥号、官职等，对于其中以官职和爵位表现的人名(PER)实体，基于对(人名，官职名，任职)这类三元组的考虑，为了维护实体标注的一致性，本研究仍旧将其标注为官职。其次是官职名(JOB)，这一类型的实体表现形式较单一，主要表现为职位名。最后是地名(LOC)和组织名(ORG)，地名(LOC)主要指地理上所定义的“地名”，如：山名、水名和地方名等。组织名(ORG)包括了国家名、氏族名、官署机构名等。

3.3.2 关系标注说明

在古汉语关系标注上，为了降低标注难度，本研究采用了结合关系触发词和上下文语义的标注方法。其中，关系触发词指的是文本中直接表达两个实体间关系的词。例如：“元舆，元褒之兄也。”，通过“兄”这一词可以得出“元舆”是“元褒”的兄长，像“兄”这种可以体现两个实体间关系的词即是关系触发词。下表3为部分关系及其相关触发词介绍。

3.4 数据标注格式

本研究在数据标注格式上，采用了王鑫等人(2021)公布的“二十四史”实体关系数据集上的标注格式，将标注后的数据存储为json文件格式，其中每条标注数据包括了如下内容：古汉语文本(text)、主体(subject)、主体类型(subject_type)、客体(object)、客体类型(object_type)、关系(relation)，具体如图1所示。

[3]https://www.gushiwen.cn/

关系类别	例句
子	守一，仁皎之**子**。
兄	中书令陈淮，徽之**兄**也。
葬于	始安忠武公温峤卒，**葬**于豫章。
弟	略、模，皆越之**弟**也。
升迁	弟晦，亦以皎故累**迁**吏部侍郎。

Table 3: 部分关系类别及其关系触发词。其中，粗体字为关系触发词

```
{"text": "又诏以太宰颙都督中外诸军事。",
"spo_list": [{"subject": "颙",
              "subject_type": "PER",
              "object": "太宰",
              "object_type": "JOB",
              "relation": "任职"
            }]
}
```

Figure 1: 数据标注格式

4 模型构建

4.1 问题描述

实体关系抽取问题具体描述如下：给定一句输入序列S，S=$(s_1, s_2, ..., s_n)$，其中，s_n表示S中的第n个词，实体关系抽取任务的目标是识别序列S中所有的主体、客体和主、客体的语义关系，并输出(主体，关系，客体)形式的三元组。

为了建模实体关系抽取任务，本文参考了Zheng等人(2021)的多任务学习思路，选择将该任务细分为：主体和客体对齐、实体和关系对齐、实体抽取三个子任务，每个任务的解释如下：(1)主体和客体对齐：目的是得到输入S中所有token间的对应分数，当该对token属于主体和客体的一部分时，其对应分数是最高的；(2)实体和关系对齐：该任务主要是通过矩阵预测输入S中所有token和任务定义的所有关系的对应分数，该对应分数表示了token和所有关系间的关联程度；(3)实体提取：目的是识别出输入S中所有的实体。为了建模三个子任务，本文采用了Zhang等人(2021)提出的全局对应矩阵的方法，为每个任务设计了对应的主体和客体全局对应、实体和关系全局对应、实体头尾全局对应三个模块(图2)，最终通过结合三个模块抽取的信息完成实体关系的联合抽取。

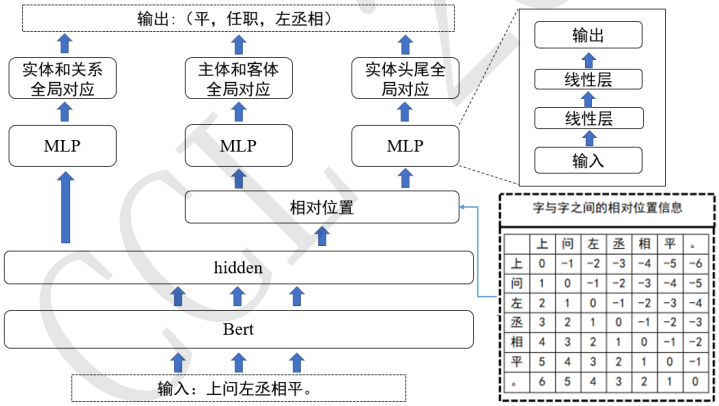


Figure 2: Bert+全局对应矩阵的模型架构

4.2 编码层

给定输入S，S=$(s_1, s_2, ..., s_n)$，s_i表示句子S中第i个词，本部分主要通过一个预训练的Bert作为模型的编码器来获得S对应的向量表示，具体如下公式所示：

$$(h_1, h_2, ..., h_n) = Bert((s_1, s_2, ..., s_n)) \quad (1)$$

其中$(h_1, h_2, ..., h_n)$表示Bert最后一层输出的隐藏层状态，n为输入文本的长度。

4.3 解码层

在本部分，本文将介绍主体和客体全局对应、实体和关系全局对应、实体头尾全局对应三个模块和相对位置信息模块的具体实现细节。

4.3.1 相对位置信息

为了缓解全局对应矩阵的解码方式存在的对实体长度约束不足的问题，本文选择引入字与字之间的相对位置信息。具体做法是：先获得两个输入token间的相对位置，然后通过将其与权重矩阵相乘获得相对位置对应的向量表示$POS_{i,j}$，之后将第i个token对应的hidden与第j个token对应的hidden进行拼接得到结果$Concat(h_i, h_j)$，最后将$POS_{i,j}$与$Concat(h_i, h_j)$相加得到最终输出$h_{i,j}^{pos}$。该过程可以公式化表示为：

$$POS_{i,j} = I_{i,j}W_{pos} + b_{pos} \tag{2}$$

$$h_{i,j}^{pos} = Concat(h_i, h_j) + POS_{i,j} \tag{3}$$

其中，I为图2中的相对位置信息矩阵，$I_{i,j}$表示输入句子中第i个token和第j个token的相对位置，h_i和h_j为输入句子中第i个和第j个词对应的hidden表示，Concat表示拼接。

	上	问	左	丞	相	平	。
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	1	0	0	0	0
。	0	0	0	0	0	0	0

(a) 主体的开始词(第一列)和客体的开始词(第一行)的全局对应矩阵

	上	问	左	丞	相	平	。
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	0	0	1	0	0
。	0	0	0	0	0	0	0

(b) 主体的结尾词(第一列)和客体的结尾词(第一行)的全局对应矩阵

Figure 3: 主体和客体全局对应

4.3.2 主体和客体全局对应

在获得了相对位置信息模块的输出后，为了建模主体和客体的全局对应任务，本文选择将主体和客体的全局对应任务细分为主、客体开始词对应和主、客体结尾词对应两个子任务，并为每个子任务设计了对应的主、客体开始词全局对应矩阵(图3(a))和主、客体结尾词全局对应矩阵(图3(b))。该模块的具体实现如下公式所示：

$$h_{i,j}^{so} = MLP(h_{i,j}^{pos}) \tag{4}$$

$$P_{i,j}^{soh} = sigmoid(h_{i,j}^{so}W_{soh} + b_{soh}) \tag{5}$$

$$P_{i,j}^{sot} = sigmoid(h_{i,j}^{so}W_{sot} + b_{sot}) \tag{6}$$

W_{soh}、W_{sot}、b_{soh} 、b_{sot} 为可训练的参数，MLP为多层感知机，$h_{i,j}^{pos}$为相对位置模块的输出。

4.3.3 实体和关系全局对应

为了建模实体和关系的全局对应任务，本文选择将其细化为主体开始词和关系的全局对应以及客体开始词和关系的全局对应两个子任务，并设置了主体开始词和关系的全局对应矩阵(图4(a))和客体开始词和关系的全局对应矩阵(图4(b))。

该模块的步骤可以形式化表示为如下公式：

$$h_i^{er} = MLP(h_i) \tag{7}$$

$$P_i^{sr} = sigmoid(h_i^{er}W_{sr} + b_{sr}) \tag{8}$$

$$P_i^{or} = sigmoid(h_i^{er}W_{or} + b_{or}) \tag{9}$$

其中，h_i为输入句子中第i个词对应的隐藏层状态表示，W_{sr}、W_{or}、b_{sr}、b_{or}为可训练的参数，MLP为多层感知机。

	...	...	任职	...	...	...	...
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	1	0	0	0	0
。	0	0	0	0	0	0	0

(a) 主体的开始词(第一列)和关系(第一行)的全局对应矩阵

	...	...	任职	...	...	...	...
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	1	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	0	0	0	0	0
。	0	0	0	0	0	0	0

(b) 客体的开始词(第一列)和关系(第一行)的全局对应矩阵

Figure 4: 实体和关系全局对应

4.3.4 实体头尾全局对应

本部分的目的是识别出输入文本中的所有实体。为了建模实体信息，本部分设计了实体头尾全局对应矩阵(图5)，通过该矩阵联合建模实体头尾信息。本模块的工作过程可以表示为如下：

$$h_{i,j}^{ht} = MLP(h_{i,j}^{pos}) \tag{10}$$

$$P_{i,j}^{ht} = sigmoid(h_{i,j}^{ht} W_{ht} + b_{ht}) \tag{11}$$

其中，W_{ht}、b_{ht}为可训练的参数，MLP为多层感知机，$h_{i,j}^{pos}$为相对位置模块的输出。

	上	问	左	丞	相	平	。
上	1	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	1	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	0	0	0	0	0
。	0	0	0	0	0	0	0

Figure 5: 实体头尾全局对应。其中，实体开始词(第一列)和结尾词(第一行)的全局对应矩阵

4.4 损失函数

本研究采用交叉熵损失函数作为模型的损失函数，模型最终的损失函数主要由L^{so}、L^{er}和L^{ht}三部分构成，其具体表示如下：$L = L^{so} + L^{er} + L^{ht}$，假设输入是由n个token组成的序列，上述公式中每部分的解释如下所示。

首先是L^{so}，其包括了L^{soh}、L^{sot}，两者具体计算过程如下公式所示：

$$L^{so} = 0.5 \times (L^{soh} + L^{sot}) \tag{12}$$

$$L^{soh} = \frac{-1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log(P_{i,j}^{soh}) + (1 - y_{i,j}) log(1 - P_{i,j}^{soh}) \tag{13}$$

$$L^{sot} = \frac{-1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log(P_{i,j}^{sot}) + (1 - y_{i,j}) log(1 - P_{i,j}^{sot}) \tag{14}$$

上述公式中，$P_{i,j}^{soh}$表示第i个token为主体开始词和第j个token为客体开始词的条件概率；$P_{i,j}^{sot}$表示第i个token为主体结尾词和第j个token为客体结尾词的条件概率。

第二十二届中国计算语言学大会论文集，第241页-第252页，哈尔滨，中国，2023年8月3日至5日。

其次L^{er}包括了L^{sr}、L^{or}，两者具体计算过程如下公式所示:

$$L^{er} = 0.5 \times (L^{sr} + L^{or}) \tag{15}$$

$$L^{sr} = \frac{-1}{n \times n^r} \sum_{i=1}^{n} \sum_{j=1}^{n^r} y_{i,j} log(P_{i,j}^{sr}) + (1 - y_{i,j}) log(1 - P_{i,j}^{sr}) \tag{16}$$

$$L^{or} = \frac{-1}{n \times n^r} \sum_{i=1}^{n} \sum_{j=1}^{n^r} y_{i,j} log(P_{i,j}^{or}) + (1 - y_{i,j}) log(1 - P_{i,j}^{or}) \tag{17}$$

上述步骤中，$P_{i,j}^{sr}$表示第i个token为主体开始词，并且与第j个关系存在关联的条件概率；$P_{i,j}^{or}$表示第i个token为客体开始词，并且与第j个关系存在关联的条件概率。

最后是L^{ht}，其为4.3.4中实体头尾全局对应矩阵对应的损失，具体的计算过程如下所示:

$$L^{ht} = \frac{-1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log(P_{i,j}^{ht}) + (1 - y_{i,j}) log(1 - P_{i,j}^{ht}) \tag{18}$$

其中，$P_{i,j}^{ht}$表示第i个token为实体开始词以及第j个token为实体结尾词的条件概率。上述公式中n为输入序列的长度，n^r为定义的关系数。

5 实验

5.1 实验的评估指标和参数设置

在实验中，本文选择Adam(Kingma and Ba, 2014)作为模型优化器，将学习率设置为5×10^{-5}，输入序列的最大长度为100，dropout设置为0.1，batch_size设置为8，epoch设置为100，主体和客体全局对应、实体和关系全局对应、实体头尾全局对应三个模块的阈值均设置为0.5，MLP中使用的激活函数是ReLU(Glorot et al., 2011)。此外，本研究以精确率(Precision)、召回率(Recall)、F1值(F1-Score)为实验的评估指标。

5.2 古汉语数据集描述

在数据集划分上，本文选择以7：2：1的比例，将古汉语数据集划分为训练集、验证集和测试集。下表为每个数据集的实体类型分布情况(表4)和三元组的分布情况(表5)描述。

实体类型	训练集	验证集	测试集
人名(PER)	8446	3167	1634
官职(JOB)	3828	1535	740
地名(LOC)	1752	664	346
组织(ORG)	474	142	104
总计	14500	5408	2824

Table 4: 各个实体类型在训练集、验证集和测试集上的样本分布情况

数据集类型	三元组数
训练集	7242
验证集	2754
测试集	1412
总计	11408

Table 5: 三元组分布情况

5.3 实验过程与结果分析

5.3.1 预训练模型选择

为了选择一个合适的预训练模型参与本文的实验，本文挑选了Guwen-Bert[4]、RoBERTa-classical-chinese(Koichi Yasuoka, 2022)、SiKuBERT、SiKuRoBERTa(Wang et al., 2022)和bert-base-chinese-ner[5]五个预训练模型，然后让它们分别与OurModel模型进行搭配组合，最后将组合后的模型分别在古汉语数据集上进行实验。其中，Guwen-Bert、RoBERTa-classical-chinese、SiKuBERT和SiKuRoBERTa是在大量古汉语语料上训练的预训练模型，而bert-base-chinese-ner则是基于命名实体识别任务训练的预训练模型，本次实验均只使用了它们最后一

[4]https://github.com/ethan-yt/guwenbert
[5]https://github.com/ckiplab/ckip-transformers

层输出的隐藏层表示。实验的最后结果如表6所示。从最后的F1值上看，bert-base-chinese-ner+OurModel的组合在古汉语数据集上的表现要优于其它组合，OurModel为本文构建的模型。

预训练模型	测试结果
Guwen-bert	65.0
RoBERTa-classical-chinese	65.5
SiKuBERT	64.9
SiKuRoBERTa	65.7
bert-base-chinese-ner	**67.0**

Table 6: 各类预训练模型在OurModel上的F1(%)值对比

5.3.2 不同方法在古汉语实体关系数据集上的性能对比

为了探究本文提出的方法和基线方法在古汉语实体关系数据集上的性能表现，本文选取了基于其它解码方式的CasRel (Wei et al., 2020)和SPN4RE(Sui et al., 2021)，以及采用矩阵解码方式的OneRel(Shang et al., 2022)、PRGC(Zheng et al., 2021)和TPLinker(Wang et al., 2020)五种基线模型，与OurModel(本文的方法)进行对比实验。为了便于对比，各个方法使用的预训练模型均为5.3.1中F1值最好的bert-base-chinese-ner，最终各个方法在古汉语数据集上的结果如表7所示。

模型名称	精确率	召回率	F1值
CasRel	48.0	33.0	39.1
SPN4RE	51.1	40.6	45.2
OneRel	62.4	47.3	53.8
PRGC	69.0	52.1	59.4
TPLinker	74.8	**58.4**	65.6
OurModel	**81.6**	56.8	**67.0**

Table 7: 五种基线模型和OurModel在古汉语数据集上的精确率(%)、召回率(%)和F1值(%)。其中，OurModel为本文构建的模型

任务名称	模型类别	精确率	召回率	F1值
r	TPLinker	88.1	**63.5**	**73.8**
	OurModel	**91.9**	59.9	72.5
(s,o)	TPLinker	77.8	**60.6**	68.1
	OurModel	**85.1**	59.0	**69.7**

Table 8: OurModel和TPLinker在不同任务上的精确率(%)、召回率(%)和F1值(%)对比。其中r表示关系，s表示主体，o表示客体

从表8的结果可知，首先，相较于基于其它解码方式的方法来说，基于矩阵解码方式的方法在古汉语数据集上均获得了更好的结果。本文认为这是由于全局对应矩阵本质上是让模型学习实体与实体或者实体与关系的内在关联，相较于其它方式来说，更能缓解实体与实体、实体与关系匹配出现的冗余问题。

其次，OurModel在古汉语数据集上的F1值比最好的基线模型TPLinker要高出1.4%。为了进一步探究OurModel在古汉语数据集中表现优于基线模型的原因，在主客体对识别和关系抽取两个任务上，本文对比了基线模型中F1值最好的TPLinker和本文提出的模型的性能，结果如表9所示。

从表8中我们还可以得知，本文的方法虽然在关系抽取任务上比TPLinker方法在F1值上低1.3%，但是在主客体对识别任务上则高出1.6%，且在主客体对识别任务上本文的方法在精确率上比TPLinker高出7.3%，于是本文推测OurModel优于其它方法的原因可能是：相对位置信息可以提升实体识别的精度。

5.3.3 消融研究

为了检验相对位置信息对OurModel最终性能的影响和探究引入相对位置信息方法的泛化能力，本文分别在两个古汉语数据集上对相对位置信息模块进行消融实验。其实验结果如表9和表10所示。

从表9和表10我们可以看出，当引入相对位置信息时，在《资治通鉴》数据集上，模型的精确率上升了9.7%，召回率上升1.8%，F1值上升了4.7%。而在“二十四史”数据

	精确率	召回率	F1值
不加相对位置信息	71.9	55.0	62.3
加相对位置信息	**81.6**	**56.8**	**67.0**

Table 9: 在《资治通鉴》数据集上，OurModel在两种情况下的精确率(%)、召回率(%)和F1值(%)对比

	精确率	召回率	F1值
不加相对位置信息	9.2	**37.6**	14.9
加相对位置信息	**12.0**	36.8	**18.1**

Table 10: 在“二十四史”数据集上，OurModel在两种情况下的精确率(%)、召回率(%)和F1值(%)对比

集(Wang., 2021)上，模型的精确率上升了2.8%，F1值上升了3.2%。从表11中可以看出，当不加入相对位置信息时，模型容易生成长度较长的错误实体。综合上述结果，本文得出：当模型缺乏相对位置信息时，会极大削弱模型对实体长度的约束，导致模型预测出长度过长的错误实体，从而使得模型的精确率和召回率下降，这也进一步验证了相对位置信息能提升实体识别的精度的假设。同时，结合表9和表10的结果，我们可知：加入相对位置信息的方法在两个古汉语数据集上均使得模型最终的性能有所提升，这证明了引入相对位置信息的方法在古汉语实体关系抽取任务上的泛化性。

输入文本	实际标注的三元组	加入相对位置信息	不加入相对位置信息
魏相州刺史中山文庄王熙，英之子也，与弟给事黄门侍郎略、司徒祭酒纂。	(纂,任职,司徒祭酒) (熙,任职, 文庄王)	(纂,任职,**司徒祭酒**) (熙,任职,**文庄王**)	(纂,任职,**司徒祭酒纂，皆为清河王**) (熙,任职,**中山文庄王**)
祜官至尚书左仆射，爵新平王。	(祜,任职,尚书左仆射) (祜,任职, 新平王)	(祜,任职,**尚书左仆射，爵新平王**)	(祜,任职,**尚书左仆射**) (祜,任职, **新平王**)

Table 11: 两种不同情况下OurModel的结果对比

6 总结与展望

本文针对目前古汉语实体关系数据集存在的标注稀疏问题，基于对《资治通鉴》语料和目前在古汉语数据标注的研究，人工标注了一份实体关系数更加丰富的古汉语实体关系数据集，并构建了结合全局对应矩阵和相对位置信息的古汉语实体关系联合抽取模型。针对全局对应矩阵的方法容易生成较长实体的问题，本文尝试引入了字与字之间的相对位置信息的方法，最后通过在古汉语数据集上的实验证明了该方法的有效性。

然而本研究目前构建的《资治通鉴》具有标注规模较小、个别关系样本数较少等特点，这对模型最终的性能产生了一定程度的影响。此外，本研究所使用的引入相对位置信息的方法目前仅在两个古汉语数据集上进行了验证，这使得本文在对该方法的泛化能力的研究上存在不足。因此，下一步，本研究将继续研究不同类别的古汉语语料，以期建立更大规模的古汉语实体关系抽取标注数据集，进一步提升古汉语实体关系抽取任务的性能。

参考文献

Brin S. 1998. Extracting Patterns and Relations from the World Wide Web. In *International workshop on the world wide web and databases.*(pp. 172-183). Berlin, Heidelberg: Springer Berlin Heidelberg.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of ACL.*

Diederik P. Kingma and Jimmy Lei Ba. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies., Volume 1 (Long and Short Papers), pages 4171–4186.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini 2017. Creating training corpora for nlg micro-planners. In *Proceedings of ACL.*

Xavier Glorot,Antoine Bordes, and Yoshua Bengio 2011. Deep Sparse Rectifier Neural Networks. *Journal of Machine Learning Research*, pages 315-323.

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Ziran Li, Ning Ding, Zhiyuan Liu, Hai-Tao Zheng, and Ying Shen 2019. Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge. In *Proceedings of ACL.*

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of ACL.*

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu 2019. DuIE: A Large-scale Chinese Dataset for Information Extraction. *Natural Language Processing and Chinese Computing: 8th CCF International Conference*, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. Springer International Publishing, pages 791-800.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1):3–26.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*. pages 1015-1024.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML-PKDD.*

Rink, Bryan, and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation.*pages 256–259.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2021. Joint Entity and Relation Extraction with Set Prediction Networks. *arXiv preprint arXiv:2011.01675.*

Xin Wang, Zijing Ji, Yuxin Shen, Qingyan Guo, Yang Sun, Guanzhong Liu, Zijun Wang, Yining Sun, and Tian Yu. 2021. C-CLUE: A Benchmark of Classical Chinese Based on a Crowdsourcing System for Knowledge Graph Construction. In*Proceedings of CCKS.*

Yu-Ming Shang, Heyan Huang, and Xian-Ling Mao. 2022. OneRel: Joint Entity and Relation Extraction with One Module in One Step. In *Proceedings of AAAI.*

Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, Kazunori Fujita. 2022. Designing Universal Dependencies for Classical Chinese and Its Application. *Journal of Information Processing Society of Japan*, 63(2): 355-363.

Zhepei Wei, Jianlin Su,Yue Wang,Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of ACL.*

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking. In *Proceedings of COLING.*

Jingjing Xu, Ji Wen, Xu Sun, Qi Su. 2017. A Discourse-Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text. *arXiv:1711.07010.*

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of ACL.*

Suncong Zheng, Feng Wang, Hong yunBao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of ACL.*

Huan Zhang, Yuan Zong, Baobao Chang, Zhifang Sui, Hongying Zan, and Kunli Zhang. 2020. 面向医学文本处理的医学实体标注规范(Medical Entity Annotation Standard for Medical Text Processing). In *Proceedings of CCL.*

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Ming Xu, and Yefeng Zheng. 2021. PRGC: Potential Relation and Global Correspondence Based Joint Relational Triple Extraction. In *Proceedings of ACL.*

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

王东波，刘畅，朱子赫，刘江峰，胡昊天，沈思，李斌. 2022. SikuBERT与SikuRoBERTa: 面向数字人文的《四库全书》预训练模型构建及应用研究. 图书馆论坛,42(06):31-43.

王一钒，李博，史话，苗威，姜斌. 2021. 古汉语实体关系联合抽取的标注方法. 数据分析与知识发现,5(9):63-74.

数字人文视域下的青藏高原文旅知识图谱构建研究 ——以塔尔寺为例

李鑫豪　　赵维纳　　赵婉亦　　李超群
青海师范大学计算机学院，西宁810001
1020603546@qq.com　　490333294@qq.com　　337081897@qq.com
lcq513@163.com

摘要

青藏地区多元的民族构成以及悠久的历史沉淀孕育出丰富且独特的青藏文化，使得这片雪域圣地焉然成为了“高原文化宝库”。然而受闭塞的交通条件和较滞后的经济水平的限制，青藏地区文旅资源的保护与弘扬工作始终处于滞后状态。本文以数字人文为导向，在提示学习框架下采用联合学习的方式对文本中实体与关系的抽取，实现低资源条件下的知识抽取，形成一套文旅知识图谱构建范式，并以全国重点文物保护单位‘塔尔寺’为代表，完整的介绍了塔尔寺知识图谱从本体设计、原始数据获取、知识抽取到可视化展示的详细流程。最终，本文所构建的塔尔寺知识图谱共包含4705个节点及17386条关系。本文的工作弥补了人文领域青藏文化的结构化数据不足的问题，同时为青藏文旅在数字人文领域的研究提供参考。

关键词： 青藏文化；提示学习；联合抽取；塔尔寺知识图谱

Research on the Construction of Cultural and Tourism Knowledge Atlas on the Qinghai-Tibet Plateau from the Perspective of Digital Humanity ——A case study of Kumbum Monastery

Xinhao Li　　Weina Zhao　　Wanyi Zhao　　Chaoqun Li
School of computingQinghai Normal University，Xining810001
1020603546@qq.com　　490333294@qq.com　　337081897@qq.com
lcq513@163.com

Abstract

The diverse ethnic composition and long history of the Qinghai-Tibet region have bred a rich and unique Qinghai-Tibetan culture. Making this snowy sacred place a treasure trove of plateau culture ”. However, due to blocked traffic conditions and lagging economic level, the protection and promotion of cultural and travel resources in Qinghai-Tibet region is always lagging behind. This paper, guided by digital humanities, implements the extraction of entities and relationships in text by means of joint learning under the prompting learning framework, achieves the extraction of knowledge from low resources, forms a set of cultural and travel knowledge graph construction paradigm, and takes the national key cultural relics protection unit ’Kumbum Monastery’ As a representative, the detailed process from ontology design, original data acquisition, knowledge extraction to visual display of the knowledge graph of Kumbum Monastery is introduced. Finally, the knowledge map of Tar Temple built in this paper contains 4705 nodes and 17386 relations. The work of this paper makes up for the shortage of structured data of Tibetan culture in the field of human culture, and provides reference for the study of Tibetan travel in the field of digital human culture.

第二十二届中国计算语言学大会论文集，第253页-第263页，哈尔滨，中国，2023年8月3日至5日。

Keywords: Qinghai-Tibet Culture , Prompt learning , joint extraction , Kumbum Monastery knowledge graph

1 引言

青藏地区拥有着多元的民族构成以及悠久的历史沉淀，定居于青藏高原的50余个民族之间产生的文化交融与碰撞，促成了青藏地区人文历史的发展演变。“丝绸之路”南线的青海道、唐蕃古道、茶马古道(刘峰贵et al., 2012)留存于河湟谷地，其沿线分布着素有“世界屋脊之珠”的布达拉宫，藏传佛教圣地塔尔寺，被誉为“海藏咽喉，茶马商都”的丹噶尔古城以及“藏式建筑的千古典范”的大昭寺。除此之外青藏高原还拥有大量的非物质文化遗产，藏戏、藏医、热贡艺术这些弥足珍贵的遗产对中华传统文化产生着长期、广泛而又深刻的影响。这片辽阔的土地所孕育出的风格独特、形式多样的民族文化、历史文化、宗教文化和民俗文化焉然构成了“高原文化宝库”。然而受闭塞的交通条件与欠发达的经济水平的限制，青藏地区的文化保护与弘扬工作处于较为滞后状态。如何开展青藏文化的保护与传承工作是引人深思的。

数字人文是将现代化数字技术融入人文学科研究的一种典型的文理交叉领域学科(翁冉et al., 2023)，具体而言就是将已有的大量数字化资料，借助计算机技术的辅助分析，从海量数据中发现隐藏在数据中的模式、知识和趋势，揭示事物发生与发展的规律，对未来发展进行预测。随着数字人文研究愈来愈受关注，青藏高原的文旅产业也迎来了新的机遇。

知识图谱（Knowledge Graph）是一种新的知识组织方式，它采用（实体，关系，实体）的三元组形式进行表达和组织(Paulheim and Heiko, 2017)，如同蛛网一般链接起海量的异构信息，得易于其天然的图结构，知识图谱在查询与推理方面拥有显著优势。当前，知识图谱在文化领域得到了初步应用，在传统文化保护与传承、旅游推广等领域发挥着越来越重要的作用。如(聂欣晗and 张亮玉, 2021)利用知识图谱对《红楼梦》、《三国演义》等经典著作进行知识组织、分析与推理，在数字人文视觉下实现对传统文化的深入挖掘。可以看出，知识图谱在文化领域有重要的应用前景。然而知识图谱在青藏文旅领域的应用还较少，本文初步探索以知识图谱为代表的数字人文技术在青藏高原文旅中的应用。

具体地，本文提出一套文旅知识图谱构建范例，并在提示学习框架下利用联合学习实现实体与关系的抽取，探索低资源情况下知识的抽取。同时，在预训练语言模型的基础上进行链接，实现不同数据源的知识融合。实验结果表明，该方法能够在少量标注数据的情况下，实现高质量的知识抽取及融合，形成文旅知识图谱。最后，利用Neo4j数据库实现了知识图谱的存储与展示。本工作的意义在于为青藏文化保护工作的开展提供了新模式，弥补了青藏文旅数据的结构化不足的问题，从数字人文的角度实现对传统文化的传承与保护。

2 相关工作

2.1 数字人文

近年来,如何利用数字化技术激发创新创造活力，推动文化旅游产业高质量发展已成为一项重要课题，研究人员对数字人文进行了不同领域及层次的研究，(孙乃荣et al., 2022)以河北省文化和旅游外宣网站切入点，针对当前网站存在的问题，提出相应解决方法，从而助力河北文旅产业的发展。(刘泽权, 2021)通过重新界定名著重译的概念、对其评价方法进行梳理与审视，并从数字人文视角出发，提出了名著重译等级评价模型及相关变量，最后以《老人与海》、《红楼梦》和《香菱学诗》为例展示了模型的操作流程。(张卫et al., 2021)以古诗文本为例，利用机器学习与深度学习实现面向汉语诗文及其鉴赏的大规模人文情感术语的自动化抽取与分析。并得出将现代鉴赏融入古诗原文可显著优化情感知识的广度与深度的结论。

基金项目：省部共建藏语智能信息处理及应用国家重点实验室自主课题基金项目面向藏文古文献知识图谱构建（2022-SKL-012）国家自然基金项目汉藏双语藏医药知识图谱构建技术研究（62266036）国家社科基金项目藏汉双语藏文古籍知识图谱构建研究（22BTQ010）
通讯作者：赵维纳

2.2 文旅知识图谱

文旅领域知识图谱的研究同样备受关注。旅游方面，(刘济源, 2019)提出一种改进的本体构建方法利用信息抽取与知识融合技术从多源异构的数据中构建出旅游知识图谱，以此为基础设计出旅游问答系统，进一步改进设计出智能旅行助手。(张宇飞et al., 2022)利用爬虫技术从网络中获取河北省旅游景区的原始数据，将景区的信息进行分析、整合，构建出河北省旅游景区知识图谱，从而解决了河北省景点数量多分布广、信息杂糅且线上搜索智能化不足的问题。文化方面，(周莉娜et al., 2019)通过调研唐诗领域知识服务需求，设计出唐诗本体模型，对唐诗领域海量数据的语义化处理和存储构建唐诗知识图谱，基于该图谱搭建了面向领域知识服务的唐诗智能服务平台。虽然文旅领域知识图谱发展迅猛，但当前涉及青藏文化领域的知识图谱鲜有开展，随着中华文化保护与弘扬的提出，构建面向青藏文化领域的知识图谱成为一件具有重要意义的工作。

2.3 知识抽取

知识抽取是从半结构化或非结构化的文本中，抽取出机器可以理解和处理的知识，它主要包含实体识别、关系抽取、事件抽取三个子任务。基于深度学习方法的知识抽取,是近年的研究热点,其能够将低层特征进行组合,形成更加抽象的高层特征。

基于深度学习知识抽取早期通常采用流水式抽取法，即实体识别与关系抽取互相独立进行。虽然流水式抽取不断优化迭代，但其始终面临着两个子任务间存在错误传播、无关联实体间产生冗余信息以及信息丢失的问题。研究人员也因此开始尝试将实体与关系进行联合抽取。

联合抽取方法是通过实体识别和关系分类联合模型,直接得到存在关系的实体三元组。(Miwa and Bansal, 2016) 首次提出了基于深度学习的联合抽取模型，其利用双向顺序结构和双向树形结构的LSTM-RNNs进行实体和关系的联合建模。虽然该模型中的关系分类子任务和实体识别子任务仅共享了编码层的双向序列LSTM表示，但它的提出为日后真正意义上基于参数共享的联合学习奠定了基础。(Zheng et al., 2017)提出了基于新的标注策略的实体关系抽取方法,将原来涉及到命名实体识别和关系分类两个子任务的联合学习模型完全变成了一个序列标注问题。(Wang et al., 2020)提出了一种名为TPLinker（Token Pair Linking）的实体关系联合抽取标注方案，该方案可在一个模型中实现单阶段联合抽取，模型不存在曝光偏差，保证训练和测试的一致性。并且同时可解决多关系重叠和多关系实体嵌套的问题。当前基于深度学习的联合抽取方式已经成为了关系抽取的主流方式。

3 数字人文视域下的青藏高原文旅知识图谱构建框架

青藏高原文旅类知识图谱的构建包括：本体构建、数据获取、模型训练、知识抽取、知识融合等流程。以下将以塔尔寺知识图谱构建为范例，分别按照步骤介绍构建流程：

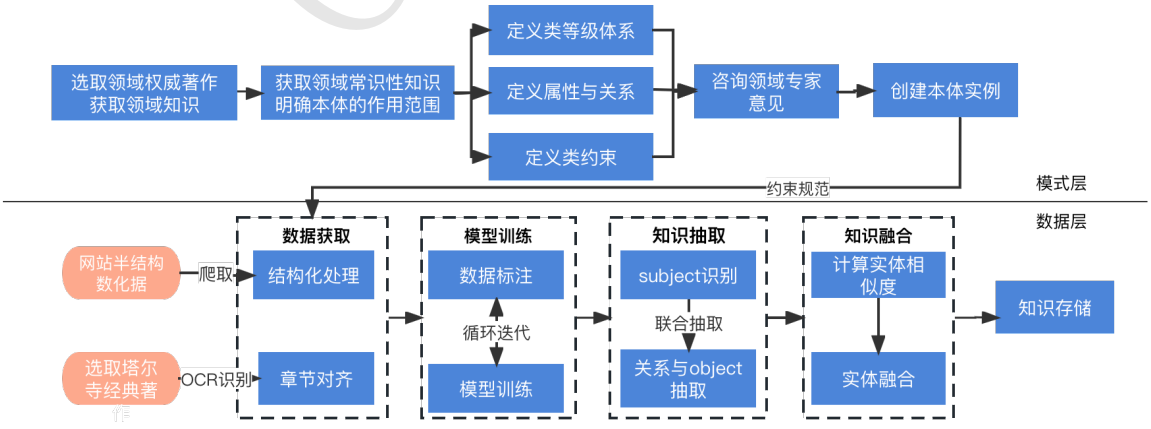


图 1: 青藏高原文旅知识图谱构建流程

3.1 数据源获取

本文以塔尔寺相关著作为数据源，利用OCR工具将《塔尔寺史话》、《圣地莲花塔尔寺》、《塔尔寺建筑艺术史》等著作的纸质文本转化为数字文本，并将不同著作中内容相关的章节进行章节整合，最后选取每个章节的前百分之二十进行标注，标注过程与训练过程迭代进行，模型的训练结果作为反馈对下一轮次的标注进行指导。为保障所构建的原始语料库尽可能完善，本文还利用爬虫技术在青海省图书馆塔尔寺知识平台和青海文化旅游厅爬取了有关青海省文化遗产的结构化数据，最终得到塔尔寺原始语料库。

3.2 本体设计

本体是对实体存在形式的描述，往往表示为一组概念定义和概念之间的层级关系，本体构建是知识图谱的构建的基础工作，本文将结合七步法(Abdellatif et al., 2017)与骨架法(傅柱et al., 2013)提出一种创新性的本体构建方式，为日后青藏文化领域相关知识图谱的模式层构建提供参考。

（1）选取要构建领域的权威著作或文献；通读著作，尤其需要针对著作内所存在的命名实体保持敏感，当一类实体频繁出现时需及时记录实体类型及其属性。以塔尔寺图谱构建过程所选取的《塔尔寺史话》为例，书中频繁出现人物、寺院、大师、佛像、法会等实体类型。

（2）获取领域常识性知识；若在通读著作的过程中遇到领域内常识性问题，需要适当的加以学习，这些领域内常识性问题往往对明确实体类型的边界，以及后续制定标注规则有着重要作用。同样以塔尔寺图谱构建过程所选取的《塔尔寺史话》为例，书中并未介绍大师、喇嘛、班禅以及活佛的区别，但这些实体类型对人物实体的层级划分以及后续标注规则的制定着重要影响。

（3）定义类及其等级体系、类属性、类约束等；通过先前的两个步骤，已经能够大致确定图谱的实体类型及其属性，在定义类的层级划分时可参考著作目录，不同类型的实体在目录中大致都会得到体现。塔尔寺知识图谱的构建过程中共定义实体31类，关系20种。

（4）咨询领域专家建议；本体构建作为图谱构建的基础工作，后续改动所造成的额外工作量十分庞大，为确保后续工作顺利进行，可向领域专家说明任务需求，咨询专家建议，对构建出的类以及关系体系进行修改。

（5）创建实例；在创建本体实例时可以利用Protégé软件，该软件拥有便捷的实体、关系以及属性的管理功能，本体实例创建完成后可进行界面化展示，直观的展示出各实体类型之间的关系。如图2展示了塔尔寺知识图谱的本体结构。

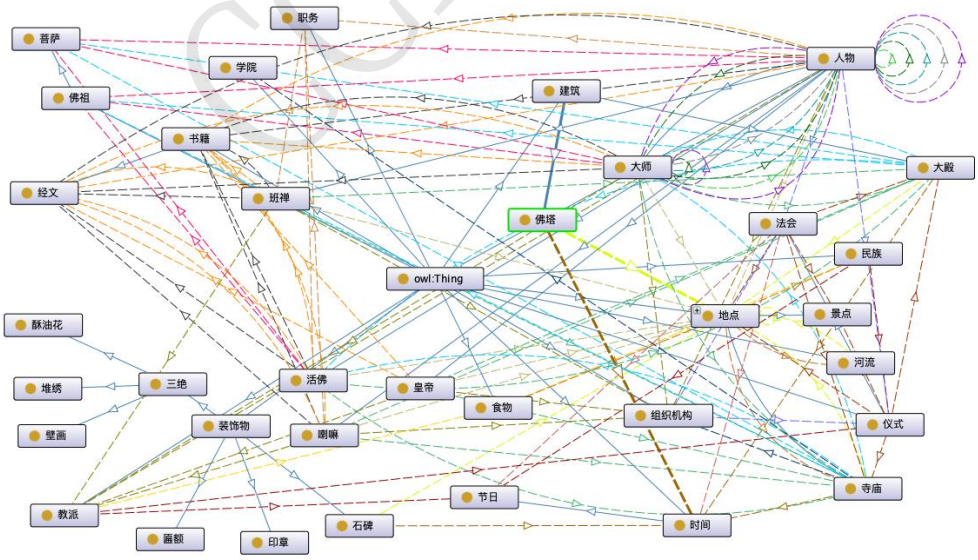


图 2: 塔尔寺图谱本体结构展示

3.3 数据标注

数据标注是知识图谱构建过程中最耗费人力与时间的环节，标注成员需具备一定的领域知识且对原始语料进行充足的了解，除此之外还要考虑不同成员的标注习惯以及词语所具有的多样性与歧义性的特性可能造成的标注差异。为此，我们提出了以下建议：在开始标注之前预先制定出一套标注规范，每一位成员充分学习标注规范并了解原始数据，利用标注平台提高标注效率。现有的标注平台种类繁多，不同平台所针对的细化任务也不尽相同，本团队选取了支持多人协同标注的LabelStudio标注平台。标注任务的前期工作包含：（1）对数据进行文本对齐以及章节划分等预标注操作；（2）将原始语料上传至标注平台并拆分为多个部分并分配给每一位标注成员；（3）在平台中创建本体构建任务所设计的31类实体和20种关系；（4）制定标注规范，对标注成员进行培训，提高标注团队结果的一致性。

图 3: LabelStudio标注示例

本团队在命名实体标注过程中制定了以下三点规范：（1）仅可标注已预先定义的31类实体；（2）同一字符串不能重叠标注；（3）除书籍实体的书名号外，尽量不标注其他符号。在关系标注过程中遵循以下两点规范：（1）仅可标注已预先定义的20类关系；（2）优先标注同一句子内的关系，若同一句子内没有关系，允许跨句标注。经过多轮迭代，最终标注成员在塔尔寺原始语料中累积标注命名实体10380个，实体关系3370条。

实体1	关系类型	实体2
人物	任职于	地点
		学院
		组织机构
		寺庙
		职务
	创建	组织机构
		学院
		仪式
		教派
	出生于	地点
		时间
	编著	经文
	供奉	佛祖
	亲属	人物

实体1	关系类型	实体2
地点	位于	地点
	属于	地点
	临近	地点
	拥有	寺庙
		组织机构
		民族
	举行	仪式
寺庙	位于	地点
	供奉	佛祖
	设立	组织机构
		学院
	拥有	大殿
		佛塔
		石碑

表 1: 部分实体与关系展示

3.4 知识抽取

知识抽取是构建知识图谱的核心环节也是难点，本图谱的抽取工作采用了一种基于文本生成的联合抽取模型，在此期间面临着以下问题：（1）数据集规模较小;（2）类别分布高度不平衡，例如“人物”与“寺庙”实体类型的占比远高于“节日”和“职务”类型;（3）三元组重叠如“塔尔寺位于青海省省会西宁市”，其中青海省和西宁市都是塔尔寺所在地。鉴于这些问题，本工作利用(Sun et al., 2019)Erine预训练模型作为编码层结合(Wei et al., 2020)指针网络进行知识联合抽取(Lu et al., 2022)从而解决了类别分布不匹配以及三元组重叠问题，并通过引入提示学习（Prompt learning）(Liu et al., 2021)策略缓解数据标注量较少的问题。

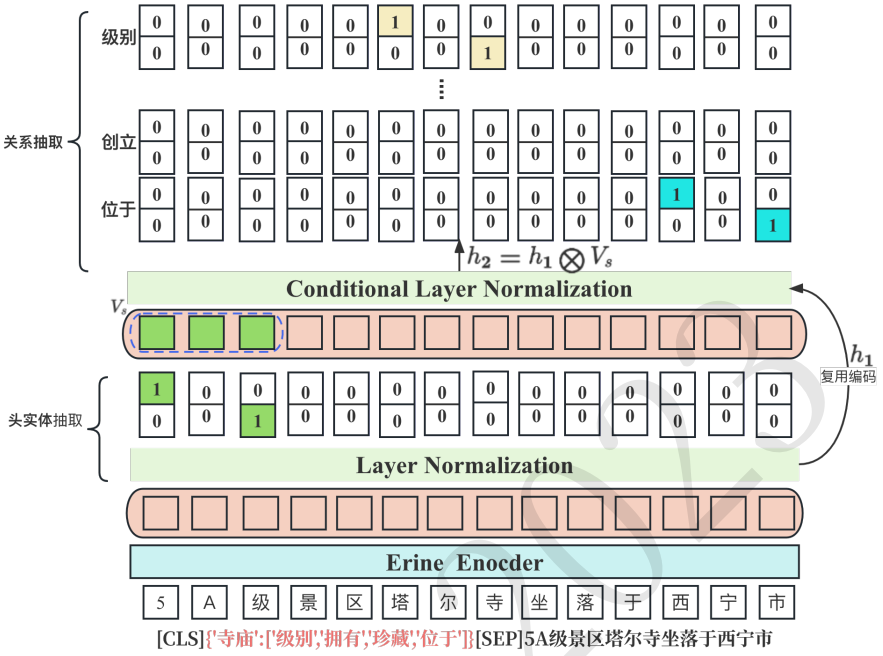


图 4: 基于提示学习与指针网络的联合抽取模型

知识抽取的流程如图4所示，首先依据需要抽取的对象构建模版，模版与文本拼接作为Encoder的输入从而获取输入语句的每一个字符的特征向量表示。之后对实体识别与关系抽取进行建模利用句子的编码信息，抽取出三元组头实体(subject)，具体实现是通过对每个token，使用两个独立的二分类器预测其是否为实体的开始（p^{s_start}）或实体的结束(p^{s_end})，并设定一个阈值，若大于阈值则标记该token为1，否则标记为0。计算如公式（1），公式（2）所示。

$$p^{s\_start} = \sigma(\omega^{s\_start} h_1^i + b^{s\_start}) \tag{1}$$

$$p^{s\_end} = \sigma(\omega^{s\_end} h_1^i + b^{s\_end}) \tag{2}$$

其中，$\omega^{(*)}$可训练权重，$b^{(*)}$ 表示可训练偏置，σ 表示sigmoid激活函数，h_1为句子编码。主体抽取后进行关系与客体的联合抽取，首先subject的编码V_s与句子编码h_1进行特征结合得到新的句子编码h_2，然后利用新的句子编码h_2计算当前subject对应于每一种关系下object的起始token。其抽取过程采用的计算公式与subject抽取计算过程相同，不再赘述。

	输入	输出
1	[CLS]寺庙[SEP]5A级景区[MASK]坐落于西宁市	塔尔寺
2	[CLS]地点[SEP]5A级景区塔尔寺坐落于[MASK]	西宁市
3	[CLS]{’寺庙’:[’级别’,’拥有’,’珍藏’,’位于’]}[SEP]5A级景区[MASK]坐落于西宁市	（塔尔寺，位于，西宁市） （塔尔寺，级别，5A级）

表 2: 模版示例

MLM（masked language model）是将一句话中的某个字符用[MASK]替换掉，而后用模型预测这句话中的每一个词，最后利用模型预测被[MASK]替换的词实际上是什么。提示学习（Prompt learning）是伴随输入，给予模型的一个提示模版，用以指导模型接下来应当要做什么任务。也就是说，Prompt learning能够将下游任务改造成预训练模型期望的样子，从而提升模型的表现。在进行知识抽取任务之前，需根据要抽取的对象，制定提示模版，如表2所示，提示模版嵌入在输入语句的头部形成“头模版”，“头模版”可充分激发语言模型的文本生成能力，提升模型的抽取表现。

模版与文本拼接组成Erine（Enhanced Representation through knowledge Integration）模块的输入。其结构与Bert相似，由Embedding层以Transformer层组成，如图6所示。Erine采用了一种如表3所示改进后的MLM策略，将原本以字为单位的Mask方法变为对整个汉语单词Mask，即屏蔽整个词语而不是屏蔽汉字。相较于Bert学习原始语言信号，Erine直接对先验语义知识单元进行建模，增强了模型语义表示能力。

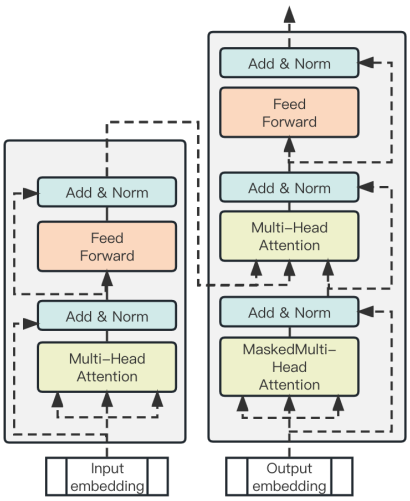


图 5: Transformer模型结构

说明	样例
语句	5A级景区塔尔寺坐落于西宁市
原MASK策略	5[Mask]级景区塔尔[Mask]坐落于西[Mask]市
WWM策略	[Mask][Mask][Mask]景区[Mask][Mask][Mask]坐落于西宁市

表 3: 全词掩码示例

实体关系抽取的标注策略包含序列标注法和指针标注法，序列标注法的天然的缺陷在于，当实体与上下文中的多个实体存在关系时，只将关系分配给最近的实体(马建红et al., 2021)，无法解决解决三元组重叠的问题。如图5所示，序列标注只能识别出塔尔寺位等级为5A级，无法继续识别出塔尔寺位于西宁市。

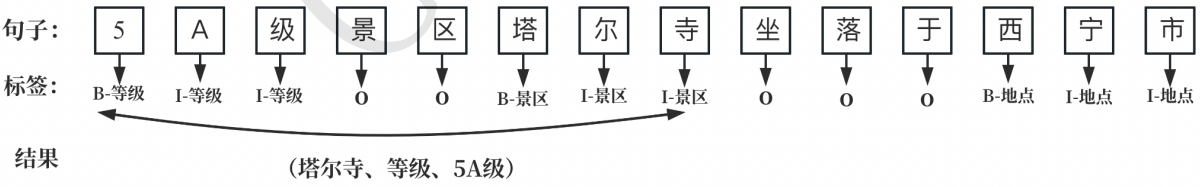


图 6: 序列标注法

本模型的指针网络中采用“0”、“1”对句子中的实体进行标注，具体如图4所示，实体的开始token和结束token标注为“1”，剩余token 标注为“0”。通过建立级联结构，可重复标注塔尔寺对应token，以此有效的解决了三元组重叠问题。

3.5 知识融合

由于本图谱原始语料来自于不同著作，不同著作中对同一个客观实体的表述存在差异，对于获取到的知识，需要进行融合加以关联。本图谱的知识融合任务主要是将不同著作中对同一实体的不同语义表述，关联到同一客观实体上，从而实现同一实体的多名语义消歧。具体的我

们首先统计出各个实体的词频，利用Bert学习实体的表示，融合word2vec的表示和编辑距离(梅筱and 刘海鹏, 2010)一起来计算两个实体的相似度。如果两实体的相似度大于设定的阈值，则将词频较低的实体融合近词频较高的实体中。

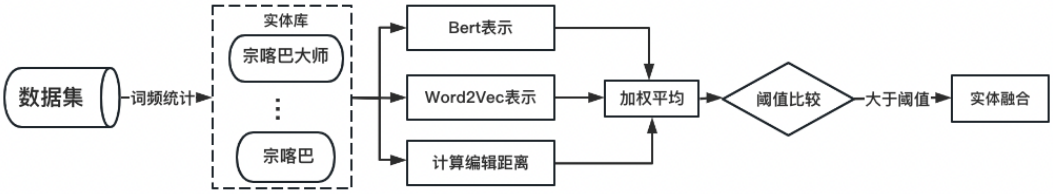


图 7: 知识融合流程

如图7所示，首先统计出实体"宗喀巴大师"与"实体宗喀巴"的频率，经过对词语的Bert表示，以及word2vec的表示和编辑距离进行加权平均，得出低频词"宗喀巴大师"与高频词"宗喀巴"的相似度大于阈值，则认为其二者表述的是客观世界的同一实体，最后将"宗喀巴大师"实体融合到词频更高的"宗喀巴"实体中。

4 塔尔寺知识图谱构建

4.1 实验

本实验采用Human-in-loop方式(Wang et al., 2021)，即数据标注与模型训练循环交替进行，每一轮标注后对模型进行一次训练，一次标注加上一次模型训练构成一轮迭代，凭借这种方式可清晰的观察出数据集标注量对实验效果的影响，并且可以依据上一轮次的实验结果，开展新一轮次更具针对性的标注工作。

本文采用精确率（P），召回率（R）以及F1-score（F1）值作为模型结果的评判指标。具体计算公式如公式（3）、公式（4）、公式（5）所示，其中，TP表示模型能正确检测出的实体个数、FP表示模型检测到的无关实体个数、FN表示模型未检测到的实体的个数。

$$P = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{5}$$

本实验使用Tesla PG503-216显卡对模型进行训练，显卡内存128G，服务器系统Linux ubuntu 5.4.0-125-generic，python版本3.7.16，训练框架paddlepaddle-gpu 2.4.1。

经过5轮迭代后。标注成员最终在塔尔寺原始数据集中共标注命名实体10380个，实体关系3370条。各实验周期的结果如表4所示。

轮次	实体标注量	关系标注量	实体抽取结果			关系抽取结果		
			P	R	F1	P	R	F1
迭代-1	4062	835	55%	62.8%	58.6%	52.2%	64.7%	57.8%
迭代-2	6015	1346	59.3%	61.2%	62.5%	60.3%	66.6%	61.5%
迭代-3	8093	1677	64.2%	60%	62.1%	68.4%	68.4%	68.4%
迭代-4	9052	2033	77.7%	87%	82.3%	64.9%	85.7%	80%
迭代-5	10380	3370	86.7%	86.1%	86.4%	80.6%	83.3%	81.9%

表 4: 各迭代阶段实验结果

实验结果表明，伴随数据标注量的逐渐增加，实验的结果也在稳步提升。最终实体抽取任务与关系抽取任务在各项评价指标下的表现都超过了80%，不同本体类型与不同关系类型之间

的抽取表现相差较大，其抽取表现与权重成正相关关系，本图谱中核心类型实体与关系的实验结果如表5所示。

实体	P	R	F1	weight	关系	P	R	F1	weight
学院	95.2%	76.9%	85.1%	3.7%	珍藏	100%	67%	43%	7.4%
民族	93.3%	56%	70%	2.1%	编著	100%	50%	67%	4.3%
寺庙	86.7%	83.3%	84.8%	11.7%	位于	94.5%	62.2%	66.7%	11.1%
法会	85.7%	66.7%	75%	8.9%	供奉	89.4%	56.2%	70%	6.2%
教派	84.6%	78.5%	81.4%	2.9%	临近	83.6%	79.7%	81.2%	19%
大师	83.3%	71.4%	76.9%	7.2%	称呼	82.9%	87.1%	85%	5.5%
时间	82.3%	80.1%	81.2%	5.7%	拥有	80.1%	66.7%	72.7%	6.6%
佛像	81.4%	84.6%	83.1%	9.7%	属于	77.7%	77.7%	77.7%	7.3%
书籍	80.6%	83.3%	82%	6.6%	生于	75%	51.7%	61.2%	2.2%
人物	77.8%	87.5%	82.3%	14.6%	亲属	70.4%	51.1%	59.3%	3.1%

表 5: 实体关系抽取结果

4.2 对比实验

为验证实验的有效性，本实验分别与流水式抽取以及联合抽取相比较。流水式抽取中采用Bert结合Bilstm以及条件随机场(CRF)作为实体抽取任务的模型，并采用Bert和MLP作为关系抽取任务的模型。联合式抽取采用先利用Global Pointer抽取subject的首尾(i, j)和object的首尾位置(i, j)，然后利用Global Pointer抽取每一种关系p的实体的头部所对应的位置(hi, hj)和实体的尾部tail所对应的位置(ti, tj)组合，最终输出交集的结果。从表6可以看出，联合抽取的表现均优于流水式抽取，GPlinker相对流水式抽取的表现有了明显提升。而本模型使得实体抽取以及关系抽取的表现进一步的提升，图8展示了对比实验在关系抽取任务中的表现。

方式	模型	任务类型	F1
Pipline	Bert+Bilstm+CRF	实体抽取	70.5%
	Bert+MLP	关系抽取	69.5%
Joint	GPlinker	实体抽取	84.9%
		关系抽取	77.7%
Joint	本文模型	实体抽取	86.4%
		关系抽取	81.9%

表 6: 对比实验结果

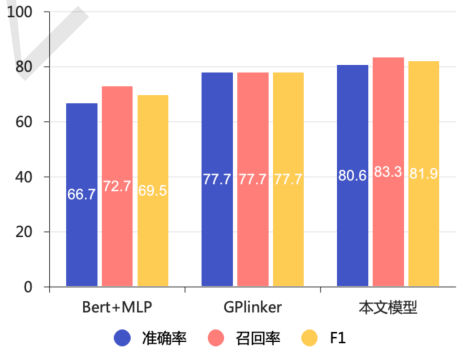


图 8: 对比实验关系抽取结果

4.3 错误分析

经过对实验结果和实验语料进的分析，可得出错误原因主要由以下两方面造成。首先是关系分布的不平衡，塔尔寺知识图谱所定义的20类关系子类型中，‘位于’和‘临近’等9种常见关系的语料数量占比达到了百分之70，其余11类关系对应语料的总和占比仅百分之30，这导致了大部分的关系和实体对的可用样例仍然较少，而神经网络模型作为典型的需要大量训练数据支撑的技术，在训练样例过少的情况下各项评价指标都受到极大影响。其次塔尔寺知识图谱原始语料的非结构化信息比较复杂，著作中包含较多的代词和长难句，这些对信息抽取都造成了负面影响。实验结果中有部分关系类型例如’珍藏’与‘编著’虽占比较少，但也取得了较好的抽取表现，这是由于该类实体在语料中集中分布于特定章节。

4.4 知识存储

本图谱利用Neo4j图数据库作为知识存储工具，实现了知识查询与更新以及界面化展示等功

能(冯俐, 2019)，可为图谱的下游互联网任务提供支持。塔尔寺部分知识在数据库中的存储形式如图8所示，经过人工校对，剔除掉6709条错误信息，最终本图谱针对塔尔寺31类实体以及20种关系所构建的知识图谱共包含4705个节点及17386条关系。

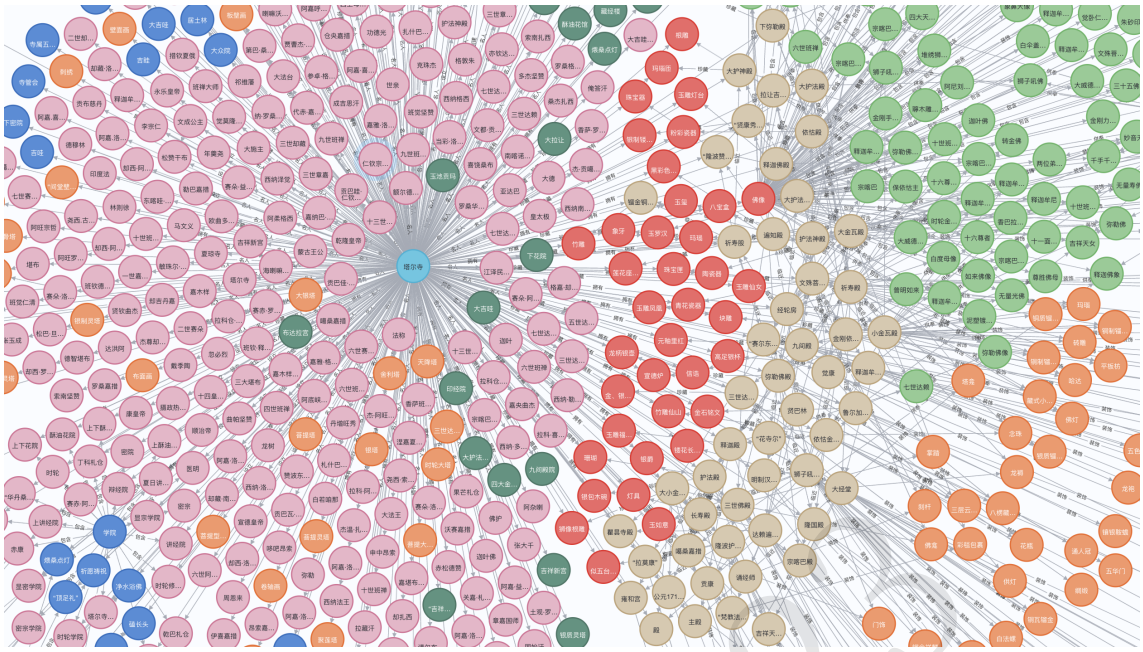


图 9: 塔尔寺知识图谱存储展示

4.5 图谱应用

伴随文旅产业的复苏，青海省的景区也再次进发出活力，2023年五一假期期间，塔尔寺景区游客量已经达到2019年同期的118.6%，面对激增的游客量，景区服务水平亟待提升。我们将通过引入更多的著作构建出足以支撑塔尔寺问答助手的知识图谱，该问答助手除了为游客介绍各个景点的风貌特征外还可以讲解塔尔寺的历史典故，并且该问答助手可以作为教育应用，帮助游客提前了解塔尔寺文化，推广塔尔寺丰厚的文化传播，实现塔尔寺文化的传承保护。

5 总结

本文的意义在于：1）立足数字人文，提出一套针对青藏文化保护工作的范例，为日后青藏文化保护工作的开展提供了新思路。2）将提示学习与基于文本生成式的实体关系联合抽取模型结和，实现了在低资源情况下对塔尔寺相关知识的抽取。3）完整的介绍了塔尔寺知识图谱的构建流程，弥补了人文领域有塔尔寺的结构化数据不足的问题，满足了互联网任务需求，实现了对传统文化的传承保护。

参考文献

M. Abdellatif, M. S. Farhan, and N. S. Shehata. 2017. Overcoming business process reengineering obstacles using ontology-based knowledge map methodology. *Future Computing Informatics Journal*, page S2314728817300296.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu. 2022. Unified structure generation for universal information extraction.

M. Miwa and M. Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures.

Paulheim and Heiko. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*.

Y. Sun, S. Wang, Y. Li, S. Feng, and H. Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*.

Z. J. Wang, D. Choi, S. Xu, and D. Yang. 2021. Putting humans in the natural language processing loop: A survey.

Z. Wei, J. Su, Y. Wang, Y. Tian, and Y. Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme.

傅柱, 王曰芬, and 孙铭丽. 2013. 本体存储技术研究综述. 情报理论与实践, 36(9):6.

冯俐. 2019. 基于neo4j图数据库构建中学语文诗词知识图谱. 陕西师范大学.

刘峰贵, 王锋, 张海峰, 周强, 陈琼, and 李春花. 2012. 青藏高原文化旅游资源开发探讨. 青海社会科学, (5):6.

刘泽权. 2021. 数字人文视域下名著重译多维评价模型构建. 中国翻译.

刘济源. 2019. 旅游领域知识图谱的构建及应用研究. Ph.D. thesis, 浙江大学.

周莉娜, 洪亮, and 高子阳. 2019. 唐诗知识图谱的构建及其智能知识服务设计. 图书情报工作, 63(2):10.

孙乃荣, 雷芳, 侯晓丹, and 陈杭. 2022. 数字人文视域下河北省文化和旅游外宣网站高质量发展对策研究. 西部旅游, (16):77–79, 8.

张卫, 王昊, 邓三鸿, and 张宝隆. 2021. 面向数字人文的古诗文本情感术语抽取与应用研究. 中国图书馆学报, 47(4):19.

张宇飞, 李腾, and 贾东立. 2022. 河北省旅游景点知识图谱的构建. 计算机与数字工程, (007):050.

梅筱and 刘海鹏. 2010. 基于编辑距离结合词性的词相似度算法.

翁冉, 何世群, 杨秀璋, and 罗子江. 2023. 国内数字人文领域热点及趋势探析. 河南图书馆学刊, 43(1):3.

聂欣晗and 张亮玉. 2021. 论四大名著的文化旅游开发. 旅游纵览, No.354(148-152).

马建红, 魏字默, and 陈亚萌. 2021. 基于信息融合标注的实体及关系联合抽取方法. 计算机应用与软件, 38(7):8.

基于互信息最大化和对比损失的多模态对话情绪识别模型

黎倩尔, 黄沛杰*, 陈佳炜, 吴嘉林, 徐禹洪, 林丕源
华南农业大学, 数学与信息学院, 广东广州, 510642
li@stu.scau.edu.cn, pjhuang@scau.edu.cn,jw_chen@stu.scau.edu.cn,
galinwu@stu.scau.edu.cn, xuyuhong@scau.edu.cn, pyuanlin@scau.edu.cn

摘要

多模态的对话情绪识别（emotion recognition in conversation，ERC）是构建情感对话系统的关键。近年来基于图的融合方法在会话中动态聚合多模态上下文特征，提高了模型在多模态对话情绪识别方面的性能。然而，这些方法都没有充分保留和利用输入数据中的有价值的信息。具体地说，它们都没有保留从输入到融合结果的任务相关信息，并且忽略了标签本身蕴含的信息。本文提出了一种基于互信息最大化和对比损失的多模态对话情绪识别模型MMIC来解决上述的问题。模型通过在输入级和融合级上分级最大化模态之间的互信息（mutual information），使任务相关信息在融合过程中得以保存，从而生成更丰富的多模态表示。本文还在基于图的动态融合网络中引入了监督对比学习（supervised contrastive learning），通过充分利用标签蕴含的信息，使不同情绪相互排斥，增强了模型识别相似情绪的能力。在两个英文和一个中文的公共数据集上的大量实验证明了所提出模型的有效性和优越性。此外，在所提出模型上进行的案例探究有效地证实了模型可以有效保留任务相关信息，更好地区分出相似的情绪。消融实验和可视化结果证明了模型中每个模块的有效性。

关键词： 多模态对话情绪识别；图卷积网络；互信息；监督对比学习

Multimodal Emotion Recognition in Conversation with Mutual Information Maximization and Contrastive Loss

**Qianer Li, Peijie Huang*, Jiawei Chen, Jialin Wu,
Yuhong Xu, Piyuan Lin**
College of Mathematics and Informatics, South China Agricultural University, China
li@stu.scau.edu.cn, pjhuang@scau.edu.cn,jw_chen@stu.scau.edu.cn,
galinwu@stu.scau.edu.cn, xuyuhong@scau.edu.cn, pyuanlin@scau.edu.cn

Abstract

Emotion recognition in conversation (ERC) is a key component for building emotional dialogue systems. In recent years, graph-based fusion methods have been proposed to dynamically aggregate multimodal context features in conversation, which improve the performance of models on multimodal emotion recognition in conversation. However, these methods do not fully preserve and utilize valuable information in the input data. Specifically, they do not retain task-relevant information from input to fusion result, and ignore the information implied by labels themselves. In this paper, to overcome

*通讯作者

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

the above issues, we propose a new model based on Mutual Information and Contrast(MMIC), for multimodal emotion recognition in conversation. The model maximizes the mutual information between modalities at input level and fusion level hierarchically, which preserves task-relevant information during fusion process and generates richer multimodal representations. We also introduce supervised contrastive learning into graph-based dynamic fusion network, which leverages the information implied by labels to make different emotions repel each other and enhances the model's ability to recognize similar emotions. Extensive experiments on two public benchmark datasets and a new Chinese dataset demonstrate the effectiveness of our proposed model. In addition, case studies conducted on the proposed model effectively confirmed that the model can effectively retain task-related information and better distinguish similar emotions. Ablation experiments and visualization results demonstrated the effectiveness of each module in the model.

Keywords: multimodal emotion recognition in conversation , graph convolutional network , mutual information , supervised contrastive learning

1 引言

情绪是人类日常交流的重要组成部分。对话中的情绪识别（ERC）旨在对话过程中自动识别和跟踪说话者的情绪状态，如图1所示。近年来，它越来越引起了自然语言处理和多模态处理领域研究者的关注。ERC具有广泛的潜在应用范围，如心理学、人机交互和社交机器人等领域。具体来说，对话系统想要和人类进行有效的情感沟通，就必须具备一定的情感能力，因此对用户情绪进行正确的识别判断显得十分关键(Zhou et al., 2018)。在ERC的背景下，充分利用现有数据中的有价值的信息是至关重要的，因为对话是一个复杂而动态的过程，情绪可以通过各种方式表达，如言语，面部表情和肢体语言等。通过充分利用这些不同形式的多模态信息，我们可以对说话者的情绪状态有更为完整的理解。

图 1. M^{3}ED数据集中的一个对话示例

社交媒体的快速增长是促使人们研究对话情绪识别的原因之一。对话情绪识别和传统情绪识别的区别在于，对话情绪识别是一个分类任务，旨在对一段对话中的话语进行情绪分类。任务的输入是一段连续的对话，输出是这段对话中所有话语的情绪。传统情绪识别是一种识别单独话语中不同人情感类型的方法。对话情绪识别比独立情绪识别更困难更有意义的地方在于，对话层面的情绪分析有助于理解说话者在整个说话过程中所涵盖主题的情绪动态变化(Zadeh et al., 2016)。此外，对话情绪识别可广泛应用于各种对话场景中，如社交媒体中评论的情感分析、人工客服中客户的情绪分析等。对话中通常蕴含了丰富的语境线索(Hu et al., 2019)，能否捕捉到语境线索对情绪识别至关重要。在以往与语境相关的研究中，对话中的文本模态是研究的重点，而这些研究常常忽略了其他模态的有效组合。在考虑多模态上下文信息的工作中，对

话通常被建模为序列或图结构，而特征连接则是通过简单的早期或是晚期融合方法(Majumder et al., 2019; Ghosal et al., 2019)实现的，缺乏对多模态信息的有效利用探索。

近年来，已经发表的著名研究是在图架构中模拟跨模态和单模态交互(Hu et al., 2021; Liu et al., 2021)，这为跟踪情绪模式的互补性提供了深刻的见解。研究(Hu et al., 2022)利用基于图的融合方法对来自不同语义空间的上下文信息进行动态建模，减少了连续聚合中的冗余信息。然而，这些方法不能充分利用数据中存在的有价值的信息，即从输入到融合结果的任务相关信息以及标签本身包含的信息。因此，这些方法无法捕获数据中底层关系和依赖关系的全部复杂性。我们认为，将关键任务相关信息从输入保留到融合结果中，可以提取和整合输入中的单模态原始数据，从而生成更丰富的多模态表示(Han et al., 2021)。此外，利用标签中蕴含的情绪信息，可以使相似的情绪之间相互排斥，从而更好地区分相似的情绪，如“沮丧”和“悲伤”，“快乐”和“兴奋” (Li et al., 2022)。

为了克服上述提及的问题，本文提出了一种基于互信息最大化与对比损失的多模态对话情绪识别模型——MMIC。为了更好地保存从输入到融合结果的任务相关信息，我们首先最大化了各模态表示之间的互信息。在通过图卷积运算得到融合结果后，我们进一步最大化它与低水平单模态表示之间的互信息。为了更充分地利用标签中隐含的信息，我们采用监督对比学习，使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥。我们在两个公共基准数据集和一个中文数据集上进行了大量实验，证明了所提出模型的有效性和优越性。此外，在所提出模型上进行的案例探究有效地证实了该模型可以有效保留任务相关信息，更好地区分出相似的情绪。消融实验和可视化结果进一步证明了该模型中每个模块的有效性。本文的主要贡献如下：

- 我们提出了一种基于互信息和对比损失的多模态对话情绪识别模型，实现了对输入数据的充分利用，提高模型在多模态对话情绪识别上的性能。
- 我们在输入级和融合级上分级最大化模态之间的互信息，更好地在融合过程中保存与任务相关的信息，从而生成丰富多模态表示。
- 我们将监督对比学习纳入基于图的融合网络，充分利用标签信息，使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥，从而增强对相似情绪的识别。
- 我们在两个英文公共基准数据集和一个新发布的中文数据集上进行了大量实验，验证了本文提出的模型取得了优于研究进展方法的效果。

2 相关工作

本文的研究通过充分保留和利用输入数据中有价值的信息，提高模型在多模态对话情绪识别上的性能。本节简要介绍相关技术方法，并阐述本文方案中融入这些技术方法的设计依据。

随着近年来社交媒体的快速进步，以及高质量拍摄设备的出现，我们见证了多模态数据的爆炸式增长，如电影、短视频等。在现实生活中，多模态数据通常由三个模态组成：视觉、声学和转录文本。一般来说，同一数据段中的不同模态往往是具有互补性的，为语义和情绪消歧提供额外的线索(Li et al., 2011)。多模态数据融合是情绪识别中至关重要的部分，模型要从所有输入数据中提取和整合信息，才能理解数据背后代表的情绪。因此，模型能否在过滤噪声的同时，生成具有带有任务相关信息的丰富多模态表示，对情绪识别十分关键。

早期多模态融合的典型代表是LSTM(Poria et al., 2017)和ICON(Hazarika et al., 2018)，它们通过拼接三种模态的特征来利用多模态信息，而不对模态之间的交互进行建模。Chen等人(Chen et al., 2017)在单词水平上进行多模态融合，对孤立的话语进行情绪识别。Zadeh等人(Zadeh et al., 2018)提出MFN来融合多视图的信息，从而很好地协调来自不同模式的特征。Hu和Liu等人(Hu et al., 2021; Liu et al., 2021)在图架构中模拟跨模态和单模态交互，很好利用了模态之间的互补性。方法(Hu et al., 2022)则利用基于图的融合方法对来自不同语义空间的上下文信息进行动态建模，减少了连续聚合中的冗余信息。然而，这些方法都缺乏对从原始输入到融合嵌入的信息流的控制，这可能会导致信息流在融合过程中丢失任务相关信息。

互信息（mutual information，MI）是信息论中的一个概念，指两个随机变量之间的关联程度。即给定一个随机变量后，另一个随机变量不确定性的削弱程度。互信息越大，说明两个随机变量之间的相关性越强；互信息为0，说明两个随机变量之间没有任何关系。互信息可以用来评价两个变量是否有关系，以及关系的强弱，其在机器学习、深度学习和数据挖掘等领域有

广泛的应用。Alemi等人(Alemi et al., 2016)首先将与MI与深度学习模型进行结合。从那时此，大量的著作(Bachman et al., 2019; He et al., 2020; Amjad and Geiger, 2019)研究并证明了MI最大化原则的好处。互信息最大化已经被证明了在减少与下游任务无关的冗余信息，保留任务相关信息方面起到有效作用(Poole et al., 2019; Han et al., 2021)。在本文的工作中，我们利用互信息最大化，分级对模态输入对以及多模态融合结果与单模态融合结果之间的互信息进行最大化，更好地过滤噪声，在融合过程中保存与任务相关的信息。

对比学习（contrastive learning，CL）(Hinton et al., 2006)是一种基于对比思想的判别式表示学习框架，已被用于图像、文本、语音等多种领域。监督对比学习（supervised contrastive learning, SCL）(Khosla et al., 2021)则是一种利用标签信息来提高对比学习效果的方法。对比学习是自监督学习，它通过构造正负样本对，让模型学习区分不同的数据表示。监督对比学习则是在一个批处理内，让同类别的样本表示彼此接近，而不同类别的样本表示彼此远离。这样可以增强模型的判别能力和泛化能力。利用好标签本身所携带的信息，我们可以对说话者的情绪状态有更为完整的理解。监督对比学习已经被证明，在充分利用标签信息的情况下，可以使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥(Li et al., 2022)。在本文中，我们将监督对比学习纳入基于图的融合网络，利用标签信息增强模型对相似情绪的识别能力。

3 多模态对话情绪识别模型

形式上，一个包含N句话语的对话可以被描述为一系列话语$\{u_1, u_2, \ldots, u_N\}$。每个话语都包含三种话语层面的特征，包括声学(a)、视觉(v)和文本(t)，它们可以表示为$u_i = \{u_i^a, u_i^v, u_i^t\}$。多模态ERC的目标是预测每个话语$u_i$的情绪标签$y_i$。

所提出的MMIC如图1所示，其包括四个基本模块：模态编码器、基于图融合的有监督对比损失、互信息最大化和情绪分类器。不同模态的编码器首先捕获对应模态的上下文和说话者特征。然后我们最大化编码后的多模态输入之间的互信息。对于每个对话，我们为每个模态构造一个全连通图，在不同模态之间连接着对应于同一话语的节点。为了结合多模态上下文特征，基于图的融合网络以动态和顺序的方式堆叠。我们将经过网络后的特征与编码后的多模态输入进行拼接，得到最终的融合特征，接着最大化融合特征与编码后的多模态输入之间的互信息。最后，利用得到的融合特征和预测的情绪标签计算有监督对比损失和交叉熵损失。

3.1 模态编码器

使用相应的情态编码器，我们为每个情态创建上下文感知的话语特征编码。对于文本模态，我们使用双向长短期记忆网络(LSTM)对顺序文本上下文数据进行编码(Hochreiter et al., 1997)。对于声音或视觉模态，我们应用了一个全连接网络:

$$\mathbf{c}_i^t, \mathbf{h}_i^c = \overleftrightarrow{LSTM}_c\left(\mathbf{u}_i^t, \mathbf{h}_{i-1}^c\right) \tag{1}$$

$$\mathbf{c}_i^\varsigma = \mathbf{W}_c^\varsigma \mathbf{u}_i^\varsigma + \mathbf{b}_c^\varsigma, \varsigma \in \{a, v\} \tag{2}$$

考虑到说话者的个人信息也会影响对话，因此，我们还采用双向GRU(Chung et al., 2014)网络来捕获同一对话相邻话语之间的自依赖性。说话者嵌入可计算为:

$$\mathbf{s}_i^\delta, \mathbf{h}_{\lambda,j}^s = \overleftrightarrow{GRU}_s\left(\mathbf{u}_i^\delta, \mathbf{h}_{\lambda,j-1}^s\right), j \in [1, |U_\lambda|] \tag{3}$$

其中$\delta \in \{a, v, t\}$. $\mathbf{h}_{\lambda,j}^s$ 是第j个p_λ，$\lambda = \phi(u_i)$的隐藏状态。在一段对话中，所有的p_λ都被表述为U_λ。

3.2 基于图融合的有监督对比损失

在前面工作(Hu et al., 2021)的基础上，我们构建了一个无向图来模拟对话，表示为$\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其中$\mathcal{V}(|\mathcal{V}| = 3N)$表示三种形式的话语节点，$\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$是连接的集合。节点通过上下文和说话者嵌入进行初始化:

$$\mathbf{x}_i^\delta = \mathbf{c}_i^\delta + \gamma^\delta \mathbf{s}_i^\delta, \delta \in \{a, v, t\} \tag{4}$$

其中$\gamma^a, \gamma^v, \gamma^t$ 是权衡超参数。边的权值$\mathcal{A}_{ij}$被计算为$\mathcal{A}_{ij} = 1 - \frac{\arccos(\text{sim}(x_i, x_j))}{\pi}$。

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。

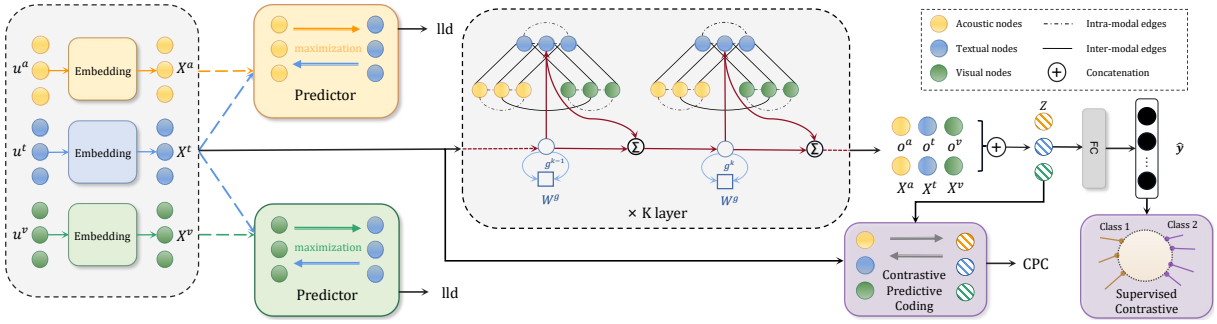


图 2. MMIC模型结构图

为了将模态间和模态内上下文信息在网络特定的语义区域内结合，我们在每一层使用图卷积操作。在(Hu et al., 2022)之后，门控方法用于动态融合会话中的多模态上下文信息:

$$\mathbf{g}^{(k)}, \mathbf{C}^{(k)} = \overrightarrow{LSTM}_c\left(\mathbf{g}^{(k-1)}, \mathbf{H}'^{(k-1)}\right) \tag{5}$$

经过改进的卷积运算(Chen et al., 2020)表示为:

$$\begin{aligned}\mathbf{H}^{(k)} = \mathrm{ReLU}(&\left((1-\alpha)\tilde{\mathbf{P}}\mathbf{H}'^{(k-1)} + \alpha\mathbf{H}^{(0)}\right) \\ &\left((1-\beta_{k-1})\,\mathbf{I}_n + \beta_{k-1}\mathbf{W}^{(k-1)}\right))\end{aligned} \tag{6}$$

其中使用了正则化的图卷积矩阵是$\tilde{\mathbf{P}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$。$\alpha$是一个超参数。$\rho$，$\beta_k$也是两个超参数。单位映射矩阵记为$\mathbf{I}_n$，$\mathbf{H}'^{(k)} = \mathbf{H}^{(k)} + \mathbf{g}^{(k)}$表示第$k$层的输出。

受到(Khosla et al., 2021)的启发，我们使用监督对比损失改进基于图的融合模块。有监督对比学习将批处理中所有具有相同标签的样本都视为正样本。在多模态对话情绪识别中，来自不同模态的特征在映射到共享嵌入空间之前被编码和融合。然而，不同模式之间的信息冗余会增加识别相似情绪的难度。通过有监督的对比学习，利用标记信息，可以增强模型对相似情绪类别的辨别能力，从而提高模型情绪识别的性能。由于数据集具有高度不平衡的特点，在批处理中有单个样本的情况下，损失计算可能会被掩盖(Poria et al., 2020)。为了解决ERC数据集高度不平衡的问题，我们创建话语隐藏状态$H_{d-\mathrm{win}}$的副本，并分离其梯度以确保稳定的参数优化。下式可表示批处理中所有样本的总监督对比损失:

$$X = \left[H_{d-\text{ win}}, \bar{H}_{d-\mathrm{win}}\right] \tag{7}$$

$$\mathrm{SIM}(p,i) = \log \frac{\exp\left(\left(X_i \cdot X_p\right)/\tau\right)}{\sum_{a\in A(i)} \exp\left(X_i \cdot X_a/\tau\right)} \tag{8}$$

$$\mathcal{L}_{\mathrm{SCL}} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \mathrm{SIM}(p,i) \tag{9}$$

其中$P(i) = I_{j=i} - \{i\}$表示与i类别相同但不包括自身的样本，$A(i) = I - \{i, N+i\}$表示多视图批处理中除自身之外的样本。$X \in \mathbb{R}^{2N\times d}, i \in I = \{1, 2, \cdots, 2N\}$表示多视图批处理中样本的索引。$\tau \in R^+$表示用于控制实例之间距离的温度系数。

3.3 互信息最大化

在输入级和融合级上进行分级互信息最大化，有利于模型捕捉不同层次模态之间的依赖关系，保留任务相关信息，生成丰富的多模态表示(Barber et al., 2004; Han et al., 2021)。因此，我们采用了分级MI最大化框架(Han et al., 2021)。

输入级互信息最大化。对于最大化多模态输入对之间的MI，我们采用了(Barber et al., 2004)来计算精确的MI下界，用$q(y \mid x)$来近似$p(y \mid x)$。计算公式为：

$$\begin{aligned} I(X;Y) =& \mathbb{E}_{p(x,y)}\left[\log \frac{q(y \mid x)}{p(y)}\right] + \\ & \mathbb{E}_{p(y)}[KL(p(y \mid x) \| q(y \mid x))] \\ & \geq \mathbb{E}_{p(x,y)}[\log q(y \mid x)] + H(Y) \\ & \triangleq I_{BA} \end{aligned} \tag{10}$$

考虑到文本模态提供了更多的信息(Arandjelovic et al., 2017)，我们优化了两种模态对(文本，声学)和(文本，视觉)的边界。$H(Y)$为Y的熵，计算采用高斯混合模型(Nilsson et al., 2002)。假设$q_\theta(y \mid x) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{\mu}_{\theta_1}(\boldsymbol{x}), \boldsymbol{\sigma}^2_{\theta_2}(\boldsymbol{x})\boldsymbol{I}\right)$(Cheng et al., 2020)。$x$决定了公式中的两个高斯分布参数$\boldsymbol{\mu}$和$\boldsymbol{\sigma}$，这些参数通常由回归多层感知器(MLP)来计算。似然最大化损失函数为：

$$\mathcal{L}_{lld} = -\frac{1}{N}\sum_{tv,ta}\sum_{i=1}^{N}\log q\left(y_i \mid x_i\right) \tag{11}$$

其中N是训练批大小。tv，ta是两个预测因子的概率之和。然后采用无参数估计$H(Y)$对不同类别(负和非负)的样本分别进行高斯混合模型建模。由$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} h_c^i$，$\hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} h_c^i \odot h_c^i - \hat{\boldsymbol{\mu}}_c^T\hat{\boldsymbol{\mu}}_c$，$H(Y)$ 被计算为(Huber et al., 2008)：

$$H(Y) = \frac{1}{4}\left[\log\left(\left(\det\left(\boldsymbol{\Sigma}_1\right)\det\left(\boldsymbol{\Sigma}_2\right)\right)\right] \tag{12}$$

其中$\boldsymbol{\mu}$是平均向量，$\boldsymbol{\Sigma}$是协方差矩阵，c表示负或非负。该层MI下界最大化的损失函数如下：

$$\mathcal{L}_{BA} = -I_{BA}^{t,v} - I_{BA}^{t,a} \tag{13}$$

融合级互信息最大化。对于融合结果与输入模态之间的MI最大化，优化的目标是融合网络生成的融合结果$Z = F\left(\mathbf{X}^t, \mathbf{X}^v, \mathbf{X}^a\right)$。我们使用一个评分函数来评估归一化预测和真实向量之间的相关性：

$$\overline{G_\phi(Z)} = \frac{G_\phi(Z)}{\|G_\phi(Z)\|_2}, \quad \overline{h_m} = \frac{h_m}{\|h_m\|_2} \tag{14}$$

其中G_ϕ是一个参数为ϕ的神经网络，Z表示融合特征，h_m表示模态m。$\|\cdot\|_2$是欧几里得范数，我们除以它以得到单位长度向量。CPC分数表示为$s\left(h_m, Z\right)$，得分越高，该模态在融合模态中越显著。$\mathbb{E}_{\mathbf{H}}$为噪声对比估计框架，正样例表示存在融合关系的单模态表示和融合表示形成的对，负样例则是同一个批次内其他样本的单模态表示和融合表示形成的对。该层的损失函数为：

$$s\left(h_m, Z\right) = \exp\left(\overline{h_m}\left(\overline{G_\phi(Z)}\right)^T\right) \tag{15}$$

$$\mathcal{L}_{\mathbf{N}}\left(Z, \mathbf{H}_m\right) = -\mathbb{E}_{\mathbf{H}}\left[\log \frac{s\left(Z, h_m^i\right)}{\sum_{h_m^j \in \mathbf{H}_m} s\left(Z, h_m^j\right)}\right] \tag{16}$$

$$\mathcal{L}_{CPC} = \mathcal{L}_{\mathrm{N}}^{z,v} + \mathcal{L}_{\mathrm{N}}^{z,a} + \mathcal{L}_{\mathrm{N}}^{z,t} \tag{17}$$

3.4 情绪分类器

在经过网络K层的融合之后，每个语句i的三种形式的表示可以进一步细化为$\mathbf{o}_i^a; \mathbf{o}_i^v; \mathbf{o}_i^t$。最后，使用分类器来预测每个话语的情绪：

$$\hat{\mathbf{y}}_i = \mathrm{Softmax}\left(\mathbf{W}_z\left[\mathbf{x}_i^a; \mathbf{x}_i^v; \mathbf{x}_i^t; \mathbf{o}_i^a; \mathbf{o}_i^v; \mathbf{o}_i^t\right] + \mathbf{b}_z\right) \tag{18}$$

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。

其中W_z和b_z是可训练的参数。为了得到任务损失，我们使用带有L2正则化的交叉熵损失:

$$\mathcal{L}_{\text{task}} = -\frac{1}{\sum_{l=1}^{L}\tau(l)}\sum_{i=1}^{L}\sum_{j=1}^{\tau(i)}\mathbf{y}_{i,j}^{l}\log\left(\hat{\mathbf{y}}_{i,j}^{l}\right)+\eta\|\Theta\|_2 \tag{19}$$

其中$c(i)$为对话中的话语数，N为对话数。$i, P_{i,j}$是对话中话语j预测情绪标签的概率分布，$i, y_{i,j}$是期望类标签。θ为l2正则化权值，三个变量都是可训练的参数。最终我们得到总加权损失来训练模型:

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{SCL} + \beta\mathcal{L}_{CPC} + \zeta\mathcal{L}_{BA} \tag{20}$$

其中α代表监督对比损失的参数。超参数β和ζ是调节MI最大化的参数。

4 实验与分析

4.1 数据集

随着社交媒体的快速发展，交互数据越来越多，其中包括一些公开的对话数据集。我们在这次实验中选用的是两个英文的公共基准数据集和一个新发布的中文数据集，即IEMOCAP，MELD和M^{3}ED。对于IEMOCAP，MELD，我们选用了和研究(Hu et al., 2021)一致的话语级特征。对于M^{3}ED，我们使用从(Zhao et al., 2022)提供的帧级特征中获得的话语级特征。三个数据集的详细数据分布如表1所示。

	IEMOCAP			MELD			M3ED		
	train	val	test	train	val	test	train	val	test
说话者数量	8		2	307		100	421	87	118
对话数量	120		31	1153		280	685	126	179
话语数量	5810		1623	11098		2610	17427	2821	4201
平均话语长度	48.4		52.4	9.65		9.3	25.44	22.39	23.47

表 1. 实验数据集的详细信息

IEMOCAP(Busso et al., 2008)包含十个不同的说话人对之间的对谈视频。它包括7433个话语和151个对话。对话中的每句话都被标注了六个类别的情绪标签中的一个，标签包括快乐、悲伤、中性、愤怒、兴奋和沮丧。我们遵循之前的研究(Majumder et al., 2019)，使用前四次训练，使用最后一次测试，并随机抽取10个训练对话作为验证分割。

MELD(Poria et al., 2019)包含多方对话，这些对话都是从《老友记》系列电视剧中收集的视频得来的，每轮对话都有两个或两个以上的说话者参与。它包含1433个对话，13708个话语和304个不同的说话者。对话中的每句话都带有七种情绪标签中的一个，标签包括愤怒，厌恶，恐惧，喜悦，中性，悲伤和惊讶。为公平的比较，我们遵循(Hu et al., 2021)的设置进行实验。

M^{3}ED(Zhao et al., 2022)包含来自56部不同电视剧的990个二元情绪对话，是中文首个大规模多模态情绪对话数据集。它包含990个对话，24449个话语和626个不同的说话者。对话中的每句话都被标注了七个类别的情绪标签中的一个，标签包括中性、高兴、惊讶、悲伤、厌恶、愤怒和恐惧。我们使用M^{3}ED中预定义的训练/验证/测试分割进行实验。

4.2 实验参数设置

在实验中，我们使用基于随机梯度下降的Adam(Kingma et al., 2017)优化器来训练我们的模型。为了避免过拟合，Dropout(Srivastav et al., 2014)被设置为0.1到0.5进行验证。具体超参数设置如下：三个数据集的GCN层数均为16层，批大小（Batchsize）设置为16，α in $\{0.8, 1.0\}$, β, ζ in $\{0.01, 0.05, 0.10\}$。特别地，由于IEMOCAP数据集中随机噪声较少，我们将ξ设为0。以下结果中的数据均为10次独立实验的平均值。所有实验均在GeForce RTX A5000 GPU上进行。

Model	IEMOCAP							MELD
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Weight-Avg-F1	Weight-Avg-F1
BC-LSTM	33.82	78.76	56.75	64.35	60.25	60.75	60.42	57.29
MFN	48.19	73.41	56.28	63.04	64.11	61.82	61.60	57.80
ICON	32.80	74.40	60.60	68.20	68.40	66.20	63.50	-
DialogueRNN	32.20	80.26	57.89	62.82	**73.87**	59.76	62.89	57.11
DialogueGCN	47.10	80.88	58.71	66.08	70.97	61.21	65.04	57.34
MMGCN	43.44	77.80	60.03	67.25	75.20	61.79	65.20	57.87
MM-DFN	42.06	79.66	65.37	68.11	74.56	67.16	67.86	58.07
COGMEN	**54.89**	79.13	64.81	60.61	74.17	57.71	65.71	-
MMIC(Ours)	44.44	**82.83***	**66.27***	**69.72***	73.38	**67.27***	**68.74***	**58.57***

表 2. 在IEMOCAP和MELD数据集上的性能对比

Model	M^3ED
	Weight-Avg-F1
DialogueRNN	51.57
DialogueGCN	45.90
MMGCN	49.28
MM-DFN	53.27
COGMEN	52.06
MMIC(Ours)	**53.65***

表 3. 在M^3ED数据集上的性能对比

4.3 实验对比模型

为了评估我们提出的MMIC模型，我们与以下模型进行比较:

- **BC-LSTM** (Poria et al., 2017):该模型在话语级利用LSTM网络捕获多模态特征。
- **MFN** (Zadeh et al., 2022): 该模型采用多视图门控记忆单元同步多模态序列。
- **ICON** (Hazarika et al., 2018): 该模型通过多层跳跃存储器提供从模态到会话的特性。
- **DialogueRNN** (Majumder et al., 2019): 该模型引入循环网络来跟踪说话人的状态和对话中的上下文。
- **DialogueGCN** (Ghosal et al., 2019): 该模型利用图形结构来组合上下文依赖关系。
- **MMGCN** (Hu et al., 2021): 该模型使用基于图形的融合模块来捕获模态内和模态间的上下文特征。
- **MM-DFN** (Hu et al., 2022): 该模型用动态图形融合模块来融合对话中多模态的上下文特性。
- **COGMEN** (Joshi et al., 2022): 该模型使用基于图的架构在对话中建模复杂的关系(局部和全局信息)。

4.4 主实验

与已有的工作(Poria et al., 2017; Majumder et al., 2019; Joshi et al., 2022)一样，我们在多模态对话情绪识别的实验中使用加权平均F1分数作为评价指标。表2和表3比较了在多模态的设置下MMIC与其他模型在三个数据集上的性能。标有*的结果表示在配对t检验下，MMIC与最先进的分数相比有统计学意义($p < 0.05$)。由于一些模型的再现结果与原论文报告的结果存在差异，为了确保结果的可靠性和准确性，我们选择使用再现结果。

数据集	IEMOCAP	MELD	M^3ED
评估方法		Weight-Avg-F1	
MMIC	68.74	58.57	53.65
-MI	68.48(↓0.26)	58.37(↓0.20)	53.42(↓0.23)
-Contrast	68.34(↓0.40)	58.36(↓0.21)	53.46(↓0.19)
-MI, Contrast	67.86(↓0.88)	58.07(↓0.50)	53.27(↓0.38)

表 4. MMIC的消融实验结果

数据集	测试集样例	w/o MI,Contrast	w/o MI	w/o Contrast	MMIC	真实标签
M^3ED	我就生气怎么了!	Sad	Angcr	Sad	Angcr	Angcr
M^3ED	顺便躲避一下小蚯蚓的唉声叹气	Neutral	Neutral	Sad	Sad	Sad
MELD	I will!	Surprise	Anger	Joy	Joy	Joy
MELD	Yes!! Yes! Yes! Yes!! That's my Dad, that's Frank! Yeah!I'm sorry I'm getting all flingy.	Joy	Surprise	Joy	Surprise	Surprise

表 5. 案例探究

根据表2和表3可知，MMIC在三个数据集上的F1分数表现，超过了所有的基准模型和目前最先进的方法。具体来说，MMIC在IEMOCAP上的平均提高为0.88%，MELD上的平均提高为0.50%，M^3ED上的平均提高为0.38%。IEMOCAP数据集上，MMIC在7个评价指标中有5个指标均优于其他方法，并且在其他指标上也同样产生了具有竞争力的结果。实验结果表明了MMIC是有效的多模态对话情绪识别模型，在中英文和各种情境下均能获得优于其他所有模型的指标分数，取得优秀的性能表现。

4.5 进一步研究

通过比较以上结果可以看出，MMIC模型取得了良好的性能。为了进一步探究模型性能提升的原因，我们首先进行了消融实验，以分析本文模型建模的不同类型的关系对模型整体性能带来的影响。然后，我们在所提出模型上进行了案例探究和可视化操作，以证实该模型可以有效保留任务相关信息和利用标签信息，更好地区分出相似取消，提高在现实对话场景中识别不同情绪的准确度。

4.5.1 消融实验

表4显示了消融实验的结果，移除了所提出模型的关键模块。当最大化互信息模块或监督对比损失模块被移除时，数据集上的结果显著下降。当两个模块都被移除时，结果进一步下降。该实验证明了两个模块在三个数据集上的有效性，以及建模的不同关系对于模型的整体性能都是有价值的。详细分析如下：

- **消去最大化互信息**：即从模型中移除最大化互信息模块。从结果中我们可以看到，在IEMOCAP数据集中F1分数下降了0.26%，在MELD数据集中F1分数下降了0.20%，在M^3ED数据集中F1分数下降了0.23%。这表明，利用互信息最大化，分级对模态输入对以及多模态融合结果与单模态融合结果之间的互信息进行最大化，可以更好地过滤噪声，避免任务相关信息在融合过程中的丢失，提高模型在多模态对话情绪识别中的性能。

- **消去有监督对比损失**：即从模型中移除监督对比损失模块。从结果中我们可以看到，在IEMOCAP数据集中F1分数下降了0.40%，在MELD数据集中F1分数下降了0.21%，在M^3ED数据集中F1分数下降了0.19%。这表明，将监督对比学习纳入基于图的融合网络，有利于充分利用标签信息，增强模型对相似情绪的识别能力，提高情绪识别的准确度。

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。

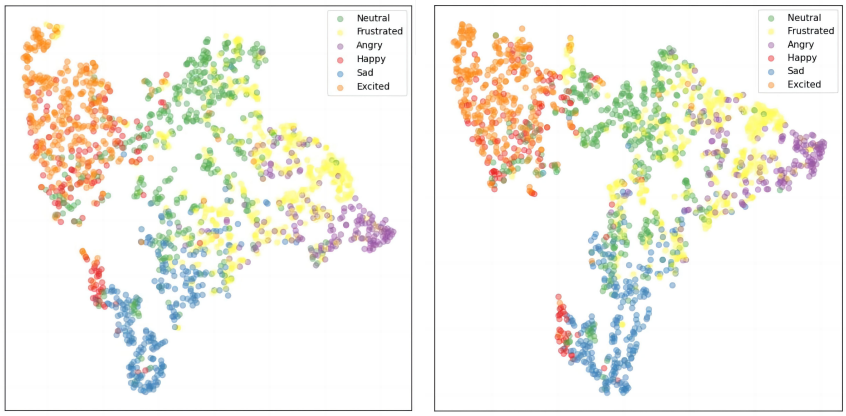


图 3. t-SNE可视化对比图

4.5.2 案例探究

我们在MELD和M^{3}ED数据集上进行了案例探究，如表5所示。可以看见，当去除了两个关键模块后，模型容易混淆两个相似的情绪，从而对说话者情绪进行错误的判别。当去除互信息模块后，模型可以准确识别案例2、3中的相似情绪，但是对案例1和4则进行了错误的识别。我们考虑这可能是由于案例1和4本身存在的信息较少或存在单独语义歧义，需要更多的上下文信息来识别情绪，而缺少了互信息模块，就导致融合信息中的任务相关信息受损，从而使模型进行了错误的判别。当去除对比损失模块后，模型可以准确识别案例1和4中的相似情绪，但是对案例2和3则进行了错误的识别。我们考虑这可能是由于案例2和3本身的信息较多或是存在两类不同的情感词语，导致语义相似度较高，使得不同模式之间的信息冗余。缺少了对比损失模块，模型无法利用标签信息来拉开相似情绪样本在表示空间的距离，从而进行错误的判别。而当两个关键模块都没有被移除时，模型可以准确判别出四个案例的情绪，这进一步证明了两个关键模块可以提高模型在多模态设置下各类情景中对话情绪识别上的性能。

4.5.3 t-SNE可视化

为了达到识别出不同相似情绪的目的，要尽量使不同类别的数据在表示空间中分开，而相同类别的数据在表示空间中靠近。这样可以有效地提高不同类别之间的耦合程度，增加相同类别之间的内聚性。耦合程度是指模块或类之间的关联和依赖的程度，内聚性是指模块或类内部元素之间的相关性。如图3所示，我们将模型的输出可视化在IEMOCAP数据集上，左边是我们模型在没有互信息最大化和监督对比损失的情况下的t-SNE可视化结果，右边是我们完整模型的结果。通过对比，我们发现我们所提出的模型使不同类别的数据在表示空间中更分离，相同类别的数据在表示空间中则更为靠近，即是模型可以更有效地提高不同类别之间的耦合程度，增加相同类别之间的内聚性，这证明了模型可以更好地区分相似的情绪。

5 总结

针对多模态对话情绪识别研究现有的问题，本文提出了一种基于互信息和对比损失的图融合模型——MMIC。为了更好地保存从输入到融合结果的任务相关信息，我们首先最大化了单模态表示之间的互信息。在通过图卷积运算得到融合结果后，我们进一步最大化它与低水平单模态表示之间的互信息。为了更充分地利用标签中隐含的信息，我们将监督对比学习纳入基于图的融合网络，使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥。我们在两个公共基准数据集和一个中文数据集上进行了大量实验，证明了所提出模型的有效性和优越性。所提出的模型经过案例探究，证实了它能有效保留任务相关信息，更好地区分相似情绪。消融实验和可视化结果证明了模型中每个模块的有效性。

致谢

本文受到广东省自然科学基金(2021A1515011864)、国家自然科学基金(71472068)、广州市智慧农业重点实验室(201902010081)、广东省普通高校特色创新项目(2020KTSCX016)、华南农业大学大学生创新训练计划项目(X202210564157)的资助。

参考文献

A. Zadeh, R. Zellers, E. Pincus, and L.P.e Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pp. 1128–1137.

X. Zhou and W. Y. Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pp. 1128–1137.

D. Hu, L. Wei, and X. Huai. 2019. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pp. 527–536.

N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the 33rd Association for the Advancement of Artificial Intelligence, AAAI 2018*, pp. 6818–6825.

D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. *Proceedings of the 9th Int. Joint Conf. Natural Lang. Process, IJCNLP 2019*, pp. 154–164.

S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.P. Morency. 2017. Context-dependent sentiment analysis in user-generated videos. *Proceedings of the 55th Annu. Meeting Assoc. Comput. Linguist. ACL 2017*, pp. 873–883.

D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. *Proceedings of Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol, NAACL-HLT 2018*, pp. 2122–2132.

D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. *Proceedings of 2018 Conf. Empirical Methods Natural Lang. Process, EMNLP 2018*, pp. 2594–2604.

J. Hu, Y. Liu, J. Zhao, and Q. Jin. 2021. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021*, pp. 5666–5675

J. Liu, S. Chen, L. Wang, and Z. Liu. 2021. Multimodal emotion recognition with capsule graph convolutional based representation fusion. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Virtual and Singapore*, pp. 6339–6343.

D. Hu, X. Hou, L. Wei, L. Jiang and Y. Mo. 2022. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore*, pp. 7037–7041.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Comput*, vol. 9, no. 8, pp. 1735–1780

J. Chung , C. Gulcehre, K.H. Cho and Y. Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *t arXiv:1412.3555*,

M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li. 2020. Simple and deep graph convolutional networks. *Proceedings the 36th International Conference on Machine Learning, ICML 2020*, pp. 1725–1735.

D. Barber and F. Agakov. 2004. The IM algorithm: A variational approach to information maximization. *Proceedings of Advances in neural information processing systems*, 16:201.

W. Han, H. Chen, and S. Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. *arXiv:2109.00412.*

R. Arandjelovic and A. Zisserman. 2017. Look, listen and learn. *Proceedings of IEEE International Conference on Computer Vision , ICCV 2017*, pp. 609–617.

M. Nilsson, H. Gustaftson, S.V. Andersen, and W. B. Kleijn. 2002. Gaussian mixture model based mutual information estimation between frequency bands in speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I–525.

M.F. Huber, T. Bailey, H. Durrant-Whyte, and U.D. Hanebeck. 2008. On entropy approximation for gaussian mixture random vectors. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 181–188.

P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin. 2020. Club:A contrastive log-ratio upper bound of mutual information. *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, pp. 1779–1788.

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. 2021. Supervised contrastive learning. *arXiv:2004.11362.*

S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea. 2023. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee and S.S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359.

S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pp. 527–536.

C.J. Zhao, T. Zhang, J. Hu, Y. Liu, Q. Jin, X. Wang, and H. Li. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. *arXiv:2205.1023.*

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:2006.11477.*

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2019. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv:1907.11692v1.*

G. Huang, Z. Liu, L.V.D. Maaten, and K.Q. Weinberger. 2017. Densely connected convolutional networks. *Proceedings of IEEE conference on computer vision and pattern recognition, CVPR 2017*, pp. 4700–4708.

A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.P. Morency. 2022. Memory fusion network for multi-view sequential learning. *arXiv:2205.1023.*

A. Joshi, A. Bhat, A. Jain, A.V. Singh and A. Modi. 2022. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. *Proceedings of the 20th Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2022*, pp. 4148–4164.

D.P. Kingma, and J.L. Ba. 2017. ADAM: A method for stochastic optimization. *arXiv:2205.1023.*

Shimin Li, Hang Yan, Xipeng Qiu. 2022. Contrast and Generation Make BART a Good Dialogue Emotion Recognizer. *arXiv:2112.11202v2.*

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y Ng. 2011. Multimodal deep learning. *Proceedings the 17th on Machine Learning, ICML 2011.* pp. 689-696

M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, and . Morency. 2017. Multimodal sentiment analysis with wordlevel fusion and reinforcement learning. *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* pp, 163–171.

A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.P. Morency. 2018. Memory fusion network for multiview sequential learning. *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence, AAAI 2018* vol 32.

B. Poole, S. Ozair, A.V.D. Oord, A. Alemi, and G. Tucker. 2019. On variational bounds of mutual information. *Proceedings of the 36th International Conference on Machine Learning, ICML 2019* pp. 5171–5180.

A.A. Alemi, I. Fischer, J.V. Dillon, and K. Murphy. 2016. On variational bounds of mutual information. *arXiv preprint arXiv:1612.00410.*

P. Bachman, R. Hjelm, and W. Buchwalter. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910.*

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* pp. 9729–9738.

R.A. Amjad and B.C. Geiger. 2020. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2225-2239.

G.E. Hinton and R. R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. science. *Science.* vol. 313, issue 5786, pp. 504-507.

N. Srivastava, G.E. Hinton, A. Krizhevsky . 2014. Dropout:A simple way to prevent neural networks from overfitting. *Mach.Learn.Res.* vol. 15, no. 1, pp. 1929-1958.

基于语义任务辅助的方面情感分析

吴肇真
新疆大学
软件学院/
乌鲁木齐
1452078416@qq.com

赵晖
新疆大学
信息科学与工程学院/
乌鲁木齐
zhmerry@126.com

谷体泉
新疆大学
信息科学与工程学院/
乌鲁木齐
gtq@stu.xju.edu.cn

曹国义
新疆大学
软件学院/
乌鲁木齐
cgy1599@163.com

摘要

方面情感分析(Aspect-Based Sentiment Analysis,ABSA)任务旨在判断一句话中不同方面的细粒度情感极性。如何有效的捕获句子的语义信息是该任务的关键。现有的大多数分类方法通过引入外部知识并设计复杂的模块来理解句子的语义信息，而忽略了外部解析器的噪音及模型的复杂化。在本文中，我们提出了一种基于语义理解的多任务学习网络，它旨在通过多任务学习从原始语料中捕获句子的语义信息。本文考虑从多任务角度出发，在具有共享参数的原始数据集中，分别提出了两个语义辅助任务：方面上下文顺序预测任务和方面上下文句法依存预测任务。然后，将辅助任务与原始的方面情感分类任务进行多任务的训练得到增强了语义理解的编码器，最后将该编码器用于方面情感分类任务。实验结果表明，模型在三个主要的公开数据集Rest14、Lap14和Twitter上的准确率和Macro-F1值都有较好的表现。

关键词： 方面情感分析；多任务；方面上下文顺序信息；方面上下文句法依存信息

Semantic Task-assisted Aspect-based Sentiment Analysis

Wu Zhaozhen
Xinjiang University
School of Software
/ Urumqi
1452078416@qq.com

Zhao Hui
Xinjiang University
School of Information
Science and
Engineering/ Urumqi
zhmerry@126.com

Gu Tiquan
Xinjiang University
School of Information
Science and
Engineering/ Urumqi
gtq@stu.xju.edu.cn

Cao Guoyi
Xinjiang University
School of Software
Urumqi/
cgy1599@163.com

Abstract

The Aspect-Based Sentiment Analysis (ABSA) task aims to determine the fine-grained sentiment polarity of different aspects in a sentence. How to effectively capture the semantic information of a sentence is the key to this task. Most existing classification methods understand the semantic information of sentences by introducing external knowledge and designing complex modules, ignoring the noise of external parsers and the complexity of models. In this paper, we propose a multi-task learning network based on semantic understanding, which aims to capture the semantic information of sentences from the original corpus through multi-task learning. This paper considers auxiliary semantic tasks from a multi-task perspective. Two auxiliary semantic tasks are proposed in the original dataset with shared parameters: an aspect and context order prediction task and an aspect and context syntactic dependency prediction task, respectively. Then, the auxiliary tasks are trained with the original aspect sentiment classification task in a multi-task to obtain an encoder with enhanced semantic understanding. Finally, the encoder is used for the aspect sentiment classification task.

第二十二届中国计算语言学大会论文集，第277页-第288页，哈尔滨，中国，2023年8月3日至5日。

The experimental results show that the model performs well in terms of accuracy and Macro-F1 values on three major publicly available datasets Rest14, Lap14, and Twitter.

Keywords: Aspect-based sentiment analysis , Multi-task , Order information of Aspect-Context , Syntactic dependency information of Aspect-Context

1 引言

情感分析又被称为意见挖掘(王婷，杨文忠, 2021)是自然语言处理(Natural Language Processing,NLP)领域的一项基础任务。方面情感分析是对实体的特定方面(属性)的情感倾向判断，本文将这样的实体的特定方面(属性)定义为方面项，它是句子中连续的方面词序列，用于情感倾向判断的词统称为情感词。一般说来，将句子中不同方面的情感类别分为积极(Positive)、中性(Neutral)和消极(Negtive)三种类别。

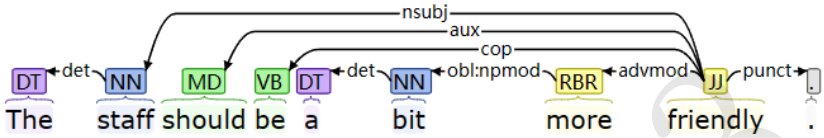


Figure 1: 依存解析样例

ABSA任务上的最近进展都与深度学习的发展息息相关。然而深度学习的基础来自于神经网络，使用神经网络来处理ABSA任务会产生这样的倾向：模型在遇到情感词时，会直接根据情感词输出情感极性，而不再更深层地理解句子的语义。例如，在句子“The staff should be a bit more friendly .”中，传统的神经网络会直接根据情感词“friendly”从而将方面项“staff ”判定为积极的情感极性，事实上这里的虚拟语气词“should”表示否定的含义，因此方面项“staff”的真实情感极性应该是消极的。另一方面，在最近的新方法中(Zhang et al., 2019; Wang et al., 2020a)，为了能够更好地提取语义特征，不得不需要一些外部的知识，例如句法依存信息，然而为了使用这种外部知识，不得不在网络中设计相应的模块，同时这些外部知识中常常包含着对于目标任务ABSA无关的信息，因而对目标任务来说是一种噪音。

考虑到上述问题，本文从多任务的角度分别设计了两个辅助任务：方面上下文顺序预测任务(Aspect-Context Order Prediction,ACOP)和方面上下文句法依存预测任务(Syntactic Dependency of Aspect-Context Prediction,SDACP)。然后在预训练模型RoBERTa(Liu et al., 2021)的基础上将辅助任务与目标任务ABSA进行多任务训练得到增强了语义理解和提取了句法结构知识的编码器,最后将该编码器用于目标任务ABSA。总的来说，本文的主要贡献如下：

(1)本文探索了多任务框架下的方面情感分析，从而定义了两种语义增强的辅助任务：ACOP和SDACP。

(2)基于RoBERTa模型，融合ACOP任务提出了RoBERTa-ACOP模型，融合SDACP任务提出了RoBERTa-SDACP模型。

(3)基于辅助任务提出的模型在公开数据集上实验，实验结果证明了本文模型的有效性。

2 相关工作

现有的方面情感分析方法分为两种：基于机器学习的方法和基于深度学习的方法。基于机器学习的方法通过考虑与任务相关的特征来完成任务，需要人工参与许多重要的特征设计。Wagner(Wagner et al., 2014; Kiritchenko et al.,)使用基于支持向量机的模型来完成方面情感分类任务。Gupta等人(Kumar Gupta et al., 2015)使用了基于粒子群优化算法来更好的提取与ABSA任务相关的特征，通过对粒子的不断迭代演化去除与任务特征无关的噪声，从而得到更高质量的特征表示，然后将特征输入条件随机场训练，得到更好表现的模型。Akhtar等

人(Akhtar et al., 2017)使用最大熵模型、条件随机场和支持向量机构建分类器，然后整合特征。Jo等人(Jo and Oh, 2011)引入无监督训练的方法提出了一种基于句子潜在Dirichlet 分布的概率生成模型。

基于深度学习的方法大致分为三个阶段。第一阶段是使用结合注意力机制的循环神经网络(RNN)来生成方面项特定的句子表示(Wang et al., 2016; Tang et al., 2016; Ma et al., 2017; Chen et al., 2017; Fan et al., 2018; Huang et al., 2018; Gu et al., 2018)，并取得了令人满意的结果。例如，Wang等人(Wang et al., 2016)提出了基于注意机制的长短期记忆网络(Long Short-term Memory，LSTM)用于方面级情感分类。Tang 等人(Tang et al., 2016)和Chen等人(Chen et al., 2017)都引入层级注意力网络来识别与给定方面相关的重要情感信息。Fan等人(Fan et al., 2018)使用多粒度注意机制来捕获方面项及其上下文之间的词级交互信息。

第二阶段是是基于图神经网络(Graph Neural Network,GNN)特别是图卷积网络(Graph Convolution Network,GCN)和图注意力网络(Graph Attention Network,GAT)的方法(Zhang et al., 2019; Wang et al., 2020a; Sun et al., 2019; Huang and Carley, 2019; Zhang and Qian, 2020; Chen et al., 2020; Liang et al., 2020; Wang et al., 2020b; Tang et al., 2020; Li et al., 2021)。Zhang等人(Zhang et al., 2019)和Sun等人(Sun et al., 2019)认为语法结构信息对于模型理解文本非常重要，都各自提出了图卷积网络。Huang等人(Huang and Carley, 2019)则在句子语法结构的基础上进一步引入注意力机制。Zhang等人(Zhang and Qian, 2020)利用全局词共现频率信息和词性信息来增强模型使得模型更好地理解文本。Chen等人(Chen et al., 2020)利用潜在推导图，在解析出的句法结构基础上获取更多的潜在关系，从而获取更多节点间的关系信息。Liang等人(Liang et al., 2020)通过注意力机制和图卷积网络在改进的依存树解析结果上取得了不错的效果。Wang等人(Wang et al., 2020a)对GAT方法进行了扩展，使用方面项和上下文之间的句法依存关系来计算注意力，由此提出了关系图注意力网络(Relational Graph Attention Network,R-GAT)。Li等人(Li et al., 2021)则同时考虑了句子的语义信息和句法信息，并融合了两种信息，提取到更好的特征。

第三阶段是大规模预训练模型的出现在ABSA任务上带来突破性的成果(Devlin et al., 2018; Xu et al., 2019; Song et al., 2019)，例如Google 的Devlin等人(Devlin et al., 2018)提出了BERT(Bidirectional Encoder Representations from Transformers)模型，该模型包括MLM(Masked Language Model)任务和NSP(Next Sentence Prediction)任务，实现了阅读理解多项任务上的突破。Xu等人(Xu et al., 2019)设计了BERT-PT，它探索了一种基于BERT模型的新型的后训练方法。Song(Song et al., 2019)进一步提出将句子的方面项附加到上下文句子中，以形成对应于文本对分类模型BERT-SPC的输入。另一方面，大规模预训练模型的出现给我们的启发不仅是多项自然语言处理任务上的突出表现，同时我们看到多任务的训练为模型的优异表现提供了坚实的基础，这为本文使用多任务方法提供了思路和方向。

3 多任务情感分析

3.1 任务定义

3.1.1 ACOP辅助任务

ACOP任务是指给出一个句子，判断该句子是否是正确的顺序，如果是乱序并识别出乱序的模式。本文提出ACOP任务的主要目的是使得模型能够理解句子的语义及方面项与上下文之间的结构信息。本文考虑了两种方式来实现ACOP任务：全局实现和局部实现。在全局实现中，本文将方面项作为边界将整个句子划分为左上下文，方面项和右上下文，然后我们基于以上三部分的全排列得到6种类型的顺序，如表1所示。

另外一种局部实现中，将方面项随机地插入到上下文中生成负样本，将句子的原始顺序作为正样本，如表2所示。由于当上下文足够长时，使用这种方式会产生非常多的负样本，给实验带来极大的计算代价，因此本文设置窗口大小(window_size)限制方面项随机插入的范围，和负样本数(neg_num)限制生成的乱序句子数量来解决这个问题。

3.1.2 SDACP 辅助任务

SDACP任务是指判断句子中每一个单词节点的关系类型，该任务是一个序列分类任务。本文提出SDACP任务的目的是希望模型能够以多任务方式去利用句法依存信息从而更好地学习句子的语义信息。本文的具体实现是在图2这样的句法依存树分析的基础上，由于生成的句

Aspect	Sentence	Output class
the food	Nevertheless the food itself is pretty good.	0
	Nevertheless itself is pretty good the food.	1
	The food nevertheless itself is pretty good.	2
	The food itself is pretty good nevertheless.	3
	Itself is pretty good nevertheless the food.	4
	Itself is pretty good the food nevertheless.	5

Table 1: ACOP任务的全局实现

Aspect	Sentence	Output class	Window size
the food	Nevertheless the food itself is pretty good .	0	0
	The food nevertheless itself is pretty good .	1	1
	Nevertheless itself the food is pretty good .	2	
	The food nevertheless itself is pretty good .	3	2
	Nevertheless itself the food is pretty good .	4	
	Nevertheless is itself the food pretty good .	5	

Table 2: ACOP任务的局部实现

法依存树不是以方面项为根(root)的，在该例中根是“was”，因此需要将依存树重新塑造成以方面项“staff”为根。然后在重塑的依存树的基础上为每一个上下文词生成句法依存标签，如表3所示，其中con_n表示上下文词与方面项不具有直接的句法依存关系，而是通过上下文词与方面项的句法依存距离为n这一关系相联系。

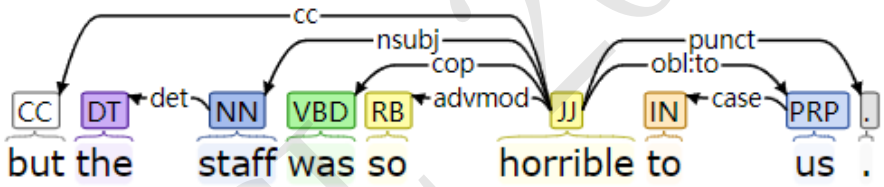


Figure 2: 句法依存树样例

在句法依存标签的生成过程中，本文使用变量cap_sd(captured syntactic distance)来控制需要捕获的句法依存距离信息范围，在表3的示例中，将该值设置为2，这意味着只捕获句法依存距离为2以内的信息(保留原来的句法依存信息)，而将距离为2以上的句法依存距离信息的上下文词统一分配给一个新的类别ncon，标点符号分配给另一个新的类别pun。

3.2 模型结构

使用辅助任务来帮助完成原ABSA任务的具体模型结构包含输入层、编码层、多任务层。模型整体架构如图3所示。

模型的具体作用过程如下：首先通过RoBERTa模型输出预训练的方面项词向量和整个句子词向量拼接，并将拼接后的向量输入多任务层训练，最后再将多任务层训练后的向量输入到分类层，输出情感极性。

3.2.1 输入层

该模型的输入是句子$\boldsymbol{S} = \{w_{s1}, w_{s2}, ..., w_{sls}\}$和相应的方面项$\boldsymbol{A} = \{w_{a1}, w_{a2}, ..., w_{ala}\}$，其中$ls$和$la$分别表示句子长度和方面项的长度。然后我们使用预训练模型RoBERTa将每个词转为一个实值向量。具体来说，对于句子$\boldsymbol{S}$，我们构造RoBERTa模型的输入“<s>” + $\boldsymbol{S}$ +“</s>”，然后将句子$\boldsymbol{S}$表示为RoBERTa隐藏状态$\boldsymbol{s} = \{s_i \mid i = 1, 2, \ldots, ls + 2\}$；对于方面项，我们构造相应的输入“<s>” + $\boldsymbol{A}$ +“</s>”，并将其转为对应的隐藏状态表示$\boldsymbol{a} = \{a_j \mid j$

Sample	but	the	staff	was	so	horrible	to	us	.
Syntactic dependency	con_2	det	root	nsubj	con_3	con_2	con_3	con_4	-
Tag	con_2	det	root	nsubj	ncon	con_2	ncon	ncon	pun

Table 3: 句法依存标签生成

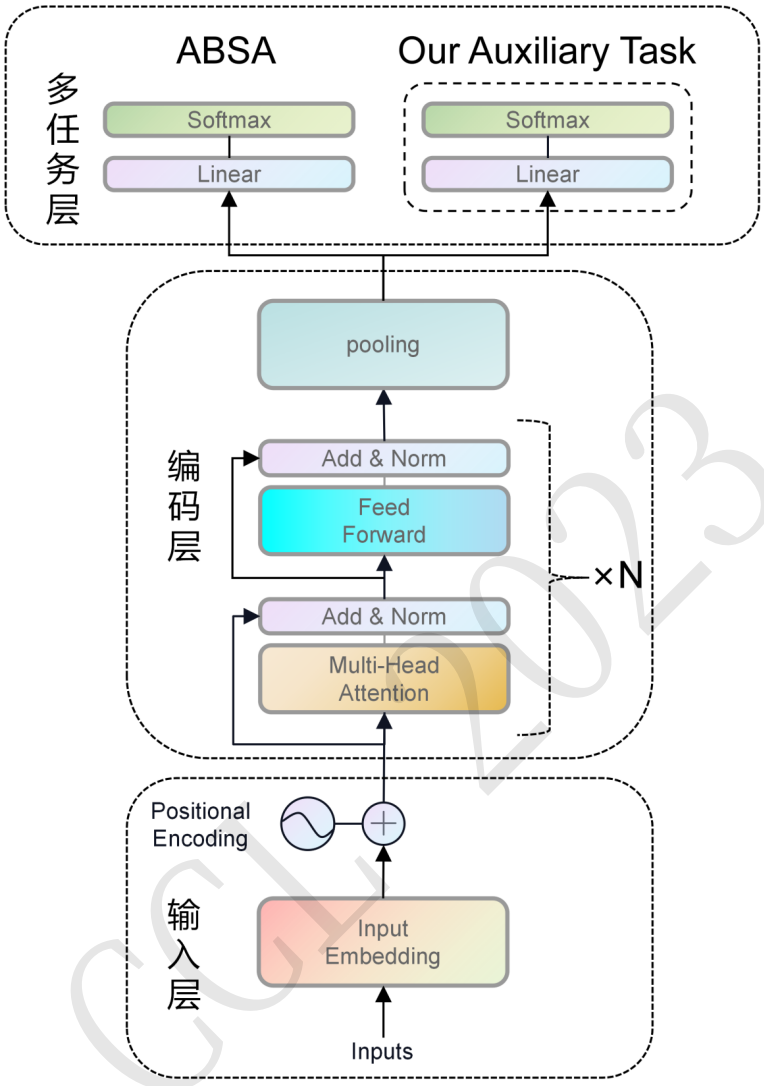


Figure 3: 模型整体架构

$= 1, 2, \ldots, la + 2\}$。为了使模型能够感知到句子的方面项，我们将上述的句子向量和方面项向量进行拼接得到感知方面项的句子表示$\boldsymbol{S^A}$，计算过程如下所示。

$$\boldsymbol{S^A} = \{s_i^a | i = 1, 2, ..., l_s, ..., l\} = \boldsymbol{s} \oplus \boldsymbol{a} \tag{1}$$

其中l=$ls + la + 4$。

3.2.2 编码层

在编码层部分需要同时对目标任务ABSA和辅助任务完成编码和池化工作。我们使用RoBERTa编码器编码,RoBERTa编码器部分像BERT模型一样仍然是基于Transformer的编

数据集	分类	积极	中性	消极	总和
Twitter	Train	1561	3127	1560	6248
	Test	173	346	173	692
Laptop	Train	994	464	870	2328
	Test	341	169	128	638
Restaurant	Train	2164	637	807	3608
	Test	728	196	196	1120

Table 4: 数据集统计

码器。RoBERTa的编码器部分，用下列公式进行概括。

$$\boldsymbol{m} = RoBERTa\_Encode\left(\boldsymbol{S^A}\right) \tag{2}$$

池化操作就是从提取到的特征中进行选择从而更好地完成后面的任务，我们直接截取第一个词的特征进行池化。在经过编码器部分后，得到特征表示$\boldsymbol{m} = \{m_i|i = 1, 2, ..., l\}$，然后进行池化得到更加重要的特征，如下公式所示。

$$\boldsymbol{r} = Pooling\left(m_i|i = 1, 2, ..., l\right) \tag{3}$$

3.2.3 多任务层

在原ABSA任务中，给出集成了方面项的句子信息$\boldsymbol{S^A}$输入到多任务层，我们得到特征表示$\boldsymbol{r}_{absa}$，对于我们的辅助任务，给出句子信息$\boldsymbol{s}$输入到多任务层，得到特征表示$\boldsymbol{r}_{aux}$，然后分别对两个任务进行分类得到相应的情感倾向概率分布，计算过程由下列公式给出。

$$\boldsymbol{p}_{absa} = softmax\left(\boldsymbol{r}_{absa}\boldsymbol{W}^{absa} + \boldsymbol{b}^{absa}\right) \tag{4}$$

$$\boldsymbol{p}_{aux} = softmax\left(\boldsymbol{r}_{aux}\boldsymbol{W}^{aux} + \boldsymbol{b}^{aux}\right) \tag{5}$$

其中$\boldsymbol{W}^{absa}$，$\boldsymbol{W}^{aux}$是可训练权重矩阵，$\boldsymbol{b}^{absa}$，$\boldsymbol{b}^{aux}$ 是偏置项。我们采用标准的交叉熵函数和L_2正则优化训练，计算过程如下所示。

$$L(\boldsymbol{p}) = -\sum_{s\in S}\sum_{c\in C}\hat{\boldsymbol{p}}log\boldsymbol{p} \tag{6}$$

$$Loss = \lambda_1 L\left(\boldsymbol{p}_{absa}\right) + \lambda_2 L\left(\boldsymbol{p}_{aux}\right) + \lambda_3\left|\left|\boldsymbol{\Theta}\right|\right|_2 \tag{7}$$

其中$\boldsymbol{S}$包含所有的训练句子，$\boldsymbol{C}$ 表示情感类别的集合，$\boldsymbol{p}$表示预测结果，$\hat{\boldsymbol{p}}$ 表示标签，λ_1，λ_2，λ_3表示正则系数，$\boldsymbol{\Theta}$表示所有的可训练参数。

4 实验与分析

4.1 数据集与参数设置

数据集：本文采用国际语义评估竞赛SemEval于2014年发布的任务4中的Restaurant、Laptop数据集(Maria Pontiki et al., 2014)以及Twitter数据集(Dong et al., 2014)。Restaurant、Laptop数据集分别是关于餐厅和笔记本电脑相关的评论，以下将分别简称为Rest14和Lap14数据集，Twitter数据集则是用户在社交平台推特上的评论数据，每一个样本数据都标注好了方面项和情感类别，情感类别包括积极(Positive)、消极(Negative)、中性(Neutral)三种类别，本文在实验中采用的评价指标是准确率Acc和Macro-F1值。数据集相关统计情况如表4所示。

参数设置：本文使用的是RoBERTa-base英文版本，该模型包含12个隐藏层、12个注意力头，每层的隐藏状态维度是768维。使用Adam(Kingma and Ba, 2014)作为模型的优化器，学习率初始化为0.000002，正则化系数λ_1设置为1，λ_2根据数据集有所不同，在主体实验中，

本文在Rest14数据集上设置为0.1，在Lap14数据集上设置为0.001，在Twitter数据集上设置为1，衰减率λ_3设置为0.01。除了衰减率的影响外，学习率在预热阶段(warm up)首先增加，然后保持在设置的学习率，最后随着当前步数的增加而降低。批量改组(Batch Shuffling)应用于训练集。RoBERTa的词汇量为50,265。所有模型的批量大小(batch size)都设置为32。针对ACOP任务局部实现中，window_size的范围是1到10，当window_size足够大时neg_num设置为5，否则我们将所有的句子作为负样本。针对SDACP任务的超参数cap_sd在主要实验中设置为5，在研究该参数影响的扩展实验中，将其设置为1到8，该任务中使用的句法依存分析的工具是spacy(Srinivasa-Desikan, 2018)。本文提出的模型通常训练不到30个epoch，就提前停止，尽管设置了100个epoch。

4.2 对比实验与分析

该小节将本文模型的表现与一些基线模型进行对比来证明本文模型的有效性。本文将对比的基线方法分为基于机器学习的方法、基于注意力的方法、图神经网络方法及预训练方法，在下面具体列出。

基于注意力模型:

ATAE-LSTM(Wang et al., 2016)：使用LSTM并结合注意力机制的情感分类模型。

IAN(Ma et al., 2017)：使用互动注意力模型识别方面项和对应上下文的联系。

MemNet(Tang et al., 2016)：使用多层记忆网络计算上下文中每个单词的情感贡献程度。

AOA(Huang et al., 2018)：使用多层注意力机制识别方面项和对应上下文的联系。

MGNet(Fan et al., 2018)：从词级识别上下文中每个单词与方面项之间的联系，是一个细粒度的情感分类模型。

TNet(Li et al., 2018)：使用CNN模型提取来自于RNN中的突出特征完成情感分类任务。

基于图神经网络模型:

ASGCN-DG(Zhang et al., 2019)：基于图卷积神经网络利用句法依存信息提升情感分类任务的效果。

BiGCN(Zhang and Qian, 2020)：利用全局词共现频率信息和词性信息来增强模型。

TD-GAT(Huang and Carley, 2019)：基于图注意力模型聚合句子中的情感信息。

R-GAT(Wang et al., 2020a)：基于图注意力模型利用句法依存信息提升情感分类任务的效果。

Dual-GCN(Li et al., 2021)：基于图卷积神经网络融合句子的语义信息和句法依存信息。

基于预训练模型:

Fine-tune BERT(Devlin et al., 2018)：在情感分类任务上微调BERT模型。

BERT-SPC(Xu et al., 2019)：方面项和上下文对作为BERT模型的句子对任务的微调BERT模型。

Fine-tune RoBERTa(Liu et al., 2021)：在情感分类任务上微调RoBERTa模型。

RGAT-BERT(Wang et al., 2020a)：基于图注意力模型利用句法依存信息并使用BERT模型向量化句子表示提升情感分类任务的效果。

实验结果如表5所示，从表中有以下观察结果。(1)基于预训练的方法在两个指标上的表现都优于大多数基于注意力的方法(例如AOA和TNet)和基于GNN的方法(例如ASGCN-DG和TD-GAT)，这表明预训练语言模型强大的语义表示能力。这一现象也表明，预训练模型已经成为NLP的趋势，其中也包括在ABSA任务上的应用。(2)具有特定任务的基于预训练的方法(例如RGAT-BERT)优于纯预训练（BERT、RoBERTa）的方法，说明将特定任务集成到预训练模型可以改善模型表现，我们对此的解释是纯预训练模型是基于整体语义从而判断出情感，而基于任务的预训练模型通常可以捕捉到句子的方面项和对应上下文的联系。这也表明细粒度的属性级ABSA相对于句子级和文档级情感分析中是更加有益的，因为方面项在情感分析中是一个关键信息。基于以上分析，本文提出两个新的辅助任务（ACOP和SDACP）并集成到预训练模型从而改善在方面情感分析上的表现。(3)尽管之前的模型表现出色，但本文的模型在准确性和F1 分数两个指标上相较于最先进的基线仍有不同程度地改善，例如，我们的模型相较于最好的基线模型（BERT-SPC、RGAT-BERT）在Rest14和Lap14数据集上的准确度和F1值分别提升了1.97%、1.76%和4.58%、5.16%，在Twitter数据集上提升了1.74%、2.38%，因此证明了ACOP任务和SDACP任务在提取方面项与上下文的结构信息和句法结构信息方面的有效性，从而增强模型的语义理解，得到更好的性能。

第二十二届中国计算语言学大会论文集，第277页-第288页，哈尔滨，中国，2023年8月3日至5日。

分类	模型	Restaurant		Laptop		Twitter	
		Acc	F1	Acc	F1	Acc	F1
Attention.	ATAE-LSTM	76.58	67.39	68.57	64.52	67.27	66.43
	IAN	76.88	68.36	70.84	65.73	68.74	67.61
	MemNet	78.12	68.99	72.32	67.03	70.19	68.22
	AOA	79.42	70.43	74.56	68.77	71.68	69.25
	MGNet	81.28	72.07	75.37	71.26	72.54	70.78
	TNet	80.69	71.27	76.54	71.75	74.93	73.60
GNN.	ASGCN-DG	80.77	72.02	75.55	71.05	72.15	70.40
	BiGCN	81.97	73.48	74.59	71.84	74.16	73.35
	TD-GAT	81.20	-	73.40	-	-	-
	R-GAT	83.30	76.08	77.42	73.76	75.57	73.82
	Dual-GCN	84.27	78.08	78.48	74.74	75.92	74.29
Pre-trained.	BERT	82.40	73.17	77.29	73.36	73.42	72.17
	BERT-SPC	84.46	76.98	78.99	75.03	74.13	72.73
	RoBERTa	84.38	77.26	78.83	74.53	74.57	74.02
	RGAT-BERT	86.60	81.35	78.21	74.07	76.15	74.88
Ours.	RoBERTa-SDACP	85.18	77.66	79.78	75.87	**77.89**	**77.26**
	RoBERTa-ACOP	**88.57**	**83.11**	**83.57**	**80.19**	77.60	76.94

Table 5: 多任务语义模型RoBERTa-ACOP和RoBERTa-SDACP在三个公认数据集上性能比较。表中的“Acc”表示准确率，“F1”表示Marco-F1 值，表中各项数值的单位均为%，以前方法中的最优结果使用下划线标记。

模型	Restaurant		Laptop		Twitter	
	Acc	F1	Acc	F1	Acc	F1
RoBERTa	84.38	77.26	78.83	74.53	74.57	74.02
RoBERTa-SDACP	85.18	77.66	79.78	75.87	77.89	77.26
RoBERTa-LACOP	86.61	80.22	82.62	78.99	77.60	76.94
RoBERTa-GACOP	88.57	83.11	83.57	80.19	77.02	75.55

Table 6: 消融结果。表中数据的单位均为%，模型中的“G”和“L”分别表示实现ACOP任务的两种方式全局实现(Globally)和局部实现(Locally)。

4.3 任务级消融实验

为了验证本文提出的辅助任务ACOP和SDACP的有效性，本文在任务级别上进行了消融实验进行对比，实验结果如表6所示。从表中可以得到，将ACOP任务和SDACP任务集成到RoBERTa模型的RoBERTa-GACOP、RoBERTa-LACOP和RoBERTa-SDACP都优于纯RoBERTa模型。例如，在Lap14数据集上，相比于纯RoBERTa模型，RoBERTa-SDACP模型在准确度和F1上分别提升了0.95%和1.34%，RoBERTa-LACOP模型在准确度和F1上分别提升了3.79%和4.46%，RoBERTa-GACOP模型在准确度和F1上分别提升了4.74%和5.66%，在另外两个数据集上也有不同程度的提升。另外，全局实现ACOP任务的RoBERTa-GACOP总体优于局部实现的RoBERTa-LACOP模型。以上分析证明了本章提出的ACOP任务和SDACP任务在提取语义方面的有效性，同时说明使用多任务方法相对于设计相应的模块引入外部知识处理目标任务ABSA确实是一个不错的选择。

4.4 参数研究

4.4.1 局部实现ACOP窗口大小的影响

在RoBERTa-LACOP模型的局部实现中，窗口大小(window_size)是一个重要的超参数，由于该参数会影响局部替换方面项和上下文片段时的范围，因此本文对窗口大小设置为1到10时

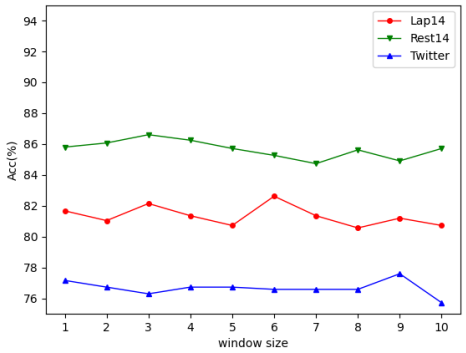


Figure 4: 窗口大小对准确率的影响

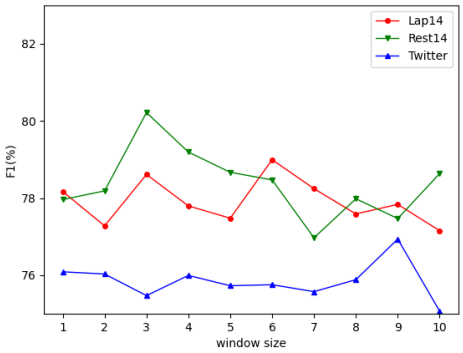


Figure 5: 窗口大小对F1值的影响

在Rest14、Lap14和Twitter数据集上进行了实验探索窗口大小参数的影响，结果如图4和图5所示。从图中，可以发现模型的性能不会随窗口大小单调变化。然而，我们观察到在Rest14数据集上窗口大小为3时达到相对最优的性能，在Lap14数据集上窗口大小为6时达到相对最优的性能，在Twitter数据集上窗口大小为9时达到相对最优的性能，这说明参数窗口大小确实会影响我们模型的性能。

4.4.2 捕捉句法依存距离的影响

在RoBERTa-SDACP模型的实现中，捕捉句法依存距离(cap_sd)是一个重要的超参数，由于该参数会影响模型捕捉句法信息的范围，因此本文做了扩展实验观察该参数设置为1到8时整个模型的表现，从而研究该参数对RoBERTa-SDACP模型的影响，结果如图6、图7和图8所示。

从图中，可以发现捕捉句法依存距离的变化确实会影响模型的性能。在三个数据集上都发现捕捉句法依存距离的参数设置为5时，模型性能达到最好的效果。本文对该实验结果可能的解释是：当该值设置得过小时，由于模型捕捉到的句法结构距离较短，不能充分地理解到观点词所表现的情感，当该值设置过大时，由于捕捉到的句法结构较大，虽然能够捕捉到观点词的情感信息，但是引入的过多的噪音会削弱模型对方面项对应的观点词的正确理解。

4.5 案例分析

为了说明本章提出的方法的有效性，本节对句子评论样本使用不同方法后的果进行比较，结果如表7所示，其中P表示积极情感，N表示消极情感，O表示中性情感。

	句子样本	方面项	情感标签	ATAE-LSTM	ASGCN-DG	RoBERTa-SDACP	RoBERTa-ACOP
1	Great food but the service was dreadful!	food, service	P,N	P,N	P,N	P,N	P,N
2	The staff should be a bit more friendly.	staff	N	P	N	N	N
3	Food was okay, just so so.	food	O	P	P	O	O

Table 7: ATAE-LSTM、ASGCN-DG、RoBERTa-SDACP和RoBERTa-ACOP的预测样例

观察表7可以得到：对于样本1，三个模型：ATAE-LSTM、ASGCN-DG、RoBERTa-SDACP和RoBERTa-ACOP都给出了正确的预测结果，这表明当样本语法结构简单且语义清楚的情况下，四种方法都能准确给出方面项对应的情感极性。对于样本2，结合了注意力机制的方法ATAE-LSTM做出了错误的判断，这是因为传统的基于注意力的方法不能识别到“should”是一个虚拟语气词，这个样本实际上表达了一个否定的含义，然而融合了语义信息的模型ASGCN-DG、RoBERTa-SDACP和RoBERTa-ACOP都能准确地给出方面项对应的情感

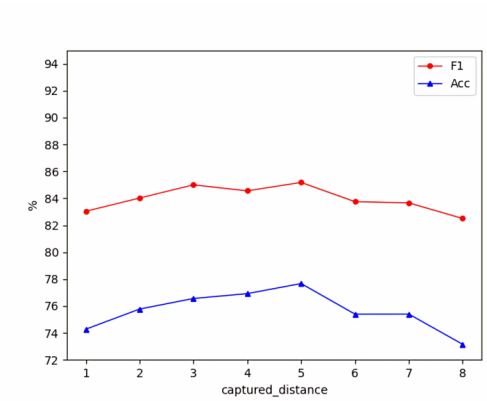


Figure 6: Rest14数据集上捕捉句法依存距离的影响

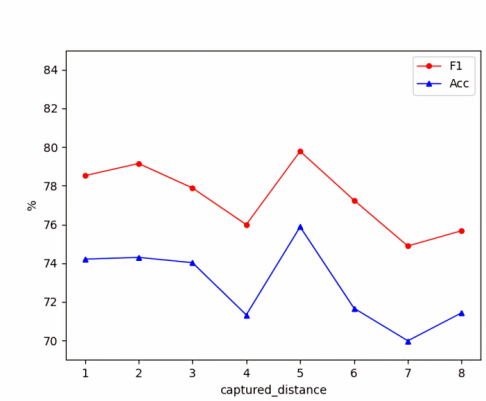


Figure 7: Lap14数据集上捕捉句法依存距离的影响

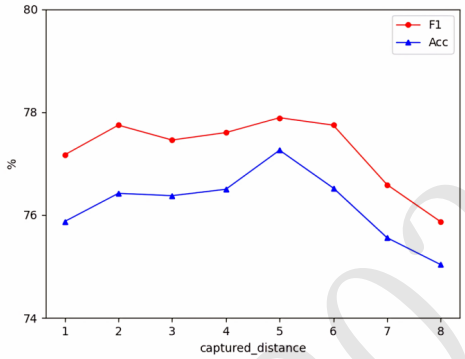


Figure 8: Twitter数据集上捕捉句法依存距离的影响

极性。对于样本3，习惯用语“just so so”削弱了“okay”情感词的积极情感，整个句子实际上表达的是一种中性的情感倾向，前两个模型ATAE-LSTM和ASGCN-DG都不能正确判断方面项对应的情感极性，只有模型RoBERTa-SDACP和RoBERTa-ACOP正确给出了情感极性，这说明了本章的模型对一些习惯用语也能提取到其语义特征。

5 总结

本文放弃了设计相应的模块引入外部知识而是从多任务的角度来解决原来的ABSA任务。本文分别提出了两个辅助任务：方面上下文顺序预测(ACOP)任务和方面上下文句法依存预测(SDACP)任务，并分别将辅助任务与原ABSA任务使用多任务方法进行训练，使得模型更好地提取句子的语义和语法特征。实验结果表明，模型在三个主要的公开数据集Rest14、Lap14和Twitter上的准确率和Macro-F1值都有较好的表现。本研究还可以从以下几个方面进一步改进:1)在ACOP任务的局部实现中，由于本文采用随机交换的方式探索方面项和上下文之间的顺序关系，引入了很多噪音，理论上会对ABSA任务带来一定的影响，因此采用一种启发式的方法来避免这些噪音带来的影响。2)探索多个任务联合辅助完成ABSA任务的可能性。

参考文献

Md Shad Akhtar, Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2017. Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems*, 125:116–135.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory

for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Chenhua Chen, Zhiyang Teng, and Yue Zhang. 2020. Inducing target-specific latent structures for aspect sentiment classification. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 5596–5607.

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

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3433–3442.

Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. 2018. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th international conference on computational linguistics*, pages 774–784.

Binxuan Huang and Kathleen M Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. *arXiv preprint arXiv:1909.02606*.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings 11*, pages 197–206. Springer.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

S Kiritchenko, X Zhu, C Cherry, and SM Mohammad. Detecting aspects and sentiment in customer reviews. In *8th International Workshop on Semantic Evaluation (SemEval)*, pages 437–442.

Deepak Kumar Gupta, Kandula Srikanth Reddy, and Asif Ekbal. 2015. Pso-asent: Feature selection using particle swarm optimization for aspect based sentiment analysis. In *Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings 20*, pages 220–233. Springer.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329.

Bin Liang, Rongdi Yin, Lin Gui, Jiachen Du, and Ruifeng Xu. 2020. Jointly learning aspect-focused and inter-aspect relations with graph convolutional networks for aspect sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics*, pages 150–161.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings*, pages 471–484. Springer.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

DG Maria Pontiki, HP John Pavlopoulos, and SM Ion Androutsopoulos. 2014. Semeval-2014 task 4: Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), Dublin, Ireland*, pages 23–24.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Bhargav Srinivasa-Desikan. 2018. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5679–5688.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6578–6588.

Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. 2014. Dcu: Aspect-based polarity classification for semeval task 4. In *SemEval@ COLING*, pages 223–229.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020a. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020b. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3540–3549.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*.

王婷，杨文忠. 2021. 文本情感分析方法研究综述. 计算机工程与应用, 57(12):11–24.

中国社会道德变化模型与发展动因探究 ——基于70年《人民日报》的计量与分析

王弘睿[1]，于东[2,*]，刘鹏远[2]，曾立英[1]
1.中央民族大学 国际教育学院，北京 100081
2.北京语言大学 信息科学学院，北京 100083
whongrui18@163.com，yudong@blcu.edu.cn，liupengyuan@blcu.edu.cn，lizzengliying@qq.com

摘要

社会道德的历时变迁研究具有重要意义。通过观察语言使用与道德变迁的历时联系，能够帮助描绘社会道德的变化趋势和发展规律、把握社会道德动态、推进道德建设。目前缺少从词汇角度、利用计算手段对大规模历时语料进行系统、全面的社会道德变迁研究。基于此，该文提出道德主题词历时计量模型，通过计量指标对1946-2015共70年的《人民日报》语料进行了历时计算与分析，观察了70年社会道德主题词的使用选择与变化。研究结果发现，道德词汇的历时使用与社会道德之间存在互动关系，反映出70年中国社会道德的历时变革与发展情况。

关键词： 社会道德；道德变迁；历时计量

The Model of Moral Change and Motivation in Chinese Society ——The Vocabulary Analysis of the 70-year "People's Daily"

Hongrui Wang[1]，Dong Yu[2,*]，Pengyuan Liu[2]，Liying Zeng[1]
1.College of International Education, Minzu University of China, Beijing 100081
2.College of Information Science, Beijing Language and Culture University, Beijing 100083
whongrui18@163.com，yudong@blcu.edu.cn，liupengyuan@blcu.edu.cn，lizzengliying@qq.com

Abstract

The study of the temporal changes in social morality is of great significance. Observing the diachronic relationship between language use and moral change can help to describe the change trend and development law of social morality, grasp the dynamics of social morality, and promote moral construction. At present, there is a lack of systematic and comprehensive research on social and moral changes in the large-scale diachronic corpus from the perspective of vocabulary and using computational means. Based on this, this paper proposes a diachronic econometric model of moral theme words, and calculates and analyzes the corpus of People's Daily from 1946 to 2015 through measurement indicators, and observes the diachronic changes and laws of social moral theme words in 70 years. The results show that there is an interactive relationship between the diachronic change in moral vocabulary and the change in social morality, which reflects the diachronic change and development of Chinese social morality in the past 70 years.

Keywords: Social Morality , Moral Change , Diachronic Measurement

*为通讯作者
基金项目： 北京市自然科学基金（4192057），中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)

1 引言

抗日战争后的70多年来，中国社会发生了巨大变革，道德观念也在不断地发展变化，社会道德的变革与发展成为具有研究价值的焦点问题。词汇作为最基本、最灵活的语言单位，记录了中国社会道德的历时变化，蕴含着民族道德观念的独特特征和丰富内涵。道德主题词指能够体现中国社会道德观念的代表性词汇，例如“诚信”“爱国”，是社会道德在文本中的集中表现。通过计算道德主题词的历时使用变化情况，能够反映社会道德的变化趋势和发展规律，帮助分析社会道德建设的重点内容和政策路线，对于把握社会道德观念动态、推进社会道德建设、服务相关政策和法律的制定与完善也有重要意义。

已有诸多研究证明，历时上词汇的使用选择与社会观念的变迁具有互动关系(Michel et al., 2011; Mihalcea and Nastase, 2012)，通过计算手段可以帮助在大规模文本中快速挖掘发生变化的相关词汇，并观察其历时发展情况。但现有社会观念的词汇计量研究多为政治领域的词汇使用研究(Paquot, 1983; 村田忠禧, 2003; 金观涛and 刘青峰, 2009)，伦理学领域的道德历时研究多为理论探讨和举例分析(柴文华, 2006; 陆远, 2014; 李建华, 2020)，缺少中国社会道德观念与词汇历时使用的关系研究。由于当代中国社会道德观念的复杂性，语料的时间跨度长、数据量大，利用计算手段、从量化角度考察大数据文本中道德主题词的动态性变化，可以为社会道德的历时变迁研究提供新的视角。

基于此，本文从用词特征探究中国社会道德历时变迁的内容与特点，提出能挖掘社会道德变化的词汇计量模型，对1946～2015年的《人民日报》语料进行了历时计算与分析，从词汇视角呈现了中国社会道德70年的历时变革与发展。本文的研究创新主要包括以下三个方面：

（1）提出道德主题词历时计量模型，通过计量指标挖掘道德主题词的历时使用变化和规律。分析计量结果证明，道德词汇的历时使用与社会道德建设之间存在互动关系，道德主题词能够展现中国社会道德的历时变革与发展情况。

（2）从宏观角度分析了70年中国社会道德历时变迁的整体规律，包括道德极性、道德强度和道德类型的变化趋势，分析蕴含其后的中国社会道德观念和道德建设路线的变革与发展趋势。

（3）从微观层面探索了70年中国社会道德的核心主题、时期特色、发展内容和变化特点等，观察了经济、政策、法律、社会事件等因素对道德主题词历时使用的影响作用。

本文为社会道德分析提供了新的研究视角，为语言计量、数字人文研究提供了新的分析维度和问题参考。

2 相关工作

2.1 社会道德变迁研究

道德是一直不断向前发展变化的，是一个善恶矛盾运动的过程(吴灿新, 2003)。社会道德变迁研究一直是伦理学的重要研究问题。

建国以来，中国社会道德变迁的研究主要包括两大方面内容，一类是国家层面下道德建设的方式、途径等的转变，主要是围绕经济、政治等与道德建设的关系来进行的，王一多(王一多, 1997)通过讨论经济生活、道德和政治法律的关系，论证了道德建设的基本途径。杜振吉(杜振吉and 王晓彦, 2007)通过观察道德变迁呈现的基本趋势，对道德建设提出了几点设想和建议。第二类内容是民众层面下道德观念的转变，柳礼泉 (柳礼泉and 刘佳, 2020)从新中国历史变迁视阈下，探究民众及社会道德观的演变。阎云翔(阎云翔, 2019)认为道德观从过去强调责任和自我牺牲的集体主义伦理，向一种强调权利和自我发展的个体主义伦理转变。但现有研究形式多为理论探讨和案例分析，缺少基于语料数据的实证研究。

2.2 词汇历时计算研究

词汇使用变化被广泛运用于反映社会变迁和建构理念，通过对词汇稳定性的测量，可以发现语言的使用规律，观察语言变化与社会文化、科学技术、政治经济发展的历时联系等。

政治话语的词汇选择与社会变迁的互动关系已有诸多论证，Thierry Paquot(Paquot, 1983)统计发现，词的使用频率能够反映各工会秉持的政治理念。村田忠禧(村田忠禧, 2003)分析表明，对关键词的计量分析可反映中共政治导向及方针政策的发展变化。金观涛(金观涛and 刘青峰, 2009)通过分析报纸期刊中观念词汇的词频变化,考察了中国现代政治术语的形成和演变。道德文本计算方面，Jing Yi Xiet等(Xie et al., 2020)通过历时词汇研究了几个世纪以来对

奴隶制和民主等概念的道德情操的历史转变，Aida Ramezanit等(Ramezani et al., 2021)探究了对于实体的道德感知如何随事件而变化，以及文本分析是否有助于抽取出导致这种道德变化的源头。

综上，目前社会道德变迁研究多为理论探讨和案例分析，且研究时段上多为讨论改革开放后的道德变化，缺少时间跨度长、系统全面的道德变迁趋势呈现。历时词汇计算研究多针对政治话语，道德词汇的历时计算仅有少量英文研究，且使用的是英语语言的道德体系，缺少汉语道德词汇的历时计算和分析研究。

3 研究设计

3.1 研究框架

语言中蕴含着丰富的道德现象，本文通过观察语料，从宏观规律发现和微观变化分析两个维度入手，根据道德变化的不同特征，提出道德历时变化的7个问题，具体如表1 所示。

分析维度	变化特征	探究问题
宏观规律发现	道德趋向 互动关系 发展动因	· 70年道德特征（极性、强度、类型）的变化趋势如何 · 道德词汇的历时使用与社会道德是否存在互动关系 · 道德词汇历时使用变化背后的发展动因是什么
微观变化分析	稳定程度	· 70年中国社会道德高频稳定的核心主题有哪些 · 不同年代的社会道德有怎样的时期特色
	变化幅度	· 70年中发生剧烈变化的是哪些道德主题 · 道德词汇的历时使用存在怎样的变化特点

Table 1: 研究框架及研究问题

这些问题覆盖了道德特征的趋势变化、道德内容的发展动因、稳定核心的道德主题以及变革发展的道德主题，体现了道德词汇与社会道德建设的互动关系，能够系统描述社会的道德变化和发展情况。

3.2 研究对象

本文选取《人民日报》语料作为中国社会道德历时变迁的研究对象，根据中文道德词典(王弘睿et al., 2021)的分类体系和数据资源对语料进行了道德主题词的识别和分类，结合伦理学研究成果进行核查，得到供历时分析的道德主题词库。

3.2.1 语料来源

报刊语料语言表述规范，内容覆盖生活的方方面面，记录了社会事件的时间特征。《人民日报》作为中国共产党中央委员会机关报，完整记录了中国政治、经济、文化等方面的历史事件，能够反映中国社会道德主题词汇的使用变化情况，具有重要的历时研究价值。本文选用1946-2015年的《人民日报》作为分析对象，经去除停用词、分词、词性标注等数据预处理工作后，得到历时70年的《人民日报》语料库(孙琦鑫et al., 2020)。

3.2.2 道德主题词的识别与分类

中文道德词典展现了中国社会的道德观念，能够反映社会道德的重点问题。本文选取的研究时段中，中国社会主流的道德观念始终是马克思主义道德观和社会主义道德观，词典词可以基本覆盖此时段的道德主题词。词典中共收纳了15,371个道德词汇，并将道德词归类为社会公德、职业道德、家庭美德和个人品德4类道德场景，每种道德场景下又划分了具体的道德行为类型，共24种(王弘睿and 于东, 2022)。

本文将这些道德词在《人民日报》历时语料库中进行道德主题词的识别和提取，保留在语料库中出现总频次数大于0的词，并匹配上中文道德词典的标签。每个词包含道德细粒度标签、历时语料中出现年份、年份对应词频和70年累计总频次数。然后，结合伦理学方面的道德历时研究成果，对提取的道德主题词进行了核查和补充，最终得到3,844个道德主题词。道德主题词库的具体数量及构成情况如表2所示。

道德场景 （总词数）	道德行为类型 （正向道德词数-负向道德词数）
社会公德 （1691）	文明礼貌/乐于助人/爱护公物/保护环境/遵纪守法 (115-444)/ (259-19)/(1-10)/(20-14)/(100-706)
职业道德 （538）	爱岗敬业/诚信经营/办事公正/热情服务/奉献社会 (129-46)/(2-85)/(57-103)/(1-9)/(76-30)
家庭美德 （117）	尊老/爱幼/男女平等/夫妻和睦/勤俭持家 (22-6)/(15-6)/(4-2)/(7-33)/(14-8)
个人品德 （1498）	爱国/奉献/勤劳/善良/宽厚/正直/自强/自律 (108-19)/(109-2)/(21-22)/(101-106)/(47-83)/(157-312)/(201-29)/(20-161)

Table 2: 道德主题词库数量及构成

3.2.3 时期划分

社会道德历时发展具有阶段性，为更好地反映道德主题词的使用变化情况，本文参考饶高琦、李宇明 (饶高琦and 李宇明, 2017)提出的分期体系，将70年划分为两层四段，并以此进行指标的计算和分析，具体划分结果如图1所示。

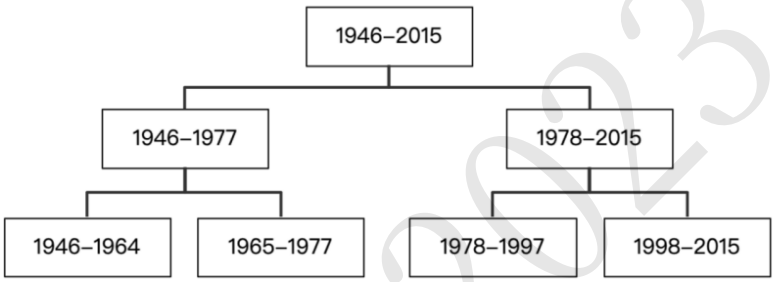


Figure 1: 70年时段划分方式

3.3 道德历时计量模型

本文构建道德历时计量模型的建模目标是挖掘道德主题词的历时使用变化情况，从而反应社会道德变迁的内容与规律。对于道德主题词w在时间点t时的道德特征m的变迁过程如式（1）所示。

$$P(m|w,t) = \frac{\sum_{d \in D_{w,t}} P_w(m|d)}{D_{w,t}} \quad (t \in t \sim t + \triangle t) \tag{1}$$

这里$D_{w,t}$是时间点t包含道德主题词w至少一次的文档集。

宏观层面的道德变化趋势计量指标又包括道德极性趋势、道德强度趋势和道德类型趋势，微观角度的计量指标包括道德稳定性、道德活跃度和道德核心指数。

道德极性趋势反应随时间t变化，道德极性MP的历时倾向性，计算如式(2)所示。

$$MP_t = \frac{4 \times PM_{f,t}}{PM_{f,t} + NM_{f,t}} + 1 \tag{2}$$

其中，$PM_{f,t}$为t年正向道德的总词频数，$NM_{f,t}$为t年负向道德的总词频数。MP_t结果值为5时，表明该年表征正向道德的程度最深；值为3时，说明该年道德表征偏中性，值为1时，说明该年道德表征偏负向。

道德强度趋势反应随时间t变化，道德强度MI的历时波动情况，计算如式(3)所示。

$$MI_t = \frac{\sum_{w=1}^{N}(F_w \times MI_w)}{\sum_{w=1}^{N} F_w} \tag{3}$$

第二十二届中国计算语言学大会论文集，第289页-第299页，哈尔滨，中国，2023年8月3日至5日。

其中，Fw表示词w在历时语料中出现的频次数，MI_w表示词w的道德强度。

道德类型趋势反应随时间t变化，道德类型MR的丰富程度变化，计算如式(4)所示。

$$MR_t = |mean_{f,t+\triangle t}/stdev_{f,t+\triangle t} - mean_{f,t}/stdev_{f,t}| \tag{4}$$

其中，$mean_{f,t}$表示t年词频的平均值，$stdev_{f,t}$表示t年词频的标准差。

衡量道德稳定性，需要考虑道德词w历时使用的平均频次及波动变化，稳定性系数S的计算公式如式(5)所示：

$$S_w = \frac{\bar{w_f}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(w_f - \bar{w_f})^2}} \tag{5}$$

其中，$\bar{w_f}$代表的是词w历时使用的平均频次。稳定性参数S_w越小，说明词w历时使用稳定性越低。

衡量道德活跃度，需要考虑词w的词频数和分布文档数。活跃度系数L的计算公式如式(6)所示：

$$L_w = \frac{F_w}{F} \cdot log\frac{|D|}{1 + |D_w|} \tag{6}$$

其中，Fw和Dw分别表示词w在历时语料中出现的频次数和语料中包含词w的文档数。F和D分别表示语料中全部词的总频次数和文档总数。道德活跃度L越高，代表该词汇在70年中高频活跃。

道德核心指数词MC既需要考虑词汇历时使用的稳定性，又要考虑词汇使用的高频活跃程度，计算公式如式（7）所示。

$$MC_w = |S_w \cdot L_w| \tag{7}$$

这些指标的计算结果能够帮助从大规模文本中挖掘出有使用变化的词汇，并展现出其历时的分布情况，从而辅助历时分析。

4 宏观趋势分析

宏观计量指标反映了中国社会道德变迁的整体趋势和方向调整。本章讨论了70年道德特征的变化趋势，包括道德极性、道德强度和道德类型三个方面。而后探究了道德词汇历时使用变化背后的发展动因，验证了道德词汇历时使用与社会道德建设的互动关系。

4.1 道德特征变化趋势

本节从道德极性、道德强度和道德类型三个道德特征维度分析了中国社会道德历时变迁的整体规律和发展趋势。

4.1.1 道德极性变化趋势

根据道德极性MP的计量公式，70年中国社会道德极性变化趋势如图2所示。可以看出，70年道德极性值均大于3，说明人民日报语料文体风格整体呈正向道德色彩。建国初期报道正负向内容相对均衡，道德色彩偏向性不大，1990年后极性值大于4，开始明显偏向正向道德色彩，最高点为举办北京奥运会的2008年，体现道德极性与时间点的交互关系。

4.1.2 道德强度变化趋势

根据道德强度MI的计量公式，按5年的时间间隔计算70年中国社会道德强度变化趋势如图3所示。道德主题词库中道德词的平均强度为1.355，可以看出，自1966年以来，人民日报语料的道德强度一直低于平均道德强度，且人民日报语料的道德强度整体呈下降趋势，说明高强度的道德词使用有所减少，道德环境向着宽容开放的趋势发展。

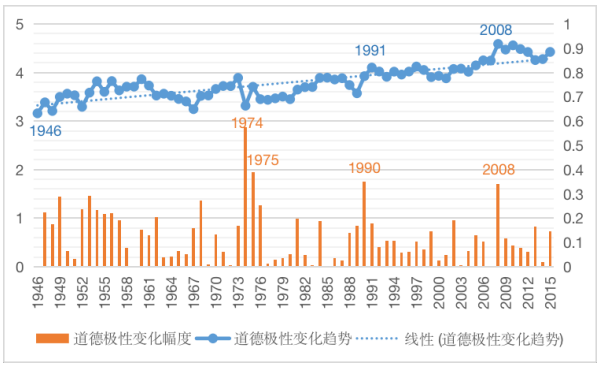


Figure 2: 70年道德极性趋势图

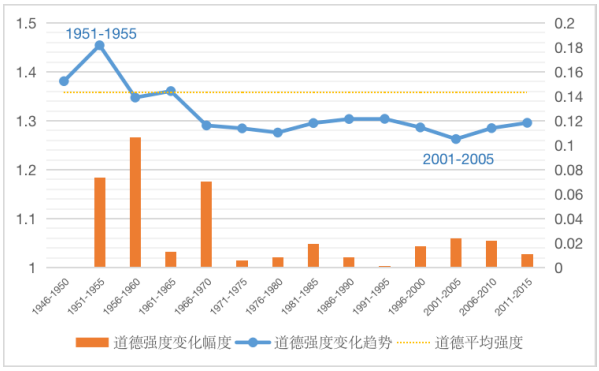


Figure 3: 70年道德强度趋势图

4.1.3 道德类型变化趋势

根据道德类型MR计量公式，70年中国社会道德类型变化趋势如图4所示。可以看出，道德类型随着时间的变化呈上升趋势，说明随着平等、开放的现代生活方式的逐步确立，普适性、多元性的道德价值观应运而生。道德类型最为一元集中的时期分布在1967-1972年间，而后开始上升，体现出社会道德规范从典范性向群众性的转变，从强调无私无畏、完全奉献，走向和谐、多元的、平民化的底线道德和基本伦理。道德类型最为多元的为1979-1983和1989-1991年间，主要由于这一时期新旧体制交汇，市场经济在肯定个体价值的同时，冲击了革命战争时期以共产主义理想为主线的道德观念，旧有的道德标准受到质疑，又尚未建立新的道德标准，社会缺乏道德共识和共同遵守的道德准则。2008年后道德类型开始有所下降，体现出随着社会经济的稳定发展和道德建设的完善，社会道德共识程度的提升。

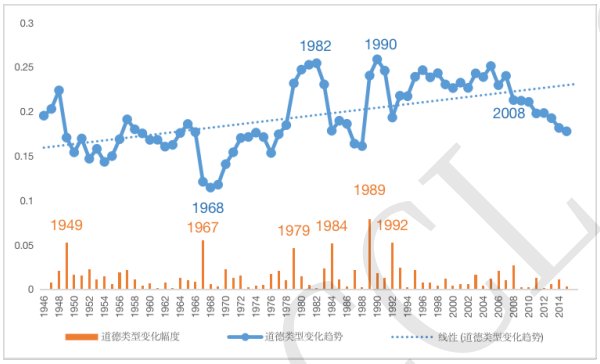


Figure 4: 70年道德类型趋势图

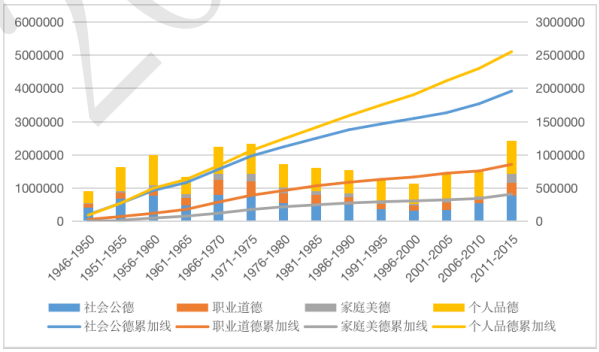


Figure 5: 70年道德场景变化图

另一方面，不同道德场景的变化程度也有所不同，如图5所示，个人品德类道德词增加最为剧烈，家庭品德和职业道德增加较为平缓。这种道德场景的波动变化体现出道德建设主题的内容选择和方向调整。中国传统的礼让、互助等美德，在当今社会仍然是重要的道德原则。而进取型、创新型人格为个体品德的发展注入了自信、勇敢、勤劳、顽强等道德要求，西方强调的契约精神、平等观念和权利意识等逐渐也得到重视，既提倡道德义务也尊重道德权利，体现出道德关系向着“平等”和“相互”的方向转变。

4.2 社会道德发展动因

本节讨论了经济、政策、法律、社会和文化等因素对道德词历时使用的影响作用，并通过相关性分析验证道德词历时使用与社会道德建设的互动关系。

4.2.1 经济因素

一个社会基本的价值观念和伦理道德，归根到底是建立在现实的经济生活的基础之上，并由

这种经济生活所决定的。经济对道德的影响一方面体现在道德词使用变化的时期分布，经济体制发生变化的时期，道德词也会随之有明显使用变化。70年中，第三时段发生变革的道德词量最多，而这一时期也正是中国经济体制改革的起始，其中发展主题占比最高，新生道德词也开始增加，体现出这一时期由于经济体制的变化，大量道德主题词由低频进入高频的使用变化，如“廉洁自律”“脱贫致富”等等。

另一方面，经济因素也会间接对道德词的使用产生影响。例如“抗震救灾”一词，与其历时使用直接相关的是地震灾害发生的年份。但将《中国环境统计年鉴-全国自然灾害情况（2000-2015年）》中地震灾害的数据与“抗震救灾”一词在语料中的对应年份词频使用斯皮尔曼相关系数进行相关性分析发现，如表3 所示，与该词历时使用变化最高度相关的是地震造成的直接经济损失，体现出经济因素对道德词的间接影响作用。

统计数据		人员伤亡		直接经济损失		灾害次数	
抗震救灾	Spearman相关系数	0.614	*	0.845	***	0.542	*
	p值	0.011		<.001		0.03	

Table 3: “抗震救灾”历时相关性分析

4.2.2 政策、法律因素

考察发现，道德主题词进入高频使用的时间点往往与政策的颁布或法律的制定有关，呈现突变性剧烈变化。例如，“精神文明”一词频序虽有过波动，但最终能够进入高频使用，是在1981年6月，中共十一届六中全会审议通过的《关于建国以来党的若干历史问题的决议》把社会主义精神文明建设归纳为社会主义现代化建设道路的十个要点之一。自此，“精神文明”才成为高频道德主题词。

通过对《中国统计年鉴（1981-2015年）》相关统计数据与道德词的历时使用进行斯皮尔曼相关性分析发现，政策和法规对道德的影响呈现两种不同趋向，一类是高度正相关、趋向高频稳定使用的道德词，一类是呈现负相关，趋向退出高频的道德词。如表4所示，“创新”一词的历时使用与科技活动人数、科学家和工程师人数呈现高度正相关，而“保护区”的历时使用与自然保护区的个数呈现高度负相关，这一点在体现道德历时变化动因的同时，也反应出道德内容变化的趋向性，在自然保护区达到一定数量后，“保护区”一词会逐渐退出高频使用。这种影响也体现出道德词历时使用与道德建设的互动关系。

	创新	科技活动人数		科学家和工程师人数	
正相关、趋向高频	Spearman相关系数	0.794	***	0.795	***
	p值	<.001		<.001	
	节能	能源加工转换效率		能源消费量	
	Spearman相关系数	0.64	***	-0.428	**
	p值	<.001		0.009	
	保护区	自然保护区个数		自然保护区面积	
	Spearman相关系数	-0.836	***	-0.193	
负相关、趋向波动	p值	<.001		0.49	
	农村贫困人口	脱贫致富		扶贫	
	Spearman相关系数	-0.583	***	-0.644	***
	p值	<.001		<.001	

Table 4: 政策因素相关性分析

4.2.3 社会、文化因素

一些词由于全球化的背景和相关社会新闻事件的发生而进入高频使用。例如，“反恐”一词在2001年产生了使用的波动，2001年9月11日，本拉登领导的基地组织对美国发动了恐怖主义袭击，造成3000多人丧生。美国随即宣布反恐战争在全球的开始。这一事件之后，“反恐”成为高频使用的道德主题词。

文化思想对道德的影响主要体现在时期主题，如“大无畏”等词，只在特殊时期高频使用，在文化运动结束后逐渐退出使用，且没有回归高频的趋势，可见文化思想对道德的影响作用主要停留在文化思想盛行时期。

还有一些词的使用变化是由于词汇历时使用的竞争关系，如随着时间的发展，“法规”一词在词汇竞争中胜出，逐渐替代“法令”一词。

综上，从宏观趋势来看，70年间中国社会的道德传播及建设整体朝正向、良性方向发展，社会道德标准更加具有包容性，呈现多元化的发展趋势，发展内容和变革趋势受到经济、政策、法律和社会事件等因素的影响，体现出道德词汇历时使用与社会道德建设的互动关系。

5 微观变化分析

本章根据前述分析框架中的微观分析维度，分析了70年中国社会的道德核心主题、时期特色、变革内容和变化特点。

5.1 社会道德核心主题

本文根据时期划分，分别计算两时段的道德核心指数MC，如图6～图7所示，y轴为词w在1946-1977时段的道德核心指数，x轴为该词在1978-2015时段的道德核心指数，气泡大小为总词频数。可以看出，词的y值低，x值高，说明其在后一时段开始高频使用，属于发展变革主题，如“改革”“犯罪”“污染”等。词的y值高，x值低，说明其在前一段高频使用，第二时段使用频率大幅下降，属于调整消亡主题，如“抓革命”“反动”等。x值和y值均较高的词为稳定使用的核心主题，如“努力”“破坏”等。

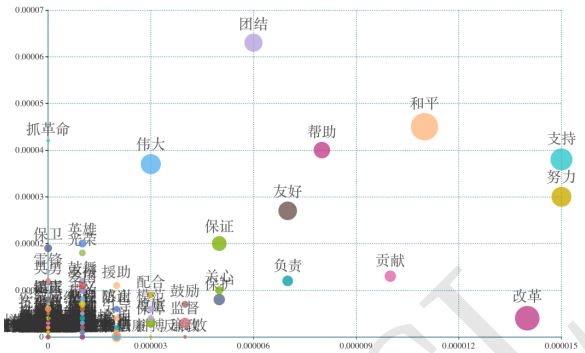


Figure 6: 正向道德核心指数图

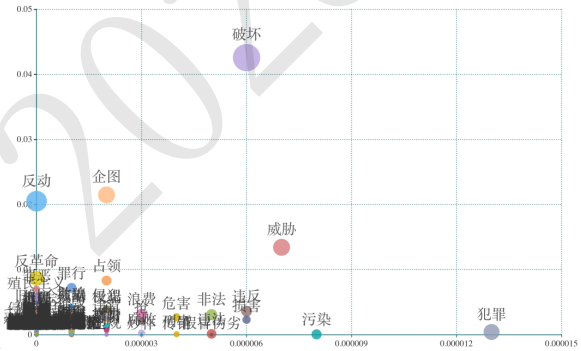


Figure 7: 负向道德核心指数图

高频稳定使用的道德主题词“和平”“团结”“帮助”“友好”等反映了中国社会道德的整体面貌和重点内容，体现出国家与社会对倡导和平、维护团结、促进友好等方面道德主题的重视与坚持，也体现出强调不断深化改革、努力奋斗的时代特性。

高频稳定道德主题词在不同时段的共现搭配词有所不同。如表5所示，“努力”一词在时段一中的高频共现词是“学习”，在时段三、四中则变为“经济”。“改革”一词在时段一中的高频共现词是“技术”，在时段三变为“经济”，时段四则变为“制度”。体现出随着时代的发展变化，原有道德主题不断增加着新的含义、新的内容和新的要求。

道德主题词	1946～1964		1965～1977		1978～1997		1998～2015	
	共现词	共现次数	共现词	共现次数	共现词	共现次数	共现词	共现次数
努力	学习	2354	改造	1432	经济	4040	经济	4019
改革	技术	3860	规章制度	719	经济	17847	制度	13805
团结	力量	7587	两国人民	4237	国家	3629	党	1874

Table 5: 四时段高频稳定道德主题词的共现词示例

5.2 社会道德时期特色

本文选取的研究时段的主流观念始终是马克思主义道德观和社会主义道德观，道德主题具有稳定性和延续性。但人类社会的发展变化有着鲜明的时间阶段性，社会历史发展的各阶段有其时代独特的内容与特征，因此社会道德也就不可避免地具有浓郁的时代色彩。由于经济结构的调整变化，这一阶段又可再分为多个时段，每个时段有自己独特的道德核心内容。

根据每个时段道德主题词的内容构成，可以透视出该阶段社会道德的时期特征。社会道德在本质上属于社会的上层建筑，社会经济决定了社会道德的基本形态。根据中国社会经济发展进程，70年可以划分为两个不同的时代：传统计划经济时代和社会主义市场经济时代，因此也就可以划分出两个时代对应的道德形态。去除两时段均高频稳定的道德主题词后，根据道德活跃度L得到两时段的道德主题词云如图9所示.

Figure 8: 道德时期特色词云

可以看出，改革开放前，传统计划经济时代的道德主题重点主要体现在爱国、奉献等方面，改革开放后，社会主义市场经济时代道德则转为强调创新、法制、环保、社会公平和精神文明建设等方面。前一时期，中国社会经历了内地全面解放、抗美援朝、镇压反革命等一系列社会改造运动，初步建立了社会主义制度，道德主题词集中体现在“爱国”这一核心点上。其中，“保卫”一词出现词频最高，其他与保家卫国相关的词汇也出现较多，如“志愿军”“英勇”“正义”“爱国主义”等。后一时期，对“创新”的强调，反映出随着社会主义市场经济体制的完善和国人受教育程度的普遍提高，人们的自主意识、竞争意识和效率意识等均有较大进步，市场经济也对人们的创新能力提出了新的要求。“精神文明”“治理”“依法”“法规”等词的高频使用，则体现出该时期对社会主义精神文明建设、社会主义法制建设、社会主义生态文明建设等方面的重视。另一方面，随着改革开放力度的加大，在经济快速发展的同时，对利益的过度追求导致道德标准的迷失。与环境污染、贪污腐败、贫富差距变大等社会现象相对应的，“环保”“廉政”“扶贫”等方面道德主题开始得到重视。

从上述道德主题词的高频使用及时间分布情况可以看出，道德核心指数和道德活跃度能够较全面地展现中国社会道德建设的核心主题和重点内容，帮助把握各时期的道德特色。

5.3 社会道德变革内容

根据稳定性系数S和道德主题词使用的变化方向，可分为进入高频使用的道德主题词，即新生、发展主题和退出高频使用的道德主题词，即调整、消亡主题。如图9所示，时间轴上方展示了四个时段中分别增加的变革型道德主题词和其进入高频使用的变化时间节点，例如，1994年，“爱心”“敬业”“反腐倡廉”等词成为高频使用的道德主题词。时间轴下方为退出高频使用的道德主题词及变化时间。

道德主题词这种强烈的历时变化，是具有重大意义的历史变革在具体词汇使用上的表现，体现了中国社会道德内容的历时选择和发展。根据道德内容划分，新生和发展的道德主题主要集中于完善社会主义法制建设、完善社会主义生态文明建设、促进社会公平正义、对弱势群体的关注和缩小贫富差距的努力等方面。

另外，一些道德主题词随着时代的变化发展，尤其是由传统计划经济转向社会主义市场经

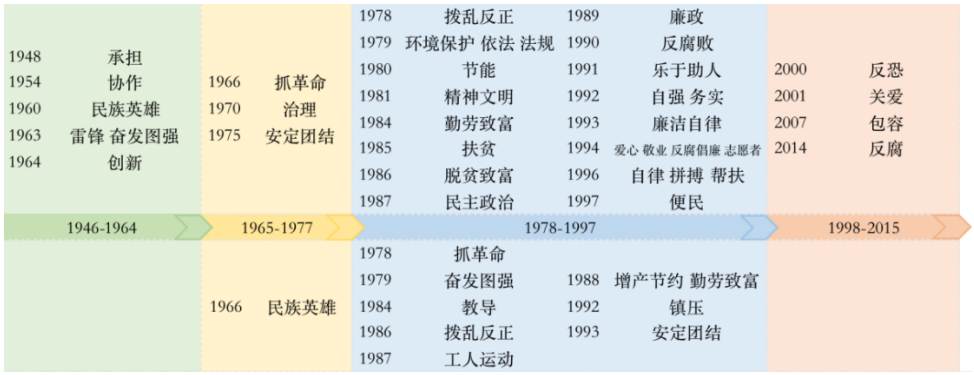


Figure 9: 社会道德变革内容

济后，不再是国家和社会面对的主要矛盾，因此也不再是道德传播的重点内容，逐渐退出高频使用，成为调整甚至消亡的道德主题。例如“工人运动”一词在1987年后退出高频使用，体现出市场经济和全球化使传统的运动形式难以适应现代社会的需要。

5.4 社会道德变化特点

道德的不变性是相对的，变化性则是绝对的。在人类社会的历史长河中，道德一直在发展变化。根据道德主题词的使用变化过程，可分为突变型使用变化、渐变型使用变化和曲折型使用变化。如“环境保护”一词进入高频使用前频序较稳定，没有明显波动。进入高频使用的变化过程短、变化情况剧烈，呈现出突变性；“精神文明”一词在进入高频使用前有多次使用频序的波动，呈现出渐变性的变化特点。

部分道德主题词仅在特定历史时期短暂高频使用，而后由于不再符合社会发展进程，退出了高频使用。如“抓革命”一词仅在第二时段被高频使用，随着文革时期的结束，退出了高频使用。这一点反映出社会道德历时发展的曲折性特点。

从道德词汇历时使用的稳定程度和变化程度进行分析发现，道德主题词的历时使用变化与社会道德的变革节点和发展方向高度相关，道德稳定性、活跃度和道德核心指数能够反映社会道德的变化特点和发展内容，帮助描绘社会道德的变迁过程。

6 总结与讨论

道德主题词的历时使用变化是社会道德变迁在词汇层面上的投射。本文选取了1946～2015共70年的《人民日报》语料，根据中文道德词典的标签体系和数据资源对语料中的道德主题词进行了识别和分类，得到供历时分析的道德主题词库。提出道德词汇历时计量模型，宏观层面的道德变化趋势计量指标包括道德极性趋势、道德强度趋势和道德类型趋势，微观角度的计量指标包括道德稳定性、道德活跃度和道德核心指数，根据计量结果观察了道德主题词使用与社会道德发展的历时联系，对70年中国社会道德的整体趋势、核心主题、时代特色、发展内容、变化特点进行了探索与分析，并观察了经济、政策、法律、社会事件等因素对道德主题词历时使用的影响作用。主要研究结论如下：

（1）道德主题词能够反映社会道德的变革与发展。70年间，中国社会道德的传播及建设整体朝良性方向发展，涉及道德行为从个体使用和组合结构上均逐渐摆脱一元集中性，呈现多元化趋势。这与中国社会政治民主化、社会主义法制的逐步完善密不可分。社会道德的多元趋势也反映出市场经济的发展带来了更加复杂多样的道德事件，对中国社会道德建设提出了新的要求。（2）道德主题词的历时使用能够体现出社会道德的核心主题和时代色彩。70年中有一直高频使用的道德主题词，如“团结”“和平”等，体现出中国社会道德的传承性。随着社会的发展，不同时期的道德重点内容又有所不同，改革开放之前更强调“爱国”与“奉献”，之后则更注重“法制”和“创新”，体现出社会道德阶段性的时期特征。（3）道德主题词也展现出社会道德的历时发展变化。有些词经历从无到有，进入了高频使用，比如“扶贫”，有些词随着社会的发展使用越来越少，比如“工人运动”。发展内容主要体现在社会主义法制建设、生态文明建设和对社会公平正义的维护和努力等内容，呈现出突变性、渐变性和曲折性的变化特点。（4）道德主题词的使用变化受经济、政策、法律、社会事件等因素的影响。其中，经济因素往往起决定性作

用，政策、法规、社会事件对道德变化的时间点有明显影响作用，文化思想对道德的影响作用主要停留在文化思想盛行时期。

道德主题词是社会道德建设和传播趋向的反映，是刻画社会整体道德传播方向、判断社会道德文明建设变化趋势的重要工具。本文从计量角度对道德主题词在各维度的呈现进行了观察，是对社会道德历时变迁进行可视化分析的一次有益尝试。当然，目前道德主题词库还不能完全覆盖所有历时的道德事件，未来计划进一步完善历时道德主题词库的数据资源，利用各时期的真实语料补充中国社会道德变迁的词条，从而获得更为精确的中国社会道德历时发展面貌。

参考文献

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 1:176–182.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–263.

Thierry Paquot. 1983. Groupe de saint-cloud: la parole syndicale paris: Puf, 1982. *Autogestions*, 14:122–123.

Aida Ramezani, Zining Zhu, Frank Rudzicz, and Yang Xu. 2021. An unsupervised framework for tracing textual sources of moral change. *arXiv preprint arXiv:2109.00608*.

Jing Yi Xie, Renato Ferreira Pinto Jr, Graeme Hirst, and Yang Xu. 2020. Text-based inference of moral sentiment change. *arXiv preprint arXiv:2001.07209*.

吴灿新. 2003. 简论道德的发展变化. 伦理学研究, 1:21–24.

孙琦鑫, 饶高琦, and 荀恩东. 2020. 基于长时间跨度语料的词义演变计算研究. 中文信息学报, 34(8):10–22.

李建华. 2020. 从适应性看道德的变化. 江海学刊, 4:48–52.

村田忠禧. 2003. 从改革开放以来的党代会政治报告的词语变化来看中共十六大的特点. 中共党史研究, 1(80-85).

杜振吉and 王晓彦. 2007. 当前中国社会道德生活的变革及其基本趋势. 当代世界与社会主义, 3:151–156.

柳礼泉and 刘佳. 2020. 新中国70 年榜样文化建构的演进图式及其启示. 伦理学研究, 4:16–21.

柴文华. 2006. 论中国现代道德变迁的一般特征. 求是学刊, 9(117-123).

王一多. 1997. 道德建设的基本途径——兼论经济生活, 道德和政治法律的关系. 哲学研究, 1:7–12.

王弘睿and 于东. 2022. 面向机器道德判断任务的细粒度中文道德语义知识库构建. 中文信息学报, 36(7):59–68.

王弘睿, 刘畅, and 于东. 2021. 面向人工智能伦理计算的中文道德词典构建方法研究. 中文信息学报, 35(10):39–47.

金观涛and 刘青峰. 2009. 观念史研究*:* 中国现代重要政治术语的形成. 法律出版社.

阎云翔. 2019. 当代中国社会道德变革的轨迹. 思想战线, 1:93–105.

陆远. 2014. 国家导向下的道德进步运动及其中国实践. 南京社会科学, 9:50–56.

饶高琦and 李宇明. 2017. 基于词汇聚类方法的现代汉语分期与分期体系构建. 中文信息学报, 31(6):18–24.

动词视角下的汉语性别表征研究 ——基于多语体语料库与依存分析

陈颖诗, 于东, 刘鹏远*
北京语言大学信息科学学院，北京*100083*
chenyingshi411@163.com, yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

摘要

动作是反映性别社会化的重要形式，研究汉语中动词的性别表征，可以找到语言构建不同性别身份的路径，即所采用的方式、形式。本文以依存句法关系为抓手，在四种语体的语料中抽取出和不同性别词构成依存结构的动词，统计出有显著性别差异的动词，并根据性别词充当的句子成分，结合语义进行了定量和定性分析。总体来看，大部分汉语动词表征是中性的，能体现性别的动词是少数，汉语作为一种承载着中华智慧且具有深厚文化底蕴的语言，对性别的表征是中立且平等的，这也体现出了我国的性别平等观念。而在表征性别的动词中，能看到构建男性和女性身份的两种不同路径。显著表征女性的动词在不同语体的语料中均多于显著表征男性的，但是表征男性的动词的语义分布则更为均衡，体现了“男性默认-女性专门”。在司法动词上，女性常常作为暴力行为的受害者，同时施害者男性却隐身了，体现了“男性主宰-女性顺从”。不同语体的动词在构建性别时体现了不同的功能，新闻塑造了较为传统的性别规范，传统和网络文学以不同的形式打破了固有的性别规范。

关键词： 动词；性别；依存分析；多语体

Gendered Representation in Chinese via Verbal Analysis —Based on a Multi-register Corpus and Dependency Parsing

Yingshi Chen, Dong Yu, Pengyuan Liu*
School of Information Science, Beijing Language and Culture University, Beijing 100083
chenyingshi411@163.com, yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

Abstract

Actions can reflect the process of gender socialization. Studying the gendered representation in Chinese verbs can reveal how language is used to construct different gender identities. Starting from the perspective of dependency relations, this paper extracts verbs that collocate with different gender words from a multi-register corpus, and statistically identifies verbs with significant gender differences. Quantitative and qualitative analyses are then conducted based on the sentence components that gender words serve, combined with their semantics. In general, Chinese verbs that reflect gender are few, and the majority of Chinese verbs are neutral in terms of gender representation. As a language that embodies Chinese wisdom and has a rich cultural

*为通讯作者
基金项目：北京市自然科学基金（4192057），中央高校基本科研业务费(北京语言大学, 23YJ080009)

heritage, Chinese representation of gender is neutral and equal, reflecting China's concept of gender equality. Among the verbs that can reflect gender, two different paths for constructing male and female identities can be observed. Female verbs are more common than male verbs in all registers, but the semantic distribution of male verbs is more balanced, reflecting a "male default - female specialization" pattern. In judicial verbs, women often appear as victims of violence while the perpetrators, men, remain invisible, reflecting a "male dominance - female submission" dynamic. Verbs in different registers have different functions in constructing gender. News shapes traditional gender norms, while traditional and online literature challenge inherent gender norms in different ways.

Keywords: verb , gender , dependency parsing , multi-register

1 引言

"花婆婆"方素珍有一首童诗《收获》，写父亲和孩子钓了半天鱼，却把鱼都放了，两手空空地回家。放走鱼的原因各不相同："鱼小弟/年纪太小""鱼爸爸/年纪太大"，放走"鱼妈妈"的理由则是"她要回家/照顾鱼爸爸/她会想念/鱼小弟"。传统社会对男女角色的分配是"男主外，女主内"，男性赚钱养家，女性则负责生育和照顾家庭。"五四运动"掀起了救国图存的浪潮，也同样掀起了性别革命、家庭革命的热浪，一部分受过教育的女性从私人家庭走入公共生活，从此开启女性作为现代社会中的独立个体，而非仅仅是传统社会中以母亲、妻子、女儿等性别化身份进入公共领域的历程(Brownell and Wasserstrom, 2002; 卢淑樱, 2020; 高晓君and 魏伟, 2022)。新中国成立之后，男女平等被写入宪法，成为基本国策；在"妇女能顶半边天"的感召下，女性不再局限于家庭内部的再生产劳动，也进入公共领域从事生产性劳动(Evans, 1997; Su et al., 2021)。性别是一种社会建构，它不是一个人所拥有的，而是在特定社会文化和社会历史文本中表现出来的属性(Angouri and Baxter, 2021)。在近百年来女性角色的重大变化中，语言也是其构建过程的一个中心特征(Butler, 1988)。语言通过不断的话语重现建构出社会角色，名词和代词能直接体现出社会对男性、女性的称呼与认知，形容词代表了对其外表和内在的期望与规约。考察动词，则能窥探出当今社会对于不同性别群体所期待的动作行为，即语言通过动词建构的不同性别身份的路径。

汉语是一种无语法上的性范畴的语言，动词不能直接体现出性别。因此，考察汉语中动词与性别的关系，就需要借助动词与性别词的搭配。传统的语料库搭配抽取主要依靠动词和性别词的位置关系，不能直接体现出两者的语法关系。比较而言，依存分析方法（Dependency Parsing）抓住的是两个词语之间的句法依存关系，在此基础上抽取，能更准确地得到具有语法关系的动词和性别词词对。此前对汉语动词和性别的研究主要依托单种语料，如微博或新闻，为了能更全面准确地考察动词和性别在语言生态中的关系，本文选择了一个包含了新闻、传统文学、网络小说和微博的大规模多语体语料库，利用依存分析工具从中抽取出男性性别词或女性性别词充当动词主语或宾语的结构，即SBV结构和VOB结构，如图1所示（依存结构的取舍见附录6）。经过统计检验，最终得到在构成结构时具有显著性别差异的四类动词：男性支配动词、男性承受动词、女性支配动词和女性承受动词。在结合动词语义分析的基础上，我们对这些显著表征某一性别的动词进行了定量和定性分析。本文的主要研究问题为：

（1）总体上，显著表征男性的动词和显著表征女性的动词体现出怎样的性别身份构建路径?

（2）不同性别支配和承受的动词有怎样的异同？体现出了怎样的性别规范?

（3）不同语体的动词在构建性别身份时存在怎样的功能差异?

2 相关工作

2.1 语言中的动词和性别

美国社会学家Cooley (1902)将"自我"概念引入社会化研究，认为个体通过与他人的接触来了解自己，把别人当作镜子来看自己的行为动作，从而产生自我感知。性别的社会化同样依赖于行为动作，而这种社会化又通过语言的动词表达出来。动词在语言研究中十分重要，它在句

第二十二届中国计算语言学大会论文集，第300页-第314页，哈尔滨，中国，2023年8月3日至5日。

法结构中活动能力最强，大部分的词类都要同它发生组合关系，还常常作为谓语出现，成为句子的结构联系和语义联系的中心。动词和性别关系的研究大致可以分为两类。

第一类直接从动词的语法上的性范畴出发。例如在俄语中，动词的过去式包含与主语一致的性别标记，由此可以直接探索俄语动词的性别分布。俄语动词词尾表明俄国社会将男性和职业活动联系在一起，而将女性和生育、家务联系在一起；社会期望女性更多地表达自己的感受和情绪，而男性比女性更容易受到坏习惯的诱惑，如吸烟饮酒；男性和快速坚定的行走具有联系，而和女性相关的动词中，有两个与巫术相关，反映了女性与超自然世界联系在一起的刻板印象(Kuznetsova, 2015)。然而，汉语中没有语法上的性范畴，不能直接根据动词的特征进行分析。（当然，性范畴同性别本来也不一定一致。）

第二类从语料库语言学（corpus linguistics）的角度，研究与不同性别词搭配的动词的差异，进而反映出性别构建路径。例如，对推特和英国网络文本语料分析发现，与女性搭配的动词多表达情感需要和生育，而与男性搭配的动词多表达赚钱、权力的含义(Herdağdelen and Baroni, 2011)；在英语语料库中，“girl（女孩）”更多地与“dance（跳舞）”等艺术类动词搭配，而“boy（男孩）”则与“run（跑步）”“swim（游泳）”和“throw（投掷）”等具有实际动作的动词搭配(Taylor, 2013)；与“girl”搭配的动词多表达特定的情绪、感情和认知状态，并多描述其穿着，且是一些暴力行为（如强奸、绑架、谋杀）的受害者，而“boy”则是物理动作和状态的主语，总体来看并不是暴力行为的受害者(Baker, 2014)。这些都表明了在英语中男性被构建成一个施加暴力、获得权力、活泼好动的形象，而女性则被构建为一个接受暴力、生育子女、富有艺术气息的形象。但是，这些语料库中的搭配方法建立在动词和性别词的位置关系上，具体的语法关系则需要人工标注，无法应用在更大规模的语料分析中。

2.2 现代汉语中的动词和性别

在已有的对现代汉语动词与性别的相关研究中，有依托博客、微博语料对“喜爱”类和“怨恨”类心理动词进行考察，研究这些心理动词在分布、句法、语义和语用等方面的性别差异(徐修倜, 2017; 詹秀红, 2017)；有对转述动词的研究，发现新闻报道中被转述的女性发言者数量少于男性，转述女性使用的动词类型不如男性的丰富(张滢，曹榛, 2013)；也有从心理语言学角度探讨汉语动作动词的性别编码，认为这种性别编码倾向对动词的加工具有启动效应、影响句子主语的选择和人们对两性性别的社会认知(段新焕, 2007)。除了针对某一类动词的，也有抓住汉语动词搭配进行的研究，比如，在语料库上通过抽取“他/r+.../v”和“她/r+.../v”的搭配得到表征男性、女性和中性的动词，发现人们认知和语言中都强烈表征男性的有与战争和政治有关的动词，将男性和政治进行隐喻关联；强烈表征女性的则有“怀孕”“嫁给”“打扮”“织”“倾诉”“哭”和“抚养”等词，这些词要么是女性独有的行为，要么是被普遍认为属于女性特征和义务的，还将女性和“感性”建立了隐喻关联(朱述承, 2021)。

汉语词汇没有语法上的性范畴，没有屈折形式，难以得到动词的性别分布，但因此也可以认为大部分的汉语动词是相对中性的，并不天然地就和某一性别共现，构建不同性别的路径也更为隐蔽。在此基础上进行的汉语动词与性别关系的研究，更能体现出社会文化通过语言使性别社会化、对各种性别群体所建立的性别规范。本文基于依存句法关系在大规模语料上进行句法分析，并抽取动词的依存结构。与搭配抽取的方法相比，一方面，能够更准确地抽取出彼此有语法关系的动词和性别词词对；另一方面，在一个句子的范围内，能够不受距离限制地抽取(邵艳秋et al.)。经过统计学检验，最终可以得到大规模语料库中全部与性别有关的动词，可以说是一种穷尽的方法，既可以对动词与性别的关系进行全面的定量分析，也可以截取有意义的语句进行深入的定性分析。

3 研究设计

3.1 多语体语料库

本文所使用的语料来源于国家语言资源监测与研究中心平面媒体中心研制开发的DCC动态流通语料库[0]，选取了四个模块的语料：新闻、传统文学、网络小说和微博。其中，新闻语料涵盖了2018年至2019年国内主流媒体报纸和地方报纸的语料，如《人民日报》《法制晚报》《北京晚报》《新疆日报》等；传统文学包含了824本经典的汉语白话文文学作品，如鲁迅的《朝花夕拾》、莫言的《丰乳肥臀》、沈从文的《边城》等；网络小说涉及了2014年至2015年公布

[0] http://cnlr.blcu.edu.cn

的都市、幻想、历史、同人、网游、武侠、悬疑等不同题材的文章；微博语料来源于“新浪微博”，爬取于2019年5月18日至20日。原始语料词数、规模及经过分词、词性标注和依存句法分析后的语料规模如表1所示。由于传统文学包括的是经典白话文文学著作，完书、流传门槛高，规模相对其他语料较小，但其用语典范，艺术水平高，对这部分语料进行分析能体现出近现代文学思想领域对男性、女性的构建。另外，传统文学语料年代跨越久远，在早期，代词“她”的使用仍不稳定，所得到的动词可能比其他语料少。

语料	词数	原始规模	分析后规模
新闻	1,824,536,626	8.27GB	18.8GB
传统文学	105,822,119	419MB	1.22GB
网络小说	2,236,385,748	8.96GB	24.7GB
微博	2,062,183,508	8.37GB	21.32GB

表 1: 所用语料及其词数和处理前后规模

3.2 性别词

汉语词汇中直接表达性别的手段为词汇性别（lexical gender）词和指称性别（referential gender）词，前者包括区别词、亲属称谓词、性别称谓词等有明显性别的词汇，后者主要指性别代词“他”和“她”。我们将这两类可以直接表达性别的词合称为性别词，并使用Li et al. (2022)构建的汉语性别词表作为本文抽取结构中性别词的部分，所使用的性别词如表2所示。

性别	性别词
男性	他，男，男士，男孩，男子，男性，先生，男人，爸爸，父亲，姥爷，儿子，男友，叔叔，哥哥，弟弟，爷爷，外公
女性	她，女，女士，女孩，女子，女性，小姐，女人，妈妈，母亲，姥姥，女儿，女友，阿姨，姐姐，妹妹，奶奶，外婆

表 2: 18对性别词

3.3 研究方法

3.3.1 抽取“性别词-动词”依存结构

如图1所示，我们首先利用哈尔滨工业大学开发的LTP平台提供的开源依存句法工具[1](Che et al., 2021)对四种语体的语料进行分词、词性标注及依存句法分析，然后将能与性别词构成SBV和VOB依存关系的动词（被标注为动词“v”）结构抽取出来，如SBV（女儿，嫁）。

3.3.2 统计检验

抽取出结构后，分别统计各语料库中动词与男性、女性性别词构成SBV、VOB依存结构的频次。考虑到原始语料中性别词与动词构成依存结构的分布可能不平衡，我们需要按照公式$cc = \frac{c}{For M}$进一步消除这一影响。具体来说，在某一语体的语料库中，以动词“怀孕”和女性性别词构成的SBV结构为例，c为“怀孕”和某一女性性别词（如“女士”）构成SBV结构的频次，用c除以能与18个女性性别词构成SBV依存结构的动词总频次F（M则表示能与18个男性性别词构成SBV依存结构的动词总频次），得到了处理后的“怀孕”的频次cc。随后，对处理后的动词频次cc进行假设检验。本文使用独立样本t检验，α水平取0.05（双尾），检验后得到在统计学意义上同男性、女性性别词构成依存结构频次显著不同的动词，在这里将这些动词称为性别表征动词，简称性别动词，其中包括显著表征男性的动词（简称“男性动词”）和显著表征女性的动词（简称“女性动词”），用以刻画动词与性别之间的关联。再根据所构成结构为SBV或VOB结构，将男性动词分为男性支配动词和男性承受动词，将女性动词分为女性支配动词和女性承受动词。

[1] https://github.com/HIT-SCIR/ltp

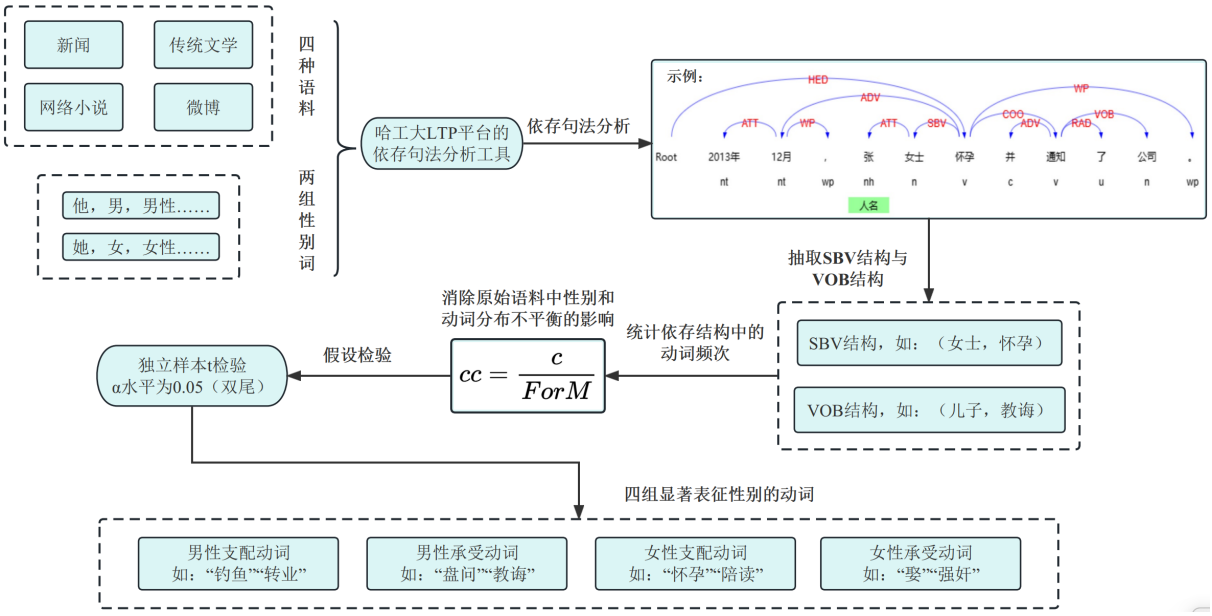


图 1: 抽取显著表征性别的动词的流程

3.3.3 语义分析

进一步，我们从语义角度出发，对性别动词的语义域进行分析，以期揭示汉语动词在表征男性和女性时，在语义层面、传递信息内容方面的差异，由此探讨汉语在描述两种性别群体时所经由的或显性或隐性的不同构建路径，及所呈现的不同性别规范。我们参考《现代汉语分类词典》(苏新春, 2013)，对性别动词进行了语义域上的分类标注。书中共有五个语义层，本文在参考该词典的基础上，将类别简化至两层。一级语义域包括"生活工作""生理活动""社交""司法"等19个类别，对于动词数量多、频次高的"生活工作"和"生理活动"两类进行了细分类，包括"婚恋""就业""生育繁殖""死亡"等21个二级语义域（见附录7）。对于部分有多个词义、被标记为"v"却不是动词的、显然有错别字或分词不合理的词，结合原始语料进行了人工校对，删去了不是动词的、分析错误的，有多个词义的动词归在对应的语义域内。

4 结果与分析

从总体来看，大部分汉语动词性别表征是中性的，能够体现出性别的动词是少数，这说明了汉语作为一种具有深厚文化历史底蕴的语言对不同性别群体表达的尊重，也体现出了我国的性别平等观念。而在能表征性别的动词中，能看到构建男性和女性身份有不同路径。下文的结果分析主要针对所得到的性别表征动词，即在统计学意义上，和男性、女性性别词构成依存结构频次显著不同的动词。

4.1 不同的性别身份构建路径：男性默认-女性专门

本文最终得到220个性别动词，其中男性动词57个，女性动词163个。四种语体中的男性动词、女性动词及其词频前5名如表3和表4所示。

新闻		传统文学		网络小说		微博	
男性	女性	男性	女性	男性	女性	男性	女性
战死 38	嫁 1498	平反 11	出嫁 188	平反 20	嫁 4588	钓鱼 82	嫁 5776
时任 8	怀孕 1142	相助 4	饰 5	便化（化为） 7	撒娇 608	服兵役 21	佩戴 122
参过（军） 8	抚养 436		探监 4	拥戴 6	出嫁 634	行医 18	点击 114
考考 7	织 348			压境 5	打闹 123	防卫 12	怀胎 93
下工 5	产 237			训示 5	私奔 111	出道 9	食 67

表 3: 词频前5名的男性支配动词、女性支配动词及其词频

新闻				传统文学				网络小说				微博			
男性		女性		男性		女性		男性		女性		男性		女性	
骑行	7	娶	451	补	5	还有	495	佩戴	18	娶	5549	闯入	13	身为	240
盘问	5	强奸	79			如	78	送入	6	好像	218	拍打	9	发布	51
教诲	5	演绎	73			携	30	置于	4	不顾	146	装作	5	感到	34
划分	5	描写	50			演	21	相传	4	携	97	敬告	5	残害	18
出席	4	暴打	31			惟有	12			飘	86	借助	5	食	17

表 4: 词频前5名的男性承受动词、女性承受动词及其词频

性别动词的频次和类型可以体现语料对这一性别表征的固化程度和语义范围。总的来说，女性动词在四个语体的语料中均远多于男性动词，这表明语料中对于女性动作行为的固化描述可能更多，即总倾向于使用某些词语表述女性的动作，而对于男性的固化描述则相对较少。从表中可以初步看出，各语料对于女性行为的描述常与“婚嫁”相关，占据女性支配动词最高频次位置的，是动词“嫁”或“出嫁”，微博语料中“嫁”的频次达到了5776次，是所有语料中频次最高的一个性别动词。这说明婚姻关系是语言在构建女性身份时显著重要的部分，在谈及女性时，我们总在关注她们“婚嫁与否”。这体现了我们所说的“男性默认-女性专门”的性别构建路径，即语言中不会特意强调男性的的某些行为动作，在谈论女性的行为动作时却集中在如婚嫁等专门领域。这一点在男性动词和女性动词的语义分布上可以进一步体现出来，如图2和图3所示。

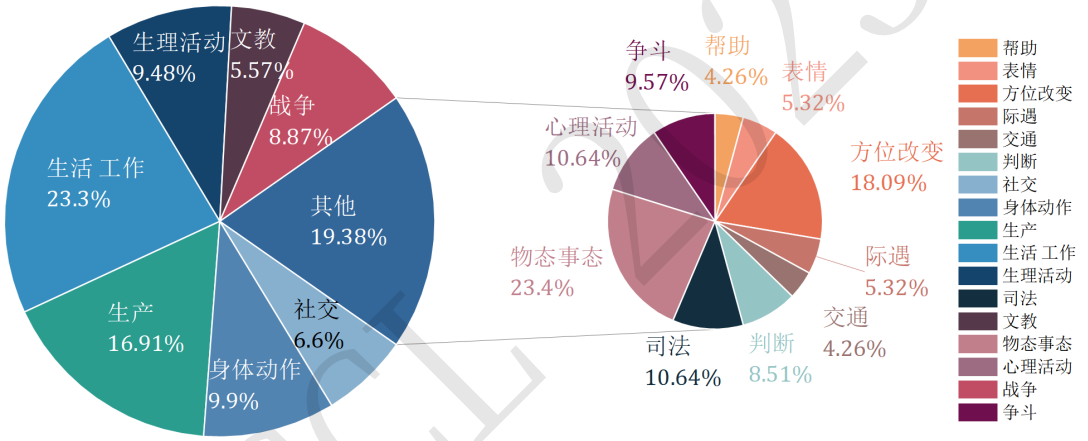


图 2: 男性动词语义域分布

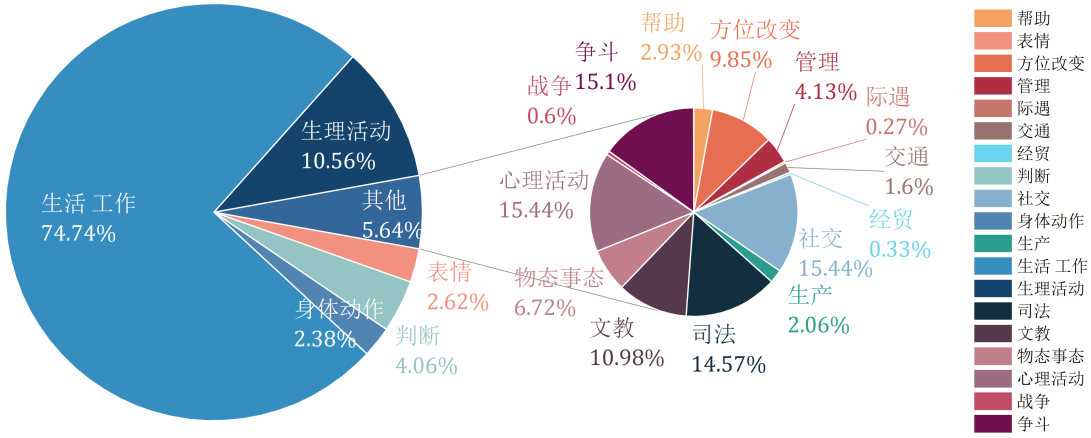


图 3: 女性动词语义域分布

从整体上看，男性动词和女性动词分布的语义域基本一致，占比最多的都是“生活工作”类动词。但男性动词分布更均衡，女性动词则更集中在“生活工作”类。“生产”类在男性动词中占比第二高，在女性动词中却非常低。这都说明语言对于男性动作的刻画更多样，对女性的描绘则集中在某些领域。语言是性别建构过程中的一个中心特征(Butler, 1988)，性别通过话语行为不断重现(Wagner, 2010)，并被社会语境中的不同参与者所接受(Speer and Stokoe, 2011)。正如de Beauvoir (1972)所指出的，女性不是生来就是女性的，而是被社会建构出来的，是社会实践塑造出女性的“第二性（second sex）”。男性动词均匀地分布在各个语义域上，说明男性的行为是社会行为的默认属性，均衡地触及到社会的方方面面；女性动词的分布集中，说明女性的行为动作是较为专门化的。这种构建“男性默认-女性专门”身份的路径是十分隐蔽的。

4.2 权力结构中的支配与承受：男性主宰-女性顺从

经统计，男性支配动词共有38个，女性支配动词共有90个，性别比为0.42；男性承受动词共有19个，女性承受动词共有73个，性别比为0.26，由此可以看出男性有更多充当支配动词主语的趋势，而相对来说，女性则承受了动词所带来的后果。前人研究语法角色和动词类别时也发现，男性往往是主动的实干者，女性更多地是男性行为的被动接受者或旁观者(Macaulay and Brice, 1997)。这也就是我们所总结的“男性主宰-女性顺从”的性别构建路径。

首先来看一类高频的、典型与女性有关的性别动词，其中“婚恋”类动词占比最高，见表5。

最典型的是频次非常高的“嫁”，它在微博语料中作为女性支配动词的频次高达5776次，作为男性支配动词的频次也达到811次，在同男性构成依存结构当中是频次最高的。一般来说，“嫁”被认为是女性才能执行的动作。如果说在统计学意义上，动词“嫁”显著与女性性别词相关时，可以认为语言总是倾向于使用“嫁”来表述女性，那么“嫁”在男性支配动词中占据如此高的频次又意味着什么呢？回归到语料中，我们发现“SBV（他，嫁）+女儿”和“SBV（父亲，嫁）+女儿”是非常重要的结构，如：“他要将女儿嫁给另一个富翁，然而女儿已经爱上了他家的一个书僮，他们相约双双逃走。”“父亲为了挽救家族企业，将年仅十八岁的‘我’嫁给了四十岁的理查德。”在主谓宾结构中，施事与受事之间的关系是主动与被动、控制与被控制的关系，这体现出父亲对女儿行为的支配，后者对父权的抗争主要是“逃婚”“私奔”这种主动逃脱出父权权力结构的方式。“父亲嫁女儿”这种用法在语言中的普遍存在，正是传统父权社会思想在现代汉语中的一种遗留(刘道锋, 2009)。

依存结构	语料来源	语义域	依存结构	例句
SBV结构	新闻	缝织	SBV（母亲，织）	除了日常的洗衣做饭、打扫卫生，母亲还织毛衣、缝衣服。
			SBV（母亲，缝制）	王世卿和大家回家探望时，母亲正在给他缝制鞋垫，闲不住的父亲正在给黄瓜浇水。
	传统文学	婚恋	SBV（女儿，出嫁）	老陈家的两个女儿都出嫁到外地，剩下老陈和媳妇，院子一年四季冷冷清清……
			SBV（她，出嫁）	桃花开的时候她就出嫁了。
	网络小说	婚恋	SBV（女儿，私奔）	白绫儿快要哭出来的说道：“你如果再不说的话，女儿就跟建邦哥私奔。”
			SBV（她，逃婚）	自从她逃婚后她的父亲维克托公爵一直派人在找她。
		表情	SBV（女孩，暗送）	一路上还有同龄的女孩暗送秋波，让庄子诺红了脸，不由得低下头加快脚步。
	微博	婚恋	SBV（女人，嫁）	女人嫁没嫁对人，看这个部位就知道
			SBV（姐姐，择偶）	结果是，姐姐择偶时更容易“在垃圾堆里找男人”，而妹妹择偶的情形要好些。
VOB结构	新闻	婚恋	VOB（她，娶）	看着她变得丑陋的双手，我决定娶她，我对她说，以后，你的手要好好保养。
	网络小说	婚恋	VOB（女儿，娶）	人刚死没到两天，冯胜就强娶了人家女儿。
			VOB（她，娶）	况且还是女皇，娶了她，就等于当了大汗。

表 5: 典型的女性动词结构及其例句

随着我国促进男女平等与妇女全面发展的社会环境进一步优化，性别平等观念增强，社会文化更加包容，“半边天”力量进一步彰显。一方面，女性能够广泛参与经济社会发展，就业领域进一步拓展，在业比例保持较高水平，18-64岁在业者中，女性占43.5%，男性占56.5%。另一方面，女性也为家庭建设做出重要贡献。调查显示，在业女性工作日平均总劳动时间为649分钟，其中有酬劳动时间为495分钟；照料家庭成员和做饭/清洁/日常采购等家务劳动时间为154分钟，约为男性的2倍，即女性承担家庭照料主要责任。尽管近十年来夫妻家庭地位更加平等，同时也应看到女性家庭照料负担重、公共服务支持不足的现象(第四期中国妇女社会地位调查领导小组办公室, 2021)。本文对语料的分析也印证了现实调查中的结果，见表6。女性和家庭领域的关联更为密切，“生育繁殖”类动词共有23个，全为女性动词，这里不仅包括“怀孕”“分娩”等只能由女性支配的动作，更包括其他如“伺候”“持家”“照料”等不局限于性别的词。“生育繁殖”类动词与女性高度相关，体现出语言对于女性角色的动作描述集中于家庭空间，对男性的角色则没有这样的限制。如，“就业”类动词共7个，其中6个为男性动词，男性和

职业领域的关联更为紧密。这些动词类别的频次虽然不及前述动词高，却也表征出较为单一的性别。婚恋、家庭是涉及两性的社会概念，但这类动词呈现出单一性别的特征，说明语言所建构的女性的存在主要依赖于家庭，这又是以男性的存在为前提的。对于男性的构建则并不如此，这就建立起了一种女性依附男性的规范："男性主宰-女性顺从"。

语义域	男性动词个数	女性动词个数	依存结构	例句
就业	6	1	SBV（父亲，行医）	祖父是太守，父亲行医，"（"）十六七岁曾随父宦历楚粤，出塞省视。……"
			SBV（他，时任）	他时任国家经贸委副主任，此前1年主掌二汽……
生育繁殖	0	23	SBV（她，怀孕）	如今，她怀孕了，周俐安排她只坐门诊，其他事情都不让她参与。
			SBV（女人，持家）	患难夫妻躲避到乡下，男人苦读女人持家……
			SBV（女人，生养）	现在男的真会呀，孩子女人生养，自己爹妈还让女人照顾，那么的力气的大老爷们真不要个脸了……
伤残	0	1	SBV（女性，大出血）	此外，女性产后大出血也可利用介入治疗，替代传统子宫切除术。
美容打扮	0	5	SBV（姐姐，瘦身）	蓉姐姐火了，瘦身成功，成了正面楷模
			SBV（女友，卸妆）	看到女友卸妆后，是什么感受？男生第一次见到女生卸妆时的心里（理）活动。
司法	2	8	SBV（男子，追捕）	广西一偷狗男子，被村民追捕持刀抢车逃跑。
			VOB（女友，劫持）	吉恩阿特金斯现年28岁，21日朝外祖母连开数枪，劫持女友，驾车逃逸。
			VOB（女孩，强奸）	嫌犯承认自己强奸和杀死了女孩。
丧葬	1	0	SBV（他，祭祖）	据计某供诉，事发当天他跟家人清明祭祖后，喝了点酒。
造福补救	2	0	SBV（父亲，平反）	去岁他父亲平反，富阳知县才照顾他进县衙，当上了书吏，这才解决了生计问题。
文教	5	10	VOB（妈妈，演出）	在剧中我还要演出年轻妈妈对于孩子又爱又纠结的复杂情感，这些对我来说都是挑战。
			SBV（他，练字）	这女生昨天顺嘴说了一句我朋友字写的丑，他就立刻开始练字，游戏也不打了……
战争	4	1	SBV（他，服兵役）	历经波折找到同案中另一位当事人家属，得知他是一位退伍军人，服兵役十几年……

表 6: 特殊语义域的动词结构及其例句

"美容打扮"类动词只同女性相关，包括"瘦身""显瘦""卸妆"等，这与一般的认知一致，人们普遍更关注女性的外表，外表是女性受到称赞的主要主题，长相和可爱对于女性来说非常重要(Eckert and McConnell-Ginet, 2003)。赞美一方面表达的是欣赏与喜爱，另一方面也是构建和规范性别秩序的手段。通过对男性、女性的不同称赞能构建起不同的社会约束，对男孩赞美其勇敢、坚韧，对女孩赞美其美丽、善良，这背后是对胆小的男孩、不美丽的女孩的忽视甚至是束缚。同样，"文教"类动词也体现出男性和女性所承担的不同的社会期望，这一类的男性动词主要是"练字""训示"等与学习教育相关的动词，女性动词则主要是"舞动""演出"等和表演相关的动词，女性仍被期待为一个被欣赏的形象，而男性则是可教育的。

如果说男性作为支配者，总是出现在主宰女性婚姻的"婚嫁"类动词的主语位置上，那么同样是影响女性的"司法"类动词中，作为施暴者的男性却"隐身"了。表达受害含义的动词均为女性承受动词，如"劫持""强奸""拐走"等，女性更常成为暴力恶性事件的受害者，但是在SBV结构中却没有施暴的性别动词。前人研究也发现(Clark, 1992)，英国报纸上关于男性暴力侵害妇女的报道中很难找到男性施暴者，"女性被强奸"的这种无主语被动句被用来转移对男性压迫女性的注意，如果再加上对女性着装暴露的描述，甚至会把强奸的责任从男性强奸犯转移到女性受害者身上(Penelope, 1990)，这都体现父权制社会下，无主语的语言表达对男性施暴者的维护，对女性受害者精神指责的伤害。这种构建"男性主宰-女性顺从"的手段就十分微妙了，一方面，在正面或中性情况下的主宰动作中，如"婚嫁"，男性可以堂而皇之地出现，但在那些负面侵害的主宰行为中，男性却又"神隐"了。另外，军人一般被视为更适合男性的职业(朱述承，苏祺，刘鹏远, 2021)，与战争有关的政治活动也被视为"纯粹的"男性活动(Wilson, 1992)，"战争"类大多是男性动词也体现出这一点。

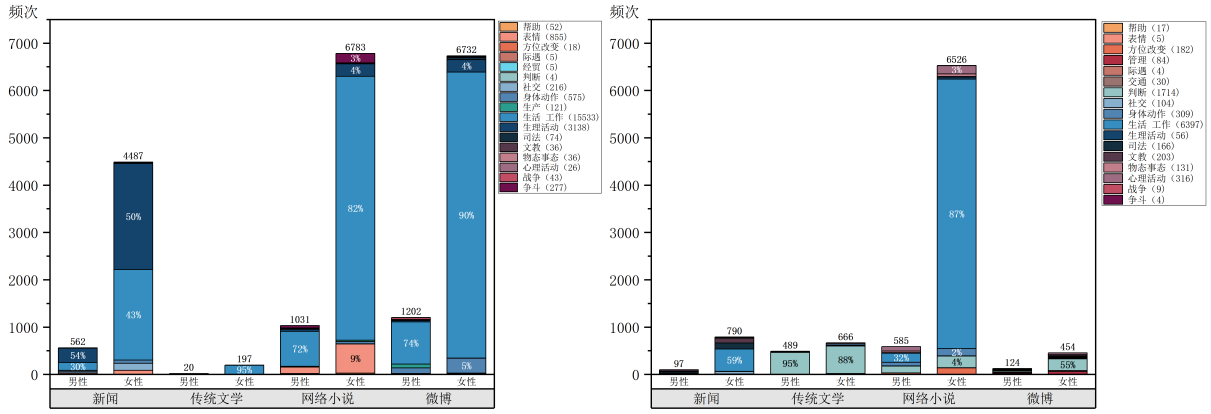
4.3 性别构建上的语体差异

语体是语言为适应不同的交际需要而形成的具有不同风格特点的表达形式，通常分为口语语体和书面语体(中国社会科学院语言研究所词典编辑室, 2016)。综合考察各语体中动词对性别的表征，能够更全面地研究社会对性别的构建。

我们首先依据SBV和VOB两种依存结构，按照四种语料对性别动词同男性、女性性别词构成结构的频次进行分类和统计，如图4所示。

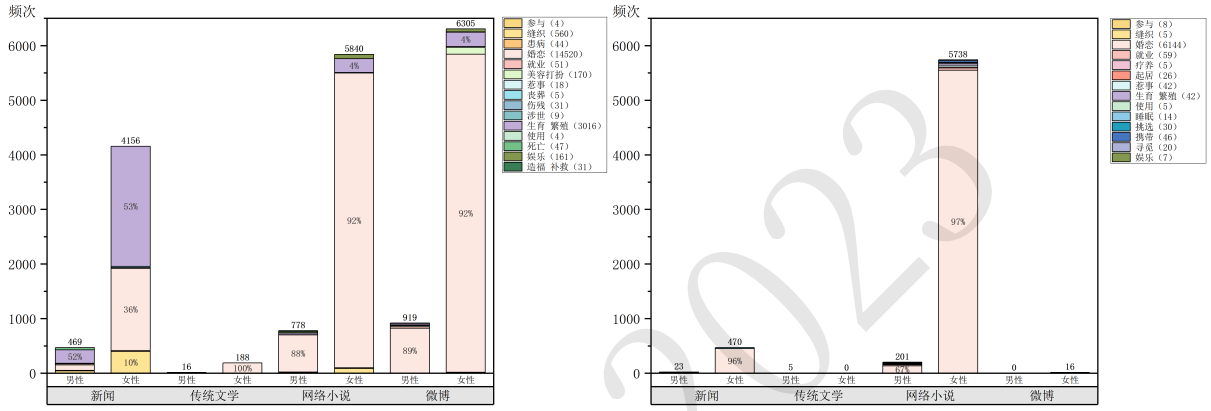
总的来看，各语料对两性动作行为的描述集中在"生活工作"和"生理活动"领域；同一语料中，动词同女性构成结构频次高于同男性的，由于女性动词比男性动词多，这是较为合理的结果。在网络小说中，女性和"表情"类动词"撒娇"构成结构达608次，可见网络小说较多描绘女性娇媚、可爱的一面，遵循传统的性别规范，将女性塑造为被欣赏的对象。

接着，我们对占比最高的"生活工作"和"生理活动"两类进行了细分类，以进行更细致的探索，如图5所示。"婚恋"类动词在更口语化的网络小说和微博语料的SBV结构中占据90%以上，这表明婚恋择偶问题可能是现今人们最关注的生活类问题。但在新闻语料SBV结构中，"生育繁殖"类动词同男性和女性群体搭配是最多的，达到50%以上，虽然对于男性和女性最关注的都



(a) 各语体中支配动词的一级语义域分布　　(b) 各语体中承受动词的一级语义域分布

图 4: 各语体中的动词一级语义域分布



(a) 各语体中支配动词的二级语义域分布　　(b) 各语体中承受动词的二级语义域分布

图 5: 各语体中“生活工作”与“生理活动”类动词的二级语义域分布

是生育，但对于女性还会关注“婚恋”及为他人“缝织”的动作，将女性的社会活动局限在家庭之内，将其形象构建为维持家庭秩序的“贤妻良母”(Eckert and McConnell-Ginet, 2003)，典型用法见表5。媒体话语对社会观念具有引领作用，对男性和女性的差异化表征，提供了不同性别角色的行为规范，且由于媒体倾向于抓人眼球，所用的较固化的言辞又进一步塑造性别秩序。传统文学的性别动词远少于其他语料，一方面语料较少会造成一定影响，另一方面，写就文学作品更需要打破常规，减少固化描写，但女性群体结构中的动词100%都是“婚恋”类，这更凸显出语言对于女性婚恋的关注，对于女儿的担忧和希望都系于一个“嫁”字，见表5。网络小说语料对于女性婚嫁的描绘同其他语料有所不同，高频出现的不仅包括“出嫁”，更包括“私奔”这种表示反抗的动词，网络小说有其表达固化的一面，仍然将女性的形象置于家庭之内，或将女性塑造为被欣赏的主体，但已经描绘出了更加勇敢自由的女性形象。

这两类动词的VOB结构较少，其中值得关注的是传统文学中的男性动词“补”，如“没有到阿贵嫂不得已要补男人白天等用的裤子的时候啦！”，这与“缝织”类女性支配动词形成对应。女性的家庭任务是由他人提出的持续需求而来的(Eckert and McConnell-Ginet, 2003)，在家庭中是满足他人需要、照料他人的角色。目前，照料研究领域的研究者们已经打破了关于照料为“私领域”的认知，指出照料跨越了“公”与“私”、“有偿”与“无偿”，结合了“市场”“国家”与“家庭”，其理论发展吸纳了马克思主义女性主义的“社会再生产”理论资源，将照料视为一种劳动，强调照料的社会价值及其与生产体系的关系(肖索未，简逸伦, 2020)。过去几十年中，伴随市场经济转型，中国的经济高速增长，而承担照料孩子与赡养老人的责任往往会减少女性就业机会(Pei et al., 2017)。在今天，如果语言对两性所活动的范围、所承担的家庭和社会责任能够更加包容，能有利于形成更平等和谐的家庭气氛，也更利于女性在职业领域发挥“半边天”的力量。

第二十二届中国计算语言学大会论文集，第300页-第314页，哈尔滨，中国，2023年8月3日至5日。

4.4 对两个特殊动词的定性分析

男性动词“考考”是一个有趣的结果：“叔叔今天考考你，1+1等于几啊。”“爷爷会时不时地考考她.....‘去,剪个侧脸出来。’”如果家庭是社会分配给女性的空间，生儿育女是女性的职责，那为什么这种对子女的小测试不是由女性提出的呢？虽然女性的职能被分配给家庭或私人领域，男性更多被分配在公共领域，但在家庭当中，女性仍不是主宰。学者研究美国家庭餐桌的谈话发现，母亲总是鼓励孩子给父亲讲述自己的一天，而父亲则负责判断评估孩子的行为和感情的正确性，在家庭权力结构中处于质疑他人行为的地位(Ochs and Taylor, 1992)。如前文所述，女性的任务“怀孕”“生养”“缝补”都是由他人的需求而来的(Eckert and McConnell-Ginet, 2003)，女性处于一种满足、回应他人需求的地位，而男性则是处于凝视、审判的主位。

女性动词“求测”也值得探究，这并不是一个收录于词典的规范的动词：“一女士拿来一丁未年男八字求测，方舟周易告诉她.....但婚姻不顺.....”“网友丁小姐求测事业，婚姻。”“求测”主要是指请求大师看八字、算卦看相等行为。语言将女性同超自然力量的连结在对俄语动词和性别的研究中也有呈现，结果中有两个与巫术有关的动词同女性相关(Kuznetsova, 2015)。女性在社会结构中是相对男性的弱者，缺少公共权力，那就更需要去寻求强大力量的庇佑，女性可以通过道德权威的建设来发展影响力(Eckert and McConnell-Ginet, 2003)，也可以寻求神灵的庇护，“狠毒的阴谋和巫术”(布尔迪厄, 2017)都可以是女性的武器。但求测的内容仍然是同婚姻相关，或许在性别越来越平等的当今社会，“嫁不出去”“嫁不到好男人”这些话题能够逐渐淡出女性的生活，女性能够在更广阔的天地中选择自己的生活方式。

5 不足与展望

本文采用Li et al. (2022)的汉语性别词表作为依存结构中性别词的部分，但这仍不能涵盖所有能表征人性别的名词，我们希望未来能补充这部分不足，具体来说：一，传统文学语料时间跨度大，代词“她”的使用仍不稳定，还需考虑其他代词和名词的性别表征，如“伊”“他”“她们”等；二，命名实体识别部分，需要构建姓名-性别数据库以及文学作品人物姓名-性别数据库，从而更准确地识别文本中的人物性别；三，共指消解部分，需要进一步研究如何准确地识别和消解文本中的共指关系。这些补充措施可以进一步提高性别表征动词的抽取准确率、数量和频次，从而更好地理解和研究文本中动词对性别的构建。

6 结语

语言在塑造文化和社会态度方面具有关键作用。汉语是几乎没有形态变化的语言，大部分动词表征性别更为隐性，因此研究汉语的性别表征动词有重要的文化和社会意义。研究结果表明，总体来看，大部分汉语动词对性别的表征是中性的，能够体现出性别的动词属于少数，这说明了汉语作为一种具有深厚文化历史底蕴的语言对不同性别群体的尊重，也体现出中华民族对于性别平等观念的尊重。而在能表征性别的动词中，发现了构建男性和女性身份的不同路径。女性动词多于男性动词，语言中对于女性动作行为的固化描述更多；语言对于两性群体最关注的行为都与“婚恋”相关，但对于女性的关注更多；男性动词的语义类别主要有“就业”“丧葬”“造福补救”和“战争”，女性的则有“生育繁殖”和“美容打扮”；女性作受害动词的宾语时，施事男性往往被隐去；动词将女性的活动范围局限在家庭，男性的则更为广阔。这是语言对于两性群体建立的不同性别规范，即“男性默认-女性专门”和“男性主宰-女性顺从”。同时，不同语体体现了不同的功能，新闻塑造了较为传统的性别规范，传统和网络文学以不同的形式打破了固有的性别规范。在以后的研究中，我们希望能够从更多角度对语言中动词和性别的关系进行更全面、合理的刻画，我们也认为在越来越平等包容的现代社会中，各性别群体都能够按照自己的能力和意愿选择适合自己的生活。

7 致谢

感谢评审专家对本文的认真审阅和指导，感谢他们提出的中肯意见和建议，这些意见和建议极大地提高了本文的立意和质量，使得本文的研究成果更加准确和可靠，也进一步确立了以社会主义核心价值观和全人类共同价值为立足点的研究范式，这将对未来的研究和实践产生积极的影响。

References

Jo Angouri and Judith Baxter. *The Routledge Handbook of Language, Gender and Sexuality*. London: Routledge, 2021.

Paul Baker. *Using Corpora to Analyze Gender*. A&C Black, 2014.

S. Brownell and J. Wasserstrom. *Chinese Femininities, Chinese Masculinities: A Reader*. University of California Press, Berkeley, 2002.

Judith Butler. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre Journal*, 40:73–83, 1988.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. N-LTP: An open-source neural language technology platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.6. URL https://aclanthology.org/2021.emnlp-demo.6.

Kate Clark. The linguistics of blame: representations of women in the sun's reporting of crimes of sexual violence. In Michael Toolan, editor, *Language, Text, and Context: Essays in Stylistics*, pages 208–226. Routledge, 1992.

Charles H. Cooley. *On Self and Social Organization*. University of Chicago Press, 1902.

Simone de Beauvoir. *The Second Sex*. New York: Routledge, 1972.

Penelope Eckert and Sally McConnell-Ginet. *Language and Gender*. Cambridge: Cambridge University Press, 2003.

H. Evans. *Women and Sexuality in China: Female Sexuality and Gender Since 1949*. Polity Press, Cambridge, 1997.

Amaç Herdağdelen and Marco Baroni. Stereotypical gender actions can be extracted from web text. *Journal of the American Society for Information Science and Technology*, 62(9):1741–1749, 2011.

Julia Kuznetsova. Are verbs politically correct? a corpus study of gender in russian verbs. *Gender & Language*, 9(3), 2015.

Jiali Li, Shucheng Zhu, Ying Liu, and Pengyuan Liu. Analysis of gender bias in social perception and judgement using chinese word embeddings. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 8–16, 2022.

Monica Macaulay and Colleen Brice. Don't touch my projectile: Gender bias and stereotyping in syntactic examples. *Language*, pages 798–825, 1997.

Elinor Ochs and Carolyn Taylor. Family narrative as political activity. *Discourse & Society*, 3 (3):301–340, 1992.

X. Pei, H. Luo, Z. Lin, N. Keating, and J. Fast. The impact of eldercare on adult children's health and employment in transitional china. *Journal of Cross-Cultural Gerontology*, 32(3): 305–320, 2017.

Julia Penelope. *Speaking Freely: Unlearning the Lies of the Fathers' Tongues*. Pergamon Press, New York, 1990.

Susan A Speer and Elizabeth Stokoe. *Conversation and Gender*. Cambridge University Press, 2011.

Qi Su, Pengyuan Liu, Wei Wei, Shucheng Zhu, and Chu-Ren Huang. Occupational gender segregation and gendered language in a language without gender: trends, variations, implications for social development in china. *Humanities and Social Sciences Communications*, 8(1):133, 2021.

第二十二届中国计算语言学大会论文集，第300页-第314页，哈尔滨，中国，2023年8月3日至5日。

Charlotte Taylor. Searching for similarity using corpus-assisted discourse studies. *Corpora*, 8 (1):81–113, 2013.

Sarah Wagner. Bringing sexuality to the table: language, gender and power in seven lesbian families. *Gender & Language*, 4(1), 2010.

Fiona Wilson. Language, technology, gender, and power. *Human Relations*, 45(9):892–898, 1992.

中国社会科学院语言研究所词典编辑室. 现代汉语词典（第7版）. 商务印书馆, 2016.

刘道锋. 《史记》嫁娶类动词的句法考察及其所反映出的性别等级. 现代语文*(*语言研究版*)*, 2009.

卢淑樱. 母乳与牛奶*:* 近代中国母亲角色的重塑*(1895—1937)*. 华东师范大学出版社, 上海, 2020.

皮埃尔·布尔迪厄. 男性统治. 中国人民大学出版社, 北京, 2017.

张滢，曹榛. 从转述动词看新闻报道中的性别偏见——以《中国日报》连续五天的叙事类新闻报道为例. 长春工业大学学报*(*社会科学版*)*, 25(138-140), 2013. ISSN 1674-1374.

徐修倜. 基于微博语料库的“怨恨”类心理动词性别差异研究. 硕士, 华中师范大学, 2017.

朱述承. 基于平面媒体语料库的汉语性别表征研究. 硕士, 北京语言大学, 2021.

朱述承，苏祺，刘鹏远. 基于语料库的我国职业性别无意识偏见共时历时研究. 中文信息学报, 2021.

段新焕. 汉语动作动词的性别编码及对认知的影响. 硕士, 华南师范大学, 2007.

第四期中国妇女社会地位调查领导小组办公室. 第四期中国妇女社会地位调查主要数据情况. 中国妇女报, 12 2021.

肖索未，简逸伦. 照料劳动与社会不平等: 女性主义研究及其启示. 妇女研究论丛, 2020(5):1–11, 2020.

苏新春. 现代汉语分类词典. 商务印书馆, 2013.

詹秀红. 基于博客语料的“喜爱”类心理动词性别差异研究. 硕士, 华中师范大学, 2017.

邵艳秋, 申资卓, and 刘世军. 基于依存搭配抽取技术的平面媒体语言监测研究. doi: 10.13451/j.cnki.shanxi.univ(nat.sci.).2019.03.16.002.

高晓君and 魏伟. 女人当家?——单身生育和性别角色的重新协商. 妇女研究论丛, No.171 (103-113), 2022. ISSN 1004-2563.

附录A.关于依存结构的取舍

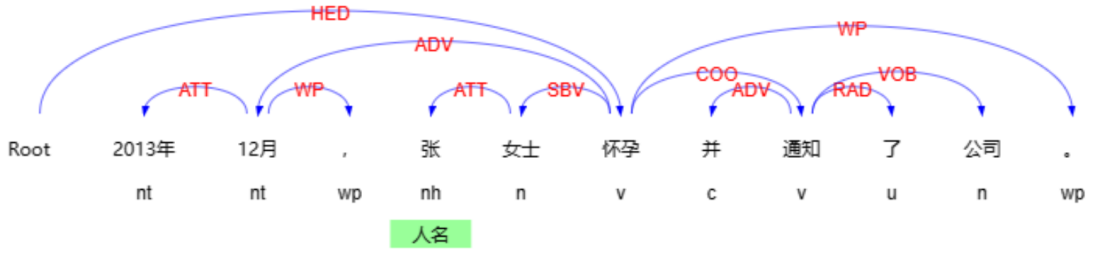


图 6: 依存图分析示例

LTP平台所提供的开源依存句法工具会将每个句子中具有依存句法关系的两个词语用依存弧连接并标出相应的关系[2](Che et al., 2021)，如图6所示，所包含的依存句法关系类型见表7。由于我们分析的是动词与性别的关系，在所抽取的结构中，除了动词以外的另一个成分需要

[2]https://github.com/HIT-SCIR/ltp

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我送她一束花(我← 送)
动宾关系	VOB	直接宾语，verb-object	我送她一束花(送→ 花)
间宾关系	IOB	间接宾语，indirect-object	我送她一束花(送→ 她)
前置宾语	FOB	前置宾语，fronting-object	他什么书都读(书← 读)
兼语	DBL	double	他请我吃饭(请→ 我)
定中关系	ATT	attribute	红苹果(红← 苹果)
状中结构	ADV	adverbial	非常美丽(非常← 美丽)
动补结构	CMP	complement	做完了作业(做→ 完)
并列关系	COO	coordinate	大山和大海(大山→ 大海)
介宾关系	POB	preposition-object	在贸易区内(在→ 内)
左附加关系	LAD	left adjunct	大山和大海(和← 大海)
右附加关系	RAD	right adjunct	孩子们(孩子→ 们)
独立结构	IS	independent structure	两个单句在结构上彼此独立
核心关系	HED	head	指整个句子的核心

表 7: LTP平台所提供的依存句法关系类型

能够体现人的性别，在这里我们使用了(Li et al., 2022)的性别词表作为能够高效抽取出性别词的部分。在LTP所抽取的各个依存结构中，主谓结构（在工具中被标注为“SBV”）和动宾结构（被标注为“VOB”）能够最直接地体现出性别词与动词之间的关系。前者可以体现出性别群体与动作的支配关系，后者则可以体现出性别群体与动作的承受关系。抓住这两个结构，我们可以直观地看出性别词是动词的发出者还是承受者，抑或是有别的特殊关系。所以在本文中我们着重关注能与性别词构成这两种语法关系的动词，以突出语料对不同性别群体的刻画。扩充性别词以及对其他关系类型深入分析的工作，需要在未来的研究中进一步完成。

附录B.动词及其语义域类别

对于性别表征动词语义上的分析，本文参考苏新春 (2013)的体系进行了分类。词典共收词条8.3万多个，主要是以科学的现代化的语料库为基础选择的通用程度较高的语文性词语。8.3万多个词条按五级语义层的分类体系编排，共有一级类9个，二级类62个，三级类508个，四级类2057个，五级类12659个，上层分类反映了整个社会生活与汉语词汇的宏阔概貌，底层分类将同义、近义、反义词语汇聚在一起，细致地反映出词语的同义、近义、反义关系。词典建构了一个词量庞大、覆盖面广、层次清楚、分类严密的词汇分类体系，可以说充分反映了现代汉语词汇的面貌，所以本文参考这一词典，根据所得到的性别动词的情况进行了简化，最终得到了19个一级语义域，语义域类别及所包含的动词见表8和表9。由于一级语义域中的“生活工作”和“生理活动”类动词数量多、频次高，我们对他们进行了细分类以进行更细致的分析，见表10和表11。表中动词后的数字表示该动词的来源语料，1-4分别表示新闻语料、传统文学、网络小说和微博语料。需要说明的是，词典中“十、帮助-B取予”类下有“b交送”和“c领取”两小类，故本文将“取出”与“交出”归在“帮助”类动词中，见表9。

第二十二届中国计算语言学大会论文集，第300页-第314页，哈尔滨，中国，2023年8月3日至5日。

一级	男性支配动词	男性承受动词
帮助	相助2	
表情		装作4
方位改变	撞击4	闯入4
管理		
际遇	升官1	
交通		卸载4
经贸		
判断	实属4	置于3
社交	受邀3，致谢4	盘问1，送入3，借助4，敬告4
身体动作	抽根烟3，修剪4，坐坐4	骑行1，佩戴3，拍打4
生产	钓鱼4	
司法	追捕4	暴击4
文教	考考1，训示3，练字4	教诲1，反串4
物态事态	便化（化为）3，变色4，散播4	相传3
心理活动	拥戴3	着迷4
战争	压境3，防卫4，服兵役4，缴械4	
争斗	开不起3，评理3	

表 8: 一级语义域分类及其男性动词（不含“生活工作”和“生理活动”类）

一级	女性支配动词	女性承受动词
帮助	帮腔3	取出2，交出2
表情	哭喊1，撒娇3，暗送3	
方位改变	凑合4	飘3，邂逅3
管理		度假4，发布4，建设4
际遇		堕落3
交通		装满1，重返4，返4
经贸	扫货4	
判断		引起1，还有2，如2，惟有2，好像3，疑似3，身为4，维持4，致使4
社交	哭诉1，陪读1，探监2，探亲3，交际4，求测4	描写1，讲讲2，示意2，挥别3
身体动作	尖叫1，轻解3，入浴3，洗衣3，点击4，佩戴4，食4，跳跳4	反锁1，携2，携3，发抖3，沐浴3，倾3，食4
生产	织布3，锄草3	
司法	卖淫1，充入3，卖淫4	暴打1，劫持1，强奸1，拐走3，残害4
文教	饰2，实拍4	演出1，演绎1，按摩2，演2，舞动3，培训4，饰4，吟唱4
物态事态	未遂3	做到1，转为1，变幻3，落成3，外加3，进行4
心理活动	当心1，闲置4	寻求1，不顾3，胆敢3，惊呆3，唯恐3，感到4，感知4
战争		占据1
争斗	斗嘴1，打闹3，投怀送抱3，附合（和）3，喧闹3，结伴4	谎称3

表 9: 一级语义域分类及其女性动词（不含“生活工作”和“生理活动”类）

第二十二届中国计算语言学大会论文集，第300页-第314页，哈尔滨，中国，2023年8月3日至5日。

一级	二级	男性支配动词	男性承受动词
生理活动	患病		
	伤残		
	生育繁殖		
	睡眠		
	死亡	毙命1，战死1，下世3	
生育繁殖	参与		出席1
	缝织		补2
	婚恋	配对3	
	就业	参过（军）1，时任1，下工1，行医4，务工4，专业4	
	疗养		
	美容打扮		
	起居		
	惹事		
	丧葬	祭祖1	
	涉世	出道4	
	使用		划分1
	挑选		
	携带		
	寻觅		
	娱乐		
	造福补救	平反2，平反3	

表 10: “生活工作”和“生理活动”类的二级语义域分类及其男性动词

一级	二级	女性支配动词	女性承受动词
生理活动	患病	便秘1，抱病3，发痴3	
	伤残	大出血1	
	生育繁殖	产1，持家1，诞下1，分娩1，抚养1，怀胎1，怀孕1， 流产1，哺乳3，持家3，诞下3，诞生3，分娩3，怀胎3， 临产3，陪护3，生养3，早产3，伺候4，分娩4，怀胎4，受孕4，生养4	诞生3
	睡眠		沉睡3
	死亡		
生育繁殖	参与	练手4	下手4
	缝织	织1，缝补1，织3，缝制4	
	婚恋	嫁1，悔婚1，出嫁2，出阁3，出嫁3，嫁3，私奔3，择偶3， 嫁4，逃婚4，择偶4	娶1，娶3
	就业		兼3
	疗养		调理4
	美容打扮	美容1，美容3，瘦身4，显瘦4，卸妆4	
	起居		整理3
	惹事	晒晒4，秀出4	染指3
	丧葬		
	涉世		
	使用	识货4	
	挑选		选出1，看上3
	携带		带有3
	寻觅		觅3
	娱乐	嬉闹3，畅游4，出游4，玩乐4	玩耍4
	造福补救		

表 11: “生活工作”和“生理活动”类的二级语义域分类及其女性动词

基于多任务多模态交互学习的情感分类方法

薛鹏[1]，李旸[2]，王素格[1,3]，廖健[1]，郑建兴[1]，符玉杰[1]，李德玉[1,3]
1.山西大学计算机与信息技术学院，山西省 太原 030006
2.山西财经大学金融学院，山西省 太原 030006
3.山西大学计算智能与中文信息处理教育部重点实验室，山西省 太原 030006
wsg@sxu.edu.cn

摘要

随着社交媒体的快速发展，多模态数据呈爆炸性增长，如何从多模态数据中挖掘和理解情感信息，已经成为一个较为热门的研究方向。而现有的基于文本、视频和音频的多模态情感分析方法往往将不同模态的高级特征与低级特征进行融合，忽视了不同模态特征层次之间的差异。因此，本文采用以文本模态为中心，音频模态和视频模态为补充的方式，提出了多任务多模态交互学习的自监督动态融合模型。通过多层的结构，构建了单模态特征表示与两两模态特征的层次融合表示，使模型将不同层次的特征进行融合，并设计了从低级特征渐变到高级特征的融合策略。为了进一步加强多模态特征融合，使用了分布相似性损失函数和异质损失函数，用于学习模态的共性表征和特性表征。在此基础上，利用多任务学习，获得模态的一致性及差异性特征。通过在CMU-MOSI和CMU-MOSEI数据集上分别实验，实验结果表明本文模型的情感分类性能优于基线模型。

关键词： 多模态融合 ； 多任务学习 ； 情感分析

Sentiment classification method based on multitasking and multimodal interactive learning

Peng Xue[1], Yang Li[2], Suge Wang[1,3], Jian Liao[1], Jianxing Zheng[1], Yujie Fu[1], Deyu Li[1,3]
1.School of Computer and Information Technology, Shanxi University,Shanxi 030006
2.Shanxi University of Finance and Economics,Shanxi 030006
3.Key Laboratory Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University,Shanxi 030006
wsg@sxu.edu.cn

Abstract

With the rapid development of social media, multimodal data has shown explosive growth. How to mine and understand emotional information from multimodal data has become a popular research direction. However, existing multimodal sentiment analysis methods based on text, video, and audio often fuse high-level and low-level features of different modalities, ignoring the differences between different levels of modal features. Therefore, this article proposes a self supervised dynamic fusion model for multitasking and multimodal interactive learning, centered around text modality and supplemented by audio and video modalities. Through a multi-layer structure, a hierarchical fusion representation of single modal feature representation and pairwise modal features was constructed, enabling the model to fuse features from different levels. A fusion strategy was designed to gradually transition from low-level features to high-level features. In order to further strengthen multimodal feature fusion, distributed similarity Loss function and heterogeneous Loss function are used to learn common and specific

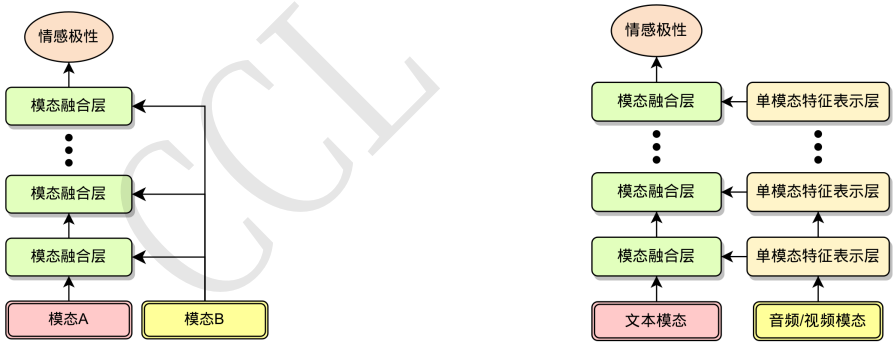
第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。

representations of modes. On this basis, multi task learning is utilized to obtain the consistency and difference features of modalities. Through separate experiments on the CMU-MOSI and CMU-MOSEI datasets, the experimental results show that the sentiment classification performance of our model is superior to the baseline model.

Keywords: Multimodal fusion , Multi-task learning , Sentiment analysis

1 引言

随着社交媒体的迅速发展，以及配备高质量摄像头的智能手机的普及，使得多模态数据（如电影、短视频等）呈爆炸式增长。多模态数据通常是由文本、视觉（视频）和声学（语音）组成。在对话系统和虚拟现实等应用领域中，如何让机器在不忽视非语言因素的情况下，能够理解多感官信息，成为保持高质量的用户交互的关键。多模态情感分析(multimodal sentiment analysis, MSA)旨在从音频、视觉和语言特征中预测情感的得分 (Soleymani et al., 2017)。例如，通过产品数据的情感得分，可以获得客户对整体产品反馈信息，通过对选民的评论数据的情感分析，可以预测潜在选民对投票的意图。对于同一时间段的不同模态数据，情感的表达通常是相互补充的，因此，对多模态数据的情感分析，可对语义和情感消歧提供技术支持。MSA的关键问题是如何将多模态数据进行融合，即对所有输入的模态数据，如何提取和整合它们的信息，用于深入理解数据的情感。目前，许多研究者已经提出了利用多模态信息进行情感分析的方法 (Zellinger et al., 2017; Liu et al., 2017; Ruder and Plank, 2018)。其中，跨模态注意力机制是使用较多的多模态融合方法，它可以通过建模不同模态之间的关系，从而强化其中的某一模态。然而，已有的研究工作较少关注将模态的高级特征的融合问题。如图1(a)所示，已有的工作通常是将一个模态的高级特征与另外一个模态的低级特征进行融合，体现了模态融合过程中的不一致性。通过这种方式的模态融合，无法获得不同模态之间的最佳映射，降低了模型的情感分析性能。因此，本文提出了多任务多模态交互学习的自监督动态融合模型(Self-supervised dynamic fusion model for multi-task and multi-modal interactive learning，MMILN)，如图1(b)所示，该模型可在模态的低级特征向高级特征转化时，从多个层次充分融合，同时使用以文本模态为中心，其余模态为辅助的方式，使模型对多个模态的情感信息进行相互关联，并有效表示。



(a)已有工作　　(b)本文工作

Figure 1: 已有工作与本文模态融合图示对比

本文的主要贡献如下:

（1）本文提出了一种多任务多模态交互学习的自监督动态融合方法，采用从低级特征渐变到高级特征的融合策略，建立了单模态特征表示与两两模态特征的层次融合表示。

（2）在模型建立时，构建分布相似性损失函数和异质损失函数，学习不同模态的共性表征以及特性表征，进一步加强多模态特征的有效融合。同时，利用多任务学习策略，获得模态的一致性及差异性特征。

项目基金： 国家自然科学基金项目(62106130, 62076158, 62072294, 62272286)；山西省基础研究计划(20210302124084)；山西省高等学校科技创新项目(2021L284)；CCF-智谱AI大模型基金（CCF-Zhipu202310）

（3）本文的模型在CMU-MOSI和CMU-MOSEI数据集上分别进行实验，实验结果表明本文模型的情感分类性能优于基线模型。

2 相关工作

对于多模态表示学习方面，主要思想在于如何减少单模态表示之间的距离差异，使得不同模态之间的差距尽可能缩小。(Zadeh et al., 2017)利用张量融合网络通过三倍笛卡尔积将单模态分解为张量，然后计算这些张量的外积作为融合结果。(Liu et al., 2018)利用低秩多模态融合将堆叠的高阶张量分解成许多低秩因子，然后基于这些因子进行高效融合。(Tsai et al., 2018)将一个推理网络和一个具有中间模态特定因素的生成网络连接起来，以促进融合过程的重建和判别损失。Yu (2021)等人利用自监督学习策略，设计了一个标签自动生成模块，并将其运用在多模态和单模态的训练任务上，以达到减小模态差异的目的。Han (2021)等人将互信息的概念引入多模态情感分析中，提出了一种最大化互信息学习框架，避免了与任务相关信息的丢失。Hazarika (2020b)等人将模态向量投影至两个不同的空间中，利用正则化组件进行共有模态特征和特有模态特征的表示学习。

在多模态情感分析领域中，更多的研究是针对多模态融合方面的。如何将来自不同模态的数据进行有效融合，是该领域面临的一个挑战性问题。由于Transformer和BERT拥有强大的特征提取能力，许多研究针对Transformer 中的自注意力模块进行改进，使得不同模态的向量可以动态交互，从而达到跨模态融合、互补学习的目的 (Tsai et al., 2019)。Rahman (2020)等人提出了多模态门控组件，使得BERT 模型在不改变结构的基础上能够动态地接受多模态信息。(Sun et al., 2022)引入了一种基于元学习的方法来学习更好的单模态表示，然后将其用于随后的多模态融合。(Sun et al., 2023)提出一个通用的、统一的框架EMT-DLFR来实现鲁棒MSA。

受以上工作的启发，本文使用自注意力机制分别对两个单模态（音频、视频）特征进行抽取，模型采用从低级渐变到高级的特征融合策略，将不同层次的模态特征进行融合，并在多模态特征融合过程中，引入权重系数，用于刻画各个模态的情感贡献度，进一步，利用相似性损失与异质损失学习不同模态间的共性表征与特性表征。

3 多任务多模态交互学习的自监督动态融合方法

3.1 模型描述

为了使不同模态的层次特征充分融合，本文提出了多任务多模态交互学习的自监督动态融合方法(MMILN)，如图2所示。该方法的核心思想是通过对两个单模态特征（音频和视频），按层次进行抽取，并将不同层次的特征与不同层次的文本模态特征进行融合，使得文本模态与另外两个单模态（音频和视频）交互学习，得到更加准确的多模态特征。模型由三个模块组成，分别是不同模态数据的层次融合模块、三种模态表示再融合模块和多任务学习模块。

在多模态情感分析任务中，我们将$X_t \in R^{l_t \times d_t}$，$X_a \in R^{l_a \times d_a}$，$X_v \in R^{l_v \times d_v}$分别作为模型的输入，其中，$l_t$、$l_a$和$l_v$分别是文本、音频和视频的序列长度，$\hat{y}_m \in R$作为情感的预测结果。在训练过程中，采用两个单模态（音频和视频）的模型，辅助多模态的表示学习，这两个单模态模型的输出分别为：$\hat{y}_a, \hat{y}_v$，它们均由自监督学习生成，用于辅助更新多模态融合的输出结果$\hat{y}_m$。

对于音频模态（X_a）和视频模态（X_v），本文使用COVAREP与Openface/Facet提取音频与视频浅层特征后，再分别使用LSTM（Long Short-Term Memory）捕获模态的时序特征，从而获得特征向量M_a和M_v。

$$M_a = LSTM(X_a; \theta_a) \in R^{d_a} \tag{1}$$

$$M_v = LSTM(X_v; \theta_v) \in R^{d_v} \tag{2}$$

对于文本模态，我们使用预训练BERT模型提取句子表示M_t。

$$M_t = BERT(X_t; \theta_t) \in R^{d_t} \tag{3}$$

第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。

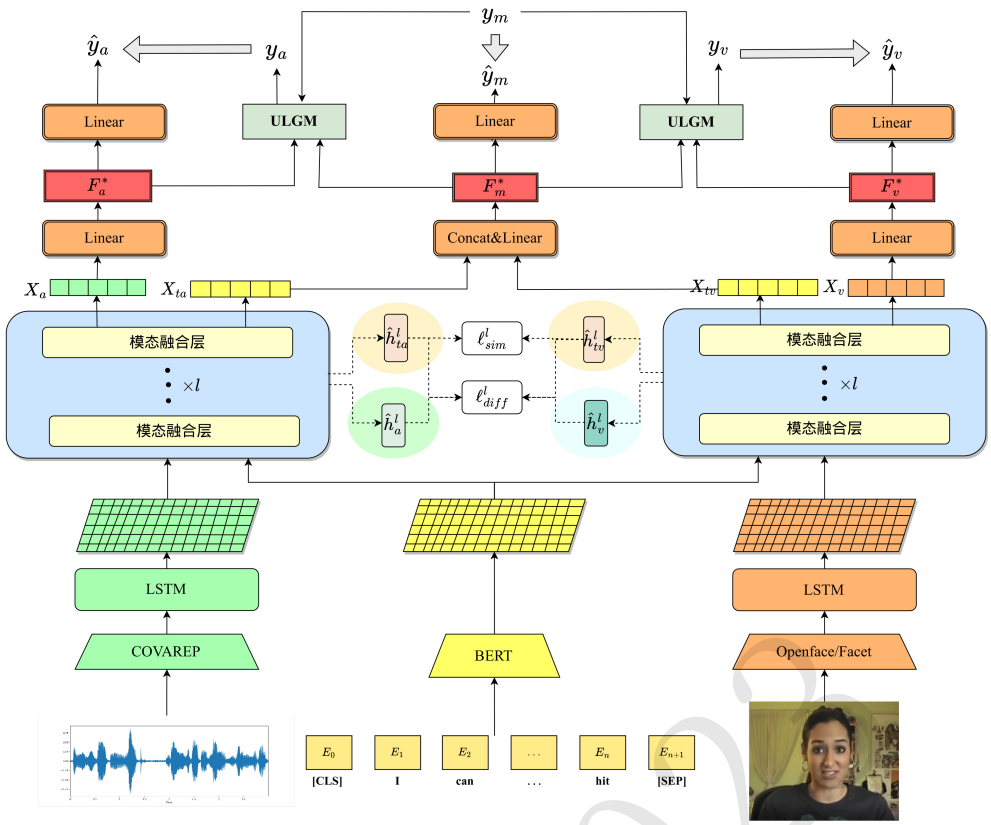


Figure 2: 模型整体结构图

利用公式(1)-(3)，获得了来自三个模态的特征表示:M_t、M_a和M_v，其中，θ_m表示模型的参数。之后我们使用卷积层统一三个模态的维度为R^{d_s} 。

$$X_m = Conv(M_m) \in R^{d_s} \tag{4}$$

其中，$m \in \{t, a, v\}$，其结果序列可以表示为：$X_m^0 = (X_{m,1}^0, X_{m,2}^0, \ldots, X_{m,k}^0)$作为模型的初始输入，k表示序列长度。

3.2 不同模态数据的层次融合模块

为了使不同模态的特征进行有效融合，我们使用多个模态融合层，并且构建了GateTransformer模型和SAG-Transformer模型。其中，前者用于单模态特征表示，建立分层网络结构对单模态特征进行从低级到高级的特征表示；后者用于两两模态特征的层次融合表示。本模块分为两个部分——单模态特征表示与两两模态特征的层次融合表示。模态融合层如图3所示。

3.2.1 单模态特征表示

本文采用以文本模态为中心，音频模态和视频模态为补充的表示方式。在本节中，我们对音频和视频模态分别进行特征表示，用于对文本模态的补充。音频（X_a^{i-1}）和视频（X_v^{i-1}）模态分别输入到各自的GateTransformer中，它们的结构完全一致，仅输入的模态数据不同，输出结果为单模态特征表示（X_n^i）与多头注意力值（$X_{mh\_n}^i$），$n \in \{a, v\}$。其中，单模态特征（X_n^i）作为下一个模态融合层的输入，多头注意力值（$X_{mh\_n}^i$）作为3.2.2节两两模态特征层次融合的输入。在得到前一个模态融合层的输出结果$X_n^{i-1} = (X_{n,1}^{i-1}, X_{n,2}^{i-1}, \ldots, X_{n,k}^{i-1})$之后，将音频模态和视频模态分别输入到GateTransformer中得到单模态特征和多头注意力结果，其公式如

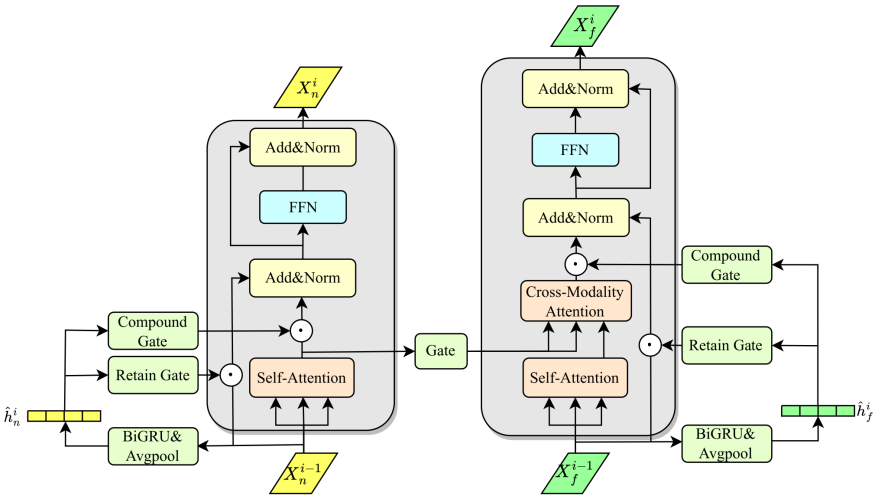


Figure 3: 模态融合层（由GateTransformer模型（左）和SAG-Transformer模型（右）组成）

下。

$$X_a^i, X_{mh\_a}^i = GateTransformer(X_a^{i-1}, X_a^{i-1}, X_a^{i-1}) \tag{5}$$
$$X_v^i, X_{mh\_v}^i = GateTransformer(X_v^{i-1}, X_v^{i-1}, X_v^{i-1}) \tag{6}$$

由于缺乏信息流控制，多头注意的操作表现出次优的性能。为了以细粒度和可控的方式改进它，我们使用GateTransformer，通过两个门——保留门g^r与复合门g^c对信息流控制。其中，保留门g^r决定了残差结构中模态信息的比例，以及复合门g^c决定了目标模态的前向传播比例。其公式如下。

$$\hat{h}_n^i = avgpool(h_n^i) = avgpool(BiGRU(X_n^{i-1}; \theta_n^i)) \tag{7}$$
$$g_r^i = \sigma(W_r^i \hat{h}_n^i) \tag{8}$$
$$g_c^i = \sigma(W_c^i \hat{h}_n^i) \tag{9}$$

其中，θ_n^i是BiGRU在第i层的参数，$W_* \in R^{d_s \times d_s}$，$* \in \{r, c\}$。

query $Q^i = W_Q^i X_n^{i-1}$，key $K^i = W_K^i X_n^{i-1}$，value $V^i = W_V^i X_n^{i-1}$用于多头注意力机制，并且使用门控机制限制单模态特征抽取的残差块与信息流。

$$X_{mh\_n}^i = MH - ATT(Q^i, K^i, V^i) \tag{10}$$
$$\widetilde{X}_n^i = LN(g_c^i \odot X_{mh\_n}^i + g_r^i \odot X_n^i) \tag{11}$$

其中，MH-ATT表示多头注意力、$\odot$表示按元素相乘、LN是层归一化。之后将注意力结果输出到前馈神经网络，并结合残差网络生成当前的模态融合层的最终结果。

$$X_n^i = LN(\widetilde{X}_n^i + FFN(\widetilde{X}_n^i)) \tag{12}$$

此模块生成的结果分别为X_a^i、X_v^i，表示单模态a（音频）、v（视频）在模态融合层第i层的输出，$X_{mh\_a}^i$，$X_{mh_v}^i$表示单模态a（音频）、v（视频）在第i层的多头注意力结果。

3.2.2 两两模态特征的层次融合表示

为了使模态特征从低级特征融合渐变到高级特征融合，我们采用分层的网络结构，将单模态特征从多个层次与文本模态进行融合。在本节中，我们利用3.2.1节单模态特征的多头注意力结果（$X_{mh\_n}^i$）与多模态特征（X_f^{i-1}）输入到SAG-Transformer，其中，$f \in \{ta, tv\}$。在得到

前一个模态融合层的输出结果$X_f^{i-1} = (x_{f,1}^{i-1}, x_{f,2}^{i-1}, \ldots, x_{f,k}^{i-1})$之后，我们将$X_f^{i-1}$与$X_{mh_n}^i$共同输入SAG-Transformer中。其公式分别如下所示。

$$X_{tv}^i = SAG - Transformer(X_{tv}^{i-1}, X_{mh\_v}^i, X_{mh\_v}^i) \tag{13}$$
$$X_{ta}^i = SAG - Transformer(X_{ta}^{i-1}, X_{mh\_a}^i, X_{mh\_a}^i) \tag{14}$$

值得注意的是:$X_{tv}^0 = X_t^0$，$X_{ta}^0 = X_t^0$。

SAG-Transformer是在GateTransformer的基础之上增加了交叉模态注意力机制，SAG-Transformer公式表示如下。

$$H_f^i = Self - Attention(X_f^{i-1}, X_f^{i-1}, X_f^{i-1}) \tag{15}$$
$$\hat{h}_f^i = avgpool(h_f^i) = avgpool(BiGRU(X_f^{i-1}; \theta_f^i)) \tag{16}$$
$$g_s^i = \sigma(W_s^i X_{mh\_n}^i) \tag{17}$$
$$\widetilde{X}_{mh\_n}^i = g_s^i \odot X_{mh\_n}^i \tag{18}$$

其中，θ_f^i是BiGRU在第i层的参数。

我们将query $Q^i = W_Q^i H_f^i$，key $K^i = W_k^i \widetilde{X}_{mh\_n}^i$，value $V^i = W_v^i \widetilde{X}_{mh_n}^i$用于多头注意力机制，并且为了限制双模态融合与残差块的信息流，定义门控机制如下:

$$g_r^i = \sigma(W_r^i \hat{h}_f^i) \tag{19}$$
$$g_c^i = \sigma(W_c^i \hat{h}_f^i) \tag{20}$$
$$r^i = MH - ATT(Q^i, K^i, V^i) \tag{21}$$
$$\widetilde{X}_f^i = LN(g_c^i \odot r^i + g_r^i \odot X_f^{i-1}) \tag{22}$$
$$X_f^i = LN(\widetilde{X}_f^i + FFN(\widetilde{X}_f^i)) \tag{23}$$

其中，$W_* \in R^{d_s \times d_s}$，$* \in \{r, c\}$，$(f, n) \in \{(ta, a), (tv, v)\}$ 也就是说$X_f^i \in \{X_{ta}^i, X_{tv}^i\}$。即，模态融合层在第$i$层的文本模态分别与音频、视频模态融合的输出表示。

3.3 三种模态表示再融合模块

在第3.2节不同模态数据层次融合模块的基础上，本节对文本与音频融合表示和文本与视频融合表示进一步融合，使模型学习到更加全面的模态互补信息。为了减少冗余，并捕获潜在模态表征，本文加入了两种损失函数，相似性损失和异质损失。

3.3.1 相似性损失

相似性损失用于学习不同模态间的共性表征，为此，设计最小化相似性损失，以减少每种模态的共享表征之间的差异。本文使用中心距差异(CMD)作为相似性损失。CMD (Zellinger et al., 2017)是一种先进的距离度量，它通过匹配两个表示的阶数矩差衡量两个表示的分布之间的差异，两个分布越相似，CMD距离会越小。

$$\begin{aligned} CMD_K(X, Y) = & \frac{1}{b-a} \parallel E(X) - E(Y) \parallel_2 \\ & + \sum_{K=2}^{K} \frac{1}{|b-a|^k} \parallel C_k(X) - C_k(Y) \parallel_2 \end{aligned} \tag{24}$$

其中，$E(X) = \frac{1}{|X|} \sum_{x \in X} x$是样本X的经验期望向量。$C_k(X) = E((x - E(X))^k)$是所有$k^{th}$阶样本的中心矩的向量。

在本文中，计算文本音频模态特征序列级隐藏表示（$\hat{h}_{ta}^i$）与文本视频模态特征融合表示（$\hat{h}_{tv}^i$）的CMD损失如下:

$$\ell_{sim}^i = CMD_K(\hat{h}_{ta}^i, \hat{h}_{tv}^i) \tag{25}$$

3.3.2 异质损失

异质损失用于学习不同模态间的特性表征。我们使用软正交约束度量异质损失，获取不同模态的特性表征。软正交约束 (Bousmalis et al., 2016; Liu et al., 2017; Ruder and Plank, 2018)，计算定义如下：

$$\ell^{i}_{diff} = \sum_{(m1,m2)\in\{(ta,a),(tv,v)\}} \| \hat{h}^{i\top}_{m1}\hat{h}^{i}_{m2} \|^2_F \tag{26}$$

3.4 多任务学习模块

表征学习是多模态学习中一个重要而富有挑战性的课题。有效的情感表达应包含一致性和差异性两部分特征。由于统一的多模态标签，现有方法在捕获差异化信息时受到限制。然而，额外标注单模态标签需要耗费大量的时间和人力。在本节中，我们使用了一个基于自监督学习策略的标签生成模块来获取独立的单模态标签。然后，对多模态和单模态任务进行联合训练，分别学习模态的一致性和差异性特征。

利用第3.2.2节两两模态特征的层次融合表示，得到了两个多模态融合表示X^l_f，$f \in \{ta, tv\}$，再利用线性层FFN，得到多模态特征表示F^*_m和预测结果$\hat{y}_m$。

$$\begin{aligned} F^*_m &= FFN(Concat(X^l_{ta}, X^l_{tv})) \\ \hat{y}_m &= FFN(F^*_m) \end{aligned} \tag{27}$$

对于第3.2.1节的单模态特征表示X^l_n,同样利用线性层FFN得到单模态特征表示F^*_n和预测结果$\hat{y}_n$，其中，$n \in \{a, v\}$，l表示模态融合层的最后一层。

$$\begin{aligned} F^*_n &= FFN(X^l_n) \\ \hat{y}_n &= FFN(F^*_n) \end{aligned} \tag{28}$$

与此同时，本文利用Yu (2021)等人设计的单模态标签生成模块(ULGM，Unimodal Label Generation Module)进行单个模态的训练。一般来说，单模态标签与多模态标签高度相关。因此，ULGM根据从模态表示到类中心的相对距离计算偏移量。

对于模态表示，我们使用L2归一化作为F^*_i和类中心之间的距离。

$$D^p_i = \frac{\|F^*_i - C^p_i\|^2_2}{\sqrt{d_i}} \tag{29}$$

$$D^n_i = \frac{\|F^*_i - C^n_i\|^2_2}{\sqrt{d_i}} \tag{30}$$

其中，$i \in \{m, a, v\}$，d_i是表示维度的比例因子，C^p_i和C^n_i分别表示正类中心和负类中心。

然后，对相对距离值进行定义，该值评估了模态表示到正向中心和负向中心的相对距离。

$$\alpha_i = \frac{D^n_i - D^p_i}{D^p_i + \epsilon} \tag{31}$$

其中，$i \in \{m, a, v\}$。ϵ是一个很小的数，为了防止除零异常。

为了得到标签和预测值之间的联系，考虑以下两种关系。

$$\frac{y_s}{y_m} \propto \frac{\hat{y}_s}{\hat{y}_m} \propto \frac{\alpha_s}{\alpha_m} \Rightarrow y_s = \frac{\alpha_s * y_m}{\alpha_m} \tag{32}$$

$$y_s - y_m \propto \hat{y}_s - \hat{y}_m \propto \alpha_s - \alpha_m \Rightarrow y_s = y_m + \alpha_s - \alpha_m \tag{33}$$

其中，$s \in \{a, v\}$。我们可以通过等权求和得到单模态标签。

$$\begin{aligned} y_s &= \frac{y_m * \alpha_s}{2\alpha_m} + \frac{y_m + \alpha_s - \alpha_m}{2} \\ &= y_m + \frac{\alpha_s - \alpha_m}{2} * \frac{y_m + \alpha_m}{\alpha_m} \\ &= y_m + \delta_{sm} \end{aligned} \tag{34}$$

第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。

其中，$s \in \{a, v\}$。$\delta_{sm} = \frac{\alpha_s - \alpha_m}{2} * \frac{y_m + \alpha_m}{\alpha_m}$表示单模态标签对多模态标签的偏移值。为了减轻由式(34)计算得到的标签不够稳定的问题，设计了基于动量的更新策略，该策略将新生成的值与历史值相结合，经过多次迭代使生成的标签值稳定。

利用两个单模态的特征表示F_n^*、多模态融合后的特征表示F_m^*和多模态标签y_m，共同输入到ULGM 模块后，生成了两个伪标签y_a,y_v，用于两个单模态任务训练过程。对于单模态子任务，本文使用L1损失，对于多模态子任务，使用均方误差构造MMILN 模型的优化目标函数。对于这两个单模态子任务，伪标签y_n与实际标签y_m的差异作为损失函数的权重，即$W_n^i = tanh(|y_n^i - y_m|)$。这意味着模型将注意力更加集中到模态特征差异性较大的样本上。具体损失函数定义如下。

$$\ell_{task} = \frac{1}{N}\sum_{i=1}^{N}((\hat{y}_m^i - y_m)^2 + \sum_{n\in\{a,v\}} W_n^i * |\hat{y}_n^i - y_n^i|) \tag{35}$$

最终，将ℓ_{sim}^l和ℓ_{diff}^l损失函数与任务损失ℓ_{task}进行组合，作为模型整体的损失函数定义如下:

$$\ell = \ell_{task} + \alpha\ell_{sim}^l + \beta\ell_{diff}^l \tag{36}$$

其中，α、β为两个超参数，决定每个正则化分量对总体损失ℓ的贡献权重。

4 实验与分析

本节包括以下三个部分:实验数据、实验设置、实验结果及分析。

4.1 实验数据

本文采用的实验数据是Zadeh等人发布的CMU-MOSI数据集 (Zadeh et al., 2016)和CMU-MOSEI数据集 (Zadeh et al., 2018b)。它们都是利用社交媒体数据得到的多模态情感分析数据集。CMU-MOSI是从93个Youtube的视频中获取的2199个独白类型的短视频片段。CMU-MOSEI包括来自5000个视频的23453个视频片段。数据的标注由人工完成，为情感的评分，分数值从-3到+3七个等级，其中，负值代表消极情感，正值代表积极情感，0分代表无情感。具体统计信息如表1所示。在实验过程中，为保证结果的公平性，本文采用已有工作 (Yu et al., 2021)对CMU-MOSI和CMU-MOSEI分别按照6:1:3与7:1:2的比例划分训练集、验证集和测试集。

Dataset	#Train	#Valid	#Test	#All
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856

Table 1: CMU-MOSI与CMU-MOSEI数据集的统计信息

4.2 实验设置

本文所提出的MMILN模型是在Yu (2021)等人的Self-MM模型的基础上改进而来。除层次融合模块的参数外，其余参数与Self-MM模型完全一致。我们在CMU-MOSI和CMU-MOSEI数据集上对本文所提出的模型及各基线模型进行了情感分析实验。分别采用分类和回归两种方法进行情感类别判别。对于分类指标，本文分别采用positive/negative和non-negative/negative计算F1-Score和二分类精度（Acc2）。对于回归指标，采用平均绝对误差（MAE）和皮尔逊相关（Corr）。除MAE外，值越高表示这项指标的性能越好。

4.3 实验结果及分析

在这一节，先将本文的模型与基线模型的性能进行对比和分析，之后进行了消融实验、模态融合层的数量对模型效果的影响实验，以验证本文所提出的模型设计的合理性。

4.3.1 与基线模型的对比及分析

为了充分验证本文模型MMILN的性能，我们与多模态情感分析任务中的多个模型进行实验比较。比较模型如下：

- TFN(Zadeh et al., 2017)：计算多维张量（基于外积），获取单模态、双模态和三模态相互作用
- LMF(Liu et al., 2018)：对TFN的改进，采用低秩多模态张量融合技术来提高效率
- MFN(Zadeh et al., 2018a)：持续建模特定视图和交叉视图交互，并通过多视图门控内存对其进行总结
- RAVEN(Wang et al., 2019)：利用基于注意力的模型，根据辅助非言语信号重新调整单词嵌入
- MFM(Tsai et al., 2018)：学习生成表征，获取模态特定的生成特征以及分类的区别表征
- MulT(Tsai et al., 2019)：扩展了具有定向两两交叉关注的多模态Transformer架构，使模态定向两两交互
- MISA(Hazarika et al., 2020b)：结合损失组合，包括分布相似性、正交损失、重建损失和任务预测损失，学习模态不变和模态特定表示
- MAG-BERT(Rahman et al., 2020)：对RAVEN在对齐数据方面改进，在Bert主干的不同层应用了多模态适配门
- SELF-MM(Yu et al., 2021)：为每个模态分配一个带有自动生成标签的单模态训练任务，目的是调整梯度反向传播
- MMIM(Han et al., 2021)：使用一个分层的互信息最大化框架来指导模型从所有模态中学习共享表示
- AMML(Sun et al., 2022)：引入了一种基于元学习的方法来学习更好的单模态表示，然后将其用于随后的多模态融合
- EMT(Sun et al., 2023)：利用统一的框架EMT-DLFR来实现鲁棒MSA，并具有更好的性能

	CMU-MOSI				CMU-MOSEI			
模型	ACC-2	F1-Score	Corr	MAE	ACC-2	F1-Score	Corr	MAE
TFN	-/80.8	-/80.7	0.698	0.901	-/82.5	-/82.1	0.700	0.593
LMF	-/82.5	-/82.4	0.695	0.917	-/82.0	-/82.1	0.677	0.623
MFN	77.4/-	77.3/-	0.632	0.965	76.0/-	76.0/-	-	-
RAVEN	78.0/-	76.6/-	0.691	0.915	79.1/-	79.5/-	0.662	0.614
MFM	-/81.7	-/81.6	0.706	0.877	-/84.4	-/84.3	0.717	0.568
MulT	81.5/84.1	80.6/83.9	0.711	0.861	-/82.5	-/82.3	0.703	0.58
MISA	81.8/83.4	81.7/83.6	0.761	0.783	83.6/85.5	83.8/85.3	0.756	0.555
MAG-BERT	84.2/86.1	84.1/86.0	0.796	0.712	84.7/-	84.5/-	-	-
Self-MM	84.00/85.98	84.42/85.95	0.798	0.713	82.81/85.17	82.53/85.30	0.765	0.530
MMIM	84.14/86.06	84.00/85.98	0.800	0.700	82.24/85.97	82.66/85.94	0.772	0.526
AMML	-/84.9	-/84.8	0.792	0.723	-/85.3	-/85.2	0.776	0.614
EMT	83.3/85.0	83.2/85.0	0.798	0.705	83.4/86.0	83.7/86.0	0.774	0.527
MAG-BERT*	82.54/84.36	82.48/84.37	0.796	0.717	82.10/85.09	82.42/84.95	0.754	0.543
MMIM*	83.67/85.52	83.56/85.47	0.800	**0.694**	77.81/82.55	78.66/82.7	0.700	0.615
EMT*	82.75/84.45	82.67/84.43	0.795	0.709	77.38/83.41	78.31/83.58	0.768	0.532
MMILN（Ours）	**84.55/86.43**	**84.49/86.42**	**0.801**	0.709	**84.67/85.86**	**84.70/85.62**	**0.774**	**0.529**

Table 2: CMU-MOSI数据集与CMU-MOSI数据集的结果

在表2上，分别展示了各模型在CMU-MOSI 和CMU-MOSEI数据集上情感分析的性能，*表示模型在相同条件下重现的结果，“/”左边表示non-negative或negative，右边表示positive或negative。由表2可以看到：

（1）本文的MMILN 模型在两个数据集上的分类性能均优于当前已有模型，且在MOSEI上的回归性能也优于其他模型，但MOSI数据集上的回归表现却并非最佳。其主要原因是MOSEI的数据量远超MOSI，使得测试实验更加稳定，也就是说模型在MOSEI上的性能可能更具有代表性和普适性。

（2）本文的MMILN 模型相较于基线模型Self-MM性能均较为突出。在CMU-MOSI和CMU-MOSEI数据集的ACC-2分别增加了0.55/0.45和1.86/0.69，F1-Score相较于Self-MM分别增加了0.07/0.47和2.17/0.32，Corr相较于Self-MM分别增加了0.003和0.009，MAE相较于Self-MM分别下降了0.004和0.001。这表明，我们所提出的从低级特征渐变到高级特征的融合策略有助于提升模型的整体性能。

4.3.2 消融实验

为了验证本文提出模型各模块的性能，我们使用CMU-MOSEI数据集设计如下的消融实验，其中减号“-”表示在这组实验中删掉的模型结构。

- -MFM：删除单模态特征表示门控机制与两两模态特征的层次融合门控机制（删除本模块会同时删除MRM）
- -MRM：删除三种模态表示再融合模块
- -GateFusion：删除多模态门控融合，仅使用单模态特征进行模态融合

模型/结构	ACC-2	F1-Score	Corr	MAE
-MRM	**85.13**/85.00	84.84/84.52	0.762	0.586
-MFM	84.01/85.75	84.20/85.62	0.771	0.574
-GateFusion	84.98/85.09	**84.78**/84.68	0.764	0.576
MMILN（Ours）	84.67/**85.86**	84.70/**85.62**	**0.774**	**0.529**

Table 3: 消融实验

从表3可以看出，本文所提模型的各个模块均发挥了一定的作用。-MRM模型相比MMILN，CMU-MOSEI数据集在positive/negative (right)方面，ACC-2下降了0.86，F1-Score下降了1.1，Corr下降了0.012，MAE增加了0.057。相应的，在non-negative/negative(left)方面，ACC-2和中F1-Score分别增加了0.46和0.14。说明MRM可以使得模型对于多模态融合特征表示有一定的提升效果，并进一步融合多模态融合特征表示，使模型学习到更加全面的模态互补信息。

-MFM模型相比MMILN，CMU-MOSEI数据集的ACC-2下降了0.66/0.11，F1-Score下降了0.50/0.00，Corr下降了0.003，MAE上升了0.045。说明MFM 与MRM通过使用门控机制可以对信息流进行有效地控制，并且使用两个损失函数可以将多模态表示进一步融合。

-GateFusion模型相比MMILN，在CMU-MOSEI数据集positive/negative (right)方面，ACC-2下降了0.77，F1-Score下降了0.94，Corr下降了0.010，MAE增加了0.047。相应的，在non-negative/negative (left)，ACC-2 增加了0.31和F1-Score 增加了0.08。说明仅使用单模态特征进行模态融合不利于模型学习到更加全面的模态信息，而通过门控机制对单模态信息进行筛选有助于提升模型的整体性能。

4.3.3 实例分析

为了展示CMU-MOSI数据集的示例，如表4所示。基本事实情感标签介于强烈消极(-3)和强烈积极(+3)之间。对于每个例子，均展示了Ground Truth和预测输出。其中，MMILN模型相对Self-MM模型，情感预测的性能更加准确。

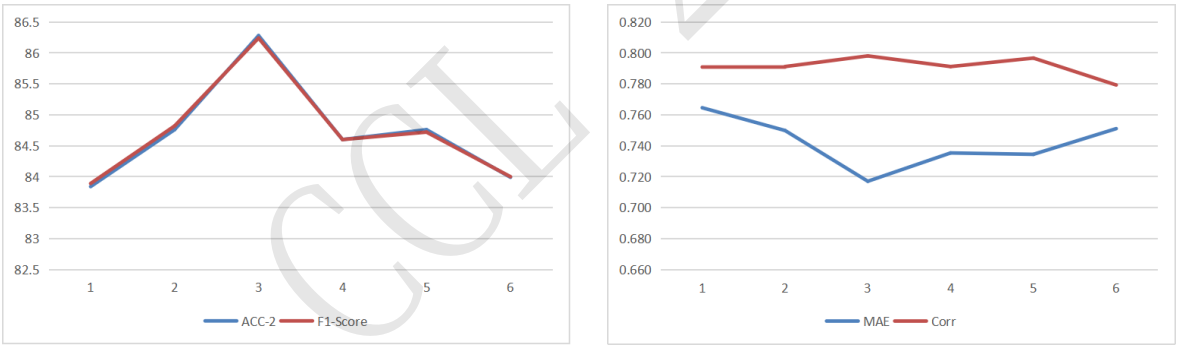
#	Spoken words + acoustic + visual behaviors	Ground Truth	MMILN	Self-MM
1	“I think you will really love this movie if you are 8.” + 强调的语音+嘴巴张大并且表情震惊	2.0	2.0	2.3
2	“It seems like the writers and creators took the easy way out.”+ 强调的语音+ 皱眉、眼角上扬	-1.6	-0.1	0.3
3	“I though that adults would appreciate” + 语调平稳+ 点头、眼睛左右看	0.4	0.7	1.2
4	“and I think its predictable up an to a point” + 语调平稳+ 表情平淡	-1.0	0.3	-0.4

Table 4: 实例分析

在例1和例3中，MMILN均相对准确的预测了情感的强度。在例2中，文本的描述相对中性，但是强调的语音和皱眉、紧张的表情帮助了MMILN模型预测的情感是负面极性。而Self-MM模型预测的情感极性是错误的，标注了正面极性，说明Self-MM模型并没有很好的利用到语音和视频模态。进一步证明了MMILN模型通过多层网络结构，采用从低级特征渐变到高级特征的融合策略是有效的。在例4中，MMILN模型将情感极性预测错误，但是，经过我们查验原始数据集后发现，原始数据的文本没有体现出情感，音频的语调平稳且说话人的表情平淡，并没有体现出情感，再一次体现出我们所提出的MMILN模型的性能比较突出。

4.3.4 模态融合层的数量对模型效果的影响实验

在本节中，我们将通过实验来研究模态融合层的数量对模型效果的影响。我们的实验在CMU-MOSI数据集上进行。其中，ACC-2与F1-Score取positive/negative值，模态融合层的层数选取1-6层。图4(a)展示的是不同层数下ACC-2与F1-Score的性能，图4(b)展示的是不同层数下MAE与Corr的性能。



(a)不同层数下ACC-2与F1-Score的性能展示　(b)不同层数下MAE与Corr的性能展示

Figure 4: 不同层数下MMILN模型性能比较

从图4(a)中可以看到，随着模态融合层的层数增加，ACC-2与F1-Score呈现先上升后下降的趋势，在层数为3时模型效果达到了最优。从图4(b)中可以看到，随着模态融合层的层数增加，Corr指标呈现波动，在层数为3和5时，模型效果较优。而MAE指标呈现出先下降后上升的趋势，总体来看，两者都在层数为3时模型效果达到了最优。这表明，我们所提出的MMILN模型对于模态融合层的层数较为敏感，提高模态融合层的层数一定程度上可以加强模态特征的融合，得到更加丰富的模态表征，并且提高模型的性能。

5 结论

本文提出了多任务多模态交互学习的自监督动态融合方法，该方法通过交互学习方式，采用从低级特征渐变到高级特征的融合策略，获得了模态间的一致性特征，再利用两个单模态特

征抽取子任务，获得模态自身的独特特征，从而对多模态融合特征进行了有效的补充和加强，较好地实现了模态间的相互作用。使用MFM和GateFusion，模型能够动态地、自适应调节不同层次的模态特征融合过程，减少与任务无关的弱模态特征信息对结果预测的干扰。为了进一步提升模型的性能，设计了模态表示再融合模块，加强了模态的融合。最后，通过实验验证，本文提出的方法加入到多任务学习的框架中之后，在测试集上性能优于主流基线模型，提升了模型的整体性能。

参考文献

Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *CoRR*, abs/1905.05812.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems*, 29.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 163–171, New York, NY, USA. Association for Computing Machinery.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: contextual sarcasm detection in online discussion forums. *CoRR*, abs/1805.06413.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020a. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020b. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. 2019. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *CoRR*, abs/1911.05544.

Y. Sun, S. Mai, and H. Hu. 2022. Learning to learn better unimodal representations via adaptive multimodal meta-learning. *IEEE Transactions on Affective Computing*, (01):1–1, may.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–17.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1823–1833, Online, November. Association for Computational Linguistics.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 949–954.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*.

第二十二届中国计算语言学大会论文集，第315页-第327页，哈尔滨，中国，2023年8月3日至5日。

基于动态常识推理与多维语义特征的幽默识别

吐妮可·吐尔逊[a], 林鸿飞*[a], 张冬瑜[b], 杨亮[a], 闵昶荣[a]
[a]大连理工大学，计算机科学与技术学院，大连，116024
[b]大连理工大学，软件学院，大连，116620
(Tunuh,11909060)@mail.dlut.edu.cn,(hflin,zhangdongyu,liang)@dlut.edu.cn

摘要

随着社交媒体的飞速发展，幽默识别任务在近年来受到研究者的广泛关注。该任务的目标是判断给定的文本是否表达幽默。现有的幽默识别方法主要是在幽默产生理论的支撑下，利用规则或者设计神经网络模型来提取多种幽默相关特征，比如不一致性特征、情感特征以及语音特征等等。这些方法一方面说明情感信息在建模幽默语义当中的重要地位，另一方面说明幽默语义的构建依赖多个维度的特征。然而，这些方法没有充分捕捉文本内部的情感特征，忽略了幽默文本中的隐式情感表达，影响幽默识别的准确性。为了解决这一问题，本文提出一种动态常识与多维语义特征驱动的幽默识别方法**CMSOR**。该方法首先利用外部常识信息从文本中动态推理出说话者的隐式情感表达，然后引入外部词典WordNet计算文本内部词级语义距离进而捕捉不一致性，同时计算文本的模糊性特征。最后，根据上述三个特征维度构建幽默语义，实现幽默识别。本文在三个公开数据集上进行实验，结果表明本文所提方法**CMSOR**相比于当前基准模型有明显提升。

关键词： 幽默识别 ； 常识推理 ； 模糊理论 ； 注意力机制

Humor Recognition based on Dynamically Commonsense Reasoning and Multi-Dimensional Semantic Features

Tuerxun·Tunike[a],Hongfei Lin* [a], Dongyu Zhang[b],Liang Yang[a], Changrong Min[a]
[a]School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
[b]School of Software Technology,Dalian University of Technology,Dalian,116620
(Tunuh,11909060)@mail.dlut.edu.cn,(hflin,zhangdongyu,liang)@dlut.edu.cn

Abstract

With the rapid development of social media,humor recognition has been a popular topic in the community of NLP.The goal of this task is to discriminate whether a given text expresses humor. Existing humor recognition methods mainly rely on the support of humor-centered theory, and use rules or designed neural network architectures to extract various humor-specific features, such as inconsistency features, emotional features, and linguistic features, etc. These methods indicate the importance of emotional information for modeling humor semantics, and also show that the construction of humor semantics depends on multi-dimensional features. However, these methods do not fully capture such emotional features within the text, ignoring implicit emotional expressions in humorous texts, which affects the accuracy of humor recognition. Therefore, we propose a novel approach named **CMSOR**, which is based on dynamic commonsense and

* 通讯作者

第二十二届中国计算语言学大会论文集，第328页-第340页，哈尔滨，中国，2023年8月3日至5日。

multi-dimensional semantic features for humor recognition. Specifically, it first makes use of external commonsense to infer latent emotions of speakers from the given text, and then leverage WordNet lexicon to calculate semantic distances from the word level, aiming to capture inconsistent features. This lexicon is also used to calculate the ambiguous features of the text. Eventually, we make use of such three kinds of humor-specific features to construct humor semantics. We conduct experiments over three publicly available benchmarks. The experimental results demonstrate that the proposed **CMSOR** is superior to the state-of-the-art baselines.

Keywords: Humor detection , Commonsense reasoning , Ambiguity theory , Attention mechanism

1 引言

幽默作为一种修辞手法，是人类交际中不可或缺的一部分，在使得人与人之间的沟通更加流畅的同时，营造了轻松愉悦的交流氛围。得益于社交媒体飞速发展所带来的海量文本数据，自然语言处理领域的文本幽默识别研究在近年来取得了长足进展。文本幽默识别的主要目标是通过计算方法来理解文本中的幽默表达并判断该文本是否为幽默。幽默识别不仅能够应用于文本生成、机器翻译以及隐喻识别等其他任务，还能够赋予机器理解幽默的能力，提升现实中人机交互的效果。因此，从文本中理解幽默产生的机制并识别幽默文本变得尤为重要。

从语言学与心理学的角度，主要存在三种观点来解释幽默的产生，分别是：优越论(Ritchie, 2009)、宽慰论(LeCun et al., 2015)以及乖讹论(Suls, 1972)。其中，优越论认为幽默是一种表达并强调自我价值与地位的方式，它强调通过取笑、讽刺或嘲笑他人来获取优越感；宽慰论认为幽默有助于缓解人们的压力和紧张情绪；乖讹论又称不一致性理论，它的表达方式通常会包含一些出人意料的不一致性，通过产生违背人们常识和期望的事物的感知，来引发人们的笑声和关注。基于上述理论，研究者们从多个角度提取文本中的幽默特征，同时通过设计不同结构的神经网络模型来学习幽默的深层次语义，基于此判断该文本是否为幽默。比如，Chauhan等人(Chauhan et al., 2022)认为幽默与情感和情绪密切相关，提出了利用Transformer和情绪感知嵌入(SE-Embedding)的多任务框架来检测幽默。Liu等人(Liu et al., 2018)基于“优越论”和“宽慰论”的观点，结合情感特征对语篇单元中的情感关系建模，证明了情感信息能更有效的解决对话幽默识别问题。Li等人(Li et al., 2020)使用“乐观幽默类型”和“悲观幽默类型”的情感极性来标注数据集中“积极”和“消极”情绪类别，采用Bi-LSTM模型结合注意力网络的方法，更好地捕捉俚语和微博表情符号在情感分析中的影响，为深入了解俚语和微博表情符号对中文情感分析提供了新视角。

从上述工作中可知，文本内蕴的情感特征对于识别幽默表达十分重要，这些工作主要通过外部词典匹配的方式来捕捉文本内的情感特征。然而，本文发现在幽默表达中很多情绪往往是隐式表达的，如表1所示，其中第二个幽默样本表达了“悲伤”或者“愤怒”的情绪，但是该样本并没有包含直接表达情绪的词汇，而是通过短语“get fired”来表达。这种方式称之为隐式情感表达。现存的幽默识别方法主要采用外部情感词典来捕捉文本内的情感信息。显然，这种方式无法有效识别出这些隐式情感表达，这降低了模型识别文本幽默的能力。

从认知角度，理解这些隐式情绪表达需要不仅需要结合上下文信息，还有充分利用外部常识。尽管现有的预训练语言模型（PLM）能够高效的捕捉文本的上下文信息，但是由于其是在大规模通用语料上训练，因此无法有效感知这些文本背后的隐式情绪。为了解决这一问题，本文提出一种动态常识与多维语义特征驱动的幽默识别方法(**C**ommonsense and **M**ulti-dimensional **S**emantics based Humor Detect**or**)，简写为**CMSOR**。该方法主要是利用外部常识，根据文本的上下文信息，动态地推断文本中的隐式情绪，并将其作为文本情绪特征的一部分，参与幽默识别。具体地，该方法首先根据文本内容利用预训练常识推理工具COMET (Bosselut et al., 2019)根据上下文信息动态推断文本的内蕴情感信息，然后将文本内容与推断出的情感信息拼接融合，通过预训练语言模型BERT进一步将显式情感融入到文本语义当中，形成显式情感增强

幽默文本	情绪信息
I used to play piano by ear, but now I use my hands.	自嘲
I can't believe I got fired from the calendar factory. All I did was take a day off!	愤怒
I don't trust people who do acupuncture. They're back stabbers.	不满

Table 1: 幽默样本以及包含的情绪信息

的文本表示。同时，利用外部词典WordNet计算语义距离以及同义词数量，分别形成文本的不一致性特征以及模糊性特征。最后，将上述三种特征进行结合，形成多维幽默语义表示，输入到分类器中，得到幽默预测结果。本文做出的贡献总结如下：

（1）本文提出了一种动态常识驱动的幽默识别方法**CMSOR**，利用外部常识动态捕捉文本的隐式情感特征，同时利用外部词典建模模糊性与不一致性特征，从多个维度构建幽默语义，实现幽默识别。

（2）本文在Pun of the Day、SemEval21以及ColBERT三个公开数据集上进行了实验，实验结果表明本文所提出的**CMSOR**模型相比于现有方法在四项评价指标上有明显提升，说明了该方法的有效性。

本文组织结构如下：第2章主要介绍幽默识别相关工作。第3章介绍文本所提出的**CMSOR**模型。第4章主要介绍实验设置以及实验结果分析。第5章为总结与展望。

2 相关工作

由于幽默表达本身的复杂性，幽默识别在近些年来一直是一项极具挑战的任务。早期的幽默识别方法主要是基于特征工程，利用统计机器学习方法作为分类器，在幽默理论的基础上设计不同的幽默特征提取方案。这些人工提取的特征包括通用语言学特征以及面向幽默的文本特征。比如，Mihalcea和Strapparova(Mihalcea and Strapparava, 2005)定义了头韵、反义词和成人俚语三种幽默特征，通过实验证明了他们在one-liner数据集中幽默识别的有效性。Mihalcea等人(Mihalcea et al., 2010)将幽默文本分为"铺垫"和"笑点"两部分，通过计算两者的语义相关性进行幽默识别。Yang等人(Yang et al., 2015)深入探讨幽默潜在语义特征，构造了四种幽默特征分别是语音特征、歧义特征、不一致性特征和情感特征。Morales和Zhai(Morales and Zhai, 2017)针对Yelp评论使用概率模型结合背景文本资源进行幽默识别。Cattle和Ma(Cattle and Ma, 2018)利用单词关联的语义关联特征进行幽默识别。上述这些工作大多是利用统计或者匹配的方法来提取文本中的浅层幽默特征，无法对于幽默的深层次潜在语义进行表示，从而限制了幽默识别的性能。

随着计算能力的进步以及社交媒体数据的爆炸性增长，深度学习在不同领域被广泛用以辅助或替代传统的特征工程。与其他领域相比，深度学习在幽默识别任务中应用较晚。这些基于深度神经网络的幽默识别方法主要是利用预训练语言模型表示文本，然后设计不同结构的神经网络实现对于幽默特征的深层次提取。比如，Bertero等人(Bertero and Fung, 2016)认为幽默情景剧是一种具有独特特点的喜剧形式，背景笑声可以视为观众对于搞笑场景的反应，自动标注这些笑声可以有效地识别笑点，便在此基础上使用长短期记忆网络（LSTM）对幽默情景剧中的对话进行建模，同时提取对话语义特征和声音特征用于识别笑点。Buenod等人(Ortega-Bueno et al., 2018)针对西班牙推文结合语言特征和基于注意力的递归神经网络进行幽默识别。Blinov等人(Blinov et al., 2019)收集大量笑话和趣味对话构造俄语数据集，并微调语言模型用于幽默识别。Justine T. Kao等人(Kao et al., 2016)提出模糊性和独特性两个特征使用语言模型识别幽默语句。Weller和Seppi(Weller and Seppi, 2019)使用Transformer架构识别幽默。Hasan等人(Hasan et al., 2019)使用循环神经网络针进行多模态幽默识别。Diao等人提出(Diao et al., 2018)一种基于不一致性、模糊性、情感因素和语言学潜在语义结构的识别模型。Fan等人(Fan et al., 2020) 基于Bi-GRU网络融合语音特征和歧义性特征进行幽默检测。Annamoradnejad和Zoghi(Annamoradnejad and Zoghi, 2020)改进Bert模型在自创建的幽默数据集ColBERT上进行实验，证实了提出模型能够有效的检测幽默。Zhang等人(Zhang et al., 2021)利用卷积神经网络结合标签转移关系提出多任务学习模型识别幽默。Ren等人(Ren et al., 2021)结合幽默和双关语识别任务，提出一种基于注意力的多任务学习模型来进行幽默检

测。Ren等人(Ren et al., 2022)提出一种基于注意力机制的神经网络来验证发音、句法与词法特征对于幽默识别任务的重要性。

与上述工作类似，本文同样考虑了情感特征在幽默表达中的重要作用。不同的是，本文为了解决幽默文本中隐式情感表达难以被词典有效识别的问题，采用动态常识推理，从文本中推断内蕴的隐式情感。并结合模糊特征与不一致特征，从多个维度对于文本的幽默语义进行刻画。

3 模型

3.1 问题描述

幽默识别任务的具体目标可以描述为：给定幽默识别训练集$\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$。其中$x_i = (w_1, w_2, \ldots, w_m)$为输入文本序列，$m$为其中包含的单词总数。$y_i \in [0, 1]$为对应的幽默标签，$N$为训练集中样本总数。幽默识别任务的目标则是学习到一个映射函数$f : \mathcal{D} \rightarrow y \in [0, 1]$，以此预测输入文本序列是否为幽默。

3.2 模型整体框架

本文所提出的幽默识别方法**CMSOR**结构如图1所示。该模型主要由三个部分组成：情感特征提取层、语义特征提取层、模糊性特征提取层。其中，情感特征提取层主要是考虑到幽默表达中存在大量隐式情感表达，利用外部常识推断文本中的隐式情感表达，充分挖掘文本中的情感特征；语义特征提取层主要是通过计算句子内部词对之间的语义关联来学习文本内部的不一致性特征；模糊性特征提取层主要是利用外部词典捕捉文本中存在歧义性的词汇，通过循环神经网络学习其模糊性特征。最后，将幽默的三个维度特征进行拼接，通过分类器，获得文本的幽默预测结果。

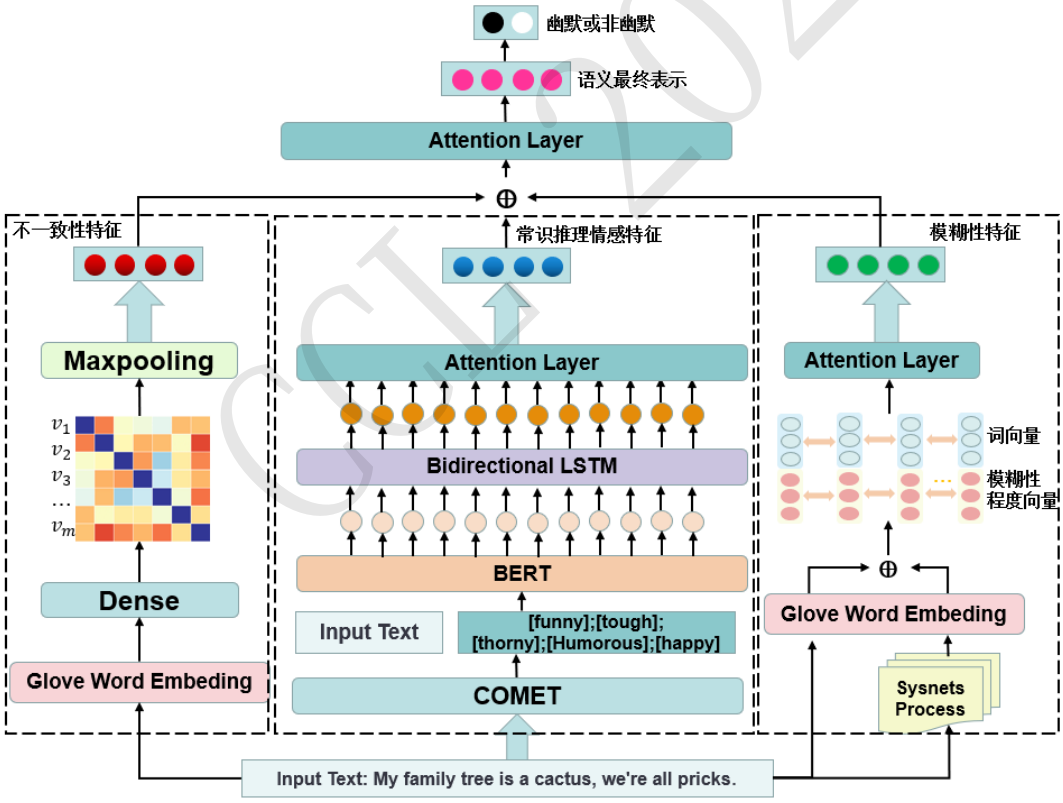


Figure 1: **CMSOR**模型结构图

3.3 外部常识驱动的情感特征提取

幽默表达与情感有着极大的关联。一些带有强烈感情色彩的词会增加受众对于作者表述的认同感，使得读者的情绪被更为充分的调动，从而达到幽默的效果(樊小超et al., 2021)。然而幽

默内存在的隐式情感表达使得通过外部词典捕捉文本情感特征变得十分困难。为了解决这一问题，本文采用预训练常识推理模块COMET(Bosselut et al., 2019)根据上下文信息动态推断文本内所蕴含的情感特征。COMET作为一种常识推理工具，在给定上下文的情况下，能够根据不同的事件关系来推理相应的结果。COMET是以Transformer为基础架构，并在ATOMIC$^{20}_{20}$(Sap et al., 2019)数据集上训练得到。该数据集共提供23种事件关系，而本文主要采用[**xReact**]这一关系。它的功能是根据上下文推断句子中主语的内心情绪，并以文本形式输出。

具体地，以幽默文本序列$x = (w_1, w_2, \ldots, w_m)$作为输入，COMET能够根据$x$推理出说话者可能的内心情绪。在这里，本文选择概率最高的前l个可能结果，并得到说话者情绪候选集$K = \{k_1, k_2, \ldots, k_l\}$。其中,$k_i$表示第$i$个情绪词。然后，将初始文本序列$x$与情绪候选集拼接，得到显式情绪增强的幽默文本序列：

$$x_e = \{w_1, w_2, \ldots, w_m, [\mathbf{SEP}], k_1, k_2, \ldots, k_l\} \tag{1}$$

其中，[**SEP**]为句子分割符。然后，本文采用BERT对于x_e进行上下文编码。其计算公式如下：

$$v_e = \mathbf{BERT}\left(x_e; \mathbf{W_0}\right) \tag{2}$$

其中，v_e为编码后得到的句子表示，$\mathbf{W_0}$为BERT的可学习参数。

一方面，BERT能够有效的捕捉上下文信息，将幽默文本x中的单词w_i与情绪候选集K中的情绪单词k_i从语义层面上关联起来，进而有效捕捉文本内的情绪特征。另一方面，BERT内的多头注意力机制能够为文本中的每个单词赋予不同的权重，通过降低与幽默文本无关的情绪词的权重，来避免引入过多噪声信息。在得到上下文编码后，采用双向长短时记忆神经网络（Bi-LSTM）对于上下文语义信息进行进一步学习，最后通过注意力机制获取潜在情感特征z_e，其计算公式如下：

$$u_e = \mathbf{Bi-LSTM}\left(v_e; \mathbf{W_1}\right) \tag{3}$$

$$z_e = \mathbf{Attention}\left(u_e; \mathbf{W_2}\right) \tag{4}$$

其中，$u_e \in \mathbb{R}^{1\times p}$为输出的幽默文本表示，$p$为Bi-LSTM的隐藏层维度。$\mathbf{W_1}$为Bi-LSTM的可学习参数,$\mathbf{W_2}$为注意力机制的可学习参数。

3.4 基于语义距离的不一致性特征提取

一些语言学研究(Lefcourt, 2001; Paulos, 2008)认为幽默的本质在于表现出两种不一致的思想或概念。同样的，Raskin等人(Raskin, 1979)也指出幽默的产生往往借助于一些有意义但含义不同或相反的词语或短语的组合，通过制造错觉或矛盾感而达到幽默的效果。比如：

例*3.1*：*I am deeply aware that I am a superficial person.*

例3.1中“*deeply*”可以翻译成“深刻”，“*superficial*”可以翻译成“肤浅”。这个句子的中文翻译是“我深刻的意识到我是个肤浅的人”，其中“深刻”和“肤浅”有相反的含义，达到幽默效果。上述例子也可以说明幽默中的不一致特征具有隐晦和抽象的特点并与深层次语义关联紧密。从听者角度，需要具有背景知识才能够推断出词汇或者短语之间的隐含关系。因此，需要引入外部知识更好地捕捉幽默的不一致性特征。

具体地，给定一个输入文本序列$x = (w_1, w_2, \ldots, w_m)$，本文首先通过预训练语言模型将文本序列中的每个词进行向量化表示并得到$\mathbf{V} = [v_1, v_2, \ldots, v_m] \in \mathbb{R}^{m\times d}$。其中，$d$表示词向量维度。然后，针对于$x$中的每个词$w_i$，利用WordNet(Miller, 1995)获取其词义特征，并得到$\mathbf{H} = [h_1, h_2, \ldots, h_m] \in \mathbb{R}^{m\times d'}$，$d'$表示其词义特征维度。将词义信息$\mathbf{H}$与深层次语义信息$\mathbf{V}$进行拼接，得到$\mathbf{V'} = [v'_1, v'_2, \ldots, v'_m] \in \mathbb{R}^{m\times(d+d')}$。为了计算词级语义不一致性，首先采用两个平行语义编码器对于文本表示V'进行压缩。编码器由全连接神经网络实现。具体计算如下：

$$\hat{\mathbf{V}} = \sigma(\mathbf{W_3}\mathbf{V'} + \mathbf{b_2}) \tag{5}$$

$$\overline{\mathbf{V}} = \sigma(\mathbf{W_4}\mathbf{V'} + \mathbf{b_3}) \tag{6}$$

其中，$\hat{\mathbf{V}}$与$\overline{\mathbf{V}}$分别表示压缩后的语义信息，$\mathbf{W_3}$与$\mathbf{W_4}$表示两个编码器中的可学习参数，σ表示激活函数。然后，对于$\hat{\mathbf{V}}$与$\overline{\mathbf{V}}$进行点积运算，获得不一致性矩阵$\mathbf{S} = \hat{\mathbf{V}} \cdot \overline{\mathbf{V}}^{\mathrm{T}}$，用以刻画幽默文本的内部不一致性特征。

3.5 基于同义词的模糊性特征提取

Reyes和Rosso(Reyes and Rosso, 2012)认为幽默是一个单词的多个含义对句子产生不同的理解，借助语义和语境的歧义来产生的。Miller和Gurevych(Miller and Gurevych, 2015)指出模糊性是幽默的关键因素，是幽默中常见的语言现象。随之Reyes等人(Reyes et al., 2012)得出结论：幽默的表达往往伴随着语义的模棱两可。如下例：

例*3.2*: *Why did the tomato turn red? Because it saw the salad dressing!*

例*3.3*: *My trip to the grand canyon cost a hole lot of money and gorged my bank account butte it was worth it.*

例3.2中"*salad*"一词既可以被解释为用于沙拉的一种酱汁，也可以表示"穿衣服"的意思，从而导致句子产生两种截然不同的意义来产生幽默效果。例3.3中，首先"*hole*"字面含义为"洞"，但在口语中也可表示为"大量"或"很多"，其次"*butte*"在字面含义为"丘陵"，但在句中被用作双关词，与"*but*"相呼应。句子通过"*hole*"和"*butte*"的双关含义，使例3.3既可描述为旅行花费了大量的钱，也可暗示这个花销像一个巨大的洞一样，吞噬了大量的资金。结合上述例子，幽默通过词汇的多个含义来创造幽默达到幽默效果。由此可见，模糊性是判断是否幽默的重要因素之一，是幽默文本的重要组成部分。综上所述，本文为提高幽默识别的性能，利用外部资源Wordnet捕获句子中的歧义词。

在WordNet数据库中，名词、动词、形容词和副词都被存储为同义词集合的形式，每一个同义词集合被称为一个Synset包含一组具有相似意义的单词。不同的Synset之间可以通过语义关系和词性关系等边相连接，这些关系可以帮助人们理解这些单词之间的联系和含义。

针对于输入文本序列$x = \{w_1, w_2, \ldots, w_m\}$，首先利用WordNet中的同义词集合Synset计算每个w_i的同义集数量n。本文认为单词的同义词数目越多，会导致句子理解存在很多歧义，从而模糊性程度就会增加，因此本文将同义集数量最多的词汇设置为最容易出现歧义的词汇，停用词在句子中不承载实际的语义信息，因此可以被移除或忽略，从同义词集合和同义词数量中删除停用词汇及其个数。针对于同义词集的数量，定义如下规则来描述每个词的模糊程度：

$$c = \begin{cases} 0 & n = -1 \\ 1 & 0 < n \leq 5 \\ 2 & 5 < n \leq 15 \\ 3 & 15 < n \leq 30 \\ 4 & n > 30 \end{cases} \tag{7}$$

得到x的模糊程度序列$C = \{c_1, c_2, \ldots, c_m\}$，其中0表示模糊程度最低，4表示模糊程度最高，对于文本中的停用词，其模糊程度统一设定为0。然后，将该序列C进行one-hot表示，得到模糊程度矩阵$\mathbf{V_c} = [c_1, c_2, \ldots, c_m] \in \mathbb{R}^{m \times d}$。将$\mathbf{V_c}$与文本表示$\mathbf{V} = [v_1, v_2, \ldots, v_m] \in \mathbb{R}^{m \times d}$通过拼接方式进行融合，并利用模糊特征编码器$\mathcal{G}_{\mathbf{fuz}}$学习包含模糊特征的文本表示，该编码器由Bi-LSTM及注意力机制实现。其计算公式如下：

$$z_f = \mathcal{G}_{\mathbf{fuz}}\left([\mathbf{V_c} \oplus \mathbf{V}]; \mathbf{W_5}\right) \tag{8}$$

其中，z_f为模糊性特征表示，p'为Bi-LSTM的隐藏层维度。$\mathbf{W_5}$为可学习参数,$\oplus$表示拼接操作。

3.6 幽默标签预测以及损失函数

在获得幽默文本的情感特征z_e、不一致性特征$z_s = \mathbf{MaxPooling}(\mathbf{S})$以及模糊性特征$z_f$之后，将三种特征通过拼接方式进行融合，得到多维度融合幽默特征$z = z_e \oplus z_s \oplus z_f$。通过注意力机制进一步学习三种特征之间的内在关联，具体计算如下：

$$Z = \mathrm{Attention}(z; \mathbf{W_6}) \tag{9}$$

其中，$\mathbf{W_6}$表示注意力机制层的可学习参数。在此基础上，将其输入到由全连接层构成的幽默分类器f_h中，获得文本x的幽默标签预测。具体计算公式如下：

$$\hat{y} = f_h(Z; \mathbf{W_7}) \tag{10}$$

其中，$\mathbf{W_7}$表示可学习的参数矩阵，$\hat{y}$表示幽默预测结果。

最后，**CMSOR**在分类中采用交叉熵（Cross Entropy）作为损失函数。其损失计算如下：

$$\mathcal{L} = -\frac{1}{\mathrm{N}} \sum_{i}^{\mathrm{N}} \left(\hat{y}_i \log y_i + (1 - \hat{y}_i) \log (1 - y_i)\right) \tag{11}$$

4 实验与分析

在本节中，将从以下四个部分介绍实验的细节：数据集、实验数据与设置、对比试验、消融实验。

4.1 数据集

为了证明方法的有效性，本文实验中使用了三个公开的数据集，其统计信息如表2所示。具体介绍如下：

- **Pun of The Day** (Yang et al., 2015)：这个数据集的构建是Yang等人通过在互联网上收集幽默文本而完成的，包括了各种类型的幽默，如双关语、笑话、俏皮话等等。为确保数据的准确性和可靠性，通过人工标注和质量控制的方式对数据进行了筛选和整理。该数据集目前广泛使用于幽默识别中。

- **SemEval 2021 Task 7-1a** (Meaney et al., 2021)：该任务是一项国际评测，Task 7子任务一是识别文本是否为幽默文本，该数据集可以用来幽默检测，本文利用Task 7子任务一涉及数据来判断是否为幽默文本。

- **ColBERT** (Annamoradnejad and Zoghi, 2020)：该数据集是一个大规模的幽默数据集，它包含了20万个来自网络的英文幽默文本，其中10万正样本由Reddite收集得到，另外10万负样本来源于新闻头条。

数据集	正样本	负样本
Pun of The Day	2403	2403
SemEval 2021 Task 7-1a	5547	3453
ColBERT	100,000	100,000

Table 2: 数据集统计信息

4.2 实验数据与设置

实验在python3.7和Kreas2.2.4环境下进行。对于本文提出的**CMSOR**模型，其中常识知识层本文采用12层的BERT-base-cased [0] 作为预训练语言模型编码输和知识，其中向量维度为768，共110M个参数；语义特征提取以及模糊性特征提取采用GloVe，维度100，词嵌入在训练的过程中固定，不在词汇表中出现的单词词使用(0.01,0.01)上的平均分布随机初始化；使用WordNet获取单词同义词集合；Bi-LSTM的神经元数量为128；Dropout为0.3；Batch大小为64；模型采用Adam Optimzation优化算法更新模型参数；采用了学习率衰减和早停机制以防止过拟合现象。此外采用准确度（Accuracy）、精确率(Precision)、召回率(Recall) 和F1值(F1-Score)作为实验结果的评价指标，并且所有实验均进行五倍交叉验证，取平均值作为实验结果。

[0]https://huggingface.co/bert-base-cased

4.3 对比试验

本文采用如下基线模型进行对比:

- **LSTM** (Graves and Graves, 2012):通过经典LSTM 模型提取幽默特征进行幽默识别。
- **Bi-LSTM**:利用可以更好的捕捉双向语义依赖关系的Bi-LSTM模型。
- **Bi-LSTM+ATT**:使用Bi-LSTM模型结合注意力机制提取幽默特征进行幽默识别。
- **CNN**:采用CNN获取幽默语句的潜在语义及模糊性特征进行幽默识别。
- **CNN+F+HN** (Chen and Soo, 2018):采用了融合人工特征的CNN和highway神经网络模型。
- **BERT** (Devlin et al., 2018):使用预训练BERT模型在幽默数据集上进行微调。
- **IEANN** (Fan et al., 2020):通过结合内部及外部注意力神经网络构建两种注意力机制，以捕捉幽默文本中的不一致性和模糊性特征。
- **ABML** (Ren et al., 2021):通过联合幽默和双关语检测的多任务学习模型进行幽默识别。
- **ANPLS** (Ren et al., 2022):通过结合发音、词汇和句法幽默特征的注意力网络，提取幽默特征进行幽默识别。

Dataset	PUN OF THE DAY			
Model	ACC	P	R	F1
LSTM	84.97%	84.02%	84.57%	84.29%
Bi-LSTM	86.11%	85.13%	85.87%	85.50%
Bi-LSTM+ATT	86.94%	87.95%	84.13%	86.00%
CNN	86.42%	83.18%	91.56%	87.17%
CNN+F+HN*	89.40%	86.60%	94.00%	90.10%
BERT	90.50%	88.75%	91.80%	90.46%
IEANN	92.24%	91.14%	92.25%	91.69%
ABML	93.18%	92.45%	92.07%	92.26%
ANPLS	92.94%	93.00%	92.55%	92.79%
CMSOR	**94.56%**	**93.47%**	**92.61%**	**93.24%**

Table 3: Pun of The Day 数据集上实验结果，*表示结果引用自对应论文，加粗表示最优实验结果。

Dataset	SemEval 2021 Task 7-1a			
Model	Acc	P	R	F1
LSTM	83.30%	83.06%	81.34%	82.44%
Bi-LSTM	84.90%	87.79%	87.64%	87.71%
Bi-LSTM+ATT	84.70%	87.62%	87.48%	87.55%
CNN	86.15%	87.20%	90.32%	89.02%
BERT	91.78%	93.29%	92.62%	92.14%
IEANN	91.03%	91.32%	92.10%	91.71%
ABML	**92.20%**	91.92%	92.77%	92.34%
ANPLS	92.06%	92.38%	93.07%	92.72%
CMSOR	92.15%	**92.67%**	**93.40%**	**93.34%**

Table 4: SemEval 2021 Task 7-1a数据集上实验结果，加粗表示最优实验结果。

Dataset	Colbert			
Model	Acc	P	R	F1
LSTM	93.60%	93.82%	94.05%	93.93%
Bi-LSTM	94.07%	94.80%	93.19%	94.08%
Bi-LSTM+ATT	95.48%	96.01%	94.84%	95.42%
CNN	94.40%	93.18%	95.81%	94.45%
BERT	95.55%	95.57%	95.47%	95.52%
IEANN	94.92%	95.33%	93.87%	94.59%
ABML	94.42%	94.39%	94.16%	94.27%
ANPLS	94.39%	94.82%	95.07%	94.94%
CMSOR	**96.23%**	**95.98%**	**96.40%**	**96.19%**

Table 5: ColBERT数据集上实验结果，加粗表示最优实验结果。

实验结果如表3，4，5所示，从表中可以得到如下结论：(1)本文提出的**CMSOR**方法在三个数据集上均取得了最好的结果，在三个数据集上的F1值相比于现存的最优结果分别提升了0.45%、0.62%、0.67%，证明了从情感、不一致性以及模糊性三个维度构建幽默语义并应用于幽默识别是有效的。(2)从表中可以看出，相比于基于CNN或者RNN的幽默识别方法，基于Transformer的方法（BERT以及**CMSOR**）在四项评价指标上有明显提升，这说明Transformer能够通过全局注意力机制更好地捕捉幽默文本的上下文信息。(3)**CMSOR**方法能够通过深度神经网络结构，在外部知识驱动下，自动构建幽默特征，相比于人工提取幽默特征（CNN+F+HN），取得了明显的提升（F1值提高3.14%）。这也验证了深度学习模型能够在幽默理论约束下学习到幽默相关特征。(4)相比于基于RNN的方法，基于CNN的方法在三个数据集上的F1值取得了明显的提升，比如在Pun of The Day数据集上，CNN+F+HN相比于BILSTM+ATTEN在F1值上提高了4.46%。这说明幽默表达可能与局部语义信息（N-gram）有着一定的关联。(5)与采用情感词典捕捉文本内部情感信息的IEANN相比，**CMSOR**在F1值上有明显提升（1.6%），这说明利用动态外部常识信息能够更准确的推断文本内部情感。(6)ABML模型在三个数据集上相比较IEANN和ANPLS，ACC值达到最高。ABML模型不仅考虑双关语的特点，还考虑了幽默和双关语之间共同的潜在语义信息。这意味着模型能够更好地理解双关语的双重含义，并将其与幽默特征联系起来，有效的增强模型对幽默的识别能力。

4.4 消融实验

为了验证**CMSOR**中不同组件的有效性，本文在三个数据集上进行消融实验，并设计以下模型变体：CMSOR-C表示仅使用情感特征；CMSOR-I表示仅使用语义不一致性特征；CMSOR-A表示仅使用模糊性特征；CMSOR-CI表示融合情感特征和语义不一致性特征；CMSOR-CA表示融合情感特征和模糊性特征；CMSOR-IA表示融合语义不一致性特征和模糊性特征。

Model	ACC	P	R	F1
BILSTM	86.11%	85.13%	85.87%	85.50%
CMSOR-C	86.53%	85.71%	86.09%	85.90%
CMSOR-I	86.32%	86.12%	85.00%	85.56%
CMSOR-A	91.48%	96.27%	86.20%	90.96%
CMSOR-CI	91.81%	92.24%	90.43%	91.33%
CMSOR-CA	92.23%	91.22%	92.61%	91.91%
CMSOR-IA	92.95%	92.06%	**93.26%**	92.66%
CMSOR	**94.56%**	**93.47%**	92.61%	**93.24%**

Table 6: Pun of The Day数据集消融实验，加粗表示最优实验结果。

三个数据集上的消融实验结果分别如表6，7，8所示。从三个表中可以得到如下结

Model	ACC	P	R	F1
BILSTM	84.90%	87.79%	87.64%	87.71%
CMSOR-C	86.60%	88.24%	90.11%	89.17%
CMSOR-I	85.70%	86.99%	90.24%	88.59%
CMSOR-A	86.30%	87.94%	90.08%	89.00%
CMSOR-CI	91.61%	91.11%	91.30%	91.21%
CMSOR-CA	91.92%	90.99%	92.17%	91.58%
CMSOR-IA	87.45%	89.59%	89.95%	89.77%
CMSOR	**92.15%**	**92.67%**	**93.40%**	**93.34%**

Table 7: SemEval 2021 Task 7-1a数据集消融实验，加粗表示最优实验结果。

Model	ACC	P	R	F1
BILSTM	93.96%	93.82%	94.05%	93.93%
CMSOR-C	94.07%	94.80%	93.19%	94.28%
CMSOR-I	94.45%	96.66%	92.00%	94.08%
CMSOR-A	95.65%	96.30%	94.90%	95.59%
CMSOR-CI	95.72%	94.90%	96.59%	95.74%
CMSOR-CA	95.86%	94.69%	97.13%	95.89%
CMSOR-IA	96.21%	**96.77%**	95.61%	**96.19%**
CMSOR	**96.23%**	95.98%	**96.40%**	**96.19%**

Table 8: ColBERT数据集消融实验，加粗表示最优实验结果。

论：(1)当分别移除情感特征（**CMSOR**-IA）、模糊特征（**CMSOR**-CI）以及不一致性特征（**CMSOR**-CA）之后，模型在SemEval 2021 Task 7-1a数据集上的四项指标均有明显下降（F1值分别下降3.57%，2.13%，1.76%），这说明三种情感特征在幽默识别任务中的有效性。然而，在Pun of The Day数据集上，当移除情感特征后，模型在召回率R上有了提升，这可能是因为BERT在学习情感增强的文本表示时，将错误的情绪信息融入到语义表示当中，所以导致该指标下降。同时，这种情况还出现在ColBERT数据集上，原因同上。(2)当只保留模糊性特征的时候，模型在Pun of The Day和ColBERT数据集上的表现相比于**CMSOR**下降的最少，这说明模糊性特征在构建幽默语义过程中相比于情感特征以及不一致性特征更加重要。然而，对于SemEval 2021 Task 7-1a数据集，情感特征更加重要。

5 参数分析

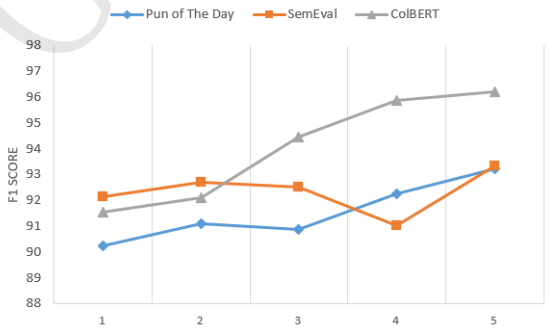


Figure 2: 不同数量的知识候选对模型性能的影响

图2展示了不用数量的常识信息对于模型性能的影响。从图中可以观察到，在Pun of The Day和ColBERT数据集上，当知识数量为1时，模型效果最差。随着候选知识数量的不断增加，模型的表现逐渐提升，并且在$l = 5$时取得最好的结果。这说明有效处理隐式情感表达对于**CMSOR**建模幽默语义具有重要作用，并且显式情感信息的增加会提升模型对于文本情感特

征的捕捉效果。对于SemEval 2021 Task 7-1a数据集而言，变化趋势于其他两个数据集不同。随着知识数量的增加，模型表现在略微提升之后，呈现出下降趋势，并且在$l = 4$时取得最差的结果，但是在$l = 5$时结果最优。这可能是因为当将知识数量增加到5时，一些样本的隐式情感表达才能够被COMET有效推理出来。

6 总结与展望

针对于现有幽默识别方法没有充分捕捉文本内部的情感特征，忽略了幽默文本中的隐式情感表达这一问题，本文提出一种动态常识与多维语义特征驱动的幽默识别方法**CMSOR**。该方法首先利用外部常识信息从文本中动态推理出说话者的隐式情感表达，然后引入外部词典WordNet计算文本内部词级语义距离进而捕捉不一致性，同时计算文本的模糊性特征。最后，根据上述三个特征维度构建幽默语义，实现幽默识别。本文在三个公开数据集上进行实验，结果表明本文所提方法**CMSOR**相比于当前基准模型有明显提升。未来，本文将尝试把常识信息应用到幽默生成、多模态幽默识别等任务当中。

参考文献

Issa Annamoradnejad and Gohar Zoghi. 2020. Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*, 1(3).

Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135.

Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski. 2019. Large dataset and language model fun-tuning for humor recognition. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4027–4032.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Andrew Cattle and Xiaojuan Ma. 2018. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1849–1858, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6752–6761.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana, June. Association for Computational Linguistics.

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

Yufeng Diao, Liang Yang, Dongyu Zhang, Linhong Xu, Xiaochao Fan, Di Wu, and Hongfei Lin. 2018. Homographic puns recognition based on latent semantic structures. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 565–576. Springer.

Xiaochao Fan, Hongfei Lin, Liang Yang, Yufeng Diao, Chen Shen, Yonghe Chu, and Yanbo Zou. 2020. Humor detection via an internal and external neural network. *Neurocomputing*, 394:105–111.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*.

Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. A computational model of linguistic humor in puns. *Cogn. Sci.*, 40(5):1270–1285.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Herbert M Lefcourt. 2001. *Humor: The psychology of living buoyantly*. Springer Science & Business Media.

Da Li, Rafal Rzepka, Michal Ptaszynski, and Kenji Araki. 2020. Hemos: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media. *Information Processing & Management*, 57(6):102290.

Lizhen Liu, Donghai Zhang, and Wei Song. 2018. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 586–591, Melbourne, Australia, July. Association for Computational Linguistics.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119. Association for Computational Linguistics, August.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538.

Rada Mihalcea, Carlo Strapparava, and Stephen Pulman. 2010. Computational models for incongruity detection in humour. In *Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010, Iaşi, Romania, March 21-27, 2010. Proceedings 11*, pages 364–374. Springer.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of english puns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–729.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Alex Morales and Chengxiang Zhai. 2017. Identifying humor in reviews using background text sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Copenhagen, Denmark, September. Association for Computational Linguistics.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. Uo upv: Deep linguistic humor detection in spanish social media. In *Proceedings of the third workshop on evaluation of human language technologies for Iberian languages (IberEval 2018) co-located with 34th conference of the Spanish society for natural language processing (SEPLN 2018)*, pages 204–213.

John Allen Paulos. 2008. *Mathematics and humor*. University of Chicago Press.

Victor Raskin. 1979. Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society*, volume 5, pages 325–335.

Lu Ren, Bo Xu, Hongfei Lin, and Liang Yang. 2021. ABML: attention-based multi-task learning for jointly humor recognition and pun detection. *Soft Comput.*, 25(22):14109–14118.

Lu Ren, Bo Xu, Hongfei Lin, Jinhui Zhang, and Liang Yang. 2022. An attention network via pronunciation, lexicon and syntax for humor recognition. *Applied Intelligence*, 52(3):2690–2702.

Antonio Reyes and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision support systems*, 53(4):754–760.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Graeme Ritchie. 2009. Can computers create humor? *AI Magazine*, 30(3):71–71.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Jerry M Suls. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. *The psychology of humor: Theoretical perspectives and empirical issues*, 1:81–100.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. *arXiv preprint arXiv:1909.00252*.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2367–2376.

Tongyue Zhang, Shaowu Zhang, Bo Xu, Liang Yang, and Hongfei Lin. 2021. 结合标签转移关系的多任务笑点识别方法(multi-task punchlines recognition method combined with label transfer relationship). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 238–247, Huhhot, China, August. Chinese Information Processing Society of China.

樊小超, 杨亮, 林鸿飞, 刁宇峰, 申晨, 楚永贺, and 张桐. 2021. 基于多维潜在语义特征的幽默识别. 中文信息学报, 35(8):38–46.

融合Synonyms词库的专利语义相似度计算研究

佟昕瑀[1]，廖佳伦[1]，路永和[2,*]
[1]中山大学信息管理学院, 广州, 中国, 510006
[2]中山大学人工智能学院, 珠海, 中国, 519082

摘要

一直以来，专利相似度计算和比较等工作都由专利审查员人工进行并做出准确判断。然而，以人工方式分析和研判专利的原创性、实用性以及是否侵权等工作需要投入大量的人力物力资源且效率较低。基于此，本文将ALBERT预训练模型用于专利的文本表示，并通过引入Synonyms近义词库增强专利文本的语义表达能力，探索一种基于语义知识库和深度学习的专利文本表示模型与相似度计算方法。实验结果表明，加入Synonyms近义词库消歧后的专利文本相似性度量的实验准确率有一定的提升。

关键词： 语义相似性；文本表示；自然语言处理；预训练模型

Patent Semantic Similarity Calculation by Fusing Synonyms Database

Xinyu Tong[1], Jialun Liao[1], Yonghe Lu[2,*]
[1]School of Information Management, Sun Yat-sen University, Guangzhou, China
[2]School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China

Abstract

Traditionally, patent examiners have relied on manual analysis to assess patent originality, practicality, and infringement, among other tasks related to patent similarity measurement and comparison. However, such analysis and judgment require significant human and material resources and low efficiency. Therefore, this paper utilizes the ALBERT pre-training model for text representation in patents and enhances the semantic expression capability of patent texts by introducing a Synonyms synonym library. It explores a patent text representation model and similarity calculation method based on semantic knowledge bases and deep learning. Experimental results demonstrate that incorporating Synonyms synonym library for disambiguation improves the accuracy of measuring patent text similarity.

Keywords: Semantic similarity , Text representation , Natural language processing , Pre-trained models

*路永和（通讯作者）：luyonghe@mail.sysu.edu.cn
本文系广东省重点领域研发计划项目“基于大数据智能的多层次知识检索关键技术研究及应用”（项目编号：2021B0101420004）的研究成果之一。

1 引言

专利作为科技文献的主要表现形式之一，从专利文本中可以获取当前创新技术，通过对专利文献进行分析，可以从中获得先进技术的细节，并进一步预测未来的技术发展趋势，以帮助相关群体进行关键的投资决策(Zhang et al., 2015)。在2021年的世界五大知识产权局（欧洲、日本、韩国、中国和美国）统计报告中指出，专利是用于保护发明创造的法律文件，专利数量被认为是衡量创新活动的重要指标，与日俱增的专利文献数量也导致了审查周期的延长(European Patent Office, 2023)。因此，各国专利局不断优化专利申请和审查程序以进一步缩短专利审查周期，从而达到鼓励创新、推动技术发展以及优化技术收益的目的。目前，专利授权以及专利侵权判定等大部分都是由专利审查员人工进行的，开展该项工作一方面需要具有不同领域背景的大量专利审查员，耗费较多的人力资源，另一方面人工判定耗时长且容易出现差错准确度难以保证(Liang and Tan, 2007)。

与其他文本相比较，专利文本是对某领域相关问题的创新深入研究，专业性强且会涉及大量领域词汇(Lee and Jeong, 2008)。专利文本的重点在于新技术的提出，因此，在技术的介绍部分往往会包含较多领域基础内容，在利用计算机判别时需要在文本的特征表达部分达到较高标准(Wen et al., 2021)。基于此，在判断专利文本的相似性时需要对领域词汇、生僻词和近义词进一步处理以提高文本的可读性。将知识库融入到专利文本相似性计算模型中，可以从文本语义理解的角度对专利文本进行更为准确的相似性计算。本文提出了一种基于Synonyms近义词库和ALBERT预训练模型的专利文本相似度计算方法，利用同义词库对专利文本的多义词、生僻词等进行解释或替换，使得文本成为更易于计算机神经网络理解的纯净文本，然后输入到预训练模型中，在下游任务中判断专利文本之间的相似度。

2 国内外研究现状

2.1 基于文本的专利相似度计算

基于文本内容的相似度计算方法中，最常用的方法是VSM模型，已有较多研究应用该方法或对其进行改进来计算专利文本的相似度。常用的方法是基于文本挖掘技术并结合专利关键词分析，计算专利文本元素的VSM加权相似度(Peng and Tan, 2010; Zhang et al., 2016; Arts et al., 2018)。

同样，Word2Vec和Doc2Vec模型也经常被用来表示专利文本，从而计算其相似度。Lee等人(2020)使用Word2Vec模型对特定产品领域的专利文本进行建模，并提出了一种产品潜在技术机会的分析方法。Xu等人(2018)用Word2Vec对专利文本进行训练和建模，并使用训练后的向量来计算专利领域词的相似度。Zhang等人(2018)以中文专利权利要求书为训练语料，用Doc2Vec模型计算专利权利要求文本的相似度，然后确定专利间的相似度。Cao等人(2018)使用Doc2Vec模型计算专利摘要相似度，并以此作为专利相似度参考。

在传统模型中加入语义信息可以有效提高模型计算专利技术相似度的准确率。具体方法包括设置不同的权重值以反映其他位置上的词的语义信息差异(Arts et al., 2021; Xia et al., 2018)，基于共享近邻聚类算法计算专利词的相似度(Jiang et al., 2016)，以及通过连接知识图谱中的专利术语和语义关系来优化专利相似度计算模型(Wang and Liu, 2022; An et al., 2021; Li et al., 2020)。通过融合专利文本特征的词位权重和领域相关性权重，并结合词向量加权方法和VSM文本表示法，将语义信息纳入专利相似度的测量中(Yu et al., 2019)。

在基于神经网络的专利相似度计算研究中，该研究将深度学习算法和专利模型树结合起来进行专利相似度计算，并采用连体LSTM算法(Mueller and Thyagarajan, 2016)进行专利应用，取得了良好的结果(Ma et al., 2018)。一些研究通过结合自然语言处理嵌入技术和最近邻相似性方法来衡量专利之间的技术相似性，将整个专利领域表现为一个技术网络(Hain et al., 2022)。基于文本的专利相似性计算方法已经发展得比较成熟，而结合深度学习的计算方法也不断提高了专利相似性计算模型的语义计算能力。

2.2 基于本体的专利相似度计算

基于本体的文本相似度计算方法的核心是基于人工构建的语义词典进行计算。目前，国际上使用最为广泛的语义词典是WordNet(Miller, 1995)，其包含了英文词汇与其实际意义的大量关系，展现了词与词之间的强大互联。中文词典方面，由梅家驹等人于1983年提出的《同义词林》(梅家驹, 1996)，以及其由哈工大进一步完善的扩展版本，是现在较为常用的语义

词典，其中包含了大量的中文词汇关系以及知识内容。此外，由董振东等人推出的HowNet（《知网》）(Dong and Dong, 2003)也是现在中文领域使用较多的语义词典。Synonyms近义词库的构建参照了《同义词词林》以及HowNet（知网）中所收录的词汇以及词汇之间的关系，但Synonyms相较于其他词典来说拥有更大的词汇量，对中文词义的解释更全面，并且更使用时更为方便。

刘影等(2011)通过对前人基于HowNet提出一种义原相似度的改进算法，实验结果表明，基于HowNet的方法可以和基于WordNet的方法取得一致的精确度；张思琪等(2017)基于WordNet本体，综合了大量前人的研究，改进了信息量IC计算模型，进而提出了两种混合式的语义相似度的计算方法，实验结果显示优于其他方法。Lu等(2023)在科技论文文本表示阶段融入WordNet词典，并以此作为后续引文推荐模型的输入，实验结果表明，融入WordNet的文本表示模型可以表达更多的语义信息，并且对后续模型输出有较好的提升效果。

现有研究证实，在文本表示阶段融入语义词典可以帮助模型学习到更多的语义信息，并且对多义词消歧有一定的帮助。基于此，本文将ALBERT预训练模型用于专利的文本表示，并通过引入Synonyms近义词库增强专利文本的语义表达能力，探索一种基于语义知识库和深度学习的专利文本表示模型与相似度计算方法。

3 专利文本表示及相似度计算模型构建

3.1 模型总体架构

本文构建的专利相似度研究模型由三个部分组成，分别是基于Synonyms的多义词消歧、基于ALBERT的文本表示和基于Softmax的相似性判断。模型结构如图1所示。

3.2 Synonyms近义词库

Synonym近义词库是一个开放在GitHub上的一个中文近义词工具包，它可以十分简便快捷的用于多种自然语言处理任务，并且对于不同领域的文本都具有很强的适应性，是一个通用型的近义词库。该工具包目前的词汇量已经达到了435,729个，包含了丰富的词语语义信息，主要可以根据其丰富的词汇量以及所包含的语义关系进行近义词检索等操作(Hai Liang Wang, 2017)。Synonym近义词库自发布以来进行了18次更新，最近一次更新时间为2022年5月5日。

随着英文语义词典高速发展，不断涌现了许多高质量词库，并且已有大量研究者将词典用于所研究的自然语言处理任务并对其进行泛化，加快了文本消歧、文本处理领域的发展，而纵观国内，相关的研究还偏少。因此，Synonyms近义词库的作者基于word2vec词向量训练了一个质量高、关系齐全的同义词库，将词语的表达规范化处理。在经典的信息检索系统中对于文档的检索都是基于严格的文字匹配操作得到的，查询算式与所匹配的文本之间字符必须是一一对应上的。而基于word2vec对大量的文本数据进行训练，可以获取到文本的上下文信息，获取上下文句子/词之间的语义关系，将词汇映射到低维的向量空间当中。因此，词汇之间的关系可以用词汇在向量空间中的距离来表示，那么度量相似性时也可以通过距离来度量。在Synonyms近义词库的构建中参照了《同义词词林》以及HowNet（知网）中所收录的词汇以及词汇之间的关系，但Synonyms相较于其他词典来说拥有更大的词汇量，对中文词义的解释更全面，并且更使用时更为方便(Hui et al., 2019)。图2是Synonyms近义词库运行时的实例，其中[$nearby_words$]是输入词的若干个近义词，以list的形式存储，并且按照与输入词的距离长度由近及远进行排列；[$nearby_words_score$]是每个近义词所对应的距离得分，该得分的区间为（0-1），越接近于1则说明与输入词越为接近；[$size$]为返回的词数量，在Synonyms中该值默认为10。

3.3 ALBERT预训练模型

自从BERT问世以来，越来越多的研究将研究目光转向了预训练模型，并且为了使得模型的训练效果更好，许多研究人员选择在模型中加大参数量以取得更好的模型表现。然而，如果无限制的去增加参数量则一定会带来一些附加问题，如参数量越大则模型训练的速度就越慢，对计算机性能的要求也越高，当模型的参数增加到一个临界点时不仅不能继续提升模型的效果，甚至会适得其反的因为参数的过大导致模型效果变差(Lan et al., 2020)。为解决模型参数量过大、内存占用过多等问题，ALBERT团队提出了因式分解嵌入层矩阵和跨层参数共享两种能够大幅减少预训练模型参数量的方法，此外还提出用Sentence-order prediction（SOP）任务代

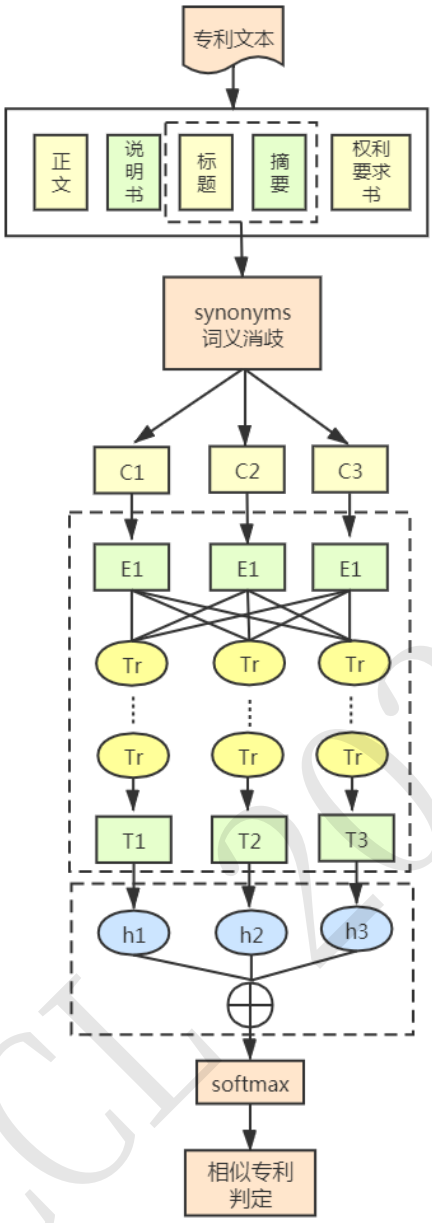


图 1: 基于Synonyms和ALBERT的专利相似度研究模型

```
Synonyms.nearby（人脸，10）=（
    [“图片”,“图像”,“通过观察”,“数字图像”,“几何图形”,“脸部”,“图象”,“放
     大镜”,“面孔”,“Mii”],
    [0.597284, 0.580373, 0.568486, 0.535674, 0.531835, 0.530095, 0.525344, 0.524009,
     0.523101, 0.516046]）
```

图 2: Synonyms近义词库实例

第二十二届中国计算语言学大会论文集，第341页-第351页，哈尔滨，中国，2023年8月3日至5日。

替BERT中的Next-sentence prediction（NSP）任务，经实验证实其在许多自然语言理解的任务和应用中取得了客观的结果。

因式分解嵌入层矩阵顾名思义可知该矩阵是对嵌入层的输入词向量做因式分解。词嵌入训练不仅仅只局限于某一部分的语境，而是将眼光放至了文本全局，使得最后生成的向量能够从本质上代表该词的全局语境。BERT 中的词嵌入层大小E和隐藏层的大小H相等，其隐藏层学习得到的是上下文相关向量。而在BERT的预训练模型中，可以将嵌入层大小与隐藏层大小根据实际任务要求和效果进行优化，比如将H设置的较大，则可以包含更多的信息，甚至可以根据需要设置为$H >> E$。此外，在自然语言处理任务中字典大小V通常达到上万，随着H的增大，而如果EaH则会导致词嵌入矩阵的无限扩大。为解决以上问题，ALBERT在BERT的基础上予以改进，通过引入数学因式分解的方式，把E和H分开设定，这样把原本较大的矩阵分解成了两个小矩阵，参数量有十分明显的减少，当$H >> E$的时候，参数削减更加明显。例如：$V = 30000$，$E = 128$，$H = 768$，则原参数量$V * H = 30000 * 768 = 23,040,000$，削减后$V * E + E * H = 30000 * 128 + 128 * 768 = 3,938,304$，参数变成了原来的1/6。

深度学习领域的重要参数压缩方式就是参数共享，其在很多深度学习神经网络中都有应用，不同的神经网络模型会根据自己实际的模型结构以及实现效果在不同的结构位置实现参数共享。而ALBERT是对Encoder中每一层之间参数共享，即多个层使用相同的参数，默认将全部的参数都进行共享，而不仅仅局限于前馈层或者注意力层。虽然该方式的性能相比于其他两种降低了1个百分点，但在参数量的减少上有着显著的效果。

在BERT的上下文判定中，不仅会考虑到两个句子之间的连贯性，还会考虑到两个句子的话题。BERT中所包含的NSP任务将注意力放在句子所表征的话题是否一致上，因此模型的判断是否为上下文任务变为了一个非常简单的主题匹配，而未考虑句子之间是否连贯，这使得NSP的任务无法学习更多的文本句子信息。因此，ALBERT中提出了SOP来取代NSP，具体来说，SOP分为正反两例，正序的判断方式与NSP一样，而反例则是将原本相邻的两个句子交换顺序，这样模型可以对句子的连贯性做出更多的吸收。在文本的实验中，ALBERT预训练模型充当着文本语义表示的角色，将专利文本输入至ALBERT模型后能够获得与其他模型相比更为丰富的语义表达信息。同时，也规避了大规模的参数导致模型占用大量计算资源。当使用ALBERT预训练模型判断两个句子在语义上是否相同时，可利用标注样本对预训练模型的参数进行微调以达到更好的效果，即通过持续的预训练不断学习数据集特征，优化下游任务结果。

4 实验及结果讨论

4.1 数据采集与预处理

4.1.1 数据采集

本文所爬取数据集来源于Google Patent上通信技术领域和文本处理领域的专利全文数据，所使用的编程语言是Python，并安装了selenium包，使用队列技术获取专利全文数据。在爬取过程中主要参考了专利审查员对于专利的标记识别，专利审查员标记的引用专利在技术上有较强的相似性(Chen, 2017; Lu et al., 2020)。除了专利审查员人工审查的内容之外，本文为扩展数据内容，将那些并没有被专利审查员判定为相似的专利对而Google自动判定为相似的专利对也纳入到了本文的数据获取队列中。当爬取的专利数量超过设定的阈值或者人工检测发现专利所涉及的主题与本文所确定的两个领域之间关联度很小的时候则立即停止对专利文本的爬取。本文一共爬取了通信技术和文本处理的两个专业领域共2620篇专利数据，构建实验数据集用于进行专利文本表示及相似度计算实验。表1是实验数据集相关字段含义描述。 由于本文的下游任务为专利相似性的判断，其为一个二元分类任务，只需要判断两个专利文本之间相似与否，因此在数据集的采集上获取了Similar字段，该字段所表示的含义为与该专利在内容上为相似专利的专利文本，该相似性判断由Google Patent所设置的程序和专利审查员共同进行判断，其中所判断的相似专利是与该专利不含有引用关系的专利文本。

4.1.2 数据预处理

文本预处理是后续进行文本表示的基础，预处理的效果会在很大程度上影响到后续相关任务的有效性和准确性。本文数据预处理中最重要的步骤是文本分词，为完成这一步，常用的方式时使用分词工具将中文文本分为最小字词，然后将其转换为词表中对应的数字id。为了取得

表 1: 实验数据集字段描述

字段名	字段含义
Patent_no	专利公开号，是表示专利唯一性的标识符之一。
Title	专利的标题。
Abstract	专利的摘要。
Citations	专利引用的专利列表，特指专利审查员的引用。
Cited	被引专利列表，即引用该专利的专利列表。
Similar	与该专利相似的专利列表，由Google Patent自动识别，不包含与该专利存在引用关系的专利。

更好的分词效果，本文首先使用NLPIR分词软件提取专利文本中的关键词，并人工筛选使用频率高、实际意义强的关键词加入用户字典辅助后续的分词过程。分词阶段采用了目前使用范围最广泛且操作方式简单的jieba分词系统。分词之后的文本数据中仍然包含了一些停用词和标点符号，这些对于我们的文本分析任务来说属于杂质，应当导入停用词表对其进行剔除。

为了便于后期的实验过程，针对本文的研究内容对专利文本进行具体处理。首先将每一个专利的编号、标题、摘要以及它相似文本的专利编号提取出来，并做简单的预处理工作。由于专利文本的标题通常较短，因此可以直接使用。而摘要部分句子比较长，但由于摘要的前半部分通常是包含了文本主题的主要信息，并且ALBERT模型的句子接受长度最多为512个字符，因此截选了摘要中的前512个字符输入至神经网络模型中。

4.2 实验流程及参数设置

本文进行专利文本表示和相似度研究构建了8组对比实验，包含了基于深度学习的神经网络模型及其对比模型。分别通过在专利文本表示及其相似度计算的各个环节采取不同的方法再进行组合从而得到本文所选择的对比实验模型。最终所得到的实验模型可归纳为两组，第一组实验为只使用ALBERT预训练模型和Softmax分类器对专利文本进行文本表示和相似度计算与使用传统向量空间模型进行文本表示和余弦相似度计算进行对比实验，第二组实验为基于Synonyms词库和ALBERT预训练模型以及Softmax分类器与上述传统模型进行比较实验。其中，将余弦相似度判断的阈值设定为0.8，若两篇文本余弦值大于0.8则判定其为相似文本，反之则不相似。每一组实验中专利文本的呈现方式都分为“标题”、“摘要”和“标题+摘要”三种。具体的对比实验流程图如下图3所示。

4.2.1 Synonyms文本消歧实验

文本消歧实验通过计算文本单词的tf-idf值加以人工筛选判断得到专利文本的前5个关键词，将关键词输入到Synonyms词典中并将其近义词追加到关键词后，由于有些词会有不止一个关键词，为了便于操作，本文选取synonyms中给定排序中的第一个近义词。筛选完成后将近义词添加到专业词汇后面。之所以采用直接添加的方式，是因为如果直接将原词替换成为其同义词可能会影响到原词在文本中所表达的意思使其不够准确。而直接在后面添加，该同义词与原词具有同样的意思表达，可以将原词的意思更为扩展便于计算机理解，并且不会影响到上下文的原意，保持句子的整体通顺。

4.2.2 ALBERT预训练模型构建

模型中的参数来源主要为在训练过程中自动获取，并且根据实际的模型运行情况结合人工经验对参数进行了必要的调整。表2列举了本文模型涉及的主要超参数类型及其描述。通过对已有ALBERT预训练模型的相关研究进行调研，以及在实验过程中反复测试调参，最终发现在本文的研究框架和内容下，每批训练数据设置为6时，可以与实验的输入长度512有较好的契合程度，并且对于数据集的遍历次数也可以达到适中的程度，模型能够达到一个较快的迭代收敛速度，同时对于计算机的内存消耗不大。实验设置迭代轮数为4，最大序列长度为512，在模型的优化器使用上选择了基于低阶矩阵估计的Adam，其实现过程限制较少，对其运行环境的性能要求不严苛，较为适合后续实验的调整和迭代。同时该算法还有利于训练的准确性和稳定性，防止其他干扰的杂质影响。

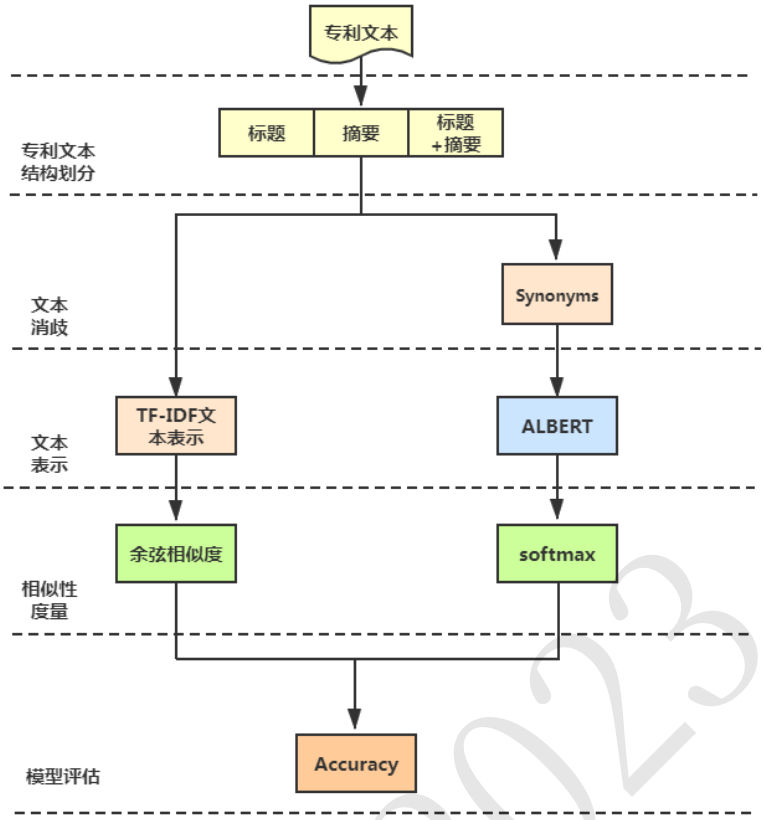


图 3: 实验模型及其对比模型

表 2: 实验参数

参数名称	参数设置	说明
Size	32000	所用词表大小
Embedding size	128	词嵌入层的大小
Hidden layer num	10	输入层的层数
Hidden size	768	输入层和池化层的大小
Hidden gro	1	Encoder的分割组数
Mul attention heads	12	多头自注意力层头数
Fully connected layer	3072	Encoder全连接层大小
Scaling	1	缩放比例
Activation function	gelu	激活函数
Dropout probability	0	全连接层dropout概率
Attention dropout probability	0	注意力层dropout概率
Max embeddings	512	最大长度
Vocab size	21128	类别数量
Train:Dev:Test	6:2:2	训练集、验证集和测试集比例

为防止过拟合的现象发生，本文在ALBERT实验代码编写过程中进行了一些调整以取得更好的实验效果。加入BatchNormld函数用以防止梯度消失，避免出现权重的过大或过小等情况；每训练50个import就会让学习率降低一定的参数值；虽然ALBERT模型本身为了减少对计算机内容的占用因此进行了移除了dropout的操作，但是实验中发现加入dropout函数对于防止过拟合存在着较好的效果，因此在本文实验过程中保留了dropout。

本文的实验所采用的系统为$Windows10_64$ 位，硬件配置为$Interi7-9700k$处理器、8GB内存，显卡型号为七彩虹$iGameGeForceRTX2080TiAdvance$。本文所有的实验所使用的编程语言均为Python，编写代码所使用的是Pycharm。其中神经网络模型部分由pytorch库中的函数搭建。

4.3 实验结果及分析

本文选择准确率accuracy来评价模型在专利文本相似度计算方面的效果。表3是具体的实验结果数据，共设置了两组对比实验，分别是：（1）ALBERT与传统的向量空间模型与余弦相似度组合算法的比较；（2）ALBERT与加入Synonyms近义词库的ALBERT模型对比。同时也比较了专利标题、摘要与标题+摘要对专利相似度分类实验的影响。

表 3: 实验结果

模型	准确率
title+tf-idf+cos	0.474763
abstract+tf-idf+cos	0.536277
title-abstract+tf-idf+cos	0.526813
title+ALBERT	0.824921
abstract+ALBERT	0.837539
title-abstract+ALBERT	0.850157
synonyms+ title+ALBERT	0.837838
synonyms+abstract+ALBERT	0.851351
synonyms+title-abstract+ALBERT	0.864864

从文本组合方面来看，用摘要部分进行文本表示和相似度匹配的准确率要比单用标题高，这说明专利文本的摘要部分包含了较多的主题信息，能够较为全面的反应出专利所表达的主题内容。而采用标题+摘要的方式所取得的效果则会更佳，这可能是因为标题和摘要之间互相补充互相配合更好的体现出了专利文本所描述的主题细节，同时两者的结合也能避免一个词存在多种意义而使得模型判断失误的情况出现。

通过第一组对比实验可以得出，将专利文本输入到ALBERT深度学习模型中进行文本表示以及相似度判断的准确率要远远高于使用传统的方式，具有30%左右的准确率提升，作为一个二分类的下游任务，ALBERT在实验采用的数据集上取得较好效果。究其原因可能在于专利文本具有较强的专业性和语言理解性，其文本中出现的内容通常为某一领域的新技术、新方法。而tf-idf文本表示方式是基于词频统计的模型，无法反映词与词、句子与句子之间的语义关系，难以有效捕捉到文本的主题信息。此外，余弦相似度的计算也只是基于向量进行简单的数学公式运算，其在计算过程中并不能考虑到相关的语义信息，导致使用tf-idf+余弦相似度来计算专利文本相似度的准确率较低模型效果较差。而使用ALBERT模型进行专利文本表示和相似度研究时，其能够通过自注意力机制关注到上下文的文本信息从而预测当前单词，可有效对文本进行表示。

从加入Synonyms近义词库消歧的结果来看，在标题、摘要、标题-摘要的三种文本组合与ALBERT预训练模型的实验中，加入了Synonyms近义词库做词义消歧的效果相比于未加入时准确率提升了1%左右，说明在加入外部知识库进行语义消歧确实能够对ALBERT预训练模型进行文本表示和下游文本相似度计算任务有一定的提升。但是从实验结果中看到，加入知识库后的提升并未达到最为理想的效果，可能有以下两点原因：（1）同义词替换与添加的方式可能会影响到句子原始语义的表达，甚至会在某些情况下反而影响到句子语言的通顺程度，不利于ALBERT预训练模型对文本特征的提取，因此对于ALBERT模型来说帮助并不大；（2）在

利用synonyms进行同义词替换和添加时并没有考虑到上下文语境，将导致多元的信息变得单一化，可能进一步模糊了句子间的边界。由于在知识库的使用上未能将其更加融入进神经网络模型的训练过程中，会在一定程度上弱化知识库的使用效果，从而导致整体的提升不大。本研究所得到的结果也可证明采用这种方式引入外部知识库仍然能够在文本语义的消歧和扩展上起到提升的效果。

5 总结

为了在大量的专利文本中找到其相似专利，便于研究者在开展研究前查询资料以及帮助专利审查员判定专利是否侵权，本文提出了一种结合语义信息和深度学习共同进行文本表示和相似度计算的方法，并验证了该方法在专利文本上应用的可行性和有效性。

本文选取了目前模型效果和性格各方面综合较为优秀的ALBERT预训练模型用于专利文本表示和相似度计算任务领域，通过ALBERT模型对专利文本特征的提取得到更为准确的专利文本表示，并连接Softmax分类器对下游的专利文本相似度计算任务做出更为精准的判断。在此基础上，引入开源的Synonyms近义词库在专利文本输入至ALBERT预训练模型之前对专利文本进行消歧以及语义扩展。最终通过采集谷歌专利库中文本处理和通信技术两个领域的专利文献验证实验效果，发现在使用Synonyms近义词库消歧后再将专利文本输入到ALBERT预训练模型中的实验效果比起只使用ALBERT预训练提升了1%，即对专利文本语义表达能力进行了一定程度的增强。

此外，实验部分根据专利文献的行文结构特点，截取了专利文本的标题和摘要两个部分进行文本组合，最终得到“专利标题”、“专利摘要”、“专利标题+摘要”三种文本组合，将其分别输入至实验模型中可以发现“专利标题+摘要”的效果要略好于其他两种组合方式，说明输入模型的文本长度可以影响模型特征提取效果。

综上所述，在专利文本表示及相似度计算领域引入预训练模型能够取得较好的实验效果，同时引入外部的知识库对文本的语义信息进行增强也能够对预训练模型起到一定的辅助作用，提升模型的效果。在未来对专利原创性、新颖性以及是否侵权等做出判断时，则可以将大部分工作任务转向由计算机程序自动处理，而减少对于人力资源的投入。

参考文献

Xin An, Jinghong Li, Shuo Xu, Liang Chen, and Wei Sun. 2021. An improved patent similarity measurement based on entities and semantic relations. *Journal of Informetrics*, 15(2):101135.

Sam Arts, Bruno Cassiman, and Juan Carlos Gomez. 2018. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84.

Sam Arts, Jianan Hou, and Juan Carlos Gomez. 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Qi Cao, Wei Zhao, Yingjie Zhang, Shujun Zhao, and Liang Chen. 2018. Comparative study of patent documents similarity detection on deep learning of doc2vec based methods. *Library and Information Service*, 62(13):74–81, 1.

Lixin Chen. 2017. Do patent citations indicate knowledge linkage? the evidence from text similarities between patents and their citations. *Journal of Informetrics*, 11(1):63–79.

Zhendong Dong and Qiang Dong. 2003. Hownet - a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824.

European Patent Office. 2023. Ip5 statistics report 2021 edition. https://www.fiveipoffices.org/sites/default/files/2023-01/IP5%20Statistics%20Report%202021_1.pdf.

Hu Ying Xi Hai Liang Wang. 2017. 中文近义词工具包synonyms.

Daniel S. Hain, Roman Jurowetzki, Tobias Buchmann, and Patrick Wolf. 2022. A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559.

Yueting Hui, Yijia Xia, Zihe Chen, and Xin Tong. 2019. Short text clustering algorithm based on synonyms, and k-means. *Computer Knowledge and Technology*, 15(1):5–6, 1.

Lixue Jiang, Duo Ji, and Dongfeng Cai. 2016. Measuring term similarity based on internal semantic role in patent text. *Journal of Chinese Information Processing*, 30(4):37–43, 1.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Bangrae Lee and Yong-Il Jeong. 2008. Mapping korea's national r&d domain of robot technology by using the co-word analysis. *Scientometrics*, 77:3–19.

Changyong Lee, Daeseong Jeon, Joon Mo Ahn, and Ohjin Kwon. 2020. Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database. *Technovation*, 96-97:102140.

Quanjia Li, Baoan Li, Xindong You, and XUeqiang Lyu. 2020. Computing similarity of patent terms based on knowledge graph. *Data Analysis and Knowledge Discovery*, 4(10):104–112, 10.

Yanhong Liang and Runhua Tan. 2007. A text-mining-based patent analysis in product innovative process. In Noel León-Rovira, editor, *Trends in Computer Aided Innovation*, pages 89–96, Boston, MA. Springer US.

Ying Liu, Li Chen, Zilin Song, Qingchao Dong, Xinghua Chen, Weixing Zhu, and Jixian He. 2011. An improved ontology-based method to measure similarity between concepts. *Journal of Nanjing University of Posts and Telecommunications(Natural Science)*, 31(6):60–66, 1.

Yonghe Lu, Xin Xiong, Weiting Zhang, Jiaxin Liu, and Ruijie Zhao. 2020. Research on classification and similarity of patent citation based on deep learning. *Scientometrics*, 123:813–839, 02.

Yonghe Lu, Meilu Yuan, Jiaxin Liu, and Minghong Chen. 2023. Research on semantic representation and citation recommendation of scientific papers with multiple semantics fusion. *Scientometrics*, 128:1367–1393, 01.

Chunyan Ma, Tong Zhao, and Hao Li. 2018. A method for calculating patent similarity using patent model tree based on neural network. In Jinchang Ren, Amir Hussain, Jiangbin Zheng, Cheng-Lin Liu, Bin Luo, Huimin Zhao, and Xinbo Zhao, editors, *Advances in Brain Inspired Cognitive Systems*, pages 633–643, Cham. Springer International Publishing.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2786–2792. AAAI Press.

Jidong Peng and Zongyin Tan. 2010. A text mining-based patent similarity measurement method and its application. *Information Studies:Theory Application*, (12):114–118, 1.

Zihong Wang and Y. Liu. 2022. Sea-ps: Semantic embedding with attention to measuring patent similarity by leveraging various text fields. *Journal of Information Science*.

Chaodong Wen, Cheng Zeng, Junwei Ren, and Yan Zhang. 2021. Patent text classification based on albert and bidirectional gated recurrent unit. *Journal of Computer Applications*, 41(2):407–412, 2.

Bin Xia, Baoan Li, and Xueqiang Lyu. 2018. Calculation of patent text similarity based on word location and semantic information. *Computer Engineering and Design*, 39(10):3087–3091, 1.

Kan Xu, Yuan Lin, Chen Qu, Bo Xu, and Hongfei Lin. 2018. Research on patent query expansion methods using word embedding. *Journal of Frontiers of Computer Science Technology*, 12(6):972–980, 1.

Yan Yu, Lei Chen, JInde Jiang, and Naixuan Zhao. 2019. Measuring patent similarity with word embedding and statistical features. *Data Analysis and Knowledge Discovery*, 3(9):53–59, 1.

Haichao Zhang and Liangwei Zhao. 2018. Judge chinese patents similarity based on doc2vec. *Technology Intelligence Engineering*, 4(2):64–72, 1.

Longhui Zhang, Lei Li, and Tao Li. 2015. Patent mining: A survey. *SIGKDD Explor.*, 16:1–19.

Yi Zhang, Lining Shang, Lu Huang, Alan L. Porter, Guangquan Zhang, Jie Lu, and Donghua Zhu. 2016. A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics*, 10(4):1108–1130.

Siqi Zhang, Weiwei Xing, and Yuanyuan Cai. 2017. A wordnet-based hybrid semantic similarity measurement. *Computer Engineering and Science*, 39(5):971–977, 1.

梅家驹. 1996. 同义词词林. 上海辞书出版社.

中医临床切诊信息抽取与词法分析语料构建及联合建模方法

王亚强[1,2,3†]，蒋文[1,2,3]，蒋永光[4]，舒红平[1,3]
[1]成都信息工程大学软件工程学院
[2]成都信息工程大学数据科学与工程研究所
[3]软件自动生成与智能服务四川省重点实验室
[4]成都中医药大学基础医学院
[†]通讯作者：yaqwang@cuit.edu.cn

摘要

切诊是中医临床四诊方法中极具中医特色的疾病诊察方法，为中医临床辨证论治提供重要的依据，中医临床切诊信息抽取与词法分析研究具有重要的临床应用价值。本文首次开展了中医临床切诊信息抽取与词法分析语料构建及联合建模方法研究，以万余条中医临床记录为研究对象，提出了一种语料构建框架，分别制定了中医临床切诊信息抽取、中文分词和词性标注语料标注规范，形成了可支撑多任务联合建模的语料，语料最终的标注一致性达到0.94以上。基于同级多任务共享编码参数模型，探索了中医临床切诊信息抽取与词法分析联合建模方法，并验证了该方法的有效性。

关键词： 中医临床切诊信息；信息抽取；词法分析；语料构建方法；多任务学习

On Corpus Construction and Joint Modeling Method for Clinical Pulse Feeling and Palpation Information Extraction and Lexical Analysis of Traditional Chinese Medicine

Yaqiang Wang[1,2,3†], Wen Jiang[1,2,3], Yongguang Jiang[4], Hongping Shu[1,3]
[1]College of Software Engineering, Chengdu University of Information Technology
[2]Institute for Data Science and Engineering, Chengdu University of Information Technology
[3]Sichuan Key Laboratory of Software Automatic Generation and Intelligent Service
[4]Department of Preclinical Medicine, Chengdu University of Traditional Chinese Medicine
[†]Corresponding author：yaqwang@cuit.edu.cn

Abstract

Pulse feeling and palpation (PFP) are the most distinctive and representative clinical diagnostic methods of traditional Chinese medicine (TCM). They provide important evidence for TCM clinical practitioners to differentiate syndromes. Extracting PFP information and analyzing its lexical features have important clinical value. In this paper, we carried out research on corpus construction and joint modeling method for PFP information extraction (IE) and lexical analysis (LA) of TCM for the first time. Based on more than ten thousand TCM clinical records, we proposed a corpus construction framework, built annotation guidelines and constructed a labeled PFP IE and LA corpus to support multi-task learning. Labeling consistency evaluated by inter-annotator agreement value for the corpora achieve more than 0.94. Moreover, we attempted to jointly model PFP IE and LA tasks of TCM with a same-level-shared multi-task model, and experimental results verified effectiveness of the joint model.

Keywords: Clinical pulse feeling and palpation information of traditional Chinese medicine , Information extraction , Lexical analysis , Corpus construction method , Multi-task learning

1 引言

辨证论治是中医学认知和治疗疾病的基本原则，四诊合参是中医辨证的基本需要(李灿东, 2021)。切诊是中医临床四诊方法[1]中极具中医特色的疾病诊察方法，是中医专家经过长期的临床实践，逐步形成并不断补充完善建立起的诊察技术，为中医临床辨证提供重要的诊断信息(谭同来, 2010)。

切诊包括按诊和脉诊两种方法。其中，按诊是中医专家用手对患者体表特定部位进行触、摸、按、叩，通过观察患者的反应，探明疾病的部位、性质和程度的方法；脉诊是中医专家运用手指切按患者的脉搏动处，体验脉动应指的形象，以确定全身脏腑功能、气血、阴阳的协调能力状况综合信息的方法(谭同来, 2010)。它们的诊察结果通常记录在中医临床记录中。

抽取中医临床记录中的切诊描述，获取其中蕴含的浅层词法信息，将为中医临床辅助辨证、中医临床医案分析等下游任务提供丰富的医学语义信息。如图1所示，由方位词“左”、专有名词“脉”和形容词“细”所构成的脉诊描述，说明了患者的左心室收缩和舒张功能存在异常，推断患者可能“气血亏虚”(杨杰, 2006)。

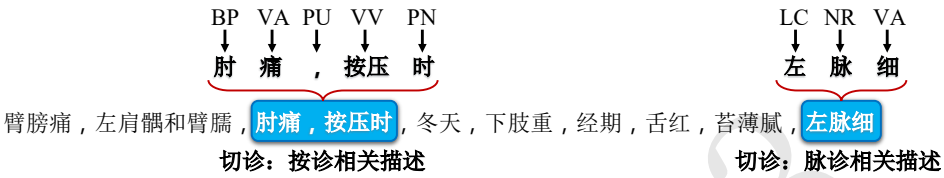


图 1. 中医临床记录中的切诊描述及其包含的浅层词法特征信息

中医临床记录中实体及其关系的信息抽取研究已广泛开展。Wang等人(2014)开展了基于统计序列标注模型的中医临床症状信息抽取研究。Ruan等人(2020)开展了基于深度学习的中医临床“病-证-药”实体关系抽取研究。此外，中医临床四诊信息抽取的研究尚处在起步阶段，2022年王亚强等人(2022)首次开展了中医临床记录四诊描述抽取研究。目前，中医临床切诊信息抽取与词法分析的研究尚未见相关报道。

中医临床切诊描述具有其领域特殊性和文本描述方式的独特性：

首先，切诊包括按诊和脉诊两种方法，两种方法的描述方式各有不同。按诊描述倾向叙述过程，用语与其它三诊描述类似，如图1中的“肘痛，按压时”，叙述了“按压过程患者肘部的疼痛状态”，且其中包含的词语均在其它三诊中常见。脉诊描述倾向描述位置和状态，如图1中的方位词“左”和形容词“细”，但其用语也有特殊性，如“脉”是脉诊描述中的专有名词。

其次，由于中医临床切诊描述的口语化特点，且通常采用简短的短语或短句描述，使其具有较强的稀疏性。如图1中所示，按诊描述“按压时肘痛”被口语化地记录为“肘痛，按压时”。此外，在本文构造的语料中，切诊信息简短，文本长度平均包含2.97个字[2]。

第三，中医临床记录具有简短性和独特的语言特色(Wang et al., 2012)，与一般领域相比，中医临床切诊描述中的词和词类定义存在不同。例如，简短的脉诊描述“脉滑”表达了“脉”往来流利，应指圆“滑”的脉象状态，其中包含“脉”和“滑”两个单字词，并且“脉”是脉诊描述中的专有名词，“滑”被用作形容词，这些词法定义在一般领域中并不常见。

最后，中医临床切诊信息抽取与词法分析任务存在类别分布不均衡的问题。在本文构造的语料中，中医临床记录的平均长度为38.90个字，而中医临床切诊描述包含的字数平均占比仅为9.30%。此外，中医临床切诊描述中长度大于2的词远少于词长为1的词数（如图2所示）。并且如图3所示，词类标签的计数分布同样存在长尾现象。

这些中医领域的特殊性和中医临床切诊描述的独特性，给中医临床切诊信息抽取与词法分析带来巨大挑战。因此，本文首次开展了中医临床切诊信息抽取与词法分析的研究，主要贡献包含以下三个方面：

1. 以“宾州中文树库分词指南”(Xia, 2000a)为基础，根据中医领域的特殊性和中医临床记录描述的独特性，建立了“中医临床切诊描述的中文分词指南”。此外，通过融合“信息处理用现

[1]中医临床四诊方法包括“望诊”、“闻诊”、“问诊”和“切诊”四种方法(李灿东, 2021)
[2]本文将中医临床记录中的标点符号也视为“字”

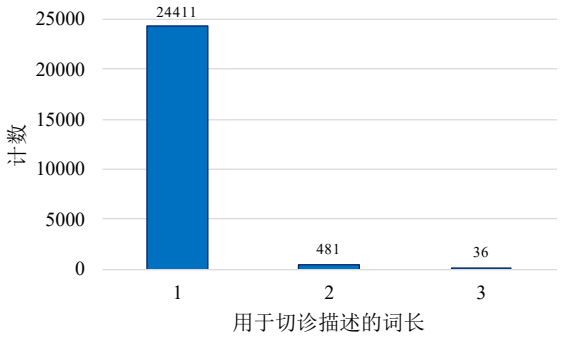


图 2. 中医临床切诊描述的不同词长计数

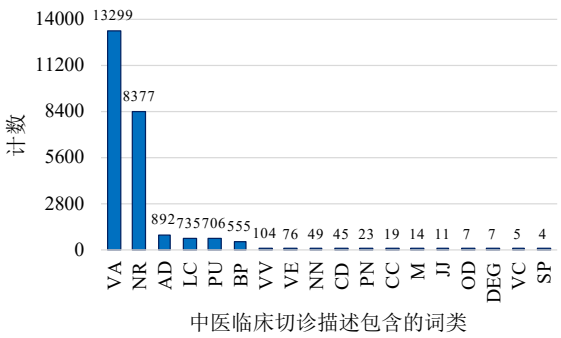


图 3. 中医临床切诊描述包含的词类计数

代汉语词类标记规范”[3]、“中文电子病历词性标注规范”(杨锦锋et al., 2016)等规范，结合中医领域知识，建立了“中医临床切诊描述的词类标注指南”。

2. 提出了一种语料构建框架，以贡献1中的两份“指南”为基础，聘请专家对前期收集的10594条中医临床记录进行了切诊描述标注、中文分词标注和词性标注，构建了可用于中医临床切诊信息抽取和浅层词法分析的多任务联合抽取语料。首次开展了面向中医临床切诊描述的中文词法分析研究，为中医临床切诊描述的语法分析以及面向中医临床记录的深层语义分析奠定良好基础。

3. 基于构建的中医临床切诊信息抽取和词法分析语料，本文将中医临床切诊信息抽取、中文分词和词性标注采用同级多任务共享编码参数方式进行多任务学习建模，开展了中医临床切诊信息抽取、中文分词和词性标注联合信息抽取研究初探，验证了多任务学习方法在中医临床切诊信息抽取、中文分词和词性标注联合信息抽取方面的有效性，发现了新问题，为未来的进一步研究指明了方向。

实验结果表明，采用本文提出的标准化语料构建框架（参见第3章描述）分别构建中医临床切诊信息抽取、中文分词和词性标注语料，均能使标注专家对标注任务的理解快速达到一致，各语料最终的标注一致性IAA（Inter-Annotator Agreement(Ron and Massimo, 2008)）值均达到0.94以上（为强一致性标注结果(Fleiss, 1981)）。该结果说明，本文设计的标注指南正确，语料构建质量高。此外，基于同级多任务共享编码参数方法构建中医临床切诊信息抽取、中文分词和词性标注联合抽取模型，能提升各任务的抽取性能，F值最高提升了2.2%（在最复杂的中医临床切诊词性标注任务上取得），说明了多任务联合建模方法是未来开展中医临床切诊信息抽取、中文分词和词性标注研究的重要方向。

2 相关工作

2.1 中医临床信息抽取

近年来，中医临床信息抽取研究受到广泛关注。Zhang等人(2022)对2010年至2021年间，中医文本信息抽取任务进行了综述，中医临床信息抽取是其中最重要的任务之一。中医临床信息抽取主要围绕中医临床记录中的实体（如疾病、症状、体征等(Liu et al., 2015; Guan et al., 2021)）信息抽取任务展开，少量研究关注实体间的关系抽取任务(Bai et al., 2022)。针对中医临床记录所包含的语言学信息（如词法信息(Jiao et al., 2018)）和临床语义信息（如中医临床四诊信息(李灿东, 2021)）抽取的相关研究较少。近期，王亚强等人(2022)首次开展了中医临床记录四诊描述抽取的研究。本文在此基础上，进一步围绕极具中医特色的中医临床切诊信息抽取与词法分析展开探索。

面向中医临床记录的词法分析研究尚未见报道，更不用说针对中医临床切诊描述的词法分析研究。通用领域的研究发现，词法分析的结果可作为补充信息，为下游任务提供丰富的语言学特征(Sagot and Alonso , 2017)，并有助于深层语义分析研究的开展(Guo et al., 2016; Kurita et al., 2017)。然而，由于中医领域的特殊性和中医临床记录描述的独特性，直接应用已有的

[3]URL: http://www.moe.gov.cn/ewebeditor/uploadfile/2015/01/13/20150113085826365.pdf

通用词法分析模型(Jiao et al., 2018; Sun et al., 2009)无法有效开展中医临床切诊描述的词法分析。因此，本文围绕中医临床切诊描述的词法分析进行了初探研究。

2.2 中医临床语料构建

中医临床信息抽取通常采用有监督学习方法实现(Zhang et al., 2022)，中医临床切诊信息抽取与词法分析研究需要带标注的语料库支撑。中医临床四诊信息抽取的研究处于起步阶段，支撑中医临床切诊信息抽取研究的语料初具规模(王亚强et al., 2022)。而中医临床切诊描述的词法分析研究尚未开展，由于中医领域的特殊性和中医临床记录描述的独特性，迫切需要利用中医临床记录构建语料以支撑相关研究。本文在前期构建的中医临床四诊信息抽取语料(王亚强et al., 2022)的基础上，构建了中医临床切诊信息抽取与词法分析语料。

语料构建耗时、费力，中医临床切诊信息抽取与词法分析语料的构建还需要中医专家的参与，这也为该任务提出了挑战。Zhang和Wang等人(2020)提出了一种面向中医临床信息抽取的语料构建框架，能有效提升构建的效率和质量。因此，本文针对中医临床切诊信息抽取与词法分析任务，改进了该语料构建框架，与中医专家共同探讨并制定了各任务的语料标注规范，构建了中医临床切诊信息抽取与词法分析语料，实验结果表明，本文所构建的语料质量高。

2.3 多任务联合信息抽取

多任务学习是一种通过相关任务之间共享表示，以提升原始任务模型泛化能力与性能的方法(Crawshaw, 2020)。中医临床切诊信息抽取、中文分词和词性标注是三项典型的相关任务，均可以采用序列标注方法建模(王亚强et al., 2022; Jiao et al., 2018)，三项任务之间具有共同的输入（即中医临床记录），可以共享输入的嵌入表示和中间层编码表示信息，最后根据任务的不同预测序列标注输出。因此，中医临床切诊信息抽取与词法分析可以采用基于多任务学习的序列标注方法(Rei, 2017; Lin et al., 2018; Fang et al., 2023)联合建模。本文将中医临床切诊信息抽取、中文分词和词性标注三项任务视为同级任务，借鉴Pham等人(2019)提出的“同级任务共享模型”对三项任务联合建模，实现三项任务通过共享表示的方式，促进各任务对应的序列标注模型的泛化能力和标注性能的提升。

3 语料构建方法

中医临床切诊信息抽取与词法分析采用有监督序列标注方法实现，有监督模型需要高质量的带标注语料的支撑。而中医临床切诊信息抽取与词法分析的研究尚处于起步阶段，相关研究所需的高质量的带标注语料尚有待构建。因此，本文基于前期研究工作所构建的中医临床四诊信息抽取语料，进一步的构建中医临床切诊信息抽取与词法分析语料。

3.1 构建方法框架

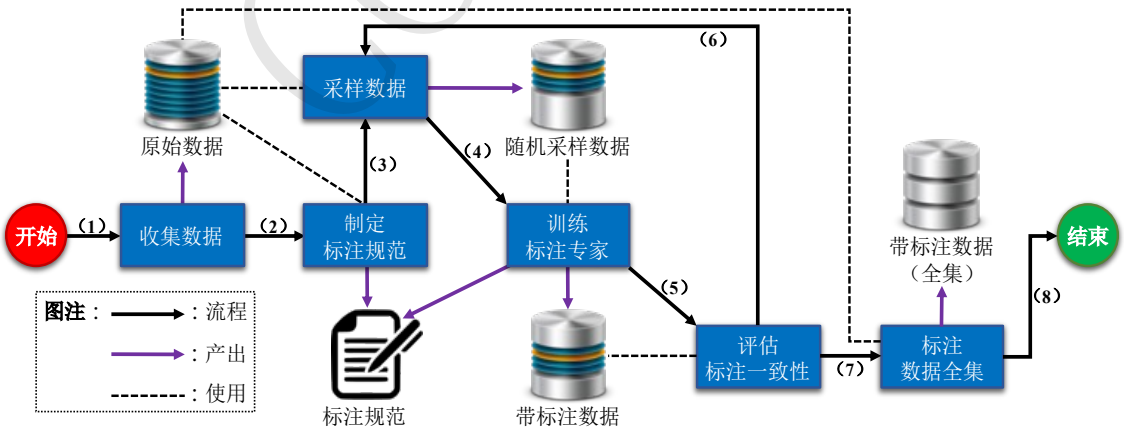


图 4. 中医临床记录切诊信息抽取与词法分析语料构建方法框架

本文借鉴Zhang和Wang等人(2020)提出的语料构建框架，分别用于中医临床切诊信息抽取、中文分词和词性标注语料的构建过程，构建方法框架如图4所示。框架中包含六个关键步骤，包括:

(1) 收集数据：根据任务需要，从中医专家日常诊疗过程中收集的中医临床记录，随机的选取指定数量的临床记录，形成原始数据。本文构造语料的10594条中医临床记录即从中医专家日常诊疗过程中收集的数据随机抽取得到。

(2) 制定标注规范：根据任务需要，以原始数据为研究对象，标注专家共同探讨，制定新的标注规范，或者借鉴并改进已有的同类任务的标注规范，制定适用于本任务的标注规范。本文采用了后者方式，分别制定了中医临床切诊信息标注规范、中文分词标注规范和词性标注规范。通常，在本步骤形成的标注规范为初稿，后期会随着对任务和数据的深入理解，对标注规范进行迭代更新。

(3) 采样数据：从原始数据中随机采样N条中医临床记录数据用于后续步骤(4)训练标注专家。通常N为一个较小的自然数，本文中$N = 100$。如需重复采样数据来反复训练标注专家，一般采用无放回采样方法，以保证未来训练时可遇见相似但不相同的待标注数据。

(4) 训练标注专家：参与制定标注规范的三位标注专家，根据当前任务的标注规范，独立地对相同的无标注随机采样数据进行标注，得到带标注数据。此外，在标注的过程中，标注专家如发现新的标注规则，将记录并在任务结束后与其他专家讨论，确认后更新标注规范。

(5) 评估标注一致性：采用IAA度量各标注专家的标注一致性，其结果用于评价专家在当前任务上的认知一致性。当IAA值小于阈值α时，将重复步骤(3)至(5)。当IAA值连续三次大于α时，认为各专家在未来独立的标注任务中，能够以相同的标准，保质地完成数据标注任务，则进入最后的第(6)步骤。对IAA的解释及本文设置参见章节3.3。

(6) 标注数据全集：各专家在达到对标注任务的认知一致后，原始数据将平均分配给各标注专家完成数据标注任务。一般地，各份数据会包含N'条相同的数据，用于计算在构建最终的带标注数据全集时，各专家的标注一致性，以确保最终的语料的质量。通常N'也是一个较小的自然数，本文中$N' = 100$。

3.2 标注规范

本文以中医临床切诊描述为案例，融合、重写和扩展同类任务的标注规范，并在语料构建方法框架中迭代完善所形成的中医临床切诊信息标注规范、中文分词标注规范以及词性标注规范。各规范主要的扩展原则如后文所述，标注规则具体内容参见(词法分析相关规范, 2023)。

3.2.1 切诊信息标注规范

本文沿用了前序工作(王亚强et al., 2022)中制定的中医临床四诊信息标注规范及其数据，构建形成了中医临床切诊信息抽取语料。在前序工作中，已形成中医临床四诊信息抽取语料，本文重点关注四诊中最具中医特色的切诊信息抽取，因此，在中医临床四诊信息抽取语料的基础上，本文统一将该语料中的其它三诊（即望诊、闻诊和问诊）的标注直接删除，最终形成中医临床切诊信息抽取语料。

3.2.2 中文分词标注规范

根据“中医四诊操作规范第4部分：切诊”国标(李灿东et al., 2021)对中医临床切诊相关概念的定义，三位标注专家融合通用领域的“信息处理用现代汉语分词规范”(靳光瑾et al., 2006)与“宾州中文树库分词指南”(Xia, 2000a)，以中医临床切诊描述为案例，重写并扩展制定了中医临床切诊描述的中文分词标注规范。

与通用领域的分词规则不同，中医临床切诊描述的中文分词规范以能够获取细粒度中医临床语义信息(Zhang et al., 2020)为成词的基本原则。因此，本文将通用领域中对人名、组织或国家名称等实体类描述的分词规则进行了修改，扩展制定了对中医切诊术语的细粒度分词规则。例如，以“脉细”为例，根据“宾州中文树库分词指南”的规则，“脉细”作为中医学实体不会被进一步分词，但为获取其中包含的细粒度中医学语义信息，“脉细”将进一步地切分为“脉”和“细”两个词语，进而为下游任务获取“脉”的状态为“细”这一细粒度中医临床修饰语义信息形成铺垫。

在中医临床切诊描述中，常见关于“大小”的描述内容，例如“周围硬块约5*5cm”，其中，“5*5”表达了中医专家通过中医按诊后，获得并记录的硬块大小信息。尽管“5”为数

词，“*”表示数学符号，在通用领域的规范中，它们均可以单独成词，并表示更细粒度的大小测算语义信息，但是为保留中医专家判断硬块大小的中医临床语义信息，将“大小”这类通常由几个通用领域的词语组成的字符串定义为词，不作细粒度切分。

3.2.3 词性标注规范

根据“中医四诊操作规范第4部分：切诊”国标(李灿东et al., 2021)对中医临床切诊相关概念的定义，三位标注专家融合通用领域的“信息处理用现代汉语词类标记规范”(靳光瑾et al., 2006)和“宾州中文树库词性标注指南”(Xia, 2000b)，以中医临床切诊描述为案例，重写并扩展制定了中医临床切诊描述的词性标注规范。

本文以语法功能作为划分词类的主要依据，并在标注的过程中，结合上下文语义综合判别词类。例如，“停”常用含义为“停止、停下”，表示“动作、行为和心理状态”，因此为动词。然而，在中医临床记录“7次1停”中，结合上下文医学语义信息，判定“停”在此处的语法功能为“动词量词”，因此被归为“量词”词类。

根据中医临床切诊描述中包含的词语特点，将专有名词等17种通用领域常见词类（表1词性标注的标签名称）保留在中医临床切诊描述的词性标注规范中。此外，单独新增了“身体部位”词类，具有表达“切按部位”以及发现“病症的部位”的中医临床语义信息，与通用领域的名词加以区分的目的。例如，在切诊描述“手热额冷”中，身体部位“手”和“额”分别体现了患者“热”和“冷”病症部位，因此需要将这类具有中医临床语义信息的词与其他通用领域的名词加以区分。

3.3 标注一致性度量方法

在构建方法框架中，IAA具有重要作用。一方面，IAA值用于衡量多位标注专家在进行相同标注任务时，对该任务认知是否达到一致。另一方面，IAA值也间接的用于评价所制定的标注规范的质量，即用多位标注专家在相同的标注规范指导下是否能够保持对相同任务的认知一致，来说明标注规范的全面性、代表性和准确性。

本文采用了Fleiss' Kappa方法来计算IAA值，其具体计算方式参见(Fleiss, 1971)。IAA值越大，说明多位标注专家在完成相同标注任务时，对该任务的认知歧义越低，即认知一致性越强，说明未来多位标注专家在完成该类任务时，能够做出一致的判断，从而保证标注的质量。反之亦然。通常，$IAA \geq 0.75$表示一致性优秀(Fleiss, 1981)。本文在构建方法框架中，将IAA的达标值α设置为0.85。并且，为避免偶然性，本文在构建方法框架中的步骤(5)中要求IAA值连续三次大于α，才能进入步骤(6)。

4 多任务联合信息抽取框架

中医临床切诊信息抽取、中文分词和词性标注是三项相关任务，它们在对输入数据的编码层面具有可共享的信息，解码后获得不同任务各自的输出结果。这是典型的多任务学习应用场景，通过相关任务之间的信息共享，提升原始任务模型泛化能力与性能(Crawshaw, 2020)。作为初探，本文自然地将Pham等人(2019)提出的“同级任务共享模型”（Same-level-Shared Model）应用到中医临床切诊信息抽取、中文分词和词性标注三项任务的联合信息抽取建模，模型框架如图5所示。

中医临床切诊信息抽取、中文分词和词性标注的多任务联合信息抽取可自然地基于序列标注方法建模。如图5所示，给定输入中医临床记录字序列$\boldsymbol{w} = \{w_1, w_2, \cdots, w_n\}$和输出标注序列$\boldsymbol{l}_j = \{l_{j,1}, l_{j,2}, \cdots, l_{j,n}\}$。其中，$n$为$\boldsymbol{w}$包含字$w_i$的数量及标签序列$\boldsymbol{l}_j$的长度，$j \in (1, T)$，$T$为多任务联合信息抽取框架中任务的数量，通常大于2，本文中$T = 3$。$w_i \in V$，V为训练数据中字符的集合，$l_{j,k} \in L_j$，L_j表示任务j的标签集。在本文实验数据中，字符包括中文字、标点、数字，中医临床切诊信息抽取（$j = 1$）、中文分词（$j = 2$）和词性标注（$j = 3$）任务的标签集如表1所示。

在联合信息抽取框架中，主要包括两部分。一是编码层，实现多任务共享的中医临床记录输入的向量表示（字的$\boldsymbol{w}_i$的独热表示$\boldsymbol{x}_i$和嵌入$\boldsymbol{e}_i$）变换和中间信息融合编码（$\boldsymbol{h}_i$）学习。二是解码层，根据任务的不同，综合利用共享的编码信息$\boldsymbol{h}_i$，通过线性变换形成序列标注推理模型的输入，最终输出预测的序列标注结果。

本文中，中医临床记录输入的嵌入表示变换采用了通用的BERT(Devlin et al., 2018)实现；为实现上下文信息融合，中间信息融合编码采用了BiLSTM(Shin and Lee, 2019)；各任务的序

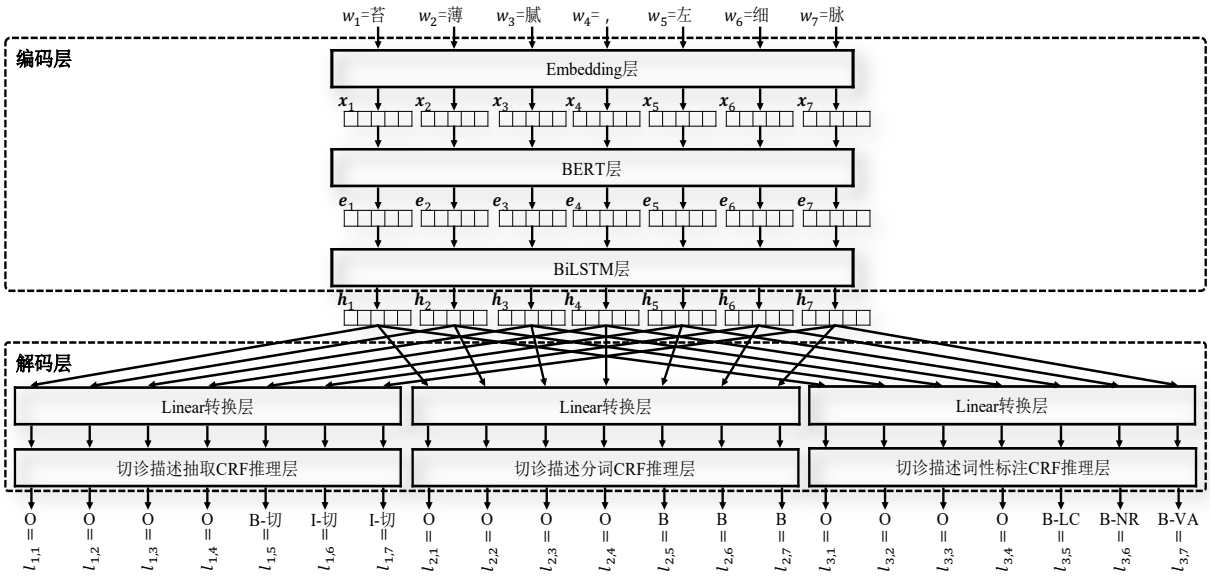


图 5. 中医临床切诊信息抽取、中文分词和词性标注联合信息抽取框架

任务名称	标签数量	标签名称(类型)
信息抽取	3	切诊描述开始位置(B-切), 切诊描述中间位置(I-切), 其他位置(O)
中文分词	3	词语开始位置(B), 词语中间位置(I), 其他位置(O)
词性标注	37	专有名词(NR), 普通名词(NN), 能愿动词(VV), 有字动词(VE), 系动词(VC), 表语形容词(VA), 名词修饰语(JJ), 基数词(CD), 序数词(OD), 量词(M), 副词(AD), 代词(PN), 方位词(LC), 身体部位(BP), 并列连词(CC), 属格标记(DEG), 句末助词(SP), 标点符号(PU), 其他(O) （**说明**：每种词类有“B”和“I”两种标签，分别表示该词类标签的开始和中间位置）

表 1. 中医临床切诊信息抽取、中文分词和词性标注任务的序列标注标签集定义

列标注结果预测采用了CRF（Conditional Random Fields(Ma and Hovy, 2016)）模型实现。

多任务学习优化的目标是最小化多个任务的加权损失函数，同时保证模型参数在多个任务中具有共享性。因此，多任务优化的目标函数为：

$$L_{global} = \arg\min_{\hat{\boldsymbol{\theta}}_1,\cdots,\hat{\boldsymbol{\theta}}_T} \sum_{j=1}^{T} \beta_j \cdot L_j(\boldsymbol{\theta}_j)$$

其中，β_j为超参数，表示任务j的损失E_j在多任务条件下的全局损失L_{global}中的贡献度，$\sum_{j=1}^{T} \beta_j = 1$。

5 实验

5.1 实验数据与模型设置

本文所构建的中医临床切诊信息抽取与词法分析语料是基于中医专家日常诊疗过程中收集的中医临床记录完成，共有10594条数据，包含了412123个字，字典大小为2453。其中，最短和最长的中医临床记录分别包含2个和311个字。中医临床记录的长度及该长度出现的频数关系如图6所示。从图中可以看出，大部分中医临床记录的长度集中在11个字至62个字之间，属于短文本。此外，中医临床记录中，片段描述之间通常采用逗号分隔，简单地以中文逗号和英文逗号为分隔统计，共包含片段描述91023条，片段描述的长度及该长度出现的频数关系如图7所示。从图中可以看出，大部分中医临床记录的片段描述的长度集中在1个字至9个字之间，属于短文本。这些特点进一步使得中医临床切诊信息抽取及词法分析联合信息抽取任务具有挑战。

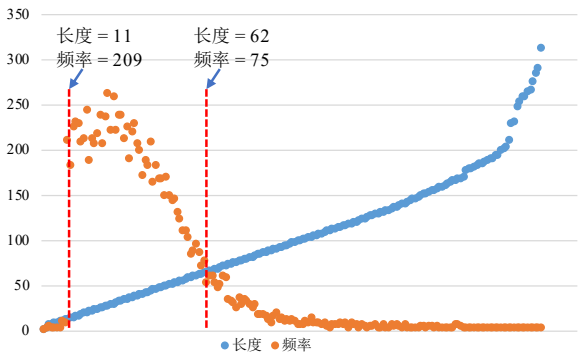


图 6. 中医临床记录的长度及该长度出现的频数关系

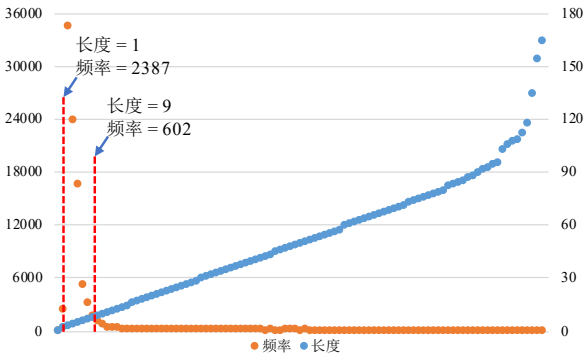


图 7. 中医临床记录中片段描述的长度及该长度出现的频数关系

本文的多任务联合信息抽取框架是基于BERT+BiLSTM+CRF改造，如图5所示，负责编码的BERT层和BiLSTM层为参数共享部分，CRF推理层分别执行各标注任务。框架的训练采用AdamW优化器，其超参数β_1和β_2分别设为0.9和0.999，为避免过拟合，Dropout设置为0.1，批量大小设置为32，Epochs设置为100，学习率设置为5e-5，最大句子长度设置为256，中间层嵌入表示向量$\boldsymbol{e}_i$和$\boldsymbol{h}_i$的大小设置为128。

5.2 语料构建效率与质量

本文基于图4所描述的中医临床切诊信息抽取与词法分析语料构建方法框架完成了中医临床切诊信息抽取、中文分词和词性标注语料三个语料的构建。如前文所述，中医临床切诊信息抽取语料基于前序工作构造的语料得到（具体方法参见3.2.1），该语料也采用图4所述框架构造得到，语料构建质量（即IAA值的相关信息参见(王亚强et al., 2022)）。

基于图4所描述的中医临床切诊信息抽取与词法分析语料构建方法框架，三位标注专家经过了三轮的标注训练，即达到对中医临床切诊描述的中文分词和词性标注两项标注任务的认知一致性，三轮标注的IAA值结果如表2所示。

	中文分词标注任务的IAA值	词性标注任务的IAA值
第一轮训练	0.9281	0.8543
第二轮训练	0.9564	0.9361
第三轮训练	0.9523	0.9381
最终标注结果	0.9450	0.9470

表 2. 中文分词和词性标注任务上的标注训练及最终标注的IAA值

从表2的结果可以看出，本文提出的中医临床切诊信息抽取与词法分析语料构建方法框架具有良好的语料构建效率，可以高效地完成三份语料的构建，经过三轮的标注训练，即能达到对标注任务的认知一致。

此外，从表2的结果还可以看出，三位标注专家在每项任务上均可以在第一轮就达到对标注任务的较高认知一致性。在中文分词标注任务中，三位第一轮训练后的IAA值达到0.9281。在词性标注任务中，三位第一轮训练后的IAA值达到0.8543。这些结果说明，本文制定的标注规范具有良好的代表性和可操作性。

进一步观察2的结果可以看出，中医临床切诊描述的词性标注任务的IAA值在不断提升，但在标注训练过程中，中医临床切诊描述的词性标注任务的IAA值始终低于中文分词标注任务的IAA值。同时观察表1可以发现，中医临床切诊描述的词性标注任务标签数量远高于其它两项任务。上述结果表明，中医临床切诊描述的词性标注任务相较于中医临床切诊信息标注任务和中文分词标注任务更有难度。

本文采用随机无放回采样策略构造每轮标注专家标注训练数据，以保证训练过程的客观性，从最终的标注结果来看，本文所构建的中医临床切诊信息抽取、中文分词和词性标注语料

质量高，各标注任务最终的IAA值均高于0.94，为强一致性标注结果(Fleiss, 1981)。从中文分词和词性标注任务训练过程获得的IAA结果来看，词性标注任务的IAA值始终在提升，说明本文制定的标注指南将词性标签给予了清晰的定义。比较特殊的是，中文分词任务的IAA值尽管第一轮训练时即达到0.9281，然而训练过程中的IAA值有波动上升的趋势。该结果说明，中医临床切诊描述的中文分词任务尽管相对简单，但由于中医临床记录描述存在不规范的现象，使其词语边界的界定常存在歧义性，指南需要持续更新。

5.3 语料特点分析

表3中分别统计了排名前十的中医临床切诊信息抽取语料中包含的频繁字和非频繁字，中医临床切诊描述的中文分词语料中包含的频繁字和非频繁字以及频繁词和非频繁词。基于表3，对比两份语料中的频繁字与非频繁字之间的差异可以明显看出，中医临床切诊描述具有中医特色，并且中医切诊在中医临床辨证论治过程中具有重要意义，“脉”、“细”、“略”等具有中医切诊特色的字频繁出现在中医临床记录中。

中医临床切诊信息抽取语料				中医临床切诊描述的中文分词语料							
频繁字	计数	非频繁字	计数	频繁字	计数	非频繁字	计数	频繁词	计数	非频繁词	计数
痛	9740	腴	1	脉	7915	眉	1	脉	7915	石	1
脉	8052	悉	1	细	4486	此	1	细	4486	皆有	1
苔	6647	络	1	弦	2093	人	1	弦	2093	一直	1
舌	6410	哽	1	弱	1123	到	1	弱	1123	好	1
薄	4726	绳	1	滑	945	皮	1	滑	945	眉	1
便	4658	死	1	沉	940	光	1	沉	940	裂纹	1
黄	4631	割	1	数	889	六	1	数	889	耳门	1
细	4564	呵	1	略	598	柔	1	略	598	下	1
干	4424	屋	1	软	585	时	1	软	585	裂	1
略	4200	回	1	平	532	或	1	平	532	凹陷性	1

表 3. 中医临床切诊信息抽取语料和中文分词语料中排名前十的频繁与非频繁字和词的统计数据

此外，从表3中“脉”、“细”、“略”等具有中医切诊特色的字的频数，在中医临床切诊描述的中文分词语料和中医临床切诊信息抽取语料中的对比发现，频数并不相等。该结果说明，中医临床切诊描述中的字存在歧义性，这为中医临床切诊信息抽取、中文分词以及词性标注任务均带来一定的挑战。

对比表3中，中医临床切诊描述的中文分词语料中的频繁字和非频繁字，以及频繁词和非频繁词可以发现，中医临床切诊描述单字成词的现象较严重，这与中医临床记录具有简短性的特点(Wang et al., 2012)有关，这一特点使得单字词发生兼类现象(陆俭明, 1994)的风险增加，一定程度上会对中医临床切诊描述的词性标注任务形成挑战。

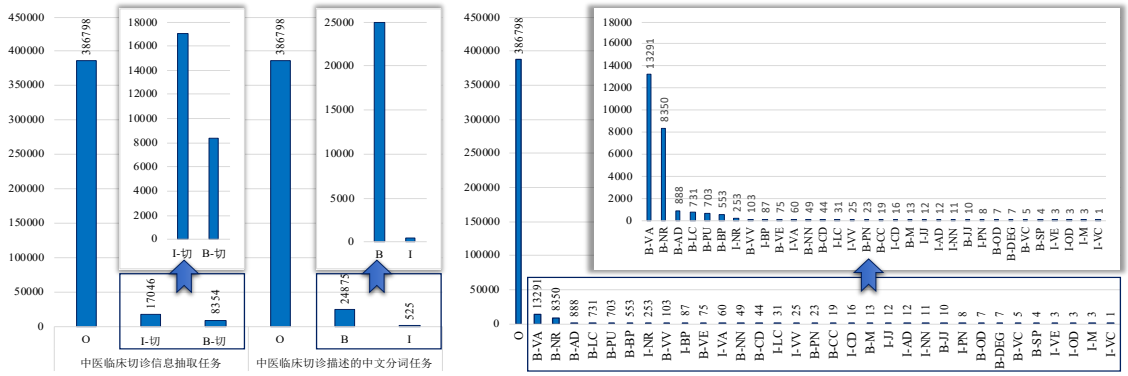


图 8. 中医临床切诊信息抽取、中文分词和词性标注语料标签分布

中医临床切诊信息抽取、中文分词和词性标注任务的标签分布如图8所示，从图中可以清晰的观察到，三项任务均面临严重的标签分布有偏问题，其中其他类标签“O”最多。并且，即使将其他类标签“O”删除，三项任务依然面临标签分布有偏的问题，长尾分布始终保持，该问题将对基于序列标注的多任务联合信息抽取带来挑战。

5.4 多任务联合信息抽取

本文评价多任务联合信息抽取结果分别采用了对象级（即切诊描述片段对象、词语对象、词性对象的整体）以及标签级的准确率（P）、召回率（R）和F度量值（F）的宏平均结果性。两类P、R和F的宏平均计算方法参见(Wang et al., 2014)，实验结果如表4所示。

	任务损失贡献度	切诊抽取任务			中文分词任务			词性标注任务		
	$(\beta_1, \beta_2, \beta_3)$[1]	P	R	F	P	R	F	P	R	F
对象级	(1, 0, 0)	93.94	93.18	93.55	-	-	-	-	-	-
	(0, 1, 0)	-	-	-	80.24	**71.01**	**74.6**	-	-	-
	(0, 0, 1)	-	-	-	-	-	-	45.42	38.18	40.33
	(1/3, 1/3, 1/3)	**94.00**	93.02	93.51	79.80	68.97	73.19	43.62	35.72	37.73
	(0.2, 0.3, 0.5)	93.98	**93.24**	**93.58**	79.80	69.86	73.95	**49.28**	**40.20**	**42.80**
	(0.1, 0.2, 0.7)	93.96	92.90	93.43	**80.25**	68.39	72.97	47.58	38.95	41.58
标签级	(1, 0, 0)	96.85	95.31	96.06	-	-	-	-	-	-
	(0, 1, 0)	-	-	-	86.19	**80.81**	**82.96**	-	-	-
	(0, 0, 1)	-	-	-	-	-	-	50.81	43.03	45.21
	(1/3, 1/3, 1/3)	**96.87**	95.25	96.04	**86.25**	79.28	82.04	47.58	39.72	41.38
	(0.2, 0.3, 0.5)	96.67	**95.50**	**96.08**	85.74	79.88	82.35	**53.19**	**45.53**	**47.41**
	(0.1, 0.2, 0.7)	96.83	95.20	96.00	85.97	78.89	81.70	52.97	42.50	45.18

[1] $(\beta_1, \beta_2, \beta_3)$被设置为(1, 0, 0)、(0, 1, 0)和(0, 0, 1)时，多任务联合信息抽取框架退化为基于BERT+BiLSTM+CRF的单任务的信息抽取模型，“1”所在维度为当前执行的抽取任务。

表 4. 多任务联合信息抽取结果

如表4所示，相对于单任务的信息抽取模型，无论是对象级还是标签级，采用多任务联合信息抽取方法，在不同的$(\beta_1, \beta_2, \beta_3)$设置下，总能取得$P$、$R$或$F$的结果提升。特别是在复杂的中医临床切诊描述的词性标注任务上，效果提升明显，在$(\beta_1, \beta_2, \beta_3)$被设置为(0.2, 0.3, 0.5)时，$F$值在对象级和标签级分别提升了2.47%和2.2%。

从表4可以看出，多任务联合信息抽取方法在中医临床切诊描述的中文分词任务上效果不明显，仅P值结果有少量提升。此外，中医临床切诊描述的词性标注任务的总体性能不高，还有巨大的提升空间，这与其任务复杂性有关，标签数量多，且中医临床记录描述有其独特性。如何有效提升多任务联合信息抽取的性能是未来有待深入研究的方向。

6 总结

切诊极具中医特色，中医临床切诊信息抽取与词法分析为基于中医临床记录的辨证论治上下游任务研究提供丰富的中医临床医学语义信息。本文首次开展了中医临床切诊信息抽取与词法分析研究，围绕中医临床切诊信息抽取、中文分词和词性标注任务，构建了万余条带标注的多任务高质量语料，为中医临床记录的自然语言处理和文本挖掘开辟了新方向。本文基于同级多任务共享编码参数的多任务学习框架，开展了中医临床切诊信息抽取、中文分词和词性标注联合信息抽取的研究初探，形成了基线结果，并发现了系列问题，为后续研究指明了方向。

参考文献

Tian Bai, Haotian Guan, Shang Wang, Ye Wang, and Lan Huang. 2022. Traditional Chinese medicine entity relation extraction based on CNN with segment attention. *Neural Computing and Applications*,34: 2739-2748.

Michael Crawshaw. 2020. Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv preprint arXiv:2009.09796*.

Jacob Devlin, MingWei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378.

J. L. Fleiss 1981. Statistical methods for rates and proportions. 2nd ed. *John Wiley and Sons*, IBSN: 978-0-471-26370-8.

Qin Fang, Yane Li, Hailin Feng, and Yaoping Ruan. 2023. Chinese Named Entity Recognition Model Based on Multi-Task Learning. *Applied Sciences*, 13(8): 4770.

Yu Guan, Huan Li, and Wenjing Xu. 2021. A Traditional Chinese Medicine Terminology Recognition Model Based on Deep Learning: A TCM Terminology Recognition Model. *Proceedings of the 6th International Conference on Big Data and Computing*, 15-20.

Zhen Guo, Yujie Zhang, Chen Su, Jinan Xu, and Hitoshi Isahara. 2016. Character-Level Dependency Model for Joint Word Segmentation, POS Tagging, and Dependency Parsing in Chinese. *IEICE TRANSACTIONS on Information and Systems*, 99(1): 15-20.

Honglan Liu, Xiaona Qin, and Bin Fu. 2015. The Symptoms and Pathogenesis Entity Recognition of TCM Medical Records Based on CRF. *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 1479-1484. IEEE.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 799-809.

Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *arXiv preprint arXiv:1807.01882*.

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2017. Neural Joint Model for Transition-based Chinese Syntactic Analysis. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1204-1214.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv:1603.01354*.

Thai-Hoang Pham, Khai Mai, Nguyen Minh Trung, Nguyen Tuan Duc, Danushka Bolegala, Ryohei Sasano, and Satoshi Sekine. 2019. Multi-Task Learning with Contextualized Word Representations for Extented Named Entity Recognition. *arXiv preprint arXiv:1902.10118*.

Marek Rei. 2017. Semi-supervised Multitask Learning for Sequence Labeling. *arXiv preprint arXiv:1704.07156*.

Artstein Ron and Poesio Massimo. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational linguistics*, 34(4): 555-596.

Chunyang Ruan, Yingpei Wu, Guang Sheng Luo, Yun Yang, and Pingchuan Ma. 2020. Relation Extraction for Chinese Clinical Records Using Multi-View Graph Learning. *IEEE Access*, 8: 215613-245622.

Benoît Sagot and Héctor Martínez Alonso. 2017. Improving neural tagging with lexical information. *Proceedings of the 15th International Conference on Parsing Technologies*, 25-31, Pisa, Italy, September, Association for Computational Linguistics.

Xiao Sun, Degen Huang, and Fuji Ren. 2009. Chinese lexical analysis based on hybrid MMSM model. *International Journal of Innovative Computing, Information and Control*, 5(12 (A)).

Youhyun Shin and Sang-goo Lee. 2019. Learning Context Using Segment-Level LSTM for Neural Sequence Labeling. *IEEE/ACM Transactions on Audio, Speech, and language processing*, 28:105-115.

Yaqiang Wang, Zhonghua Yu, Yongguang Jiang, Yongchao Liu, Li Chen, and Yiguang Liu. 2012. A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. *Journal of Biomedical Informatics*, 45(2):210-223.

Yaqiang Wang, Zhonghua Yu, Li Chen, Yunhui Chen, Yiguang Liu, Xiaoguang Hu, and Yongguang Jiang. 2014. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study. *Journal of Biomedical Informatics*, 47:91-104.

Fei Xia. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0).

Fei Xia. 2000. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). *IRCS Technical Reports Series*, 38.

Tingting Zhang, Yaqiang Wang, Xiaofeng Wang, Yafei Yang, and Ying Ye. 2020. Constructing fine-grained entity recognition corpora based on clinical records of traditional Chinese medicine. *BMC Medical Informatics and Decision Making*, 20(1):1-17.

Tingting Zhang, Zonghai Huang, Yaqiang Wang, Chuanbiao Wen, Yangzhi Peng, and Ying Ye. 2022. Information Extraction from the Text Data on Traditional Chinese Medicine: A Review on Tasks, Challenges, and Methods from 2010 to 2021. *Evidence-Based Complementary and Alternative Medicine*, 2022.

靳光瑾, 肖航, 郭曙伦, 富丽, 章云帆, 于桂英, 陈玉泉, 王立. 2006. 信息处理用现代汉语词类标记规范. 中华人民共和国国家质量监督检验检疫总局; 中国国家标准化管理委员会, GB/T 20532-2006.

李灿东. 2021. 中医诊断学. 中国中医药出版社

陆俭明. 1994. 关于词的兼类问题. 中国语文, 28-34.

李灿东, 陈研, 郭宇博, 苏祥飞, 王天芳, [朱文锋], 郑进, 顾星, 王洋, 林雪娟, 甘慧娟, 李宇涛. 2021. 中医四诊操作规范　第4部分：切诊. 国家市场监督管理总局; 国家标准化管理委员会, GB/T 40665.4-2021.

谭同来. 2010. 中华医学切诊大全. 山西科学技术出版社.

王亚强, 李凯伦, 蒋永光, 舒红平. 2022. 基于批数据过采样的中医临床记录四诊描述抽取方法. 第*21*届全国计算语言学大会论文集, 611-622.

杨杰. 2006. 基于脉动信息获取的中医脉诊数字化、可视化探讨. 北京中医药大学

杨锦锋, 关毅, 何彬, 曲春燕, 于秋滨, 刘雅欣, 赵永杰. 2016. 中文电子病历命名实体和实体关系语料库构建. 软件学报, 27(11): 2725-2746.

中医临床切诊描述词法分析相关规范. 2023. *https://github.com/xx-Jiangwen/Guideline.*

大规模语言模型增强的中文篇章多维度阅读体验量化研究

孙嘉黛, 汤思怡, 王诗可, 于东*, 刘鹏远
北京语言大学/ 信息科学学院
北京市海淀区学院路15号, 100083
{suesunegg,tangsiyi0805,shikewang98}@gmail.com
yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

摘要

现有的文本分级阅读研究往往从文本可读性的角度出发，以离散的文本难度等级的形式为读者推荐阅读书目。目前，仍缺少一种研究读者在阅读过程中产生的多方面、深层次阅读体验的体系结构。对此，我们调研了读者在阅读中文篇章过程中产生的不同阅读体验，提出了中文篇章多维度阅读体验的量化体系。我们将阅读过程中呈现的连续性的阅读体验归纳为多种类别，并在此基础上构建了中文篇章多维度阅读体验数据集。同时，我们探究了以大规模语言模型为基础的ChatGPT对阅读体验的量化能力，发现其虽具备强大的信息抽取和语义理解能力，在阅读体验的量化上却表现不佳。但我们发现大规模语言模型所蕴含的能力能够以知识蒸馏的方式协助深层属性的量化，基于此，我们实现了大规模语言模型增强的中文篇章多维阅读体验量化模型。模型在各维度阅读体验上的平均F1值达到0.72，高于ChatGPT的Fewshot结果0.48。

关键词： 阅读体验；大规模语言模型；知识融入

Quantitative Research on Multi-dimensional Reading Experience of Chinese Texts Enhanced by Large Language Model

Jiadai Sun, Siyi Tang, Shike Wang, Dong Yu*, Pengyuan Liu
Beijing Language and Culture University / Faculty of Computer Science
15 Xueyuan Road, Haidian District, Beijing, 100083
{suesunegg,tangsiyi0805,shikewang98}@gmail.com
yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

Abstract

Existing studies on graded reading often recommend reading materials to readers in the form of discrete difficulty levels from the perspective of text readability. However, there is still a lack of a systematic framework that can capture the multi-faceted and deep reading experiences of readers during the reading process. To address this issue, we investigate the different reading experiences of readers during the process of reading Chinese texts and propose a quantitative framework for multidimensional reading experiences in Chinese discourse. We group the continuous reading experiences that emerged during the reading process into multiple categories and constructe

*为通讯作者
基金项目：教育部人文社会科学基金项目(19YJCZH230)；中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)

a dataset of multidimensional reading experiences in Chinese discourse. Additionally, we explore the ability of ChatGPT, based on Large Language Models, to quantify reading experiences and find that while it has strong information extraction and semantic understanding capabilities, it performs poorly in quantifying reading experiences. However, we find that the capabilities inherent in Large Language Models can assist in the quantification of deep attributes through knowledge distillation. Based on this, we implement an enhanced model for quantifying multidimensional reading experiences in Chinese discourse using a Large Language Model. The model achieves an average F1 score of 0.72 on various dimensions of reading experience, which is higher than ChatGPT's Fewshot result of 0.48.

Keywords: Reading Experience , Large Language Models , Knowledge Intergration

1 引言

在全民阅读氛围浓厚的当下，阅读推广活动正在广泛开展，中小学生等青少年儿童在被推广人群中的比重最高。由此，为青少年推荐适合他们阅读的优秀读物是专家和学者一直关注的话题。为了适应匹配不同阅读水平的青少年读者，分级阅读也成为近年来研究的重点(Rao et al., 2021)。目前，分级阅读通常只考虑文本难度，即文本可读性(Dale and Chall, 1948)。但阅读也需要深度和广度，在保障难度适配的同时增加分级阅读文本的多样性、提升读者在阅读过程中的体验感也是阅读研究的新兴方向(池春红, 2022)。何应艳 (2016)认为阅读体验即为读者在理解文字的情况下，通过阅读产生的深层次、多元的感受。在此基础上，我们总结了阅读体验的两个特征，一是深层次，二是多维度。因此，阅读体验的概念可以扩展为读者在阅读过程中对文本的不同维度产生的多元感受。与文本难度相比，阅读体验结合了文本客观因素、读者心理因素两方面，它更加重视阅读过程中读者的主观感受。

文本可读性研究往往从字数、词数、句法结构、篇章结构等浅层语言特征对文本进行分析，忽视了文本的一些深层语义内涵，也忽视了读者与文本的互动性。图1展示了两个来自同年级语文课文的不同文段，两个文段在文字难度上基本相同，即在可读性上的等级相同，但两者给读者带来的阅读体验却截然不同。文段1通过优美的语言表达，给读者较高的文采体验；而文段2用词平实，却讲述了一个具有哲理性的故事，更加引人深思。融合阅读体验进行文本分析、分级阅读、读物推荐等工作，可以从深层与浅层多个角度提取出文本特征，使读者能更快速、准确地感受文章情感，欣赏文章的优美文字，领悟文章的内涵和主旨(杨春and 张秋月, 2022)。

文段1：道两旁的法国梧桐树，掉下了一张张金黄金黄的叶子。这一张张闪着雨珠的叶子，一掉下来，便被紧紧地粘在湿漉漉的水泥道上。我走在院墙外的水泥道上。水泥道像铺上了一块彩色的地毯，这是一块印着落叶图案的，闪闪发光的地毯，从脚下一直铺到很远很远的地方，一直到路的尽头……每一张法国梧桐树的落叶，都像一个金色的小巴掌，熨帖地、平展地粘在水泥道上。它们排列得并不规则，相反，很凌乱。然而，这更增添了水泥道的美。我一步一步小心地走着，我一张一张仔细地数着。我穿着一双棕红色的小雨靴。你瞧，这多像两只棕红色的小鸟，在秋天里变得金黄的叶丛间，愉快地欢跳着、歌唱着……

文段2：孙中山想，这样糊里糊涂地背，有什么用呢？于是，他壮着胆子站起来，问："先生，您刚才让我背的这段书是什么意思？请您给我讲讲吧！"这一问，把正在摇头晃脑高声念书的同学们吓呆了，课堂里霎时变得鸦雀无声。先生拿着戒尺，走到孙中山跟前，厉声问道："你会背了吗？""会背了。"孙中山说着，就把那段书一字不漏地背了出来。先生收起戒尺，摆摆手让孙中山坐下，说："我原想，书中的道理，你们长大了自然会知道的。现在你们既然想听，我就讲讲吧！"先生讲得很详细，大家听得很认真。后来，有个同学问孙中山："你向先生提出问题，不怕挨打吗？"孙中山笑了笑，说："学问学问，不懂就要问。为了弄清楚道理，就是挨打也值得。"

图 1. 可读性相同但阅读体验不同的文段

现阶段对阅读活动的探索仅仅着眼于文本可读性的定量研究或基于机器学习、深度学习模型的语义研究，而忽略了其他深层阅读体验的挖掘，基于此形成的分级阅读和文本推荐在广泛性、多样性及可解释性上有所欠缺。对于文本浅层性质的研究大多指向各类可读性公式，许多可读性公式对多种浅层语言特征进行了定量分析(程勇et al., 2020; 刘苗苗et al., 2021)；对于文本深层阅读体验的研究缺失，和文本深层属性相关的研究大多以领域内的分类任务为方向，并

不涉及阅读过程体验感的探索。作为阅读活动的主体，读者的阅读感受能够直观地反映阅读活动中他们的收获，探究读者在阅读过程中基于文本产生的深层次阅读体验能够在分级阅读的基础上融入更多有价值的阅读信息。因此，中文文本多维度阅读体验体系缺失的问题亟待解决。在这样的背景下，我们结合调研分析，将具备连续性的多维阅读体验归纳为特定的类别，并由此构建了中文篇章多维度阅读体验量化体系和数据集。同时，我们发现大规模语言模型出色的信息抽取和语义理解(Bang et al., 2023)能力能够使其表达的自然语言信息能很好地融入模型中以提高模型性能，但在具备多样性的阅读体验量化任务上的表现却不佳。因此，我们提出了以大规模语言模型辅助多维度阅读体验量化的方法，并采用了蒸馏大模型知识的方式来完成这一过程。

本文的贡献主要包括以下三个方面：第一，我们提出了中文篇章多维度阅读体验的量化体系，并基于此构建了中文篇章多维度阅读体验数据集，为衡量中文篇章的阅读体验提供了理论基础，数据集公开在https://github.com/blcunlp/CMRED；第二，我们探究了大规模语言模型ChatGPT(Dai et al., 2022)在阅读体验上的能力，发现ChatGPT虽然能以自然语言的形式出色地完成抽取阅读体验相关表达的工作，但在多维度阅读体验数据集上体现的量化能力较差；第三，基于探究的结果，我们提出了大规模语言模型增强的中文篇章多维度阅读体验量化方法，利用了ChatGPT在语义层面的信息抽取能力，以三种不同的方式在模型中融入来自ChatGPT的知识并通过实验验证了方法的有效性。

2 相关工作

文本作为作者向读者传递信息的载体，有着许多不同的属性，在自然语言处理领域讨论的文本属性多集中在文本可读性领域。Dale and Chall (1948)将文本的可读性定义为文本易于理解和阅读的程度和性质，通常被狭义理解为文本的难度。早期的可读性研究尝试将文本词汇层面和句子层面的如词长、常用词比例、句长等浅层特征进行量化，并以此类特征对文本的难度进行回归预测。常用的可读性公式有Reading Ease Score (Flesch, 1948) 、Grade Level (Kincaid et al., 1975)、Lexile Framework (Smith and others, 1989)等。然而，可读性公式缺少对语义、篇章层面的深层语言特征的考量(McNamara et al., 2010)，仅以统计学上的相关浅层特征作为构建基础，有效性值得商榷(Schriver, 2000)。此后随着计算机学科的发展，可读性领域也引入了机器学习、深度学习等新方法。此类方法依靠心理语言学、语料库语言学等理论支撑，以计量的方法表示出更多语言特征，通过模型对文本的难度等级进行预测。常用的机器算法有决策树(Decision Tree，DT)、随机森林(Random Forest，RF)、逻辑回归(Logit Regression，LR)、最近邻(K Nearest Neighbor，KNN)、支持向量机(Support Vector Machine，SVM)等(Wu et al., 2018)。深度学习方法则旨在运用深层神经网络的特点，使得文本深层的特征得以表示。Lee et al. (2021)的研究表明在模型构建时，同时考虑计量语言特征和深度语言特征的方法可以使模型有更好的表现。

上述可读性研究仅考虑了文本的客观特点，而阅读活动需要读者的亲身参与，文本所描写的事物、表达的感情、蕴含的哲理都会对读者产生影响。读者在阅读文本后产生的情感连接、引发的思考等也是影响阅读活动的重要因素。何应艳 (2016)提出了阅读体验的概念，认为读者在理解文字的基础上，通过阅读产生的深层次、多元的感受被称为阅读体验。阅读体验反映了作者所传达的审美意识，文本与科学美、社会美、自然美、艺术美等方面结合，体现了阅读体验的多维性，从而打破了用文本难度这一单一指标来评价文本的局限。廖圣河and 陈怡瑾 (2021)提出读者在阅读过程中的体验是沿着对文字符号、艺术形象、思想意识的深入理解逐步产生的，在调动感受力充分阅读文本得到情感、哲理、道德、人文等方面的体验后，读者的人文、科学素养和共情能力得以提升。Tracy (2017)提出阅读体验源自读者参加阅读活动的主观感受，阅读活动能够带给读者认知、情操、价值与分享四个方面的感受，因此其获得的阅读体验也分为与之对应的四种。谭继雄 (2020)则关注了读者的个体区别，认为不同的读者对阅读有着不同的需求，而需求的不同也会带来不同的阅读体验，阅读体验最终可以分为认知、审美、价值和社交四个维度。因此，我们将探究读者在阅读过程中产生的基于文本的阅读体验，把读者的阅读感受映射到文本特征上，从而使主观感受具象化，使连续属性离散化。同时，考虑阅读体验的文本在表示上也更加丰富饱满，提供了文本评价的新思路。

第二十二届中国计算语言学大会论文集，第364页-第376页，哈尔滨，中国，2023年8月3日至5日。

3 多维度阅读体验理论构建与数据建设

3.1 多维度阅读体验理论构建

由于目前对于阅读体验的研究较少，尚没有工作从多维度考虑读者在阅读过程中产生的阅读体验。因此，我们将构建一个完整的文本阅读体验的评价体系，并在此体系下完成量化阅读体验的目标。由于文本分级的重点在于篇章的分类，我们研究的对象是中文篇章，对中文篇章多维阅读体验的量化任务定义为如下形式：G_i=Classify$_{\mathrm{i}}$($text$)。其中，G_i表示某个阅读体验i的最终量化结果，Classify$_{\mathrm{i}}$指对指标i的分类操作。我们经过多方调查完成了阅读体验指标的选择和体系的构建。读者在阅读过程中产生的阅读体验具有多元、广泛的特点，因此我们以问卷的形式招募读者进行了广泛调查。问卷中，事先给出了可能与阅读体验相关的各种因素，读者需要阅读篇章后选择对阅读影响较大的因素。通过多方面比较和筛选，最终选定了理解性体验、文采性体验、道德性体验、思维性体验和情感性体验五个较有代表性的指标作为中文篇章多维度阅读体验指标体系的基础构成。

在阅读过程中，读者对文本的理解程度会第一时间影响阅读的体验感，过难或过简单的文本都不利于读者对文本的进一步理解，因此可理解性是必须考虑的重要指标。在我们的体系中，理解性体验按照之前可读性工作中的划分方式将小学六个年级分为三个学段，每两个年级定义为一个学段，即一二年级为学段一，三四年级与五六年级依次递增，分别标注为“0”、“1”、“2”三个类别，数字越大表示文本越不容易理解。Li et al. (2022)将文采分为“高文采”与“低文采”两个类别，我们参照该工作，以“0”表示读者对该文本的文采性体验较低，以“1”表示文采性体验较高。Wang et al. (2020)将词语的道德性分为正向道德、负向道德、中性、被动四类，在阅读文本过程中，读者会对文本中的人物形象、行为模式产生自己的道德取向，这样的道德取向性体验可以在文本中起到教育警示的作用。我们结合本次的标注语料，将文本的道德性体验分为正向、负向与中性三类，分别用“1”、“0”、“-1”表示。具有哲理性、思辨性的文本能够对读者阅读后的长期感受、价值观形成影响，我们将读者在阅读过程中感受到的关于这些方面的体验定义为思维性体验，分为两类，以“0”表示无思维相关的体验，“1”表示有思维相关的体验。读者阅读过程中感受到的不同强度的情感能够在不同程度上唤起读者的共情能力，因此情感性体验也是体系中的重要因素。我们把情感性体验分为三个类别，分别用“0”、“1”、“2”表示情感性弱、情感性适中、情感性强。最终的阅读体验类别标签情况如表1所示。

指标名称	类别数量	类别标签
理解性体验	3	0, 1, 2
文采性体验	2	0, 1
道德性体验	3	-1, 0, 1
思维性体验	2	0, 1
情感性体验	3	0, 1, 2

表 1. 阅读体验各指标对应类别情况

3.2 多维度阅读体验数据集构建

当前没有统一的数据集用于评价读者在阅读过程中产生的阅读体验，因此我们采用人工标注的方式构建一个全新的中文篇章多维度阅读体验数据集CMRED（**C**hinese **M**ultidimensional **R**eading **E**xperience **D**ataset）。该数据集内包含篇章及其对应的多维阅读体验类别。将来，也可以通过人工标注或机器辅助的方式扩充该数据集的数量，或增加其他的指标以完善对阅读体验的探究。

我们使用骆香莹 (2022) 构建的语文教科书语料库中的小学教材作为标注语料，其中教科书语料库包含四个出版商的语文教材版本，分别为北师大版、人教版、苏教版以及义务教育课程标准实验教科书部编版，共计776篇课文。由于低年级的文本数量较少，为了保证数据基本平衡，我们又融入了绘本教材。绘本为学龄前儿童用于学习基础汉字使用的图书，文本内容较为简单，可与低年级的文本划为同一等级。我们选取了两个口碑较好、教材内容有权威性的品牌，奕阳绘本和亿童绘本，从中选择了208篇内容作为标注文本。最终总计984篇文本作为本次

标注工作的基础数据。为了确保文本长度不会显著影响阅读体验，我们以500字为准对文本进行切分。切分以句为单位，同时保证切分点位于句子末尾。最终，我们得到了1880篇由部分段落构成的不超过500字的新篇章，并随机选取其中的1200篇用于本次标注工作。

数据标注包括标注前准备和数据正式标注两个阶段。在标注前准备阶段，我们使用“锚点标注法”(唐玉玲et al., 2022)进行本次标注，具体规则如下：对于每个指标中每一类别，提供两个篇章作为参照锚点。锚点由两名语言学专业的硕士共同商讨决定，选择最具有代表性的两个篇章作为参考标准。标注者在标注过程中与锚点进行比较，如果相似则标记相同类别，反之则考虑与锚点间比较的倾向性，是偏向更高级别还是更低级，如果介于两者之间则标记为较低等级。例如对于三分类的情感性阅读体验，判断其强度没有类别“2”那么强烈，但又比类别“1”强烈，最终仍然将该篇章归为类别“1”。这样的要求是为了统一标注上的歧义，方便后续进行结果整理。初始数据标注由五名硕士研究生完成，在正式标注开始之前，我们对标注员进行了统一的培训，明确了标注任务和标注规则，并让标注员对比锚点文本进行了试标注，为了进一步强化理解本次标注工作。在正式标注阶段，每名标注员需独立完成1200条数据的标注，即每条数据要被标注5次。标注周期为20天，每日对标注员的工作进行检查，以保证标注质量。

标注结束后，我们对五名标注员的标注结果进行了检查以及合并。对于文采性体验和思维性体验，我们按照多数投票原则选择最终类别；对于理解性体验、道德性体验和情感性体验指标，会出现选择各标签类别的标注员人数为2:2:1的情况，无法通过多数投票法选出最终类别，此时，我们让第六名标注员对该样本进行二选一的类别选择，确定标注等级。

最终我们收集到了1200条标注结果，数据集CMRED的情况如表2所示。

指标名称	理解性体验			文采性体验		道德性体验			思维性体验		情感性体验		
类别	0	1	2	0	1	-1	0	1	0	1	0	1	2
文段数量	304	518	378	1045	155	39	859	302	1082	118	418	650	132

表 2. 数据集CMRED指标及其类别分布

4 方法

4.1 大规模语言模型增强的多维度阅读体验量化整体架构

当前，具备海量知识的超大规模语言模型获得了瞩目，如何利用大模型中的知识开展工作成为了一个值得研究的课题。由于我们的任务和语义知识的相关性，我们提出了基于大规模语言模型增强的多维阅读体验量化方法。多维阅读体验的某些指标之间存在相辅相成的关系，例如文采性阅读体验往往会影响读者对篇章的理解性阅读体验，因此我们使用多任务协同计算的方式来实现基础模型，我们的方法的整体架构如图2所示。

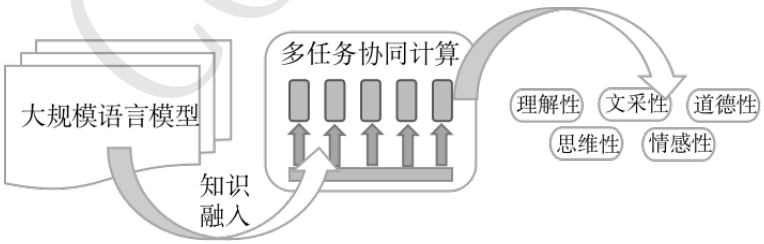


图 2. 多维度阅读体验量化整体架构

在此架构中，整个量化过程主要通过两步实现，第一步是通过合适的方式获取来自大规模语言模型的知识，第二步是将这些知识融入多任务协同计算的模型中。这两个步骤的最终输出将给读者呈现多维度的阅读体验量化结果，为读者阅读文本提供参考性指标，并能在后续的应用中实现更多有价值的工作，例如从文库中匹配与某一篇文本相似的文章或适宜某一类读者阅读的文章。

4.2 多任务协同计算方法

数据集CMRED在五个层面上对篇章的阅读体验进行了指标量化，量化的过程实际上为分

类任务，最终目标是能够给出文本在各维度体验上的分类值。文本的阅读体验并非相对于文本独立存在，而是相互影响、交融，多任务学习(Crawshaw, 2020)能够合理地通过共享参数提高数据的利用效率，通过共享表示减少过拟合，因此我们使用多任务协同计算的方法来实现阅读体验量化，将5种属性的量化任务以表示共享的方式构建统一量化结构，以较小的参数量实现了模型结构。损失函数如下：

$$Loss_m = \sum_{i=1}^{N} \beta_i (\mathrm{FL}_{\mathrm{softmax}}(encoder\_out))_i \tag{1}$$

其中，$\mathrm{FL}_{\mathrm{softmax}}$表示多分类任务下的Focal Loss，Focal Loss是Lin et al. (2017)提出的用以解决样本不平衡的一种损失函数。β表示每项任务损失的系数，在我们的方法中，为平衡每一项分类结果，所有β的取值均为1。

4.3 大规模语言模型增强方法

在多维度阅读体验量化体系中，一些短语能够在第一时间影响读者对某个维度的阅读体验。例如，当读者读到“伟大”、“好人好事”等表达时，会立刻触发心中对文本描述的人或事物的道德判断，随即根据文本具体内容产生更多有关道德的阅读感受。这样的表达会在一定程度上影响读者阅读下文时的关注点，也会让读者对前文内容作出短时的总结，因此我们将这种涉及到文本阅读体验的短语称为“阅读体验触发语”。阅读体验触发语是影响读者对阅读属性体验判断的关键信息，在模型中融入这些短语也会有利于模型进行量化。

ChatGPT是大型预训练语言模型InstructGPT(Ouyang et al., 2022)的后继模型，作为带有对话界面的通用语言模型，具备较强的信息抽取能力和语义理解能力。我们采用了合适的提示方式，以获取来自ChatGPT的语言知识。我们使用ChatGPT抽取指定文本的文采性、道德性、思维性、情感性相关的触发语。由于文本的理解性体验很难通过短语的方式呈现，在我们的工作中并没有标注涉及理解性体验的触发语。对于ChatGPT标注的阅读体验触发语，我们进行了预处理和人工评估，发现其标注结果和人工标注的结果基本一致。

我们共采用了三种技巧在模型中融入来自ChatGPT的知识，分别是前缀标签化提示法、标签联合监督法和标签线性化监督法。图3展示了以上三种方式在模型中融入知识的过程，整个过程分为三个阶段进行。第一阶段是利用ChatGPT进行阅读体验触发语的挖掘，第二阶段是对阅读体验触发语进行标签化，第三阶段是在基线模型中融入触发语优化，接下来我们将分别介绍以上三种基于大模型增强的阅读体验量化方法的实现过程。

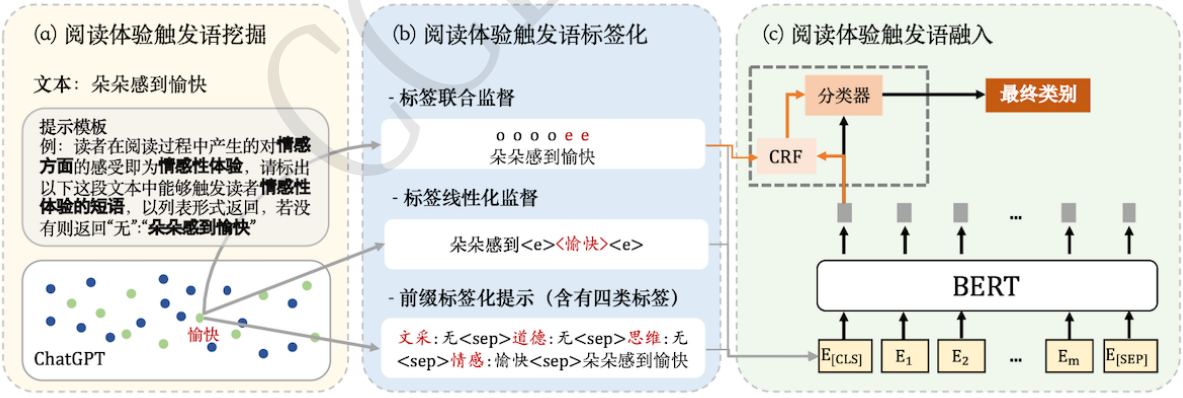


图 3. 基于大模型增强的中文文本多维阅读体验量化流程

前缀标签化提示 在提示工程的新范式(Liu et al., 2021)下，预训练语言模型的上下文学习能力可以充分体现，借鉴该思路对提示的认知，我们提出以标签化提示的方式，在模型的输入上添加提示信号。具体的形式如下：[**标签类型:标签词**][原始文本输入].其中，标签类型包括文采、道德、思维、情感四种，输入模型时，每一种标签类型之间以分隔符号分隔。相比起提示工程从预训练语言模型中直接得到对应的输出，我们的提示方式旨在将强信号融入输入的文

本，将该信号作为模型需要学优先习的表示之一，因此需要在模型的输入中强调重点关注的标签类别和标签词语。由于该信号的强调发生在文本的初始，能起到提示信号的作用，为模型学习输入文本提供先决条件。

标签联合监督 在原始的方法中，我们用多任务学习的方法构建了量化模型的统一框架，但该框架仅在输入和指标间形成了映射关系，对文本表示的可解释性较差。对此，我们采取序列标注的方式，对文本中序列$x_1x_2...x_m$进行基于阅读体验触发语的标注得到对应的新序列，以g代表文采性、m代表道德性、p代表思维性、e代表情感、o代表普通文本，即：$a_1a_2...a_m$=**A**($x_1x_2...x_m$)。其中A表示标注操作。对于标签冲突的情况，我们采取随机的方式选择最终标签。在模型中，我们使用CRF(Lafferty et al., 2002)作用于基线模型的输出结果，以得到序列标注的损失。就整体模型而言，我们将此序列标注任务视作是多任务协同计算中的一环，增设一个模型通道融入多任务协同计算框架。该通道后的损失函数如下所示：

$$Loss_a = -\log(\sum_{t \in T} \mathrm{e}^{score(t)}) - score(t_G) \tag{2}$$

$$score(t) = \sum_{i}^{L_t} \mathrm{CRF_i}[t_{i-1}, t_i] \tag{3}$$

其中，$Loss_a$是维特比损失，t_G表示真实的标注序列，T表示所有可能的标注序列。$score_t$定义了序列标注t的得分等于每个标注得分的总和，其中L表示序列长度。该方法增加的全序列监督信号能使阅读体验触发语以上下文位置关联的方式融入文本表示中，为后续的分类提供重点信号和位置信息。标签联合监督方法的最终损失是：

$$Loss = Loss_m + Loss_a \tag{4}$$

标签线性化监督 Ding et al. (2020)提出以线性化的方式把文本数据和序列标注变成语言模型的输入数据，从而通过语言模型的输出得到更多的数据以进行数据增强，用于解决低资源情况下的下游任务。由于序列标注的多任务学习方法以完全独立的通道实现对文本表示的监督，会造成一定的信息损失，因此我们将标签线性化的方法迁移至大规模语言模型增强中用于量化任务的表示，不同的是，我们将文采性、道德性、思维性、情感性体验的原始触发语看作是特殊标记，而其他的词看作是普通文本，将特殊标记添加在原始触发语在文本中位置的左右两端。标签线性化之后，模型的输入中既有自然语言，又有特殊标记所带来的深层属性信号，为后续的阅读体验量化任务提供完整、有效的信息。和标签联合监督法相同，我们以g代表文采性、m代表道德性、p代表思维性、e代表情感性、o代表普通文本，对于长度为l的序列$x_1x_2...x_l$，将标签融入到该序列中，即：$m_1m_2...m_{l+k}$=**Linear**($x_1x_2...x_l$)。其中，*Linear*表示标签线性化过程，k表示增加的阅读体验触发语标签数量，融入触发语标签后，序列长度由l变为$l+k$。

5 实验与结果

5.1 实验设计

FewShot下对ChatGPT的能力的探究 ChatGPT是由OpenAI推出的带有图形化界面的基于大规模语料训练的自然语言处理工具(Dai et al., 2022)，我们在此大模型工具上使用FewShot方法测试了其对阅读体验指标的量化能力。

单任务独立计算 沿用传统的文本分类方法，采用预训练语言模型Bert(Devlin et al., 2019)分别对5个阅读体验指标单独建模和微调。

多任务协同计算 采用多任务学习的方法对5个阅读体验指标进行协同计算，此实验和单任务独立计算实验相同，是基于Bert模型实现的。同时，我们在此实验中使用了在第4节中提出的大模型增强方法，完成了在多任务协同计算架构下的大规模语言模型增强的多维度阅读体验量化。

5.2 实验数据与评价指标

我们将数据集CMRED中的1200条样本按照8:1:1的比例分为训练集、验证集和测试集。在实验中，epoch设置为20，学习率设置为1e-5。为了能够准确衡量数据集类别不平衡情况下的模型性能，我们使用精确率（P）、召回率（R）和F1值（macro-F1）作为模型的评价指标。

5.3 主实验分析

ChatGPT在许多任务，例如错误信息检测、问答的FewShot结果令人惊喜(Bang et al., 2023)，但尚未有人在深层属性量化领域探索ChatGPT的能力，因此我们使用数据集CMRED探究ChatGPT在阅读体验量化上面的表现，具体的提示过程在附录A中例举。表4中是ChatGPT对CMRED测试集的FewShot结果。从表中我们可以看到，ChatGPT在各个阅读体验的量化整体上并没有获得较为准确的结果，甚至在思维性这个二分类任务上的表现较差。这可能是由于ChatGPT并不擅长以较为严苛的标准量化深层属性，而仅在上下文语义中理解文本、生成更多的文本解读，导致大规模语言模型中蕴含的知识并不能直接以量化的形式呈现。思维性作为较难定义的量化指标，更加依赖于人类的主观感受，因此ChatGPT在思维性体验指标的量化上效果更差。图4展示了针对某一个篇章，ChatGPT给出的错误思维性体验标签，可以看出，ChatGPT“有理有据”地回答了文本的思维性标签，但它在量化时受到了量化任务要求的影响，倾向于过度理解文章内容，从而导致混淆了“哲理性、思辨性”的含义。

文本X：我国古代有一位女英雄，名叫花木兰。那时候，北方经常发生战争。一天，朝廷下达了紧急征兵的文书。木兰见到上面有父亲的名字，焦急万分。她想：父亲年老多病，难以出征；弟弟又小，还不够当兵的年龄。自己理应为国为家分忧。她说服了家人，女扮男装，替父从军。木兰告别了亲人，披战袍，跨骏马，渡黄河，过燕山，来到了前线。在多年征战中，她为国立下了赫赫战功。木兰胜利回乡后，脱下了战袍，穿上了心爱的女装。将士们前来探望她，这才惊讶地发现，昔日英勇善战的花将军，竟是位文静俊美的姑娘。
提示1：文本中蕴含哲理内容或者思辨性内容则标注思维性体验值phi为1，若不含有则标注为0。以下是我给你的几个示例，[……]，[……]，[……]，[……]，[……]，请先学习这些示例，以完成下一步工作。
提示2：根据学到的标注方法和规则，对以下文本标注phy值：“[文本X]”。

ChatGPT：phi为**1**。这篇文本表达了一个女英雄因国家需要而舍弃自己的安逸生活去从军，最终立下赫赫战功的故事。这个故事蕴含的哲理是，为国家、家庭和正义而奋斗是值得尊重和赞扬的。

图 4. ChatGPT错误标注样例

表3展现了图4中的文本的真实标签、ChatGPT预测结果和我们的标签线性化监督方法下模型的预测结果。可以看到，ChatGPT不善于分清这五种阅读体验的界限和各个量化值的含义，容易对体验值造成错误的标注，在本条标注中，对文采性、思维性和情感性的体验都产生了偏差，而我们的方法得到的预测值相对而言准确率更高。

模型名称	理解性	文采性	道德性	思维性	情感性
实际标注标签	1	0	1	0	0
ChatGPT	1	1	1	1	1
Bert+线性	1	0	2	0	0

表 3. 模型预测样例

同时，我们采用多任务协同计算的架构实现了阅读体验量化任务，并验证了三种大模型增强的方法对于模型性能的提升效果。基于Bert的多任务协同计算方式不仅统一了整个框架从而减少了参数量，由表4我们也可以对比发现其在各维度的阅读体验量化任务上的F1值均高于单任务的结果。由此我们可以得出结论，多任务协同计算的方式比单任务独立计算的效果更好，这证实了阅读体验彼此之间存在的正向影响。

模型名称	理解性	文采性	道德性	思维性	情感性
	P/R/F1	P/R/F1	P/R/F1	P/R/F1	P/R/F1
ChatGPT	0.56/0.45/0.40	0.68/0.76/0.69	0.45/0.50/0.42	0.58/0.70/0.43	0.48/0.48/0.45
Bert单任务	0.70/0.69/0.68	0.74/0.74/0.72	0.42/0.47/0.44	0.84/0.62/0.67	0.49/0.52/0.50
Bert多任务	0.73/0.73/0.72	0.84/0.74/0.77	0.48/0.52/0.49	0.77/0.73/**0.75**	0.55/0.52/0.53
Bert+提示	0.77/0.75/0.75	0.78/0.74/0.76	0.49/0.54/0.51	0.84/0.62/0.66	0.66/0.59/0.60
Bert+联合	0.71/0.71/0.71	0.87/0.80/0.83	0.89/0.62/**0.67**	0.45/0.50/0.47	0.63/0.56/**0.68**
Bert+线性	0.79/0.78/**0.78**	0.86/0.86/**0.86**	0.81/0.60/0.60	0.78/0.69/0.73	0.75/0.60/0.64

表 4. 模型在各维度阅读体验上的分类表现

第二十二届中国计算语言学大会论文集，第364页-第376页，哈尔滨，中国，2023年8月3日至5日。

从表中不难发现，不论采取哪种策略，融入ChatGPT的知识的多任务协同计算方法的效果都在一定程度上比多任务协同计算的基础模型的效果更好。其中，标签联合监督法和标签线性化监督法分别在不同类别的量化任务中表现优于前缀标签化提示法，我们认为这可能是由于前两者以不同方式实现了整个文本在对应位置上的标记融合，这能够更好地帮助模型学习上下文中的深层属性，而前缀标签化提示法仅给予了输入信号，却缺乏标签的位置信息，使得模型无法学习到标签和自然语言之间的对应位置关系。另外，实验结果表明，在采用大模型增强方法的情况下，尽管多数体验的量化效果有所提升，但思维性体验的量化效果却比只采用多任务协同计算架构的效果要差。图5展示了基础的多任务架构和标签线性化监督方法在各维度指标上的F1值，可以看到除了思维性体验指标，其他指标均在线性方法上得到了F1值的提升，我们将在下一小节中讨论这一点。

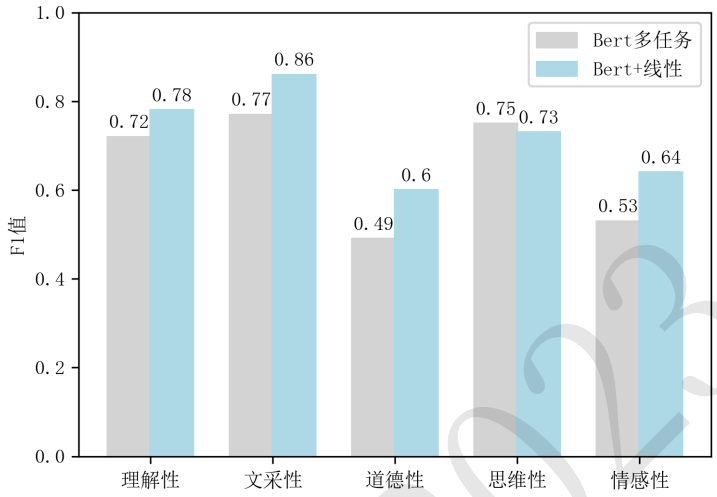


图 5. 多任务基础架构和标签线性化方法下各指标结果的F1值

5.4 探究实验分析

(1) 大模型增强对思维性体验量化的影响

基于ChatGPT增强的方法降低了多任务协同计算架构下对于思维性阅读体验的量化效果。对此，我们推测可能是思维性阅读体验的抽象性导致其在量化上更难实现，且由于该指标具备相对的独立性，对思维性的特征学习需要借助文本整体的上下文含义。因此，加入的标签越多、对其他体验的学习越深入，对思维性体验的学习越容易造成混淆。

我们在多任务协同计算的架构下对思维性阅读体验量化展开了单独的研究，在标签线性化监督法上分别探究了仅添加文采性触发语、道德性触发语、情感性触发语或思维性触发语本身的情况下思维性阅读体验的表现，表5展示了该实验下思维性阅读体验的计算结果。从表中我们可以发现，文采性阅读体验、道德性阅读体验标签对思维性阅读体验计算的结果产生了负面影响，这可能是由于文采性和道德性阅读体验标签同思维性阅读体验的联系较少，添加这两类标签导致了多余噪声的产生。再者，思维性体验作为一种具备开放性的属性，是一种无限集合，然而文采性和道德性相关的触发语能够在一定程度上被划分为有限的几种类型，有限集合的知识融入对无限集合相关的指标量化有损。

模型名称	P	R	F1
全部标记	0.78	0.69	0.73
仅文采性标记	0.76	0.62	0.65
仅道德性标记	0.71	0.61	0.64
仅情感性标记	0.77	0.73	0.75
仅思维性标记	0.71	0.61	0.64

表 5. 其他触发语在标签线性化监督法中对思维性阅读体验的影响

此外，我们发现思维性标签本身也易对结果产生负面的影响，这可能是由于思维性阅读体验非常依赖于长距离的上下文判断，导致标注的阅读体验触发语具有多样性、特殊性的特点，因此此类标签的添加不利于模型的泛化性。图6是ChatGPT标注的思维性体验触发语的示例，可以看到，思维性体验触发语和文本本身的相关性极强且跨度较大，ChatGPT对思维性的理解较为灵活，在两个不同的文本中标注的表达完全不同。

标注1：无论如何，我不能使家乡的孩子们失望，我终于拿起了笔。请原谅，我今年不能回家乡，并不是不愿意看望你们，正相反，我多么想看见你们天真的笑脸，多么想听见你们歌唱般的话语，但是我没有体力和精力支持这样一次长途的旅行。那么，就让这封信代替我同你们见面吧。不要把我当作什么杰出人物，我只是一个普通人。我写作，不是我有才华，而是我有感情，对我的祖国和同胞有无限的爱，我用作品表达我的这种感情。我今年八十七岁，我**思索**，我**追求**，我终于明白**生命的意义**在于奉献而不在于享受。我在回答和平街小学同学们的信中说：“我愿意再活一次，重新学习，重新工作，让我的生命开花结果。”

标注2：认识大自然文字的人，立即会说，它不是人搬来的，而是冰搬来的。那些冰块从寒冷的北方“爬”过来，沿路把大大小小的石块带着一起走。这是好久好久以前的事了，当时这儿根本就没有森林。周围的森林是后来才长起来的。要学会**认识大自然的文字**，从小就应当到树林里或者田野上走走，注意观察。假如有什么不明白的地方，应再到书里去寻找，看那里有没有解释。你还应该去请教有学问的人：这是什么石头？这是什么树？总是**坐在家里的人，永远不会懂得大自然的文字**。”

图 6. ChatGPT对思维性体验的标注样例

(2) 思维性体验量化任务和其他任务之间的关系

我们在探索中发现了思维性体验的特殊性，其特殊性不仅在于知识表示形式上的特殊，更在于体验形成过程的特殊。我们认为思维性体验同其他任务不同，需要进行跨度较大的上下文学习与总结，因此我们探索了思维性体验量化任务和其他任务之间的关系。我们分别用多任务协同计算基础架构和标签线性化监督方法实现了去除思维性量化任务的实验，表6是实验结果的展示。

从表中我们可以看出，去除思维性任务后，其他任务的F1值有所提升。但在多任务协同计算的基础架构下，思维性任务的效果要好于单独建模思维性的量化效果。实验结果表明，与其他任务的联合学习对思维性任务的量化有一定的帮助，但思维性任务会造成其他任务量化性能的下降。这说明思维性体验除了在知识上与其他任务不同，在表示上也和其他任务有异，该任务需要在长上下文中学习对应的总结性特征，从而对其他体验的量化产生干扰。而其他任务的知识虽然与思维性无关，但量化过程中能够产生的多维度总结性表示，对思维性体验的量化有益。

模型名称	理解性	文采性	道德性	思维性	情感性
	P/R/F1	P/R/F1	P/R/F1	P/R/F1	P/R/F1
Bert单任务	-	-	-	0.84/0.62/0.67	-
Bert多任务	0.73/0.73/0.72	0.84/0.74/0.77	0.48/0.52/0.49	0.77/0.73/0.75	0.55/0.52/0.53
无思维性	0.79/0.78/0.78	0.86/0.86/0.86	0.52/0.52/0.52	-	0.64/0.58/0.60
线性+无思维性	0.78/0.79/0.76	0.88/0.87/0.87	0.67/0.67/0.67	-	0.76/0.60/0.62

表 6. 去除思维性体验任务后各维度体验上的分类表现

6 结论

针对中文文本分级和读物推荐中存在的标准单一、体系缺失的现状，我们提出了中文篇章多维度阅读体验量化体系，以完善中文文本分级及推荐的依据和标准。基于此，我们构建了中文篇章多维度阅读体验数据集CMRED。我们又提出了三种融入ChatGPT知识的方式来实现基于大模型增强的阅读体验量化方法，分别是前缀标签化提示法、标签联合监督法和标签线性化监督法。这三种增强方式在后续实验中的表现均超出了基线模型，证明了增强方式有效性的同时，也说明了大模型知识对于阅读体验量化任务的正向作用。同时，我们使用构建的体系及数据集探究了ChatGPT在阅读体验量化上的表现，发现ChatGPT更擅长自然语言的理解和生成，但不善于对深层的特征进行量化。未来，我们将会继续探索更多阅读体验的指标，同时在基于大模型增强的方法上作出其他尝试，而对于探究实验中发现的思维性指标特殊性的持续研究，也是我们未来工作的重点。

参考文献

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online, November. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

J. Lafferty, A. Mccallum, and Fcn Pereira. 2002. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *proceedings of icml*.

Bruce W Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

Yi Li, Dong Yu, and Pengyuan Liu. 2022. CLGC: A corpus for Chinese literary grace evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5548–5556, Marseille, France, June. European Language Resources Association.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Danielle S McNamara, Max M Louwerse, Philip M McCarthy, and Arthur C Graesser. 2010. Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Simin Rao, Hua Zheng, and Sujian Li. 2021. 阅读分级相关研究综述(a survey of leveled reading). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 689–702, Huhhot, China, August. Chinese Information Processing Society of China.

Karen A Schriver. 2000. Readability formulas in the new millennium: what's the use? *ACM Journal of Computer Documentation (JCD)*, 24(3):138–140.

Dean R Smith et al. 1989. The lexile scale in theory and practice. final report. *analysis of variance*.

Daniel G Tracy. 2017. Libraries as content producers: How library publishing services address the reading experience. *College & Research Libraries*, 78(2):219.

Hongrui Wang, Chang Liu, and Dong Yu. 2020. 面向人工智能伦理计算的中文道德词典构建方法研究(construction of a Chinese moral dictionary for artificial intelligence ethical computing). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 539–549, Haikou, China, October. Chinese Information Processing Society of China.

Siyuan Wu, Jianyong Cai, Dong Yu, and Xin Jiang. 2018. A Survey on the Automatic Text Readability Measures(文本可读性的自动分析研究综述). 中文信息学报, 32:1–10, December.

何应艳. 2016. 小学语文阅读教学中美育的实施. 硕士, 云南师范大学.

刘苗苗, 李燕, 王欣萌, 甘琳琳, and 李虹. 2021. 分级阅读初探:基于小学教材的汉语可读性公式研究. 语言文字应用, (2):116–126.

唐玉玲, 张宇飞, and 于东. 2022. 结合深度学习和语言难度特征的句子可读性计算方法. 中文信息学报, 36(2):29–39.

廖圣河and 陈怡瑾. 2021. 小学生语文阅读感受力的构成和培养. 语文建设, (20):19–23.

杨春and 张秋月. 2022. 文本阅读体验应是语文课堂教学的发端与根本. 中华活页文选:教师, (7):3.

池春红. 2022. 激发阅读兴趣,增强阅读体验,促进深度学习——初中语文名著阅读教学的有效性探究. 亚太教育, (21):91–94.

程勇, 徐德宽, and 董军. 2020. 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究.语言文字应用. 语言文字应用, (1):132–143.

谭继雄. 2020. 公共图书馆阅读体验的影响因素研究——基于青少年群体阅读风格视角. 图书馆学刊, 42(8):68–72.

骆香莹. 2022. 基于多元语言特征与眼动特征的中文可读性自动评估. 硕士学位论文, 北京语言大学.

A 思维性体验的提示模板和ChatGPT响应

提示1：文本中蕴含哲理内容或者思辨性内容，能引发读者哲思的，则标注思维性体验值phi为1，若不含有则标注为0。以下是我给你的几个示例，示例1：[文本：这时候，我清清楚楚听见老师说，按照进化论的观点，母腹中的胎儿再现了从简单生命进化成人的过程。当时教室里安静得出奇，大家都默不作声。我忽然想起了自己的发现，情不自禁地笑出了声音。老师狠狠地瞪了我一眼。“费奥多罗夫！.....你笑什么？再笑就从教室里出去！”“奥尔加·伊万诺夫娜，我.....我想起了自己的发现.....”教室里一阵笑声。奥尔加伊万诺夫娜气得脸色苍白，大步朝我走来。“费奥多罗夫！.....你立刻从教室里出去！.....”我的脸由于困窘和羞愧一下子涨得通红。这时候我意识到，老师误解了我的笑声，以为我的笑不怀好意。幸亏她没有容我解释，不然的话，同学们听见我说自己三年前就发现了进化论，还不笑塌房顶！不过，被轰出教室，站在外面，我倒想出了一条和自我安慰的理由，我明白了——世界上重大的发明与发现，有时还面临着受到驱逐和迫害的风险。此文本的phi值为1。]，示例2：[文本：春雨，像春姑娘纺出的线，轻轻地落到地上，沙沙沙，沙沙沙.....田野里，一群小鸟正在争论一个有趣的问题：春雨到底是什么颜色的？小燕子说：“春雨是绿色的。你们瞧，春雨落到草地上，草就绿了。春雨淋在柳树上，柳枝也绿了。”麻雀说：“不对，春雨是红色的。你们瞧，春雨洒在桃树上，桃花红了。春雨滴在杜鹃丛中，杜鹃花也红了。”小黄莺说：“不对，不对，春雨是黄色的。你们看，春雨落在油菜地里，油菜花黄了。春雨落在蒲公英上，蒲公英花也黄了。”春雨听了大家的争论，下得更欢了，沙沙沙，沙沙沙.....此文本的phi值为0。]，请先学习这些示例，以完成下一步工作。

ChatGPT回答：根据您提供的示例，示例1中的文本涉及到思辨性内容和哲理思考，因此可以标注为思维性体验值(phi) 为1。这是因为文本描述了主人公对自己的发现以及对被误解的感受的思考，涉及到了对人类发明与发现所面临的风险和迫害的理解。示例2中的文本主要描述了小鸟们对春雨颜色的争论，没有明显的哲理或思辨性内容，因此可以将其标注为思维性体验值(phi) 为0。这段文本更多地展现了小鸟们之间的对话和辩论，而不是探讨深层次的思考或哲学观点。根据上述解释，示例1的phi值为1，示例2的phi值为0。

提示2：根据学到的标注方法和规则，对以下文本标注phi值：东汉的时候，有个青年叫陈蕃，他虽志存高远，喜好读书，但生活上却很懒散，连自己的书房也很少收拾打扫。有一天，一位叫薛勤的人来拜访陈蕃的父亲，正巧他父亲外出不在，只有陈蕃在家。陈蕃见来了客人，就从自己的书房里出来接待，告诉客人父亲现在不在家。没料到薛勤一边与他寒暄，一边随意地走进了陈蕃的书房。薛勤一进书房，着实吓了一跳。屋子里乱七八糟，桌子上书呀，杂物呀，什么都有，上面满是灰尘；地上到处是垃圾；墙壁止有不少蜘蛛网。屋里不仅肮脏，空气也很污浊。薛勤看到后，不禁皱起了眉头。他对陈蕃说：“年轻人，你的屋子怎么弄得这么乱？为什么不把屋子打扫干净呢？”没想到陈蕃却满不在乎地说：“大丈夫活在世上，要干的是轰轰烈烈的大事业，要扫除的是天下一切不平之事，哪里会花心思去清扫小小的一间屋子呢？”他说得理直气壮，很有些得意的神气。听着陈蕃这样的回答，薛勤心里暗想：此人虽年少而有大志，但连小事都不愿意做，又怎么能做成大事呢？于是他盯着陈蕃，反问道：“年轻人，你连一间小小的屋子都不扫，又怎么去扫天下呢？”薛勤走后，陈蕃沉思起来，觉得他的话很有道理。

ChatGPT回答：根据您的示例，我将对文本进行标注。这篇文本的phi值为1。这篇文本含有哲理，通过薛勤的质问，告诫读者做事要有条理，不要忽略小事。

第二十二届中国计算语言学大会论文集，第364页-第376页，哈尔滨，中国，2023年8月3日至5日。

融合文本困惑度特征和相似度特征的推特机器人检测方法*

王钟杰 张朝文 丁文琪 付雨濛 单丽莉 刘秉权
哈尔滨工业大学
{wzhongjie07, zhangzhaowen02, dingwenqi2020, cherry1232023} @163.com
{shanlili, liubq} @hit.edu.cn

摘要

推特机器人检测任务的目标是判断一个推特账号是真人账号还是自动化机器人账号。随着自动化账号拟人算法的快速迭代，检测最新类别的自动化账号变得越来越困难。最近，预训练语言模型在自然语言生成任务和其他任务上表现出了出色的水平，当这些预训练语言模型被用于推特文本自动生成时，会为推特机器人检测任务带来很大挑战。本文研究发现，困惑度偏低和相似度偏高的现象始终出现在不同时代自动化账号的历史推文中，且此现象不受预训练语言模型的影响。针对这些发现，本文提出了一种抽取历史推文困惑度特征和相似度特征的方法，并设计了一种特征融合策略，以更好地将这些新特征应用于已有的算法模型。本文方法在选定数据集上的性能超越了已有的基准方法，并在人民网主办、传播内容认知全国重点实验室承办的社交机器人识别大赛上取得了冠军。

关键词： 推特机器人检测 ； 预训练语言模型 ； 文本困惑度分析 ； 文本相似度分析

Twitter robot detection method based on text perplexity feature and similarity feature

ZhongjieWang ZhaowenZhang WenqiDing YumengFu LiliShan BingquanLiu
Harbin Institute of Technology
{wzhongjie07, zhangzhaowen02, dingwenqi2020, cherry1232023} @163.com
{shanlili, liubq} @hit.edu.cn

Abstract

The goal of the Twitter robot detection task is to determine whether a Twitter account is a real person account or an automated robot account. With the rapid iteration of automated account impersonation algorithms, it becomes increasingly difficult to detect the latest categories of automated accounts. Recently, pre-trained language models have shown excellent performance in natural language generation tasks and other tasks. When these pre-trained language models are used for automatic Twitter text generation, they pose significant challenges for Twitter robot detection tasks. This study found that the phenomenon of low perplexity and high similarity has always appeared in historical tweets from automated accounts in different eras, and this phenomenon is not affected by the pre-trained language model. In response to these findings, this paper proposes a method for extracting perplexity and similarity features from historical tweets, and designs a feature fusion strategy to better apply these new features to existing algorithm models. The performance of the method in this paper on the selected dataset exceeded the existing benchmark method, and won the championship in the social robot recognition competition which hosted by People's Daily and undertaken by the National Key Laboratory for Content Awareness Communication.

Keywords: twitter robot detection , pre-trained language model , text perplexity analysis , text similarity analysis

* 基金项目：国家重点研发计划（项目编号：2021YFF0901600）；中央高校基本科研业务费专项资金资助（项目编号：2022FRFK0600XX）

1 引言

推特社交机器人检测的目的是给定一名用户的历史推文和关系网络等信息（如图 1所示）判断其是真人账号还是自动化机器人账号。目前已经有一些推特机器人检测工作讨论了如何利用各种用户信息，如利用元信息进行推特机器人检测(Efthimion et al., 2018; Hayawi et al., 2022)；利用历史推文信息进行推特机器人检测(Derhab et al., 2021; Lundberg et al., 2019; Cresci et al., 2017)；利用关系网络信息进行推特机器人检测(Feng et al., 2022; Feng et al., 2021b)。然而，这些工作在检测最新一代自动化账号上的表现仍有进步的空间(Cresci, 2020)，其中一部分原因是这些工作对于历史推文信息的利用程度不够，导致自动化账号在更新其拟人算法后，这些检测算法原本所利用的信息可能会失效甚至误导检测结果。本文在分析具有代表性的几代自动化账号的基础之上发现，自动化账号的历史推文信息始终拥有较低的困惑度和较高的相似度，并且不随拟人算法的更新而变化。因此，为了更好地检测最新一代的自动化账号，本文提出了一种抽取历史推文困惑度特征和相似度特征的方法，并设计了一种特征融合策略，以更好地将这些新特征应用于已有的算法模型。

推特机器人检测工作的难点在于机器人拟人算法的快速迭代特性。Cresci (2020)总结了过去十年社交媒体机器人的发展情况，研究证据表明这些机器人之间存在着一种进化机制，进化的机器人通过协同完成任务，即一个群组的机器人被同一个人所控制，通过协同发布信息，从而实现更好的伪装。基于这一进化特点，最近的推特机器人检测算法更加注重关系网络信息，通过对这些关系网络信息进行建模，来提升模型检测效果(Feng et al., 2022; Feng et al., 2021b)。然而这些最新的检测工作也带来了新的问题。首先，关系网络信息的建模对收集数据提出了更高的要求，算法要求同时收集用户和相关用户的社交信息才能对关系网络信息完成建模。其次，这些基于关系网络信息的检测算法对关系网络结构十分敏感，如果关系网络过于稀疏或者当自动化账号故意增加与真人账号的互动之后，那么基于关系网络信息的检测算法可能会引入过多噪声，从而造成检测结果的误判。

研究表明，在推特机器人检测算法中，相对于其他信息，历史推文信息拥有更高的贡献权重。从直觉上来看，人们首先会根据一则账号发布的推文内容来怀疑其是否为自动化账号，如果产生怀疑，人们会进一步调查该账号的相关资料以及其与他账号的相关互动，以做出更准确的判断(Feng et al., 2021a)。第一代、第二代的机器人受技术限制，发布的推文内容很容易被人类识别。后来，由于大规模预训练语言模型的出现以及其在文本生成领域的成功应用(Devlin et al., 2018; Brown et al., 2020)，许多基于此项技术的推特机器人生成的推文内容已经很难被人类识别。但是这些生成的推文内容都有一个共同的特点，即相较于真实的人类推文，这些生成推文拥有更低的文本困惑度。这主要是因为早期的生成推文由于技术限制，往往只具有简单的句法结构或语义信息，而基于大规模预训练语言模型的生成推文则是受到其优化目标的影响，导致其更偏向于生成拥有更低文本困惑的推文内容。除此之外，相较于真实的人类推文，这些生成的推文内容之间往往具有更高的相似度，这主要是由推特机器人的任务性质所决定的，这些机器人只需要完成特定的目的，所以它们的推文只集中在某些特定话题下，而真实用户则更关注快速迭代的热点话题。

基于上述分析，本文在设计推特机器人检测算法模型时加入了文本困惑度特征和相似度特征。具体来说，本文利用Radford et al. (2018)提出的模型计算用户历史推文的文本困惑度，为每一个推特用户构建一个困惑度向量，并且统计了这些困惑度向量的五项数值特征，即最大值、最小值、均值、方差、总和，进而通过困惑度向量和其统计特征构建推特用户的困惑度特征。本文还利用Devlin et al. (2018)提出的模型计算这些推文的嵌入向量，再计算这些嵌入向量两两之间的相似度，得到一个相似度矩阵，最后统计相似度矩阵的均值，将其作为推特用户的相似度特征。本文利用一个简单的三层MLP模型验证了特征提取方法的有效性，此外，本文结合Feng et al. (2021b)提出的模型，提出并验证了一种特征融合策略。

本文的主要成果：

（1）分析用户历史推文信息的特点，提出了一种抽取推文文本困惑度和相似度特征的方法，并在数据集上验证了特征抽取方法的有效性。

（2）结合BOTRGCN模型(Feng et al., 2021b)，提出了一种特征融合策略，并在数据集上验证了特征融合策略的有效性。

（3）利用新的特征和特征融合策略，在一定程度上解决了机器人拟人算法快速迭代的问题。

（4）本文方法在选定数据集上的性能超越了已有的基准方法，并且在人民网主办、传播内容认知全国重点实验室承办的社交机器人识别大赛上取得了冠军。

2 相关工作

利用传统机器学习的推特机器人检测方法可以检测出早期的自动化账号。这些工作从数据中手工提取特征，包括元信息手工特征(Efthimion et al., 2018; Hayawi et al., 2022)、历史推文信息手工特征(徐帅帅et al., 2017; Derhab et al., 2021; Lundberg et al., 2019; Cresci et al., 2017)、关系网络信息手工特征(张玄and 李保滨, 2022)等,这些特征会被送入SVM、决策树、无监督聚类等传统机器学习算法中进行推特机器人检测。这些传统机器学习工作在提取特征时消耗了昂贵的资源，但是在检测最新一代的自动化账号时，并没有表现出很好的效果，这主要是因为提取的手工特征对账号信息的挖掘程度不够，当推特机器人更新其拟人算法后，已挖掘的特征会失效，甚至导致结果的误判。

深度学习技术的出现，使得研究者们能够进一步挖掘已有账号信息的特征。当CNN、LSTM以及后续的GPT(Radford et al., 2018)、BERT(Devlin et al., 2018)等深度学习模型出现之后，在传统机器学习中需要手工提取的特征逐渐被深度学习模型输出的嵌入向量所代替。Wei and Nguyen (2019)等人提出了第一个用深度学习模型进行词嵌入来完成推特机器人检测工作的模型，并在Cresci-2017数据集上达到了93%的准确度。Chen et al. (2022)等人则是利用预训练语言模型BERT对用户推文内容进行词嵌入，并论证了BERT得到的历史推文信息嵌入向量对推特机器人检测效果的提升有很大帮助。这些基于深度学习的模型和基于传统机器学习的模型之间的效果对比，说明了如果能够进一步挖掘已有账号信息的特征，就能够进一步提升模型的检测效果。

Cresci (2020)指出目前推特上活跃的自动化账号是第三代自动化账号，它们通过协同工作来实现更好的伪装。Kipf and Welling (2016)首次提出了图卷积的概念，CNN和RNN能够帮助我们从二维和一维的欧式空间数据中提取相应的特征，而图卷积的出现，则使我们能够从不规则的图结构中提取相应的特征。为了能够对最新一代自动化账号的异常协同工作进行识别，一些研究者将图卷积技术应用于用户关系网络信息的建模，他们将用户信息视为节点，将用户之间各种类型的互动关系视为不同类型的边，并用图卷积技术更新关系网络中节点和边的信息(Ali Alhosseini et al., 2019; Feng et al., 2022; Feng et al., 2021b)。然而这种基于图卷积技术的推特机器人检测工作也引入了新的问题，比如，对关系网络信息进行建模需要收集额外的数据，以及最终检测效果过度依赖于用户关系图的结构完整性。从本质上来看，这些对异常协同工作进行建模识别的检测算法相当于引入了新的特征，与充分挖掘已有账号信息的特征属于两种不同的思路。

在充分挖掘已有账号信息方面，还有一些研究者的工作值得关注。如Lei et al. (2022)提出了一种利用多模态技术挖掘历史推文信息中的语义不一致性信息，进而提升推特机器人检测效果的算法模型，该模型不仅关注历史推文信息中的文本信息，还关注推文中的图片和音视频信息。此外，ChatGPT[0]的出现使得由机器生成的文本更加难以和人类真实的文本作区分。Guo et al. (2023)提出了一种用于检测ChatGPT生成文本的工具，其核心思想是相较于人类真实的文本，由文本生成模型生成的文本有更低的文本困惑度。Gehrmann et al. (2019)也在他们的工作中论证了文本生成模型在解码阶段总是选择预测概率较高的词输出，这进一步导致了通过文本生成模型生成的文本拥有更低的文本困惑度。这些检测ChatGPT生成文本的工作，也为推特机器人检测工作带来新的思路。

综上所述，充分挖掘已有账号信息特征和引入新的特征都可以提高推特机器人检测的效果。但考虑到以下现实:

（1）引入新的特征会带来新的问题。

（2）历史推文信息比其余账号信息蕴含更多有助于检测的内容，且历史推文信息还没有被充分地利用。

（3）各种文本生成模型在推文自动生成上被广泛应用。

[0]https://chat.openai.com/

本文提出了一种抽取推文困惑度特征和相似度特征的方法，并设计了一种特征融合策略。新的特征以及特征融合策略兼顾了两种不同的问题解决思路，同时也在一定程度上对以上现实做了回应。实验结果也证明，引入的两个基于历史推文信息的特征比前人所提取的特征更有效，并且融合了这两者特征的检测模型能够在实验数据集上表现出更优秀的结果。

3 融合文本困惑度和相似度特征的推特机器人检测方法

3.1 任务定义

推特机器人检测任务旨在通过用户信息判断一个用户是真人账号还是自动化机器人账号，用户信息包含的内容如下。

个人简介信息：$B = \{b_i\}_{i=1}^{L}$，其中b_i代表一个单词。个人简介信息是一段由用户编写的文本，用来对用户所持有的账号做简单说明，如一些天气预报账号可能会在其个人简介中说明该账号是一个转发天气信息的自动化账号。

历史推文信息：$T = \{t_i\}_{i=1}^{M}$，其中$t_i = \left\{w_1^i, ..., w_{Q_i}^i\right\}$，$w_{Q_i}^i$代表一个单词。历史推文信息是用户在推特上进行活动产生的主要信息之一，用户往往通过发布推文来表达自己的观点。

元信息：$P = \{P^{num}, P^{cat}\}$，其中P^{num}表示数值元信息，P^{cat}表示分类元信息。元信息是用户创建时，推特平台为其分配的一些固有信息。数值元信息是指一些由数字构成的元信息，如：账号已激活时间、昵称长度等；分类元信息是指一些由布尔变量构成的元信息，如：账号是否经过认证；账号是否使用默认头像等。

关系网络信息：$N = \left\{N^f, N^t\right\}$，其中$N^f = \left\{N_1^f, ..., N_u^f\right\}$，表示该用户的$u$个追随者用户，$N^t = \{N_1^t, ..., N_v^t\}$，表示该用户的$v$个关注用户。关系网络信息构成一个社交网络，这些网络表明了哪些用户具有相似的话题爱好。

3.2 模型整体架构介绍

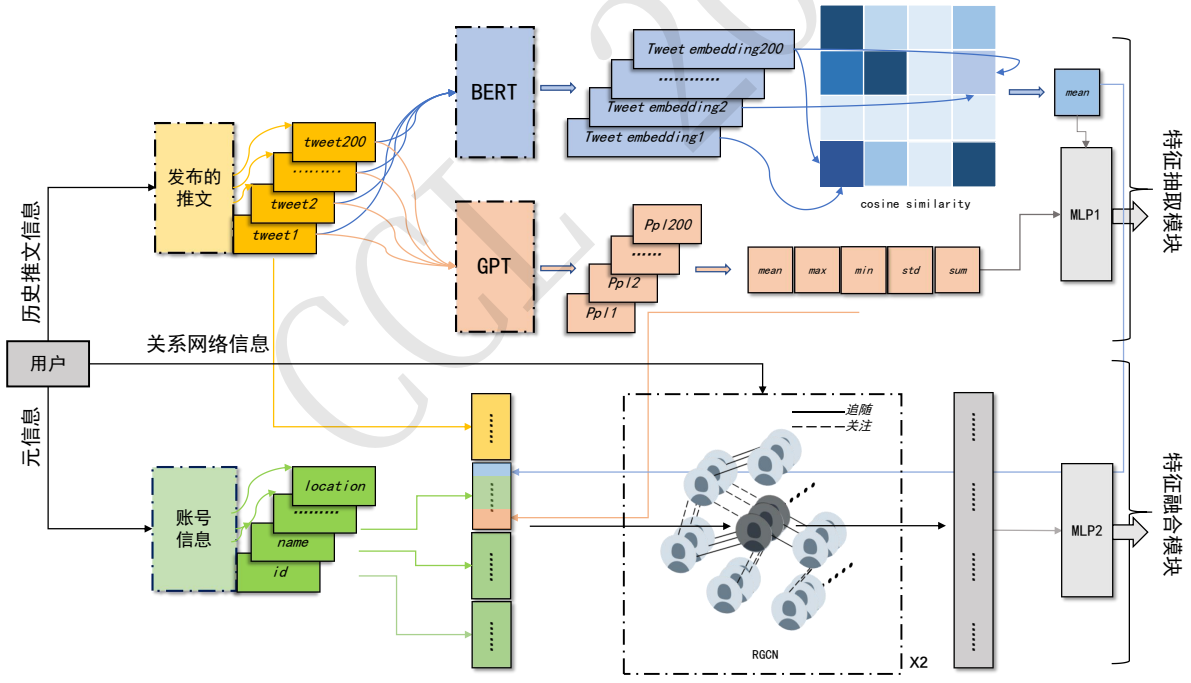


Figure 1: 模型架构图

本文提出了一种融合文本困惑度特征和相似度特征的推特机器人检测模型，如图 1所示。该模型主要由两部分组成：1）文本困惑度特征和相似度特征抽取；2）特征融合。第一部分展示了文本困惑度特征和相似度特征的抽取过程，第二部分则展示了特征融合策略，利用该策略，检测模型将抽取出的特征与BOTRGCN模型(Feng et al., 2021b)进行了融合。

各部分的具体工作流程如图 2所示。主要工作流程由三个阶段组成，分别是：特征抽取阶段、图处理阶段以及结果预测阶段。特征抽取阶段，模型对于输入的每个用户信息进行特征抽取；图处理阶段，模型通过抽取的特征初始化关系图中的每一个节点，通过关系网络信息初始化图中的每一条边，再通过图卷积技术更新图中的每一个节点，最后将每一个节点的特征送入预测网络；结果预测阶段，模型通过图卷积输出的某一节点的特征判断该节点是真人账号还是自动化账号。后续的 3.3章节和 3.4章节将对模型的工作细节进行更详细的阐述。

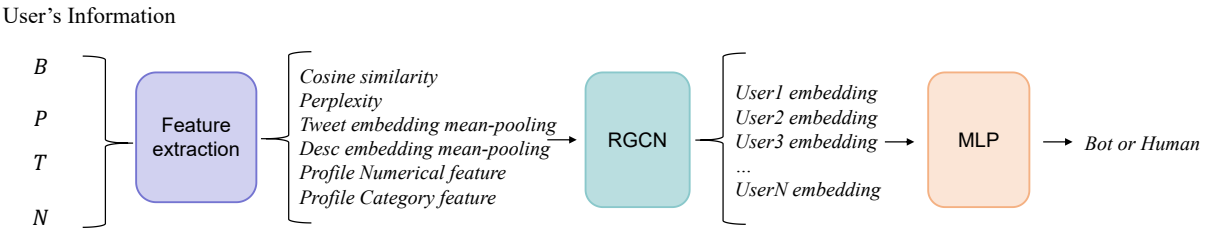


Figure 2: 算法工作流程图

3.3 特征抽取

文本困惑度特征：困惑度（perplexity）常被用来衡量一个预训练语言模型的能力，其计算过程见公式 (1)。显然，一个优秀的预训练语言模型应该在其相应的测试集上拥有更低的困惑度。

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}} \tag{1}$$

本文使用GPT-1预训练语言模型(Radford et al., 2018)计算用户历史推文信息的困惑度，具体的说，检测模型将历史推文信息依次输入至GPT-1中计算其困惑度，计算完某一用户的所有历史推文信息困惑度，可以得到困惑度向量l'，$l' \in R^{1\times 200}$。检测模型还会计算l'的最大值、最小值、方差、均值和总和这五个统计特征，将其分别记做l'_{max}、l'_{min}、l'_{std}、l'_{mean}、$l'_{sum} \in R$，最终检测模型抽取的困惑度特征$l = concat(l'_{max}, l'_{min}, l'_{std}, l'_{mean}, l'_{sum})$，$l \in R^{1\times 5}$。

文本相似度特征：相似度（similarity）常被用来衡量文本内容之间的语义相似性，常用的计算方法有余弦相似度、Jaccard相似度等。本文选用余弦相似度作为计算公式，计算过程见公式 (2)。

$$\cos\theta = \frac{\vec{A}\cdot\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|} = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{2}$$

本文使用BERT预训练语言模型(Devlin et al., 2018)计算这些用户历史推文信息的文本相似度，具体的说，检测模型将历史推文信息依次输入至BERT，使用BERT输出的[CLS]向量作为一条历史推文信息的嵌入向量表示，再计算这些嵌入向量之间的余弦相似度，进而得到相似度矩阵C'，$C' \in R^{200\times 200}$。检测模型还计算了C'的均值C'_{mean}，$C'_{mean} \in R$，将$C = C'_{mean}$，作为最终的相似度特征。

粗粒度的历史推文信息特征：BOTRGCN对历史推文信息进行了粗粒度的特征提取，具体的说，BOTRGCN将历史推文信息依次输入至预训练语言模型RoBERTa(Liu et al., 2019)，对于每一条推文，通过对RoBERTa输出的[TOKEN]向量进行平均池化得到该条推文的嵌入向量，再将推文的嵌入向量进行平均池化，得到历史推文信息特征f_t，$f_t \in R^{1\times 768}$。

实际上，推特中的推文涉及英语、中文在内的多种语言，除此之外，其语言习惯也受一些网络文化影响。因此，本文还尝试了用TwHIN-BERT(Zhang et al., 2022)对历史推文信息进行粗粒度的特征提取。相较于RoBERTa，TwHIN-BERT由推特上的七十亿条推文数据训练而成，并且具备处理多语种的能力，这使其能够更好地提取历史推文信息特征。$f_{t'}$，$f_{t'} \in R^{1\times 1024}$就是检测模型通过TwHIN-BERT提取的历史推文信息特征。

个人简介信息特征：BOTRGCN将个人简介信息输入至RoBERTa，通过对RoBERTa输出的[TOKEN]向量进行平均池化得到个人简介信息特征f_b，$f_b \in R^{1\times768}$。

元信息特征：对于元信息，BOTRGCN将这些离散的数字或布尔变量输入至一个三层的MLP模型，通过MLP进行编码，得到两个稠密的嵌入向量f_{num}，$f_{num} \in R^{1\times32}$；$f_{cat}$，$f_{cat} \in R^{1\times32}$，其分别表示数值元信息特征和分类元信息特征。

3.4 图操作和结果预测

图初始化操作和特征融合策略：BOTRGCN在对关系网络信息进行建模时，将用户视为节点，将用户之间的关注与被关注关系视为边，从而构建出一张异质图，其中节点特征由 3.3中提取的特征构成。具体的说，针对某一用户节点x_i，其初始化特征$x_i^0 = concat(MLP_1(f_b), MLP_2(f_t), f_{num}, f_{cat})$，$x_i^0 \in R^{1\times128}$，$MLP_1$和$MLP_2$为两个三层的MLP模型，通过它们进行编码，对$f_b$，$f_t$进行降维处理。

在BOTRGCN的工作基础之上，本文还将l'_{max}、l'_{min}、l'_{std}、l'_{mean}、l'_{sum}、C视为六个分类元信息，拼接至原元信息后，从而使得后续得到的f_{cat}能够融入困惑度特征和相似度特征。

图节点更新操作：采用RGCN(Schlichtkrull et al., 2018)更新图中节点特征，更新过程见公式 (3)，其中Θ_{self}和Θ_r为RGCN卷积算子需要学习的参数。

$$x_i^{(l+1)} = \Theta_{self} \cdot x_i^{(l)} + \sum_{r\in N}\sum_{j\in N_r(i)} \frac{1}{|N_r(i)|}\Theta_r \cdot x_j^{(l)}, x_i^{(l+1)} \in R^{1\times128} \tag{3}$$

结果预测：在预测时，BOTRGCN将RGCN的最后一层输出输入至头部网络，得到最终结果$\hat{y}$，$\hat{y} \in R^{1\times2}$（见公式 (4)）。训练时的损失函数采用交叉熵损失函数，并且采用$L2$正则化防止模型过拟合，Y为真实标签集合（见公式 (5)）。

$$\hat{y}_i = softmax\left(W_O \cdot x_i^L + b_O\right) \tag{4}$$

$$L = -\sum_{i\in Y}[y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)] + \lambda\sum_{w\in\Theta} w^2 \tag{5}$$

4 实验设计与分析

4.1 数据集

实验使用人民网主办、传播内容认知全国重点实验室承办的社交机器人识别大赛提供的数据作为主要数据集。该数据集收集了推特平台上的10500个账号信息，实验按照8：1：1的比例划分训练集、验证集、测试集，数据集各部分的样本分布见表 1，其中每条数据的格式如表 2所示。

所属集合	机器人样本数	真人样本数	总样本数
训练集	1714	6686	8400
验证集	204	846	1050
测试集	191	859	1050

Table 1: 样本分布

数据域	样例
个人简介信息	Ambition is preiceless
历史推文信息	[@tha..., @STU..., ...]
元信息	{id : ..., name : ..., ...}
关系网络信息	{following : [...] , follower : [266181184, ...]}

Table 2: 数据集格式

需要具体说明的是，每条数据中的历史推文信息被组织成列表的形式，每个列表包含了该账号发布的二百条历史推文；元信息被组织成字典的形式，按照填充值数据类型的不同，元信息可以分为数值元信息（见表 3）和分类元信息（见表 4）；关系网络信息被组织成字典的形式，以$follower$（粉丝）为例，其对应的列表中包含了该账号所有粉丝的id。

数值元信息特征	备注
followers	关注者数量
followings	追随者数量
favorites	点赞数量
statuses	状态数量
active_days	账户激活时间
screen_name_length	网名长度

Table 3: 数值元信息特征

分类元信息特征	备注
protected	账号是否受保护
geo_enabled	是否允许定位
verified	账号是否经过验证
contributors_enabled	是否允许有贡献者
is_translator	是否是翻译人员
is_translation_enabled	账号信息是否允许被翻译
profile_background_tile	背景图片是否平铺
profile_user_background_image	背景图片url
has_extended_profile	是否启用扩展资料
default_profile	是否使用默认个人资料
default_profile_image	是否使用默认个人资料图片

Table 4: 分类元信息特征

本文还统计了每条数据实际收录的$following$和$follower$的平均数量，并将其和数值元信息中记录的$followings$和$followers$作对比。两者实际收录的平均长度为0.58和0.64，而两者在数值元信息中记录的平均长度为3142和101027。这说明对关系网络信息进行建模时，收集一个用户的相关用户信息的代价是昂贵的。

4.2 实验设置

本文使用RoBERTa和TwHIN-BERT作为编码器对历史推文信息进行编码；使用distilRoBERTa作为编码器对个人简介信息进行编码；使用GPT-1和BERT分别提取历史推文信息的困惑度特征和相似度特征。在反向传播时，训练轮数设置为500轮；由于对关系网络建模需要全部训练样本，因此batchsize设置为训练集样本数量大小；学习率设置为0.001；$L-2$正则化系数λ设置为0.005；dropout设置为0.3；最大序列长度设置为512；并使用AdamW优化器进行梯度更新。

本文采用宏F1作为评价指标，计算如公式 (6)至公式 (8)所示，TP、FP、TN、FN分别为真阳性、假阳性、真阴性、假阴性样本。

$$F1_{marco} = \frac{2 \times precision \times recall}{precision + recall} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

4.3 实验结果与分析

本文分别设计实验，验证了困惑度特征和相似度特征相较于前人总结的特征的有效性；以及特征融合策略的有效性。

为了验证困惑度特征和相似度特征相较于前人总结的特征的有效性，本文将这些特征分成四组（如表 5所示）。

分组编号	特征	备注
group1(g1)	数值元信息特征	同表 3
group2(g2)	分类元信息特征	同表 4
group3(g3)	g1+g2	/
group4(g4)	困惑度和相似度特征	/

Table 5: 特征分组

本文试图用最简单的模型研究这些特征的有效性，因此实验直接将这四组特征输入至一个三层MLP模型中，考虑到特征维数的不同，实验只对MLP的输入维数进行了改动，并没有改动MLP的深度和宽度，实验结果如表 6所示。

分组编号	$\mathbf{F1_{macro}}$
g1	0.4420
g2	0.4500
g3	0.4688
g4	**0.4741**

Table 6: 特征有效性实验结果

表 6的实验结果显示，检测模型抽取的困惑度特征和相似度特征在简单模型上显著优于前人总结的特征及其特征的组合。这也证明了用户的历史推文信息比元信息蕴含更多的价值。

为了验证融合策略的有效性，本文在BOTRGCN的工作基础之上融合了文本困惑度特征和相似度特征，实验结果如表 7所示[1]。

算法模型	第一次	第二次	第三次	第四次	第五次	平均$\mathbf{F1_{macro}}$
BOTRGCN	0.8536	0.8539	0.8570	0.8531	0.8538	0.8543
BOTRGCN+g4	0.8445	0.8368	0.8324	0.8426	0.8485	0.8410
BOTGAT+$f_{t'}$	0.8236	0.8238	0.8254	0.8199	0.8196	0.8225
BOTGAT+$f_{t'}$+g4	0.8363	0.8390	0.8267	0.8298	0.8376	0.8339
BOTRGCN+$f_{t'}$	0.8583	0.8556	0.8577	0.8458	0.8594	0.8554
BOTRGCN+$f_{t'}$+g4	**0.8639**	**0.8612**	**0.8595**	**0.8570**	**0.8631**	**0.8610**

Table 7: 融合策略有效性实验结果

表 7的实验结果显示，融合了困惑度和相似度特征的BOTRGCN在实验数据集上取得了更好的表现。除此之外，采用TwHIN-BERT对历史推文信息进行编码也提升了检测模型的性能表现。

为了研究困惑度和相似度特征的鲁棒性，本文还采取BOTGAT(Veličković et al., 2017)作为基础模型架构，并融合困惑度和相似度特征进行实验验证，可以看到困惑度和相似度特征在不同的基础模型架构上也有良好的表现。

此外，本文分析实验结果发现BOTRGCN在初始化图节点时存在特征融合瓶颈。具体的说，BOTRGCN初始化图节点只是简单地拼接不同类型的特征，没有完全发挥它们的潜力。如表 8所示，理论上用户的个人简介信息和历史推文信息是相互独立的，但是在融合f_t和f_b特征

[1]由于PyG库中graph算子的影响，即使固定随机种子，也无法完全固定实验结果，因此本文在同一随机种子上重复实验了五次，后续涉及graph算子的实验同理

时，相较于只使用f_t特征，模型预测效果没有明显提升，这说明只是简单地拼接两个不同类型的特征并没有挖掘出f_b特征的全部潜力，为了更好地提升模型效果，需要研究者们设计出新的特征融合策略。

4.4 消融实验

为了分析个人简介信息特征、数值元信息特征、分类元信息特征、粗粒度历史推文信息特征、文本困惑度特征和文本相似度特征对最终分类结果的影响，实验在用户关系图初始化时，只选用某一个（组）特征初始化节点，从而探究其对最终结果的影响权重，实验结果如表 8所示。

选取特征(组)	第一次	第二次	第三次	第四次	第五次	平均$\mathbf{F1_{macro}}$
f_b	0.7293	0.7223	0.7234	0.7199	0.7264	0.7243
f_{num}	0.5507	0.5617	0.5623	0.5627	0.5534	0.5582
f_{cat}	0.4500	0.4500	0.4500	0.4500	0.4500	0.4500
f_t	0.8445	0.8445	0.8375	0.8438	**0.8438**	0.8428
f_t+f_b	**0.8511**	**0.8531**	**0.8585**	**0.8446**	0.8411	**0.8497**
g3	0.5446	0.5453	0.5598	0.5559	0.5624	0.5536
g4	0.6954	0.6986	0.7402	0.6988	0.7042	0.7074

Table 8: 单一特征实验结果

表 8的实验结果显示，当只采用某一个（组）特征时，f_t、$g4$这些基于历史推文信息的特征对最终结果的正面影响显著大于基于其他用户信息的特征。这验证了本文在前面的论述，即判断一个用户是否是自动化账号时，其发布的推文往往更有价值。除此之外，结合表 6，可以发现与MLP模型相比，结合BOTRGCN模型并不会使f_{cat}特征对检测效果有明显提升，这或许是因为目前的自动化账号更加倾向于盗窃真人账号的分类元信息，而不是通过自动化程序生成，从而使得模型无法捕捉到同一组自动化账号之间分类元信息的相似性。

5 总结与展望

本文为了解决推特机器人拟人算法快速迭代，以及最新一代推特机器人检测算法对数据要求高、对图网络结构敏感的问题，提出了一种融合文本困惑度特征和相似度特征的推特机器人检测算法。本文认为，在进行推特机器人检测时，用户的历史推文信息相较于其他的用户信息，会为检测工作提供更多有价值的内容，因此，本文提出一种抽取推文文本困惑度和相似度特征的方法，旨在充分挖掘历史推文信息的价值。实验结果验证了提取特征的有效性；以及这些特征对最新一代推特机器人检测算法有明显的效果提升。本文分析了数据集中的数据，发现这些数据具有两个特点：多语种和语言习惯网络化。为了更好地捕捉文本特征，本文采用由七十亿条推特数据训练的多语种预训练语言模型TwHIN-BERT代替RoBERTa作为检测模型的编码器。除此之外，通过实验验证，本文还发现目前的算法在对不同类型的特征进行融合时采取的策略过于简单，这会导致特征融合瓶颈现象。虽然本文没有详细讨论，但仍值得注意的一点是，推特机器人种类的多样性(张玄and 李保滨, 2022)也是推特机器人检测算法需要应对的问题，像僵尸粉丝机器人这种类型的自动化账号可能很少发布推文，导致抽取出的历史推文信息特征质量较低，进而影响模型的性能表现。

综上所述，本文认为未来的推特机器人检测工作有如下几个发展方向：

- 提出更好的特征融合策略。目前的工作已经总结出各种类型的用户特征，但是忽视了对于特征融合策略的设计，更好的特征融合策略或许可以更好地利用已有的各种特征。

- 考虑不同种类的推特机器人。大部分推特机器人以发布各种有害推文为主，但是研究工作也不能忽视几乎不发布推文的自动化账号，如僵尸粉丝机器人等。同时，有些自动化账号发布的推文对人们的日常生活有益，如天气预报机器人等，未来的研究工作或许需要考虑更详细的分类标准，将这些有益的自动化账号剔除出自动化账号的检测范围。

- 更充分地挖掘历史推文信息。历史推文信息中还有许多有价值的内容可供挖掘，比如推文中的多模态信息。

参考文献

Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, and Christoph Meinel. 2019. Detect me if you can: Spam bot detection using inductive representation learning. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 148–153.

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

Yeyang Chen, Mondher Bouazizi, and Tomoaki Ohtsuki. 2022. Social robot detection using roberta classifier and random forest regressor with similarity analysis. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 6433–6438. IEEE.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15(4):561–576.

Stefano Cresci. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Abdelouahid Derhab, Rahaf Alawwad, Khawlah Dehwah, Noshina Tariq, Farrukh Aslam Khan, and Jalal Al-Muhtadi. 2021. Tweet-based bot detection using big data analytics. *IEEE Access*, 9:65988–66005.

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

Phillip George Efthimion, Scott Payne, and Nicholas Proferes. 2018. Supervised machine learning bot detection techniques to identify social twitter bots. *SMU Data Science Review*, 1(2):5.

Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. 2021a. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4485–4494.

Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. 2021b. Botrgcn: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 236–239.

Shangbin Feng, Zhaoxuan Tan, Rui Li, and Minnan Luo. 2022. Heterogeneity-aware twitter bot detection with relational graph transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3977–3985.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Kadhim Hayawi, Sujith Mathew, Neethu Venugopal, Mohammad M Masud, and Pin-Han Ho. 2022. Deeprobot: a hybrid deep neural network model for social bot detection based on user profile data. *Social Network Analysis and Mining*, 12(1):43.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Zhenyu Lei, Herun Wan, Wenqian Zhang, Shangbin Feng, Zilong Chen, Qinghua Zheng, and Minnan Luo. 2022. Bic: Twitter bot detection with text-graph interaction and semantic consistency. *arXiv preprint arXiv:2208.08320*.

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

Jonas Lundberg, Jonas Nordqvist, and Mikko Laitinen. 2019. Towards a language independent twitter bot detector. In *DHN*, pages 308–319.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Feng Wei and Uyen Trang Nguyen. 2019. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In *2019 First IEEE International conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*, pages 101–109. IEEE.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*.

张玄and 李保滨. 2022. 微博环境中的机器人账户检测综述. 中文信息学报, 36(12):1–15.

徐帅帅, 戴新宇, 黄书剑, and 陈家骏. 2017. 基于无指导学习的微博评论分析方法. 中文信息学报, 31(2):179–186.

差比句结构及其缺省现象的识别补全研究

周鹏飞[1]，**曲维光**[1,2,4,*]，**魏庭新**[3]，**周俊生**[1]，**李斌**[2]，**顾彦慧**[1]
(1.南京师范大学计算机与电子信息学院/人工智能学院，江苏省南京市210023;
2.南京师范大学文学院，江苏南京210097;
3.南京师范大学国际文化教育学院，江苏南京210097;
4.南京师范大学中北学院，江苏丹阳212334;
*通讯作者，Email：wgqu_nj@163.com)

摘要

差比句是用来表达两个或多个事物之间的相似或不同之处的句子结构，常用句式为“X比Y+比较结果”。差比句存在多种结构变体且大量存在省略现象，造成汉语语法研究和自然语言处理任务困难，因此实现差比句结构的识别和对其缺省结构进行补全非常有意义。本文采用序列化标注方法构建了一个差比句语料库，提出了一个能够融合字与词信息的LatticeBERT-BILSTM-CRF模型来对差比句结构自动识别，并且能对缺省单位进行自动补全，实验结果验证了方法的有效性。

关键词： 差比句结构；神经网络；缺省补全；信息融合

A Study on Identification and Completion of Comparative Sentence Structures with Ellipsis Phenomenon

ZHOU Pengfei[1]，**QU Weiguang**[1,2,4,*]，**WEI Tingxin**[3]，**ZHOU Junsheng**[1]，**LI Bin**[2]，**GU Yanhui**[1]
(1.School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, Nanjing , Jiangsu 210023, China;
2.School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
3.International College for Chinese Studies, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
4.Zhongbei College, Nanjing Nornal University, Danyang , Jiangsu 212334, China;
*Corresponding, Email: wgqu_nj@163.com)

Abstract

Comparative sentences are sentence structures used to express the similarities or differences between two or more things. The common pattern is ”X compare to Y+comparison result.” There are various structural variants and they often involve ellipsis, which poses challenges in Chinese grammar research and natural language processing tasks. Therefore, it is of great value to recognize the comparative structures and complete their ellipse. In this paper, first we constructed a comparative sentence corpus using a sequential tagging method. Then we proposed a LatticeBERT-BILSTM-CRF model that can integrate the information of Words and Phrases to identify the comparative, and complete the ellipsis units automatically. Experimental results confirm the effectiveness of our approach.

Keywords: comparative sentences structure , neural network , ellipse completion , information fusion

1 引言

差比句是汉语语法中的一种特殊句式，因其简洁明了的表达方式和丰富多样的语义功能，被广泛应用于各个领域的语言表达。差比句在比较句的研究中占据重要的地位。现在通常认为一个完整的差比句需要四个要素：比较主体（记作X）、比较基准（记作Y）、比较标记和比较结果（记作R）。并且按照现代汉语语序应该表现为“X+比较标记+Y+W”。吴颖菲(2020)提到差比句有以下几种类型:

例1：摩托车的车速比自行车的车速更快。

例2：小明的身高跟小李的身高比更高。

例3：当今的公务员考试相比几年前的已经有了很大进步。

例4：你吃的饭比我多。

其中例1-2就是典型的完整差比句，如例1其中“摩托车的车速”为比较主体（X）,“比”为比较标记，“自行车的车速”为比较基准（Y），“更快”为比较结果（R）。例2中“跟.....比”作为比较标记。例3-4虽然句子结构很完整，但是其结构内部存在省略，如例3中的“几年前”作为比较基准（Y）省略了“公务员考试”这一部分内容，这类句子称为省略差比句。能够正确识别出这类差比句中的省略部分并将其进行补全，对于自然语言处理任务如机器翻译、信息抽取等工作都有重要的意义。然而，差比句的多种结构给机器识别差比句结构及自动缺省补全带来很大的挑战。

目前，对于差比句的研究主要集中在语言学领域。而在自然语言处理领域，对于差比句结构识别的研究相对较少。这是因为差比句的结构相当特殊，传统的句法识别方法并不适用。此外，目前还缺乏专门针对差比句的语料库，而且差比句中存在大量省略现象，通用的解析器方法往往只能进行省略判别而无法进行补全，导致效果不尽如人意。虽然其他一些特殊结构缺省补全工作已经取得了不错的效果(侍冰清等, 2019; 施寒瑜等, 2020)，但是它们的规则模型不符合差比句中的省略特点，因此并不适用于差比句。

本文的贡献如下:

1. 设计并制定了一套语料库标注规范，构建了一个包含多种类型差比句的语料库。该语料库包括5800条句子，其中涵盖了具有完整成分的差比句以及比较主体和比较基准先行语省略的差比句。

2.提出一个LatticeBERT-BILSTM-CRF模型来实现判别差比句类型并对差比句成分进行识别，该模型在BERT模型基础上加入了词的信息，引入新的位置编码和预训练任务，将字和词的信息融合输入BERT。模型在对差比句类型判别和差比句结构识别任务上取得了很好的效果，差比句结构识别任务F值达到了91.4%。

3.根据生成的句子标签，设计一套规则找到差比句中存在缺省的部分、识别出可补全的部分，并进行插入补全。

2 相关工作

2.1 差比句研究现状

差比句的语义功能丰富多样，包括了比较、强调、排比、修辞等多种功能。在研究差比句的语义方面，研究者通常会探讨差比句的语义特征、语义类别、语义关系等问题。例如，刘丹青(2004)从语言类型学的角度入手，结合Greenberg(1963)的研究结果，得出结论：汉语差比句的基本要素包括性质属性的主体、表示性质属性的形容词、基准和比较标记，分析在进行类型学调查时需要关注的差比句的要素。

差比句的语法结构由差异部分和比较部分组成，其表达形式多样。在研究差比句的语法结构方面，研究者通常关注其句子成分、成分的排列方式和缺省现象等问题。例如，任海波(1987)、邵敬敏，刘焱(2002)等人分别从句子成分的角度出发，探讨了差比句中的主语、谓语、宾语、定语和状语等成分在语法结构中的位置和作用，提出了差比句中成分排列的一些基本规律。在研究差比句的句法特征方面，研究者主要关注其语序、语法功能和句子结构等方面。例如吴颖菲(2020)通过对比分析汉语差比句的句子结构，发现汉语差比句中还存在一些仅仅含有差比语义但结构不固定的句子类型，引出了对多类非典型差比句的研究。

基金项目：国家社会科学基金重大项目(21&ZD288); 国家自然科学基金(62277031)。

这些研究者虽然对差比句句法语义做了大量研究，但均是从词典、公共语言数据集上做出的研究。未能构建专门差比句语料库进行研究，并且也未能与自然语言处理任务结合。

2.2 省略结构研究现状

省略结构是指在句子中省略某些成分而不影响句子的完整性和语义表达的现象，是语言学中一个重要的研究领域。在语法学层面，研究者通常会探讨省略结构的形式和语法特征。例如，赵元任(2002)等人研究了汉语省略结构的形式，提出了省略现象的三种类型，即语义省略、语用省略和结构省略，并探讨了各种省略现象的规律和特征，分析了不同省略形式的句法特征和语法功能。Marie Mikulová(2014)对英语和其他语言中省略结构的形式和语法特征进行分析，并提出了"链式理论"(Chain Theory)来解释省略结构的产生。

刘依欢(2020)在探讨汉语省略现象的抽象语义表示时，重点分析了省略句中省略内容和上下文之间的语义关系。她发现，大多数句子恢复省略成分依赖于先行语，先行语是指在句中的上下文中出现的与省略成分相同的词语。省略成分存在一个对应的先行语，省略现象与指代类似，其本质是回指。张伟男等(2009)采用决策树分类算法对中文对话中的省略现象进行判别，而杨国庆等(2011)则基于句子的基本结构，提出了省略三元规则进行省略识别。侍冰清(2019)研究了语义省略中的"的"字结构，发现其省略现象比较普遍，并对"的"字结构中的各个成分进行识别和补全，利用神经网络在构建的语料库上取得了87.1%的整体F值。施寒瑜(2020)同样采用神经网络的方法研究汉语数量名短语中的缺省成分，并通过联合学习方法，在构建的语料库上取得了85.07%的F值。

综上，差比句缺省形式多样、句法结构特殊，以往研究未能专门对差比句中的缺省现象进行系统性研究，并且目前尚未有对差比句内部缺省现象的补全工作。

3 差比句语料库构建

3.1 差比句语料库来源

本文选取1998年1至3月《人民日报》新闻语料以及北京语言大学语料库中心(BLCU Corpus Center)作为差比句标注的基础语料，此外还有一部分差比句来自CTB中文宾州树库，共计10300条，其中人工标注了5800条包含完整和省略的差比句供实验使用。

3.2 差比句语料库的构建

本文针对差比句的标注问题，使用了BIO 标注体系。该体系的优点在于能够很好地处理实体的边界问题。标注的基本组成结构为"比较主体(X) + 比较标记(C) + 比较基准(Y) + 比较结果(R)"。然而，如果仅将X、Y、C、R 作为标签进行标注，难以处理一些基本结构出现省略的情况。例如，"哥哥身高比弟弟更高"，这句话中的"弟弟"作为比较基准存在省略。若仅有四个标签，只能将其标注为Y，无法让机器学习到缺省的位置。另外，可以看到这类句子可以通过文中的先行语进行补全，而先行语作为补全的成分也需要一个标签。因此，本文对基本结构进行更加细粒度的拆分，引入标签X1、Y1 和K，分别标记缺省的比较主体、缺省的比较基准、先行语（句子中出现的与省略成分相同的词语）。这样做的好处是可以很好地处理省略的情况，提高标注的准确性和效率。因此最后的标签为X(比较主体) 、Y(比较基准)、C(比较标记)、R(比较结果)、X1(缺省比较主体)、Y1(缺省比较基准)、K(先行语)。

3.2.1 完整差比句结构的标注

完整的差比句结构需要同时满足下列情况：句子中出现完整的差比句结构且比较主体和比较基准两个比较对象属于同一个范畴。下面以具体例子分析完整差比句标注方法:

例5：今年工厂的实际投资额$_X$比$_C$去年工厂的实际投资额$_Y$高$_R$。

在例5中"工厂的实际投资额"和"店铺的实际投资额"比较的是同一个范畴，比较的主题都是"实际投资额"。因此"工厂的实际投资额"作为完整的比较主体标注为X，"店铺的实际投资额"作为完整的比较基准标注为Y，另外这句话还有完整的比较标记和比较结果，因此例5为完整的差比句，完整的差比句不应该存在X1,Y1标签。因为该句子不存在标签X1,Y1，所以不需要进行补全，具体标注示例可见表1。

工	厂	的	实	际	投	资	额	比	店
B-X	I-X	I-X	B-X	I-X	I-X	I-X	I-X	B-C	B-Y
铺	的	实	际	投	资	额	高	。	
I-Y	I-Y	B-Y	I-Y	I-Y	I-Y	I-Y	B-R	O	

Table 1: 完整差比句标注示例

3.3 缺省差比句结构的标注

在差比句中，省略现象多数是由于比较主体与比较基准的比较范畴不对应所导致的。例如，在例句6中，“厦门气温”是完整的比较主体，因此将“厦门”标记为比较主体中的修饰成分X，而将“气温”标注为先行语K。而“南京”作为比较基准存在省略，因此将其标记为“Y1”，以此表示省略。另外，“比”标注为比较标记C，“更高”则标注为比较结果R。在例句7中，省略成分的先行语是比较基准中的一部分，即“我的成绩”的中心语为“成绩”。此时，“你”并非完整的比较主体，因此将其标记为X1；“我的成绩”作为比较基准是完整的，因此将“我的”标记为修饰，标记为Y，而将“成绩”标记为比较主题，标记为可供比较主体缺省补全的先行语K。具体的标注示例可以参见表2。

例6：厦门$_X$气温$_K$比$_C$南京$_{Y1}$更高$_R$。

例7：你$_{X1}$比$_C$我$_Y$的成绩$_K$优秀$_R$。

厦	门	气	候	比	南	京	温	和	。
B-X	I-X	B-K	I-K	B-C	B-Y1	I-Y1	B-R	I-R	O

你	比	我	的	成	绩	优	秀	。
B-X1	B-C	B-Y	I-Y	B-K	I-K	B-R	I-R	O

Table 2: 缺省差比句标注示例

3.4 差比句语料库统计

本文依据3.1，3.2 节中方法进行标注，数据的选择考虑均衡每种差比句类型比例，让模型能够充分学习到不同类差比句的特征，数据集划分类型如表3所示。为确保标注数据质量，先制定详细标注规范，提供相应培训确保标注人员掌握规范。试标阶段对两名标注人员进行一致性统计，以评估其标注准确性和一致性。统计结果显示标注人员一致性存在差异时，进行调整，以保证正式标注的准确性和一致性。此外，对标注数据反复校验和核对，确保标注数据质量和准确性。最终标注数据达到较高一致性和准确性水平，为后续实验提供可靠数据基础。

句子类型	标注数量	样本样例
完整差比句	3000	今年产量比去年产量高
缺省差比句	2300	现在的生活比以前更好
非差比句	2000	他是一位正直的人

Table 3: 语料库成分

此外，还对不同种类的比较标记类型做了统计如表4所示。

比较标记	比	跟	和	相对	于	跟...比	过
数量（句）	2006	64	13	63	22	116	16
占比（%）	87.22	2.78	0.57	2.74	0.957	5.04	0.66

Table 4: 不同类型比较标记分布

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

4 差比句结构识别与缺省补全模型设计

本文将差比句结构识别视为缺省补全的一个子任务，通过识别出的标签信息进行缺省补全。基于序列化标注任务建模，解决差比句结构识别。由于差比句结构类型特殊，每个字都对应一个结构标签，容易出现边界识别错误。由此使用Lattice结构将词信息加入到预训练模型BERT来解决此问题。模型如图1所示，由四个层级构成：1. Lattice 结构转换层，该层用于将文本数据转换为lattice形式，从而支持Lattice-based模型的输入；2. Lattice-BERT编码层，将经过Lattice转换层处理后的Lattice序列作为输入，进行编码，得到一系列的向量表示；3.BI-LSTM层，用于从Lattice-BERT的输出中提取特征，以便更准确地预测每个位置的标签；4.CRF层, 用于将经过BI-LSTM层提取的特征序列转换成标注序列。它通过考虑相邻标记之间的关系，并对相邻标记之间的转移施加约束，来解决标记之间的依赖关系和标记边界问题，从而确保结构识别的完整性。

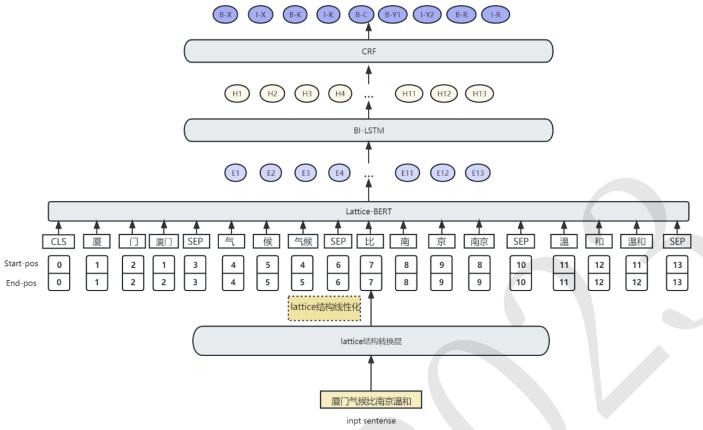


Figure 1: 模型结构组成图

4.1 Lattice结构转换

在预处理阶段，本文将中文文本转换为Lattice结构，转化过程通过以下步骤实现：

1.分词：首先使用字符词汇创建一个BERTTokenizer，再使用BERTTokenizer解析词典，获取每个单词的字符序列，然后测试字符范围是否与序列匹配。如果匹配，将其视为一个词。词汇的大小可以计算为字符词汇的大小加上单词词汇的大小。

2.构建词图：将分词得到的词语构建成一个有向无环图（DAG）。DAG的节点代表词语，边表示两个词语之间的关系。比如“厦门气候比南京温和”这句话中包含五个词语“厦门”，“比”，“南京”，“气候”“温和”则可以构建一个DAG，其中节点包括“厦”，“门”，“厦门”，“比”，“南”，“京”，“南京”，“气”，“候”，“气候”，“温”，“和”，“温和”，同时建立从“厦”指向“门”，厦门指向“夏”和“门”两个节点，同时又指向“气”节点，如此一直到最后，如图2 所示。

3.填充标签：将第三章设计的七个不同标签，采用BIO的标注方法对差比句各个结构进行标注。

4.线性化：将转换的lattice结构图线性化作为输入，如图2所示直接将词格图中各粒度的信息“拍平”，得到图1中输入的线性序列。

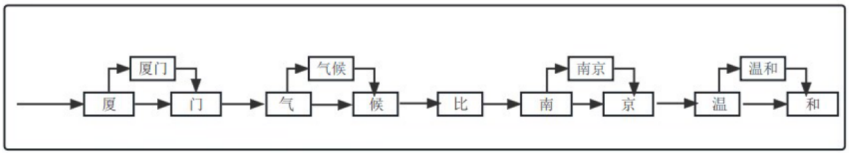


Figure 2: Lattice结构生成图

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

4.2 文本表示层

中文以词为语义基本单位，因此传统中文NLP都需要先进行分词处理，但是分词的难度很大，经常需要处理一些歧义和多义情况。在预训练模型中，中文模型通常采用字作为输入，这种输入会忽略中文的粗粒度信息，词语对于中文的语义理解来说至关重要。Lattice结构可以让BERT在预训练中学习到字和词的信息。

差比句结构识别任务对于普通序列化标注任务有如下特点：1、差比句结构标签分布很密集，几乎每个字都有一个对应的标签。这意味着对于每个字，其上下词信息对于正确预测它的标签非常重要；2、在差比句中，有时候会出现两个或多个实体的首尾相接或者互相嵌套的情况。因此词信息在差比句结构识别任务中尤为重要，lattice结构可以将输入融合词的信息(Matthias Sperber et al., 2019)，因此本文利用Lattice结构图输入到预训练模型BERT来得到一个同时包含字和词信息的文本表示。本文使用的方法是将Lattice结构图中各粒度的信息“拍平”，得到一个线性序列。然而，“拍平”词格的输入会导致重复和冗余的问题，进而影响位置编码的适应性。除此之外在“拍平”之后，原先二维的复杂图结构信息也会有所损失。为了解决这两个问题，模型设计了新的注意力机制。首先改进了BERT的绝对位置编码，式1里的P_i代表相对位置长度，由P^S 和P^E做差得到，P^S表示当前输入token的开始位置，P^E 表示结束的位置。式2将token的起始位置的绝对位置编码拼接，进行attention 操作，从而得到相对位置编码。模型除了保留原始BERT中的位置编码，还加入了词格输入。由于词格输入的每一项长度是不固定的，因此引入头尾位置，对应图1中的start-pos和end-pos。

然而，光是绝对位置编码所提供的信息还不够充足，因为在理论上对绝对位置编码的限制只有一点，即不同位置的编码不同。这样会忽略了很多信息，比如:位置1和2的距离与位置5和6的距离应该一样，位置1和3的距离比位置4和10的距离要小等等。在绝对位置编码的设计上只能让BERT隐式地“学习”。因此后续还加入相对位置编码以及针对层叠问题加入层叠编码信息，来表示token 之间的相对距离。式3中的$\tilde{\alpha}_{ij}^{l}$第一项是字的表示得到的attention score，第二项是绝对位置编码，相对位置编码，层叠编码相加，相对位置编码为P^E，P^S之差，层叠编码根据这两个token 起始相对位置的不同，两个token 可以分成下列七种关系(Yuxuan Lai et al., 2021)T.1自身；T.2在左边，且无重叠；T.3在左边，且有重叠；T.4包含关系；T.5被包含关系；T.6在右边，且有重叠；T.7在右边，且无重叠。

$$P_i = \left| P^E - P^S \right| \tag{1}$$

$$\tilde{\epsilon} = \epsilon_i^{in,0} + P_i \tag{2}$$

$$\tilde{\alpha}_{ij}^{l} = \alpha_{ij}^{l} + f(i,j) \tag{3}$$

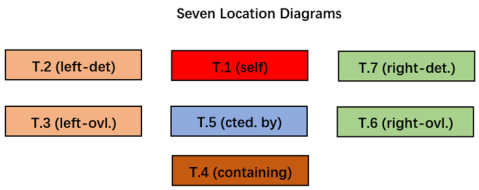


Figure 3: 七种位置关系

此外，Lattice-BERT引入了一些新的预训练任务：Masked Segment Prediction来取代原先的Masked Language Modeling（MLM）任务。MSP任务的训练数据是单个句子，因此每个训练样本都是单个句子的segment，其目标是预测输入句子中哪些部分是连续的。在MSP任务中，将输入句子切分为多个segment，并在其中随机mask掉一些segment，然后让模型预测哪些segment是连续的。这个任务的目的是让预训练模型学习到segment之间的连续性关系，进一步提升对于长文本的建模能力，以进一步提高中文自然语言处理的性能。

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

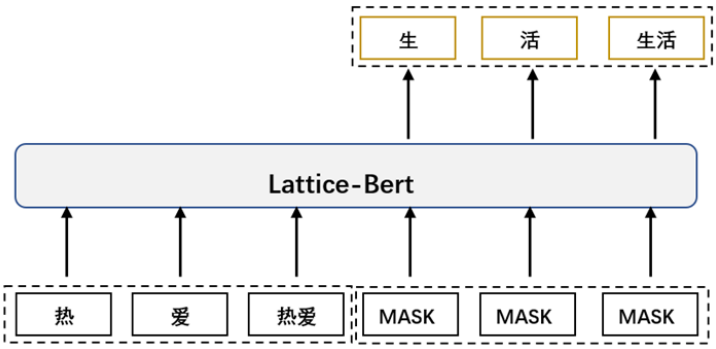


Figure 4: MSP任务mask示例

4.3 特征提取层

BI-LSTM（Bidirectional Long Short-Term Memory）是一种循环神经网络（RNN）的变体，它在自然语言处理领域广泛应用于序列标注任务，如命名实体识别（NER）和情感分析等。相比传统的单向LSTM，BI-LSTM不仅考虑了当前时刻之前的信息，还考虑了当前时刻之后的信息，从而更好地捕捉了序列中的上下文信息。BI-LSTM通过在正向和反向两个方向上分别处理输入序列，得到两个独立的隐藏状态向量，然后将这两个向量按位置相加，作为输出。这样可以使每个位置的输出同时考虑当前位置之前和之后的上下文信息。

相比于CNN，BI-LSTM具有更强的序列建模能力。Lattice-BERT的输入是嵌套的lattice结构，如果使用CNN对嵌套的结构进行建模，需要将lattice结构“拍平”成一维的序列，这样就可能损失掉一些与嵌套结构相关的信息。而BI-LSTM能够对序列中的上下文信息进行建模，更加适合处理嵌套结构。

相比于LSTM，BI-LSTM在提取特征时能够同时考虑上下文的信息。LSTM是一种单向模型，只能考虑当前时刻之前的信息，而BI-LSTM是一种双向模型，能够同时考虑当前时刻之前和之后的信息，因此能够更好地捕捉上下文信息。

因此，将BI-LSTM与Lattice-BERT结合能够充分利用Lattice-BERT的字和词级别信息，同时对序列中的上下文信息进行建模，提高序列化标注任务的准确性。

4.4 输出层

模型采用CRF作为序列标注的输出层，它可以通过整合上下文信息和标签约束来提高模型的性能。在Lattice-BERT模型中，CRF层用于将经过BI-LSTM层提取的特征序列转换成标注序列。它通过考虑相邻标记之间的关系，并对相邻标记之间的转移施加约束，来解决标记之间的依赖关系和标记边界问题，从而提高模型在序列标注任务中的性能。

4.5 缺省补全

缺省补全解析器基于句子中的先行语和缺省结构进行补全，即通过插入省略结构来实现缺省部分的补全。具体地，本文利用核心思想，即先找到先行语，然后在省略结构中插入先行语K，如果遇到K标签，X1或Y1则说明存在缺省内容，X1或Y1标签后面的位置就是先行语K需要插入的位置。为了实现这一目标，本文设计了一套补全规则流程图，如图5所示。该流程图具有一定的纠错功能，如果K标签和X1或Y1标签没有同时出现，则输出error并且需要进一步检查输出结果的准确性。

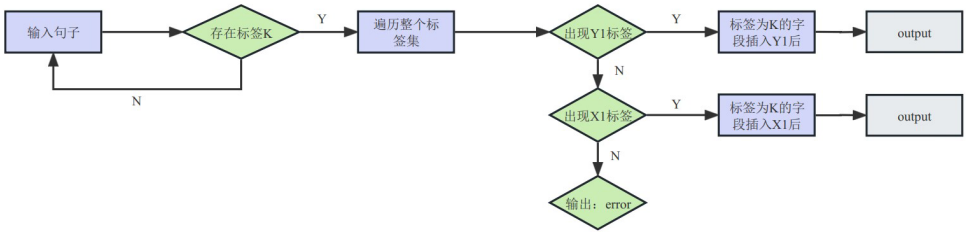


Figure 5: 缺省补全流程图

例如，当图6的第一个输入作为缺省补全解析器的输入时，解析器会首先分析句子的标签。在分析过程中，发现“气候”的标签为K，意味着“气候”是比较基准。随后，解析器会遍历整个标签集，发现“南京”的标签为Y1，说明比较基准部分存在缺省。为了补全缺省部分，先行语需要复制并补充在比较基准“南京”之后。最终，句子变成了“厦门气候比南京气候温和”，并被加入典型差比句语料库中。解析器具有标签错误检测功能，对于第二个输入，解析器检测到“成绩”为K标签，因此开始遍历标签集。然而，在遍历过程中，解析器发现句子中不存在X1和Y1标签，这意味着标签生成出现错误，句子无法被正确补全。因此，这个输入被视为含有错误生成标签，并输出到错误标签样本类进行人工处理。

通过这样的解析过程，缺省补全解析器可以对含有缺省部分的句子进行自动补全。对于能够正确补全的句子，可以将其加入典型差比句语料库中，以便后续的分析和应用。而对于无法正确补全的句子，则需要进行人工处理，以确保整个模型的准确性和可靠性。

Figure 6: 缺省补全解析器输入样例

5 实验

5.1 数据集划分和模型参数设置

实验将数据集由完整差比句、省略差比句、非差比句三类句型组成，按照8：1：1的比例划分为训练集、验证集和测试集，再将句子随机打乱。数据集如表5所示。其中训练集用于模型学习，拟合分类器的参数（普通参数、神经元的权重等）；测试集用于调整模型的参数，比如：确定隐藏单元数、确定神经网络结构和复杂程度的参数（超参数：如隐藏层数、每一层的神经元数等）；验证集用于测试模型的表现。

类别	训练集	测试集	验证集
完整差比句	2400	300	300
省略差比句	1840	230	230
非差比句	1000	500	500

Table 5: 数据集划分

本文采用Lattice-BERT预训练模型chinese_laBERT-tiny-std-512，超参数具体设置如表6。

超参数	参数值
Epochs(迭代次数)	12
batch size（单批次处理数量）	30
Dropout（丢弃率）	0.01
Learning Rate（学习率）	1e-5
Segment Size（每段最大token数量）	64
Embedding Size（输入最大token数量）	128

Table 6: 实验参数设置

5.2 评价方法

评价方法使用准确率P、召回率R和F1值对模型的性能进行评价，计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{4}$$

第二十二届中国计算语言学大会论文集，第388页-第398页，哈尔滨，中国，2023年8月3日至5日。

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{6}$$

5.3 实验结果与分析

本文将编码层和特征提取层分别引入不同的模型与我们构建的模型作为比对。其中包括：BERT+BILSTM+CRF、LatticeBERT+CNN+CRF、LBERT+LSTM+CRF。整体实验结果如表7所示。实验结果显示，本文提出的模型在P、R和F1三个指标上都优于基线模型。可以看出，仅使用BILSTM+CRF时，F值仅为77.28，当加入BERT编码层时，性能得到提升可以达到88.19，而将编码层换成Lattcie-BERT可以进一步提升性。在编码层选择用Lattice-Bert的基础下，特征提取层选用BILSTM可以使得性能达到最优值91.57。这说明本文的模型能够更准确地识别标签，并且相较于基线模型具有更好的综合性能。

从实验结果可以看出，相较于常用的BERT编码器Lattice-BERT实验结果上有明显的优势。因为一般我们常用差比句中几乎每个词都会打上一个标签，因此容易错误地将距离近的标签作为错误生成，尤其体现在先行语和与其构成完成比较主体（基准）的其他内容上。BERT模型没有词的信息很容易扩大或缩小生成先行语的标签，从而造成缺省补全的困难，引入词的信息明显可以帮助模型理解差比句的各个结构，使词与词之间位置信息更加明确。

本文的模型采用了多层Lattice-BERT、BILSTM和CRF等技术，能够有效地解决结构识别任务中的嵌套标签问题，同时还具有更好的语义表示能力。这些技术的组合使得本文的模型在处理结构识别任务时能够更加全面地考虑句子中字和词的信息，从而提高了识别的准确率和召回率。

实验模型	P(%)	R(%)	F1(%)
BILSTM+CRF	78.34	76.25	77.28
BERT+BILSTM+CRF	87.63	88.76	88.19
LatticeBERT+CNN+CRF	90.16	89.24	89.70
LatticeBERT+LSTM+CRF	91.31	90.34	90.82
本文模型	**92.39**	**90.76**	**91.57**

Table 7: 差比句结构识别结果

表8展示了用于缺省补全的先行语识别结果。由于先行词通常存在于比较基准和比较主体之间，因此会造成先行语和缺省比较主体（基准）标签之间的边界信息和位置信息识别困难。从而经常造成生成的标签之间相互越界。观察表8可以发现，相对于其他模型，各模型对于先行词的识别整体水平普遍下降。然而，在本文模型中，通过在编码过程中引入词的信息和多种位置编码的信息，使得本文模型受到的影响最小。这说明本文模型的识别更加稳定，同时也显著提高了后续缺省补全的准确率。

实验模型	P(%)	R(%)	F1(%)
BILSTM+CRF	73.25	71.51	72.37
LatticeBERT+CNN+CRF	86.63	87.16	86.89
LatticeBERT+LSTM+CRF	88.47	87.56	88.01
BERT+BILSTM+CRF	88.65	88.37	88.51
本文模型	**91.23**	**90.41**	**90.82**

Table 8: 省略差比句中先行词识别结果

5.4 消融实验

本文通过使用Softmax层代替CRF层；减去BILSTM层分别做了两次实验，如表9所示。

实验模型	P(%)	R(%)	F1(%)
Ours	**92.39**	**90.76**	**91.57**
Ours-BILSTM	88.70	88.24	88.47
Ours-CRF	91.85	90.10	90.97

Table 9: 差比句结构识别消融实验结果

实验发现，两个消融实验在P、R和F1三个指标上相较于本文模型均有所下降。原因是虽然Lattice-BERT编码已经可以通过引入词信息，很好地表示结构特征。但由于输入内容是一张“拍平”后的图，输入过于冗长因此会丢失一些长距离信息。而BILSTM层的加入能够更好地学习长距离依赖关系且BILSTM很适合处理嵌套关系信息，由此加入BILSTM作为特征提取层可以很好缓解由于输入过长而导致信息丢失的问题。此外加入CRF是由于其可以自动学习序列标注的规律且能够更好地处理复杂的标注模式，包括多个标签之间的依赖关系和标签转换模式，从而使系统更加准确和稳定。基于将这些模块结合，能够更好地处理差比句识别和补全任务，实验结果表明，我们所选取的另外两层模块每一层对于整体模型都有一定提升。

5.5 缺省补全实验

本实验采用了两种基线模型，分别是BILSTM+CRF和BERT+BILSTM+CRF，这两种模型生成的标签被送入缺省补全解析器中，用于对比本文设计的解析器的性能。实验结果见表10。由于LatticeBERT+BILSTM+CRF解决了先行语标签生成容易和存在缺省标签互相“越界“问题，从而能够精准识别出补全的关键标签—先行语和缺省比较主体（基准）。实验结果表明LatticeBERT+BILSTM+CRF模型与本文设计的解析器适应度最高，P、R、F1值均达到了最高水平。

Model	P（%）	R（%）	F1（%）
BILSTM+CRF+缺省补全解析器	68.77	75.28	71.88
BERT+BILSTM+CRF+缺省补全解析器	80.34	81.50	80.92
LatticeBERT+BILSTM+CRF+缺省补全解析器	**86.49**	**88.21**	**87.34**

Table 10: 缺省补全各模式实验结果

解析器生成的错误可以分为三类，如表11所示。第一种类型是模型仅生成了K标签，而没有生成X1或者Y1。当该标签被送到解析器时，解析器无法找到合适的位置插入先行语，从而导致错误。第二种类型的错误是模型生成的K标签覆盖了比较主体或比较基准的过多部分，在补全后形成的句子语义混乱。第三种类型的错误是模型生成的比较主体和比较基准标签存在中断，同样会在补全后形成语义混乱的句子。这三种类型的错误都与模型标签生成的准确性直接相关。对于这样的错误案例分析错误的原因如下:1.词典构建不完善，在分词时出现错误进而导致生成错误的Lattice结构。2.数据集规模仍相对较小，对于一些句子类别语料库内较少，因此模型学习时有一定困难。后续会继续在这两个问题上做出相应改进。

原句子生成补全标签	结果类型	输出句子
若长此以往，以后$_X$的情况$_K$会比现在$_{Y1}$更差。	正确输出	若长此以往，以后的情况会比现在的情况更差。
两者斟酌，前者$_X$严重程度$_K$固然比后者$_Y$轻。	错误1	Error
太平洋面积$_X$是其余三大洋的总和$_K$，比北冰洋$_{Y1}$大十四倍。	错误2	太平洋面积是其余三大洋的总和，比**北冰洋面积是其余三大洋的总和**大十四倍。
1993年$_X$新钢利润$_K$比上$_{Y1}$年下降88.9%	错误3	1993年新钢利润**比上新钢利润年**下降88.9%

Table 11: 补全转换示例

6 结语

本文通过神经网络对差比句结构及其缺省情况进行了研究，利用BIO方式标注的数据集进行模型训练，采用Lattice-BERT与BILSTM结合的复合模型成功识别差比句结构并补全缺省部分。未来，本文将继续研究不规则的差比句并进一步优化模型以提高其性能。具体来说，我们将尝试通过篇章级别的学习，让模型脱离先行语的限制对差比句的所有省略结构进行识别。

参考文献

侍冰清, 戴茹冰, 曲维光, 顾彦慧, 周俊生, 李斌, 徐戈, 史胜旺. 2019. 基于组合神经网络的语义省略“的”字结构识别. 北京大学学报(自然科学版), 2019, 55(01): 75-83.

施寒瑜, 曲维光, 魏庭新, 周俊生, 顾彦慧. 2022. 基于组合深度模型的现代汉语数量名短语识别. 南京师大学报(自然科学版), 2022, 45(01): 127-135.

刘丹青. 2004. 差比句的调查框架与研究思路. 现代语言学理论与中国少数民族语言研究. 北京: 民族出版社, 2004: 1-22.

邵敬敏, 刘焱. 2002. 比字句强制性语义要求的句法表现. 汉语学习, 2002（5）: 3-7.

任海波. 1987. 现代汉语”比”字句结论项的类型. 语言教学与研究, 1987(4):91-103.

吴颖菲. 2020. 汉语非典型差比句的研究与教学. 华东师范大学.

赵元任. 2002. 赵元任语言学论文集. 北京：商务印书馆，2002：61-72.

刘依欢. 2020. 基于抽象语义表示的省略现象研究. 南京师范大学.

张伟男, 张宇, 刘挺. 2009. 基于决策树的中文对话省略句判别.中国中文信息学会会议论文集, 2009:315-322.

杨国庆, 孔芳. 2011. 基于规则的中文缺省识别研究. 计算机科学, 2011(12): 255-258.

邓依依, 邬昌兴, 魏永丰, 等. 2021. 基于深度学习的命名实体识别综述. 中文信息学报, 2021, 35(9): 30-45.

郑远汉. 1998. 省略句的性质及其规范问题. 语言文字应用, 1998(02): 12-19+3.

李艳惠. 2005. 省略与成分缺失. 语言科学, 2005(02): 3-19.

Yuxuan Lai, Yijia Liu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2021.

Lattice-BERT: Leveraging Multi-Granularity Representations in Chinese Pre-trained Language Models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1716–1731.

Greenberg J H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. Universals of language, 1963: 73-113.

Marie Mikulová. 2014. Semantic Representation of Ellipsis in the Prague Dependency Treebanks. In Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014), 125–138.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-Attentional Models for Lattice Inputs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1185–1197.

Devlin J, Chang M W, Lee K, and Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, 2019: 4171-4186.

Jason P. C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 357-370.

Yue Zhang and Jie Yang.2018. Chinese NER Using Lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics,1554-1564.

Song K, Tan X, Qin T, et al. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. International Conference on Machine Learning, 5926-5936.

基于框架语义场景图的零形式填充方法

王俞智[1,‡]，李茹[1,2,⋆,†]，苏雪峰[1,3,‡]，闫智超[1,‡]，李俊材[1,‡]
[1]山西大学 计算机与信息技术学院，山西 太原 030006
[2]山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006
[3]山西工程科技职业大学现代物流学院，山西 晋中 030609
[‡]{1007079197, 455375251, 751824801, 1251972979}@qq.com
[⋆]{liru}@sxu.edu.cn

摘要

零形式填充是在篇章上下文中为给定句子中的隐式框架语义角色找到相应的填充内容。传统的零形式填充方法采用pipeline模型，容易造成错误传播，并且忽略了显式语义角色及其填充内容的重要性。针对上述问题，本文提出了一种端到端的零形式填充方法，该方法结合汉语框架网信息构建出框架语义场景图并利用GAT对其建模，得到融合了显式框架元素信息的候选填充项表示，增强了模型对句中隐式语义成分的识别能力。在汉语零形式填充数据集上的实验表明，本文提出的模型相较于基于Bert的基线模型F1值提升了9.16%，证明了本文提出方法的有效性。

关键词： 零形式填充；框架语义场景图；GAT

A Null Instantiation Filling Method based Frame Semantic Scenario Graph

Yuzhi Wang[1,‡], Ru Li[1,2,⋆,†], Xuefeng Su[1,3,‡], Zhichao Yan[1,‡], Juncai Li[1,‡]
[1]School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006,China
[3]School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology, Jinzhong, Shanxi 030609,China
[‡]{2280493770, 455375251, 751824801, 1070913573, 1055342647}@qq.com
[⋆]{liru}@sxu.edu.cn

Abstract

Null Instantiation Filling aims to find the corresponding filling content for the implicit semantic roles of a given sentence in the context of the discourse. Traditional Null Instantiation Filling methods use a pipeline model that could cause error propagation and ignore the importance of explicit semantic roles.To address these issues, this paper proposes an end-to-end Null Instantiation Filling method that combines the Chinese FrameNet information to construct a Frame Semantic Scenario Graph and uses GAT to model it. The proposed method obtains candidate filling item representations that integrate explicit frame element information, enhancing the model's ability to recognize implicit semantic components. Experimental results demonstrate that the proposed model compared to the baseline model , the F1 value has increased by 9.16%.

Keywords: Null Instantiation Filling , Frame Semantic Scenario Graph , GAT

1 引言

* 基金项目：基于语言认知机理的汉语框架语义计算研究（61936012）；中新语言智能国际联合实验室（202204041101016）；山西省1331工程项目
† 通讯作者 Corresponding Author

汉语框架网（Chinese FrameNet，CFN） (You and Liu, 2005)是以Fillmore的框架语义学 (Fillmore et al., 1976)为理论基础，以汉语真实语料为依据，参照伯克利大学的框架语义网（FrameNet，FN） (Baker et al., 1998)构建的汉语框架语义知识库，包括框架库、句子库和词元库，其相关术语如表1 (Li, 2012)所示。汉语框架语义分析是基于汉语框架语义知识库提出的任务 (Shi et al., 2014)，零形式填充（Null Instantiation Filling，以下简称NIF）是汉语框架语义分析的子任务之一，它旨在将句子中目标词所触发的特定框架下的隐式语义角色与其在篇章上下文中的填充内容（若存在）联系在一起。NIF有助于机器对篇章的正确理解，对多跳阅读理解、篇章级语义分析等任务具有重要意义。

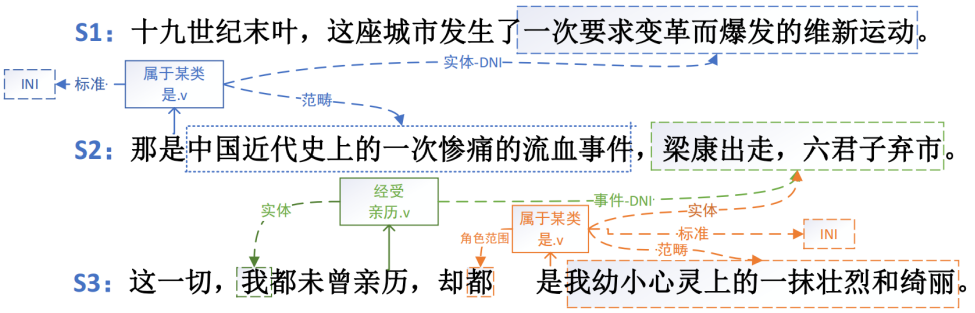


图 1: 语料示例

术语	定义	“属于某类”框架示例
框架	与一些激活性语境相一致的结构化范畴系统，是存储在人类认知经验中的图示化情境	属于某类：该框架表示某个实体属于某个范畴
核心框架元素	核心框架元素是一个框架在概念理解上必有的语义角色，显示出框架的个性	Entity（实体）：某个具体范畴的实体实例 Category（范畴）：某个实体所属的一种概括性的类型或者类别 Criteria（标准）：决定实体应归于哪个范畴的特征
非核心框架元素	非核心框架元素并不显示框架的个性，表达时间、空间、原因、目的等外围语义成分	Scop_of_role（角色范围）：该框架元素指实体的数量范围。
词元	能够触发特定框架的词或者短语	如“是”、“归于”、“作为”、“算作”是“属于某类”框架的词元

表 1: 汉语框架语义知识库术语及示例

传统的框架语义角色标注是基于句子级的，只能为句子中显式表达的语义论元分配语义角色，而忽略了一些隐含在篇章上下文中未显式表达的语义论元，这些未显式表达的语义论元对应的核心框架元素被称为零形式框架元素，简称为零形式（Null Instantiation，NI）。按照缺失的核心框架元素在语义理解上的不同解释类型，零形式被分为无定的零形式（Indefinite Null Instantiation, INI）和有定的零形式（Definite Null Instantiation, DNI）。INI指缺失的核心框架元素不影响人们对语篇的正确理解，且在上下文中没有特定的填充内容。DNI是指句子范围内缺失的核心框架元素在上下文中能够明确找到对应的填充内容。

NIF是已知零形式框架元素所属框架以及该框架在句子中的显式框架元素及其填充内容，为有定的零形式框架元素在上下文中找到相应的填充内容。如图 1 所示，S2中目标词“是”触发“属于某类”框架，框架信息如表1所示，核心框架元素“范畴”在S2中有填充内容，该框架元素称为显式框架元素，而核心框架元素“实体”和“标准”在句子范围内没有显式表达，但是可以推

断出“实体”是由前一句中的“一次要求变革而爆发的维新运动”所填充，即标记为DNI，而核心框架元素“标准”的缺失不影响对语篇的正确理解，因此标记为INI。

零形式填充任务的形式化表示如式(1)所示：

$$f = argmax_{c_i \in C} P(c_i|NI, EX, EC) \tag{1}$$

其中，NI指的是目标词触发的框架中所有核心框架元素集合$E = \{e_1, e_2, ..., e_n\}$减去句子中显式核心框架元素集合$EH = \{e_1, e_2, ..., e_k\}$剩下的部分，即$NI = E - EH$；$EX = \{e_1, e_2, ..., e_m\}$是句子中所有显式框架元素集合，即$EH \subseteq EX \subseteq E$，$EC = \{ec_1, ec_2, ..., ec_n\}$是显式框架元素填充内容集合，$C = \{c_1, c_2, ..., c_n, \phi\}$是零形式框架元素$NI$的候选填充项集合，称为**候选语集合**，如图1所示例句中，S2中“属于某类”框架中的零形式框架元素“实体”的候选语包括“十九世纪”、“一次要求变革而爆发的维新运动”等。若某个零形式框架元素在上下文中找不到填充内容，则认为该零形式框架元素为INI。

现有的零形式填充方法采用pipeline模型，先对零形式类型分类，之后通过手动设计与DNI框架元素候选语关联的信息作为特征，学习候选语的表示。这些方法存在以下不足：1）零形式分类的结果会影响零形式填充的效果，造成错误传播。2）在学习候选语表示时没有考虑到显式框架元素及其填充内容的信息以及他们之间的框架语义关系。

针对以上问题，本文设计了一个端到端的基于框架语义场景图的零形式填充方法，该方法采用基于词汇级跨度的方式获取零形式框架元素的候选语集合，并充分利用显式框架元素及其填充内容的信息，提升了零形式填充模型的性能。另外，本文为了降低候选语过多对模型性能产生的影响，根据候选语与目标词的语义相关性提出了一种候选语剪枝方法。

本文的主要贡献包括：1）提出了一种端到端的模型结构，有效减少了零形式填充的错误传播；2）构建出框架语义场景图并利用图注意力网络对该框架语义场景图建模，得到融合了句子中显式框架元素及其填充内容以及框架语义信息的候选语表示；3）在CFN零形式填充数据集上的实验结果表明，本文提出的方法有效提升了零形式填充模型的性能。

2 相关工作

国际语言学会议ACL在2010年举办了关于“LinKing Event and Their Participants in Discourse”的语义评测 (Ruppenhofer et al., 2010)，该任务要求参赛者在已标注语义角色的语料上识别出有定的零形式并在上下文中为其找到相应的填充内容，使得零形式填充任务受到了广泛关注。现有的零形式填充方法主要分为基于规则与统计的方法以及基于词嵌入模型的机器学习方法。

早期对于零形式填充的研究结合规则与统计的方法，借助外部系统工具进行零形式填充实验。 (Tonelli et al., 2010)借助文本蕴涵识别系统VENSES (Delmonte et al., 2008)得到每个词元不同的标注模式，在训练语料中查找与其相似的谓词论元结构，找到后与其对比得出缺失的论元，最后通过计算缺失框架元素与候选语的相似度寻找其填充内容。 (Chen et al., 2010)融入统计方法，扩展了SEMAFOR1.0 (Das et al., 2010)工具，利用语义角色与候选语的相关性与相似性得分完成零形式填充任务，但其性能仍然较低。

随着机器学习的发展， (Gerber et al., 2010)在NomBank 语料库中，使用实体指代方法，结合句法、语义特征，完成零形式填充任务。 (Silberer et al., 2012)将该问题看作一个指代消解任务，将实体链作为候选填充项，结合语义角色标注和共指消解的特征，构建了有监督的机器学习模型。 (Wang et al., 2013)在传统的零形式填充特征上引入中心词信息和框架信息。 (Laparra et al., 2013)总结研究了传统指代消解所使用的特征，并尝试着将其应用到零形式的填充任务中。 (Wu et al., 2016)首先利用规则与过滤的方法进行零形式识别，再选取相关的语义特征，建立最大熵分类模型，实现了零形式的分类和填充。 (Li et al., 2018)提出了基于SVM的零形式分类并结合框架关系与框架语义特征，提升了零形式填充任务的效果。 (Zhang et al., 2020)从数据非平衡的角度出发,对非平衡数据进行平衡化处理,融入语义相似性特征及框架元素间的映射关系,提升零形式填充效果。但是，上述通过构造分类特征对零形式框架元素进行填充的方法，需要人工构造分类特征，效率低下且性能较低。

基于语义图的方法已经应用于自然语言处理的各种任务中。 (Pan et al., 2020)为了捕获问题生成任务中输入文档的全局结构，为输入文档构造语义图，并引入基于注意力机制的门控图神经网络学习语义图的表示。 (Guan et al., 2021)基于FrameNet构建了语义场景图和词关系

第二十二届中国计算语言学大会论文集，第399页-第409页，哈尔滨，中国，2023年8月3日至5日。

图，并设计图融合模块来为摘要生成任务获得丰富的语义表示。 (Zheng et al., 2022)通过构造框架知识图和框架语义图将框架语义解析转换为增量图构造问题，以加强子任务之间的交互和论元之间的关系。

早期的零形式填充方法通过计算框架元素与候选语的相似性与相关性来为DNI框架元素寻找填充内容。以机器学习为主的零形式填充方法采用两阶段法，先识别零形式类型，再进行零形式填充；并基于词嵌入模型通过手动设计与候选语关联的特征来学习候选语的表示。这些方法忽略了显式框架元素及其填充内容的重要性并且容易造成错误传播。因此本文提出了一种端到端的零形式填充模型，该模型使用Bert预训练模型获取候选语的上下文表示，并通过构建框架语义场景图来融合显式框架元素及其填充内容的信息以及框架语义信息。

3 基于框架语义场景图的零形式填充模型

框架语义分析就是从句子中把完整的语义结构抽取出来。而已有的框架语义角色标注都是基于句子级的，只能识别出目标词所在句子内的框架语义角色，不能将语义结构完整的表达出来，因此零形式填充对于汉语框架语义分析具有重要意义。显式框架元素及其填充内容对于DNI框架元素在上下文中找到填充内容起着很重要的作用，如图1S2中显式框架元素“范畴”的填充内容“中国近代史上的一次惨痛的流血事件”指的是一个“流血事件”，为其在上下文中寻找零形式框架元素“实体”的填充内容“一次要求变革而爆发的维新运动”提供了线索。另外，CFN中框架、框架元素以及框架元素填充内容是一个网状关系，因此本文使用图模型来融合显式框架元素信息以及框架语义信息，进而根据图中已知的显式框架元素及其填充内容信息以及框架语义信息推断出未知的部分。

考虑到候选语过多会对模型运行效率产生一定的影响，并且会引入噪声信息，所以本文根据目标词与候选语的语义相关性提出了一种候选语剪枝方法。

本文提出的基于框架语义场景图的零形式填充模型结构如图2所示，该模型包括：1）输入层：获取零形式填充模型所需的输入文本、框架语义信息以及零形式框架元素的候选语集合；2）编码层：利用Bert对上下文序列编码，并对每个框架和框架元素进行唯一编码；3）特征构造层：在编码层输出的基础上得到目标词、候选语以及显式框架元素填充内容的上下文表示；设计出一个候选语剪枝得分函数，选取前top-k个作为候选语集合；4）图注意力层：根据框架语义信息构造框架语义场景图，并使用GAT对框架语义场景图建模，得到融合了显式框架元素及其填充内容以及框架语义信息的候选语表示；5）预测层：将GAT学习到的候选语表示通过分类器分类，为DNI框架元素找到它的填充内容。

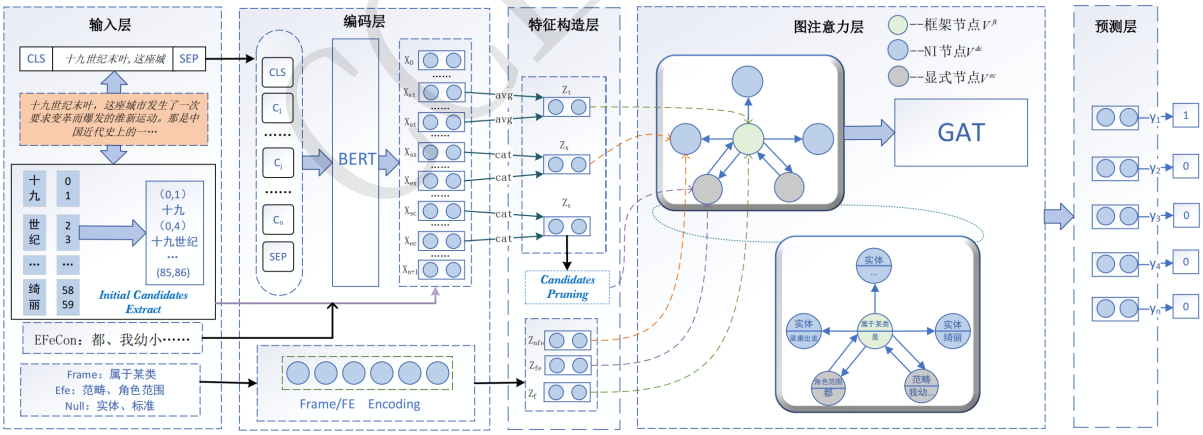


图 2: 模型结构图

3.1 输入层

输入层提供零形式填充模型所需的输入文本、框架语义信息以及零形式框架元素候选语集合。输入文本为原始的文本数据，把它输入零形式填充模型以获取目标词、显式框架元素填充

内容以及候选语的上下文表示。框架语义信息包括目标词所属框架名称、显式框架元素名称以及零形式框架元素名称。

本文针对中文数据更偏重以词为基本句法成分的特点，采用基于词汇级跨度的方法选取零形式框架元素的候选语集合，即以词为基本单位枚举出句子中所有的跨度$C=\{c_1,c_2,...,c_n\}$。考虑到目标词所在跨度以及已经是显式框架元素填充内容的跨度不可能是零形式框架元素的填充内容，所以在枚举出的跨度集合中排除掉这些跨度，同时还把候选语长度限制在max_length以内。另外，由于INI框架元素的填充内容为空，所以最后得到零形式框架元素候选语集合为$C=\{c_1,c_2,...,c_n,\phi\}$。

3.2 编码层

3.2.1 上下文编码

Bert以无监督的方式对未标记的语料库进行训练，能够学习语言中隐含且丰富的文本语义。其体系结构是一种多层双向TransformerEncoder，相比传统的Transformer拥有双向编码能力，可以更好的捕捉上下文信息获得动态词向量表示，具有更深的层数和并行性，进一步增加词向量模型泛化能力，充分提取到了字符级、词级、句子级和句间等特征。因此本文用Bert来进行上下文编码。

将$[CLS]+\{E_1,...,E_i,...,E_n\}+[SEP]$作为Bert模型输入，编码层将输入中的每一个字符编码成字符嵌入$E_{token}(t_i)$、分段嵌入$E_{seg}(t_i)$和位置嵌入$E_{pos}(t_i)$三个向量，将三个向量相加输入BERT预训练模型得到输出向量$X\subseteq R^{n\times d}$，如公式(2)、(3)所示。

$$E_i=E_{token}(i)+E_{seg}(i)+E_{pos}(i) \tag{2}$$

$$X=Bert(E_0,...,E_i,...,E_{n+1}) \tag{3}$$

3.2.2 框架信息编码

本文对目标词触发的框架、显式框架元素以及零形式框架元素进行唯一编码。如式（4）-（5）所示，f_{id}表示目标词触发框架f的id，fe_{id}表示某个框架元素fe的id。

$$Z_f=Embedding(f_{id}) \tag{4}$$

$$Z_{fe}=Embedding(fe_{id}) \tag{5}$$

3.3 特征构造层

3.3.1 上下文表示

将目标词开始位置信息记为st，结束位置信息记为et，目标词的上下文表示如公式（6）所示；将某个显式框架元素填充内容的开始位置信息记为sx，结束位置信息记为ex，长度记为$width_x$，则该填充内容的上下文表示如公式（7）所示，其中cat表示拼接函数；同理，初始候选语集合中某个候选语的上下文表示可以表示为公式（8），其中，sc、ec、$width_c$分别是该候选语的开始位置信息、结束位置信息以及长度信息。

$$Z_t=\frac{X_{st}+X_{et}}{2} \tag{6}$$

$$Z_x=cat[X_{sx};X_{ex};width_x] \tag{7}$$

$$Z_c=cat[X_{sc};X_{ec};width_c] \tag{8}$$

3.3.2 候选语剪枝

为了降低候选语过多对模型性能产生的影响，本文设置了如公式（9）所示的候选语剪枝得分函数，该得分函数反映了候选语与目标词的语义相关性，最后选取前top-k个跨度构成候选语集合。

$$S(Z_t,Z_c)={Z_t}^{\top}WZ_c+S_c(Z_c)+S_t(Z_t)+\phi(Z_t,Z_c) \tag{9}$$

$$S_c(Z_c)=w_c^{\top}F_c(Z_c) \tag{10}$$

$$S_t(Z_t)=w_t^{\top}F_t(Z_t) \tag{11}$$

其中，Z_c为初始候选语编码，Z_t为目标词编码，F为一个前馈神经网络，W、w_c、w_t为可学习的参数，$\phi(Z_t,Z_c)$为初始候选语与目标词的距离特征。

3.4 图注意力层

3.4.1 框架语义场景图构建

本文采用如图3所示的框架语义场景图$G=(V,E)$来融合显式框架元素及其填充内容以及框架语义信息，其中V是节点集合,E是边集合，$(v_i,v_j)\in E$表示节点v_i到v_j之间存在有向边。

如图3所示：图G的共有三种类型的节点：分别为框架节点V^{ft}，显式节点V^{ec}以及候选语节点V^{dc}，即$V^{ft}\subset V$，$V^{ec}\subset V$，$V^{dc}\subset V$。其中，V^{ft}表示融合了框架与目标词信息的节点，V^{ec}表示融合了显式框架元素及其填充内容信息的节点，V^{dc}表示融合了DNI框架元素以及该框架元素的候选语信息的节点。

为了使候选语表示融合丰富的框架语义信息以及显式框架元素及其填充内容信息，本文连接不同类型的节点形成三种不同类型的边：①框架节点V^{ft}到显式节点V^{ec}的边e^{tc}；②显式节点V^{ec}到框架节点V^{ft}的边e^{ct}；③框架节点V^{ft}到候选语节点V^{dc}的边e^{fd}；值得注意的是，图3中的虚线边仅表示框架与框架元素之间的关系，并不在边集合E中。由此可知，框架节点与每一个显式节点之间都是双向边，而框架节点与每一个候选语节点之间都是单向边，只存在框架节点到候选语节点的边，这样可以使候选语节点在融合框架信息以及显式框架元素及其填充内容信息的同时，防止不同候选语节点之间相互干扰。

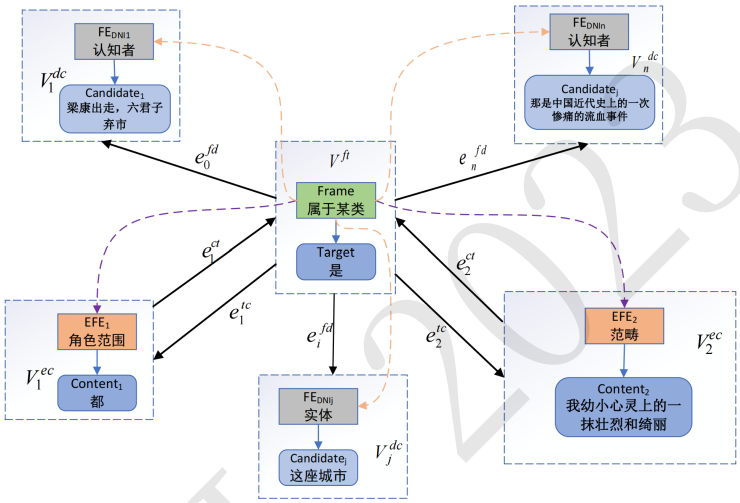


图 3: 框架语义场景图示例

3.4.2 框架语义场景图编码

为了更好的融合显式框架元素及其填充内容的信息以及框架语义信息，本文使用GAT来更新框架语义场景图的节点信息，得到聚合了显式框架元素及其填充内容信息、目标词与框架信息以及零形式框架元素信息的候选语表示。其公式表示如式（12）所示。

$$R_G=GAT(Z,A) \tag{12}$$

其中，$A\in R^{m\times m}$表示图的邻接矩阵，$A\in\{0,1\}$表示节点V_i到节点V_j之间是否存在有向边。$Z\in R^{m\times d}$表示节点的初始特征矩阵。框架节点V^{ft}初始编码由框架编码Z_f与目标词编码Z_t拼接而成；显式节点V^{ec}初始编码由显式框架元素编码Z_{fe}与显式框架元素填充内容编码Z_x拼接而成；候选语节点V^{dc}初始编码由零形式框架元素编码Z_{feNI}与候选语编码Z_c拼接而成。

3.5 预测层

将图编码层获取的某个候选语表示$R_G(c_i)$进行线性变换和非线性激活得到该候选语作为零形式框架元素填充内容的概率p_i，如公式（13）所示，其中L代表线性变换层。最后取概率最大的位置作为当前预测的结果类别。

$$p_i=Sigmoid(L(R_G(c_i))) \tag{13}$$

本文采用二分类交叉熵损失作为分类损失，如公式（14）所示，其中p_{real}表示真实样本类别分布，p_i表示模型预测出的样本类别分布。

$$Loss = \sum_{i=1}^{n} -(p_{real}\log(p_i) + (1 - p_{real})\log(1 - p_i)) \tag{14}$$

本文通过阈值t来判定零形式框架元素是否为INI，若某个零形式框架元素概率最大的候选语仍小于阈值t，则认为该零形式框架元素为INI，不需要寻找填充内容。

4 实验设计与分析

4.1 实验数据

本文使用的零形式填充数据来源于CFN数据库，语料主要来源于阅读理解和人民日报，其中涉及天文、历史等14 个领域。本文使用山西大学人机协同标注系统进行人工语义角色标注，共包含19285条数据，涉及到538个框架，并按照8:1:1的比例分配得到训练集、验证集和测试集，如表2所示。

	数据总数	框架数
训练集	11571	512
验证集	3857	438
测试集	3857	436

表 2: 数据集分布

4.2 评价指标及实验环境

本实验使用准确率(Precision)，召回率（Recall），F1值作为评价指标。本实验的实验环境为pytorch1.11.0+cu102，python版本为3.6.2，GPU为TeslaP100-PCIE-16GB。

4.3 参数设置

本实验参数包括train_batch_size、学习率、Bert最大序列长度、GAT隐层维度、GAT层数、候选语选取最大数量以及判断零形式框架元素是否为INI的阈值，具体参数设置如表3所示。

参数名	参数值
train_batch_size	16
learning_rate（学习率）	1e-6
max_seq_length（Bert最大序列长度）	512
GAT_hidden_size（GAT隐层维度）	768
GAT_Layers（GAT层数）	3
top-k（候选语选取最大数量）	80
max_length（候选语最大长度）	30
t (阈值)	0.5

表 3: 参数设置

4.4 实验结果及分析

4.4.1 零形式填充对比实验

为了验证本文提出方法的有效性，本文设置了如下对比实验，实验结果如表4所示。

（1） (Wang et al., 2013)、 (Wu et al., 2016)、 (Zhang et al., 2020)三组对比实验,这组对比实验都是通过手动设计与候选语关联的特征来学习候选语的表示，并且未利用预训练模型，上下文特征抽取能力较弱；

（2）本文设计了一种基于Bert的基线模型：通过Bert预训练模型得到候选语的初始上下文编码Z_c，之后将零形式框架元素编码$Z_{fe^{NI}}$与候选语编码Z_c拼接，得到候选语表示Z_{cf}，最后选取得分最高的候选语作为DNI框架元素的填充内容。

（3）将每个显式框架元素编码Z_{fe}与其填充内容编码Z_x拼接后得到显式向量Z_e^i，将框架编码Z_f与目标词编码Z_t拼接后得到框架向量Z_{ft}；在基线模型的基础上设置了不同的方式来融合显式向量$Z_e = \{Z_e^0, Z_e^2, ..., Z_e^n\}$，以增强候选语的表示，进而验证本文提出的融合方法的有效性：

①BertAvg:将框架向量以及所有显式向量Z_e求平均后向量$Avg(Z_e)$加到候选语表示Z_{cf}中：

$$Z_c^{new} = Z_{cf} + Z_{ft} + Avg(Z_e) \tag{15}$$

②BertAtt：将所有显式向量Z_e输入Attention机制，给每个显式向量赋予不同的重要性后加到候选语表示Z_{cf}中：

$$Z_c^{new} = Z_{cf} + Z_{ft} + Attention(Z_e) \tag{16}$$

（4）为了证明本文提出的端到端的模型有效减少了错误传播，本文设计了一种管道模型（BertPipline）先对零形式框架元素类型分类，再利用本文提出的框架语义场景图对分类为DNI的框架元素寻找填充内容。

（5）本文提出的基于框架语义场景图的零形式填充方法（BertFSG）

Model	P(%)	R(%)	F1(%)
(Wang et al., 2013)	31.93	12.76	18.23
(Wu et al., 2016)	39.15	27.36	32.21
(Zhang et al., 2020)	38.76	48.80	44.34
Bert	83.90	42.42	54.02
BertAvg	86.63	42.98	55.05
BertAtt	81.13	46.88	56.93
BertPipeline	47.42	50.46	46.40
BertFSG	**87.74**	**52.62**	**63.18**

表 4: 零形式填充对比实验结果

表4的实验结果显示，本文提出的模型BertFSG优于之前的方法，相比 (Wang et al., 2013)的方法和 (Wu et al., 2016)的方法F1值分别提升了44.95%和30.97%，相比 (Zhang et al., 2020)的方法F1值提升了18.84%；从基于Bert的基线模型和第（1）组对比实验的对比结果可知，Bert可以更充分的利用上下文信息，具有更强的信息表达能力；BertAvg、BertAtt与基于Bert的基线模型相比，F1值分别提升了1.03%、2.91%,由此可以说明，显式框架元素对于零形式填充的重要性；BertAvg与BertAtt的对比实验结果可知，每个显式框架元素及其填充内容对于零形式填充有不同的重要性；本文提出的模型BertFSG与基于Bert的基线模型以及BertAvg、BertAtt相比，F1值分别提升了9.16%、8.13%、6.25%，证明了本文提出的利用框架语义场景图融合显式框架元素及其填充内容的方式对于零形式填充的有效性；从BertPipeline与BertFSG的对比实验结果可知，零形式类型分类结果会对零形式填充结果产生很大的影响，证明本文提出的端到端的零形式填充模型的可靠性。

4.4.2 超参数分析

本小节探讨了候选语最大长度max_length、阈值t以及候选语选取最大数量top-k对实验结果的影响。表5为阈值t为0.5，top-k为80时，max_length对实验结果结果的影响，表6为max_length为40，top-k为80时阈值t对实验结果的影响，表7为max_length为40，阈值t为0.5时，top-k对实验结果的影响。

max_length	P(%)	R(%)	F1(%)
10	82.10	49.47	59.18
20	86.28	49.52	60.29
30	87.74	**52.62**	**63.18**
40	**88.08**	51.18	62.12
50	83.56	47.78	58.42
60	81.02	46.05	56.42

表 5: max_length对实验结果影响

t	P(%)	R(%)	F1(%)	t	P(%)	R(%)	F1(%)
0.10	63.77	43.55	49.42	0.50	**88.08**	**51.18**	**62.12**
0.20	79.67	44.63	54.41	0.60	86.92	44.82	55.09
0.30	87.11	48.50	58.49	0.70	86.53	43.23	53.84
0.40	87.70	49.05	59.00	0.80	83.08	41.23	46.97

表 6: 阈值t对实验结果影响

top-k	P(%)	R(%)	F1(%)
50	85.89	50.24	61.06
60	85.38	50.83	61.28
70	87.55	50.96	61.79
80	**88.08**	**51.18**	**62.12**
90	87.67	49.58	61.18
100	86.29	47.67	58.84

表 7: top_k对实验结果影响

从表5的实验结果可知，随着max_length的增加，F1值呈升高趋势，当max_length为30时，模型效果达到最佳，当max_length继续增大时，F1值逐渐减小。由此可知，候选语最大长度太长或太短都会对模型效果产生影响。候选语最大长度太长时，零形式框架元素候选语数量会增多，给零形式填充模型带来了噪声干扰；候选语最大长度太短时，会导致候选语集合覆盖率降低，进而影响零形式填充的效果。从表6的实验结果可知，t从0.1到0.5时,F1值逐渐增大，当t取0.5时，模型效果达到最优，随着t继续增大，F1值逐渐下降。当阈值t太小时，会使部分INI错被模型当作DNI；当阈值t太大时，会使部分DNI被错误识别。由此可见，阈值t太大或太小都会影响模型性能。从表7的实验结果可知，top-k从50到80时，F1值逐渐增大，当top-k取80时，模型效果最好，随着top-k继续增大，F1值逐渐下降。

4.4.3 消融实验

为了验证不同模块的有效性，本文分别将候选语剪枝（CSE）模块以及在框架语义场景图中去掉显式节点（EFE），来检测模型性能的变化，具体结果如表8所示。

从表8的实验结果可以看出：

（1）把CSE模块去掉后，模型F1值下降了4.38%，说明候选语过多对基于跨度的模型性能会产生一定的影响，也证明了本文提出候选语剪枝方法的有效性。

（2）把EFE模块去掉后，模型F1值下降了3.36%，说明了显式框架元素及其填充内容对零形式填充的重要性。

第二十二届中国计算语言学大会论文集，第399页-第409页，哈尔滨，中国，2023年8月3日至5日。

Models	P(%)	R(%)	F1(%)
-CSE	85.78	47.74	58.80
-EFE	85.12	49.41	59.82
-All	84.88	45.96	57.22
Ours	**87.74**	**52.62**	**63.18**

表 8: 消融实验结果

4.4.4 案例分析

本小节从零形式填充数据中选取了一条进行分析，如图4所示S2中目标词“给”触发“给予”框架，该框架包含“捐赠者”、“接受者”、“转移体”三个核心框架元素以及“方法”等非核心框架元素，框架元素“接受者”、“转移体”、“方法”为显式框架元素，核心框架元素“捐赠者”在S2中没有填充内容，因此需要为其在篇章上下文中寻找填内容。本文模型通过构建框架语义场景图融合显式框架元素信息，根据显式框架元素及其填充内容线索，利用框架语义场景图推断给予“接受者”“转移体”的“捐赠者”是“该公司”，进而成功找到有定零形式框架元素“捐赠者”的填充内容。而当在框架语义场景图中把显式框架元素信息去掉之后，模型将零形式框架元素“捐赠者”的填充内容填充为错误答案“产品良好的质量保证”。由此可见，显式框架元素信息的融入，增强了候选语的表示，有效提升了零形式填充的效果。

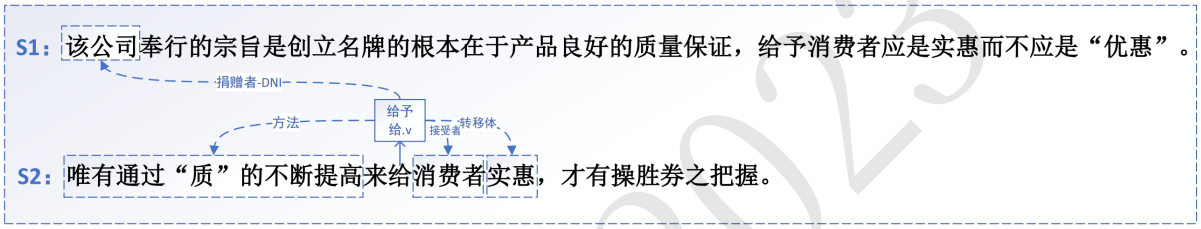


图 4: 案例分析示例图

5 总结

本文针对汉语进行了零形式填充的研究，提出了一种基于框架语义场景图的零形式填充模型，相较于先前工作使用的词嵌入模型，该模型使用Bert增强了模型抽取特征的能力，可以获得更为丰富的语义信息，本文结合框架语义信息构建框架语义场景图，并通过GAT对该框架语义场景图编码，得到融合了显式框架元素及其填充内容以及框架语义信息的候选语表示。实验结果证明，本文的方法相较于当前最好的模型F1值提升了18.84%，验证了该填充方法的可行性。零形式填充任务目前还面临着数据规模有限、数据不平衡等问题，如何利用ChatGPT等大模型扩充数据规模、解决数据不均衡性将是后续的研究重点。

参考文献

You, Liping and Liu, Kaiying. 2005. *Building chinese framenet database*. Natural Language Processing and Knowledge Engineering, pages:301-306.

Fillmore, Charles J and others. 1976. *Frame semantics and the nature of language*. Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech, 280:20-32.

Baker, Collin F and Fillmore, Charles J and Lowe, John B. 1998. *The berkeley framenet project*. COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics, pages:86-90.

石佼, 李茹, 王智强. 2014. 汉语核心框架语义分析. 中文信息学报, 28(6):48-55.

李茹. 汉语句子框架语义结构分析技术研究. 山西大学, 2012.

Veličković, Petar and Cucurull, Guillem and Casanova, Arantxa and Romero, Adriana and Lio, Pietro and Bengio, Yoshua. 2017. *Graph attention networks.* arXiv preprint arXiv:1710.10903.

Ruppenhofer, Josef and Sporleder, Caroline and Morante, Roser and Baker, Collin and Palmer, Martha. 2010. *SemEval-2010 Task 10: Linking Events and Their Participants in Discourse.* Proceedings of the 5th International Workshop on Semantic Evaluation, pages:45-50.

Tonelli, Sara and Delmonte, Rodolfo. 2010. *VENSES++: Adapting a deep semantic processing system to the identification of null instantiations.* Proceedings of the 5th International Workshop on Semantic Evaluation, pages:296-299.

Delmonte, Rodolfo. 2008. *Computational Linguistic Text Processing - Lexicon, Grammar, Parsing and Anaphora Resolution.* New York: Nova Science.

Chen, Desai and Schneider, Nathan and Das, Dipanjan and Smith, Noah A. 2010. *SEMAFOR: Frame Argument Resolution with Log-Linear Models.* Proceedings of the 5th International Workshop on Semantic Evaluation, pages:264-267.

Das, Dipanjan and Schneider, Nathan and Chen, Desai and Smith, Noah A. 2010. *Probabilistic Frame-Semantic Parsing.* Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages:948-956.

Gerber, Matthew and Chai, Joyce. 2010. *Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates.* Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages:1583-1592.

Silberer, Carina and Frank, Anette. 2012. *Casting Implicit Role Linking as an Anaphora Resolution Task.* *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages:1-10.

Ning Wang, Ru Li, Zhangzhang Lei, Zhiqiang Wang, Jingpan Jin. 2013. *Document Oriented Gap Filling of Definite Null Instantiation in FrameNet.* Proceedingsof Chinese Computational Linguistics and Natural Language Processing Based on Natu-rally Annotated Big Data, pages:85-96.

Laparra, Egoitz and Rigau, German. 2013. *Sources of Evidence for Implicit Argument Resolution.* Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, pages:155-166.

武娟, 李茹, 王智强. 2016. 汉语篇章中零形式的识别与消解. 中文信息学报, pages:9-15.

雷章章, 王宁, 李茹. 2013. *FrameNet* 中有定的零形式识别. 中文信息学报, pages:107-113.

李茹, 郭倩. 2018. 基于汉语框架语义关系的零形式识别与消解. 山西大学学报: 自然科学版, pages:41-49.

张月平，李茹，王元龙，柴清华，武宇娟，关勇. 2020. 汉语语篇零形式识别与填充方法研究. 计算机工程, pages:79-86.

Pan, Liangming and Xie, Yuxi and Feng, Yansong and Chua, Tat-Seng and Kan, Min-Yen. 2020. *Semantic Graphs for Generating Deep Questions.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages:1463-1475.

Guan, Yong and Guo, Shaoru and Li, Ru and Li, Xiaoli and Zhang, Hu. 2021. *Integrating Semantic Scenario and Word Relations for Abstractive Sentence Summarization.* Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages:2522-2529.

Zheng, Ce and Chen, Xudong and Xu, Runxin and Chang, Baobao. 2022. *A Double-Graph Based Framework for Frame Semantic Parsing.* Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages:4998-5011.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages:4171-4186.

基于FLAT的农业病虫害命名实体识别

任义[1],沈洁[*1],袁帅[1]
1.沈阳建筑大学，计算机科学与工程学院，辽宁沈阳 110168
renyi@sjzu.edu.cn,1007404126@qq.com,yuanshuai@sjzu.edu.cn

摘要

针对传统命名实体识别方法中词嵌入无法表征一词多义及字词融合的模型存在特征提取不够准确的问题，本文提出了一种基于FLAT的交互式特征融合模型，该模型首先通过外部词典匹配获得字、词向量，经过BERT预训练后，通过设计的交互式特征融合模块充分挖掘字词间的依赖关系。另外，引入对抗训练提升模型的鲁棒性。其次，采用了特殊的相对位置编码将数据输入到自注意力机制，最后通过CRF得到全局最优序列。本文模型在农业病虫害数据集上识别的准确率、召回率、F1值分别达到了93.76%、92.14%和92.94%。

关键词： 命名实体识别； 农业病虫害； 对抗训练； 特征融合； 自注意力机制

Named Entity Recognition of Agricultural Pests and Diseases based on FLAT

Yi Ren[1],Jie Shen[*1],Shuai Yuan[1]
1.School Of Computer Science And Engineering, Shenyang Jianzhu University
Shenyang, Liaoning 110168, China
renyi@sjzu.edu.cn,1007404126@qq.com,yuanshuai@sjzu.edu.cn

Abstract

Aiming at the problem that feature extraction is not accurate enough in the traditional named entity recognition method in which word embedding cannot represent the polysemy of a word and word fusion model, this paper proposes an interactive feature fusion model based on FLAT. The model first obtains word and word vector through external dictionary matching. After BERT pre-training, The interactive feature fusion module is designed to fully explore the dependency relationship between words. In addition, adversarial training is introduced to improve the robustness of the model. Secondly, a special relative position encoding is used to input the data into the self-attention mechanism, and finally the globally optimal sequence is obtained by CRF. The identification accuracy, recall rate and F1 value of the model in the agricultural pest and disease data set reached 93.76%, 92.14% and 92.94%, respectively.

Keywords: Named entity recognition , Agricultural pests and diseases , Adversarial training , Feature fusion , Self-attention mechanisms

1 引言

*通讯作者
国家自然科学基金（62073227）；辽宁省教育厅基金（LJKZ0581，LJKZ0584）

目前我国农业的发展受到多种因素制约，除了水涝干旱等自然灾害外，农业病虫害是农业生产最为常见的问题之一。面对海量的非结构化的农业病虫害相关文本数据，人们无法快速准确获取到农业病虫害的防治信息。为了更好地解决我国农业生产实践中遇到的病虫害问题，信息化防治农业病虫害成为提高农业生产效率、增加农业产量的重要手段。农业病虫害命名实体识别任务能够帮助准确识别和分类出相关命名实体。

命名实体识别(王颖洁 et al., 2023)又称实体抽取，主要分为基于规则、机器学习和深度学习的方法。基于规则的方法需要大量领域专家来构建规则，人工及时间成本太高，可迁移性差(Xu K et al., 2019)。基于机器学习的方法主要包括隐马尔科夫模型(Morwal S, 2012)、支持向量机(Isozaki H and Kazawa H, 2002)、最大熵模型(Saha S K et al., 2009)和条件随机场(Lafferty J et al., 2001)。但是，基于机器学习的方法依赖人工制定的特征模板，不具备领域通用性。近年来，随着深度学习的发展，基于深度学习的命名实体识别方法，因其具有能够从数据中自主学习特征而无需人为设定，逐步成为中文命名实体识别的主流模型。刘新亮(2021)等提出一种基于BERT-CRF模型的命名实体识别方法，完成对生鲜蛋供应链领域的命名实体识别。郭知鑫(2021)等提出基于BERT-BiLSTM-CRF的实体识别模型，对法律文本中的案件实体进行智能识别。郭军成(2021)等基于BiLSTM-CRF模型融入BERT层，实现中文简历命名实体识别。由于上述方法未考虑到词向量特征对实体识别效果的影响，Y.Zhang(2018)首次提出了Lattice-LSTM模型，该模型利用字向量和词向量信息，在公开数据集上取得了较好的结果。基于Lattice的模型难以充分利用GPU的并行计算，推理速度通常较慢，X.Li(2020)提出FLAT模型，它将Lattice结构转换为平面结构，在保留Lattice原有信息的基础上提高了并行计算的能力。在以往的研究中，字词融合大多采用不同特征表示向量（如字符向量、词向量）的拼接或累加的方式提取信息，这就造成了不同特征表示之间的相互依赖关系被忽略。

近年来，基于通用领域的命名实体识别已经相对成熟，由于缺少公开标注的数据集，针对农业病虫害领域的命名实体识别研究仍处于探索阶段，现阶段国内外只有少数学者针对农业领域开展了一定研究。李想(2017)等、张剑(2018)等提出基于条件随机场的方法，对农作物、病虫害、农药进行实体识别。郑泳智(2021)等将BERT模型和BiLSTM-CRF模型相结合，实现对农业病虫害领域的命名实体识别。目前，极少数的研究将字词融合运用到农业病虫害领域的命名实体识别中。基于上述问题，本文提出了一种基于FLAT的交互式特征融合模型，对字级嵌入和词级嵌入特征向量进行交互学习，并加入对抗训练以提升模型的鲁棒性。

2 数据处理与标注

2.1 数据获取

本文通过数据获取、数据标注两个步骤，建立农业病虫害领域的数据集。农业病虫害领域命名实体识别目前还没有公开可使用的数据集，本文节选了百度百科有关农业病虫害的信息作为文本语料初始数据。通过数据清洗、去噪、去冗等预处理，保证数据的可靠性。接下来就是标注标签工作，由于是自己定义的标签类别，所以需要人工手动标注，而实体的标注需要大量特定领域的知识，从而又增加了标注的难度。另外，本研究通过查阅资料和咨询专家的方法，基于现有研究本文进一步将疾病类别划分为更细粒度的实体，分别为“病害”和“虫害”。此外，与农作物相关的实体，如防治药剂和为害症状等也被考虑在内。经过数据清洗后的文字约6万字，包含1476个句子，其中有1968个农业病虫害实体。

2.2 数据标注

本文预先定义好的命名实体类别包括病害、虫害、防治药剂、防治方法、为害症状、为害地区、作物，实体定义和样例如表1。通过使用Brat标注工具(Mahanazuddin S et al., 2021)，对获取的语料进行人工标注。本文采用BIO规则对语言序列进行标注，其中B（Begin）描述句子中每个命名实体的开始位置；I（Internal）描述命名实体除起始位置外的其他部分，O（Other）用来描述句子中其他非预先定义好的实体。标注中为了更好地识别命名实体的类别信息，本文将类别信息与BIO规则进行融合，类别信息如病害实体用（-Disease）表示、为害地区实体用（-Area）表示等。以句子“水稻云形病主要发生在长江流域”为例，其序列标注如图1所示。

实体	定义	样例
病害	农业病害名称	水稻纹枯病、赤霉病
虫害	农业虫害名称	褐飞虱、白粉虱
防治药剂	防治农业病虫害的药剂学名、俗名、生物防治药剂名	井冈霉素、多菌灵
防治方法	防治农业病虫害的农业防治方法、生物防治方法等	加强肥水管理、检疫、水旱轮作
为害症状	农业病虫害危害作物的特征	失水青枯、枯萎霉烂
为害地区	发生农业病虫害的地区	长江流域、江苏、海南
作物	农作物名称及品种	水稻、小麦、临稻6号、辽粳326

表 1: 实体定义和样例

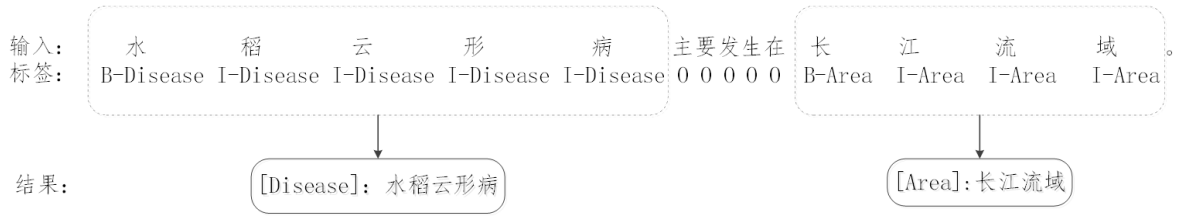


图 1: 语料序列标注示例

3 本文模型

本文提出了一种基于交互式特征融合与对抗训练的农业病虫害命名实体识别模型。对于输入序列$X = \{x_1, x_2, x_3, ..., x_n\}$,将$x_i$与外部词典匹配得到句子中的潜在词汇向量$d_i$，$x_i$通过BERT-WWM层生成具有丰富信息的字向量$c_i$，$c_i$与词典特征$d_i$拼接得到向量$w_i$，$w_i = c_i \oplus d_i$。随后，$w_i$与$d_i$进行交互式特征融合，接着对融合之后的向量进行对抗训练。使用相对位置编码将信息输入到自注意力机制，为了提升模型的性能，引入残差连接、归一化和前馈神经网络，最后通过线性层和条件随机场得到全局最优的标注结果。整体模型架构如图2所示。

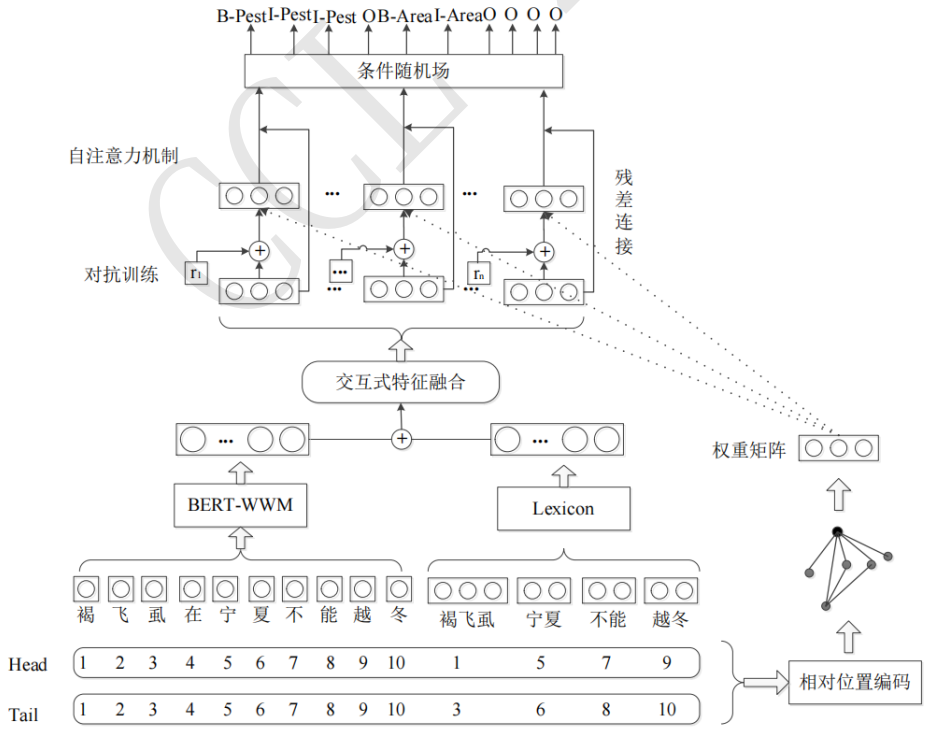


图 2: 总体模型架构

第二十二届中国计算语言学大会论文集，第410页-第419页，哈尔滨，中国，2023年8月3日至5日。

3.1 交互式特征融合

为了解决特征融合时不同特征之间的相互依赖关系被忽略的问题，而传统的字词融合只是通过向量的简单累加来达到融合的目的，本文提出一种交互式特征融合机制，如图3所示。

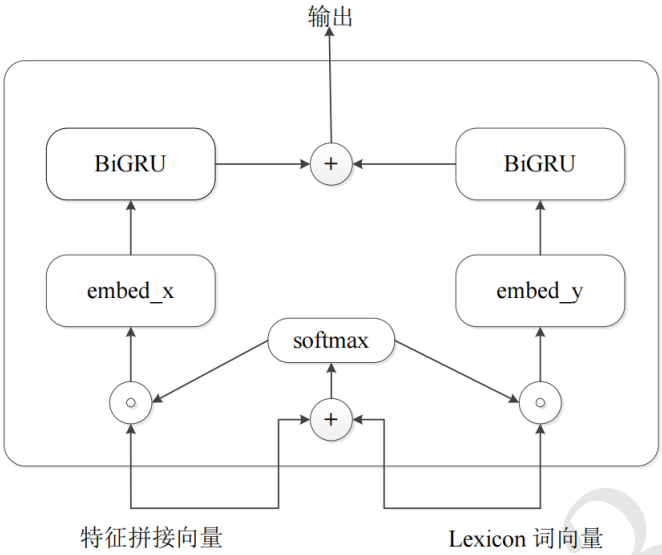


图 3: 交互式特征融合结构图

该机制通过一种交互的方式将不同特征进行充分融合。上述提及的d_i和w_i先经过简单融合，经过softmax函数筛选特征，之后再与w_i按元素相乘，这样通过以w_i特征向量为主体融合字词特征向量的部分重要特征得到$embed_x$，同理可得$embed_y$，其数学原理可以概括为下式(1)-(2)。

$$embed_x = w_i \odot \text{softmax}(w_i + d_i) \tag{1}$$

$$embed_y = d_i \odot \text{softmax}(w_i + d_i) \tag{2}$$

本文引入softmax激活函数更新不同特征向量的权重，从而提高有用信息的比重，丰富输入信息的特征表示。式中，$\odot$表示哈达玛积。

将特征交互得到的$embed_x$和$embed_y$分别经过BiGRU(Bidirectional Gated Recurrent Units)训练后进行融合，进一步丰富输入序列的语义表示，最终实现了整个交互式特征融合，其过程可概括为式(3)。

$$output = \text{BiGRU}(embed_x) + \text{BiGRU}(embed_y) \tag{3}$$

通过BiGRU层特征提取后，可以更加充分捕获上下文之间的关系。GRU工作原理的详细计算公式如式(4)-(7)所示。

$$r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)}) + b_{hr}) \tag{4}$$

$$z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)}) + b_{hz}) \tag{5}$$

$$n_t = \tanh(W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{(t-1)} + b_{hn}) \tag{6}$$

$$h_t = (1 - z_t) * n_t + z_t \odot h_{(t-1)} \tag{7}$$

式中，x_t表示t时刻的输入信息，$h_{(t-1)}$表示$t-1$时刻的隐藏状态，h_t表示t时刻的隐藏状态，W与b分别为权重矩阵与偏置项，σ为sigmoid非线性变换函数，tanh为激活函数，$\odot$为哈达玛积，r_t为重置门，z_t为更新门，n_t为候选隐藏状态。更新门与重置门通过sigmoid函数将值压缩在0到1之间。重置门用于决定前一时刻的状态是否需要保留，更新门则显示了保留前一时刻状态的程度。候选隐藏状态包含了t时刻的输入x_t的信息和对$t-1$时刻的隐藏状态$h_{(t-1)}$选择性的

保留了信息。

简单的GRU不能充分利用文本的上下文信息，本文设计使用BiGRU网络模型来提取序列信息的关键特征。BiGRU由两层GRU组成，分别是前向GRU与后向GRU，将他们各自得到的隐藏层状态拼接得到最终的隐藏层状态，详情见如下公式(8)-(10)。

$$\overrightarrow{h_t} = \overrightarrow{\mathrm{GRU}}(x_t) \tag{8}$$

$$\overleftarrow{h_t} = \overleftarrow{\mathrm{GRU}}(x_t) \tag{9}$$

$$h_t = (\overrightarrow{h_t}, \overleftarrow{h_t}) \tag{10}$$

式中，$\overrightarrow{h_t}$、$\overleftarrow{h_t}$分别为前、后向GRU的隐藏层状态，$\overrightarrow{\mathrm{GRU}}(x_t)$为前向GRU的计算过程，$\overleftarrow{\mathrm{GRU}}(x_t)$为后向GRU的计算过程，隐藏状态$h_t$为序列信息的特征表示。

3.2 对抗训练

对抗训练最早应用于计算机视觉领域，指在训练样本中添加一些可能导致误分类的微小扰动，并使神经网络适应这种改变，现已逐渐引入到自然语言处理领域中。本文利用对抗训练在保留原始数据的基础上加入对抗样本来发挥部分数据增强的作用，提高模型识别边界模糊实体的能力，进而提升模型在农业病虫害命名实体识别时的性能及鲁棒性。

对抗训练的数学原理可以概括为式(11)。

$$\min_{\theta} E_{(x,y)\sim D}\left[\max_{r_{adv}\in\Omega} L\left(\theta, x + r_{adv}, y\right)\right] \tag{11}$$

它可以看成由2部分组成，分别是内部扰动最大化和外部误差损失最小化。其中，r_{adv}表示在输入样本中添加的扰动，Ω表示扰动的空间范围，L表示损失函数，θ表示模型参数，x表示输入样本，y表示样本的标签，D表示输入样本的空间分布。

常用的对抗训练算法有FGM(Fast Gradient Method)、PGD(Projected Gradient Descent)和FreeLB(Free Large-Batch)，由于本文模型的计算量较大，本文选用训练速度较快的FGM对抗训练算法，其扰动r_{adv}的计算方法如式(12)-(13)所示。

$$r_{adv} = \epsilon \cdot \left(\frac{g}{\|g\|_2}\right) \tag{12}$$

$$g = \bigtriangledown_x(L(x, y, \theta)) \tag{13}$$

式中，ϵ为超参数的小有界范数，$\|g\|_2$表示梯度的$L2$范数，g为损失函数关于x的梯度。得到对抗样本X_{adv}如式(14)。

$$X_{adv} = x + r_{adv} \tag{14}$$

对抗样本会模仿标签中数据集的自然误差，使模型更能容忍模型参数波动带来的变化，从而提高模型的鲁棒性。在生成对抗样本之后，将交互式特征融合之后的向量与对抗样本一起送入自注意力机制训练。

3.3 相对位置编码

H.Yan(2019)指出，位置和方向信息在命名实体识别任务中非常重要。对于lattice中的两个span如x_i和x_j，本文使用4种相对距离来表示x_i和x_j之间的关系，其计算公式如式(15)-(18)所示。

$$d_{ij}^{(hh)} = head[i] - head[j] \tag{15}$$

$$d_{ij}^{(ht)} = head[i] - tail[j] \tag{16}$$

$$d_{ij}^{(th)} = tail[i] - head[j] \tag{17}$$

$$d_{ij}^{(tt)} = tail[i] - tail[j] \tag{18}$$

式中，$d_{ij}^{(hh)}$表示x_i的$head$与x_j的$head$之间的距离，$d_{ij}^{(ht)}$表示x_i的$head$与x_j的$tail$之间的距离，$d_{ij}^{(th)}$表示x_i的$tail$与x_j的$head$之间的距离，$d_{ij}^{(tt)}$表示x_i的$tail$与x_j的$tail$之间的距离。举例来说，如图2，假设“越冬”与“褐飞虱”分别为x_i与x_j，则$d_{ij}^{(ht)} = head[i] - tail[j] = head[\prime越\prime] - tail[\prime虱\prime] = 6$，同理可得$d_{ij}^{(hh)}$=8，$d_{ij}^{(th)}$= 9，$d_{ij}^{(tt)}$= 7。

最终的相对位置编码通过这4种相对距离经过简单的非线性变换得到，具体计算公式如式(19)。

$$R_{ij} = \text{ReLU}(W_r(p_{d_{ij}^{(hh)}} \oplus p_{d_{ij}^{(th)}} \oplus p_{d_{ij}^{(ht)}} \oplus p_{d_{ij}^{(tt)}})) \tag{19}$$

式中，W_r为可学习参数，$\oplus$表示拼接运算，p_d的计算与原始的Transformer相同，计算公式如下。

$$p_d^{(2k)} = \sin(\frac{d}{10000^{2k/d_{model}}}) \tag{20}$$

$$p_d^{(2k+1)} = \cos(\frac{d}{10000^{2k/d_{model}}}) \tag{21}$$

式中，d表示$d_{ij}^{(hh)}$、$d_{ij}^{(ht)}$、$d_{ij}^{(th)}$、$d_{ij}^{(tt)}$，k表示位置编码的维度索引。

本文使用改进后的自注意力机制，用新的注意力打分函数代替原本的缩放点积模型，公式如下。

$$\text{Attention}(A, V) = \text{softmax}(A)V \tag{22}$$

$$A_{ij} = W_q^T E_{x_i}^T E_{x_j} W_{k,E} + W_q^T E_{x_i}^T R_{ij} W_{k,R} + u^T E_{x_j} W_{k,E} + v^T R_{ij} W_{k,R} \tag{23}$$

$$[Q, K, V] = E_x[W_q, W_k, W_v] \tag{24}$$

式中，Q表示查询，K表示键，V表示值，d_{head}是多头注意力机制中每个头的维度，E是Embedding层，W、u、v为可学习参数，R_{ij}表示相对位置编码。

4 实验结果与分析

为验证本文模型的有效性，对自建的数据集按照训练集、测试集、验证集为6:2:2比例进行划分，验证集用于模型训练及优化，三个数据集无重叠交叉，因此测试集的训练结果可以作为模型性能的评价指标。

4.1 实验设置

实验采用Ubuntu操作系统，运行环境为RTX3080 GPU，内存为10G。模型所用的优化算法为SGD(Tian Yingjie et al., 2023)，为了缓解过拟合的问题，引入Dropout机制，模型参数设置如表2所示。

超参数设置	数值
隐藏单元数	8
多头注意力机制$head$个数	8
输入维度	160
学习率	6e-4
epoch	20
batch size	2
Dropout	0.5

表 2: 实验参数设置

4.2 评价指标

本文采用准确率（Precision，P）、召回率（Recall，R）以及准确率和召回率的调和平均数F1值（F1-score，F1）作为评价指标，其数值越高代表模型效果越好，各评价指标的计算公式如式(25)-(27)所示。

$$P = \frac{TP}{TP + FP} \times 100\% \tag{25}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{26}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{27}$$

式中，TP表示准确识别出的实体个数，FP表示错误识别出的实体个数，FN表示数据集中存在但未被识别出的实体个数。

4.3 对比实验

本文采用的主要对比模型如下：

BERT-BiLSTM-CRF：该模型是各领域命名实体识别基于字符的主流模型，主要由BERT-base-chinese预训练模型获取字向量，后输入到 BiLSTM 中对句子上下文进行建模，然后使用CRF进行解码。

BERT-IDCNN-CRF：与上述模型类似，其中IDCNN相比BiLSTM能够实现并行运算。

BERT-ADV-BiLSTM-CRF：在BERT-BiLSTM-CRF模型的基础上引入对抗训练，增强模型的泛化性。

BERT-FLAT：基于Transformer设计了一种巧妙位置编码来融合Lattice结构，可以无损引入词汇信息。

LatticeLSTM：加入了字词融合，避免因分词错误导致实体识别错误，是基于词汇增强的中文实体识别方法。

Radical-BiLSTM-CRF：采用BiLSTM-CRF神经网络，该网络同时利用字符级和部首级radical-level表示。

CNN-BiRNN-CRF：采用五笔画输入法获得笔画级表示，并将其与预训练的字符嵌入相连接，以探索字符的形态和语义信息，CNN用于提取N-gram特征。

LRCNN：改进LatticeLSTM，用CNN代替LSTM以提升性能，使用rethinking机制，通过高层特征的语义来优化词向量权重。

NFLAT：NFLAT对FLAT进行了解耦，第一阶段使用InterFormer融合词的边界和语义信息，第二阶段使用Transformer对上下文进行词汇信息编码，最后使用条件随机场作为解码器来预测序列标签，达到了SOTA。

本文在自建的农业病虫害数据集上设置五组模型分别进行实验，对比实验结果如表3所示。在模型训练中，本文的模型具有更加优异的表现。

模型	准确率/%	召回率/%	F1值/%
BERT-BiLSTM-CRF	83.95	71.53	77.24
BERT-IDCNN-CRF	80.47	72.24	76.13
BERT-ADV-BiLSTM-CRF	81.82	77.58	79.64
BERT-FLAT	88.65	91.09	89.85
本文模型	93.76	92.14	92.94

表 3: 对比实验结果

由表3可知，基于BERT-BiLSTM-CRF的识别效果优于BERT-IDCNN-CRF，这是因为IDCNN只能获取局部特征，而BiLSTM能够获取上下文全局特征。基于BERT-ADV-BiLSTM-CRF的识别效果较BERT-BiLSTM-CRF又有了较大的提升，表明了对抗训练的引入可以提高模型对实体边界的识别能力，可以证明对抗训练在命名实体识别任务中的有效性。

实验表明BERT-FLAT模型对于农业病虫害的识别效果相较其它模型有质的进步，所以本文在BERT-FLAT模型的基础上进行改进，本文模型的准确率、召回率、F1值分别提升了5.11%、1.05%和3.09%。

为了避免本文模型在自建的农业病虫害数据集上的实体识别结果具有偶然性以及证明模型在通用领域仍具有较好的表现，因此，在SIGHAN Bakeoff 2006的MSRA数据集上也进行了实验，MSRA数据集是微软亚洲研究院提供的较为权威的命名实体识别数据集，其规模如表4所示。

种类	训练集（K）	测试集（K）
句子	46.4	4.4
字符	2169.9	172.6
实体	74.8	6.2

表 4: MSRA数据集介绍

MSRA数据集包含三种实体，分别是人名、机构名、地名。MSRA数据集命名实体识别的实验结果如表5所列。Radical-BiLSTM-CRF模型同时利用字符级和部首级表示来提取特征，F1值为90.95%；CNN-BiRNN-CRF模型融入笔画特征以探索字符的形态和语义信息，F1值为91.67%；Lattice-LSTM在字向量的基础上引入词向量信息，且不会受到分词错误的影响，F1值为93.18%；LR-CNN在Lattice-LSTM的基础上使用rethinking机制来优化词向量权重同时提高了运行效率，F1值为93.71%；NFLAT去除self-attention的冗余计算，提高模型的性能和效率，F1值达到了94.55%。实验结果表示，本文模型在MSRA数据集上的表现更好，将F1值提升了0.97%。

模型	准确率/%	召回率/%	F1值/%
Radical-BiLSTM-CRF(Dong C et al., 2016)	91.28	90.62	90.95
CNN-BiRNN-CRF(Yang F et al., 2018)	92.04	91.31	91.67
Lattice-LSTM	93.57	92.79	93.18
LR-CNN(Tao Gui et al., 2019)	94.50	92.93	93.71
NFLAT(Wu S et al., 2022)	94.92	94.19	94.55
本文模型	95.39	95.65	95.52

表 5: MSRA数据集命名实体识别实验结果

4.4 消融实验

为了证明本文提出的模型中交互式特征融合与对抗训练这两部分的有效性，表6列出了模型在去除这两部分之后在自建农业病虫害数据集的性能。

模型	准确率/%	召回率/%	F1值/%
本文模型	93.76	92.14	92.94
BERT-FLAT-交互式特征融合	92.95	92.40	92.67
BERT-FLAT	88.65	91.09	89.85

表 6: 消融实验结果

由表6可知，在去除对抗训练和交互式特征融合后，模型的性能均有所下降：

（1）在去除对抗训练后，综合指标F1值下降了0.27%，这是由模型识别边界模糊实体的能力变弱、泛化性变差引起的。

（2）在（1）的基础上再去除交互式特征融合后，综合指标F1值下降了2.82%，这是因为特征融合变成了不同特征的简单累加，完全忽略了它们之间的相互依赖关系，语义特征无法得到充分表示。基于（1）、（2）所述，可以证明本文模型中交互式特征融合与对抗训练的有效性。

5 结语

由于农业领域缺乏公开的语料库，本文首先构建了农业病虫害命名实体识别的数据集。对于农业领域中文命名实体识别任务，考虑到用字词融合的方法来获取丰富的语义表示，在FLAT的基础上加入交互式特征融合模块以充分提取字词间的依赖关系，此外加入对抗训练来提升模型的鲁棒性和泛化性。下一步研究重点将集中在以下两方面：一是将自建的农业数据集进一步扩充以及修正或增强存在的噪音误差，以提升模型的识别效果。二是在扩充规模的数据集进行显著性验证。三是继续研究特征融合模块，引入更加丰富的特征信息。

参考文献

Dong C, Zhang J and Zong C. 2016. Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages. *Character-based LSTM-CRF with radical-level features for Chinese named entity recognition*, ICCPOL 2016, Kunming, China, December 2—6, 2016, Proceedings 24. Springer International Publishing, 2016: 239–250.

Isozaki H and Kazawa H. 2002. International Conference on Computational Linguistics-volume. *Efficient Support Vector Classifiers for Named Entity Recognition*, 1–7.

Lafferty J, McCallum A and Pereira F. C. 2001. ICML. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, 282–289.

Lin Y, Shen S and Liu Z. 2016. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). *Neural relation extraction with selective attention over instances*, 2124–2133.

Mahanazuddin S, Shaymaa A L S and Shorabuddin S. 2021. Studies in health technology and informatics. *DeIDNER corpus: Annotation of Clinical Discharge summary notes for named entity recognition USING BRAT TOOL*, 281–432.

Morwal S . 2012. *Named Entity Recognition using Hidden Markov Model (HMM).* Int.j.nat.lang.comput,1(4):15–23

Saha S K, Sarkar S and Mitra P. 2009. Journal of Biomedical Informatics. *Feature selection techniques for maximum entropy based biomedical named entity recognition*, 42(5):905–911.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. In Proceedings of the 28th International Joint Conference on Artificial Intelligence. *Cnn-based chinese ner with lexicon rethinking*, AAAI Press, 4982–4988.

Tian Yingjie, Zhang Yuqi and Zhang Haibin. 2023. Mathematics. *Recent Advances in Stochastic Gradient Descent in Deep Learning*, 11(3).

Wu S , Song X and Feng Z. 2022. arXiv. *NFLAT: Non-Flat-Lattice Transformer for Chinese Named Entity Recognition*,2205.05832.

Xiaonan L, Hang Y A N and Xipeng QIU. 2020. Association for Computational Linguistics. *FLAT: Chinese NER Using Flat-Lattice Transformer*, 6836–6842.

Xu K, Yang Z G and Kang P P. 2019. Computers in Biology and Medicine. *Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition*, 108(22):122–132.

Yan H, Deng B and Li X. 2019. arXiv preprint arXiv. *TENER: adapting transformer encoder for named entity recognition*,1911.04474.

Yang F, Zhang J and Liu G. 2018. Natural Language Processing and Chinese Computing: 7th CCF International Conference. *Five-stroke based CNN-BiRNN-CRF network for Chinese named entity recognition,NLPCC 2018, Hohhot, China, August 26—30, 2018, Proceedings, Part I 7.* Springer International Publishing, 2018: 184–195.

ZHANG Y and YANG J. 2018. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics，Association for Computational Linguistics. *Chinese NER using lattice LSTM*, 1554–1564.

郭军成,万刚,胡欣杰. 2021. 计算机应用. 基于*BERT*的中文简历命名实体识别, 41(S01):5.

郭知鑫,邓小龙. 2021. 北京邮电大学学报. 基于*BERT-BiLSTM-CRF*的法律案件实体智能识别方法, 44(4):129.

李想,魏小红,贾璐. 2017. 农业机械学报. 基于条件随机场的农作物病虫害及农药命名实体识别, 48(S1):178–185.

刘新亮,张梦琪,谷情. 2021. 农业机械学报. 基于*BERT-CRF*模型的生鲜蛋供应链命名实体识别, 52(S01):7.

王颖洁,张程烨,白凤波,汪祖民,季长清. 2023. 计算机科学与探索. 中文命名实体识别研究综述, 17(02):324–341.

张剑,吴青,羊昕旖. 2018. 计算机与现代化. 基于条件随机场的农业命名实体识别, 2018(1): 123-126.

郑泳智,吴惠,朱定局. 2021. 计算机与数字工程. 基于荔枝和龙眼病虫害知识图谱的问答系统, 49(12):2618–2622.

基于结构树库的补语位形容词语义分析及搭配库构建*

田思雨 北京语言大学 1416092064@qq.com

邵田 北京语言大学 shaotian2017@163.com

荀恩东† 北京语言大学 edxun@126.com

饶高琦‡ 北京语言大学 raogaoqi@blcu.edu.cn

摘要

在形容词充当补语的粘合式述补结构[1]中，通常以两个谓词性成分连用（"形容词+形容词"、"动词+形容词"）的形式出现，由于这一结构没有形式标记，为计算机自动识别该结构带来了较大的难度，同时，形容词充当补语并不是其最基本、典型（作定语、谓语）的用法，在语言学界与计算语言学界也没有受到足够的关注。因此，该文以补语位的形容词为研究对象，从大规模的句法结构树库中抽取形容词直接作补语的述补结构，并通过编程和人工校验的方式对语料进行降噪，对补语位形容词进行穷尽式检索，得到补语位形容词词表，进一步对补语位形容词的语义进行细分类，构建相应的语义搭配库。不仅可以提升句法切分的正确率，为深层句法语义分析提供语义信息，也可以为语言学本体的相关研究提供参考。

关键词： 补语位形容词；语义分类；语义搭配库

Semantic analysis of complementary adjectives and construction of collocation database based on structural tree library

Siyu Tian Beijing Language Culture University 1416092064@qq.com

Tian Shao Beijing Language Culture University shaotian2017@163.com

Endong Xun Beijing Language Culture University edxun@126.com

Gaoqi Rao Beijing Language Culture University raogaoqi@blcu.edu.cn

Abstract

The adjective-complement bonded declarative-complement structure, which usually appears in the form of two predicative components in succession ("adjective + adjective", "verb + adjective"), poses a greater difficulty for automatic computer recognition of this structure because of its lack of formal markers, and because the adjective-complement is not its most basic and typical (as a definite article and predicate) usage, and has not received sufficient attention in linguistic and computational It has not received enough attention in the linguistic and computational communities. Therefore, this paper takes adjectives in the complement position as the object of study, extracts the descriptive-complement structure of adjectives as complements directly from a large-scale syntactic structure treebank, programmatically and manually verifies the corpus for noise reduction, performs exhaustive retrieval of adjectives in the complement position, obtains a

* 本课题为北京语言大学研究生创新基金资助项目（中央高校基本科研业务费专项资金）（23YCX135）与国家自然科学基金"中文意合图的表征与生成方法研究（62076038）"

† 通讯作者:荀恩东(edxun@126.com)

‡ 通讯作者:饶高琦(raogaoqi@blcu.edu.cn)

[1]指不带形式标记"得、个、不"的述补结构

lexical list of adjectives in the complement position, further subcategorizes the semantics of adjectives in the complement position, and constructs a corresponding semantic collocation database. It can not only improve the correct rate of syntactic cut and provide semantic information for deep syntactic semantic analysis, but also provide reference for the related research on linguistic ontology.

Keywords: Complement adjective , Semantic classification , Semantic collocation library

1 引言

补语位形容词是指在述补结构中充当补语的形容词，包括粘合式述补结构中“动词/形容词（谓语）+形容词（补语）”和组合式述补结构中“动词/形容词（谓语）+得+形容词（补语）”两类。前者在语法上可能构成主谓、动宾、述补等结构，语义上可能表示支配、修饰等关系，句法语义关系较为复杂，在句法语义分析时较难准确识别。由于后者具有明显的形式标记，计算机在识别时没有较大的难度，因此不在本文的研究范围之内。例如：

（1）这台机器**操作简单**。
（2）祖国江山**显得壮丽**。
（3）有些问题一定要**考虑清楚**。
（4）再多钱也会被**挥霍干净**。

在例（1）-例（4）中，加粗词语的词性序列均为“动词+形容词”，但是其句法与语义关系却不相同。例（1）中的“这台机器操作简单”是主谓谓语句，其中“操作简单”为形容词作谓语的主谓结构是句子的谓语结构，“操作”为动词，“简单”为形容词，词性序列与述补结构相似，但该搭配为形容词作谓语的主谓结构。在例（2）中，“显得壮丽”与述补结构的词性序列也相同，但该搭配为形容词作宾语的动宾结构。在例（3）中，“考虑清楚”为述补结构，“清楚”表预期的结果。而例（4）中的述补结构“挥霍干净”，表示一种非理想结果。

因此，本文将研究对象定位于形容词作补语的粘合式述补结构，从句法结构树库中抽取“中心语（形容词/动词）+形容词”结构，并通过编程、人工等方式对语料进行消岐，获取补语位形容词词表及其搭配语料。进而以消岐后的语料为基础，对补语位的形容词进行两种语义划分：其一为结合形容词和情感词的分类体系，对补语位形容词进行细致的语义分类；其二为根据形容词补语在述补结构中的语义，对形容词补语进行语义分类，并进一步构建补语位形容词语义搭配库。形容词本身的语义分类能够为相关研究提供细致的近义词类聚，比如可以为深度学习等提供语义支持，也可以为信息检索、情感分析、关系抽取、机器翻译等提供语义聚类。形容词补语的语义分类能够更好地观察形容词在述补结构中的作用及分布状况，从而更好地识别整个述补结构组块的语义及感情色彩。此外，语义搭配库的构建也可以提高句法语义分析中在识别述补结构时的准确率，并提供更丰富的语义信息。

2 研究现状

关于补语位形容词，语言学界暂无专门研究，但零散出现于对述补结构、形容词的众多研究中。早期语言学界并不认为形容词能够充当“补语”，如黎锦熙（1924）等。二十世纪五十年代开始，逐步明确了对核心谓词后的形容词成分的界定，确认了形容词可以直接跟在动词或形容词后作补语，如王力（1943）、吕叔湘、朱德熙（1952）、丁声树等（1961：11）等，朱德熙（1982）进一步根据有无“得”字，把前一类述补结构叫作“组合式述补结构”，后一类叫作“粘合式述补结构”。八十年代以来，对于述补结构的专门研究较多，且偏向于语义平面的研究，王还（1984）认为形容词作补语多描述既成事实；马庆株（1986）对程度补语进行三级分类；陆俭明（1990、2001）详细探讨了“VA了”的语法意义，并指出其表达语法意义的三种情况；马真、陆俭明（1997）详尽地列举了作结果补语的形容词，并说明了其语义的分布情况。进入二十一世纪以后，对于补语位形容词的研究更加细致、全面，比如李锦姬（2003）详细说明了结果补语、程度补语的语法意义及用法；刘振平（2007）对单音节形容词构成动结式述补结构

的情况进行了穷尽式考察；朱文文（2008）根据形容词的语义特征对形容词进行二级分类，进而总结了形容词作补语的句位意义；蔡丽（2011）总结出程度补语的五个特点，并从六种角度对程度补语进行划分；刘从文（2012）根据语义将形容词分为三类，考察了其作补语的入位情况，并对其句法、语义、语用及认知进行分析。但以上学者均针对某一类补语位的形容词或状补位对比进行深入研究，即无大规模语料库支撑的实证研究，也未对补语位形容词的全貌进行研究。

搭配这一概念自提出以来，受到了国内外学者的广泛关注。Firth（1957）首次提出搭配的概念，将其引入语言学研究的领域，认为搭配整体是表达一种意义的方式。Sinclair（1966）对搭配进行了更深入地研究，指出搭配是两个及以上的词在句子中共现。国内学者早期对搭配现象的探索主要集中于对词语搭配性质的界定，如刑公碗（1978）、林杏光（1990、1991）等。随着语料库语言学的兴起，语料库搭配研究受到了语言习得、语言分析等领域学者的极大关注。在搭配知识抽取方面，提出了基于词共现和分布值的统计方法和基于框架和规则的方法，比如，孙茂松等（1997）提出了以搭配强度、离散度、尖峰为指标的搭配判断算法；卫乃兴（2002）提出了基于数据的方法和数据驱动的方法；Kilgarriff等(2004)通过词性标签构建搭配规则建立了Word Sketch Engine系统。在搭配库建设方面，学界从语法结构、二语习得等角度构建了不同的搭配库，蔡丽（2017）基于留学生的使用情况和偏误分析，构建了程度补语与述语常用搭配库；邢丹等（2020）基于BCC语料库构建了介词结构搭配库；王雨（2022）基于《国际中文教育中文水平等级标准》构建了国际中文教育词语搭配知识库。以上搭配知识抽取方法以及搭配库构建成果，对本文的搭配知识抽取以及搭配库应用具有较大的借鉴意义。

综上，理论语言学对于形容词作补语的语言现象进行了句法、语义的深度探讨，探究了形容词作补语的句位意义，考察了形容词在述补结构中的语义表达，对本文的搭配知识构建、语义分类标准具有重要的参考价值。但同时对补语位形容词，尤其是形容词作补语的粘合式述补结构也存在缺乏专门研究、缺乏数据库支持的不足。此外，计算语言学领域暂无针对无形式标记的形容词作补语的知识建设工作，但对于搭配知识抽取的方法以及搭配库建设的研究成果众多，对本文的构建工作具有重要的借鉴意义。

3 数据基础

本文的语料来源是北京语言大学句法结构树库（卢露等，2020），针对形容词充当补语的粘合式述补结构进行语料的抽取，并对语料进行消岐。

3.1 树库资源

北京语言大学句法结构树库规模大、领域多、质量高。目前已有浅层结构分析树库总字数约1300万字，小句70余万句，语料涉及报刊、文学、百科、新闻、综合等领域文本1万余。该句法树库将句子分析为块状组合序列，由构成句子基本结构的句法成分、起衔接作用的衔接成分以及表附加性语义的辅助成分构成。其特点为以组块状短语结构树为句法表示，直接根据各组块的性质及功能，标注句子骨架，突出中心词信息。

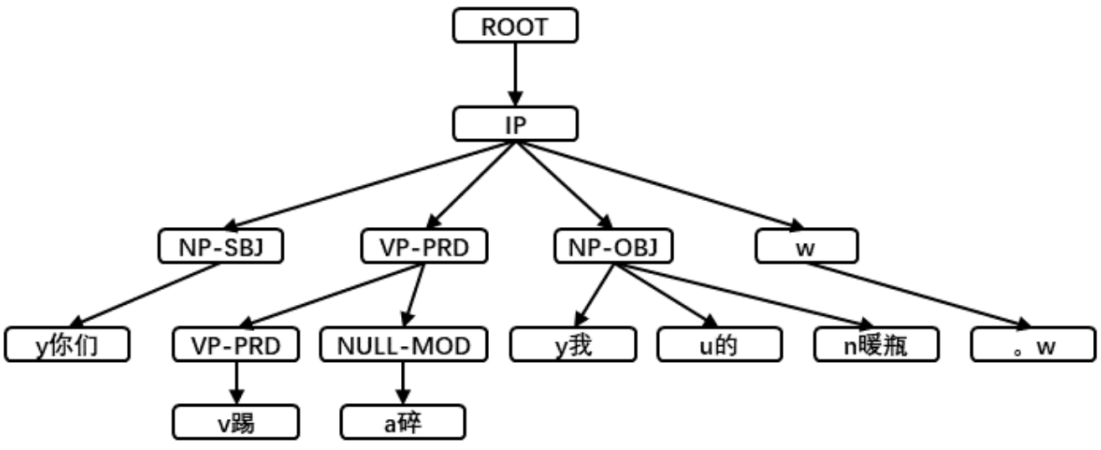


Figure 1: “你们踢碎了我的暖瓶”结构图

该树库组块句法属性标记包含组块的词性信息、功能信息以及边界信息等（见图1）。树库在对句子进行浅层结构标注时，以“谓词-论元”结构为中心的句子骨架，标注最大长度的主语块、宾语块、状语块、补语块以及主谓宾结构和状中补结构，能够准确地识别组块功能、边界及关系。

3.2 补语位形容词的抽取与消歧

本小节主要介绍从北京语言大学句法结构树库中抽取补语位形容词语料并对语料进行消歧的过程。

3.2.1 抽取过程

本文所使用的北京语言大学句法结构树库是以组块状短语结构树为句法表示，经过分词词性标注的高质量树库。因此，可以利用树库的结构信息和词性信息，结合补语位形容词构成的述补结构的句法形式构建检索式。一方面，由于直接检索“中心语（动词、形容词）+形容词”存在较多述宾结构的歧义；另一方面，在修饰性组块内部进一步限制词性信息对检索精度影响较大。因此，在进行语料的抽取时，首先将述语组块的中心语（动词、形容词）与[NULL-MOD](修饰性组块)'紧连于述语块内；其次，为了考察补语位形容词的音节形式特征，再按照音节数目构建检索式：单音节动词+修饰性组块、双音节动词+修饰性组块、单音节形容词+修饰性组块、单音节形容词+修饰性组块。

同时在检索层面进行消歧。由于检索式所抽取的述补结构，未限制后部修饰性成分的词性信息，因此在检索层面可以利用词性等信息排除其他类型的补语。不是形容词充当的补语类型有趋向补语、数量补语、时处补语。参考黄伯荣、廖序东（1991）的《现代汉语》以及实际语料，整理出需要批量排除的形式特征，如表1 所示：

修饰性组块	形式特征
介宾结构	于、中、入、里、内、在、至、到、在、为、给
趋向补语	来、去、上、下、进、出、回、起、起来
数量补语	0、1、2、3、4、5、6、7、8、9、一、二、三、四、五、六、七、八、九、十
时处补语	0、1、2、3、4、5、6、7、8、9、一、二、三、四、五、六、七、八、九、十、两、半
其他	过、着、有、了、开、住、成、是、而

Table 1: 后部修饰组块降噪

3.2.2 数据处理

经过检索层面降噪后的语料基本为形容词充当补语的述补结构。由于述补结构的形式特征有带形式标记（组合式述补结构）和不带形式标记（粘合式述补结构）两种。本文通过编程的方法，去除带有形式标记的述补结构之后，对抽取到的形容词充当补语的述补结构进行形式分类。将补语位形容词按照音节划分，与前部中心语相互组合。最后得到的分类结构一共有8种:v1+a1;v1+a2;v2+a1;v2+a2;a1+a1;a1+a2;a2+a1;a2+a2 [1]

3.2.3 语料消歧

由于分词词性标注与组块标注错误的影响，在检索层面进行消歧之后，所抽取的结果中仍包含一些不属于形容词作补语的述补结构。因此，本文利用词典与人工两种方式对检索结果进行进一步消歧。

首先，校对词性信息，进行词典消歧。在本文抽取的形容词作补语的述补结构中，中心语包含动词、形容词，补语为形容词。分别将这些词与《现代汉语词典》（第七版）中的形容词、动词进行对照，核对语料中的词性信息。具体分为两步：（1）校对中心语（动词、形容词）和补语的词性是否正确，如出现不为词或词性错误的情况，则删除该条语料。（2）校对述补结构是否正确，如所抽取粘合式补语在词典中为一个词，则删除该检索结果。

其次，核对述补结构，进行人工排查。由于字典中词语的义项较多，通过词典消歧后仍可能出现歧义的结构。因此，需要进一步通过人工判断述补结构的准确性。分为四种情况：（1）

[1]注：v1为单音节动词、v2为双音节动词、a1为单音节形容词、a2为双音节形容词

述补结构中，形式标记前后的词性符合要求，但实际上并不是述补结构，比如“显得富丽”，其中“显得”为动词，该结构为动宾结构。但由于“显”和“富丽”符合动词、形容词词性，仍保留在检索结果中。（2）所抽取的结果中频次较少的述补结构，其词性信息符合条件，但由于树库所涉及的语料涉及微博、对话等生活文本，语料中可能出现错误用法，比如“行走得性急”。（3）一些词出现了新用法，但词典暂未收录的该义项。（4）动词、形容词中存在许多兼类词，词典消岐不能识别出检索结果中的具体用法，造成错误检索与错误词性的漏排。以上四种情况均需要人工进行排查。本文以《现代汉语词典》（第七版）中的形容词、动词义项为标准，参考《汉语形容词用法词典》，依据上下文信息，对抽取语料进行逐个对照排查，删去歧义语料。

经过以上消岐步骤后，一共得到述补搭配14436个，补语位形容词词表包含657个形容词，其中单音节218个，双音节439个，大大超过了前人以往研究成果中的数量。

4 补语位形容词的语义分类

词在未进入语法结构之前就存在静态的、共性的意义，词汇意义是句法表现的基础，决定着句法表现的意义。本节参考现有研究成果，结合补语位形容词在述补结构中的词汇意义，对前文得到的补语位形容词词表进行语义细分，并对分类结果进行分析。

语义一级类（数量）	语义二级类（数量）	例词	例句
空间(112)	尺寸(10)	小	帽子给你做小了
	外观(43)	碎	你们踢碎了我的暖瓶
	光彩(30)	绿	他的脸都吓绿了
	方位(6)	高	最后一瞬间她把枪口抬高了
	距离(23)	远	两人牵着手跑远了
时间(13)	相对(4)	早	哥哥今天起床起早了
	历时(2)	久	年糕被我煮太久了
	速度(5)	慢	小徐不是因为报表做慢了被炒的
	年龄(2)	大	父母这么辛苦把我拉扯大
形态(74)	样态(67)	软	我腿都吓软了
	声音(7)	高	我抓起遥控器，把电视的音量调高
状态(146)	正面(60)	均匀	你要保证孔内涂均匀
	负面(64)	累	今天跑累了
	中性(22)	紧	用软木塞把它盖紧
情绪(20)	正面(9)	爽	今天把我吃爽了
	负面(11)	烦	家里人都听烦了
评价(210)	正面(151)	清楚	有些问题一定要考虑清楚
	负面(56)	错	我晕到连家门都走错
	中性(3)	杂	书看太杂了不好
性格(44)	正面(31)	仔细	她对事物观察仔细
	负面(13)	圆滑	要把事情做圆滑
频数(13)	频次(2)	频繁	你平时要练勤一些
	数量(11)	齐全	考试的东西备齐全
程度(32)	轻(4)	轻	学校对这事处理轻了
	重(28)	死	我真是委屈死了

Table 2: 一级类、二级类分类结果

4.1 分类标准与体系

分类是语言研究的基本方法，也是大多语义研究的基础性工作。对于形容词的语义分类，各学者的观点不尽相同。黎锦熙（1924）将形容词分为性质、形体、状态和程度形容词。刘月华等（1982）根据描述对象、感情色彩，将形容词分为描写动作者、描写动作、正向形容词、负向形容词。王惠等（2006）将形容词划分为事性值、物性值（量化属性值、模糊属性值、颜色）、人性值（年龄、品格、关系、境况）、空间值（一维值、二维值、三维值）、时间值。

鲁川（2010）提出了“属性值”和“属性域”，将具有属性值的词划分为现象、性质、境况、数量四个大类，每个大类下划分众多次类，如现象大类下分“空间、时间、气候、感觉四个小类。后两者对形容词划分较为细致、层次分明，因此本文将借鉴二人的分类结果对补语位的形容词进行语义分类。

此外，补语位形容词中包含大量的情感词，因此本文也参考了情感词的研究成果。台湾大学NTUSD的情感词典将情感词划分为积极、消极两类。李军（2008）也将情感词划分为积极、消极。HowNet情感词典将情感词分为正面评价、正面情感、负面评价、负面情感、主张、程度级别六类。徐琳宏等（2008）在国外比较有影响的Ekman的6大类情感分类体系的基础上进行二次分类，共分为7 大类（乐、好、怒、哀、惧、恶、惊）21 小类。本文参考前人的分类系统，以结构树库抽取的语料为数据基础，对补语位形容词的语义分类结果进行完善。

参考前人对形容词以及情感词的分类方法和结果，对上文得到的659个补语位形容词进行人工的语义分类。在分类时将意义相近、色彩相似的词合并，尽量得到较为科学且便于应用的分类结果。

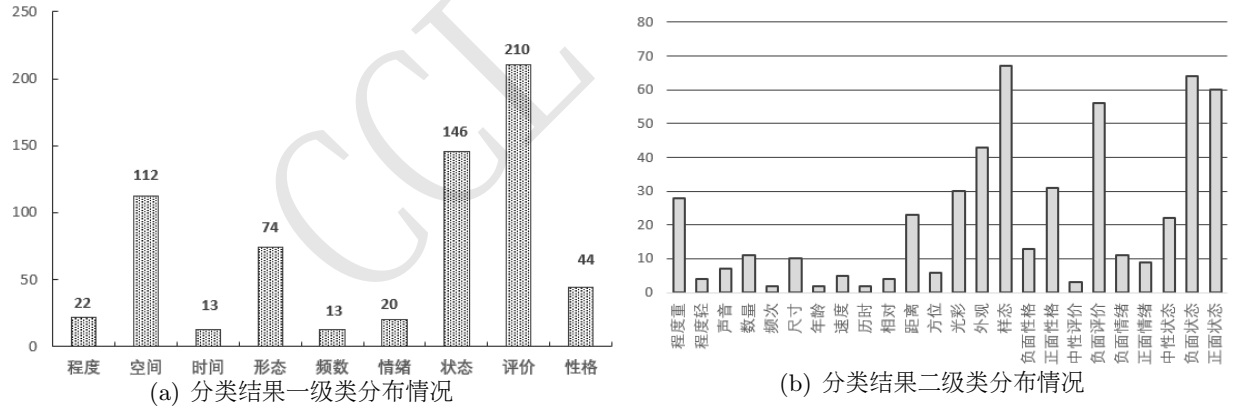
4.2 分类结果

经过多轮迭代后，得到补语位形容词三级语义分类体系，包括9个一级类，26个二级类，91个三级类。分类结果如表2所示，由于三级类较多，表中只列举一级类、二级类以及具有代表性的三级类。

该语义分类结果较为细致，层级分明，不仅可以为深层句法语义分析提供更多、更细致的信息，也可以为语言学研究提供细致的数据支持。

4.3 结果分析

根据以上对补语位形容词语义的三级划分结果，进一步对其数据进行统计分析。首先，通过对一级类数据进行统计（如图2(a)所示）。可以看出形容词作补语时，具有评价类意义的形容词更容易进入述补结构，评价类形容词共有210个，占比达32%，这与形容词中包含大量情感词有一定的关联。而表状态义的补语位形容词的数量也不在少数，占比达22%。说明人们倾向用形容词作补语来表达事物状态。此外，表程度的补语位形容词数量较少。说明形容词作补语的述补结构较少能进入表程度的形容词。



(a) 分类结果一级类分布情况　　(b) 分类结果二级类分布情况

Figure 2: 分类结果分布情况

其次，通过对二级类补语位形容词数据进行统计分析（如图1(b)），我们可以明显地观察到表正面意义的评价类形容词作补语的频次较多，共有151个，占比23%。此外，负面状态类（10%）、正面状态类（9%）、负面评价类（8%）以及样态类（10%）四类补语位形容词的频次相近，且占较大比例。

综上，可以得出补语位形容词语义主要分布在评价类语义中，且在感情色彩上，正面色彩的形容词要远高于负面色彩，表明在形容词作补语的用法中，人们更倾向于用补语来表达正面感情。其次，对空间、时间、样态等具有一定客观性的语义类进行细致划分，能够使语义分析更加准确、细致。

第二十二届中国计算语言学大会论文集，第420页-第432页，哈尔滨，中国，2023年8月3日至5日。

5 形容词补语的语义分类

考察述补结构中形容词补语的语义，不仅可以更好地识别形容词所在述补结构的意义，也可以更好地反映出不同意义形容词在述补结构中所起的作用，为机器理解补语位形容词语义提供参考。本节参考现有研究成果，对前文得到的形容词作补语的述补结构搭配进行考察，细分补语位形容词在搭配中的语义，并对其结果进行分析。

5.1 分类标准及准则

在整个述补结构中形容词所表达的语义要从述补结构的语义分类中获得。从一般的补语分类中，粘合式补语分布在趋向补语、程度补语、结果补语、数量补语、时处补语中。由于趋向补语、数量补语、时处补语中的补语指向较为明确，且在对语料考察中皆存在形容词作程度补语和结果补语的情况。因此，本文将研究对象锁定为程度补语和结果补语中形容词的语法意义分类。马真、陆俭明（1997）将形容词作结果补语的语义概括为四种，预期结果的实现、非理想结果的出现、自然结果的出现、预期结果的偏离，并提出表不同语义的决定因素是形容词的性质、动词的性质、述语动词所表示的行为动作对作补语的形容词所表示的性质的制约作用、语境。其所提出的分类结果得到众多学者认可，本文也将参考其分类。

李锦姬（2003）研究了现代汉语全部补语，在对整体性述补结构考察中将补语分为有量补语和无量补语，并进一步在无量补语之下细分语义为自然结果、预期结果、非理想结果；并阐述了有量补语主要表达心理预期结果的偏离量。该研究和马真、陆俭明（1997）存在很多相合之处，且有更详细地分类及说明。此外，该文认为程度补语为非典型性述补结构中表程度的量，并对其进行详尽阐述。因此，本文着重参考其分类成果。

以上文得到的14436个搭配为基础对其进行人工的语义分类。参考前人对形容词作补语述补结构的语法意义分类，以真实搭配中体现的语义为基础，进行人工标注。

5.2 分类结果

经过多轮迭代后，本文将程度补语非为2类，结果补语先分有量补语和无量补语2类，再分5类。分类结果如表3所示。此外，本文还对语义三级类中形容词本义的分布情况进行了考察，由于数量较多，放入附录中展示。

语义一级类（数量）	语义二级类（数量）	语义三级类（数量）	例句
结果补语(11922)	有量补语(1563)	心理预期结果的偏差-指向名词(521)	气球吹大了不好卖
		心理预期结果的偏差-指向动词(1042)	演唱会结束迟了
	无量补语(10359)	自然结果	照片存放久了不免泛黄
		预期结果(7095)	有些问题一定要考虑清楚
		非理想结果(1864)	他把这件事解释偏了
程度补语(32)	轻(4)	程度轻(4)	这个故事还是写轻了
	重(28)	程度重(28)	他真是傻透顶了

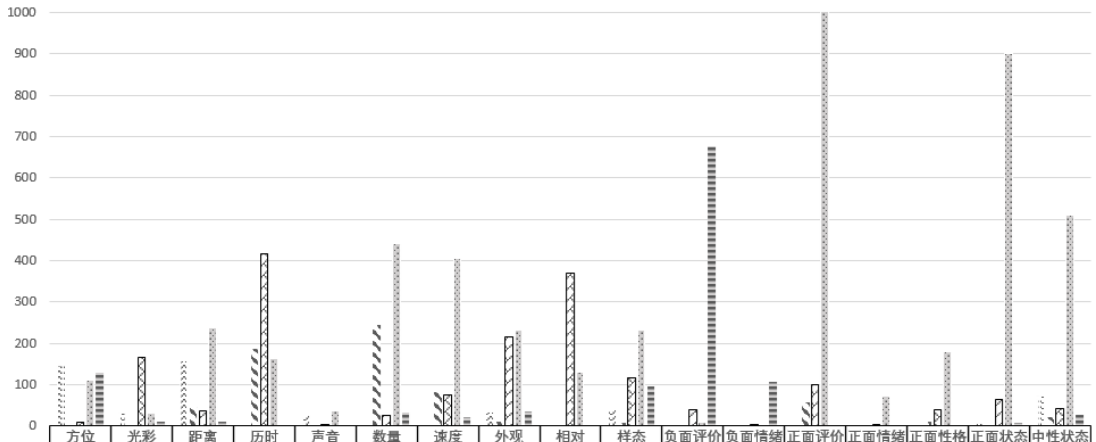
Table 3: 补语位形容词在述补结构中的语义分类结果

5.3 结果分析

根据以上对补语位形容词语义的三级划分，可以对其数据进行进一步统计分析。首先，通过对一类数据进行统计，可以看出形容词作结果补语的搭配远多于程度补语，形容词作补语时多为结果补语。其次，对二级类进行统计，结果补语中有量补语为1563，无量补语为10359。可以得出形容词作结果补语时多分布在无量补语，表达动作或变化引起的结果，较少表达心理预期结果的偏离量。

我们对形容词本义中能够表达五种语义的情况进行统计，但由于部分形容词本义数量较少，只选取较有代表性的种类（如图3[2] ）。从图中我们可以看出具有褒贬意义的形容词（即正面评价、负面评价、正面状态、负面状态等）数量较多。即具有褒贬色彩的形容词更容易进入述补结构。此外，从分布上看，具有褒贬意义的形容词在述补结构中所表达的意义较为聚集，多集中分布在某一种语义上，而其他具有客观意义的形容词的分布则较为分散。因此可以推出

[2]图3中为了更好地观察数据的对比结果，将预期结果中正面评价类数据3338排除在外



	方位	光彩	距离	历时	声音	数量	速度	外观	相对	样态	负面评价	负面情绪	正面评价	正面情绪	正面性格	正面状态	中性状态
心理预期结果的偏差（名）	146	30	157	0	24	0	0	34	0	40	0	0	0	0	0	6	72
心理预期结果的偏差（动）	0	0	46	186	0	245	82	9	0	8	1	0	57	0	10	3	20
自然结果	8	166	35	416	2	25	75	214	369	116	40	4	100	4	40	64	42
预期结果	110	30	237	164	35	442	406	233	129	231	9	1	3338	73	180	902	511
非理想结果	129	9	10	1	2	32	22	35	0	96	677	106	2	0	0	7	26

Figure 3: 形容词所在述补结构的语义的分布情况

具有褒贬义的形容词在述补结构中有其固定的表达倾向。述补结构的整体意义受褒贬义形容词的影响较大。而表客观意义的形容词在述补结构中对整体的影响较小。

综上，我们可以得出补语位形容词主要分布在结果补语中，且无量补语能够容纳更多种类、数量的形容词，而有量补语的选择则较为受限。其次，在形容词作补语的述补结构中，人们更倾向于表达预期结果，且正面评价类形容词出现率极高。此外，具有褒贬意义的形容词对述补结构的整体意义影响较大。

6 补语位形容词搭配库构建及分析

本节以补语位形容词的三级语义分类体系为基础，根据消岐后的真实语料搭配情况，构建补语位形容词的语义搭配库，并对其数据进行进一步统计分析。

6.1 构建搭配库

由于词的意义较为灵活，而搭配的意义较为固定。补语位形容词在不同的搭配情境中，其表达的语义可能不同，比如：“理解透了”和“失望透了”，“透”在前一搭配中表示“深入、透彻”，而在后一搭配中则表程度深。因此，我们希望构建语义搭配库来将语义分类更好地对接句法语义分析。本小节通过消岐后的搭配语料构建，包含述补搭配的语法信息，同时按照语义分类体系，为其添加语义信息，具体为：以中心语（动词、形容词）为中心，根据中心语和补语位形容词搭配的频次为顺序，并在补语位形容词后添加其语义类型信息。

6.2 搭配库数据分析

根据搭配库中中心语和补语位形容词的音节搭配信息，我们对其音节搭配情况进行统计。本文对粘合式述补结构进行抽取时，划分了8种音节组合情况，经过消岐后，形容词作中心语带形容词补语的语料较少，“A1+A1”、“A2+A2”的音节组合无合格语料，共得到14436个搭配。因此本文只讨论6种音节组合情况，如下表6所示。此外，我们进一步对单音节形容词和双音节形容词的组合情况进行统计，如表7所示。

从两表中可以得到，“V1+A1”格式搭配数量最多，占比35.54%。且“V1/ V2/ A2+A1”格式，占比60.03%。可以看出单音节形容词作补语的能力更强，其次“V2+A2”格式占比较高，为25.87%，体现了汉语韵律语法的音节搭配规律，双音节与双音节搭配具有普遍优势。此外，“A1+A2”、“A2+A1”格式的占比极小，说明形容词作中心语后带补语位动词的情况较少。

6.3 搭配库数据应用

本文所构建的搭配知识库对形容词作补语的粘合式述补结构进行了穷尽式检索，包含了完

音节搭配	数量	占比
V1+A1	5130	35.54%
V1+A2	2012	13.94%
V2+A1	3276	22.69%
V2+A2	3734	25.87%
A1+A2	24	0.17%
A2+A1	260	1.80%
总计	14436	100%

Table 4: 音节搭配数量及占比

音节搭配	数量	占比
V1/ V2/ A2+A1	8666	60.03%
V1/ V2/ A1+A2	5770	39.97%
总计	14436	100%

Table 5: 单双音节补语位形容词搭配对比

整的搭配数据和形容词以及述补结构的语义信息。首先，穷尽的搭配知识能够为机器自动识别无形式标记的形容词补语述补结构提供可靠的数据支撑，提升句法切分准确度。丰富的语义信息能够帮助计算机进行感情色彩的判定，进而更准确地理解自然语言。其次，对于语言学本体研究而言，详细的补语位形容词词表及搭配信息能够帮助我们验证或发现语言规律。形容词语义分类和在述补结构中的语义分类的对比，为进一步考察补语位形容词用法及规律提供了查考数据。最后，该搭配库还能够为对外汉语教学提供搭配知识支持。详细的搭配及语义知识不仅能够为智能教学设计、自动出题提供搭配知识库，也能够为对外汉语教师的教学活动提供知识支持。

7 展望

本文基于大规模语料库构建了补语位形容词词表以及补语位形容词粘合式述补结构搭配库，形容词词表包含659个补语位形容词，搭配库包含14436个搭配。通过对形容词语义以及形容词补语的语义进行细致分类，为搭配库添加了丰富的语义信息。其中形容词语义包括9个一级类，26个二级类，91个三级类，形容词补语的语义类别包括2个一级类、4个二级类、7个三级类。该搭配库数据详尽、语义信息丰富，可用于机器识别、智慧教育、自动出题以及语言研究等方面。但本文研究仍存在可进一步研究之处。首先，本文搭配资源的语义仅考察了中心语与形容词补语，未考虑副词对述补结构语义的影响。其次，由于时间原因，本文构建的搭配库缺乏对频次的考察与统计，而搭配频次也是语言研究和智慧教育的重要参考。此外，本文对于搭配库的实际应用效果没有进行详实地说明。

因此在未来的工作中，我们将从本文所抽取的述补结构中提取存在副词的搭配，进一步考察副词对该搭配语义的影响。其次，针对缺少频次信息的问题，我们将通过大规模语料库，收集各个搭配的频次信息。最后，我们将把信息丰富的搭配知识，放入智慧教学平台中，考察其在应用中的实际效果，为对外汉语教师的教学设计以及智慧教学自动出题提供搭配知识库。

参考文献

黎锦熙. 1924. 新著国语文法. 上海:商务印书馆，.

王力. 1943. 中国现代语法. 北京:中华书局

吕叔湘.朱德熙1952. 语法修辞讲话. 北京:中国青年出版社.

丁声树. 1961. 现代汉语语法讲话. 北京:商务印书馆.

朱德熙. 2008. 语法讲义. 北京:商务印书馆.

王还. 1984. 汉语的状语与“得”后的补语和英语的状语. 语言教学与研究,(04)61-66+29

陆俭明. 1990. 述补结构的复杂性——《现代汉语补语研究资料》序. 语言教学与研究,(01)13-20

陆俭明. 2001. ”*VA*了”述补结构语义分析补议——对读者意见的回复. 汉语学习,(06)78-80.

马庆株. 1986. 含程度补语的述补结构. 北京: 北京大学出版社.

马真,陆俭明. 1997. 形容词作结果补语情况考察*(一)*. 汉语学习, (1):3-7.

马真,陆俭明. 1997. 形容词作结果补语情况考察*(二)*. 汉语学习, (04):14-18.

马真,陆俭明. 1997. 形容词作结果补语情况考察*(三)*. 汉语学习, (06):7-9.

李锦姬. 2003. 现代汉语补语研究. 上海:复旦大学学位论文.

刘振平. 2007. 单音形容词作状语和补语的对比研究. 北京:北京语言大学学位论文.

朱文文. 2008. 现代汉语形容词状补语序选择机制研究. 北京:北京语言大学学位论文.

蔡丽. 2011. 现代汉语中程度补语的范围及类别. 宁夏大学学报(人文社会科学版),(04)9-14+32.

刘从文. 2013. 现代汉语形容词状补位对比研究. 北京:上海师范大学学位论文.

Firth,J.R 1957. *Papers in Linguistics 1934-1951*. London:Oxford University Press.

Sinclair,J 1966. *Beginning the study of lexis*. ln Memory of J.R.Firth,410-430.

邢公畹. 1978. 语词搭配问题是不是语法问题*?*. 安徽师大学报(哲学社会科学版),(04)77-84+65.

林杏光. 1990. 词义分类、词语搭配、语言教学. 北京语言学院出版社.

林杏光. 1991. 论词义分类和词语搭配. 中国人民大学学报,(05)77-82 .

孙茂松,黄昌宁,方捷. 1997. 汉语搭配定量分析初探. 中国语文,(01)29-38.

卫乃兴. 2002. .现代汉语补语研究. 复旦大学博士论文.

李敏. 2002. 基于语料库和语料库驱动的词语搭配研究. 当代语言学,(02)101-114+157.

Kilgarriff A ,Rychlý, Pavel,Smrž, Pavel,et al. 2004. *The Sketch Engine*. London:Oxford University Press.

蔡丽. 2017. *AEIC Academic Exchange Information Centre(China)*. Proceedings of 2017 2nd International Conference on Humanities Science and Society Development (ICHSSD 2017)（Advances in Social Science, Education and Humanities Research VOL.155）Atlantis Press.

邢丹. 2020. 基于大规模语料库状中搭配库构建与应用. 北京语言大学学位论文.

王雨,肖叶,荀恩东等. 2022. 服务国际中文教育的词语搭配知识库建设. 语言文字应用,(02)26-37.

卢露,矫红岩,李梦,荀恩东. 2020. 基于篇章的汉语句法结构树库. 自动化学报,(01)2911-2921.

中国社会科学院语言研究所词典编辑室. 2016. 现代汉语词典（第七版）. 北京:商务印书馆.

郑怀德,孟庆海. 2016. 汉语形容词用法词典. 北京:商务印书馆.

黄伯荣,廖旭东. 1991. 现代汉语. 北京:商务印书馆.

刘月华. 1982. 状语与补语的比较. 语言教学与研究,(01)22-37.

王惠,詹卫东,俞士汶. 2006. ”现代汉语语义词典”的结构及应用. 语言文字应用,(1)134-141.

鲁川. 2010. 知识工程语言学. 北京:清华大学出版社.

李军. 2008. 中文评论的褒贬义分类实验研究. 北京:清华大学硕士学位论文.

徐琳宏,林鸿飞,赵晶. 2008. 情感语料库的构建和分析. 中文信息学报,(01)116-122.

附录:

A.1 检索式实例

AddTag(”begs”,”于;中;入;来;去;上;下;进;出;回;开;起;过;着;有;了;过;至;到;成;为;给;是;住;在;里;而;0;1;2;3;4;

AddTag(”ends”,”于;中;入;在;里;下;上;内”)

Condition(”len($1)=1;mid!=[modalverbs];end($1)!=[ends];beg($2)!=[begs];end($2)!=[ends]”)

Handle0=GetAS(”—v”,””,””,””,””,””,”0,1”,””,””,””)

Handle1=GetAS(”$NULL-MOD”,””,””,””,””,””,”0,1”,””,””,””)

Handle2=JoinAS(Handle0,Handle1,”Link”)

Handle3=GetAS(”$VP-PRD”,””,””,””,””,””,””,””,””,””,””)

Handle4=JoinAS(Handle3,Handle2,”SameRight”)

Handle=Freq(Handle4,”$Q”)

Save(Handle,”v1.txt”)

其中，AddTag 的功能为构建一个词表。Condition指的是检索式的一些限制条件。!=表示不等于。$1表示对中心语进行限制，len表示长度.$2表示对修饰性组块进行限制，beg表示组块开头，end表示组块结尾。因此第一行的意思就是限制作述补结构中心语的动词为单音节动词，且不以词表中的词为开头或结尾。

Handle表示一个句柄，以序号不同标明不同的原子查询项。GetAS表示对一个原子查询项进行限制，v表示限制该原子查询项的词性为动词，后面的数字表示的是对该原子项的位置的描述，$NULL-MOD限制该语块为修饰组块。Handle0与Handle1表示得到了词性为动词的原子查询项和标签为$NULL-MOD的原子查询项。

JoinAS表示两个原子查询项在组装时的顺序和关系类型。Link表示两个原子查询项在组装时的关系是接连出现的，即中间没有其他成分。SameRight表示两个原子查询项之间的关系为右对齐。因此Handle2表示的是动词和修饰性组块接连出现的情况；，Handle4表示的是动词和修饰性组块在一个述语块内出现，且二者之间是右对齐的情况。

Freq表示输出检索式的频次信息。

A.2 补语位形容词分类结果

1.空间

(1)尺寸

大、小、长、短、薄、厚、矮、高大、庞大、厚实

(2)外观

烂、碎、细、粗、瘦、丰满、直、弯、圆、扁、平、斜、陡、方、立体、稀烂、烂糊、粉碎、细碎、微细、纤细、粗大、粗壮、窈窕、瘦小、丰腴、胖、臃肿、肥、平直、僵直、弯曲、鼓、尖、凹、滚圆、瘪、干瘪、平坦、平实、平整、歪斜、陡峭

(3)光彩

红、白、绿、黑、黄、青、紫、蓝、乌、灰、亮、暗、通红、白皙、苍白、惨白、黑暗、黝黑、金黄、焦黄、枯黄、铁青、明亮、明朗、亮堂、透亮、暗淡、灰暗、昏暗、浅

(4)方位

高、低、正、偏、歪、反

(5)距离

远、近、深、浅、紧、宽、松、窄、稀、稀疏、稀少、稀松、稀薄、深邃、紧凑、宽敞、宽松、宽广、宽大、宽阔、广阔、松弛、蓬松、狭窄

2.时间

(1)相对

晚、迟、早、紧

(2)历时

久、长

(3)速度

慢、缓慢、迟缓、迅速、飞快

(3)年龄

老、大

3.形态

(1)样态

生、熟、软、硬、轻、沉、老、臭、净、湿、滑、脏、嫩、热、凉、暖、酥、咸、苦、甜、辣、涩、油、黏、鲜、胀、粗糙、干、浑、透明、浓、淡、烂熟、松软、软和、绵软、坚硬、僵硬、硬实、轻盈、轻薄、轻快、清澈、潮湿、平滑、光滑、圆滑、细嫩、白嫩、柔嫩、热乎、烫、滚热、冷、凉快、清凉、凉爽、温暖、暖和、脆、酥脆、苦涩、黏糊、新鲜、毛糙、干爽、干燥、干涩、浑浊、稠、清淡

(2)声音

高、低、响、响亮、洪亮、大、小

4.状态

(1)正面

[1]牢固：牢、结实、坚实、坚固[2]均匀：均、匀、匀净[3]舒适：舒坦、舒畅、舒服、清闲、安心、安稳、幸福、香甜[4]平和：平静、从容、温和、温顺、安静[5]柔和：柔顺、柔软[6]流利：顺、顺溜、顺口、顺嘴、顺当、溜、通[7]真实：真切、确实、现实、具体[8]滋润：红润、圆润、健康[9]扎实：实、饱、满、饱满、瓷实[10]清醒：精神、清白、清新[11]齐：整齐

(2)负面

[1]累：乏、疲劳、疲倦、疲惫、困、笨重[2]身体状态：疼、痛、晕、饿、秃、哑、沙哑、聋、僵[3]冷淡：冷漠、麻、木、麻木、沉默、消极、低迷、冷清[4]呆：楞、迷糊[5]凶猛：凶、凶狠[6]乱：混乱、花[7]破：蔫、干枯、枯萎、馊、黄、酸、枯、旧、焦、废、钝[8]疯：野、痴[9]繁忙：仓促、匆忙、忙碌[10]惨重：惨烈[11]红眼：眼红、眼馋、馋

(3)中性

[1]密集：密、紧、密实、浓密、缜密[2]严：严肃、严格、严密、严厉[3]分散：零散、零碎[4] 完：好、干净

5.情绪

(1)正面

[1]爽：爽快、痛快[2]乐：兴奋、愉快、高兴、快乐、快活

(2)负面

[1]心切：急、急切[2]紧张：慌[3]暴躁：烦、倦[4]郁闷：沉闷、闷

6.评价

(1) 正面

[1]优秀：好、突出、棒、牛、佳、绝[2]周全：周密、周到、齐备、圆[3]完满：圆满、完美、完善、成熟[4]妥当：妥、妥善、稳妥、妥帖[5]对：正确、合理、合法、标准、规整、规范、值[6]精细：细、精、细腻、详尽、详细、细致、精致[7]强：壮、强劲、强大、强壮、强健、壮大、壮实[8]稳：稳固、稳当、稳定[9]美：漂亮、好看、美丽、美艳、靓丽、俊秀、靓、酷、甜、乖[10]清楚：明显、明白、清晰、明确、明晰、明朗、白、直、亮[11]准：准确、精确[12]充足：充分、足[13]火：红、兴旺、红火、旺盛[14]生动：活、鲜活、精彩、逼真[15]顺利：通畅、通顺、顺手、顺畅、畅通[16]容易：简单、轻松、方便、通俗[17]积极：活跃[18]亲密：紧密、密切、亲密、亲切、熟[19]灵活：灵敏、敏锐、利索、巧妙[20]深刻：透彻、通透[21]平衡：平等、匀称[22]整洁：洁净、清洁、白净、干净、工整、端正[23]舒心：顺眼、顺耳、对路[24]协调：和谐、融洽、和睦[25]华丽：高级、高贵、亮丽、隆重、优雅[26]富裕：富有、富[27]熟练：娴熟、专业、高效、熟[28]犀利：尖锐、锋利

(2)负面

[1]坏：烂、孬、糟、糟糕、贱、难听、难看、讨厌、黑[2]刁：拧、贼、抠、骄傲、可怕、残忍[3]虚：空、模糊、渺茫、玄[4]腻：腻歪、腻味、烦、滥[5]弱：脆弱、衰弱、软弱[6]错：异常、过火[7]枯燥：无聊[8]庸俗：糙、俗气、俗、平庸、平淡、丑[9]呆傻：糊涂、愚蠢、愚钝、笨、痴呆、呆板、傻、迟钝[10]艰难：复杂、麻烦、难、困难、穷、苦、危险

(3)中性

[1]贵[2]杂[3]便宜

7.性格

(1)正面

[1]勤快[2]活泼：幽默[3]温柔：随和[4]勇敢：坚强[5]认真：仔细、谨慎、严谨[6]大方：坦白、慷慨[7]聪明：精明、机灵、机敏[8]爽快：干脆、果断[9]谦虚：虚心、低调[10]沉稳：稳

重、可靠[11]正经：客气、正直

(2)负面

[1]懦弱[2]猥琐[3]懒：懒惰[4]狡猾：圆滑[5]粗心：马虎[6]调皮：淘气、幼稚[7]刻薄：偏激、敏感

8.频数

(1)频次

勤、频繁

(2)数量

多、烂、广、少、全、全乎、齐全、完全、完整、丰富、广泛

9.程度

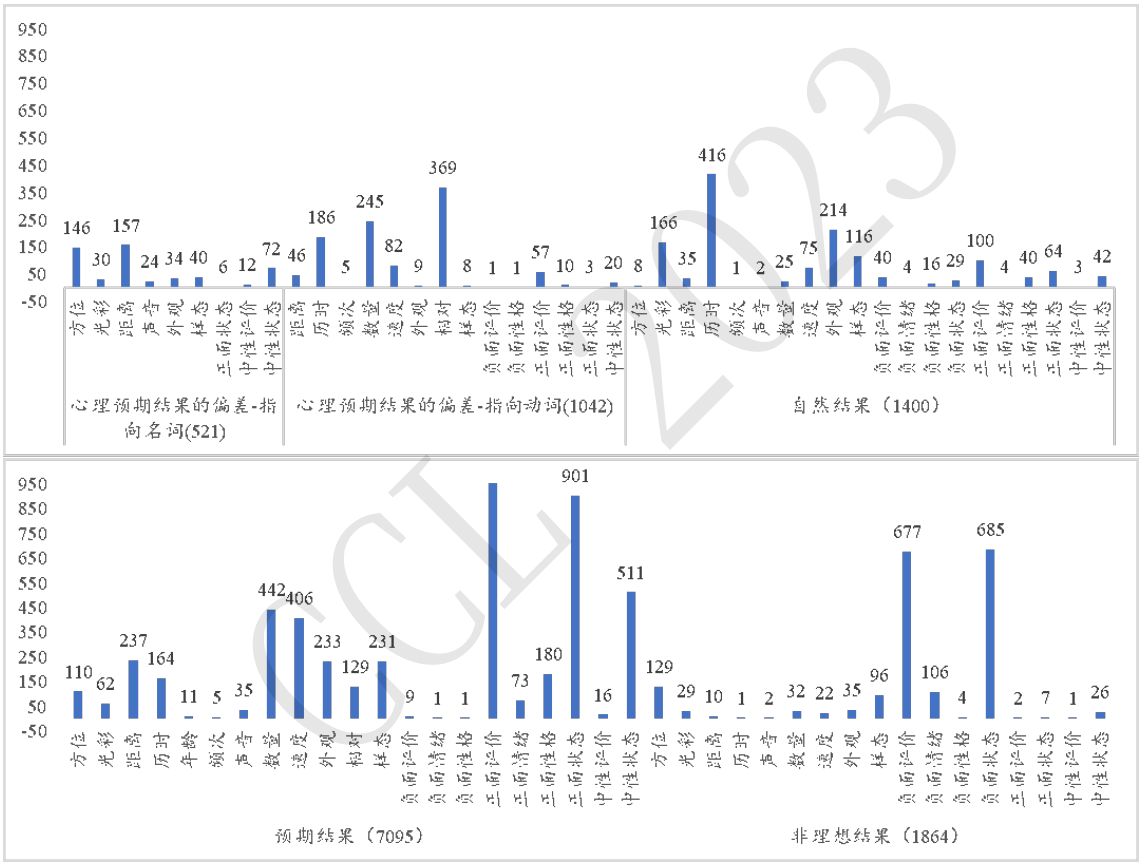
(1)轻

轻、浅、少许、轻微

(2)重

透、死、惨、深、疯、满、重、狠、甚、光、猛、绝、暴、坏、傻、严重、厉害、彻底、激烈、透顶、深厚、沉重、大发、大、远

A.3 形容词补语语义三级类中形容词本义的分布情况



基于BiLSTM聚合模型的汉语框架语义角色识别

曹学飞
山西大学
自动化与软件学院
caoxuefei@sxu.edu.cn

李济洪　王瑞波
山西大学
现代教育技术学院
{lijh; wangruibo}@sxu.edu.cn

牛倩
山西大学
自动化与软件学院
niuqian@sxu.edu.cn

摘要

目前，基于神经网络的汉语框架语义角色识别模型的性能依然较低，考虑到神经网络模型的性能受到超参数的影响，本文将超参数调优和模型预测性能的提升统一到基于BiLSTM的聚合模型框架下解决。使用正则化交叉验证进行实验，通过正则化条件约束训练集和验证集的分布差异，避免分布不一致带来的性能波动。将交叉验证得到的结果进行众数投票，以投票后的结果对不同的超参数配置进行评估，并选择若干种没有显著差异的超参数配置构成最优的超参数配置集合。然后将最优的超参数配置集合对应的子模型进行聚合，构造汉语框架语义角色识别的聚合模型。实验结果显示，本文方法的性能较基准模型显著提升了9.56%。

关键词： 框架语义角色识别；正则化交叉验证；聚合模型

Chinese Frame Semantic Role Identification Based on BiLSTM Aggregation Model

Cao Xuefei
School of Automation and Software Engineering, Shanxi University
caoxuefei@sxu.edu.cn

Li Jihong　Wang Ruibo
School of Modern Education Technology, Shanxi University
{lijh; wangruibo}@sxu.edu.cn

Niu Qian
School of Automation and Software Engineering, Shanxi University
niuqian@sxu.edu.cn

Abstract

The performance of Chinese frame semantic role identification model based on neural network is still low. Considering that the performance of neural network model is affected by hyperparameters, this paper unifies the improvement of performance and hyperparameter tuning of neural network to an aggregation model. Experiment is carried out by using regularized $m\times2$ cross-validation, and the distribution difference between the training set and the validation set is constrained by regularization conditions to avoid performance fluctuations caused by distribution inconsistency. The results obtained by cross-validation are voted, and different hyperparameter configurations are evaluated with the results of majority voting, and several hyperparameter configurations without significant differences are selected to form the optimal set of hyperparameter configurations. Then, the submodels corresponding to the optimal hyperparameter configuration set construct an aggregation model for Chinese framenet semantic role identification. Experimental results show that the performance of the

基金项目：国家自然科学基金(61806115, 62076156)

proposed method is significantly improved by 9.56% compared with the benchmark model.

Keywords: Frame semantic role identification , Regularized cross-validation , Aggregation model

1 引言

语义角色标注是语义分析中的关键技术，通过标注句子中的各种语义角色，有助于计算机准确提取句子中蕴含的语义依赖关系，从而为机器翻译、信息检索等系统提供语义支持。

汉语框架语义知识库(Chinese FrameNet，简称CFN) (刘开瑛, 2011)是以框架语义学 (Fillmore, 1976)为理论基础，以FrameNet (Baker et al., 1998)为重要参照，以真实汉语语料为事实依据构建的一个汉语词汇语义知识库，是进行汉语语义分析的一种重要资源。

框架语义学认为，词语的语义是由该词语激起的框架所描述的场景来表示的，场景中的各种参与者就是该词语所支配的语义角色。例如，对于句子“农民购买优质农用物资”，词语“购买”可以激起“商品交易”这一框架，支配两个语义角色，其中“农民”担任“购买”的“买者”角色，“优质农用物资”担任“购买”的“商品”角色。框架语义学不仅使用框架来体现词语的意义，并通过框架之间的依存关系来描述词语之间的语义关系 (王瑞波等, 2017)。进而，句子乃至篇章的语义就可以由词语激起的框架、框架之间的关系以及框架与句子中的语义角色之间的关系来表达，这就为文本的理解提供了丰富的形式化的语义信息 (宋毅君等, 2014)。

框架语义学在特定的框架下理解词语，其语义角色的描述变的更加细化和丰富，因此，CFN中的语义角色相对较多，例如，“陈述”框架就有14个语义角色。在进行汉语框架语义角色标注时，需要标注的语义角色的种类和数量都远远大于其它知识库下的标注，导致了汉语框架语义角色自动标注的难度更大。但是，丰富的语义角色可以提供更加丰富的语义信息，有助于提升计算机正确处理信息的能力。因此，构建高精度的汉语框架语义角色标注模型就成为面向汉语框架语义分析中的一个关键环节。

目前，汉语框架语义角色标注模型的精度还较低，李济洪等 (2010)经过深入分析，认为汉语框架语义角色标注的难点在于语义角色的自动识别。因此，在后续的工作中，研究者们将汉语框架语义角色标注分为语义角色识别和语义角色分类两步进行，并主要针对语义角色识别展开研究。

随着深度学习的兴起，神经网络模型也广泛应用于汉语框架语义角色识别。王瑞波等 (2017)基于词、词性等特征的分布式表示，使用一种多特征融合的神经网络构建汉语框架语义角色识别模型，并采用Dropout (Nitish et al., 2014)技术来改进模型的训练过程，最终得到了70.54%的F值。在同样的语料上，党帅兵 (2015)抽取了词特征、词性特征、位置特征、目标词特征、相邻词的组合特征、相邻词性的组合特征、基本块特征，以及词、词性和位置三者之间的两两搭配特征等多种词层面特征，在基于深层神经网络的汉语框架识别模型上得到了72.89%的F值。曹学飞等 (2022)构建了基于BiLSTM (Graves and Schmidhuber, 2005)的深度神经网络模型，使用词特征、词性特征、目标词特征和目标词的位置特征，采用3×2交叉验证 (Wang et al., 2014)进行实验，得益于BiLSTM优良的表示学习能力以及对特征的优化设计 (Cao et al., 2019)，该模型得到的F值达到了77.72%，显著高于之前其他工作的结果。

然而，针对汉语框架语义角色识别这一任务，上述基于神经网络的研究也面临以下两个问题。

- 神经网络模型的性能依赖于超参数的配置 (Reimers and Gurevych, 2017)，如何选取好的超参数使得模型性能达到最优并且稳健，即超参数调优是汉语框架语义角色识别的一个关键问题。传统的超参数调优通常是将数据集切分为训练集、验证集和测试集，将每一种不同的超参数配置看作一个独立的模型，在训练集上训练，在测试集上进行测试，检验某一种超参数配置下的模型的性能评价指标是否提高。但是，数据集的随机切分方式可能导致训练集和测试集的分布差异较大，使得模型的预测性能并不稳定，常常得到不可靠的超参数比较的结论。

- 由于计算资源的制约，早期的相关工作都是在CFN的例句库的一个子集(包含25个框架、6692条标注例句，以下简称小语料)上展开，近年来，为了和之前的方法进行对

比，研究者们延续了对这一小语料的使用。而CFN目前已标注的例句库包含了约4万条例句(以下简称大语料)，涉及205个框架，在大语料上，汉语框架语义角色识别的F值最高仅为65.31% (曹学飞等, 2022)，性能仍然较低，这是由于在框架语义学中，不同框架下的语义角色的种类和数量都不同，大语料中包含了205个框架，远大于小语料中的25个框架，这导致了框架语义角色识别的难度更大。因此，在CFN的大语料上提升框架语义角色识别的性能是目前需要解决的另一个重要问题。

基于以上分析，本文在CFN的大语料上进行汉语框架语义角色识别研究，并将上述两个问题统一到一个基于BiLSTM的聚合模型的框架下解决。由于在CFN的大语料上，针对汉语框架语义角色识别任务，曹学飞等 (2022)的工作得到了目前最优的性能，本文将该方法作为实验对比的基准方法，因此，选择了该方法中采用的BiLSTM神经网络模型来构建汉语框架语义角色识别的聚合模型。①对于第一个问题，即超参数调优。与传统的调优方法选出一个最优超参数配置不同，本文提出了应该选择若干个性能上没有显著差异的超参数配置构成“最优的超参数配置集合”。具体来讲，首先基于正则化交叉验证(Regularized $m\times2$ Cross Validation，简记为$m\times2$ RCV)进行实验，$m\times2$ RCV是一种带约束切分的m次2折交叉验证的模型训练与验证方法，它通过分布差异度量函数来均衡训练集、验证集的分布差异，避免训练集、验证集分布不一致带来的性能波动。对每一种不同的超参数配置，$m\times2$ RCV可以训练得到$2m$个模型(简记为子模型)，并可得到全部语料中每一条标注例句的m个预测序列，对m个预测序列进行众数投票，以投票结果作为最终的预测序列，并和真实的标记序列比较，从而评估不同超参数配置的性能。然后，将所有的超参数配置按照评估性能从高到低排序，再次进行增量式的投票，目的是选择“最优的超参数配置集合”，即对每一条标注例句，依次将排序后的h个超参数配置对应的$h\times m$个预测序列再次进行众数投票，检验投票后得到的性能是否进一步提升，如果性能提升，则h递增1继续以上操作，如果不再提升，则此时h个超参数配置构成了“最优的超参数配置集合”。②对于上文提到的第二个问题，即在大语料上汉语框架语义角色识别性能的提升，本文方法充分利用了每一种超参数配置在调优阶段得到的子模型，将“最优的超参数配置集合”中的h个超参数配置对应的$h\times2m$个子模型分别在公共测试集上进行测试，然后采用众数投票对$h\times2m$个结果进行聚合，从而构造汉语框架语义角色识别的投票聚合模型(见图1)。实验结果表明，本文提出的方法能在实现超参数调优的同时，还能通过对多个子模型的高效聚合，提升模型的预测性能和稳健性。

本文的组织结构如下：第2节描述了汉语框架语义角色识别任务，并介绍了本文实验使用的BiLSTM神经网络模型；第3节详细描述了本文提出的聚合模型；第4部分介绍了实验的相关设置；第5部分给出了实验结果及分析；最后总结了全文，并给出了下一步的研究方向。

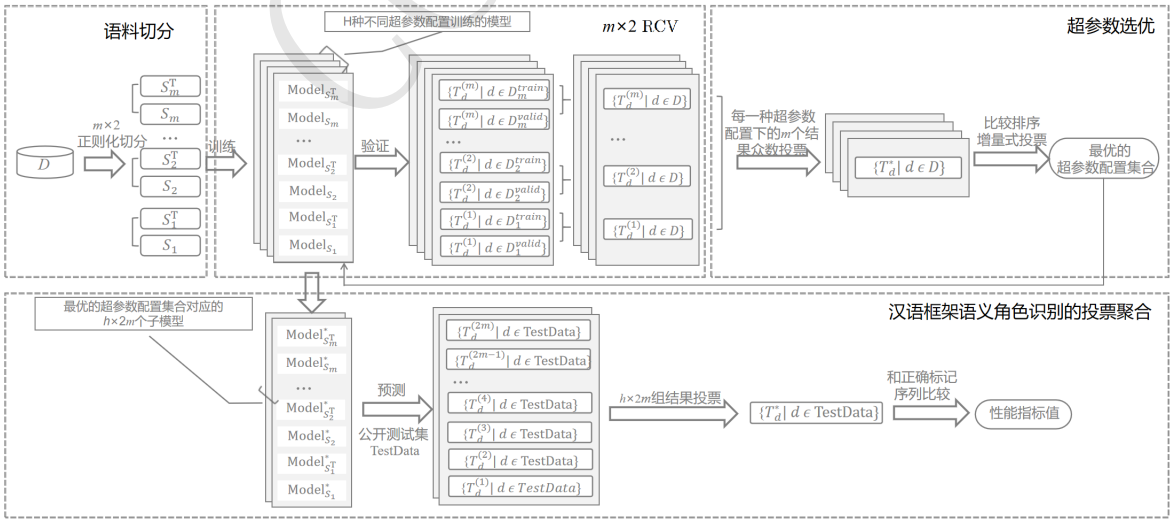


图 1：汉语框架语义角色识别的聚合模型

2 汉语框架语义角色识别任务

汉语框架语义分析包括框架识别和框架语义角色标注两部分，其中，框架识别指识别出句子中能够激起框架的成分(目标词)，并自动标注其所属的框架; 框架语义角色标注又可分为角色识别和角色标注两个步骤，即首先识别句中的哪些成分可以构成目标词所支配的语义角色，然后再对识别出的语义角色进行类别标注 (李济洪等, 2010)。

2.1 角色识别任务的形式化描述

将一条汉语句子看作是一个以词为单位的长度为N的序列，记为$S = w_1, w_2, \cdots, w_N$，$w_i$表示句子中的第$i$个词，给定句子中的目标词$w_t$，对句中每个词标记一个合适的标签$t_i$，$t_i \in \{\text{B, I, O}\}$表示句子$S$中第$i$个词对应的语义角色的边界标签，其中，B标签表示对应的词是一个语义角色块的开始词，I标签表示对应的词是一个语义角色块的中间词或结尾词，O标签表示对应的词不属于任何一个语义角色块。这样可以得到一个标签序列$T = t_1, t_2, \cdots, t_N$，基于$T$可以重构出句子$S$中的语义角色块，从而将语义角色识别转化为一个序列优化问题：$T^* = \arg\max_T \mathrm{P}(T = t_1, t_2, \cdots, t_N)$，针对该优化问题，本文构造了一个基于BiLSTM的神经网络模型进行求解(见图2)。

2.2 基于BiLSTM的语义角色识别模型

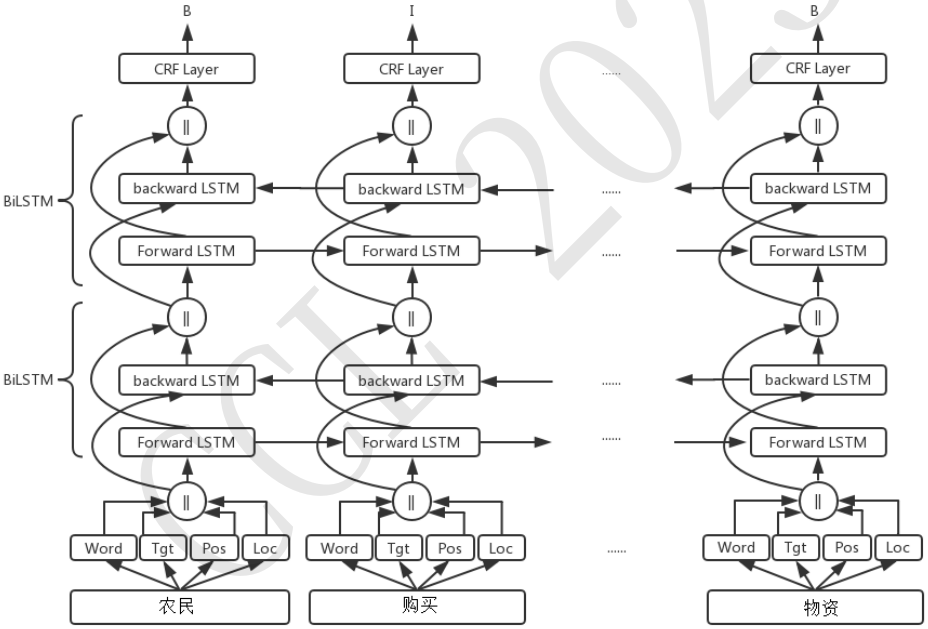


图 2: 基于BiLSTM的汉语框架语义角色识别模型

图2所示的模型包含三部分：输入层，BiLSTM层和CRF层。在输入时，将一条汉语句子看做以词为基本单位的一个序列送入模型，经过BiLSTM层的学习，在CRF层输出一个带有B、I、O标签的序列。基于输出的标签序列重构句子中的语义角色块进而完成语义角色的自动识别。

3 汉语框架语义角色识别的聚合模型

本文提出的基于BiLSTM的汉语框架语义角色识别的聚合模型如图1所示，包括以下四个模块。

第二十二届中国计算语言学大会论文集，第433页-第443页，哈尔滨，中国，2023年8月3日至5日。

3.1 语料切分

超参数选优，本质上是对不同超参数配置下的模型性能的比较，而$m\times2$交叉验证方法是常用的模型比较方法 (Wang et al., 2017; 王瑞波等, 2019)。$m\times2$交叉验证是指将数据集做m次随机切分，实施m次2折交叉验证。研究发现，基于随机切分的交叉验证进行模型比较，所得出的结果并不可靠 (Bergkirkpatrick et al., 2013; Rodríguez et al., 2010)。特别是，针对汉语框架语义角色识别任务，李济洪等 (2010)通过对CFN语料的分析发现：不同框架所支配的语义角色的种类和数量是不同的，如果对CFN语料随机切分，可能会导致框架、语义角色的分布不同。例如，包含了较多语义角色的标注例句切分在训练集中，而含有较少语义角色的标注例句切分在验证集中，或者某些框架对应的标注例句全部切分到训练集(验证集)中，这显然会增大性能评价指标的方差，在模型比较时产生不可靠的结论。

为解决语料的随机切分可能带来的性能波动，本文在$m\times2$交叉验证的基础上设计正则化条件约束训练集、验证集的分布差异，使得切分得到的训练集、验证集的分布较为均衡。具体来讲：假定语料D由$|D|$条标注例句组成，即$D=\{d_1,d_2,\cdots,d_{|D|}\}$，$D$上某次对半切分记为$\{D^{train},D^{valid}\}$，其中，$D^{train}\cup D^{valid}=D$，$D^{train}\cap D^{valid}=\emptyset$；$m\times2$交叉验证的切分集可记为：$P=\langle S_i,S_i^T\rangle$，其中，$S_i=(D_i^{train},D_i^{valid})$，$S_i^T=(D_i^{valid},D_i^{train})$，$i=1,2,\cdots,m$。$\langle S_i,S_i^T\rangle$为一个切分对，$S_i^T$为$S_i$的对折切分，某个切分下的$D_i^{train}$被称为训练集，$D_i^{valid}$为验证集。对于汉语框架语义角色识别任务，本文引入两个正则化切分条件：①框架在D_i^{train}和D_i^{valid}之间的差异要尽可能一致；②任意两个不同的D_i^{train}和$D_i^{valid}(i\neq j)$之间重叠的标注例句要尽量少且一致。

对于第一个正则化条件，本文使用离散随机变量分布一致性检验的卡方统计量来度量训练集和验证集之间框架的分布差异，设随机变量F表示框架名，其取值为离散集合$\{f_1,f_2,\cdots,f_J\}$，则框架在训练集和验证集上的分布差异，可用如下卡方统计量来度量：

$$\chi^2=\sum_{j=1}^{J}\frac{n^{train}(r_j^{train}-r_j)^2+n^{valid}(r_j^{valid}-r_j)^2}{r_j},\tag{1}$$

式(1)中，n^{train}和n^{valid}为其在训练集及验证集上出现的频次，r_j、r_j^{train}和r_j^{valid}为第j种取值f_j在数据集、训练集及验证集上的频率。实验中以单位自由度的差异度量函数χ^2/J来作为训练集和验证集上框架的分布差异度量指标，并经验性的选择指标值小于等于1的切分。对于第二个正则化条件，王瑞波等 (2019)已证明了任意两个不同的D_i^{train}和$D_j^{train}(i\neq j)$之间重叠的标注例句约为总例句数的1/4时，可以减小$m\times2$交叉验证的方差，本文采用王瑞波等 (2019)给出的方法来满足这一正则化条件。经过上述正则化处理，可以使得训练集与验证集中在框架分布上均衡，得到更为稳健的实验结果。

3.2 $m\times2$ RCV实验

在$m\times2$交叉验证的基础上通过正则化条件约束训练集、验证集的分布差异，称之为正则化交叉验证(简称$m\times2$ RCV) (王瑞波等, 2019)。

对$m\times2$ RCV的某个切分对$\langle S_i,S_i^T\rangle$，可以基于S_i的训练集D_i^{train}训练得到模型Model_{S_i}，然后对S_i的验证集D_i^{valid}中的例句进行预测，得到预测序列集合$T_d^{(i)}$，其中$d\in D_i^{valid}$。由于S_i^T为S_i的对折切分，将训练集和验证集交换进行实验，可基于D_i^{valid}训练得到模型$\text{Model}_{S_i^T}$，进而得到D_i^{train}中所有例句的预测的标记序列集合。因此，对每一种不同的超参数配置，$m\times2$ RCV可以训练得到$2m$个子模型，同时得到全部语料（训练集和验证集）中每条标注例句的m个预测的标记序列：$T_d^{(i)}$, $i=1,2,\cdots,m$, $d\in D$。

3.3 超参数调优

采用众数投票分别对语料中所有标注例句的m个预测序列进行投票聚合(以预测标记为单位投票)，假定对某条例句中的第n个词的m个预测标记的投票结果为t_n^*，投票准则如下：

$$t_n^*=l_{\arg\max_j\sum_{i=1}^{m}\mathbb{I}(t_n^i=l_j)},\tag{2}$$

第二十二届中国计算语言学大会论文集，第433页-第443页，哈尔滨，中国，2023年8月3日至5日。

其中$i=1,2,\cdots,m$，l_j为标记集合里的第j个标记，$\mathbb{I}(\cdot)$为指示函数。使用投票得到的的标记序列和该例句的真实标注序列进行比较，评估某种超参数配置下的模型的预测性能，进而可以比较得到不同超参数配置的优劣。

实际上，若干种不同的超参数配置对应的模型，其预测性能可能并没有统计意义上的显著差异。因此，本文提出了可以选择h个没有显著差异的超参数配置构成“最优的超参数配置集合”。对于h的取值，本文也给出了一种简单有效的方法：将所有的超参数配置按照评估结果从高到低排序后再次进行增量式的众数投票，即根据排序结果，依次将h个超参数配置对应的$h\times m$个预测序列进行投票，如果投票后性能进一步提升，则h递增1继续以上操作，直到性能不再提升，则此时的h个超参数配置构成了“最优的超参数配置集合”。

3.4 子模型聚合

在$m\times 2$ RCV下，超参数调优阶段每一种超参数配置均会训练得到$2m$个子模型，“最优的超参数配置集合”中的h种超参数配置可以得到$h\times 2m$个子模型，对这些子模型进行聚合构造汉语框架语义角色识别的聚合模型，即将这些子模型分别在CFN的公共测试集上进行测试，然后对$h\times 2m$个结果采用众数投票，将投票结果作为聚合模型的预测结果。

4 实验设置

4.1 语料

本文选用山西大学开发的汉语框架语义知识库(CFN)的例句库作为实验语料。该例句库包含了约4万条标注好的汉语句子，并且提供了按照约8:1:1比例切分的训练集(31526条例句)、验证集(3947条例句)和测试集(4022条例句)。针对汉语框架语义角色识别，曹学飞等 (2022)在这一语料上得到了目前最优的性能，本文将该方法作为实验对比方法，也同样采用上述切分的测试集评估本文提出的聚合模型性能。实验时，将CFN例句库提供的训练集和验证集合并，然后按照第3节的描述进行$m\times 2$ RCV进行实验(m取15)。

4.2 模型的特征及超参数设置

使用BiLSTM神经网络进行汉语框架语义角色识别时，通常可以设置一些特征丰富例句中词的信息。曹学飞等 (2022)对句子中的每个词添加了4个候选特征，包括当前词特征、当前词的词性特征、句子中的目标词特征和当前词相对目标词的位置特征(当前词在目标词左边或右边)，将该四个候选特征连同另外2个BiLSTM模型的设计选项(BiLSTM的层数和是否添加CRF层)统一看做需要调优的超参数。本文沿用了同样的设置，详细说明见表1。

表 1: 特征及超参数设置

特征	取值	说明
词	R_100, G_100	分别用随机的100维向量或GloVe (Pennington et al., 2014)预训练的100维向量来表示词
目标词	10，20	分别用随机的10维或随机的20维向量来表示目标词
词性	-，20	不使用当前词的词性特征，或用随机的20维向量来表示当前词的词性
位置信息	-，10	不使用当前词的位置特征，或用随机的10维向量来表示位置信息
BiLSTM层数	1，2	模型只采用1层BiLSTM 网络或采用2层堆叠的BiLSTM
CRF	0，1	0表示模型顶层不添加CRF层，1表示模型添加CRF层

4.3 实验的其它设置

图2所示的BiLSTM模型采用了和曹学飞等 (2022)同样的设置，包括：使用随机梯度下降算法进行训练，进行了100次的迭代，每次使用10条例句对参数进行更新(即batch-size为10)，初始学习率为0.015，学习率衰减系数为0.05；dropout rate为0.5，BiLSTM层的节点数为200。

此外，考虑到$m\times 2$ RCV对每种超参数配置都需要进行$2m$次实验，如果对表1中所有可能的超参数配置进行完全实验，所需的计算量较大，例如，本文实验中m为15，则完全实验次数为$15\times 2\times 2^6$。因此，为减少实验次数，提高超参数选优效率，本文采用$L_8(2^7)$正交表从所有可

表 2: $L_8(2^7)$正交表设计的8种超参数组合

实验号	列号						
	1(词)	2(目标词)	3(词性)	4(位置信息)	5(BiLSTM层数)	6(CRF)	7
1	R_100	10	20	-	2	1	-
2	R_100	10	20	10	1	0	-
3	R_100	20	-	-	2	0	-
4	R_100	20	-	10	1	1	-
5	G_100	10	-	-	1	1	-
6	G_100	10	-	10	2	0	-
7	G_100	20	20	-	1	0	-
8	G_100	20	20	10	2	1	-

能的超参数配置中选择相关性较低的超参数配置进行调优，即从2^6种中选出了8组有代表性的超参数配置来安排实验，实验次数减少为15×2×8。正交表的设计见表2（每一个实验号所在的行表示一种不同的超参数配置）。

4.4 性能评价指标

对于汉语框架语义角色识别任务，通常采用按识别语义角色块的P、R和F值作为模型性能的评价指标，定义如下：

P=模型识别正确的语义角色块个数(TP)/模型识别出的语义角色块总个数(TP+FP)，

R=模型识别正确的语义角色块个数(TP)/原有语义角色块总个数(TP+FN)，

F=2×P×R/(P+R)。

5 实验结果及分析

我们首先按照传统调优的方式，调优选出一种最优超参数配置(5.1节)，然后将该超参数配置对应的子模型进行聚合，初步说明聚合模型的优势(5.2节)。进一步我们选出了多个超参数配置构成“最优的超参数配置集合”(5.3节)，最后基于该集合中的多个超参数配置对应的子模型构造聚合模型(5.4节)。

5.1 最优超参数配置

对语料中的每一条例句，15×2 RCV可以得到15个预测序列，表3中的“均值”(标准差)表示某一种超参数配置下，对15×2 RCV的15组预测序列分别进行评估得到的均值和标准差(F值)，“投票结果”表示将每条例句对应的15个预测的标记序列以标记为单位进行众数投票，得到该例句投票后的标记序列，然后对语料中所有例句的投票后的标记序列进行评估的结果。

表 3: 不同超参数配置下的预测结果(%)

实验号	均值	标准差	投票结果	投票−均值
1	69.46	0.45	74.65	5.19
2	62.86	0.61	68.39	5.53
3	63.31	0.77	69.73	6.42
4	61.31	0.70	68.06	6.75
5	62.90	0.58	67.42	4.52
6	63.83	0.85	69.69	5.86
7	62.38	0.40	67.71	5.33
8	70.75	0.43	**74.97**	4.22

表3结果显示，实验号8对应的超参数配置在15×2 RCV下投票得到了最高的F值，在传统的超参数调优中，记其为最优的超参数配置。表3的最后一列显示，8组不同超参数配置上的投票

结果相比均值都有显著的提升，最低提升了4.22%，平均提升了5.48%，这也可以说明投票聚合方法在汉语框架语义角色识别上的有效性。

5.2 基于最优超参数配置的聚合模型

5.2.1 聚合模型的性能

本节讨论基于5.1节给出的最优超参数配置下的聚合模型，将8号超参数配置(最优超参数配置)对应的15×2个子模型分别在CFN的测试集上进行测试，再对15×2个结果进行众数投票，以最终的投票结果度量该聚合模型在汉语框架语义角色识别任务上的性能。实验结果如图3和表4所示。图3的结果显示，聚合模型的预测结果达到了74.87%(F值)，优于任意一个子模型单独测试得到的结果(均值为70.39%)。

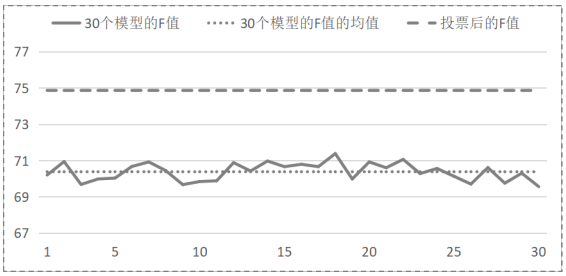


图 3: 基于最优超参数配置的聚合模型的性能

表 4: 基于投票的框架语义角色识别结果

	P	R	F
均值	71.38	69.46	70.39
聚合	77.10	72.76	**74.87**
曹学飞等 (2022)	64.16	66.49	65.31

表4的结果显示，与曹学飞等 (2022)的结果相比，本文的方法大幅提升了9.56%。与15×2个子模型测试结果的均值相比，聚合模型得到的P和R也都有明显的提高。进一步分析可知，R的提高得益于聚合模型可以得到更多的“识别正确的语义角色块”(TP)，见图4。P的提高是由于聚合模型在提高“识别正确的语义角色块”(TP)个数的同时，可以减少“识别错误的语义角色块”(FP)的数目，见图5。

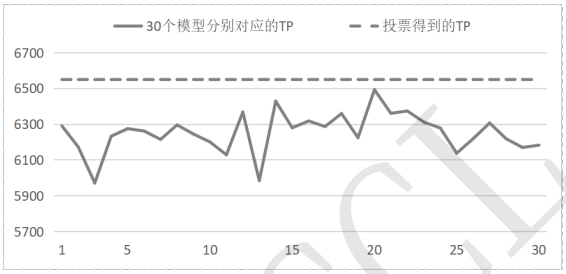


图 4: 模型投票前后TP的对比

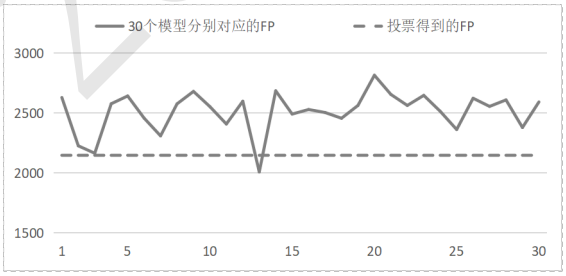


图 5: 模型投票前后FP的对比

5.2.2 显著性检验

Wang and Li (2019)给出了一种适用于m×2交叉验证下计算P、R和F值的分布函数及置信区间的方法，并说明了可以从两个结果的置信区间有无交叉来判别差异的显著性，当两个结果在1-α置信区间上没有交叉时，那么二者在置信水平α下有显著差异。本文利用该方法分别计算了聚合模型和曹学飞等 (2022)的P、R以及F值的置信区间。表5为显著性水平α=0.05下P、R和F值的置信区间的结果对比，显然，聚合模型获得的性能提升相比曹学飞等 (2022)的结果是显著的。

表 5: 置信区间的对比

	聚合	曹学飞等 (2022)
P	[76.19, 77.98]	[63.16, 65.14]
R	[71.83, 73.67]	[63.49, 65.47]
F	[74.14, 75.57]	[63.50, 65.13]

第二十二届中国计算语言学大会论文集，第433页-第443页，哈尔滨，中国，2023年8月3日至5日。

5.2.3 聚合模型的稳健性

本节我们从置信区间宽度和多次重复实验两个方面探讨聚合模型的稳健性。

(1)置信区间宽度的比较

我们分别计算了最优超参数配置(8号超参数配置)对应的15×2个子模型在测试集上预测得到的F值的置信区间，图6为所有子模型以及聚合模型对应的置信区间宽度值排序后的折线图，显然，聚合模型(投票)的置信区间宽度小于任一个参与投票的子模型的置信区间宽度，而置信区间的宽度越小，意味着对应模型的稳健性越好。

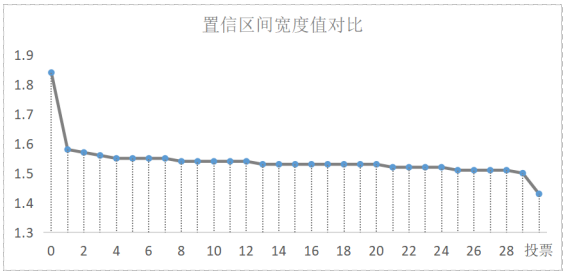


图 6: 置信区间宽度值对比

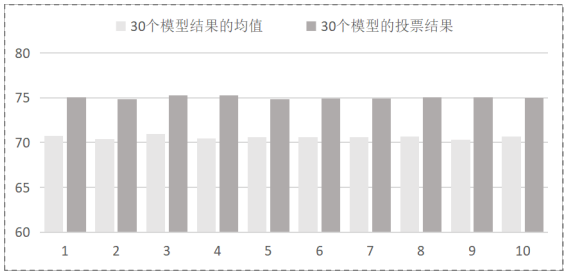


图 7: 10次重复实验结果

(2)多次重复实验的结果比较

为了进一步验证聚合模型的稳健性，本文按照如下方式在CFN的语料集进行了10折交叉验证，通过10次重复实验的结果来比较说明：

- 将CFN语料切分10份，每次选取其中9份组合，然后按照本文方法在9份组合的语料上进行15×2 RCV，然后投票选择一种最优超参数配置
- 对每次调优得到的一种最优超参数配置对应的15×2个子模型进行聚合，即将它们分别在剩下的1份语料上测试，对15×2个测试结果进行投票用来评估最终性能。

如图7所示，本文的方法可以得到非常稳健的输出结果，10次重复实验得到的聚合模型的结果(F值)，均显著优于其对应的15×2个子模型预测结果的均值，且10次聚合结果的标准差仅为0.15%。

5.3 最优的超参数配置集合

本文基于m×2 RCV进行实验，而m的选择一定程度上影响了聚合模型最终的性能，如图8所示，随着m的增大(见图8的横轴坐标)，聚合模型的预测性能在一个较大的增幅后($m < 7$)，逐渐进入一个平缓的增加态势。如果继续增大m($m > 15$)，即使能够继续保持缓慢增加的态势，也可能获得较小的性能增益，但无疑会大大增加超参数调优的时间，带来更大的计算开销。

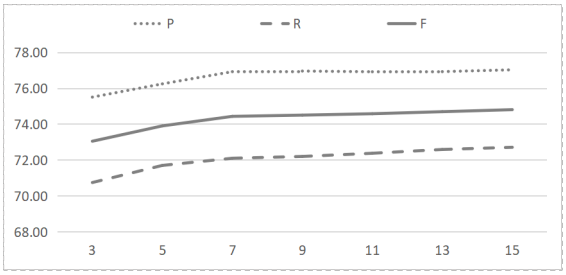


图 8: m的不同取值下投票聚合得到的结果

表 6: 构造“最优的超参数配置集合”时的投票结果

	P	R	F
8	76.97	73.07	74.97
8 & 1	77.48	73.05	**75.20**
8 & 1 & 3	76.92	72.29	74.53

直观上来讲，m×2 RCV保证了任意两次切分得到的训练集之间的重叠例句(样本)的数目尽可能少且一致，这也就意味着参与聚合的子模型来自不同的训练样本，拥有不同的预测能力。m的增大又意味着参与投票的子模型的数量增加，从而提升了聚合模型的性能。我们可以换个思路，是否可以不增大m但又可以增加参与投票的子模型的数量。考虑到若干种不同的超

参数配置对应的模型，其预测性能可能并没有统计意义上的显著差异。因此，本文进一步提出了在超参数调优时，可以选择多个没有显著差异的超参数配置构成“最优的超参数配置集合”，而不仅仅是选择一种最优超参数配置。

以表3的结果来说明“最优的超参数配置集合”的选择，8种超参数配置的优先排序如下“8 > 1 > 3 > 6 > 2 > 4 > 7 > 5”，“最优的超参数配置集合”的初始值即为8号超参数配置。然后，按照优先顺序进行增量式的投票聚合，即将在15×2 RCV时得到的1号超参数配置对应的预测序列和8号超参数配置对应的预测序列一起进行投票，再和正确标记序列对比，从表6的结果可知，本次增量投票可以得到了更高的F值，因此，当前“最优的超参数配置集合”应该包含8号和1号两种超参数配置。按照优先次序，继续使用3号超参数配置对应的预测序列进行增量式投票，表6的结果显示F值不再增加。这样我们可以得到最终的“最优的超参数配置集合”，即包含了8号和1号两种超参数配置。

5.4 基于最优的超参数配置集合的聚合模型

将“最优的超参数配置集合”中的每一种超参数配置在m×2 RCV时得到的子模型共同作为聚合模型的子模型，这样可以大大增加参与投票的子模型的数量。本文实验中，我们将“最优的超参数配置集合”中8号和1号超参数配置对应的子模型(2×15×2个)在公共测试集进行测试，并将预测结果进行众数投票，得到的结果如表7所示。显然，基于“最优的超参数配置集合”的聚合模型的结果(P、R和F)均比基于“最优超参数配置”的聚合模型的结果更好，虽然F值仅提高了0.23%，但二者之间有显著性差异的置信度达到了71.78%[0]。

表 7: 基于“最优的超参数配置集合”的聚合结果

编号	TP	FN	FP	P	R	F
8	6550	2452	1946	77.10	72.76	74.87
8 & 1	6575	2427	1918	**77.42**	**73.04**	**75.16**

表 8: 调优时，不同超参数配置的置信区间

8号	1号	3号
[74.86, 75.35]	[74.40, 74.90]	[69.48, 69.99]

分析发现，“最优的超参数配置集合”中的8号和1号超参数配置，在调优时分别对应的投票结果没有显著差异(即表3中实验号8和1对应的投票结果)。表8给出了显著性水平α=0.05下二者的F值的置信区间对比，结果显示，8号和1号的置信区间有交叉部分，这意味着这两种超参数配置并无显著差异，而它们又都和3号超参数配置有显著差异，因此，本文的“最优的超参数配置集合”实质上是超参数调优时得到的没有显著差异的若干个最好的超参数配置，进而可以将它们在m×2 RCV时训练得到的子模型聚合，通过增加性能上没有显著差异的子模型的数量提升聚合模型的性能。

6 总结

本文提出了一个基于BiLSTM的汉语框架语义角色识别的聚合模型，首先，采用m×2 RCV进行试验，可以尽可能降低实验结果对语料切分的敏感性，进而对m组结果进行众数投票，以投票结果来评估不同超参数配置的优劣。然后根据超参数配置的优先次序继续进行增量式投票，得到“最优的超参数配置集合”。最后利用“最优的超参数配置集合”对应的所有子模型构建汉语框架语义角色识别的聚合模型。与基准方法相比，该聚合模型可以显著提升汉语框架语义角色的识别性能，说明了该聚合模型的有效性。此外，本文创新性的提出了超参数调优时应该选择多种性能上没有显著差异的超参数配置构成“最优的超参数配置集合”，这样可以增加参与聚合的子模型的数量，从而提升聚合模型的性能。本文使用基于BiLSTM的神经网络模型作为子模型进行聚合，一方面是受计算平台性能的限制，另一方面是为了和基于BiLSTM的基准方法更好的对比。理论上，任何一种神经网络模型都可以作为该聚合模型框架中的子模型，因而，利用BERT等预训练模型作为子模型构建汉语框架语义角色识别的聚合模型也是我们下一步的研究计划。

参考文献

刘开瑛. 2011. 汉语框架语义网构建及其技术研究. 中文信息学报, 25(6):46–52.

[0]置信度的计算采用了 (Wang and Li, 2019)中的贝叶斯检验方法。

Charles J. Fillmore. 1976. *Frame semantics and the nature of language.* Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280:20–32.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. *The Berkeley FrameNet Project.* Association for Computational Linguistics, 86–90.

王瑞波，李济洪，李国臣，杨耀文. 2017. 基于*Dropout*正则化的汉语框架语义角色识别. 中文信息学报, 31(1):147–154.

宋毅君, 王瑞波，李济洪. 2014. 基于条件随机场的汉语框架语义角色自动标注. 中文信息学报, 28(3):36–47.

李济洪, 王瑞波, 王蔚林. 2010. 汉语框架语义的自动标注. 中文信息学报, 21(4):597–611.

Srivastava Nitish, Hinton Geoffffrey, Krizhevsky Alex, Sutskever Ilya, and Salakhutdinov Ruslan. 2014. *Dropout: a simple way to prevent neural networks from over-fifitting. Journal of Machine Learning Research*, 15(1):1929–1958.

党帅兵. 2015. 基于词分布表征的汉语框架语义角色识别研究. 山西大学.

曹学飞, 李济洪, 王瑞波, 牛倩, 王钰. 2022. 基于稳健设计的双向长短期记忆神经网络模型的调优方法. 应用概率统计, 38(3):317–332.

Alex Graves and Jurgen Schmidhuber. 2005. *Framewise phoneme classifification with bidirectional LSTM and other neural network architectures.* Neural Networks, 18:602–610.

Yu Wang, Jihong Li Ruibo Wang and Huichen Jia. 2014. *Blocked 3×2 cross-validated t-Test for comparing supervised classifification learning algorithms.* Neural Computation, 26(1):208–235.

Xuefei Cao, Jihong Li, Ruibo Wang, Yu Wang, Qian Niu and Junfeng Shi. 2019. *Calibrating GloVe model on the principle of Zipf's law.* Pattern Recognition Letters, 125(7):715–720.

N Reimers and I Gurevych . 2017. *Reporting score distributions makes a diffffference: performance study of LSTM-networks for sequence tagging.* Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 338–348.

Ruibo Wang, Yu Wang, Jihong Li, Xingli Yang and Jing Yang. 2017. *Block-Regularized m × 2 Cross-Validated Estimator of the Generalization Error.* Neural Computation, 29(2):519–554.

王瑞波, 王钰, 李济洪. 2019. 面向文本数据的正则化交叉验证方法. 中文信息学报, 33(5):54–65.

Bergkirkpatrick T, Burkett D and Klein D. 2013. *An Empirical Investigation of Statistical Significance in NLP.* IEEE Geoscience & Remote Sensing Letters, 13(3):457–461.

Rodríguez J D, Perez A, Lozano J A. 2010. *Sensitivity analysis of k-fold cross validation in prediction error estimation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3):569—575.

Pennington J, Socher R and Manning C. 2014. *GloVe: global vectors for word representation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543.

Ruibo Wang and Jihong Li. 2019. *Bayes Test of Precision, Recall, and F1 Measure for Comparison of Two Natural Language Processing Models.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4135–4145.

由L2到L1的跨语言激活路径研究——基于词汇识别的ERP数据*

杨思琴
中国人民大学文学院
ysq44@outlook.com

江铭虎
清华大学人文学院
jiang.mh@tsinghua.edu.cn

摘要

跨语言词汇激活模型是当下语言认知与计算研究的热门话题。本研究运用事件相关电位技术(event-related potentials, ERPs) 探索了二语学习者在识别二语(second language，简称L2)词汇时激活母语(native language，简称L1) 词汇表征的路径。研究设计了隐性启动范式来开展两个实验，通过观察被试能否感知只有激活L1词汇表征才能发现的对译词重复情况这一隐性条件来推测激活结果。脑电结果显示，实验一的被试在执行语义判断任务时，对译词重复与否产生了显著的N400差异，这表明被试经由概念表征激活了L1词汇表征，进而证明了激活路径Path-1（L2->L1）的存在；实验二的被试在执行书写形式判断任务时，在没有语义启动的情况下，同样感知到了对译词这一隐性条件，这表明他们可以由L2词汇表征直接激活L1词汇表征，从而证明了激活路径Path-2（L2->L1）的存在。总体而言，词汇识别过程中从L2词汇表征到L1词汇表征的激活路径与修正层次模型（the Revised Hierarchical Model, RHM）描绘的词汇产出过程的激活路径类似。据此，本研究推测，尽管大脑在词汇识别和词汇产生过程中采用不同的处理机制，但在跨语言词汇激活过程中，它们依然存在某些共通之处。

关键词： 词汇；激活；路径；模型；N400

Cross-lingual Activation Path from L2 to L1——Based on ERP Data during Word Recognition

Yang Siqin
Renmin University of China
School of Liberal Arts
ysq44@outlook.com

Jiang Minghu
Tsinghua University
School of Humanities
jiang.mh@tsinghua.edu.cn

Abstract

Cross-lingual lexical activation models are a hot topic in current linguistic cognition and computation research.This study used the ERPs to explore the pathway by which L2 learners activate their L1 lexical representations while recognizing their L2 words. The implicit priming paradigm was designed to carry out two experiments, inferring the activation condition by observing whether participants can perceive the implicit condition of repetition of translated words that can only be found after activating their L1 lexical representation. EEG results showed that when participants in Experiment 1 performed the semantic relatedness judgment task, there was a significant difference in N400 between whether the translated words were repeated or not, which indicated that

*本课题获得国家自然科学重点基金项目(62036001)的支持

participants activated their L1 lexical representation through the conceptual representation, proving the existence of Path-1 (L2->L1); when participants in Experiment 2 performed the orthographic judgment task, they also perceived the implicit condition of word translation without semantic priming, which suggested that their L1 lexical representation could be directly activated from their L2 lexical representation, thus proving the existence of Path-2 (L2->L1). Overall, the activation pathway from L2 lexical representations to L1 lexical representations during word recognition is consistent with that of word production described by the RHM. Accordingly, this study speculated that although the brain adopts different processing mechanisms in the process of word recognition and word production, they still have some commonalities during the cross-lingual word activation.

Keywords: Lexical , Activation , Path , Model , N400

1 引言

不同语言之间的激活状态（简称跨语言词汇激活）是当下认知语言学和心理语言学关注的焦点，同时，跨语言词汇激活模型也是语言认知与计算研究的热门话题。以往国内外语言学家们对此不遗余力地开展了众多研究，发现并多次验证了二语学习者在识别L2词汇时，会自动激活L1词汇。然而，遗憾的是，他们对词汇识别时的激活路径则少有涉及。鉴于激活路径亦是心理词典模型结构中的关键细节，本研究将对此开展进一步探索，试图为词汇识别过程中的跨语言词汇激活模型提供一些新的见解。

语言学家威尔金斯曾言：“如果没有语音和语法，我们所能表达的内容寥寥无几。但若没有词汇，我们则很难传递任何信息(Wilkins, 1972)。”特别是对于二语学习者而言，词汇学习是语言学习过程中关键且费时的环节。并且，在早期阶段，L2词汇的掌握也倾向于依赖L1词汇的使用。于是，在研究跨语言词汇之间关系的话题上，语言学家们从不同角度提出了跨语言词汇激活模型。其中，最为经典的三个模型分别是Kroll (1994) 提出的RHM，Dijkstra (2002) 提出的双语交互激活模型+（Bilingual Interactive Activation Model, BIA+）和Grosjean (1988) 提出的词汇访问双语模型（Bilingual Model of Lexical Access, BIMOLA）。

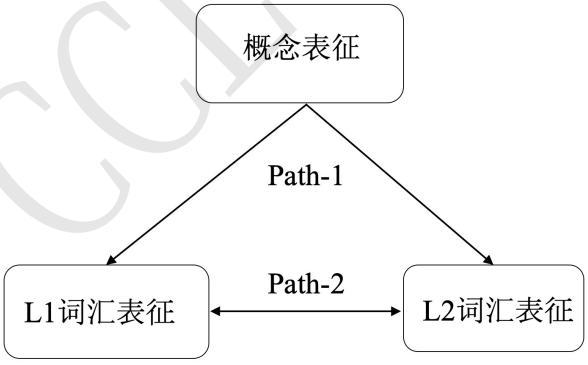


图 1: RHM两条路径示意图

RHM在描述词汇产生时明确区分了概念表征和词汇表征，并描绘了L1 词汇表征和L2词汇表征之间的两条路径, 如图1, (1) 经由概念表征间接激活的路径(Path-1)，(2) 直接激活的路径(Path-2)。然而，RHM是否适用于词汇识别目前尚未可知。BIA+则基于词汇识别而提出，它认为双语词汇信息的激活是从最基础的信息(字形和字音)到(最终)语义的自动激活，词汇激活不仅受到另一种语言词汇书写形式的影响，还受到跨语言的语音和语义表征的影响。遗憾的是，该模型的构建仅基于拥有同种书写形式的不同语言，不足以阐释其它语言之间的激活情

况。BIMOLA描述的激活过程与BIA+类似。值得关注的是，它对语言模式进行了更清晰的交代：单语语言模式下的目标语言网络会被激活得比较强烈，而非目标语言网络被激活的程度则相对较弱；但在双语模式下，两种语言网络可以同时被激活。

通过对RHM、BIA+和BIMOLA三个模型的分析，本研究发现，不同词汇表征之间的激活路径在语言模式得到控制的前提下还有待进一步探究。以往的研究分析表明，RHM清晰地反映了双语心理词典的结构，即二语学习者不同语言的词汇是分别表征的，但它们的概念是共同表征的(李荣宝，彭聃龄，郭桃梅, 2003; Zeelenberg and Pecher, 2003; Li et al., 2009; Perea et al., 2021)。例如，彭聃龄等曾经运用脑电技术探索英语和汉语两种语言词汇的表征情况。结果显示：反映语言书写形式加工的P190脑电波在汉英两种词汇加工过程中存在明显的差异；而反映语义加工的N400脑电波在不同语言的词汇中则没有显著的差别(李荣宝，彭聃龄，郭桃梅, 2003)。综合来看，该研究结果清晰地展现了在词汇记忆中，词汇表征与概念表征的分离，以及不同语言词汇表征之间的区分。

基于此，本研究推测L2词汇表征和L1词汇表征之间也存在两条关联路径，Path-1和Path-2，但这两条路径是否可以作为词汇之间的激活路径还尚未可知。以往研究表明，当二语学习者在二语模式下阅读L2词汇时，大脑会自动激活母语词汇中的对译词(Thierry and Wu, 2007; Wu and Thierry, 2012; 肖巍，倪传斌, 2016; 肖巍, 2018)，这为从L2词汇表征到L1词汇表征的激活结果提供了实验证据。在接下来的研究中，本研究将重点观测从L2词汇表征到L1词汇表征的两条路径[Path-1（L2->L1）和Path-2（L2->L1）]是否可以激活。

2 研究问题

为了控制语言模式从而营造单语模式，避免双语模式下两种语言同时激活给跨语言词汇激活路径的观测带来干扰，本论文计划采用只呈现目标语言的隐式启动范式作为词汇呈现的方式。在该范式中，存在一个只有翻译成非目标语言后才能觉察到的隐性因素（例如，window-curtain，窗户-窗帘）。实验通过观测被试能否觉察到该隐性因素（例如，窗），来推测非目标语言是被否激活。

针对Path-1（L2 -> L1），本研究将通过开展实验一，运用语义判断任务来检测它的激活情况。因为在语义判断任务中，被试只有调用了词汇的概念意义，才能得出正确的反应。所以本研究假设，如果语义判断任务下的L1词汇表征被激活，那么，该激活过程必然会经过Path-1。鉴于L1词汇概念的激活可能会同时经过两条路径，而经过Path-1的激活就必然会启动概念表征。因此，为了探索Path-2（L2->L1），本研究将采用一个抑制概念意义加工的书写形式判断任务来开展实验。Valdes (2005) 和Spruyt (2009) 发现，语义启动可以通过对被试注意力的分配进行调节，控制甚至抑制。Mari-Beffa (2005) 也发现，当被试的实验任务与语义启动无关时，大脑的控制机制将抑制语义的激活。因此，本研究将设计一个引导被试关注词汇书写形式特征而非语义的任务，简称为书写形式判断任务。它要求被试判断连续呈现的英语词对（启动词和目标词）中是否有且只有一个英语单词中有两个元音字母相邻，其中，满足以上情况的有两种（如interview-reporter，第一个单词中存在两个相邻的元音字母；arm-shoulder，第二个单词中存在两个相邻的元音字母）。不满足也有两类（如bedroom-wheel，都存在两个相邻的元音字母；network-tennis，都不存在两个相邻的元音字母）。本研究通过观察在语义没有启动的情况下L1词汇表征的激活情况，来推测该激活是否经过Path-2。

鉴于事件相关电位（event-related potential, ERP）具有毫秒级的分辨率，能更精确地记录数毫秒内大脑神经活动的细微变化，本研究将把它作为实验的研究技术。其中，作为最广泛的参考脑电成分之一(Silva-Pereyra et al., 1999; Laszlo and Federmeier, 2011; Tiedt et al., 2020)，N400可以反映大脑对词汇的概念和形式加工。例如：当两个词语先后呈现给被试时，不存在语义关联的词对将比存在语义关联的词对诱发更负的N400波幅(Thierry and Wu, 2007; Jia et al., 2013; Friesen et al., 2016; Yang et al., 2021b)。类似的，不存在形式关联的词语（例“兔子-课桌”）将比存在形式关联的词语（例“小说-小孩”）诱发较大的N400波幅(Thierry and Wu, 2007)）。另外，当多个与N400相关的语言现象同时出现的时候，N400有可能会出现叠加和覆盖的现象(Ye et al., 2006; Yang et al., 2021a)。据此，本研究将该脑电成分作为分析词汇加工的特征参数。

3 实验一

3.1 语料设计

表 1: 实验一语料设计范例

	汉语对译词首字重复		汉语对译词首字不同	
	启动词	目标词	启动词	目标词
语义相关	S+R+		S+R-	
	school	academic	coffee	milk
	学校	学术	咖啡	牛奶
语义无关	S-R+		S-R-	
	airport	chance	honesty	convenience
	机场	机会	诚实	方便

实验一使用了120组英文单词词对（240个单词）作为实验材料，包含2个变量：语义相关性和对译词首字重复情况，如表1所示。2×2组成4组实验材料（Semantic：S+代表语义相关，S-代表语义无关；Repetition：R+代表汉语对译词首字重复，R-代表汉语对译词首字不重复）。S+R+代表语义相关词对且汉语对译词首字重复的词对；S+R-代表语义相关词对且汉语对译词首字不重复的词对；S-R+代表语义无关词对且汉语对译词首字重复的词对；S-R-代表语义无关词对且汉语对译词首字不重复的词对。实验材料的范例如表1所示。其中，每组包含30个英语单词词对。

实验一从词频、词长、具体度、L2与对译词的一致性和词对的相关性五个方面对实验语料进行控制。由于单词以成对的形式出现，因此在控制词频时，研究一将词汇呈现的顺序（简称为词序）也作为一个因素纳入分析。针对词频，本研究以Van-Heuven (2014) 新改进的英语词频数据库为参考。以语义相关性、对译词首字重复情况和词序为三因素的方差分析结果显示，任何交互作用和主效应均不显著($ps > 0.5$)。对于词长，本研究计算了作为语料的英语单词的字母数。以语义相关性、对译词首字重复情况和词序为三因素的方差分析结果显示，任何交互作用和主效应均不显著($ps > 0.5$)）。为了确保英文单词翻译和中文词语的一致性。本研究邀请了16名大学生对英语词语进行翻译，并观测他们在看到英文单词时，首先想到的中文翻译是否与本研究中采用的中文词语相同。除了汉字拼写错误以外，每个词语的一致性均在75%及以上。

此外，本研究招募了15名大学生使用5点李克特量表对英语词对的语义相关程度进行了评估。其中，5分代表语义相关性最高，1分代表语义相关性最低。将评估数据收集完之后，研究一将语义相关性和对译词首字重复情况这两个条件作为两因素进行统计分析。结果显示，语义相关性的主效应显著[$F(1, 116) = 4124.094, p < 0.001, \eta_p^2 = 0.973$]，此外，语义相关性和对译词首字重复也存在交互效应[$F(1, 116) = 5.448, p = 0.021, \eta_p^2 = 0.045$]。配对t检验显示，S+R+的语义相关性高于S-R+的语义相关性[$t(29) = 31.148, p < 0.001$],和S-R-的语义相关性[$t(29) = 38.300, p < 0.001$]；S+R-的语义相关性高于S-R+的语义相关性[$t(29) = 62.058, p < 0.001$], 和S-R-的语义相关性[$t(29) = 100.116, p < 0.001$]。但是，S+R+的语义相关性与S+R-的语义相关性差异不显著[$t(29) = -1.282, p = 0.210$]。值得注意的是，S-R+的语义相关性高于S-R-的语义相关性[$t(29) = 2.414, p = 0.022$]。

在具体实验中，由于满足实验方案要求的英语词对数量有限，实验中的所有材料重复呈现2次，以保证测量次数。一共有240个试次。在呈现的过程中，所有的实验材料均以伪随机的顺序呈现。

3.2 被试

实验一一共招募了22位大学生参加了此次实验。所有的被试皆为大陆学生。其中2位被试的脑电数据无法提取和使用。最终有20位被试（9位男生）的数据进入后续分析。他们的年龄范围在19岁～26岁之间。其中，女性的平均年龄是22.27岁，男性的平均年龄是21.22岁。根据爱丁堡的惯性测试，参加实验的被试都属于右利手(Oldfield, 1971)。另外，在实验之前，所有被试的

视力或矫正视力均正常。他们在参加实验时身体健康，且没有任何神经或精神疾病的病史。所有被试均获得了《赫尔辛基宣言》（Helsinki Declaration）的知情同意。为了控制英语水平，在选取的被试中，所有被试都参加了大学英语六级考试，且分数都在550分以上。所有被试都获得了等额的现金报酬，并同意了当地伦理委员会批准的实验方案。

3.3 实验流程

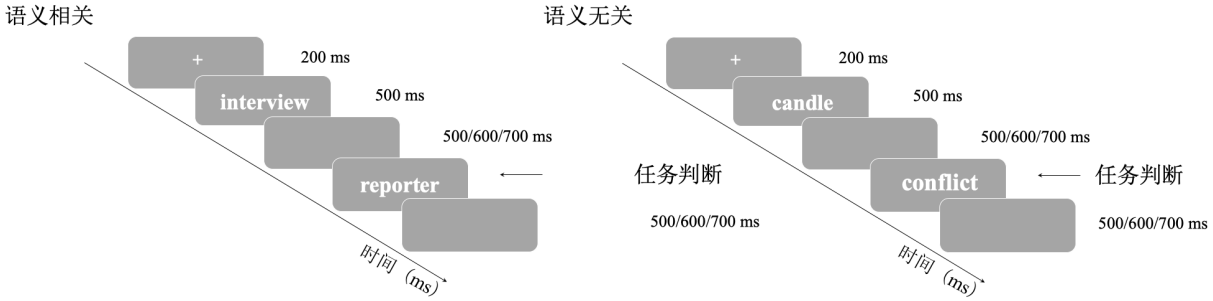


图 2: 实验一试次的流程图范例

被试坐在离电脑屏幕80 cm左右的椅子上完成操作。本研究运用心理学软件工具E-Prime 2.0将任务要求和实验材料均呈现在电脑屏幕上。屏幕背景为灰色，而作为实验材料的英语词对均以白色34号的字体呈现。被试被告知实验一的操作任务（即判断前后呈现的词语语义是否相关）后开始实验。首先是练习模块。屏幕向被试呈现了20对英语词对，包括与正式实验材料类似的四种类型分别5对。被试在练习实验任务的过程中，只有正确率达到了90%及以上之后可以进入接下来的正式实验。实验试次的流程图范例如图2所示，每个试次的流程如下：首先，一个提示符号“+”呈现在屏幕中央，持续时间为200 ms，提醒被试集中注意力；其次，第一个英语单词出现在屏幕中央，停留时间为500 ms；之后是一个空白屏幕，停留的时长是500、600或700 ms中的任意一个；最后，第二个词出现在屏幕上，直到被试对此反应后才消失；被试作出反应后，空白屏幕再次出现。此处空白屏幕的停留时间是随机的，可能停留200 ms，也可能停留300 ms 或400 ms。到此为止，一个实验任务便完成了。所有词对均以伪随机顺序呈现。研究结果的数据分析仅基于正式实验。

3.4 数据采集与分析

实验一使用62 Ag/AgCl 电极弹性帽来记录被试在参加实验过程中的脑电数据。该电极帽上配备着国际10-20电极放置系统（Easycap；Brain Products GmbH，Gilching，Germany）。设备中的FCz和AFz电极分别用作参考电极和接地电极。两个眼电极用于测量垂直和水平眼电图，分别放置在左眼下方和右眼外侧(Picton et al., 2000)。在准备正式实验前，电极阻抗保持在10kΩ以下。在Brain Vision Recorder软件（Brain Products，Munich，Germany）的控制下，研究使用带通为0.01-100Hz的BrainAmpDC放大器系统（Brain Products GmbH）记录被试的EEG数据。电生理记录的过程严格遵循以往已发表并被公认的流程(Yang et al., 2021b)。

在收集完数据之后，使用Brain Vision Analyzer软件（Brain Products，Munich，Germany）来分析EEG数据，分析的过程分别是调整参考、眼电矫正、过滤、分割、基线校正、伪影剔除和平均。调整参考：脑电图数据被重新引用到关联的乳突，然后在预处理后引用到所有脑电图通道的平均值。眼电矫正：使用嵌入在Brain Vision Analyzer软件（Brain Products，Munich，Germany）中的基于独立成分分析的程序校正眼部伪影。过滤：EEG数据从0.1到35Hz离线进行带通滤波。分割：从目标词开始前200 ms到开始后800 ms的EEG被分割成主要观测并呈现在脑电图上的窗口（200 ms目标前基线）。伪影剔除：剔除可能由肢体等外部运动引起的伪影（电压超过±80μV）波段。平均：将处理好的脑电进行叠加平均。

本研究需要分析的数据包括行为数据和脑电数据。其中，行为数据包括反应时间和反应正确率。反应时间取的是每个被试在同种条件下所有反应时间的中位数（由于样本容量较大，之

所以选择中位数，是因为其排除有效被试中的个别极端数据，更能反映整体水平）。反应正确率是同种条件下正确反应的次数与该条件下所有反应次数的比例。

对于脑电数据，本研究主要观测和分析Cz、C1、C2、C3、C4、FCz、FC1和FC2这8个电极记录的N400成分。N400成分被认为是反映词汇语义相关性和重复启动的脑电成分(Thierry and Wu, 2007; Wu and Thierry, 2010; Lau et al., 2013; Yang et al., 2021b)。该电极簇是通过观测N400的脑电地形图分布来确定的。N400的分析窗口为300～500 ms。该时间窗口与之前研究所取的时间窗口近似(Thierry and Wu, 2007; Wu and Thierry, 2010; Liang and Chen, 2014)。本研究的数据分析采用的N400 是分析窗口中电极簇（包含8个电极点）的平均振幅。

数据分析采用的是双因素（语义相关性×汉语对译词首字重复情况）重复测量方差分析。当自由度大于1时，则采用Greenhouse-Geisser 法校正。如果交互作用显著，则继续进行简单效应分析或t检验。

3.5 实验结果

3.5.1 行为数据结果

实验一记录了被试的正确率和反应时间。S+R+的平均正确率为90.1%（SD = 1.2%）。S+R-的平均正确率为96.1%（SD = 1.1%）。S-R+的平均正确率为96.7%（SD = 0.7%）。S-R-的平均正确率98%（SD = 0.5%）。每种条件的平均正确率均在90% 以上，保证了数据的有效性。

反应时间从第二个英语单词（目标词）出现的时候开始记录。S+R+的反应时间为756 ms（SD = 28）。S+R-的反应时间为722 ms（SD = 26）。S-R+的反应时间为794 ms（SD = 32）。S-R-的反应时间为740 ms（SD = 29）。研究采用双因素（语义相关性×汉语对译词首字重复情况）重复测量方差分析方法分析反应时间。结果显示，语义相关性有接近显著的主效应[$F(1,19) = 3.315, p = 0.084, \eta_p^2 = 0.149$]。汉语对译词首字重复的主效应显著[$F(1,19) = 19.366, p < 0.001, \eta_p^2 = 0.505$]。

3.5.2 脑电数据结果

实验一8个电极点各自的N400波幅（ERP）、N400平均波幅（N400 Amplitude）和脑电地形图（The Scalp Topographies of N400）如图3所示。分析结果显示，语义相关性的主效应显著[$F(1,19) = 34.745, p < 0.001, \eta_p^2 = 0.646$]，并且，语义相关性×汉语对译词首字重复的交互作用显著[$F(1,19) = 5.343, p = 0.032, \eta_p^2 = 0.219$]。在语义相关的情况下，汉语对译词首字引发的N400有显著差异，其中，汉语对译词首字重复条件引发的波幅比不重复条件引发的波幅更负[$F(1,19) = 7.09, p = 0.015$]。但是，在语义无关的情况下，统计结果没有表现出显著差异[$F(1,19) = 0.35, p = 0.560$]。在汉语对译词首字重复的词对中，语义无关条件引发的脑电波幅明显比语义相关条件引发的脑电波幅更负[$F(1,19) = 10.96, p = 0.004$]，在汉语对译词首字不重复的词对中也如此[$F(1,19) = 43.80, p < 0.001$]。

3.6 讨论

实验一结果中的正确率均在90% 以上，保证了数据的有效性。行为实验结果显示，被试在语义相关性上表现出了接近显著的反应时间差，在汉语对译词首字重复与不重复的两种情况下也表现出了明显的反应时间差。据此可推测，被试在进行语义判断过程中不仅觉察到了语义相关性因素，而且也激活了英语词对的汉语对译词。这契合(Thierry and Wu, 2007; Wu and Thierry, 2010; 肖巍，倪传斌, 2016; 肖巍, 2018)的研究结果。

在脑电数据结果中，只有在语义相关的情况下，汉语对译词首字重复的变量才会引发显著的N400差异，在语义无关的情况下则没有显著的差异。值得注意的是，在语义相关的词对中，汉语对译词首字重复条件引发的波幅比不重复条件引发的波幅更负。显然，这是汉语对译词首字重复导致的波幅差异。但是，在语义无关的词对中，无论汉语对译词首字是否重复，都会因为语义无关而引发一个显著的N400效应。并且，语义无关引发的N400波幅比汉语对译词首字重复引发的波幅明显更大。由以往N400成分意义的综述可知，N400是脑与语言研究中反映语义信息加工和反应重复启动的脑电成分(Thierry and Wu, 2007; Jia et al., 2013; Friesen et al., 2016; Yang et al., 2021b)。当多个与N400相关的语言现象同时出现时，N400有可能会出现叠加和覆盖的现象(Ye et al., 2006; Yang et al., 2021a)。因此，本研究推测，此处母语对译词首字重复条件引发的N400很可能与语义无关引发的N400重合在一起了。

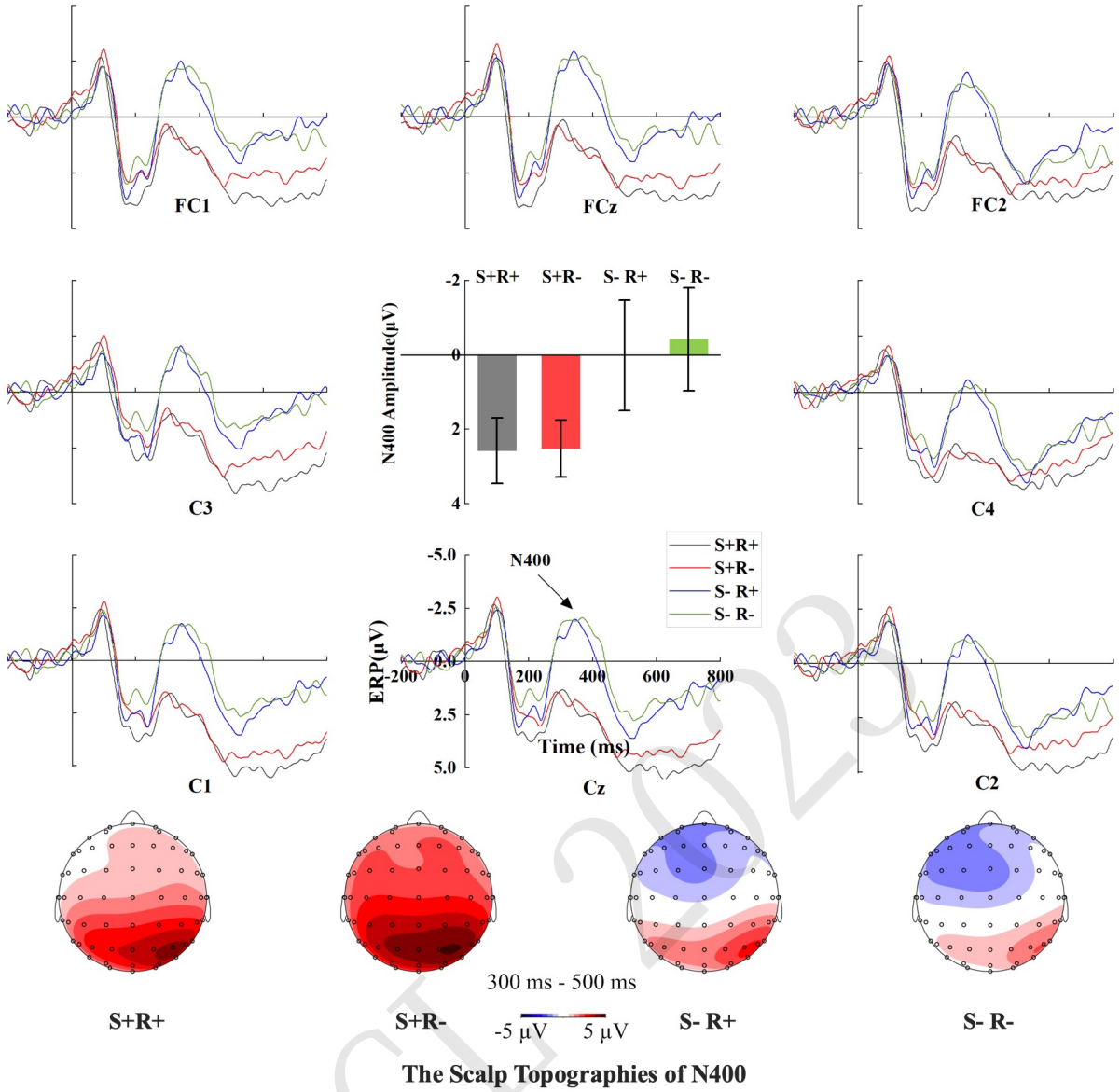


图 3: 实验一脑电结果图

总之，结合反应时间的结果来看，当被试在加工二语词汇的语义内容时，大脑可能自动激活了它在母语中的词汇。并且，结合实验一的假设和实验设计可知，该激活过程必然经过不同语言共享的概念表征。因此，本研究推测，从L2词汇表征到L1词汇表征的激活过程可能经过Path-1（L2->L1）。

4 实验二

4.1 语料设计

实验二共设置了3个变量，分别是语义相关性，对译词首字重复情况和元音字母相邻情况。如表2，2×2×2组成8组实验材料（Semantic：S+代表语义相关，S-代表语义无关；Repetition：R+代表汉语对译词首字重复，R-代表汉语对译词首字不重复；Adjacent：A+代表有且只有一个英语单词中有两个元音字母相邻，A-代表不满足有且只有一个英语单词中有两个元音字母相邻的情况）。S+R+A+：语义相关词对、汉语对译词首字重复、有且只有一个英语单词中有两个元音字母相邻；S+R+A-：语义相关词对、汉语对译词首字重复、不满足有且只有一个英语单词中有两个元音字母相邻的情况；S+R-A+：语义相关词对、汉语对

表 2: 实验二语料设计范例

	汉语对译词首字重复				汉语对译词首字不同			
	元音字母相邻		元音字母不相邻		元音字母相邻		元音字母不相邻	
	启动词	目标词	启动词	目标词	启动词	目标词	启动词	目标词
语义相关	S+R+A+		S+R+A-		S+R-A+		S+R-A-	
	education	classroom	British	English	bone	blood	bus	driver
	教育	教室	英国	英语	骨头	血液	巴士	司机
语义无关	S-R+A+		S-R+A-		S-R-A+		S-R-A-	
	fruit	level	happen	fever	pencil	society	century	plant
	水果	水平	发生	发烧	铅笔	社会	世纪	植物

译词首字不重复、有且只有一个英语单词中有两个元音字母相邻；S+R-A-: 语义关词对、汉语对译词首字不重复、不满足有且只有一个英语单词中有两个元音字母相邻的情况；S-R+A+: 语义无关词对、汉语对译词首字重复、有且只有一个英语单词中有两个元音字母相邻；S-R+A-: 语义无关词对、汉语对译词首字重复、不满足有且只有一个英语单词中有两个元音字母相邻的情况；S-R-A+: 语义无关词对、汉语对译词首字不重复、有且只有一个英语单词中有两个元音字母相邻；S-R-A-: 语义无关词对、汉语对译词首字不重复、不满足有且只有一个英语单词中有两个元音字母相邻的情况。

由于实验二仅关注语义相关性和对译词首字重复情况这两个条件的行为数据和脑电数据，元音字母条件的设置仅仅是为了制造一个加工书写形式的任务而非研究目的，因此，实验二没有考虑它的统计数据。与实验一类似，实验二也分别从词频、词长、具体度、L2与对译词的一致性和词对的相关性五个方面对实验语料进行了控制。

4.2 被试

实验二一共招募了23位大学生作为被试。所有的被试皆为大陆学生。其中，3位被试的脑电结果干扰信号太多，数据无效。最终有20位被试（10位男生）的数据进入最终的统计分析。他们的年龄范围在19岁～26岁之间。其中，女性的平均年龄是22.1岁，男性的平均年龄是21.3岁。根据爱丁堡的惯性测试结果，参加实验的被试都是右利手(Oldfield, 1971)。另外，在实验之前，所有被试的视力或矫正视力均正常。被试在参加实验时身体健康，且没有任何神经或精神疾病的病史。所有被试均获得了《赫尔辛基宣言》（Helsinki Declaration）的知情同意。为了控制被试整体的英语水平，研究记录了他们的语言成绩。被试都参加了大学英语六级考试，且分数均在550分以上。所有被试都获得了等额的现金报酬，并书面同意了当地伦理委员会批准的实验方案。

4.3 实验流程

除了实验任务以外，实验二的实验流程与实验一的实验流程大体相似。在实验二中，被试被要求执行书写形式判断任务，即判断连续呈现的英语词对（启动词和目标词）中是否有且只有一个英语单词中有两个元音字母相邻。

4.4 数据采集与分析

实验二的数据采集与分析方法与实验一一致。

4.5 实验结果

4.5.1 行为数据结果

实验二记录了被试的正确率和反应时间。S+R+的平均正确率为90.1%（SD = 1.2%）。S+R-的平均正确率为94.6%（SD = 1.2%）。S-R+的平均正确率为95.9%（SD = 1.2%）。S-R-的平均正确率92.8%（SD = 1.1%）。每种条件的平均正确率均在90% 以上，保证了数据的有效性。反应时间从第二个英语单词（目标词）出现的时候开始记录。S+R+的反

应时间为1159 ms（SD = 88）。S+R-的反应时间为1169 ms（SD = 102）。S-R+的反应时间为1087 ms（SD = 87）。S-R-的反应时间为1118 ms（SD = 80）。双因素（语义相关性×汉语对译词首字重复情况）重复测量方差分析的统计结果显示，只有语义相关性的主效应显著[$F(1,19) = 9.141, p = 0.007, \eta_p^2 = 0.325$]。

4.5.2 脑电数据结果

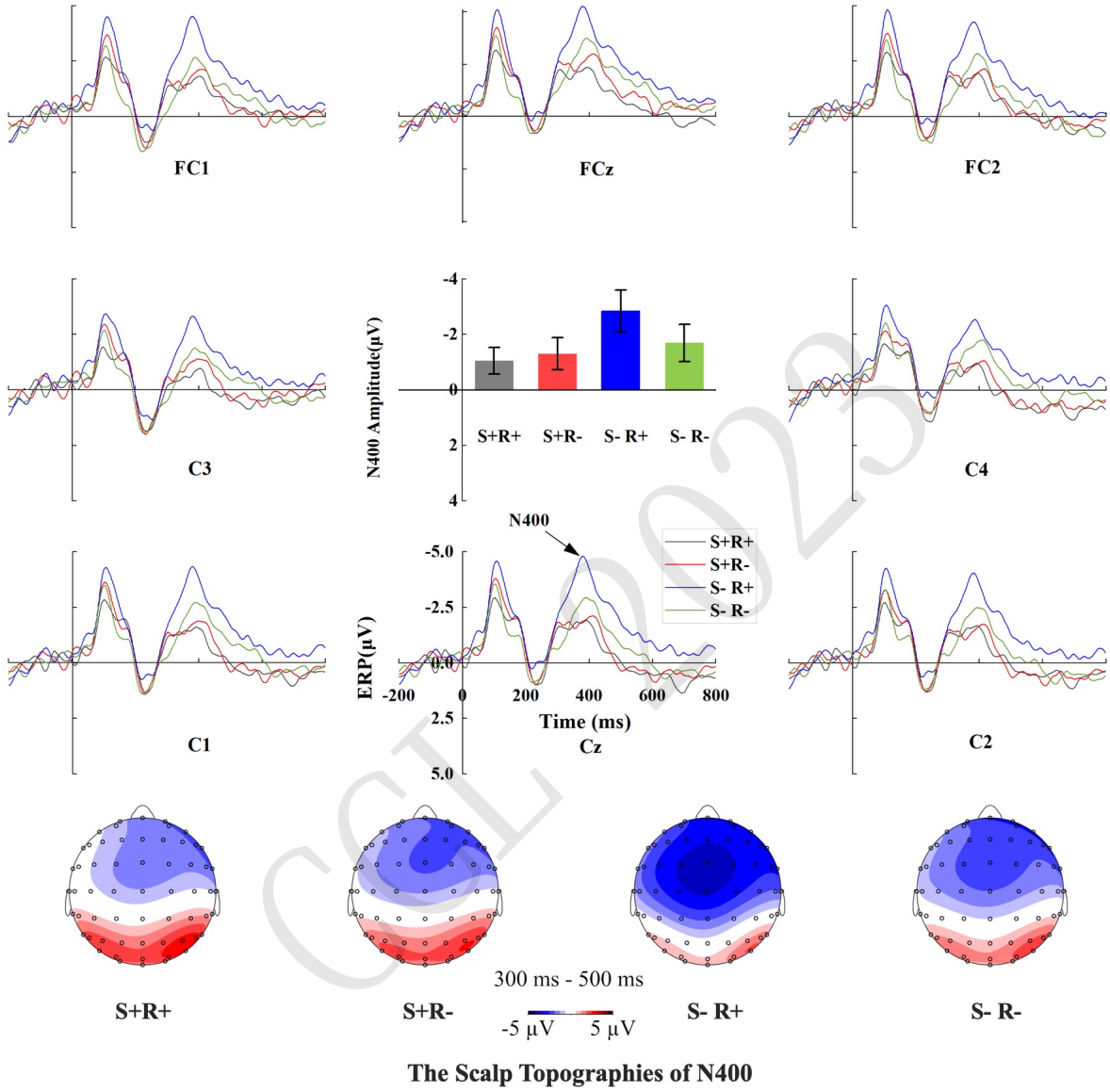


图 4: 实验二脑电结果图

实验二8个电极点各自的N400波幅（ERP）、N400平均波幅（N400 Amplitude）和脑电地形图（The Scalp Topographies of N400）如图4所示。分析结果显示，语义相关性与汉语对译词首字重复情况的交互作用显著[$F(1,19) = 4.594, p = 0.045, \eta_p^2 = 0.195$]，语义相关性的主效应也显著[$F(1,19) = 6.901, p = 0.017, \eta_p^2 = 0.266$]。在对译词首字重复的情况下，语义相关性差异显著[$F(1,19) = 9.330, p = 0.007$]，其中，语义无关词对引发的N400比语义相关词对引发的N400更负。然而，在对译词首字不重复的情况下，语义相关性差异则不显著[$F(1,29) = 0.67, p = 0.423$]。

4.6 讨论

实验二邀请了中英二语学习者作为被试，让他们运用本研究设计的书写形式判断任务来判断前后出现的英语词对中的元音相邻情况。每种条件的正确率均在90% 以上，保证了数据的有效性。其中，在行为实验结果中出现了语义相关性的主效应。但是，在脑电结果中，语义相关性与汉语对译词首字重复情况产生了显著的交互效应。并且，脑电结果显示，对译词的不同情况对语义启动产生显著的影响。具体而言，在对译词首字不重复的情况下，反映在N400波幅上的语义相关性差异不显著。而在对译词首字重复的情况下，反映语义相关性的N400波幅则呈现出了显著的差异。鉴于对译词这一隐性因素给语义启动带来的显著影响，本研究推测，实验二中的被试在二语模式下对L2词汇执行书写形式判断任务时，可能首先激活了L2词汇在L1中的对译词，而这正是经由Path-2（L2->L1）的激活路径。此外，当汉语对译词首字不重复时，该任务对语义启动的抑制效果更加显著。

此外，通过对比对译词首字重复条件下的不同情况不难发现，在在对译词首字重复的情况下，被试可能激活了共享的概念表征。但是，对概念表征的激活路径，是从二语词汇表征直接到概念表征，还是从二语词汇表征到母语词汇表征再到概念表征，目前尚未可知。但值得关注的是，在对译词首字不重复的条件中，被试却没有任何迹象表明其激活了概念表征。对比以上两种情况，其变量不是二语词汇而是母语词汇。因此，本研究从推测，在对译词首字重复的条件中，当被试在没有激活母语的词汇表征时，可能与对译词首字不重复的条件一样，也没有激活概念表征。然而，当被试激活了母语的词汇表征之后，觉察到了对译词首字重复的特殊存在，可能在此之后继续激活了概念表征。

为了解释实验中出现的再激活现象，本研究引入(Collins and Elizabeth, 1975)描述的语义加工的扩散-激活理论进行分析。该理论认为，大脑中词汇记忆可以分为概念网络和词汇网络（包括词汇的书写形式），并对激活过程进行了如下描述：当某个词汇被大脑处理时，它在大脑中的激活则以递减的梯度沿着网络路径扩散。这种激活正如来自一个源的信号，它在向外扩散过程中会逐渐衰减。值得注意的是，如果外界的刺激是一个变量，交叉的激活则需要触发到一个具体的阈值。如果对不同来源的刺激求和，当交叉点的总和达到阈值时，大脑则可能重新评估网络中产生交叉点的路径，再进行激活。基于此，本研究推测，在实验过程中，之所以在对译词首字重复的条件下出现了词汇语义的再激活，可能是因为对译词首字重复这个条件对于被试而言是在激活过程中发现的新变量，这个的新变量出现刺激了大脑重新评估网络中产生交叉点的路径，从而进行再激活。

综上所述，本研究在实验二中发现，当被试在识别词汇过程中执行书写形式判断任务时，可能经由Path-2（L2->L1）激活了母语词汇表征。

5 总结

本研究运用ERP技术在二语模式下观测到了二语学习者从L2词汇到L1对译词两条路径的激活情况。在实验一中，被试执行的是英语语义判断任务。在实验二中，被试执行的是英语书写形式判断任务。结果显示，在两种任务下，被试都可能激活母语词汇的对译词。此外，实验一的结果证明了激活路径Path-1（L2->L1）的存在；而实验二的结果则表明激活路径Path-2（L2->L1）的存在。该研究结果弥补了BIA+和BIMOLA在描述词汇识别时在跨语言词汇激活路径方面的缺失，给词汇识别过程中跨语言词汇激活路径的模型结构提供了新的证据。并且，研究发现，词汇识别过程中从L2词汇表征到L1词汇表征的激活路径与RHM描绘的词汇产出过程的激活路径类似。据此，本研究推测，尽管大脑在词汇识别和词汇产生过程中采用不同的处理机制，但在跨语言词汇激活过程中，它们依然存在某些共通之处。这种共通之处是否与心理词典和双语心理词典在大脑中的储存方式存在内在的联系，还有待进一步探索。总体而言，本研究的实验结果也进一步印证了二语激活母语的跨语言现象(Degani and Tokowicz, 2013; Mishra and Singh, 2016; Costa et al., 2017)。

近期，Dijkstra (2018) 构建了一个跨语言词汇激活的计算模型—— Multilink。作为跨语言研究领域中语言认知与计算结合的重要成果，该模型集成了RHM和BIA+的基本理论，在单语和双语词汇决策、单词命名和单词翻译生成等任务中模拟不同长度和频率的同源词和非同源词的识别和生成。未来，囊括跨语言词汇激活路径的理论模型亦有望被开发成计算模型，将诸如此类的模型理论转化为具体可实现的流程。

参考文献

Collins A M. and Elizabeth F L . 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Costa A. , Pannunzi M. , Deco G. , and Pickering M J. . 2017. Do bilinguals automatically activate their native language when they are not using it? *Cognitive Science*, 41(6):1629–1644.

Degani T. and Tokowicz N . 2013. Cross-language influences: translation status affects intraword sense relatedness. *Memory and Cognition*, 41(7):1046–1064.

Dijkstra T. and van Heuven W J B. . 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3):175–197.

Dijkstra T. O. N. , Wahl Alexander , Buytenhuijs Franka , Van Halem Nino , Al-Jibouri Zina , De Korte Marcel , and RekkÉ Steven . 2018. Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22(04):657–679.

Friesen D C. , Oh J. , and Bialystok E . 2016. Phonologically-mediated meaning activation in monolinguals and bilinguals evidence from homophone effects in erp. *Linguistic Approaches to Bilingualism*, 6(3):262–289.

Grosjean François . 1988. Exploring the recognition of guest words in bilingual speech. *Language and Cognitive Processes*, 3(3):233–274.

Jia X. , Wang S. , Zhang B. , and Zhang J. X. . 2013. Electrophysiological evidence for relation information activation in chinese compound word comprehension. *Neuropsychologia*, 51(7):1296–301.

Kroll J F. and Stewart E . 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2):149–174.

Laszlo S. and Federmeier K. D. . 2011. The n400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, 48(2):176–186.

Lau E. F. , Gramfort A. , Hamalainen M. S. , and Kuperberg G. R. . 2013. Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. *Journal of Neuroscience*, 33(43):17174–81.

Li , Mo L. , Wang L. , Luo R M. , and Y. X . 2009. Evidence for long-term cross-language repetition priming in low fluency chinese–english bilinguals. *Bilingualism: Language and Cognition*, 12(1):13–21.

Liang Lijuan and Chen Baoguo . 2014. Processing morphologically complex words in second-language learners: The effect of proficiency. *Acta Psychologica*, 150:69–79.

Mari-Beffa P. , Valdes B. , Cullen D. J. , Catena A. , and Houghton G. . 2005. Erp analyses of task effects on semantic processing from words. *Cognitive Brain Research*, 23(2-3):293–305.

Mishra Ramesh Kumar and Singh Niharika . 2016. The influence of second language proficiency on bilingual parallel language activation in hindi-english bilinguals. *Journal of Cognitive Psychology*, 28(4):396–411.

Oldfield Richard C . 1971. The assessment and analysis of handedness: The edinburgh inventory - sciencedirect. *Neuropsychologia*, 9(1):97–113.

Perea M. , Labusch M. , and Marcet A. . 2021. How are words with diacritical vowels represented in the mental lexicon? evidence from spanish and german. *Language, Cognition and Neuroscience*, pages 457–468.

Picton T.W. , Bentin S. , Berg P. , Donchin E. , Hillyard S.A. , Johnson JR. R. , Miller G.A. , Ritter W. , Ruchkin D.S. , Rugg M.D. , and Taylor M.J. . 2000. Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, 37(2):127–152.

Silva-Pereyra J. , Harmony T. , Villanueva G. , Fernández T. , Rodrıguez M. , Galán L. , Dıaz-Comas Lourdes. , Jorge B. , Fernández-Bouzas A. , Marosi E. , and Reyes A . 1999. N400 and lexical decisions: automatic or controlled processing? *Clinical Neurophysiology*, 110(5):813–824.

Spruyt Adriaan , Houwer Jan De , and Hermans Dirk . 2009. Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory and Language*, 61(1):37–54.

Thierry G. and Wu Y. J. . 2007. Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12530–12535.

Tiedt H. O. , Ehlen F. , and Klostermann F. . 2020. Age-related dissociation of n400 effect and lexical priming. *Scientific Reports*, 10:e20291.

Valdes B. , Catena A. , and Mari-Beffa P. . 2005. Automatic and controlled semantic processing: a masked prime-task effect. *Conscious Cogn*, 14(2):278–95.

van Heuven W. J. , Mandera P. , Keuleers E. , and Brysbaert M. . 2014. Subtlex-uk: a new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, 67(6):1176–90.

Wilkins D A. . 1972. *Linguistics in language teaching*. Edward Arnold, London.

Wu Y. J. and Thierry G. . 2010. Chinese-english bilinguals reading english hear chinese. *Journal of Neuroscience*, 30(22):7646–51.

Wu Y. J. and Thierry G. . 2012. Unconscious translation during incidental foreign language processing. *Neuroimage*, 59(4):3468–73.

Yang S Q. , Cai Y Y. , Xie W. , and Jiang M H . 2021a. Semantic and syntactic processing during comprehension: Erp evidence from chinese qing structure. *Frontiers in Human Neuroscience*, 15:e701923.

Yang S Q. , Zhang X C. , and Jiang M H. . 2021b. Bilingual brains learn to use l2 alliterations covertly like poets: Brain erp evidence. *Frontiers in Psychology*, 12:e691846.

Ye Z. , Luo Y J. , Friederici A D. , and Zhou X. . 2006. Semantic and syntactic processing in chinese sentence comprehension: evidence from event-related potentials. *Brain Research*, 1071(1):186–96.

Zeelenberg René and Pecher Diane . 2003. Evidence for long-term cross-language repetition priming in conceptual implicit memory tasks. *Journal of Memory and Language*, 49(1):80–94.

李荣宝，彭聃龄，郭桃梅. 2003. 汉英语义通达过程的事件相关电位研究. 心理学报, 35(3):309–316.

肖巍. 2018. 二语词汇加工中的一语自动激活：来自中国英语学习者的证据. 清华大学出版社, 北京.

肖巍，倪传斌. 2016. 二语词汇加工中的一语自动激活:来自中国英语学习者的证据. 外语教学与研究, 48(2):236–248.

汉语语义构词的资源建设与计算评估

王悦[1,2], 刘扬[1,2]*, 梁启亮[1,3], 王涵思[1,2]
[1]北京大学计算语言学教育部重点实验室，北京100871
[2]北京大学计算机学院，北京100871
[3]北京大学信息科学技术学院，北京100871
{wyy209,liuyang}@pku.edu.cn
lql_eecs@qq.com; whs1900014165@163.com

摘要

汉语是一种意合型语言，汉语中语素的构词方式与规律是描述、理解词义的重要因素。关于语素构词的方式，语言学界有语法构词与语义构词这两种观点，其中，语义构词对语素间关系的表达更为深入。本文采取语义构词的路线，基于语言学视角，考虑汉语构词特点，提出了一套面向计算的语义构词结构体系，通过随机森林自动标注与人工校验相结合的方式，构建汉语语义构词知识库，并在词义生成的任务上对该资源进行计算评估。实验取得了良好的结果，基于语义构词知识库的词义生成BLEU值达25.07，较此前的语法构词提升了3.17%，初步验证了这种知识表示方法的有效性。该知识表示方法与资源建设将为人文领域和信息处理等多方面的应用提供新的思路与方案。

关键词： 汉语语素；汉语语义构词；资源建设；词义生成

Construction of Chinese Semantic Word-Formation and its Computing Applications

Yue Wang[1,2], Yang Liu[1,2]*, Qiliang Liang[1,3], Hansi Wang[1,2]
[1]Key Lab of Computational Linguistics (MOE), Peking University, Beijing 100871
[2]School of Computer Science, Peking University, Beijing 100871
[3]School of Electronics Engineering and Computer Science, Peking University, Beijing 100871
{wyy209,liuyang}@pku.edu.cn
lql_eecs@qq.com; whs1900014165@163.com

Abstract

Chinese is a paratactic language, where the ways and rules of its word-formation play an important role in describing and understanding the meanings of words. There are two perspectives on morphemes and word-formation in linguistics: grammatical word-formation and semantic word-formation, with the latter indicating a deeper relationship between morphemes. In this paper, following the perspective of semantic word-formation, we propose a set of computing-oriented semantic word-formation labels based on characteristics of Chinese, build a Chinese semantic word-formation knowledge-base by combining random forest automatic labeling and manual verification, and evaluate the resource on the task of definition generation. Experimental results show that definitions generated from the semantic word-formation knowledge-base achieve a BLEU value of 25.07, which is 3.17% higher than previous grammatical

*通讯作者
基金项目：国家自然基金项目（62036001）、国家社科基金项目（18ZDA295）

word-formation approach. These findings confirm the effectiveness of our knowledge representation and resource construction, which may provide new insights into and solutions for a variety of tasks in humanities and computing applications.

Keywords: Chinese morphemes , Chinese semantic word-formation , Resource construction , Definition generation

1 引言

汉语是一种意合语言，普遍认为，汉语遵循从语素、词、短语到句子的层级结构。其中，语素作为汉语中最小的音义结合体，是构词的基本单位(朱德熙, 1982; 尹斌庸, 1984)。汉语中的语素在构词中非常活跃，处于重要的地位(徐枢, 1990)，语素义的组合在一定程度上能体现词义(符淮青, 1981)。汉语的词汇系统庞大，且不断产生新词、新义，但作为构词基本单元的语素，其数量与意义是相对稳定的，高达93%的汉语语素均为单音节语素，且87%以上的语素在构词时会保持意义不变(苑春法与黄昌宁, 1998)。而且，相对于句法分析和理解方面的工作(Xue and Palmer, 2003; You and Liu, 2005)，当前，词法分析方面的研究和资源建设还比较欠缺。在NLP领域，此前对词义进行表征的主流方法是利用词间信息(Zheng et al., 2013; Gui et al., 2018)，即通过目标词的上下文环境对词义进行表征。注意到语素构词的重要性和相关研究与开发的欠缺，我们希望将语言学观点引入到计算中，从语义构词的角度对汉语的词义进行表征。

关于语素构词的方式，在语言学界，主要有语法构词和语义构词这两种不同观点。

语法构词的观点认为，汉语构词的原则和汉语造句的原则基本一致，可以用主谓、动宾等句法结构标签对构词结构进行分类(陆志韦, 1964; 郭绍虞, 1979; 赵元任, 1980)。董秀芳(2011)认为，现代汉语中二字词的前身是古代汉语中单字词的自由句法组合，考虑形式上的构造性，也支持语法构词的看法。语法构词的结构体系较为简单，便于标注及计算处理(傅爱平, 2003)，因此，此前对语素构词的研究更多地关注语法构词(康司辰等, 2020)。

语义构词的观点认为,字与字之间是按语义关系组成词(刘叔新, 1990)，使用施事、受事等语义标签来分析构词语素之间的关系(朱彦, 2003)。周荐(2003)指出，在语素组合时，起决定作用的因素是它们是否能在意义、习惯上相搭配。相对而言，语义构词更符合汉语社团的思维模式(徐通锵, 1997)，对词内语素间关系的描述更为深刻(朱彦, 2003)，具有天然优势，采取该观点研究语素构词会带来一些额外的收益。但语义构词的结构较为复杂，且目前尚没有相对清晰的标准(傅爱平, 2003)，因此，当前对语义构词知识表示的研究还比较欠缺，从该角度研究语素构词也是一项有挑战性的工作。

注意到这些情况，我们希望在语法构词结构的基础上，提出一套新的语义构词结构体系；从语素出发，通过自动标注与人工校验相结合的方式构建语义构词知识库；并通过词义生成任务，对该知识表示的有效性进行计算评估。语义构词的体系规范、标注方法与资源建设，将对人文领域和信息处理等多方面的应用提供新的思路与方法。

本文结构如下：在引言中，介绍了语素构词的重要性和不同的路径方式；在第2节中，从汉语语义构词的理论探讨和相关语言资源研发等角度出发，对前人的相关工作做梳理和评述；在第3节中，结合汉语语素构词的特点，提出了一套面向理解和计算的汉语语义构词结构标签；在第4节中，基于上述标签集，对汉语中的二字词通过随机森林算法做结构标签的自动标注并进行人工校对，构建汉语语义构词知识库；在第5节中，在词义生成任务上对该知识表示的有效性进行计算评估，分析实验结果，并在新词数据集上做进一步的验证；在结语部分，总结本文方法与数据成果，并展望后续可能的研究领域与方向。

2 相关工作

2.1 汉语语义构词理论探讨

汉语语义构词关注的是构词语素之间的关系。在当代词汇语义学的研究中，生成词库理论(Pustejovsky, 1995)受到了极大的关注，被誉为是精细的形式化分析手段(Geeraerts, 2010)。基于该理论，词项的表征包括论元结构、事件结构、物性结构和词汇类型结构这四个层面。其中，论元结构为概念整合网络中的动词心理空间提供明晰的语义框架，物性结构为名词心理空间提供明晰的知识框架(王笑, 2017)。

论元结构采用语义角色（也称论元角色）来实现描述。它多以谓词性成分为核心，描述与中心谓词相关的语言成分在事件中所扮演的参与者角色，是语言学家对句子中有关结构成分之间意义关系的一种分类方式，可以为计算机处理语言提供较为充分的语义知识。但从经验上来说，想要系统、一致地给所有动词的全部配项标明语义格几乎是不可能的(Dowty, 1991)，因此，语义角色标签的划分至今仍然没有相对明晰的标准。现有的汉语语义角色标签划分方式有多种，它们多关注于句子级语义分析:袁毓林基于计算需求(2002)提出一种了包含4个层级、17种论元角色的结构体系，并描述了这些语义角色的定义和句法特征;刘茂福和胡慧君(2013)在袁毓林标签的基础上做了归并调整;鲁川(2010)将语义角色分为中枢语义角色和周边语义角色，并提出26种中枢语义角色和26种周边语义角色;宋衡等(2023)在鲁川研究的基础上，对周边语义角色做了一些调整，并将其分为主要周边语义角色和辅助周边语义角色两类，提出了一套包含60种语义角色的分类体系。在构词方面，顾阳和沈阳(2001)将论元结构引入构词理论研究中，朱彦(2003)利用论元结构，深入语义底层对4263个复合词进行分析研究。但由于构词与造句的差异，即词结构短小、无法容纳多个（或内部结构复杂的）论元，且词的上下文语境有不确定性，句子级的语义角色体系难以直接迁移到构词上。另外，汉语中的一些词不包含谓词性成分，无法用论元结构描述，单纯的论元结构在词内语义结构分析中带有一定的局限性。

物性结构是关于词汇本体知识的描述体系。与论元结构不同，它以名词为核心，使用各种物性角色从多个方面说明名词与其相关事物、事件和属性的关系。现有的物性角色标签划分方式有多种：Pustejovsky(1995)提出形式角色、构成角色、施成角色、功用角色这四种基础物性角色，后来又补充提出了规约化属性(Pustejovsky and Jezek, 2008)。袁毓林(2013)根据汉语名词在真实文本中的搭配情况，提出了10种物性角色；张念歆和宋作艳(2015)在研究形名组合定中式复合词时，在基础的四类物性角色上细分了小类；宋作艳等在研究名名(宋作艳等, 2015)、动名(宋作艳, 2022)复合词的物性结构时，在四种物性角色的基础上将规约化属性引入汉语，并在五个物性角色大类下细分小类。由于前人的研究通常只局限在特定的词性结构上，物性结构标签的划分至今同样没有相对明晰的标准，且这些研究多停留在理论语言学领域，面向计算应用的研究还比较匮乏。与论元结构的缺陷类似，汉语中的一些词也不包含体词性中心，无法用物性结构描述，单纯的物性结构在词内语义结构分析中也带有一定的局限性。

在词法分析中，还有一些其它结构不能被论元结构或物性结构覆盖。其中，包括属于单纯词、前（后）缀式和并列式等结构的词。前两类的词中至少有一个语素不表义，因此找不出论元或物性角色，后者则是因为两个语素的地位平等，找不到中心语素，所以也无法使用论元结构或物性结构来描述。在名+名联合式复合词上，王晗(2013)研究了《现代汉语词典》中的1130个此类复合词，并按前、后语素的意义关系把它们分为同义联合、相关联合和反义联合这三类，按语素和词的关系分为重合类、组合类、融合类这三类。刘明珠(2020)在同义联合、相关联合和反义联合这三类的基础上，补充了远义并列类，在语素和词的关系上则分为互注、选取、融合、合取这四类。而对于动+动的连动式，李可胜和满海霞(2013)提出了毗邻、聚合和加合这三种事件结构。

2.2 汉语构词相关语言资源研发

从中文信息处理的实践来看，此前的语义分析多集中在句子级别，对语素构词的研究还比较欠缺。在语言知识工程方面，目前具有较大影响的几项典型资源如下：

苑春法和黄昌宁(1998)的“汉语语素数据库”以语素描写和构词分析为核心，覆盖了6763个常用汉字的17470个语素项信息，包括语素义、语法类、构词方式等信息，并对这些语素项构成的43097个二字复合词进行了构词结构分析和语素项的绑定，并初步总结了汉语语素构词的规律。但在语素项上，仅仅形成了一个离散的集合而没有形成关联体系，缺乏面向整个语言系统的意义关联，难以满足现实的计算需求。

亢世勇(2004)的“汉语义类信息库”覆盖了6763个常见汉字的17430个字位（可理解为语素）的释义和词性，并与《同义词词林》（以下简称《词林》）中的语义分类体系进行了绑定。在此基础上，继续对52366个二字词中的每个字进行义类标注和简单释义，建立了“汉语语义构词信息库”。这两项工作对字位和二字词进行了归类并形成了积极的意义关联，其归类以现有《词林》为标准，存在着语素义与词义的因果参照问题，结构的合理性有待商榷。

吉志薇和冯敏萱(2015)提取了《现代汉语词典》（以下简称《现汉》）中的2268个词素并标注了每个词素所属的义类，构建了“词素-义类数据库”。在此基础上，标注和统计了8984个二字词的词素意义和词素间的词化意义（可理解为语素义和构词结构），构建了二字词语义描写

体系，并应用于二字未登录词的理解。目前其收录的二字词均包含前50高频词素，采样不均衡且数据规模过小，难以满足全局数据上的计算需求。

刘扬等(2018)研究开发了“汉语概念词典”，提取和编码《现汉》中全部8514个汉字的20855个语素释义，每个语素义具有一个唯一的编码，如“雄$_{1_05_01}$”代表“雄”字的某个语素释义；以这些全局信息为依据，进而采用“同义语素集”来表征“语素概念”，建立了“语素概念体系”。在此基础上，对《现汉》中41474个二字词的全部52108个义项赋予唯一的词条编码，进一步描述这些词的语法构词结构，实现了语法结构下的语素与其语素义的严格绑定，以此来诱导和表达汉语词义(陈龙等, 2019)，并在词义生成等任务上实现了应用(康司辰等, 2020; Zheng et al., 2021a; Zheng et al., 2021b)。其构词结构标签共分16种，分别为定中、联合、述宾、状中、单纯、连谓、后缀、述补、主谓、重叠、方位、介宾、名量、数量、前缀与复量(郑画等, 2022)。“汉语概念词典”注重表征构词语素与语素义的绑定，采取的是语法构词而并非语义构词，这些系统性的工作也为进一步提炼和标注语义构词标签提供了条件和可能。

3 面向计算的汉语语义构词结构体系

基于语义构词的技术路线和计算需求的现实考虑，我们首先从单音节语素构成的现代汉语二字词入手，在此前工作的基础上，构建了一套包含论元角色、物性角色和其它标签的汉语语义构词结构体系。其具体情况如下：

3.1 汉语语义构词的论元结构

对于谓词性中心的汉语二字词，在袁毓林(2002; 2013)等研究的基础上，考虑到二字词长度过短无法同时出现多个（或内部结构复杂的）论元，以及尽量消除在缺乏上下文情况下可能出现的结构歧义，我们归并了受事与对象等形式相近的论元结构，并加入了内容与事量，提出了一套包含14种论元角色的论元角色结构体系。其定义如表1所示：

论元角色	论元角色描述	示例
施事	自主性动作、行为的施行者	**人**造　**雷**鸣
感事	非自主性心理感觉的主体	**头**疼　**心**酸
主事	性质、状态或变化性事件的主体	**年**轻　**身**故
受事	动作、行为的承受者	吃**饭**　**皮**试
结果	动作、事件造成的影响、产物	扩**大**　录**像**
系事	在事件中与主事对应，表达主事的属性、类别	有**名**　当**官**
内容	言语行为、信息传递或心理活动的内容	讲**课**　求**婚**
工具	动作、行为所凭借的工具	**枪**毙　**珠**算
材料	动作、行为所用的材料，事件过程中所凭借和消耗的物品	**水**解　**铁**打
方式	动作、行为所使用的方式、方法	**周**游　**上**调
原因	动作、行为发生的原因	**仇**杀　**惊**醒
时间	动作、行为所发生的时间	**日**用　**春**训
空间	动作、行为所发生的空间	**家**访　**野**营
事量	事件所涉及的数量、频率、幅度	**多**疑　三**思**

Table 1: 论元角色与示例（示例中字体加粗部分为论元角色）

3.2 汉语语义构词的物性结构

对于体词性中心的汉语二字词，此前研究多局限于一种词性的组合。我们综合魏雪和袁毓林(2013)、宋作艳(2022)、张念歆和宋作艳(2015)等在名名、动名、形名组合方面的研究，结合汉语语素构词的特点，提出了一套包含17种物性角色的物性角色结构体系。其定义如表2所示：

3.3 汉语语义构词的其它结构

论元角色与物性角色适用于构词的前、后语素都表义且具有明确中心的情况，显然，它们无法覆盖全部的汉语二字词。对于前、后语素地位均等或至少一个语素不表义的汉语二字词，我们另设了单独的语义标签。其定义如表3所示：

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

物性角色	物性角色描述	示例
材料	物性角色是制成核心名词的材料	**木**板　**纸**钱
数目	物性角色是核心名词表示事物的数量	**单**亲　**七**彩
整体	物性角色是核心名词所属的整体	**果**皮　**羊**毛
领属	物性角色是核心名词的领主	**沙**俄　**人**情
成分	物性角色是核心名词的组成成分	**雨**点　**字**幕
上位	物性角色是核心名词的上位概念，核心名词是其中的一种	鲤**鱼**　氧**气**
外形	物性角色是核心名词的外在表现	蒜**黄**　**方**阵
评价	物性角色是对核心名词的主观评价	**真**品　**奸**商
单位	物性角色是作为核心名词的单位的量词	云**朵**　马**匹**
时间	物性角色是核心名词所表示事物所处的时间	**唐**诗　**早**饭
空间	物性角色是核心名词所表示事物所处的空间	**山**寨　**壁**画
方位	物性角色是方位词，整体词义指核心名词的某个方位	江**南**　**后**面
用途	物性角色是核心名词所表示事物用来做的事情	**玩**具　**鱼**网
用法	物性角色是使用核心名词所表示事物的方法	**挂**钟　**吊**灯
职能	物性角色是核心名词所表示事物（多为人）所从事的工作	**农**民　**教**师
施成	物性角色是核心动词所表示是事物的产生方式	**烤**鸭　**配**方
状态	物性角色是核心名词所表示事物的所处状态或常规活动	**沸**水　**飞**鸟

Table 2: 物性角色与示例（示例中字体加粗部分为物性角色）

其它结构	结构描述	示例
重合	词义与前、后语素义均相同或相近	错误　奶奶
组合	词义与前、后语素义均有关，前、后语素义不同且地位平等	左右　花草
顺序	词义与前、后语素义均有关，前、后语素是先后发生的动作或事件	签收　判断
偏义	词义仅与其中一个语素义有关	老**师**　**灿**烂
单纯	这个词是单独的语素	葡萄　端木

Table 3: 其它语义结构与用例（偏义结构示例中字体加粗部分为中心语素）

对于论元结构和物性结构无法覆盖汉语二字词的情况，第一种可能是，前、后语素均表义且地位相等，在语法结构上，这些词通常为并列结构类型。在语义结构上，则根据词义与语素义、前语素义与后语素义之间的依赖关系，可以进一步把它们分为重合、组合和顺序结构；第二种可能是，二字词中有且仅有一个语素与词义关系较强，这类词既包含“老师”“兔子”这样前（后）缀式的词，也包含因一个语素义脱落导致词义偏向另一个语素义的联合式的词，如“国家”“灿烂”等；第三种可能是，词义与前、后语素义均没有关联，比如一些单纯词，包括“沙发”“端木”“葡萄”等。在语言学中，通常把这些词视为独立语素。我们重点关注在现代汉语中占绝大多数的单字语素，为了计算上的形式一致性，对于这些单纯词，视构成该词的前、后语素为空语素，标注单纯结构。

4　汉语语义构词的资源建设

4.1　面向计算的汉语语义构词知识表示

表4展示了“汉语概念词典”中既有的语法构词描述信息。其中，对于多义词的不同义项，视为不同的词条分别标注。示例中的前（后）语素义指该词中前（后）语素的语义，用构词语素和语素义的绑定来做表达。我们希望从这些信息入手，通过自动标注与人工校验相结合的方式构建汉语语义构词知识库。王洪君(2000)曾指出，现代汉语中绝大多数的双音节复合词可以用句法结构的形式理解，因此，语法结构信息的应用有助于达成语义结构的识别，对语义结构的自动标注有极大的帮助。我们要开展的资源建设工作是将既有的语法结构信息拓展为语义结构信息，该信息包括语义构词结构和中心语素位置，其中，语义构词结构指前、后语素之间的语义关系，中心语素位置指在词义表征中占核心地位的语素位置。

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

词	上天$_1$	上天$_2$	上天$_3$
词义	上升到天空	用作婉辞，指人死亡	迷信者指主宰自然和人类的天
语法结构	述宾	述宾	定中
前语素	上$_{2_14_01}$	上$_{2_14_02}$	上$_{1_06_01}$
前语素义	由低处到高处	到；去	位置在高处的
后语素	天$_{1_12_01}$	天$_{1_12_11}$	天$_{1_12_10}$
后语素义	天空	迷信者指神佛仙人所住的地方	迷信者指自然界的主宰者
例句	人造卫星～	（无）	～保佑

Table 4: “汉语概念词典”中既有的语法构词描述信息

我们采用自动标注与人工校验相结合的方式对每个词的语义构词结构进行标注。标注工作由三名标注人员完成，初始抽取6000个词条，由两名标注人员独立标注语义构词结构和中心语素位置；依据这些标注信息，采用随机森林对“汉语概念词典”中的其余二字词的语义构词结构和中心语素位置进行自动标注；最后，由第三人对全部标注结果进行人工校验，作为最终的标注结果。

4.2 基于CART决策树和随机森林的语义构词结构自动标注

由于人工标注难以实现较大的样本覆盖，考虑到语义构词结构的分类多达33类，部分结构存在样本偏少、难以有效覆盖的问题，这种情况下，主流的深度学习算法无法有效捕捉这些特征。借鉴魏雪和袁毓林(2013)等对名名组合词构建释义模板的工作思路，该方法类似于人工构建决策树，且有较好的可解释性，因此，我们采用CART决策树和随机森林算法对汉语语义构词结构进行初步标注。

决策树的分类特征为：$features = \{fm, mor_1, mor_2, sim\}$，这是为每个待标词输入决策树的特征集。在该特征集中，1)fm为语法构词结构；2)$mor_i = \{m_{i1}, m_{i2}, \ldots, m_{i11}(i = 1, 2)\}$为第$i$个语素的嵌入向量表示。我们使用语素概念体系层级结构中的语素概念路径信息对每个语素进行向量表征：m_{ij}表示第i个语素在该体系的第j层下的子节点序号；3)为了衡量不同语素对词义贡献的大小，增加相似度信息$sim = \{s_{1*}, s_{2*}, s_{12}, \frac{s_{1*}}{s_{2*}}\}$。其中，$s_{1*}, s_{2*}, s_{12}$分别表示前语素与词、后语素与词、前语素与后语素之间的相似度，$\frac{s_{1*}}{s_{2*}}$则是用于衡量前、后语素对词义贡献的比重，这主要是考虑不同的语义结构对语素义贡献的侧重不同。比如“猪獾”和“鲤鱼”，它们的语法结构均为定中，前、后语素都属于动物类，但语义结构明显不同，“猪獾”是“外形像猪的獾”，侧重于后语素，而“鲤鱼”是“品种属于鲤的鱼”，侧重于前语素。注意到决策树算法每次只选择一个特征进行划分的特点，因此，把前、后语素对词义贡献的比重单独作为一个维度，便于决策树提取这一特征。

我们将语法构词结构、语素嵌入向量和相似度信息拼接起来，作为每个待分类词的特征集输入决策树。为了避免过学习，另进行了决策树剪枝和语素嵌入向量降维。决策树剪枝涉及树的最大深度、每个节点的最小样本数等因素，并将最大深度和最小叶节点样本数均设为9；语素嵌入向量降维则实验性截取每个语素嵌入向量的前n个维度，这主要考虑语素概念体系中偏底层节点的粒度过细，特征不明显，易对计算造成干扰。我们集成采用随机分类特征生成的100棵CART决策树，构成随机森林，随机森林仅对语素嵌入向量降维、不做剪枝。然后将决策树算法和随机森林算法的分类结果进行对比，以此选择最优模型。为了充分利用数据，使用10折交叉验证来衡量分类的准确率。

我们用上述算法对汉语的语义构词结构进行自动标注的分类，分类结果如表5所示，其中，语素嵌入向量为7维的随机森林效果最好，33分类下的准确率达74.26%。增大或减小向量的维数均导致准确率下降：维度较大时，向量的后端会出现大量零值，失去了语素义表征的价值；而维度较小时，意义粒度的表征过粗，无法表达相对精确的语素义。此外，为了验证语法结构信息和相似度信息在算法中的作用，我们也做了消融实验，在随机森林算法上分别去掉这些信息。结果表明，在最优的7维语素嵌入向量上，去掉语法构词结构信息，分类的准确率下降了28.86%；去掉相似度信息，分类的准确率下降了1.50%。这验证了在分类算法中采纳语法构词结构信息和相似度信息的重要性。

模型/语素嵌入向量维度	11	8	7	6
随机基准模型	3.18	3.00	2.82	3.10
决策树（不剪枝）	70.65	70.53	71.07	72.50
决策树（剪枝）	73.35	73.52	74.02	73.98
随机森林	73.52	73.72	**74.26**	73.63
随机森林w/o fm	55.92	56.00	55.40	54.80
随机森林w/o sim	72.91	73.05	72.76	72.68

Table 5: 不同向量维度下决策树算法与随机森林算法的分类准确率

4.3 汉语语义构词知识库构建

使用效果最好的随机森林算法，我们对未经人工标注的46108个二字词条进行自动标注，并在自动标注的基础上进行人工校验。此外，考虑到一些词的语义结构相同但中心语素位置不同，除了结构信息外，根据中心语素位置，我们给出了用于词义表达的语素义序列。语素义序列的生成规则为：对使用论元结构的词，将核心谓词排列在前，论元排列在后；对使用物性结构的词，将核心名词排列在前，物性角色排列在后；对偏义式的词，将与词义关联较强的语素排列在前，与词义关联较弱的语素排列在后；对并列式与单纯词，语素义依然按照在词中出现的顺序排列。表6给出了新的语义构词知识表示的示例：

词	结构	语素义序列	语素义1	语素义2
植树	受事	<植$_{1_04_01}$，树$_{1_04_01}$>	栽种	木本植物的通称
谣传$_2$	受事	<传$_{1_08_03}$，谣$_{1_03_02}$>	传播	谣言
竹器	材料	<器$_{1_05_01}$，竹$_{1_02_01}$>	器具	竹子
花束	单位	<花$_{1_18_02}$，束$_{1_05_02}$>	可供观赏的植物	用于捆在一起的东西
灿烂	偏义	<灿$_{1_01_01}$，烂$_{0_00_00}$>	光彩耀眼	<空语素>
诞生	重合	<诞$_{1_02_01}$，生$_{1_10_01}$>	诞生	生育；出生

Table 6: 语义构词知识表示示例

我们在“汉语概念词典”的基础上，将原有的语法构词结构替换为上述语义构词结构与语素义序列信息，构建了汉语语义构词知识库，知识库中的词相关信息包括：词，语义构词结构、语素义、语素义序列、词义和例句。我们的知识库涵盖了41474个二字词的52108个义项，基本实现了对《现汉》中二字词的覆盖。

5 汉语语义构词的计算评估

在语义构词思路和资源建设的有效性验证方面，Noraset(2017)提出的词义生成任务是一种评估知识表示质量的恰当且自然的方式，并有直观、良好的可解释性。该任务的目标是依据给定的词相关信息，由机器自动生成针对该词的新的释义文本，此前也被用于生成词向量的质量评估。需要指出的是，我们做语义构词结构自动标注并不使用词的释义文本信息，且最终的词义表示知识中也不直接包含词的任何释义文本信息，用词义生成任务来做计算评估是相当严苛的一项考验。

5.1 词义生成的基础模型

当前的词义生成模型依据注入特征通常分为三类：第一种是基于预训练词向量的，包括Noraset(2017)的SG模型，由于注入的特征单一，其生成效果较差，且无法区分多义词的不同义项；第二种是在词向量的基础上追加语料，包括Gadetsky(2018)基于AdaGram和注意力机制的模型和Ishiwatari(2019)的LOG-CaD模型，在词向量之外增加了上下文向量信息，该方法的有效性严重依赖上下文质量；第三种是基于知识库的算法，包括基于HowNet义原的AAM、SAAM模型(Yang et al., 2020)和基于“汉语概念词典”中语法构词结构与语素义标注

的DeFT模型(Zheng et al., 2021a)。在这种情况下，DeFT模型也是验证新的语义构词知识表示的适合的比较基准。

5.2 基于语义构词的改进模型

DeFT模型的输入序列是(w, fm, mor_1, mor_2, C)。其中，w是词形信息，fm是构词结构，mor_1与mor_2分别为前、后语素的语素义，C是例句信息。使用的特征包括如下五项：词向量w、字向量$ch = (ch_1, ch_2)$、语法结构fm、语素义向量$mor = (mor_1, mor_2)$、例句向量C。该模型中使用的语法结构是单向的，在我们构建的知识库中，语义结构是双向的：如“受事”标签既有可能表示前语素是后语素的受事，也可能表示后语素是前语素的受事，二者的支配关系由语素义序列决定。语素义序列是按照语素义的重要性排列的，并非词中的原始语素顺序，因此，在特征的注入上有两种考虑：第一种是依据模型结构调整数据格式，将语素义按照原始的语素顺序排列，这种方法简单快捷，但对少量语义标签难以判断前、后语素的语义关系，造成相关信息的混淆与损失；第二种是对模型进行改进以适应数据格式，对于语素义，按此前界定的语素义序列直接做输入，如果语素义序列中后语素在前，则对字向量信息也进行反向处理。改进后的模型结构如图1所示：

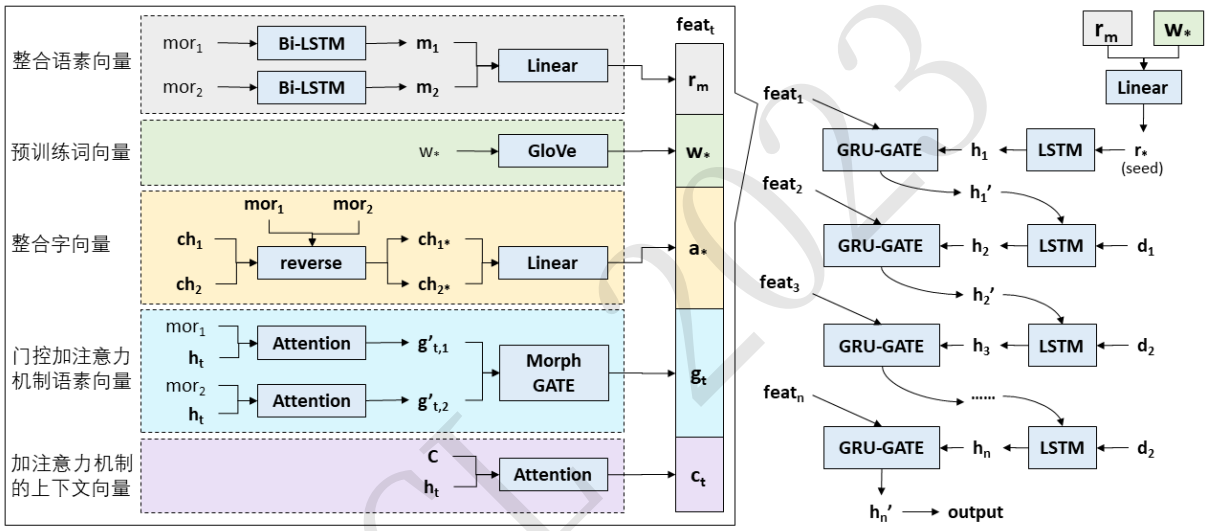


Figure 1: 改进后的DeFT+reverse模型结构图

其中，Linear是对每个构词结构表达特异性的线性层，MorphGATE通过一个线性层和一个sigmoid层对两个语素向量做权重分配，字向量和词向量使用预训练的fastText(Bojanowskiet al., 2017)。对改进后的模型，mor_1与mor_2是按语素义序列规约的语素义向量，$ch1$与$ch2$分别是词中的第一、第二个字向量，reverse函数判断字向量与语素义向量是否匹配，如不匹配则切换两个字向量的顺序。

依据“汉语概念词典”中既有的语法构词结构信息与新构建知识库中的语义构词结构及中心语素位置信息，按模型要求输入的数据格式做统一处理，我们分别建立服务于词义生成的语法构词数据集和语义构词数据集，将它们按8:1:1的比例分为训练集、测试集和验证集，并确保每个二字词在不同构词结构下出现在同样功能的集子里，且多义词的所有义项也都出现在同样功能的集子里。

在参数设置上，使用fastText词向量对词信息进行初始化，词向量的维度为300，Bi-LSTM的隐层大小为300，训练批次大小为64，随机失活率0.2。优化器使用Adam，学习率初始化为0.001。在每个epoch结束时计算验证集上的BLEU值，若出现连续6个epoch无提升则学习率乘3；为避免过学习，若连续12个epoch效果无提升则停止训练，保存验证集上BLEU值最高的作为最终模型。

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

5.3 实验结果与分析

我们分别在三种模型与输入的组合上进行词义生成实验：第一种是使用语法构词结构的DeFT，以此作为基准模型；第二种是使用语义构词结构的DeFT模型，不考虑语素义序列的规约；第三种是使用语义构词结构的DeFT+reverse改进模型，考虑语素义序列的规约。利用这三种模型和不同的构词结构信息进行训练，并在相同的测试集上进行测试。词义生成的示例如表7所示：

词	**理财**	**雨具**
词义	管理财物或财务	防雨的用具
语法结构	述宾	定中
语义结构	受事	用途
语素义序列	<理$_{1_07_04}$，财$_{1_02_01}$>	<具$_{1_03_01}$，雨$_{2_02_01}$>
前语素义	管理；办理	从云层中降向地面的水
后语素义	钱和物资的总称	用具
例句	当家～｜～之道	（无）
DeFT-语法结构	对资产的资产	可以做的用具
DeFT-语义结构	对一定的资金、资金等	防雨的用具
DeFT+reverse-语义结构	管理资金	防雨的用具

Table 7: 词义生成结果示例（中间六行为词的相关信息，后三行为藉此产生的词义生成结果）

对上述三种模型的生成结果进行分析：在“理财”一词中，语义结构为“受事”。一般而言，“受事”标签下的核心谓词位置不能完全确定，在不考虑语素义序列的规约时，仅使用“受事”标签无法区分核心谓词及其论元，难以捕捉有效的语义特征。考虑了语素义序列的规约后，生成结果得到了进一步的提升。在“雨具”一词中，语义结构为“用途”，“用途”标签下的核心名词几乎全部为后语素，语素义序列的规约对生成结果的影响较小。相比较而言，语法结构里的“定中”标签过于宽泛，该标签下词的语义结构存在极大差异，难以捕捉到有效的语义特征，导致生成的词义缺乏限制、过于笼统。

为定量评价词义生成的结果，使用自动评测指标BLEU进行评估，结果如表8所示。易见，在其它注入信息相同的情况下，仅把语法构词结构替换成语义构词结构，BLEU值就能获得1.56%的提升，即便不考虑语素义序列信息带来语义关系的混淆和损失。对模型进行改进，增加语素义序列信息，BLEU值达到25.07，较语法构词结构提升了3.17%。这些，初步验证了语义构词思路和资源建设的优势。

模型	**BLEU**
DeFT-语法结构*	24.30
DeFT-语义结构	24.68(+1.56%)
DeFT+reverse-语义结构	**25.07(+3.17%)**

Table 8: 词义生成结果评估（*为基准模型，最佳结果加粗表示）

5.4 在新词上的推广探讨

为了验证语义构词知识表示的可推广性，我们继续在新词上进行词义生成评估。新词的特点和难度在于它催生了新的词形、词义，并可能衍生出新的语素义界定，这也为新词、新义的理解和计算带来了挑战。我们在郑画等(2022)构建的新词数据集的基础上，结合词义生成任务的需求和新的语义构词资源做改进：使用全局数据上重新训练过的随机森林，对所有新词的语义构词结构进行自动标注；考虑原数据集中的新词释义与《现汉》中词的释义在长度上的差异，使用GPT3.5对新词释义进行适当简化；同时，考虑GPT简化释义与《现汉》释义在行文

风格上的差异，请一名标注人员参照《现汉》给出新词的人工释义。在改进后的新词数据集上，我们使用上述三种模型进行词义生成评估，其结果如表9所示：

模型	BLEU(GPT简化释义)	BLEU(人工释义)
DeFT-语法结构*	13.76	18.59
DeFT-语义结构	14.98(+8.87%)	20.50(+10.27%)
DeFT+reverse-语义结构	**15.98(+13.89%)**	**22.19(+19.37%)**

Table 9: 新词的词义生成结果评估（*为基准模型，最佳结果加粗表示）

实验结果表明，考虑了语素义序列信息的改进模型在GPT简化释义、人工释义上的BLEU值分别为15.98、22.19，相较于语法结构信息分别大幅提升了13.89%、19.37%，此试验结果也符合主实验中的总体趋势，且提升幅度十分显著。这进一步验证了语义构词路线的优势和资源建设的有效性，表明本文的知识表示与标注算法可以进一步推广到新词上。

但另一方面，对比主实验中的生成结果，上述三个模型在新词上的BLEU值均有所降低。我们猜测，导致这一现象的原因可能有如下三个方面：1)新词中存在谐音应用和新语素义衍生等问题，使得相关语素无法在《现汉》中找到对应而被标为“空语素”。如“美眉”是“妹妹”的谐音而不是“美丽的眉毛”，“潮妈”中的“潮”指“新潮的”，但在目前的《现汉》中，“潮”字并没有这些语素义界定。事实上，在新词数据集中，有高达11.1%的词中会出现“空语素”的情况，有效语素义表征的缺失导致这些词的词义生成结果不够理想。例如“潮妈”，三个模型的词义生成结果均体现了“妈”的语素义，而曲解或忽略了“潮”的语素义；2)大量新词中存在转喻、隐喻等非字面义的情况(陈龙等, 2019)。如“草根”指“社会中的中低收入群体”而非“草本植物的根部”，“孩奴”指“因为孩子的养育成本而感到经济压力的父母”而非“孩子的奴隶”，这种低语义透明度的状态削弱了词和词义之间的直接联系。因此，目前生成的词义往往倾向于选择词的字面义，例如“孩奴”，三个模型的生成结果均接近于字面义，无法体现非字面义；3)GPT简化释义的初始来源为中文维基百科，与训练集中的《现汉》词的释义风格不同，这种风格差异会导致BLEU值的降低。例如“辅警”的GPT简化释义为“辅助警察，是协助正规警察提供额外警察力量的人员”，而在《现汉》中，与其结构相同、意义相近的“巡警”释义为“巡逻、维持治安的警察”，二者风格有较大差异。目前最优模型生成的结果是“指辅助工作的警察”，这与《现汉》更为相似，而与GPT简化释义存在风格差异，倾向导致BLEU值偏低。使用人工释义降低释义风格的影响后，最优模型的BLEU值达22.19，较主实验仅降低了2.88。

以上情况表明，面对历时变化的语言应用，现有词典中语素的语义空间划分存在一定的缺憾，无法反映并覆盖新词中可能衍生出的新语素义。注意到语义构词知识表示在新词上取得的显著提升，我们认为在语义构词标注工程的基础上，通过新的计算性手段的导入，有可能推测出新词衍生出的新语素义，该路径将为汉语的语言文字研究和词典编纂提供帮助。

6 结语

考虑汉语语素构词的特点，在前人研究的基础上，我们提出了一套面向计算的汉语语义构词结构体系；以此为指导，通过自动标注和人工校验相结合的方式构建了语义构词知识库；我们将新的知识表示应用于词义生成评估，取得了良好的效果：基于语义构词的词义生成BLEU值达25.07，较此前的语法构词提升了3.17%，显示了语义构词思路和资源建设的优势。同时，为了验证语义构词知识表示的可推广性，进一步将其应用于新词的词义生成，较语法构词的提升幅度也十分显著。

在后续工作中，我们计划将语义构词知识表示推广到汉语的多字词上，并利用资源建设成果进一步提升语义构词自动标注的准确性，以便更好地服务于人文领域和信息处理等多方面的应用，如词典编撰与浏览、汉语教育与研究、词向量训练及应用、词义消歧、语素义消歧、未登录词识别及语义预测等，为这些应用提供新的路径与方法。

参考文献

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transaction of the Association for Computational Linguistics*, 135-146.

David Dowty. 1991. Thematic Proto-Role and Argument Selection. *Language*, 67(3):547-619.

Artyom Gadetsky, Ilya Yakubovskiy, Dmitry Vetrov. 2018. Conditional Generators of Words Definitions. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 266–271.

Dirk Geeraerts. 2010. *Theories of Lexical Semantics*. Oxford University Press, New York.

Tao Gui, Qi Zhang, Jingjing Gong, Minlong Peng, Di Liang, Keyu Ding and Xuanjing Huang. 2018. Transferring from formal newswire domain with hypernet for twitter POS tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2540-2549.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3467-3476.

Thanapon Noraset, Chen Liang, Larry Birnbaum, Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3259-3266.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Mass.

James Pustejovsky, Elisabetta Jezek. 2008. Semantic Coercion in Language: Beyond Distributional Analysis. *Italia Journal of Linguistics*, 20(1):181-214.

Nianwen Xue, Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. *Proceedings of the Second SIGHAN Workshop*, 47-54.

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

Liping You, Kaiying Liu. 2005. Building Chinese FrameNet database. *Conference on Natural Language Processing and Knowledge Engineering. New York: IEEE*, 301-306.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, Yang Liu. 2021. Decompose, fuse and generate: A formation-informed method for chinese definition generation. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5524-5531.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, Yang Liu. 2021. Leveraging Word-Formation Knowledge for Chinese Word Sense Disambiguation. *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021*, 918-923.

Xiaoqing Zheng, Hanyang Chen, Tianyu Xu. 2013. Deep learning for Chinese word segmentation and POS tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 647-657.

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

董秀芳. 2011. 词汇化：汉语双音词的衍生与发展. 商务印书馆, 北京.

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

傅爱平. 2003. 汉语信息处理中单字的构词方式与合成词的识别和理解. 语言文字应用, (04):25-33.

葛本仪. 2001. 现代汉语词汇学. 山东人民出版社, 济南.

顾阳, 沈阳. 2001. 汉语合成复合词的构造过程. 中国语文, (02):122-133+191.

郭绍虞. 1979. 汉语语法修辞新探. 商务印书馆, 北京.

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

第二十二届中国计算语言学大会论文集，第456页-第467页，哈尔滨，中国，2023年8月3日至5日。

康司辰, 虞梦夏, 刘扬. 2020. 基于平行周遍原则的汉语未登录词的知识表示与预测. 中文信息学报, 34(08):23-31.

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

李可胜, 满海霞. 2013. VP的有界性与连动式的事件结构. 现代外语, 36(02):127-134+218.

刘茂福, 胡慧君. 2013. 基于认知与计算的事件语义学研究. 科学出版社, 北京.

刘明珠. 2020. 现代汉语NN型并列式复合词的生成机制探究. 西北大学硕士论文, 西安.

刘叔新. 1990. 汉语描写词汇学. 商务印书馆, 北京.

刘扬, 林子, 康司辰. 2018. 汉语的语素概念提取与语义构词分析. 中文信息学报, 32(02):12-21.

鲁川. 2010. 知识工程语言学. 清华大学出版社, 北京.

陆志韦. 1964. 汉语的构词法. 科学出版社, 北京.

宋衡, 曹存根, 王亚, 王石. 2023. 一种改进的汉语语义角色分类体系与标注实践. 中文信息学报, 37(01):16-32.

宋作艳, 赵青青, 亢世勇. 2015. 汉语复合名词语义信息标注词库：基于生成词库理论. 中文信息学报, 29(03):27-33+43.

宋作艳. 2022. 基于构式理论与物性结构的动名定中复合词研究——从动词视角到名词视角. 世界汉语教学, 36(01):33-48.

徐枢. 1990. 语素. 人民教育出版社, 北京.

徐通锵. 1997. 语言论：语义型语言的结构原理和研究方法. 东北师范大学出版社, 长春.

王晗. 2013. 现代汉语名+名联合式双音复合词研究. 山东大学博士论文, 济南.

王洪君. 2000. 汉语语法的基本单位与研究策略. 语言教学与研究, 02:10-18.

王笑. 2017. 物性结构与论元结构视域下汉语语义构词研究——以a+b=c类双音合成词为例. 鲁东大学硕士论文, 烟台.

魏雪, 袁毓林. 2013. 基于语义类和物性角色建构名名组合的释义模板. 世界汉语教学, 27(02):172-181.

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

袁毓林. 2002. 论元角色的层级关系和语义特征. 世界汉语教学, 03:10-22+2.

袁毓林. 2013. 基于生成词库论和论元结构理论的语义知识体系研究. 中文信息学报, 27(06):23-30.

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

张念歆, 宋作艳. 2015. 汉语形名复合词的语义建构：基于物性结构与概念整合理论. 中文信息学报, 29(06):38-45.

赵元任. 1980. 中国话的文法. 香港中文大学出版社, 香港.

郑画, 刘扬, 殷雅琦, 王悦, 代达劢. 2022. 基于词信息嵌入的汉语构词结构识别研究. 中文信息学报, 36(05):31-40+66.

周荐. 2003. 论词的构成、结构和地位. 中国语文, (02):148-155+192.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

朱彦. 2003. 汉语复合词语义构词法研究. 华东师范大学博士论文, 上海.

基于多尺度建模的端到端自动语音识别方法

陈昊，张润来，张裕浩，高成浩，许晨，马安香，肖桐*，朱靖波
东北大学计算机科学与工程学院自然语言处理实验室，沈阳，中国
methanechen@126.com, me_henrychang@163.com, yoohao.zhang@gmail.com,
{gaochrishao, xuchenneu}@outlook.com,
{maanxiang, xiaotong, zhujingbo}@mail.neu.edu.cn

摘要

近年来，基于深度学习的端到端自动语音识别模型直接对语音和文本进行建模，结构简单且性能上也具有显著优势，逐渐成为主流。然而，由于连续的语音信号与离散的文本在长度及表示尺度上存在巨大差异，二者间的模态鸿沟问题是该类任务一直存在的困扰。为解决该问题，本文提出了多尺度语音识别建模方法，该方法从利用细粒度分布知识的角度出发，建立多个不同尺度形式的文本信息，将特征序列从细粒度的低层次序列逐步对齐预测出文本序列。这种逐级预测的方式能够有效降低预测难度，缓解模态鸿沟带来的影响，并通过融合不同尺度下特征，提高语料信息的丰富性与完整性，进一步增强模型推理能力。本文在LibriSpeech小规模、大规模和TEDLIUM2数据集上实验，相比基线系统词错误率平均降低1.7、0.45和0.76，验证了方法的有效性。

关键词： 多尺度建模；特征融合；语音识别

An End-to-End Automatic Speech Recognition Method Based on Multiscale Modeling

Hao Chen, Runlai Zhang, Yuhao Zhang, Chenghao Gao, Chen Xu, Anxiang Ma, Tong Xiao*, Jingbo Zhu
NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
methanechen@126.com, me_henrychang@163.com, yoohao.zhang@gmail.com,
{gaochrishao, xuchenneu}@outlook.com,
{maanxiang, xiaotong, zhujingbo}@mail.neu.edu.cn

Abstract

End-to-end automatic speech recognition models based on deep learning that directly model speech and text have become mainstream due to their simple structure and remarkable performance. However, a persistent challenge in such tasks is the modality gap between continuous speech signals and discrete text, which arises from the significant differences in length and representation scale between the two modalities. To address the problem, this paper proposes a multi-scale speech recognition modeling method that builds multiple scales of text information from the perspective of using more fine-grained distribution knowledge. This progressive prediction approach effectively reduces the difficulty of prediction, mitigates the impact of the modality gap. The approach also enhances the model's inference capability, enriches and complements the information in the speech data by fusing features from different scales. Our method is effective on LibriSpeech small-scale, large-scale, and TEDLIUM2 datasets, showing an average reduction in word error rates of about 1.7, 0.45 and 0.76 compared to baseline.

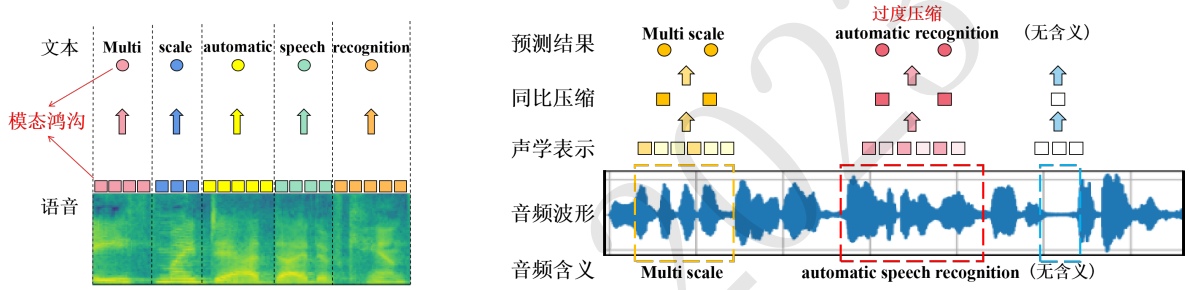
Keywords: Multiscale modeling , Feature fusion , Speech recognition

*通信作者：肖桐（xiaotong@mail.neu.edu.cn）

1 引言

自动语音识别(Automatic Speech Recognition, ASR)任务旨在将连续输入的语音信号转换为相应的输出文本，广泛应用于会议演讲、智慧办公、智能系统等日常生活领域。传统的语音识别系统需要预先对语音信号进行处理并提取特征，联合训练好的声学模型、语言模型、发音词典共同寻找由特征序列决定的最优状态序列，从而得到识别结果(Huang et al., 2001)。然而该方法不仅需要大量的人工对不同语言发音知识进行总结，随之也带来了模型更新困难、泛化能力弱、降噪能力不强等问题，进而难以满足复杂交流场景的需要。近年来，随着深度学习的兴起，神经网络模型在多个人工智能任务均取得了重要成就，在语音识别任务中基于端到端自动语音识别模型(End-to-End Automatic Speech Recognition, E2E ASR)(Graves and Jaitly, 2014)通过直接建模语音到文本的映射，模型结构简洁且大大简化了训练过程，在众多工作中已被证明能够取得更加优异的性能，逐渐成为主流。

然而，这种基于端到端语音到文本的模型在对音频进行建模时存在一个自然的问题：语音通过连续的声音信号进行传递，而文本是通过离散的符号序列进行传递，在图1(a)中可以发现，相同含义的语音和文本之间无论是序列长度还是内容的表示尺度均存在较大的差异，这种模态鸿沟(Modality Gap)(Fang et al., 2022)无疑给该类任务造成了巨大的困扰。例如，在语音处理中通常以帧级别为最小单元，即使文本处理任务中最小单元使用字符级别，二者序列长度上仍存在着数十倍的差距，同时，基于帧级别的特征信息并不足以预测出字级别文本信息，这两个问题导致了模型难以将二者进行准确对齐，从而影响预测结果。



(a) 语音与文本的模态鸿沟　　(b) 不同音频片段所需压缩方式不同

图1.语音与文本模态差异

为了减轻端到端自动语音识别模型输入与输出之间差距带来的压力，一种简单的方式是将音频特征信息进行压缩，进而对齐到文本的建模粒度上，这一定程度上能够减轻模态差异造成的干扰。在计算机视觉领域中同样面临着模态鸿沟的问题，研究者们采用金字塔结构(Fan et al., 2021)帮助模型逐渐地从图像中抽取有用信息进而对齐到需要的文本粒度上。然而，语音识别任务由于文本中不同词或者字的发音长度不尽相同，这要求金字塔结构需要根据不同的文本调整压缩比例，无疑是十分困难的。这个原因也导致了直接使用金字塔结构进行降采样会使得音频特征中重要信息被压缩。图1(b)从音频波形的角度可直观看到语音不同片段需要的降采样策略不能完全一致，采用统一的压缩策略会使得信息集中的地方被过度压缩。针对这一现象，本文展示了金字塔结构（如图2）在语音识别任务中存在的过度压缩问题，进而提出多尺度语音识别建模方法。该方法从利用更细粒度的分布知识的角度出发，建立多个不同尺度形式的文本信息供模型学习，利用细粒度层次上下文知识指导粗粒度数据的处理，进而防止模型在金字塔结构处理过程中一些低层次信息被过度压缩。本文为待识别的语音特征建立对应的词级别文本以外，同时建立对应的音素级别、字符级别文本信息共同参与训练，使模型在合适的粒度上进行对齐。这种逐级预测的方式不仅缓解了语音与文本之间粒度差距过大难以对齐的问题，并且能够通过融合不同尺度空间下的文本信息，使得语料信息更为丰富完整，缓解语音数据稀缺的问题(Zhang et al., 2022b)，例如字符以及子词级别文本侧重语料语义的理解，音素级别则更适应于声学信号的表达，二者信息能够有效互补，弥补了模型对音频过度压缩带来的损失，进一步改善自动语音识别效果。本文在LibriSpeech小规模和大规模数据以及TEDLIUM2数据集上进行了实验，本文方法相比基线系统词错误率平均约降低1.7、0.45和0.76，验证了所提出方法的有效性。

本文主要有如下贡献：

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

（1）本文利用语言发音的先验知识设计了多尺度文本金字塔结构，并发现金字塔结构在语音识别任务中存在的过度压缩问题。

（2）本文提出多尺度语音建模方法，引入子词、音素、字符等细粒度级别信息，并利用连接主义时序分类预测不同尺度的对齐效果，实现模型从细粒度的低层次序列逐步对齐预测出词序列，缓解了语音文本间模态鸿沟问题。

（3）本文提出多尺度特征融合方法，引入更多元化特征，有效补充相同语义下基于不同尺度特点的信息，提高信息完整性与丰富性，改善了由于压缩而导致的语义信息丢失的问题。

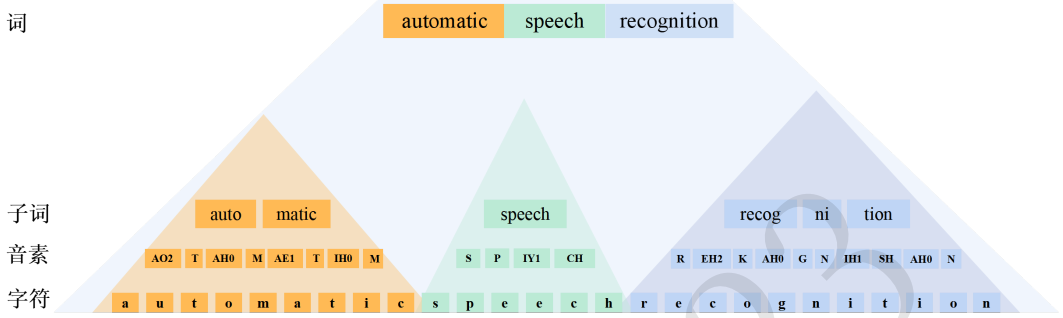


图2.多尺度文本金字塔结构

2 相关工作

语音识别技术的发展可追溯到上世纪五十年代，从最开始的模板匹配阶段到统计模型阶段，早期研究普遍集中在隐马尔科夫模型(Hidden Markov Model, HMM)(Baum et al., 1970)与统计模型例如高斯混合模型(Gaussian Mixture Model, GMM)(Dempster et al., 1977)的结合，由于模型泛化能力弱、算法复杂度高且数据质量难以保证等问题，该类模型在日常对话、新闻播报等场景下的识别率只能达到80%左右，应用普遍受到限制。在2010年后通过引入深度神经网络(Deep Neural Networks, DNN)(LeCun et al., 2015; Goodfellow et al., 2016)后发现，DNN能够将相邻帧建模到一起作为输入，更适合处理语音信号这种复杂信号。这种基于输入与输出的端到端建模方法大大简化了建模过程，通过目标函数直接训练神经网络模型，并对语音识别过程进行优化，提高语音识别精度。目前主流端到端模型主要包括连接主义时序分类(Connectionist Temporal Classification, CTC)(Graves et al., 2006)、循环神经网络转换器(Recurrent Neural Network Transducer, RNN-T)(Chorowski et al., 2014; Chorowski et al., 2015; Chan et al., 2016)和基于注意力机制的端到端语音识别(Attention Based Encoder-Decoder, AED)(Rao et al., 2017)等，然而不同端到端模型并不尽相同，例如CTC通过引入Blank机制，借鉴隐马尔科夫和动态规划思想实现标签硬对齐，而基于注意力机制的方法则通过序列模型实现输入与输出的软对齐，不同特点的模型在不同工作中均取得显著优势。2017年Transformer(Vaswani et al., 2017)架构问世，将Transformer应用于语音领域也进一步取得了明显的提升效果(Zhou et al., 2018; Dong et al., 2018)。

在语音任务中，通常以连续的帧级别特征作为标准的模型输入，而由于音频本身具有信息稀疏以及长序列的特点，导致生成的特征序列比对应文本序列长的多。这种过长序列不仅会导致捕获长距离依赖关系变的更为困难，注意力机制分配也会受到影响。为减少音频序列长度，目前多通过对特征序列进行降采样的方式(Guo et al., 2020)，通过堆叠多层卷积操作将多个相邻位置进行压缩，该方法操作简单且计算代价较低，但忽略了音频本身信息分布不均以及音频与文本之间的差异，会导致特征信息的过度采样以及遗漏问题。随着降采样策略不断发展(Xu et al., 2023b)，计算机视觉领域中提出了“金字塔”策略(Fan et al., 2021)解决类似问题。相关研究发现通过在更低分辨率下关注图像重要特征不仅能够减少计算需求，同样也可以帮助模型理解上下文以指导高分辨率下的处理(Rosenfeld and Thurston, 1971; Burt and Adelson, 1987)。

在语音任务中，为了促进端到端语音识别任务中词级别的表示学习，可通过让模型逐步学习难度提升的抽象语言序列(Higuchi et al., 2021)，为目标文本构建更细力度尺度信息，将特征序列从细粒度的低层次序列逐步对齐到子词序列，直至最后预测出单词级别序列，Jelinek and F. (1976)的工作也表明逐步提高语言信息的抽象水平对于训练语音识别模型是一种合理的方式。

除了对特征序列进行相关处理外，语音特征表示自身对语音下游任务同样具有显著的影响(Deng et al., 2013)，单一特征往往难以包含语音中语言、情感、韵律等多元化信息，近年来研究者们开始探索将多种特征经过融合用于语音下游任务。由于不同声学特征可通过不同角度对语音任务起到帮助作用，故而语音任务中可通过不同特征融合互补发挥各自优势，该方式保留了大部分信息，但同时也增加了特征维度。在先前工作中，Yoon et al.(2018; 2019)通过将韵律特征与MFCC特征融合应用于语音情感识别任务；袁文浩et al. (2021)将时域特征与频域特征融合实现语音增强；Zhang et al. (2022a)通过融合不同频率下Mel滤波器组特征在说话人识别任务中达到了最先进的性能。

3 本文方法

3.1 基线

模型结构 在基线中，本文主要采用Transformer模型用作端到端语音识别任务，与标准Transformer模型结构基本一致，在编码器输入时，通过堆叠两层步长为2的卷积模块将输入的语音特征进行压缩，将长度压缩为原先1/4，以此降低序列长度，节约计算资源，模型总体采用12层编码器，6层解码器的结构。

连接主义时序分类 连接主义时序分类(CTC)是一种在符号序列上训练而不需要对齐的递归网络方法。传统的基于隐马尔科夫模型的深度神经网络语音识别模型都需要预先建立输入语音特征与输出标签之间的对应关系，耗时耗力且难以保证对齐的准确性。不同于传统的深度神经网络语音识别模型，CTC是一种端到端的模型训练方法，通过将模型的输出层进行扩展，使输出文本和标签建立对应关系，模型能够直接对输入的语音特征进行训练，从而输出预测序列的概率。Xu et al. (2023a)将CTC应用在语音任务的监督学习取得了显著效果。

给定输入序列X=[x_1,x_2,x_3...,x_T]以及对应的标签数据Y=[y_1,y_2,y_3...,y_U]，分别对应自动语音识别任务中音频特征序列以及文本序列，CTC返回给定输入序列X的所有可能Y的输出分布，根据输出概率输出最有可能的结果。令$p(l|x)$表示输入为x，输出为序列l的概率，$p(l|x)$形式化定义如下：

$$P(l|x) = \sum_{\pi \in F^{-1}(l)} p(\pi|x) \tag{1}$$

其中，$\pi \in F^{-1}(l)$代表所有经过F^{-1}变换（将神经网络输出的原始预测序列转换为最终输出序列）后是l的路径π，y_t^k表示π路径下t时刻的概率值，对于任意一条路径π都有：

$$P(\pi|x) = \prod_{t=1}^{T} (y_t^k) \tag{2}$$

虽然音频序列远远大于文本序列，但是由于引入了空白位置对齐的方式，这种CTC建模方式能够实现两者的一一对应。这种特性使得本文可以根据当前音频的建模粒度，将音频序列对齐到相应的文本表示粒度，如字符、音素、子词等。

3.2 金字塔结构

金字塔结构是一种在计算机视觉中广泛使用的图像处理方法，通过将图像分解为不同尺度的子图像，对图像进行多尺度分析。这种结构的核心在于将降采样模块适当地加入到模型结构中，使其逐步地聚合信息，之后再利用其他的模块对压缩之后特征进行处理。由于在信息聚合的过程中，特征长度在逐渐减少，而每个位置上的信息逐渐增加，最后得到表示更有利于尺度更小的文本任务。如在图3中，本文在编码端中插入了一些降采样模块，这些模块通过聚合相邻单元的信息获得更细粒度的音频特征表示。为便于模型对于进入多尺度层向量表示序列的处理，本文采用层标准化(Layer Normalization)(Ba et al., 2016)的方式，规范向量输出，利用正则化的方式，增强模型对于不同尺度数据的适应性，提高模型训练的速度和性能，降低了过拟合的风险。这个结构作为语音**金字塔结构**模型。

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

为了确定多尺度降采样比例，本文在LibriSpeech-100h数据集上统计了各个尺度文本与输入特征序列的最小长度比（见表1），根据长度关系可发现，输入序列长度至少为字符以及音素级别序列长度4倍，至少为子词级别序列12倍，基于此本文设计了在每次对齐前引入一层卷积神经网络的策略，将序列压缩为原先长度的二分之一，降低模型训练压力并且避免二者差距过大导致性能下降。同时针对编码器输入时的两层卷积优化为一层，在最终对齐到文本输出时总共实现8倍压缩，控制了编码器端对向量序列的压缩比例。

前文已经提到，语音信号是一种时间序列信号，相邻时间点的采样值存在强相关性和依赖性(Rabiner and Juang, 1993)。因此利用这种结构会导致语音中一些小尺度的信息如字符级别将会被过度地压缩，进而导致了关键信息的丢失，影响语音预测结果。

不同尺度文本关系	最小长度比
输入序列/字符级别序列	4.20
输入序列/音素级别序列	4.86
输入序列/子词级别序列	12.31

表1.LibriSpeech-100h数据集不同尺度文本最小长度比

3.3 多尺度语音建模

基于此，本文在每次对特征序列利用卷积神经网络降采样之前，尝试及时地将音频信息对齐到对应的文本粒度，以防止对应尺度的信息在之后的处理中丢失。同时本文引入了不同文本尺度信息，丰富语音整个建模过程。根据先验知识，由于模型在中间位置层时可以同时获得全局和局部的特征信息，具有足够的高层抽象特征和低层原始特征，能够为多尺度信息的提取提供足够的基础(Lin et al., 2017)。如果将多尺度信息引入到其他层中，可能会降低模型的表现能力或者增加模型的复杂度。如图3所示，本文在编码器端中间第六、九层以及编码端的输出位置分别引入字符级别，音素级别以及子词级别的信息，通过在不同的尺度上对输入序列进行建模来捕捉输入序列中更加细致的信息，该结构为本文提出的**多尺度语音建模**方法。考虑到编码端在建模过程中已经有足够的跨模态抽象特征，可以为对应尺度的处理提供足够的信息，因此该方法直接对指定层的输出向量表示序列进行记录，并利用CTC函数辅助该序列与相应尺度文本进行对齐，之后向量序列经过压缩送到下一种尺度的编码端处理或者送到解码端，整个方法模拟了从细粒度的低层次序列逐步对齐到子词序列的过程。

具体而言，整个对齐过程可分为三个阶段：假设输入模型的特征序列为X，进入编码器端第i层特征序列表示为X_i，字符级别多尺度文本为Y_c，音素级别多尺度文本为Y_p，子词级别文本为Y_w，则三个阶段的CTC对齐损失函数可以分别表示为：

1.特征序列对齐到字符级别多尺度文本的CTC对齐损失函数：

$$L_c = -\log P(Y_c|X_6) \tag{3}$$

2.经过字符级别对齐后的特征序列，对齐到音素级别多尺度文本的CTC对齐损失函数：

$$L_p = -\log P(Y_p|X_9) \tag{4}$$

3.经过音素级别对齐后的特征序列，对齐到子词级别多尺度文本的CTC对齐损失函数：

$$L_w = -\log P(Y_w|X_{12}) \tag{5}$$

其中$P(Y_c|X_6)$、$P(Y_p|X_9)$、$P(Y_w|X_{12})$表示给定输入相应语音特征语音特征序列的条件下，该序列与对应尺度文本字符序列匹配的概率。三个阶段的对齐损失函数可以合并为一个多尺度损失函数L_{ms}：

$$L_{ms} = \alpha * (L_c + L_p + L_w) \tag{6}$$

其中α是超参数，作为缩放因子调节损失大小。通过最小化总损失函数L_{ms}，可以优化特征序列与多尺度文本之间的对齐过程，进而将多尺度信息的引入到建模过程中。这种方法能够提高语音识别的准确率和鲁棒性，使得模型同时捕捉输入序列中的全局特征和局部特征，从而增强模型的表现能力。

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

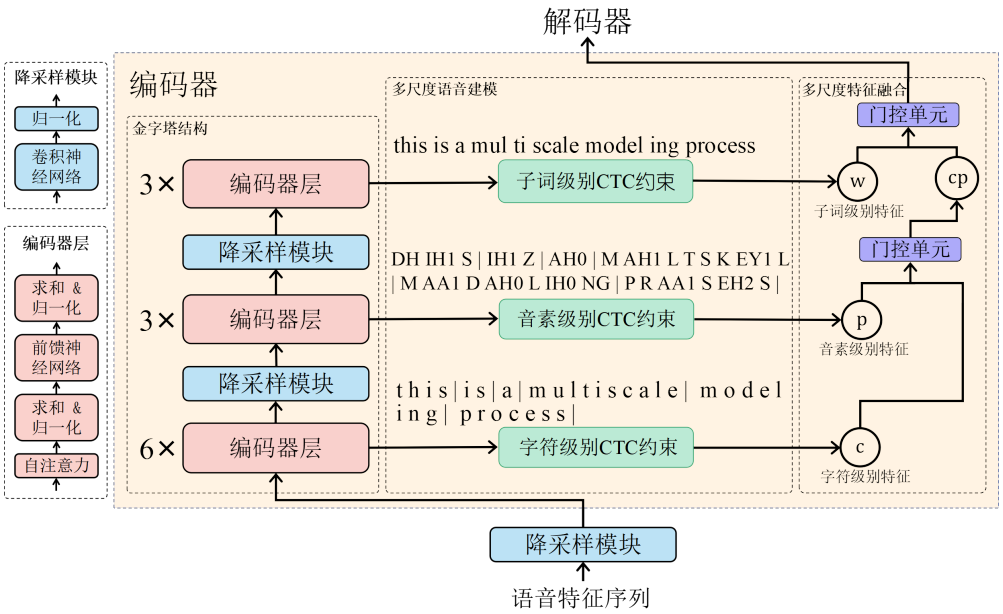


图3.多尺度语音识别建模方法

3.4 多尺度特征融合

为进一步探讨多尺度特征对端到端语音识别任务的促进作用，本文将不同尺度文本特征序列根据一定方式进行融合，通过引入多元化特征改善压缩过度导致的信息丢失问题。在特征融合上，本文主要考虑从长度以及维度两个角度融合多尺度特征。长度融合的问题在于不同尺度的特征长度并不一致，本文依旧采用了高效的卷积网络将其他尺度的特征长度对齐到子词级别的特征。

虽然长度不一致问题能够通过卷积进行处理，但是也难以保证得到不同尺度的特征之间是能够相互帮助，而不是给其他级别特征引入噪音进而降低了特征质量。因此在特征维度融合的角度，考虑到不同尺度特征在训练过程中贡献并不完全相同，可通过引入门控机制来解决这个问题。门控循环单元(Gated Recurrent Unit, GRU)(Hochreiter and Schmidhuber, 2016) 是一种常用的循环神经网络模型，它通过门控单元来控制信息的流动和捕捉长期依赖关系，帮助网络动态的调整每个特征的权重，从而更灵活地融合特征。基于此本文提出了一种基于门控的多尺度特征融合方法，将字符、音素和子词级别的特征进行融合，使得各种尺度之间的特征相互帮助。如图3中，通过逐级地两两融合特征，将其作为**多尺度特征融合**方法。

具体而言，首先将字符级别的特征F_c和音素级别的特征F_p通过一个门控模块进行融合，其中门控机制使用了一些可学习的参数，这些参数通过$sigmoid$函数进行非线性变换，从而控制门的打开程度，并通过预设融合比例实现特征融合的目的，得到融合后的特征F_{cp}，如下所示：

$$F_{cp} = W_{cp} \odot F_c + (1 - W_{cp}) \odot F_p \tag{7}$$

$$W_{cp} = a \cdot \sigma(arg_{cp}) + b \tag{8}$$

其中W_{cp}为字符级别特征与音素级别特征融合时的门控模块，σ表示sigmoid激活函数，arg_{cp}为定义的字符音素特征融合可学习参数，共同参与神经网络模型训练。a为门控参数的sigmoid函数的偏置，b则是为了保证门控参数的范围，以避免过于偏向某个特征而导致融合结果出现偏差，二者均为可调节的超参数用于控制不同特征的融合比例。

接着，将融合后的特征和子词级别特征以相同的方式再次融合，得到最终的特征表示F，如下所示：

$$F = W_{cpw} \odot F_{cp} + (1 - W_{cpw}) \odot F_w \tag{9}$$

$$W_{cpw} = a \cdot \sigma(arg_{cpw}) + b \tag{10}$$

其中，W_{cpw}为字符音素融合特征以及子词级别特征融合时的门控模块，a与b与公式8一致，arg_{cpw}为定义的多特征融合可学习参数。

本文使用门控模型对不同级别的特征进行加权融合，以避免低质量特征的干扰。在每次融合前，利用卷积神经网络对低层特征实现两倍压缩，将待融合的两个特征序列统一到相同的长度，对融合后的特征进行层标准化操作，以缓解梯度消失和梯度爆炸的问题。

4 实验

4.1 数据处理

本文实验主要基于LibriSpeech数据集(Panayotov et al., 2015)100小时子集、960小时完整数据集以及TEDLIUM2数据集(Rousseau et al., 2014)，在相应数据集上训练模型并使用标准验证集和测试集对模型训练结果进行评分。

特征提取 音频信号中存在着许多不同的特征，机器通过分析语音找到特征对应的特征参数，从语音中提取出能够有效反映关键特征参数的特征向量序列的过程就是特征提取，目前常用的几种特征参数包括：Mel 频率倒数系数(Mel-Frequency Cepstral Coefficients, MFCC)、基于滤波器库的Fbank (Log-Mel-Filter Bank)特征、线性预测分析(Linear Prediction Coefficient, LPC)等。基于人耳结构特点，Stevens and S. (1936)提出符合人耳听觉特性的Mel尺度，在Mel尺度下，对三角滤波器组输出取对数即可得到Fbank特征，Fbank特征具有二维结构，可以通过引入卷积神经网络进行处理，本文主要采用Fbank 特征进行实验。

尺度处理 本文为训练建立多尺度信息时，主要根据目标文本（本文以英语为例）建立了字符级别与音素级别文本，字符级别旨在将目标文本中所有单词均变为以单个字符为基本单元的形式，例如“speech”则对应为“s p e e c h”；在处理音素级别时，本文利用CMU发音词典[1](Carnegie Mellon University Pronouncing Dictionary)，将单词转为相应的发音音素格式，例如“speech”对应为“S P IY1 CH”，CMU词典是一本面向北美英语的开源机器可读发音词典，该词典音素集主要包含39个音素以及超过134000个单词及其发音，具有从单词到发音的映射，并依然在不断更新，在语音识别和合成领域发挥了巨大作用。而部分未出现在CMU词典中的单词，由于不能转换为对应的音素形式，本文利用谷歌开发的开源工具sentencepiece[2]先对词级别目标文本构建词典以及模型，利用该模型对音素级训练语料中仍然存在的单词级别文本进行进一步切分，将单词切分为更细粒度的语义单元。在模型训练时，本文将三个尺度文本信息通过统计词频的方式统一到大小为10000的同一张词表，以此降低训练代价。

速度扰动 由于端到端模型常常需要大规模数据进行训练，在小规模数据集时表现往往受到限制，因此数据增强相关方法也被广泛研究。速度扰动(Speed Perturb)(Liberman and Mattingly, 1985)是一种数据增广方法，主要通过对音频播放速度的调整，得到更快语速或更慢语速的语音数据，从不同语速的语音数据中分别提取特征序列，以达到训练数据的扩充。本文为LibriSpeech-100h以及TEDLIUM2数据集语音分别设置0.9、1.0、1.1倍语速，将训练数据扩充为原数据集三倍，弥补小规模数据集可能导致的训练不充分问题。

4.2 实验设置

本文实验主要基于Fairseq的S2T框架[3]，并在该框架基础上完成本文提出的多尺度建模方法。模型配置上，除了前文介绍的12层编码器，6层解码器架构外，各层隐层变量维度均为256，前馈网络维度均为2048，多头注意力头数为4，dropout为0.1。学习率最大阈值为2e-7，学习率预热迭代次数为10000，并采用inverse sqrt对学习率动态调整。使用Adam优化算法，其中两次估计指数衰减率分别为β_1=0.9，β_2=0.98，使用标签平滑率为0.1的交叉熵损失作为目标函数。在引入金字塔模型时，本文在第六层和第九层加入降采样模块。每次降采样均采用步长为2，卷积核大小为5的卷积层实现。在训练时，利用CTC损失函数辅助模型训练，对于本文中涉及到多重CTC损失（即L_w、L_p、L_c）权重α设为0.2，在特征融合时，特征融合比例参数a=0.2，b=0.4。训练结束时，对训练最后十轮模型参数进行平均，采用大小为5的束搜索算法(Jelinek, 1980)进行解码。本文通过词错误率(Word Error Rate, WER)来评价自动语音识别的效果。

[1]http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[2]https://github.com/google/sentencepiece
[3]https://github.com/pytorch/fairseq

4.3 实验结果

在本文的主要实验中，以LibriSpeech以及TEDLIUM2数据集实验得到的结果如表2所示。本文使用fairseq的端到端语音到文本模型(Wang et al., 2020)作为基线模型，并在此基础上引入了所提出的多尺度语音建模以及多尺度特征融合方法，本文模型参数量大小均为30M。此外，本文还对比了Higuchi et al. (2021)、Andrusenko et al. (2022)在相同数据集上的相关工作。实验结果表明，本文基线系统相比先前工作已经有了一定提升。然而，当直接采用金字塔结构时，性能下降较为明显，验证了前文介绍的金字塔结构容易对语音信息造成过度压缩的问题。本文提出的多尺度语音建模方法在LibriSpeech100小时测试集上的WER相对于基线分别降低了0.66和1.38，在TEDLIUM2测试集降低了0.68，在LibriSpeech960小时测试集other上降低了0.36，从而证明了多尺度语音建模方法的有效性，而在clean数据集上，由于语音信号的质量较高且干扰较少，传统的单尺度模型已经能够有效地识别语音，使用多尺度模型可能反而增加模型的复杂度，对于更加复杂的数据集，多尺度模型可能会比单尺度模型更有效。此外，本文引入了多个尺度特征共同参与训练，并采用基于多尺度特征融合的方法来进一步提升识别性能。实验结果表明，在LibriSpeech-100h测试集上WER表现比基线分别低1.37和2.03，在LibriSpeech-960h测试集上比基线低0.24和0.66，在TEDLIUM2测试集比基线低0.76，WER相对于多尺度建模方法降低了一定的程度，这证明了多尺度特征融合对于多尺度建模过程的进一步推动作用。

方法	LibriSpeech-100h				LibriSpeech-960h				TED LIUM2
	Dev		Test		Dev		Test		
	clean	other	clean	other	clean	other	clean	other	Test
Higuchi	11.50	24.80	11.80	25.50	4.20	10.00	4.50	9.90	10.70
Andrusenko	10.40	27.10	10.70	27.10	3.70	10.10	3.70	9.90	-
本文基线	9.68	23.37	10.46	23.28	4.17	9.66	4.41	9.22	10.58
金字塔结构	10.28	22.52	11.97	24.63	4.20	9.39	5.16	9.35	10.71
多尺度语音建模	8.96	21.26	9.80	21.90	4.13	8.92	4.72	8.86	9.90
多尺度特征融合	**8.39**	**21.23**	**9.09**	**21.25**	**3.86**	**8.69**	**4.17**	**8.56**	**9.82**

表2.主实验结果

值得注意的是，本文对比了LibriSpeech小规模和大规模数据集的识别精度，实验结果表明，在小规模数据集上，本文方法比传统方法取得了更好的效果。根据分析，一个可能的原因可能是由于小规模数据集的数据量较少，模型容易出现过拟合的情况，同时数据中的噪声和变化也较大，这会影响模型的泛化能力和识别精度。而本文方法能够在不同尺度下提取不同粒度的语音特征，并通过多层CTC对齐实现特征序列到字符、音素和子词三级对齐，从而更好地捕捉语音信号的结构和特征，提高模型的鲁棒性和泛化能力，因此在小规模数据集上取得更好的效果。而在大规模数据上，传统方法往往已经能够很好地处理数据的分布和噪声，同时具有更高的样本覆盖率和更多的数据多样性，导致了本文方法效果提升更微弱。

与其他人工作对比，也不难发现本文方法在多项指标上均有优势，说明多尺度建模方法利用多个尺度的特征信息，能够更全面地捕捉语音信号的特征，并通过层次化的处理逐渐融合和传递上下文信息，可以更好地捕捉语音信号中的上下文关系，以及同时处理多个尺度的特征，对噪声和干扰具有更好的抗干扰能力，具备更出色的语音建模能力。而在LibriSpeech960小时clean数据集上，我们方法略差于Andrusenko et al. (2022)，主要考虑为该工作中使用的Conformer结构在语音任务中更好地处理时间关系、全局上下文和位置关系，因此相对于传统的Transformer结构，更适用于语音识别任务，在后续工作中我们也将在Conformer结构上展开进一步实验。

5 实验分析

5.1 多尺度语音建模消融实验

为了验证不同尺度CTC都能对语音建模起到帮助作用，本文在LibriSpeech100小时数据集上展开实验（见表3），通过分别去掉字符级别、音素级别、子词级别CTC约束进行实验，可发现任意一种CTC约束的去除都会对语音识别的性能产生不同程度的影响。这说明三个级别

的CTC都能够在该任务中提供有价值的信息，验证了本文提出的三种级别CTC存在各自的尺度优势，可以帮助模型逐级学习语音和文本之间的映射，从而缓解模态鸿沟问题。同时，不难发现去掉子词级别CTC会对实验结果产生最大的影响，这是因为子词级别CTC最接近于语音识别的输出，可以更好地对应识别结果的最终目标，使得模型能够更准确地预测完整的单词，在语音识别任务中起着更为重要的作用。

方法	LibriSpeech-100h			
	Dev		Test	
	clean	other	clean	other
多尺度语音建模	8.96	21.26	9.80	21.90
-字符级别CTC约束	9.02	22.53	9.93	22.24
-音素级别CTC约束	9.68	22.27	10.21	22.29
-子词级别CTC约束	9.59	21.92	10.44	22.77

表3.多尺度语音建模消融实验（“-”表示在原始方法上进行处理）

5.2 多尺度特征融合消融实验

为了验证本文在处理特征融合的合理性，本文在LibriSpeech100小时数据集上展开实验（见表4），当不使用门控单元而将三者在维度上直接融合送入解码器时，可以看到WER在测试集中平均高出1.24，说明这种简单的融合方式不仅难以有效将不同特征进行互相补充，反而给原本特征引入了噪声，导致识别效果下降。同时，为了证明本文提出的多元化特征能够改善压缩导致的信息丢失问题，本文分别去掉字符特征、音素特征进行实验，发现当仅使用两种特征用门控方式融合时，WER在测试集上的表现也都有不同程度的上涨，这也证明了三个特征融合的合理性与必要性，融入多个特征能够帮助模型对于语义信息的理解。一个有意思的现象是，特征融合在other数据集中的表现并不如clean数据集，甚至在验证集other上融合后的特征表现反而更为劣势，分析可知，LibriSpeech other数据集比clean数据集更加复杂和多样化，包含更多的噪声和变化。当融合多尺度特征时，这些噪声和变化的影响可能被强化，导致在验证集上表现反而不如未融合的特征。相比之下，clean数据集相对更简单，噪声和变化较少，融合特征能够更好地捕捉到语音的变化，利于提高性能。

方法	LibriSpeech-100h			
	Dev		Test	
	clean	other	clean	other
多尺度特征融合	8.39	21.23	9.09	21.25
-门控单元	9.36	21.82	10.47	22.35
-字符特征	8.70	20.92	9.37	21.32
-音素特征	8.83	21.13	9.36	21.28

表4.多尺度特征融合消融实验（“-”表示在原始方法上进行处理）

5.3 基于不同建模方式预测结果实验

为了更直观体现出本文方法对于语音识别结果促进作用，本文在表5中展示了不同方法对于同一条语句的识别输出结果。在基线模型中，语音识别错误两个词语，WER为4.16，而在金字塔结构中，正如前文分析的存在过度压缩导致信息丢失的问题，从而识别出的信息也存在缺失，并且由于信息缺失进而加剧了预测错误率。在本文提出的多尺度语音建模方法中可以看到，在识别单词“word”时并没有像基线任务中错误预测成“world”，而二者在发音规则中极为接近，这说明多个尺度的文本能够帮助模型丰富语料信息，进一步理解语义，提升预测效果，而通过多尺度特征融合，更是帮助该语句预测准确率达到100%，说明融入的音素特征使得模型加强了对声学信号的理解，解决了该句基线方法中的所有预测错误。

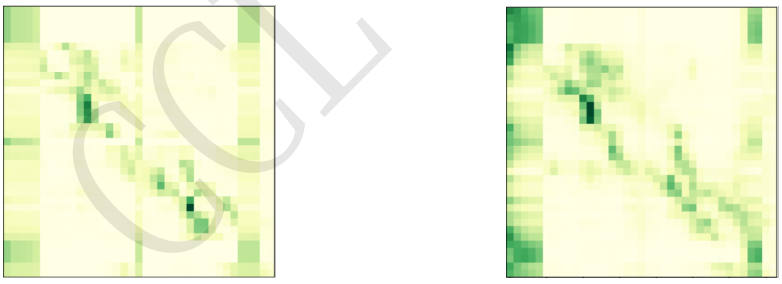
第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

原文	who were a mere handful against an army should he be untrue at once to his love to country to his word should he give to his cowardice the pretext of patriotism but this was impossible and if the phantom of his father was there in the gloom.	
方法	语音识别结果	WER
本文基线	who were a mere handful against an army should he be untrue at once to his love to country to his world(错误) should he give to his cowardice the pretext of patriot ism but this was impossible and if the fathom(错误) of his father was there in the gloom.	4.16
金字塔结构	who were a mere handful against an army should(缺失) he had and(错误) untrue at once to his love to country to his world(错误) should he give to his cowardice the pretext of(缺失) patriot ism but this was impossible and if the phantom of his father was there in the gloom.	14.6
多尺度语音建模	who were a mere handful against an army should he be untrue at once to his love to country to his word should he give to his cowardice the pretext of patriot ism but this was impossible and if the fathom(错误) of his father was there in the gloom.	2.08
多尺度特征融合	who were a mere handful against an army should he be untrue at once to his love to country to his word should he give to his cowardice the pretext of patriot ism but this was impossible and if the phantom of his father was there in the gloom.	0

表5.比较基于不同建模方式语音识别结果

5.4 多尺度建模注意力可视化分析

本文从注意力机制的角度进一步分析本文方法的有效性，通过图4比较基线模型和本文方法模型在注意力可视化方面的表现，可以发现本文提出的多尺度模型在这一方面表现更优异。在基线模型的注意力可视化结果中，颜色较浅，表示模型关注的区域较少，可能存在遗漏的信息，而多尺度模型的注意力可视化结果则更为深色，表明模型能够更充分地关注输入信号的不同部分，能够更好地捕捉语音信号中的重要特征，从而提高语音识别的性能。并且在多尺度方法图中一些较远距离单元色块颜色也更为深色，这说明多尺度方法能够更好地捕捉到长距离之间的关系，从而理解上下文信息，提高语音识别模型的性能。



(a) 基线模型　　(b) 本文方法

图4.注意力权重分布比较

6 结论

本文提出一种多尺度语音识别的建模方法，通过构建不同尺度下的文本信息，并利用CTC实现逐级对齐预测，从低层次的细粒度序列最终预测出完整文本序列，从而有效缓解了模态鸿沟问题。同时，本文还融合了不同尺度下的特征，加强了训练语料的丰富性与完整性，进一步提高了模型的推理能力。本文在LibriSpeech小规模和大规模数据以及TEDLIUM2数据集上进行了实验，结果显示本文方法相比基线系统，词错误率平均约降低1.7、0.45和0.76。后续的实验分析表明，多尺度语音建模和多尺度特征融合都促进了模型性能的提升。

在未来的工作中，我们将专注于如何在训练过程中更准确地评估不同尺度特征对模型训练的贡献，采用更加灵活的方法获取不同尺度约束比例和特征融合方法，进一步提高多尺度语音识别系统的在不同数据集上的表现，并探索多尺度建模在各种语音相关任务中的潜力。

第二十二届中国计算语言学大会论文集，第468页-第479页，哈尔滨，中国，2023年8月3日至5日。

致谢

感谢国家自然科学基金（62276056）；国家重点研发计划项目；科技部科技创新2030—“新一代人工智能”重大项目（2020AAA0107904）；辽宁省自然科学基金（2022-KF-16-01）；云南省科技厅科技计划项目（202103AA080015）；中央高校基本科研业务费项目（N2216016、N2216001、N2216002）；111引智基地（B16009）的资助。

参考文献

Andrei Andrusenko, Rauf Nasretdinov, and Aleksei Romanenko. 2022. Uconv-conformer: High reduction of input sequence length for end-to-end speech recognition. *CoRR*, abs/2208.07657.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171.

P. J. Burt and E. H. Adelson. 1987. The laplacian pyramid as a compact image code. *Readings in Computer Vision*, 31(4):671–679.

W. Chan, N. Jaitly, Q. Le, and O. Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *Eprint Arxiv*.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Neural Information Processing Systems*.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. 2013. Recent advances in deep learning for speech research at microsoft. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8604–8608. IEEE.

L. Dong, X. Shuang, and X. Bo. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

H. Fan, B. Xiong, K. Mangalam, Y. Li, and C. Feichtenhofer. 2021. Multiscale vision transformers.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. Stemm: Self-learning with speech-text manifold mixup for speech translation. *arXiv preprint arXiv:2203.10426*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

A. Graves, S Fernández, and F. Gomez. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *ACM*.

P. Guo, F. Boyer, X. Chang, T. Hayashi, and Y. Zhang. 2020. Recent developments on espnet toolkit boosted by conformer.

Y. Higuchi, K. Karube, T. Ogawa, and T. Kobayashi. 2021. Hierarchical conditional end-to-end asr with ctc and multi-granular subword units. *arXiv e-prints*.

Sepp Hochreiter and Jürgen Schmidhuber. 2016. Learning to forget: Continual prediction with lstm. *Neural Computation*, 28(10):2451–2471.

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. 2001. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Jelinek and F. 1976. Continuous speech recognition by statistical methods. *Proc IEEE*, 64(4):532–556.

Frederick Jelinek. 1980. Interpolated estimation of markov source parameters from sparse data. In *Proceeding of the Workshop on Pattern Recognition in Practice*, pages 381–397.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Alvin M Liberman and Ignatius G Mattingly. 1985. The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Tsung Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. *IEEE Computer Society*.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc.

Kanishka Rao, Hasim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

A. P. Rosenfeld and M. Thurston. 1971. Edge and curve detection for visual scene analysis. *IEEE Transactions on Computers*.

Anthony Rousseau, Paul Deléglise, Yannick Esteve, et al. 2014. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*, pages 3935–3939.

Stevens and S. S. 1936. A scale for the measurement of a psychological magnitude: loudness. *Psychological Review*, 43(5):405–416.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *arXiv*.

C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq.

Chen Xu, Xiaoqian Liu, Xiaowen Liu, Qingxuan Sun, Yuhao Zhang, Murun Yang, Qianqian Dong, Tom Ko, Mingxuan Wang, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2023a. Ctc-based non-autoregressive speech translation. *CoRR*, abs/2305.17358.

Chen Xu, Yuhao Zhang, Chengbo Jiao, Xiaoqian Liu, Chi Hu, Xin Zeng, Tong Xiao, Anxiang Ma, Huizhen Wang, and Jingbo Zhu. 2023b. Bridging the granularity gap for acoustic modeling. *CoRR*, abs/2305.17356.

S. Yoon, S. Byun, and K. Jung. 2018. Multimodal speech emotion recognition using audio and text. *IEEE*.

S. Yoon, S. Byun, S. Dey, and K. Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. *IEEE*.

J. Zhang, W. Yan, and Y. Zhang. 2022a. A new speech feature fusion method with cross gate parallel cnn for speaker recognition. *arXiv e-prints*.

Yuhao Zhang, Chen Xu, Bojie Hu, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2022b. Improving end-to-end speech translation by leveraging auxiliary speech and text data. *CoRR*, abs/2212.01778.

Shiyu Zhou, Linhao Dong, Shuang Xu, and Bo Xu. 2018. Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. *Springer, Cham*.

袁文浩, 时云龙, 胡少东, and 娄迎曦. 2021. 一种基于时频域特征融合的语音增强方法. 计算机工程, 047(010):75–81.

基于血缘关系结构的亲属关系推理算法研究

卢达威
中国人民大学文学院/中国人民大学数字人文研究院
北京市海淀区中关村大街59号
wedalu@163.com

杨思琴*
中国人民大学文学院
北京市海淀区中关村大街59号
yang-sq19@mails.tsinghua.edu.cn

摘要

以往的亲属关系推理系统，对推理的正确性无法保证，对复杂的亲属关系推理容易出错；而且难以解决多个亲属关系作为已知条件的亲属关系推理问题。本文在卢达威等（2019）的基础上，首先将推理规则和推理过程形式化和算法化；进而与基于一阶谓词逻辑的推理系统进行了对比，发现基于血缘关系结构的亲属关系推理在知识表示方法和推理规则方面都存在优势，主要表现在于执行效率更高，以及在编写和核查规则时更不容易出错；最后讨论了亲属关系推理算法的时间复杂度问题，发现该推理系统为是线性时间复杂度。本文的算法及其有效性分析得到了实验结果的支持。

关键词： 血缘关系结构 ； 亲属关系推理 ； 知识表示 ； 推理系统

A Study on Kinship Inference Algorithm Based on Blood Relationship Structure

LU Dawei
School of Liberal Arts, RUC /
School of Digital Humanities of RUC
wedalu@163.com

YANG Siqin*
School of Liberal Arts, RUC
yang-sq19@mails.tsinghua.edu.cn

Abstract

The correctness and completeness of previous kinship inference systems are not guaranteed, which leads to error-prone inferences when the kinship is complex. Moreover, these systems are difficult to infer with multiple kinships as known conditions. Based on Lu et al. (2019), this paper first formalizes and algorithmizes the inference rules and processes. When compared with kinship inference systems based on first-order predicate logic, it is found that the system based on blood relationship structure has advantages in both knowledge representation methods and reasoning rules, particularly in the high efficiency and accuracy of writing and checking rules. Finally, the time complexity problem of the kinship inference algorithm is discussed, revealing that the time complexity of this system is linear. These studies contribute to the practicality of kinship inference.

Keywords: blood relationship structure , kinship inference , knowledge representation , inference system

本文得到国家社会科学基金青年项目“汉语话题延续与转换机制及其计算模型研究”(18CYY030)的资助，特此谢

1 引言

亲属关系推理（kinship inference）是知识推理的经典问题。也因其经典性，容易让人忽略其复杂性。表面上亲属关系推理是语言中亲属称谓词之间的词汇概念的推理。但是，常见的词汇关系刻画手段，例如同义—反义关系（synonymy–antonymy），上位—下位关系（hyperonym–hyponym），整体—部分关系（holonym–meronym）等，都不能反映亲属称谓词之间的实质内涵（卢达威等，2019）。乔姆斯基（2015: 13）认为，亲属体系具有很多数学系统的性质；理解亲属关系还需要用上人类解决形式问题的算术能力。人类学家们虽然早就注意到亲属关系的复杂性（如邓巴（2016）、哈维（2006）等），但是就推理而言，许多语言的亲属称谓系统比较简单，造成对亲属关系推理的困难习焉不察。对汉语来说，由于亲属称谓系统相对复杂，反而很好地揭示了亲属关系推理的复杂性。特别是，当文本中涉及多个亲属之间的关系时，推理起来更加困难。例如：

例（1）已知：小明是我的表哥，老张是我的舅舅。问：老张和小明是什么关系?

这种亲属关系推理涉及多个句子、多个人物，是文本中亲属关系推理的常见形式。从亲属称谓的定义看，“舅舅”有两种可能，“母亲的哥哥或弟弟”。“表哥”的可能性则更多，可以是“母亲的兄、弟、姐、妹的年纪比我大的儿子”也可以是“父亲的姐、妹的年纪比我大的儿子”。例（1）的推理是比较复杂的，答案是“老张可能是小明的父亲、伯父、叔叔、舅舅或者舅母的兄弟。”推理时，除了考虑生育关系、婚姻关系这两种基本的血缘关系外，还涉及性别、长幼等属性等。

2022年11月，OpenAI公司推出了对话式通用人工智能工具ChatGPT，对自然语言处理领域产生了广泛的影响。作为一种生成式人工智能（AI generated content，AIGC）应用，ChatGPT在语言理解、语言生成、知识及常识推理等方面的能力都展现出了卓越的类人表现。但也有缺点，其中之一就是生成信息真实性、准确性不足，以及类似于例（1）这样复杂的亲属关系推理，还无法准确处理（如Figure1）。

Figure 1: ChatGPT对例（1）的两次回答

2023年4月，国家网信办发表了《生成式人工智能服务管理办法（征求意见稿）》，对生成式人工智能服务提出了一系列规范要求和指导意见，其中包括要求生成的内容应该真实准确。“真实”要求防止生成虚假信息，“准确”则包括信息推理不出错。确保生成内容真实准确可以有两种途径：一是从模型上改进，让模型尊重事实、正确推理；二是在必要的时候，对生成的内容包括事实和推理进行二次核查，此时可采取外接核查系统或人工干预的方法。面对亲属关系推理这种推理复杂但问题界定清晰的情况，在必要时外接一个专用的微型的推理系统不失为一个可行的选择。

忱。

第二十二届中国计算语言学大会论文集，第480页-第490页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

学界中，已有不少多个条件下的亲属关系推理系统研究，如靳小龙等（2001）、王树西等（2003）、陈振宇等（2009，2010a，2010b）等。卢达威等（2019）从亲属关系知识表达和推理规则两方面梳理了已有研究，指出现有系统在亲属关系的知识表示方面不能反映亲属关系实质，在推理规则方面缺乏正确性保证，一旦遇到关系较远的推理，或遇到例（1）这种比较复杂的推理，很容易出错。该文基于社会学研究，区分了血缘关系和亲属称谓词两个推理层次，认为以往亲属关系推理都是基于亲属称谓词的推理，而推理的实质应该是血缘关系的推理；推理系统应该先将亲属称谓词映射为血缘关系，基于血缘关系推理完成后，再映射为亲属称谓词。

但是，卢达威等（2019）存在一定的不足：在知识表示方面，该文虽然重新设计了基于血缘关系的知识表示方法，但是没有对这种知识表示的合理性作充分的说明。在推理算法方面，该文只有推理的元规则及推理过程示意图，并未对推理算法展开形式化描写，也缺乏论证和推导，其正确性也未能予以证明。另外，无论是对已有系统还是对该文提出的方法，都缺乏对推理过程的时间复杂度等的效率分析。

本文拟在卢达威等（2019）的基础上，首先将推理规则和推理过程形式化和算法化；进而与基于一阶谓词逻辑的推理系统进行对比，讨论它们与基于血缘关系结构的亲属关系推理的差异；最后讨论亲属关系推理算法的时间复杂度问题；并开展相应的实验验证。

2 亲属关系推理过程和推理规则的形式表达

本节以“小明是我的表哥，老张是我的舅舅，问：老张是小明的什么亲戚？”为例，探讨亲属关系的推理规则问题。其中推理规则分为元规则和具体规则。元规则指导整个推理过程，具体规则配合元规则实现具体的推理任务。这是使用传统的符号系统的方法来构建的系统。区别于基于统计或生成式的模型，对于小型符号系统，我们要求逻辑分析无误，做到100%的推理准确率。

2.1 基础概念

为更好说明亲属关系推理的形式表达，我们将卢达威等（2019）所定义的血缘关系结构和亲属关系式作为基础概念。

1.血缘关系结构

血缘关系结构是通过婚姻关系、生育关系为纽带，以“父亲”“母亲”“孩子”组成的小家庭为基础单元的递归结构。血缘关系结构中，为方便推理，定义了四种基本关系：婚姻关系、生育关系、被生育关系、兄弟姐妹关系。通过四种“基本血缘关系”，构建出血缘关系结构图（如Figure2）。血缘关系结构图是推理的基础。

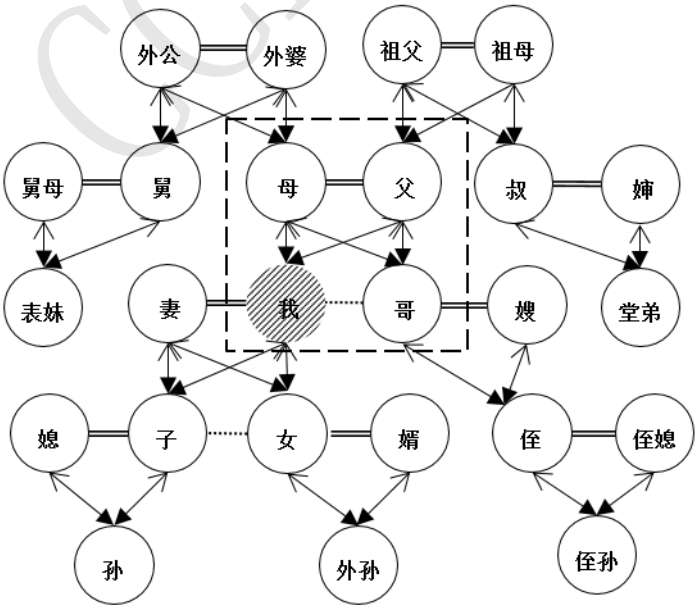


Figure 2: 血缘关系结构图（局部）

2.亲属关系式

卢达威等（2019）没有使用传统的一阶谓词逻辑来表示亲属关系，而是重新构造了亲属关系的知识表示，称为“亲属关系式”，如下：

$$R = s(x_0) + \sum_{i=1}^{n} f(x_i, y_i)$$

R为亲属关系式，其中，s(x_0)是起始标记，代表自己，x_0是性别属性；f(x,y)代表四种“基本血缘关系”之一，这四种“基本血缘关系”包括婚姻关系m(x)、被生育关系p(x)、生育关系b(x,y)和兄弟姐妹关系c(x,y)；参数x是性别属性，x∈{1, -1, 0}，1为“男”、-1为“女”、0为“两可/未知”；参数y是长幼属性，y∈{1, -1, 0}，1为“长”、-1为“幼”、0为“两可/未知”。该公式的含义是，任何一个亲属关系均可以表达为：从自身出发的若干种基本血缘关系的排列。

2.2 亲属关系推理的元规则

基于此亲属关系知识表示和血缘关系结构图，亲属关系推理元规则可以分为四步：

（1）查找定义。利用亲属关系定义，在血缘关系图上查找已知条件的亲属表达式。如上例，以“我”为中心在血缘关系图上找到“小明”结点（我的表哥）和“老张”结点（我的舅舅）。其中“表哥”路径有多个，应该分多种情况计算。

（2）连接。将血缘关系图中所求的两结点连接起来。亲属关系式有向的，例如，所求以“小明”为出发点，则应首先要调用逆关系运算规则，将“我→小明”的有向连接（即“小明是我的...亲戚”），变成“小明→我”的有向连接（即“我是小明的...亲戚”）；然后，以“我”结点作为桥梁，形成“小明→我→老张”的有向连接，得到一条从“小明”到“老张”的亲属关系式（即“老张是小明的...亲戚”）。

（3）约简。寻找所求两结点的最短路径。通过执行亲属关系式的约简规则，删除路径中不必要的关系，求出“小明→老张”的最短路径。

（4）匹配。根据最短路径上的基本血缘关系序列，匹配亲属关系定义，得到“老张”对“小明”的亲属称谓。

在以上亲属关系推理元规则中，主要涉及两类规则：一类是步骤（2）中亲属关系的逆关系运算规则；另一类是步骤（3）中寻找最短路径的约简规则。以下分别展开说明。

2.3 亲属关系的逆关系运算规则

亲属关系是有向的。血缘关系结构图中的两个结点A和B，从A到B所经过的血缘关系，与从B到A所经过的血缘关系是不同的。主要体现在三个方面：亲属关系式中基本血缘关系序列排列相反；生育关系b和被生育关系p不对称；基本关系中属性参数的含义不同。因此，我们要对逆关系运算专门定义。设原亲属关系式R为：s(0)+f(x1)+g(x2)+h(x3)，其中f, g, h是四种基本血缘关系之一，如Figure3，逆关系运算分三步：

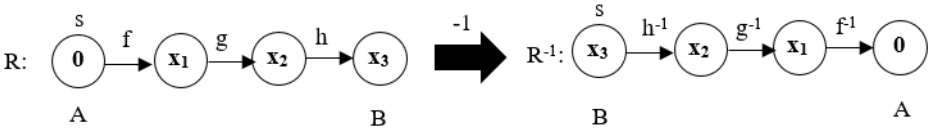


Figure 3: 亲属关系式R的逆关系运算

（1）逆关系R^{-1}中，将原亲属关系式R的基本血缘关系的排列f+g+h倒置，即h+g+f。

（2）对每个基本血缘关系分别求逆：婚姻关系m和兄弟姐妹关系c的逆关系是自身，生育关系和被生育关系互为逆关系。即m^{-1}=m，c^{-1}=c，b^{-1}=p，p^{-1}=b。

（3）重新安排各基本血缘关系的属性。性别属性取值为基本血缘关系的终点结点的性别属性值，即本例中，将求逆前的f(x1), g(x2), h(x3)分别变成f^{-1}(x0), g^{-1}(x1), h^{-1}(x2)。对于长幼属性，则进行属性值取反，原值为0时不变。

经过以上三步，亲属关系式R的逆关系式R^{-1}为：s(x3)+h^{-1}(x2)+g^{-1}(x1)+f^{-1}(x0)。

2.4 亲属关系式约简规则

2.4.1 约简规则完备性保证

所谓约简，就是在亲属关系式中，用尽可能少的基本血缘关系来表达。亲属关系式的约简，是亲属关系推理中最重要的一步。亲属关系式的约简的完备性，是基于以下两个发现作为保证的：

（1）关系的约简只可能出现在“父亲”“母亲”“孩子”组成的“血缘关系基础单元”之内。跨越基础单元的两个关系不存在被约简的可能。比如“父亲的父亲（p+p）”不在同一基础单元内，不能被约简。

（2）所有约简总能规约为两个关系之间的约简，不存在必须包含3个关系或以上的约简。例如，“兄弟的父亲的妻子（c+p+m）”，可以通过两两约简，先将“兄弟的父亲（c+p）”约简成“父亲（p）”，再将“父亲的妻子（p+m）”约简成“母亲（p）”。

以上两个特征说明：只要充分刻画一个血缘关系基础单元内两两关系之间的约简规则，就能够保证整个亲属关系式能够通过这些规则，达到最简。

2.4.2 8种约简规则

对于4种基本血缘关系，两两组合共有16种组合，其中，有8种两两基本关系组合恰好落在同一个血缘关系基础单元中，可被约简成一种基本血缘关系；有8种不在同一个基础单元中，不可被约简。分类如Table1所示。

Table 1: 可被约简的和不可被约简的两两亲属关系组合

可被约简的两两亲属关系	不可被约简的两两亲属关系
m+m	m+p
m+b	m+c
b+p	b+m
b+c	b+b
p+m	p+p
p+b	p+c
c+p	c+m
c+c	c+b

Table1右栏是不能约简的8种基本血缘关系组合，暂时不在我们的讨论范围。左栏是8种可约简的基本血缘关系组合，恰好覆盖了一个血缘关系基础单元内Figure2）所有可能的两两关系组合。左栏的两个基本血缘关系均可以约简为一个或0个基本血缘关系，约简规则详列于Table2。（注：在详细约简条件中，“→”号左边的关系和属性是前提条件，右边是约简的结果；f(x)是前一个基本血缘关系，表示约简时会受到前一个基本血缘关系f的性别属性x的影响；w, x, y, z表示属性值，可以是1、-1、0；a, e表示非0的属性值，只能是1、-1；*表示约简时不需要考虑的属性值）

由于约简细节较为复杂，我们仅以公式（2.3）b+p → 0 / m为例。该公式的字面含义是，“(我的)孩子的父/母（b+p）”可能是“我自己（0）”或者“(我的)配偶（m）”。根据公式中生育关系b、被生育关系p以及前置关系f的性别属性，分为4种情况：

公式（2.3-1）：f(a) + b(*, *) + p(a) → f(a)

表示若前置关系f的性别已知为a（a不为0），父/母亲p的性别也已知为a（即两者相等且不为0），则“(f的)孩子的父/母”必然是“（f）自己”，性别为a。

公式（2.3-2）：f(a) + b(*, *) + p(-a) → f(a) + m(-a)

表示若前置关f的性别已知为a，p的性别也已知为-a（即性别恰好相反），则“(f的)孩子的父/母”必然是“(f的)配偶”，且性别与p相同。

公式（2.3-3）：f(a) + b(*, *) + p(0) → f(a) / f(a)+m(-a)

表示若前置关f的性别已知为a，但p的性别属性值为0（两可），则“(f的)孩子的父/母”可能是“（f）自己”，性别为a，也可能是“(f的)配偶”，性别为-a。

公式（2.3-4）：f(0) + b(*, *) + p(x) → f(x) / f(-x)+m(x)

表示若前置关f的性别未知，则无论p的性别如何，“(f的)孩子的父/母”都有两种可能：可能

Table 2: 亲属关系约简规则

序号	规则简写	公式含义	公式编号	详细约简条件
1.	m+b→b	(我的)配偶的孩子 →(我的)孩子	(2.1)	$m(*)+b(x,*) \rightarrow b(x,*)$
2.	m+m→0	(我的)配偶的配偶→(我)	(2.2)	$m(x)+m(y) \rightarrow 0$ 且 x = -y
3.	b+p→0/m	(我的)孩子的父/母亲 →(我)或(我的)配偶	(2.3)-1) -2) -3) -4)	$f(a)+b(*,*)+p(a) \rightarrow f(a)$ $f(a)+b(*,*)+p(-a) \rightarrow f(a)+m(-a)$ $f(a)+b(*,*)+p(0) \rightarrow f(a) / f(a)+m(-a)$ $f(0)+b(*,*)+p(x) \rightarrow f(x) / f(-x)+m(x)$
4.	b+c→b	(我的)孩子的兄弟姐妹 →(我的)孩子	(2.4)	$b(*,y)+c(x,z) \rightarrow b(x,k(y,z))$ $k(x,y)=\begin{cases}1, & x=y=1\\ -1, & x=y=-1\\ 0, & \text{other}\end{cases}$
5.	p+b→0/c	(我的)父/母亲的孩子 →(我)或(我的)兄弟姐妹	(2.5)-1) -2) -3) -4)	$f(a)+p(*)+b(-a,x) \rightarrow f(a)+c(-a,x)$ $f(x)+p(*)+b(y,e) \rightarrow f(x)+c(y,e)$ $f(0)+p(*)+b(x,0) \rightarrow f(x) / f(0)+c(x,0)$ $f(a)+p(*)+b(x,0) \rightarrow f(a) / f(a)+c(x,0)$
6.	p+m→p	(我的)父/母亲的配偶 →(我的)父/母亲	(2.6)	$p(x)+m(y) \rightarrow p(y)$ 且 x = -y
7.	c+p→p	(我的)兄弟姐妹的父/母亲 →(我的)父/母亲	(2.7)	$c(*,*)+p(x) \rightarrow p(x)$
8.	c+c→0/c	(我的)兄弟姐妹的兄弟姐妹 →(我)或(我的)兄弟姐妹	(2.8)-1) -2) -3)	$f(a)+c(*,x)+c(-a,y) \rightarrow f(a)+c(-a,k(x,y))$ $f(w)+c(*,a)+c(z,a) \rightarrow f(w)+c(z,a)$ $f(w)+c(*,x)+c(z,y) \rightarrow f(z) / f(w)+c(z,k(x,y))$ $k(x,y)=\begin{cases}1, & x=y=1\\ -1, & x=y=-1\\ 0, & \text{other}\end{cases}$

注：“规则简写”栏中的基本亲属关系与“详细约简条件”栏中的波浪线对应。

是“（f）自己”，此时性别与p一致，为x；或可能是“（f的）配偶”，性别与p一致，并推论出f的性别与p相反，为-x。

以上4个子公式的属性组合以及覆盖了公式（3.3）f+b+p所有可能的性别属性组合，没有遗漏。其他的约简公式与此类似。

2.5 亲属关系式的约简流程

对于亲属关系式R=s+f1+f2+...+fn（f代表基本血缘关系）约简流程大致描述如下。

从左到右逐一扫描亲属关系式R中当前一个及后续一个基本血缘关系（如f_i+f_{i+1}），约简情况有三种：

（1）若不能约简，则直接执行下一个两两基本血缘关系组合（如$f_{i+1}+f_{i+2}$）；

（2）若能够约简，则约简并以约简后的结果为起点，扫描后续两两基本血缘关系组合；

（3）若执行步骤（2）时产生新了亲属关系式分支，则记录下来，完成当前亲属关系式约简后再对新的分支进行约简（重复步骤1-3）。

以上约简流程和Table2的8个约简规则，有效保证了亲属关系式能最大程度约简，从而求得两个结点之间的最短路径。然而，除了以上两类规则外，推理过程中，还有一些细节需要定义相应规则。如亲属关系式连接规则和属性核查方法等，篇幅所限，不作赘述。

2.6 亲属关系推理过程举例

下面以“已知：小明是我的表哥，老张是我的舅舅，问：老张和小明是什么关系？”为例，说明推理过程和规则的使用方法。根据2.2节推理的元规则（①查找—②连接—③约简—④匹配）流程如Figure4。

详细解释如下。

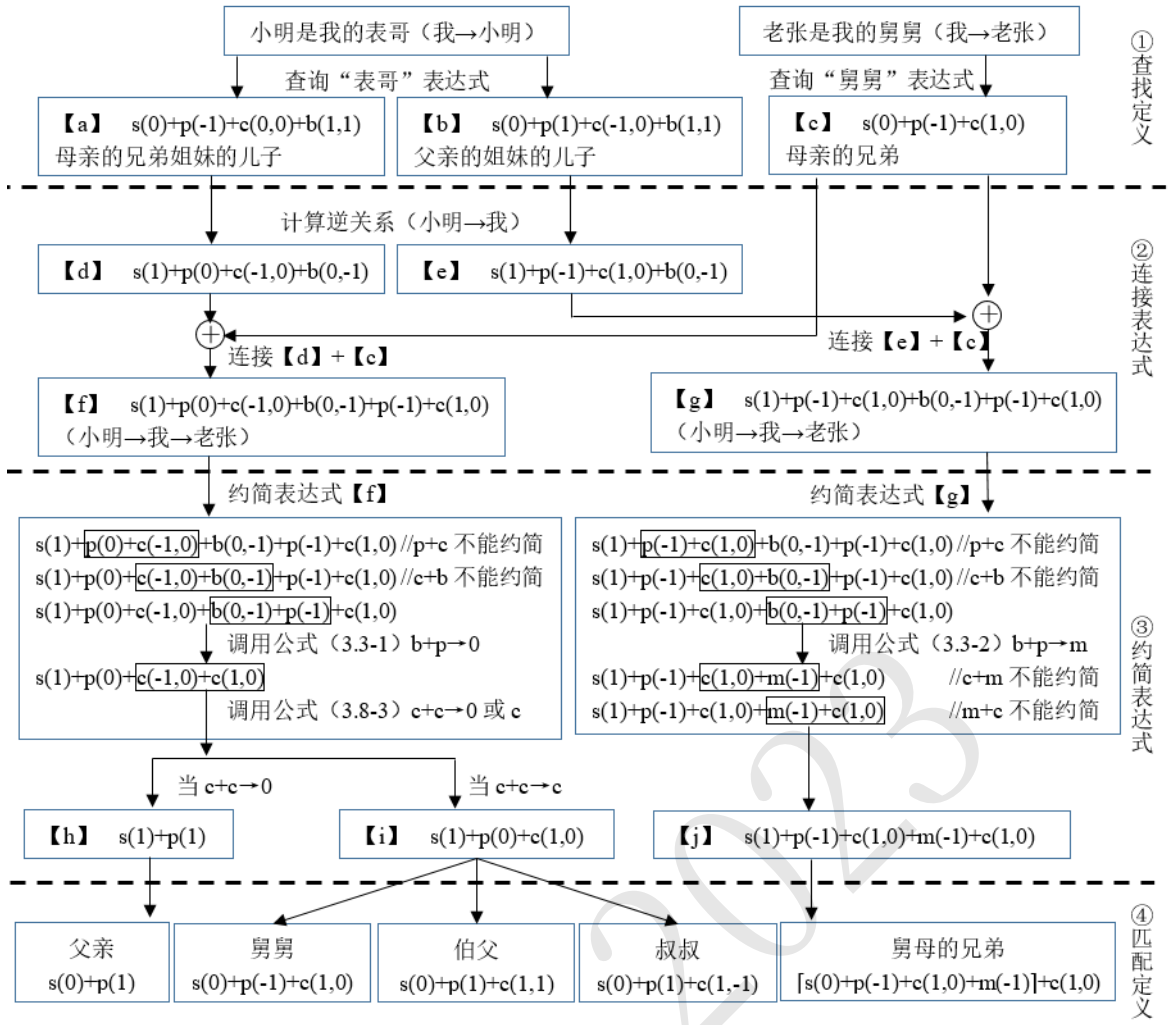


Figure 4: 亲属关系推理流程

（一）查找定义：表哥和舅舅

查询亲属关系式可知，“表哥”的亲属关系式有两种（Figure4的公式【a】和【b】），“舅舅”的亲属关系式有一种（公式【c】）。

（二）连接表达式：连接小明和老张

我们以“老张是小明的XXX亲戚”为所求，则“小明”是连接的起点，“老张”是终点。形成“小明→老张”的亲属关系式。“根据题目，“小明”和“老张”的媒介是“我”，所以连接的最终结果是“小明→我→老张”。分为逆关系变换和连接两步。

逆关系变换：已知条件中，关于“表哥”的定义有两个，【a】和【b】，其方向都是“我→小明”的，要首先调用逆关系规则，将【a】和【b】分别变成【d】和【e】，使亲属关系的方向从“我→小明”变成“小明→我”。

连接：利用亲属关系式连接规则将【d】和【e】“小明→我”与【c】“我→老张”连接起来，形成【f】和【g】“小明→我→老张”。

（三）约简表达式：约简小明到老张的亲属关系式

由于“表哥”的亲属关系定义关系有两种，因此连接后“小明→老张”的亲属关系式也有两个，要分别约简。

对于【f】s(1)+p(0)+c(-1,0)+b(0,-1)+p(-1)+c(1,0)，先从左到右两两基本血缘关系开始约简，过程如下：

（1）检查p(0)+c(-1,0)，查找约简规则发现p+c不能被约简。

（2）检查下一组两两关系c(-1,0)+b(0,-1)，发现c+b也不能约简。

（3）检查下一组两两关系b(0,-1)+p(-1)，发现b+p是可以约简的，根据约简规则（2.8），

第二十二届中国计算语言学大会论文集，第480页-第490页，哈尔滨，中国，2023年8月3日至5日。

由于p的性别属性值-1与前置关系c的性别属性值相同，应该调用公式（2.8-1），删除b(0,-1)和p(-1)，意义是“某人（女）孩子的母亲”是“她自己”。

（4）因为原关系经过了约简删除，重新核查c(-1,0)+c(1,0)，发现c+c是可以简化的，根据约简规则（2.8），前置关系的性别不确定（值为0）时，应该调用公式（2.8-3），产生两种可能亲属关系式，一种是删除c(-1,0)和c(1,0)，指“某人的姐妹的兄弟”是“某人自己（男）”；另一种是删除c(-1,0)，指“某人的姐妹的兄弟”是“某人（性别不确定）的兄弟”。

（5）新产生的两条亲属关系式均已达到关系式末尾，也没有新的关系式需要约简。约简结束。

（6）经过约简，产生了两条约简结果，即：s(1)+p(1)和s(1)+p(0)+c(1,0)，对应公式【h】和【i】。

对于【g】，约简流程相似，只是调用的规则稍有不同，此处不在赘述。约简后的亲属关系式是公式【j】：s(1)+p(-1)+c(1,0)+m(-1)+c(1,0)。

（四）匹配亲属关系定义

经过约简，得到“小明→老张”的亲属关系式有三个，【h】【i】【j】。

通过匹配亲属关系式的定义可知，【h】是“父亲”，即“老张是小明的父亲”。

而【i】的字面含义是“老张是小明的父亲或母亲的兄弟”，根据亲属关系定义有多种可能。“母亲的兄弟”是“舅舅”，“父亲的哥哥”是“伯父”，“父亲的弟弟”是“叔叔”。

再看【j】。其含义是“母亲的兄弟的妻子的兄弟”，这已经超出了一般亲属关系定义的范围。根据最大匹配原则，“母亲的兄弟的妻子”可以匹配为“舅母”。所以，【j】用两段亲属关系来描述，匹配为“舅母的兄弟”。

由此，我们知道，当“小明是我的表哥”，“老张是我的舅舅”，那么“老张”有可能是“小明”的“父亲”“舅舅”“伯父”“叔叔”或“舅母的兄弟”。

3 基于血缘关系结构的亲属关系推理的优势

3.1 亲属关系知识表示的优势

我们以“舅舅”为例，分析亲属关系式与一阶谓词逻辑的不同。“舅舅”亲属关系式为：s(0)+p(-1)+c(1,0)，而使用传统的一阶谓词逻辑，可能表示为：亲子(z1，x)∧亲子(z2,z1)∧亲子(z2,y)∧女(z1)∧男(y)。

首先，一阶谓词逻辑的表达一般需要将经过的结点（z1，z2，z3）假设出来，并作为变量。在推理时，需要对这些参数进行合一运算，而合一运算时可能涉及若干回溯，增加了推理的时间复杂度；而亲属关系式则直接使用了基本血缘关系，隐藏了结点，故不需要假设名称，也不需要进行合一运算。

第二，亲属关系式将属于结点的性别属性、长幼属性等附着到基本血缘关系中，更直观，使得编写和检查亲属关系定义的效率更高，错误率更低。

第三，一阶谓词逻辑一般只用一个特征表示生育关系，如“亲子(x,y)”，而亲属关系式区分了生育关系b(x,y)和被生育关系p(x)，实际上二者是不对称的。从上节Table2的公式（2.3）和（2.5）的区别可知，“我的孩子的父/母亲”（b+p）是“我”或“我的配偶”，而“我的父/母亲的孩子”（p+b）是“我的兄弟姐妹”。这一区别，亲属关系式可以通过b和p的顺序来表达出来。而对于只有“亲子(x,y)”的一阶谓词逻辑，“我的孩子的父/母亲”应表示为“亲子(x,z)∧亲子(y,z)”，“我的父/母亲的孩子”表示为“亲子(z,x)∧亲子(z,y)”，二者的区别隐含在结点参数的顺序中。这增加了计算和错误检查的难度。当关系复杂或条件众多时，容易引起失误。

第四，亲属关系式能够让所有亲属关系称谓的集合变成一个“可列集”。理论上，只要对四种基本血缘关系及其属性，逐一排列和穷举，就能无遗漏地列出人类所有的亲属关系式。对于推理系统来说，就能够明确亲属关系推理系统的推理范围，构造一个在此范围内的完备的亲属关系知识体系，甚至发现人类知识的盲区。例如，虽然《亲属称呼辞典》中列举了10大类亲属333小类的亲属称谓，但是依然无法确认这是不是亲属关系的全部。如果我们利用亲属关系，就能确定一定的推理范围，逐一列出此范围内的亲属称谓词，就能一一指出在此范围内，有哪些亲属关系是缺乏称谓的，可能会引起对人类学等方面的思考。

3.2 亲属关系推理规则的优势

一般基于一阶谓词逻辑的亲属关系推理，其实质是基于亲属称谓词进行推理的。具体而

言，其推理的过程需要不断与亲属称谓词的定义进对比和进行合一运算，直到找到合适的亲属关系作为答案。这样的做法，虽然大多数情况能计算出正确答案，但一来效率低，时间复杂度高；二来，若亲属关系称谓的定义不完备，很可能因缺乏定义而无法判断当前的逻辑表达式是需要回溯还是推理失败，从而导致推理系统出错。而基于一阶谓词逻辑的知识表达，无法解决亲属关系知识体系的完备性问题，往往靠人的经验来列出亲属关系称谓的定义。当关系较远的时候，人的经验很可能不足，造成定义的不完备。三来，当需要推理的关系复杂的时候，不能保证这类推理的有效性。

也有的系统不是直接匹配，而是先进行逻辑表达式的约简（如陈振宇，2010）。问题是，无法保证提出的约简规则是否覆盖了所有可以约简的情况。这是由于约简目标不明确造成的。

与基于血缘关系结构的推理则不同，基于血缘关系结构的推理系统先把所有亲属关系都映射到血缘关系结构图中，然后在血缘关系结构上进行推理；而且，这样的推理是有明确约简目标的，即找到两个结点的最短路径。推理过程中，我们先论证了所有关系的约简都能归结为两两基本亲属关系的约简；进一步，我们找到了两两关系约简的所有情况，并有针对性设计了约简规则。最后，再把两个结点的最短血缘关系路径映射到亲属称谓词中。

这样做有几个好处：一是，由于有明确的约简目标，能够检验亲属关系的推理结果的准确性；二是效率高，由于推理过程是基于血缘关系结构图的，不需要反复进行比对亲属关系称谓词的合一运算；三是，即使偶有亲属关系称谓词的定义不完备，也不影响推理的过程，因为此时的血缘关系以及最简，无论是否存在亲属称谓的定义，都可以结束推理。不过实际上，由于亲属关系式的可列性，定义的完备是可以保证的。

4 亲属关系推理的时间复杂度分析

对基于血缘关系的亲属关系推理的时间复杂度，我们对元规则的四个步骤（定义、连接、约简、匹配）逐一分析。由于这几个步骤是串行的，推理过程的时间复杂度取决于其中时间复杂度最大的部分。

（一）查找定义的时间复杂度

查找定义是将亲属称谓词转换为亲属关系式，这一过程只需要查表即可，因此是常数时间复杂度的，记为O(1)。

（二）连接的时间复杂度

连接操作是将不同的条件，根据起始结点和所求目标，将不同的亲属关系式进行连接。个别亲属关系式需要先进行逆关系操作。逆关系操作和连接操作都是常数时间复杂度的，总体还是常数时间复杂度，记为O(1)。

（三）约简的时间复杂度

约简操作是对目标亲属关系式求最短路径。由于我们发现亲属关系式总可以通过两两约简达到最简状态，因此，这一过程是不需要回溯的，而且每一次约简，只需要匹配2.3节的约简规则，进行约简操作，这个过程是常数时间的。就是说，只需要对目标亲属关系式的每个基本血缘关系扫描一遍，即可得到最简亲属关系式。对于N个基本血缘关系的亲属关系式，时间复杂度为O(N)。

（四）匹配亲属称谓词的时间复杂度

根据最简亲属关系式，我们需要在亲属称谓定义表中找到匹配的亲属称谓。这一过程也是常数时间复杂度的，记为O(1)。

综上所述，四个过程的时间复杂度为O(1)+O(1)+O(N)+O(1)，则推理系统的时间复杂度是O(N)，即线性时间复杂度。N是未约简的亲属关系式中基本血缘关系的个数，事实上N的规模非常有限。整个推理系统的复杂度是很低的。而且空间复杂度也很低，仅需要存储亲属称谓的定义集和推理规则集，推理过程中使用的空间使用量是一个常数，非常适合于外接到任意系统进行亲属关系的实时推理。

5 亲属关系推理实验

在前文的理论分析的基础上，本节将开展亲属关系推理的实验，对推理的正确性和运行效率进行检验。

首先，我们构造了一个亲属关系称谓集。为了描写亲属关系称谓集，我们定义先“N阶亲属关系”的概念。“N阶亲属关系”指该称谓的亲属关系式中基本亲属关系数量为N（如Table3）。

Table 3: 各阶亲属关系

亲属关系类型	数量	亲属称谓举例	亲属关系式	含义
一阶亲属关系	16	女儿	s(0)+b(-1, 0)	女儿
二阶亲属关系	35	祖父	s(0)+p(1)+p(1)	父亲的父亲
三阶亲属关系	86	堂哥	s(0)+p(1)+c(1,0)+b(1,1)	父亲的兄弟的儿子（比我年纪大）
四阶亲属关系	58	堂妹夫	s(0)+p(1)+c(1,0)+b(-1,-1)+m(1)	父亲的兄弟的女儿（比我年纪小）的丈夫

基于N阶亲属关系的概念，亲属关系称谓集中，我们定义了一阶亲属关系称谓16个，包括配偶（丈夫、妻子）、双亲（父亲、母亲）、孩子（儿子、女儿）、兄弟姐妹等；二阶亲属关系称谓35个，如祖父、孙子、舅舅、女婿等；三阶亲属关系称谓86个，如曾祖父、曾孙、表兄、亲家等；四阶亲属关系称谓58个，如堂舅、表嫂、姑表舅父、姨表姨母等。由于四阶亲属关系已经很多称谓不是很熟悉了，所以不再定义五阶亲属关系。值得说明的是，一阶和二阶已经无遗漏地穷举了所有亲属关系表达式。可以穷举、令所有亲属关系变成一个可列集，正是亲属关系式的优势之一。而三阶和四阶则选择了一些典型的为人熟知的亲属关系进行定义。总的来说，本文的亲属关系称谓集共定义了195个亲属关系称谓。

第二，我们将所有关系的两两组合进行全排列，如“爸爸的堂哥”“堂妹夫的表姐”等，则195个称谓共构造出195×195=38025个亲属关系推理式。

第三，使用本文的算法对这38025个亲属关系式进行推理。对于推理结果，我们随机抽取了10%进行人工验证，即3802个推理结果。这些推理结果包括两类，如果推理后得到的亲属称谓在我们的亲属称谓集中，则直接匹配亲属称谓名；若超过了亲属关系称谓集的范围，则先按最大程度匹配亲属关系称谓，再在此基础上进一步匹配其余部分。例如，“父亲的曾祖母”，由于推理结果是“曾祖父的母亲”。

由于是基于规则的推理，经检验，推理结果达到100%的正确率。这38025个亲属关系推理在普通个人计算机中（i7-4710MQ，2.5GHz的CPU，16G内存）运行时间为7秒左右。

可见，本文亲属关系推理的正确性和有效性得到了实验的支持。

6 结语

亲属关系推理是一个经典的常识推理问题，看似简单实质复杂。以往的亲属关系推理系统，或者不能解决文本中多个亲属关系作为已知条件的亲属关系推理问题，或者对复杂的亲属关系推理容易出错，且计算量大，难以实用。

本文基于卢达威等（2019），使用了传统的符号系统的方法来构建亲属关系推理系统。以基于血缘关系的亲属关系式为基础，着重构建了推理规则和推理过程的形式系统；并探讨了本文的基于血缘关系结构的亲属关系推理系统，相对于一阶谓词逻辑的推理系统的优势；还分析了时间复杂度等理论问题。本文虽然使用传统的规则方法，但从理论上能保证100%的推理正确率，并通过专门设计的亲属关系表达方式和推理规则，保证了推理的高效性。相当于在理论上和实践上彻底解决了亲属关系推理的问题。

本文所构造的亲属关系推理系统，是一个封闭、准确、小巧、离线的系统，能够独立推理，也能够非常方便得接入ChatGPT等语言处理系统中，为系统提供精确推理的服务。

值得说明的是，本文的亲属关系推理是基于“血缘关系”这一人类共有的关系构建的，因此是跨语言的。不仅汉语亲属关系推理适用，其他语言只要修改亲属关系定义部分，就可直接进行推理。另外，本文的亲属关系推理，实际上是利用血缘关系构造了一个亲属关系概念的语义空间，该语义空间有四个维度：m, b, p, c，实现了语义空间内的运算。这为其他语义问题的推理，提供一个借鉴。

参考文献

陈振宇, 袁毓林, 张秀松, 周强. 2009. 亲属关系的逻辑意义及其自动推理. 计算机工程与应用, 45(16): 43–47.

陈振宇, 袁毓林, 张秀松, 周强. 2010a. 一种基于大知识库的亲属关系自动推理模型. 中文信息学报, 24(3): 117-124.

陈振宇, 袁毓林. 2010b. 汉语亲属关系的语义表示和自动推理. 中国语文, (1): 44-56.

邓巴著（余彬译）. 2016. 人类的演化. 上海: 上海文艺出版社.

官赛萍, 靳小龙, 贾岩涛, 王元卓, 程学旗. 2018. 面向知识图谱的知识推理研究进展. 软件学报, 29(10): 2966-2994.

哈维兰著（瞿铁鹏，张钰译）. 2006. 文化人类学*(*第*10*版*)*. 上海: 上海社科院出版社.

靳小龙, 魏旺强. 2001. 基于常识的亲属关系推理模型. 计算机工程与应用, 37(17): 83-85.

卢达威, 袁毓林. 2019. 基于血缘关系结构图的亲属关系推理系统研究与实现. 中国社会科学, (11): 25-43.

诺姆·乔姆斯基著（曹道根，胡鹏志译）. 2015. 语言的科学. 北京: 商务印书馆.

王树西, 刘群, 白硕. 2003. 一个人物关系问答的专家系统. 广西师范大学学报*(*自然科学版*)*, (1): 31-36.

基于深加工语料库的《唐诗三百首》难度分级

黄宇宇[1] ， 陈欣雨[1] ， 冯敏萱[12*] ， 王禹诺[1] ， 王蓓原[1] ， 李斌[12]
[1] 南京师范大学文学院
[2] 南京师范大学语言大数据与计算人文研究中心
hyuyuz@163.com

摘要

为辅助中小学教材及读本中唐诗的选取，本文基于对《唐诗三百首》分词、词性、典故标记的深加工语料库，据诗句可读性创新性地构建了分级标准，共分4层，共计8项可量化指标：字层（通假字）、词层（双字词）、句层（特殊句式、标题长度、诗句长度）、艺术层（典故、其他修辞、描写手法）。据以上8项指标对语料库中313首诗评分，建立基于量化特征的向量空间模型，以K-means聚类算法将诗歌聚类以对应小学、初中和高中3个学段的唐诗学习。

关键词： 《唐诗三百首》 ； 语料库 ； 难度分级 ； 诗句可读性 ； 文本长度

The difficulty classification of ' Three Hundred Tang Poems ' based on the deep processing corpus

Yuyu Huang[1] ， Xinyu Chen[1] ， Minxuan Feng[12*] ， Yunuo Wang[1] ， Beiyuan Wang[1] ， Bin Li[12]
[1] College of Arts Nanjing Normal University
[2] Center of Language Big Data and Computational Humanities
hyuyuz@163.com

Abstract

In order to assist the selection of Tang poetry in primary and secondary school textbooks and reading books, based on the deep processing corpus of word segmentation, part of speech and allusion markers of ' 300 Tang poems ', this paper innovatively constructs a grading standard according to the readability of verses, which is divided into 4 layers, a total of 8 quantifiable indicators : font layer (interchangeable words), word layer (double-word words), sentence layer (special sentence pattern, title length, verse length), art layer (allusions, other rhetoric, description techniques). According to the above eight indicators, 313 poems in the corpus are scored, and a vector space model based on quantitative features is established. The K-means clustering algorithm is used to cluster the poems to correspond to the Tang poetry learning of primary school, junior high school and senior high school.

Keywords: ' 300 Tang poems ' , Corpus , Difficulty classification , Readability of verse , Text length

*通讯作者
基金项目：江苏省社科基金项目(20JYB004)；古籍工作重点课题(22GJK006)；国家语委项目(YB145-41)；深圳爱阅基金会(儿童国学经典读物的分级阅读研究)

1 引言

唐诗数量丰富、艺术成就高，是中国优秀文化成果。据统计，我国的唐诗选本约有六百多种，而《唐诗三百首》是其中影响最大、流传最广、读者最多的选本，“风行海内，几至家置一编”，被视为唐诗入门启蒙读物首选(王东旭, 2013)。《唐诗三百首》是清乾隆蘅塘退士孙洙所编写的，本意是用来作为家塾读物训育儿童。《唐诗三百首》的选编理由并不只是诗歌的艺术成就，还考虑到了受众的接受情况、应试的需要、情感的“温柔敦厚”和教材的整体等，因此成就了《唐诗三百首》雅俗共赏、老少皆宜的高品位普及选本的地位。然而，《唐诗三百首》按体裁分章节进行编写，并未对诗歌进行难度分级，可能存在不符合学习者认知规律等问题。

根据学习认知规律，不同年龄段人群的阅读认知水平存在差异，学习积累古诗词是一个由繁到简的过程，不符合认知阶段的过难或过易的诗歌都可能会影响学习积极性，降低学习效果。而分级阅读是基于各个年龄阶段人群的认知水平，选择、提供适合不同年龄阶段阅读需要的文本。柳明烨(2021)因为阅读符合认知发展规律，因此有利于提高阅读古诗的效率，发展学习能力。

国内外对分级阅读的研究成果较为丰富，但是国内分级研究关注的多为现代汉语文本，对古代汉语文本的分级研究极少。针对文言文分级，张秋玲(2010)从2010年开始构建了文言文难易度评量的数学模型，首次为文言文难易度评量提出路径，经过两次修订，张秋玲等(2022)数学模型的相关系数提高了0.112，提高了数学模型评量文言文难易度的信效度。白瑞芬(2017)根据儿童心理学将分级阅读分为5个阅读阶段，提出了目前国学经典改编的问题和改进措施。针对古诗词分级，研究则更加少，柳明烨(2021)计算了课内外诗词的可读性，基于部编版小学语文教材的古诗词分级确立古诗词分级标准，对课外古诗词按照年级进行重新判别，具有开创性。但是对“诗句可读性”的指标制定颗粒度不够，不能较全面地覆盖特征。

因此，本文以赵昌平的《唐诗三百首全解》为底本(2006)，以基于条件随机场模型（CRF）的随园古汉语分词与词性标注系统深化加工处理，校勘与标注形成共计24540字、17896词规模的初始加工语料。后在此基础上标记唐诗的各类句式、修辞，建成《唐诗三百首》深加工语料库。基于该深加工语料库，本文量化影响可读性的维度，细化标注规则，设定特征权重，制定更为科学的古诗词分级标准和算法。

2 字层分级

2.1 通假字

《唐诗三百首》所使用的字都较为简单，生僻字较少。但是，唐诗诗句较为短小，距今时代久远，与目前的语言面貌不同，存在较多通假字、古今字和异体字现象，这些用例不符合当前人们的语言使用习惯，因此，字层面是影响《唐诗三百首》可读性的重要因素。

狭义的通假字与古今字和异体字有着千丝万缕的联系，但是并不等同。通假字是本有其字的假借，如“蚤”与“早”；古今字指同一个字在不同时代用不同字表示，如“莫”和“暮”；异体字指音义皆同、形体不同的两个字，如“泪”和“淚”(卢烈红, 2007; 李国英, 2007)。广义的通假字包括狭义通假字、古今字和异体字。本文为了便于标注和后续分级处理，使用的是广义通假字的定义。

影响通假字难度的因素包括出现次数和出现频率。出现次数指的是在一首古诗中，广义通假字出现的个数，出现个数越多，难度越大。出现频率指的是在古汉语文献中，该通假字的使用频率，使用频率越高，难度越小。赋分规则如下：出现次数上，一首古诗中出现N次通假字则计为N分。出现频率上，以使用频率为标准，将古汉语词义数据库中的通假字划分难度级别，难度分数为1-4分。再将出现次数和出现频率按照4：6的权重进行加权计算，根据所得分数将古诗分为0-4分。

3 词层分级

3.1 双字词

魏晋时期，汉字出现大量单音节词双音化的现象，发展到唐代，双字词数量较为丰富。双字词的语素按照组合规则划分，包括并列、偏正、主谓等关系，意义与两个语素也不尽然相同，双字词可能会发展出不同的意义。因此，理解双字词也是影响古诗阅读难度的重要因素，选定从语境的角度对诗歌这种较为封闭的文本中出现的双字词的特定语义理解进行难度分析。

第二十二届中国计算语言学大会论文集，第491页-第500页，哈尔滨，中国，2023年8月3日至5日。

立足于语境同语言的关系，可以分出“言内语境”(linguistic context)和“言外语境”(extra-linguistic context)两大类。言内语境，即文章或言谈中的话题的上下文或上下句，一般来说，对话语的理解依据是上文，听话人或读者对上文或上句作出推理，说话人然后又进一步说明，这种说明又成为听话人理解说话人意图的依据。各种语言语境的正确把握，对正确理解话语有着重要的作用。在有些场合，较小的语言环境不能解决问题，必须考虑较大的非语言环境。非语言环境指话语所发生的语言之外的环境，非语言环境也可称为情景上下文，它从各个方面影响着词的意义，如社会背景、语言情景、具体事件以及讲话方式等等。对各种言外语境的正确把握，对正确理解话语有着重要的作用。

在对《唐诗三百首》双字词进行理解难度的标分时主要以是否需由要“言外语境”和需要“言外语境”的程度进行0、1、2三个分数阶段的划分。其中分数0是指诗中出现的双字词在“言内语境”即可较为通顺的进行意义上的理解，不需要“言外语境”即可理解的双字词，这里所指的“言内语境”也就是诗中话题的上下文或上下句，一般来说，对话语的理解依据是上文，听话人或读者对上文或上句作出推理，说话人然后又进一步说明，这种说明又成为听话人理解说话人意图的依据。各种语言语境的正确把握，对正确理解话语有着重要的作用。分数0在三个分段中是数量最多的，共计2865个；分数1的双字词则是需要“言外语境”辅助理解，多有基于“言外语境”的引申含义只依靠“言内语境”会对造成理解偏差歧义，在有些场合，较小的语言环境不能解决问题，必须考虑较大的非语言环境。非语言环境指话语所发生的语言之外的环境，非语言环境也可称为情景上下文，它从各个方面影响着词的意义，如社会背景、语言情景、具体事件以及讲话方式等等，共361个；分数2的双字词是基于“言内语境”基本不能达意，必须借助“言外语境”进行理解的双字词，共56个。

Table 1: 《唐诗三百首》双字词各等级数量

双字词等级描述	等级划分	《唐诗三百首》中双字词数量
在“言内语境”即可较为通顺的进行意义上的理解	0	2865
需要“言外语境”辅助理解	1	361
基于“言内语境”基本不能达意，必须借助“言外语境”进行理解的双字词	2	56

4 句层分级

4.1 特殊句式

文言特殊句式，一般指的是文言文中不同于现代汉语表达习惯的某些特殊的句式，主要有判断句、被动句、省略句和倒装句等。唐诗诗句字数较短，每句多在5-7个字之间，且受到篇幅和格律的限制，出现省略句和倒装句的用例较多。另外，省略句省略句子成分，倒装句改变句子顺序，对唐诗可读性的影响较大。故将特殊句式中省略句和倒装句作为难度分级的指标。

常见的省略句类型有五种，包括省略关联词、省略介词、省略动词、省略比喻词和互文省略。前两种省略类型对意义影响不大，如“我歌月裴回，我舞影零乱”是“我歌（而）月裴回，我舞（而）影零乱”（李白《月下独酌》）的省略关联词形式。省略动词对意义影响一般，如“经冬犹绿林”是“经冬犹绿（满）林”（张九龄《感遇四首二》）的省略动词形式。后两种省略类型对意义影响很大，如“万事随转烛”是“万事（如）随转烛”（杜甫《佳人》）的省略比喻词形式。常见的倒装句类型有五种，包括主谓倒装、宾语前置、状语后置、定语后置和主宾倒装。所有倒装句类型对意义影响一般，如“碧玉妆成一树高”是“碧玉妆成一高树”（贺知章《咏柳》）的定语后置倒装。《唐诗三百首》特殊句式如图1所示。

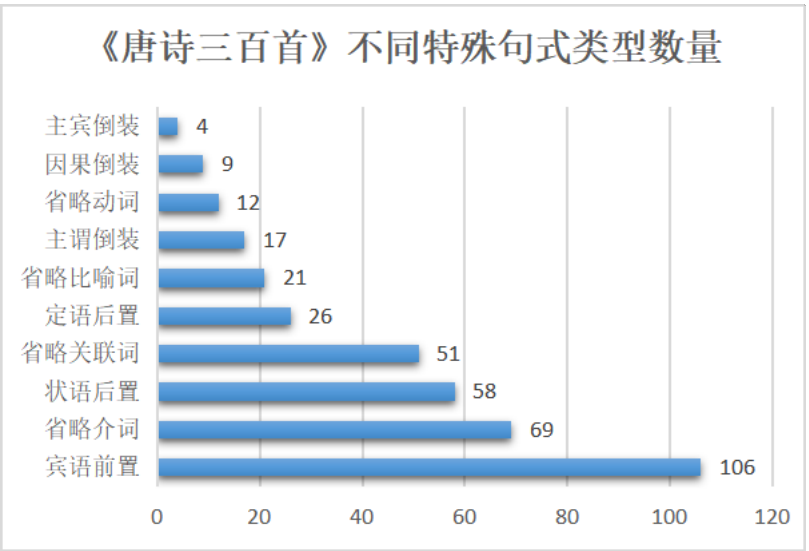


Figure 1: 《唐诗三百首》不同特殊句式类型的数量图

影响特殊句式难度的因素包括出现次数和对意义理解的影响程度。出现次数指的是在一首古诗中，特殊句式（省略句和倒装句）出现的个数，出现个数越多，难度越大。对意义理解的影响程度指的是该种句式类型（如省略关联词）对意义理解的影响程度，影响越大，难度越大。赋分规则如下：出现次数上，一首古诗中出现N次特殊句式（省略句或倒装句）则计为N分。对意义理解的影响程度上，将影响程度不大的类型计作1分（省略关联词和省略介词），将影响程序一般的类型计作2分（省略动词以及倒装句的五种类型），将影响程度很大的类型计作3分（省略比喻词和互文省略）。再将出现次数和对意义理解的影响程度，按照4：6的权重进行加权计算，根据所得分数将古诗分为0-4级。

4.2 诗句长度

为了便于统计，本文的诗歌长度指：除标题以外，一首诗歌的字数（不含标点符号）。《唐诗三百首》以体裁分章节，而体裁规定了每句诗的长度和句子数量，进而形成了每首诗的诗歌长度，诗歌长度对诗歌可读性影响很大。统计《唐诗三百首》的诗歌长度，如图2所示。可以发现，诗歌多为四种诗歌长度类型：20字（37首）、28字（60首）、40字（91首）、56字（55首）。从体裁来看，20字的诗歌为五言绝句（29首）和五绝乐府（8首），28字为七言绝句（51首）和七绝乐府（9首），40字为五言律诗（80首）、五言古诗（9首）和五古乐府（2首），56字为七言律诗（53首）、七言古诗（1首）和七律乐府（1首）。

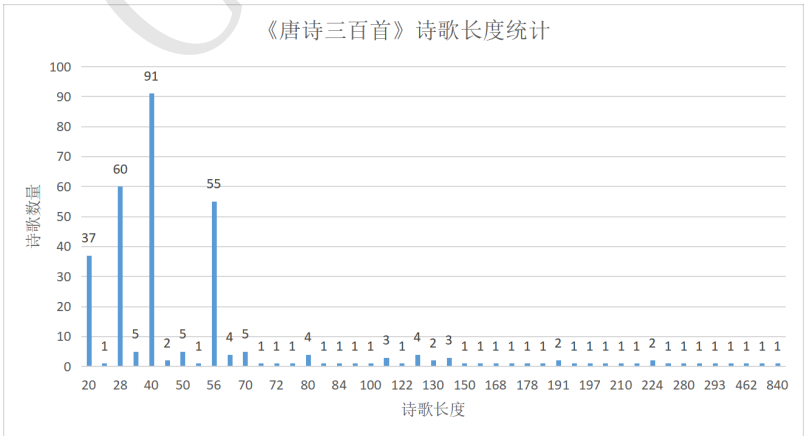


Figure 2: 《唐诗三百首》诗歌长度统计

再对人教版小学语文教材中唐朝诗人所写的古诗进行统计，如表2所示，小学低年级课文

诗歌长度多为20字，小学中年级所选的诗歌长度多为28字，小学高年级仍以20字和28字古诗为主，出现了部分40字和56字古诗。

Table 2: 人教版小学语文教材唐朝古诗诗歌长度统计

人教版小学语文教材唐朝古诗诗歌长度统计			
年级段	诗歌长度	诗歌数量	占该年级段比重
小学低年级（一二年级）	18	1	5%
	20	13	68%
	28	5	26%
小学中年级（三四年级）	20	4	24%
	28	13	76%
小学高年级（五六年级）	20	4	24%
	28	9	53%
	40	3	18%
	56	1	6%

基于对《唐诗三百首》中的诗歌长度分布统计和小学阶段唐诗诗歌长度的统计，本文将20字及以下的诗歌划分为1分，将21-28字诗歌划分为1.5分，将29-40字诗歌划分为2分，将41-56字划分为2.5分，将57字-100字诗歌划分为3分，将101-200字划分为3.5分，将200字以上划分为4分。

4.3 标题长度

标题长度对诗歌难度也有一定影响，标题长，理解难度越大。对《唐诗三百首》的标题长度（不含标点符号）进行统计，得到如图3结果。诗歌标题长度在1-50个字之间，并集中在2-6个字之间，标题长度在2-6个字诗歌占《唐诗三百首》总数量的比重为76%。

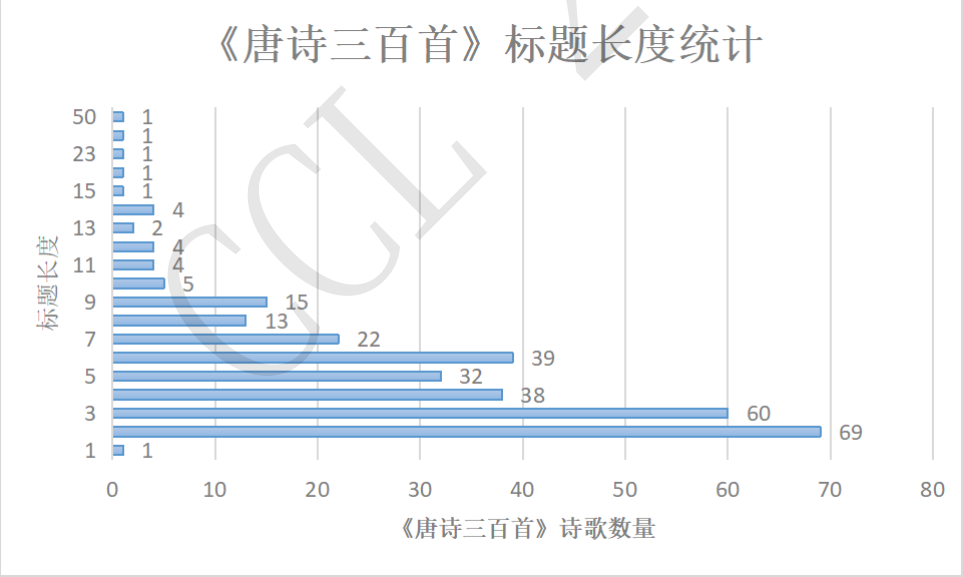


Figure 3: 《唐诗三百首》标题长度统计

再对人教版小学语文教材中唐朝古诗的标题长度进行统计。小学低年级的标题长度在1-7个字之间，标题长度为2个字的最多。小学中年级的标题长度在1-9个字之间，标题长度为2个字的最多。小学高年级的标题长度在1-10个字之间，标题长度为4个字的最多。基于《唐诗三百首》的标题长度统计和人教版小学语文唐朝古诗的诗歌长度，设计了标题长度的分级标准。标题长度为1-4个字，为1分；标题长度为5-6个字，为2分，标题长度为7-10个字，为3分；标题长度为10个字以上，为4分。

5 艺术层分级

5.1 典故

“典故”在《现代汉语词典》（第七版）的解释是“诗文等所引用的古书中的故事或词句。”也就是说，典故是古诗文中作者引用的古代故事或有出处的词句。从典故的内容角度进行分类，典故可以分为“事典”和“语典”。“事典”指的是在诗文作品中引用的事件性典故，通常包括上古神话、传说故事、历史故事、宗教故事等。“语典”指的是在诗文作品中引用的语言性典故，如前人所说的能够溯源的话语、诗词曲赋中的词汇短语、文学作品中的成语俗语等。从引用典故的方式角度进行分类，可将典故分为明用和暗用两类。明用典故是指作者在进行文学创作时，直接引用历史典故或简单概述历史故事。暗用典故是指作者在行文中对于典故的使用比较隐蔽，不像明用典故那样有直接的引用痕迹，而是通常将引用的典故消融在作品的字里行间，曲折内敛地表述情感。

在标注典故的过程中，以赵昌平《唐诗三百首全解》(2006)作为主要参考，并辅之以唐诗三百首语料库和搜索引擎，遵循尽可能多标注的原则，将典故全部标出。典故标注内容包括六类，分别为典故数量、典故类型、出处朝代、出处典籍和典故内容。标注结果显示：《唐诗三百首》收录的314首诗中共有71首诗用典，其中每首诗的用典数量如图4所示，其中一首诗中最多的典故数量为9个。

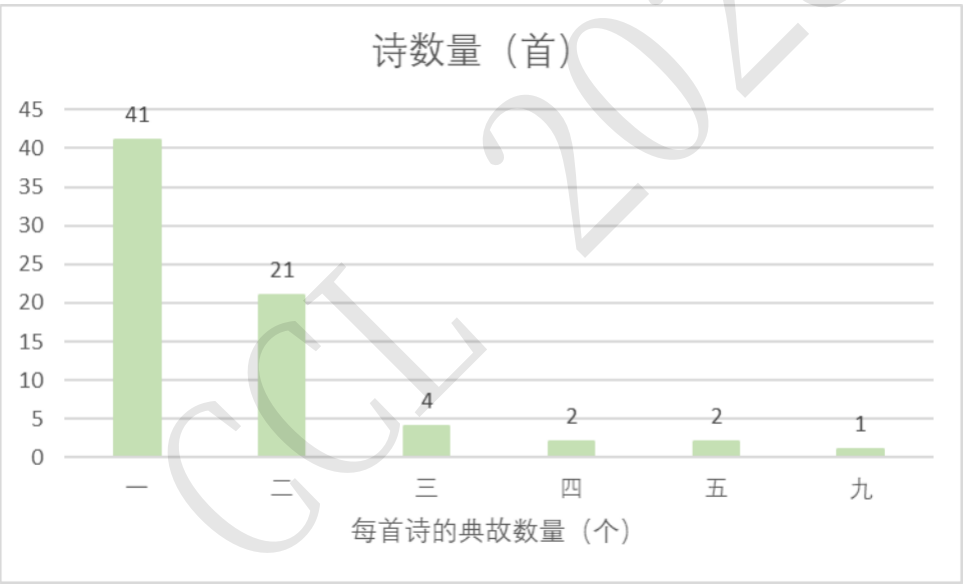


Figure 4: 《唐诗三百首》典故数量统计

影响典故难度的因素包括内容难度、出处难度和数量难度。第一，典故内容难度分类包括典故类型（事典和语典）和用典类型（明用和暗用）。根据试标结果，语典的流传度高于事典，理解难度上，事典＞语典；明用典故与暗用典故相比，更为直白，理解难度上，明用＞暗用。四个维度相较，暗用难度最高，因此对暗用赋以更高的分数。第二，出处难度包括出处朝代和出处典籍。根据BCC语料库统计出每个朝代和典籍出现的次数，出现次数越多，越容易理解，再根据次数分别划为四个等级。第三，数量难度指的是每首古诗出现的典故总数，典故越多，难度越大。三个维度的评分标准如表3所示。再将三个维度通过向量空间模型转化成特征值，设置在各个特征可读性难度均最高的理想诗句向量，计算诗句向量和理想诗句向量的欧式距离，从而得出典故的难度级别。

第二十二届中国计算语言学大会论文集，第491页-第500页，哈尔滨，中国，2023年8月3日至5日。

Table 3: 典故难度三维度评分标准

内容难度	**分数**	**语典**	**事典**
	明用	1	2
	暗用	3	4
出处难度	**典籍出现次数**	**分数**	
	1-100	1	
	100-500	2	
	500-1000	3	
	＞1000	4	
数量难度	**每首诗典故数量**	**分数**	
	1	1	
	2-3	2	
	4-5	3	
	9	4	

5.2 其他修辞

修辞手法是为提高表达效果，用于各种文章或应用文，在语言写作时表达方法的集合。古诗的修辞手法较为丰富，常见的如比喻、拟人、隐语和借代等。因为典故单独作为指标，此处其他修辞指除用典以外的修辞。与直接描写不同，修辞手法或增加喻体（如“大漠沙如雪，燕山月似钩。”——李贺《马诗》），如或将物拟人化（如“举杯邀明月，对影成三人。”——李白《月下独酌》），或用“借体”代替“本体”（如“朱门酒肉臭，路有冻死骨。”——杜甫《自京赴奉先县咏怀五百字》，不一而足。

影响修辞难度的因素主要是修辞数量，即一首诗歌中，修辞数量越多，阅读难度越大。另外，因为诗歌中使用比喻的修辞手法非常常见，数量很多，且不同喻体对诗歌可理解性影响很大，因此又根据对比喻中的喻体熟悉度，将其细分为三个难度级别（1-3级，从最不熟悉到最熟悉）。1级为最不熟悉，是在具体的情境中才可以理解的诗句，往往是暗喻或借喻（如“锦瑟无端五十弦，一弦一柱思华年”——李商隐《无题》）。3级为最熟悉，是在如今的日常语言中依然使用，或者很好理解的比喻（如“天边树若荠，江畔舟如月。”——孟浩然《秋登兰山寄张五》）。2级难度介于1级和3级之间（如“波澜誓不起，妾心井中水。天阶夜色凉如水，坐看牵牛织女星”——孟郊《列女操》）。

5.3 描写手法

描写就是作者对人物、事件和环境所作的具体描绘和刻画。描写方法：是用生动形象的语言把人物、事件、景物具体描绘出来的一种手法，给读者以身临其境的感觉。描写手法包括白描、象征、衬托、烘托、渲染等，另外诗歌中比较独特的描写手法包括比兴、主客位移等。

影响描写手法难度的因素包括描写手法数量和各个描写手法的理解难易度。描写手法数量越多，理解难度越大。各个描写手法的理解难度越高，诗歌整体难度越大。描写手法可以分为正面描写和侧面描写，侧面描写更委婉，理解难度高于正面描写。因此对属于正面描写的手法加权赋分为1分，侧面描写和特殊的描写手法（如主客位移）加权赋分为2分。

6 分级算法及验证

6.1 分级算法

本文利用AHP层次分析法计算上述8项指标的权重。AHP层次分析法（Analytic Hierarchy Process）是对于定性的决策问题进行定量化分析的一种方法。如表4，针对通假字、特殊句式、修辞、描写手法、典故、双字词、诗句长度、标题长度构建8阶判断矩阵进行AHP层次法研究，分析得到特征向量和权重值。结合特征向量可计算出最大特征根(8.299)，接着利用最大特征根值计算得到CI值(0.043)，结合判断矩阵阶数得到RI值，计算CR值（CR=CI/RI），并且进行一致性判断。本次针对8阶判断矩阵计算得到CI值为0.043，针对RI值查表为1.410，因此计算得到CR值为0.030＜0.1，意味着本次研究判断矩阵满足一致性检验，计算所得权重具有一致性。

Table 4: AHP层次分析法计算权重结果

AHP层次分析结果				
项	特征向量	权重值	最大特征值	CI值
通假字	0.35	0.0438	8.299	0.043
特殊句式	1.308	0.1635		
修辞	1.513	0.1891		
描写手法	1.513	0.1891		
典故	1.308	0.1635		
双字词	0.35	0.0438		
诗句长度	1.308	0.1635		
标题长度	0.35	0.0438		

利用AHP层次分析法设定好权重后，使用向量空间模型(VSM: Vector Space Model)，将诗句可读性的各个特征转换为标注体系，建立基于计量特征的向量空间模型。每首诗由一个维度为8的向量表示，一个计量特征代表一个维度，每一特征项都对应一个权重，对应维度的权重为该计量特征的学习优先级别。这样一个向量可用它含有的特征项及其特征项所对应的权重所表示。

并利用相似度计算的方法，设置在各个计量特征层面上，诗句可读性难度均为最高的理想诗歌（即八个维度都为最高分4分），则其各个维度均达到满分，以该理想诗歌为标准，计算其他诗歌与它的相似度，相似度越高的诗歌，理解难度越大，诗歌可读性越低。采用计算欧氏距离的方法测量各诗歌向量与理想诗歌间的距离。

$$D(x,y) = \sqrt{\sum_{i=1}^{m} (x_m - y_m)^2}$$

计算完成后，对分值进行降序排序，得分越低的诗说明和难度最高的理想诗歌越接近，即难度越大，排在越后边，得分越高的诗说明和难度最高的理想诗歌距离越远，即难度越低，排在越前边。

分级意味着同一级别内部差异不大，与其他级别则存在较大的差异。《唐诗三百首》作为蒙学教材，可以按照学习阶段来区分，因此可以分为小学阶段、初中阶段以及高中阶段三个等级。以每首诗与理想诗歌的欧式距离为依据，采用K-means聚类算法对《唐诗三百首》进行聚类，聚类数为3，结果如表5所示。从表可以看出：最终聚类得到3类群体，其中小学阶段131首，初中阶段127首，高中阶段55首，此3类群体的占比分别是41.85%，40.58%，17.57%。整体来看，3个阶段的诗歌分布较为均匀，整体说明聚类效果较好。

Table 5: 不同类别诗歌及欧式距离取值范围

类别	欧式距离取值范围	诗歌数量
1	10.66±0.42	131
2	9.41±0.38	127
3	7.79±0.80	55

6.2 难度等级验证

根据分级算法，《唐诗三百首》难度最高和最低的前十首如图5、图6所示。

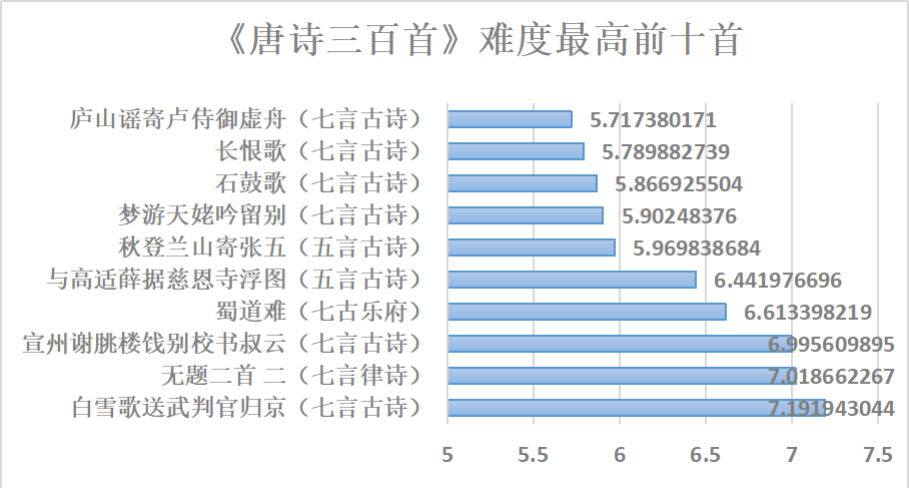


Figure 5: 《唐诗三百首》难度最高前十首

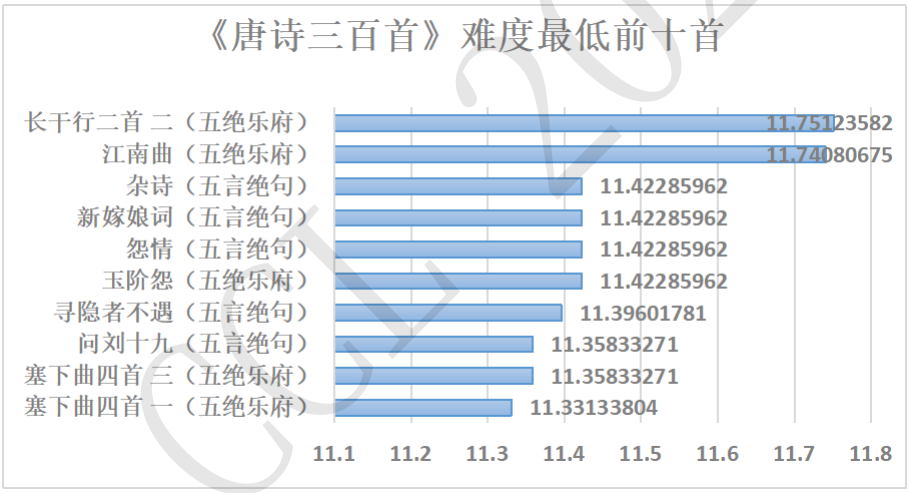


Figure 6: 《唐诗三百首》难度最低前十首

再将中小学阶段语文教材中出现的《唐诗三百首》诗篇与划分的等级进行比较，语文教材里小学阶段的诗歌，除了《山居秋暝》和《过故人庄》被划分在2类别（初中阶段），其他均在1类别（小学阶段）。教材初中阶段的诗歌，除了《宣州谢楼饯别校书叔云》、《白雪歌送武判官归京》和《兵车行》被划分为3类别（高中阶段），《送杜少府之任蜀州》被划分为1类别（小学阶段），其他均在2类别（初中阶段）。教材高中阶段的诗歌，如表6，有1篇被归为1类（小学阶段），有7篇被归为2类（初中阶段），有5篇被归为（高中阶段）。整体来看，分级结果与教材是贴合的。

第二十二届中国计算语言学大会论文集，第491页-第500页，哈尔滨，中国，2023年8月3日至5日。

Table 6: 不同类别诗歌及欧式距离取值范围

高中语文教材与《唐诗三百首》重合的诗歌篇目	诗人	体裁	分级分数	分级类别
《客至》	杜甫	七言律诗	10.34201827	1
《登高》	杜甫	七言律诗	9.96939296	2
《旅夜书怀》	杜甫	五言律诗	9.729834535	2
《燕歌行并序》	高适	七古乐府	9.708459972	2
《登岳阳楼》	杜甫	五言律诗	9.292056666	2
《将进酒》	李白	七古乐府	9.153705262	2
《琵琶行》	白居易	七言古诗	8.90408068	2
《蜀相》	杜甫	七言律诗	8.722441275	2
《无题》	李商隐	七言律诗	8.673868804	3
《锦瑟》	李商隐	七言律诗	7.944846913	3
《蜀道难》	李白	七古乐府	6.613398219	3
《梦游天姥吟留别》	李白	七言古诗	5.90248376	3
《长恨歌》	白居易	七言古诗	5.789882739	3

7 结论和展望

为了实现《唐诗三百首》的难度分级，本文在考察影响古诗阅读的因素之后，从4个层面设计了8个计量指标，并有针对性地细化标注规则，以诗歌为单位对《唐诗三百首》进行标注，形成向量空间模型，在利用AHP层次分类进行权重设定，最终利用K-means聚类形成3类难度级别。3类难度级别分别对应小学阶段、初中阶段和高中阶段，其中小学阶段有131首诗歌，初中阶段有127首诗歌，高中阶段有55首诗歌。经过与中小学语文教材进行验证，发现本论文分级算法是精密可行的。

然而，本研究仍有值得改进和发展之处，首先，各个指标权重设定时，主观判断的比重较大，因此仍需寻找各个指标重要程度的理论依据，或通过调查实践，了解各指标的重要程度。其次，虽然我们设计了8个指标，但是仍然有发展空间，可以扩充其他指标，如情感的分级。最后，分级阅读最终目的是推广阅读，在未来的工作中，我们会致力于将分级指标继续完善，助力古诗阅读的推广。

参考文献

卢烈红. 2007. 古今字与同源字、假借字、通假字、异体字的关系. 语文知识, (1):45–48.

张秋玲, 牛青森, and 赵宁宁. 2022. 中学语文教科书文言选文难易度评量模型检验. 语言文字应用, (3):49–61.

张秋玲. 2010. 文言文“浅易”的语词特征研究——以百年来初中教科书中的文言选篇为研究对象. 语言文字应用, (3):115–122.

李国英. 2007. 异体字的定义与类型. 北京师范大学学报*(*社会科学版*)*, No.201(46-50).

柳明烨. 2021. 基于部编版小学语文教材的古诗词分级标准研究. Ph.D. thesis, 南京师范大学.

王东旭. 2013. 论《唐诗三百首》与语文教材的古诗选编. Ph.D. thesis, 东北师范大学.

白瑞芬. 2017. 国家经典少儿读物分极改编的问题与思路. 编辑学刊, (3):102–106.

赵昌平. 2006. 唐诗三百首全解. 上海：复旦大学出版社.

基于RoBERTa的中文仇恨言论侦测方法研究

饶晓俊[1]，张仰森[1,2]，彭爽[3]，贾启龙[1]，刘雪阳[1]
[1]北京信息科技大学智能信息处理研究所/ 北京
[2]国家经济安全预警工程北京实验室/ 北京
[3]东北师范大学文学院/ 吉林长春
{raoxiaojun588,zhangyangsen}@163.com, shuangpeng@nenu.edu.cn,
allonlon@outlook.com, luxuryshxly@bistu.edu.cn

摘要

随着互联网的普及，社交媒体虽然提供了交流观点的平台，但因其虚拟性和匿名性也加剧了仇恨言论的传播，因此自动侦测仇恨言论对于维护社交媒体平台的文明发展至关重要。针对以上问题，构建了一个中文仇恨言论数据集CHSD，并提出了一种中文仇恨言论侦测模型RoBERTa-CHSD。该模型首先采用RoBERTa预训练语言模型对中文仇恨言论进行序列化处理，提取文本特征信息；再分别接入TextCNN模型和Bi-GRU模型，提取多层次局部语义特征和句子间全局依赖关系信息；将二者结果融合来提取文本中更深层次的仇恨言论特征，对中文仇恨言论进行分类，从而实现中文仇恨言论的侦测。实验结果表明，本模型在CHSD数据集上的F1值为89.12%，与当前最优主流模型RoBERTa-WWM相比提升了1.76%。

关键词： 中文仇恨言论；文本分类；RoBERTa；TextCNN；BiGRU

Chinese Hate Speech detection method Based on RoBERTa-WWM

Rao Xiaojun[1], Zhang Yangsen[1,2], Peng Shuang[3], Jia Qilong[1], Liu Xueyang[1]
[1]Institute of Intelligent Information, Beijing Information Science &Technology University / Beijing
[2]National Economic Security Early Warning Engineering Beijing Laboratory / Beijing
[3]College of Arts, Northeast Normal University / Changchun, Jilin
{raoxiaojun588,zhangyangsen}@163.com, shuangpeng@nenu.edu.cn,
allonlon@outlook.com, luxuryshxly@bistu.edu.cn

Abstract

With the popularity of the Internet, social media provides a platform for exchanging views, but intensifies the spread of hate speech due to its virtual and anonymous nature. Therefore, automatic detection of hate speech is crucial to maintain the civilized development of social media platforms. To solve the above problems, a Chinese Hate Speech Dataset–CHSD and a RoBERTa-CHSD model which is trained on the dataset are proposed. The RoBERTa pre-trained language model is used to serialize Chinese hate speech and extract the text feature information. Then, the TextCNN model and Bi-GRU model are respectively connected to extract multi-level local semantic features and dependency information between sentences. The two results are fused to extract deeper hate speech features in the text, and Chinese hate speech is classified, so as to realize the detection of hate speech. Experimental results show that the F1 value of the proposed model on CHSD corpus is 89.12%, which is 1.76 percentage points higher than that of the current best mainstream model RoBERTa-WWM model.

Keywords: Chinese Hate Speech , Text Classification , RoBERTa , TextCNN , BiGRU

第二十二届中国计算语言学大会论文集，第501页-第511页，哈尔滨，中国，2023年8月3日至5日。

1 引言

随着互联网的快速普及，社交媒体逐渐成为人们交流观点的最重要途径。但网络空间具有虚拟性和欺骗性，一些异常用户可以借助社交平台轻而易举地发布各种歧视言论和仇恨言论，因此社交媒体常常成为仇恨言论的爆发地。与此同时，海量数据的产生也带来了社交平台难以监管的问题，因此自动侦测仇恨言论日渐成为一个急需解决的问题。

根据现有的法律规定和通用共识，联合国 (2019)将仇恨言论定义为“因为个人或群体的身份（即他们的宗教、族裔、国籍、种族、肤色、血统、性别或其他身份因素）而攻击他们或对他们使用贬损或歧视性语言的任何言论、文字或行为交流。”。仇恨言论会对目标对象造成心理伤害，排斥、分裂不同的社会群体，严重的情况可能会引发社会暴动，从而对社会秩序造成伤害。因此，自动检测仇恨言论对于净化网络环境，维护社会和平具有重要意义。

为了解决仇恨言论的自动侦测问题，一个可靠的、通用的基准是加速该方向深入研究的必要基础。目前，常用的仇恨言论数据集有Wulczyn et al. (2017)提出的WTC，Zampieri et al. (2019)提出的OLID，Xu et al. (2020)提出的BAD等，但这些工作大多针对英文领域，中文领域由于缺乏完善的数据集和可靠的检测方法，中文仇恨言论侦测问题还有待进一步研究。

针对以上问题，提出了一个中文仇恨言论数据集CHSD（Chinese Hate Speech Dataset），包含17430条文本，主题覆盖种族、性别和地域。此外为了更深层次地提取仇恨言论特征，融合BiGRU提取全局特征和TextCNN提取多层次局部特征信息的特点，提出RoBERTa-CHSD模型来对中文仇恨言论进行侦测，实验结果表明RoBERTa-CHSD模型对于中文仇恨言论侦测的有效性。我们的数据和代码开源于https://github.com/RXJ588/CHSD。

2 相关工作

2.1 仇恨言论数据集

目前英文领域对于仇恨言论的研究非常丰富，研究范围涉及二分类到多标签分类再到多级分类任务。二分类任务方面，ElSherief et al. (2018)将从Twitter收集到的27,330条语料做是否为仇恨言论的二分类，提出英文仇恨言论数据集Peer to Peer Hate。多分类任务方面，Waseem (2016)将针对仇恨言论的类型做多分类标注，涉及类别有性别主义，种族主义和其他主义。多级分类的特点是多级注释，采用更细粒度的方案，Gomez et al. (2020)从种族、性别、性取向、宗教信仰四个方面做仇恨类型标注，在此基础上进一步标注了被攻击的对象群体。Nobata et al. (2016)区分了安全语言和辱骂性语言，又将辱骂性语言标记为仇恨言论、贬损或亵渎。Basile et al. (2014)采用三层二分类对仇恨性、攻击性和目标（个人/群体）进行标注。

目前国内相关数据集主要是面向侮辱性、性别对立、社会偏见等领域的，如表1所示。Tang et al. (2020)提出了一个分类冒犯语言的数据集COLA，主要用来对侮辱性语言，反社会语言和非法语言进行分类。Jiang et al. (2022)提出了第一个中文性别主义的数据集SWSR来识别性别相关的滥用语言。Zhou et al. (2022)提出了中文对话偏见数据集CDIAL-BIASDATASET，研究了对话中对目标群体的内隐态度。Deng et al. (2022)提出了中文冒犯语言数据集COLD，主要针对中文领域的冒犯性言论做了一个二分类任务。

Table 1: 中文领域仇恨言论相关数据集

数据集	年份	研究范围	大小
COLA	2020	侮辱性语言、反社会语言、非法语言	18k
SWSR	2022	性别相关的语言滥用	16k
CDIAL-BIASDATASET	2022	对话中的社会偏见	28k
COLD	2022	性别，种族和地域的冒犯语言	37k

总体来看，中文领域相关数据集要么话题覆盖比较单一，要么仅仅在研究文本的冒犯性或者文本的偏见表达，而仇恨言论是结合某项社会偏见的冒犯性表达，当前还缺少该类中文数据集，因此针对性别、种族和地域相关话题，提出了一个仇恨言论数据集CHSD。

基金项目：国家社会科学基金重大项目（21&ZD287）；

2.2 仇恨言论侦测

根据仇恨言论的定义，其侦测识别任务大致可以分为两类：判断文本是否为仇恨言论的侦测任务，以及对仇恨言论类型的判断的识别任务。本质上，这两个任务都属于文本分类任务。近年来文本分类技术的快速发展也为仇恨言论的侦测提供了有力的支撑，目前关于仇恨言论的研究方法主要有基于传统机器学习和基于深度学习两种方法。

传统机器学习是采用特征工程的方法获取文本特征后，再利用分类器来做仇恨言论的分类。常用的分类器有线性回归模型（LR）(Ousidhoum et al. (2019))、支持向量机（SVM）(Agarwal and Sureka (2015))、朴素贝叶斯（Naive Bayes）(Abozinadah et al. (2015))等。这种做法在特征的把握上忽略了序列关系，故不能充分利用文本上下文的信息，同时在复杂文本特征的提取上效果较差。

随着深度学习和预训练大模型的快速发展，越来越多的深层神经网络被运用到仇恨言论检测任务中。Badjatiya et al. (2017)提出使用卷积神经网络和长短期记忆网络与传统机器学习相结合的方式对Twitter仇恨言论进行侦测，结果表明深度学习模型的效果明显优于传统机器学习模型。Park and Fung (2017)提出使用混合模型进行言论滥用侦测的方法。考虑到Twitter推文的特殊性，需同时对字符级别和单词级别的特征进行学习，故结合了CharCNN与WordCNN得到混合CNN模型，用以提取不同级别的特征。实验结果显示，HybridCNN模型比单独的模型效果更加优秀。卢欣 (2019)针对中文微博数据，总结了7种语言特征，并将这种特征作为CNN网络的输入进行训练，特征输出与词向量经过CNN网络后的输出相结合，最终得到侦测结果。实验结果表明深度学习框架可以大幅度提升侦测精度，加入额外文本特征信息后的效果能得到进一步的提升。

相较于传统机器学习方法可能会出现的问题，深度学习方法则展现了更强的泛化能力，故其在仇恨言论侦测相关任务中已经成为研究主流。尽管通常深度学习的效果较好，但它仍然会受到数据的限制与词向量的影响，尤其在中文领域仇恨言论资源还很匮乏，具有进一步探索研究的意义。

3 RoBERTa-CHSD模型

针对中文仇恨言论侦测的问题，提出了RoBERTa-CHSD仇恨言论侦测模型。首先利用预训练模型RoBERTa-WWM学习文本语义特征，考虑到RoBERTa-WWM只是简单地做词嵌入的工作，难以充分考虑文本的内部语义特征信息，因此将得到的词嵌入特征再分别输入到TextCNN和BiGRU中，融合TextCNN得到的多层次局部语义特征和BiGRU得到的句子间的依赖关系信息，有效提取文本中更深层次的仇恨言论特征，从而提升仇恨言论侦测模型的性能。RoBERTa-CHSD模型结构如图1所示。

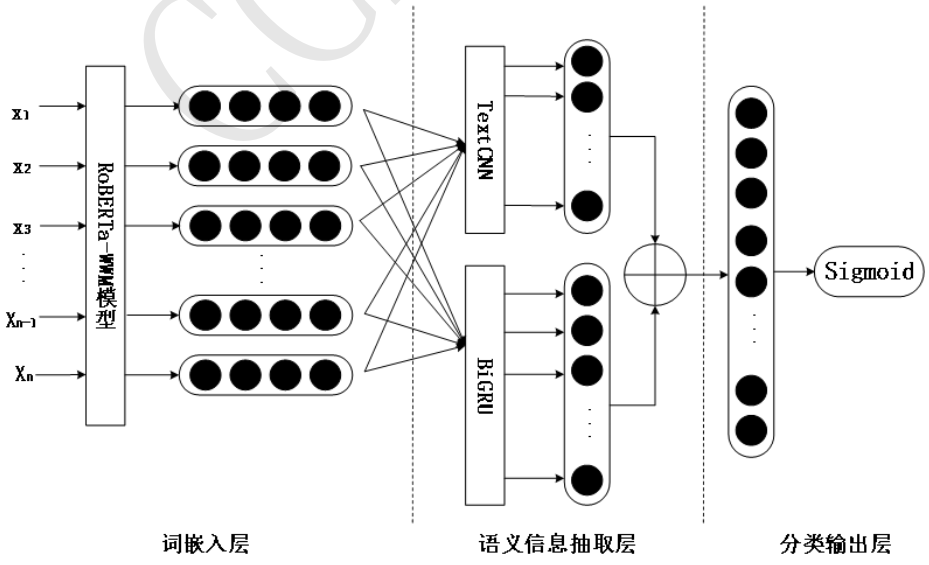


Figure 1: RoBERTa-CHSD模型结构

3.1 词嵌入层

词嵌入层是将输入的文本信息通过编码转换为相应的向量。采用RoBERTa-WWM预训练语言模型Liu et al. (2019)来对输入文本进行词嵌入表示。对于一条文本$S = [x_1, x_2, x_3, ..., x_n]$，构建文本S的词向量、句子嵌入和位置向量，将这3个向量的加和$E = [e_1, e_2, ..., e_n]$作为RoBERTa-WWM模型的输入，输入处理流程如图2所示。

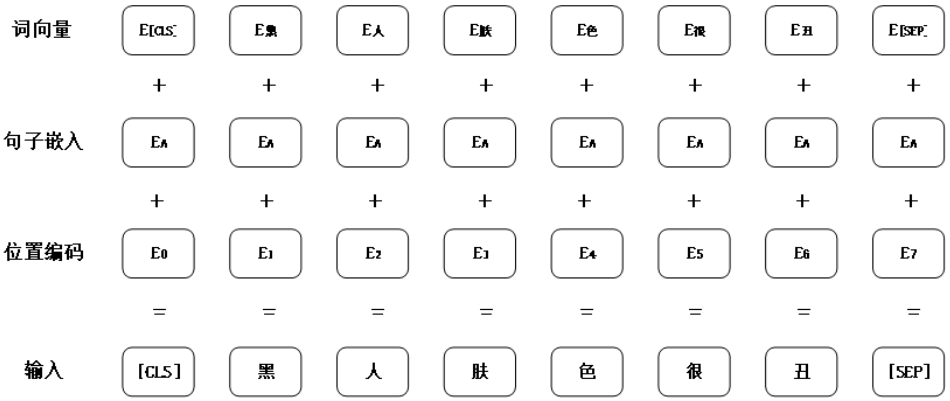


Figure 2: 输入处理流程

其中，词向量为one-hot编码后单词对应的向量表示；句子嵌入指用于分割多个句子的向量；位置向量为序列增加位置信息，保持顺序性表示。

再经过RoBERTa-WWM中的Transformer模块训练即可得到文本S的动态语义表示$V = [v_1, v_2, v_3, ..., v_n] \in R^{n \times d}$,其中$n$为输入长度，$d$为词向量维度768。RoBERTa-WWM模型结构如图3所示，中间层表示12层双向Transformer特征提取器。

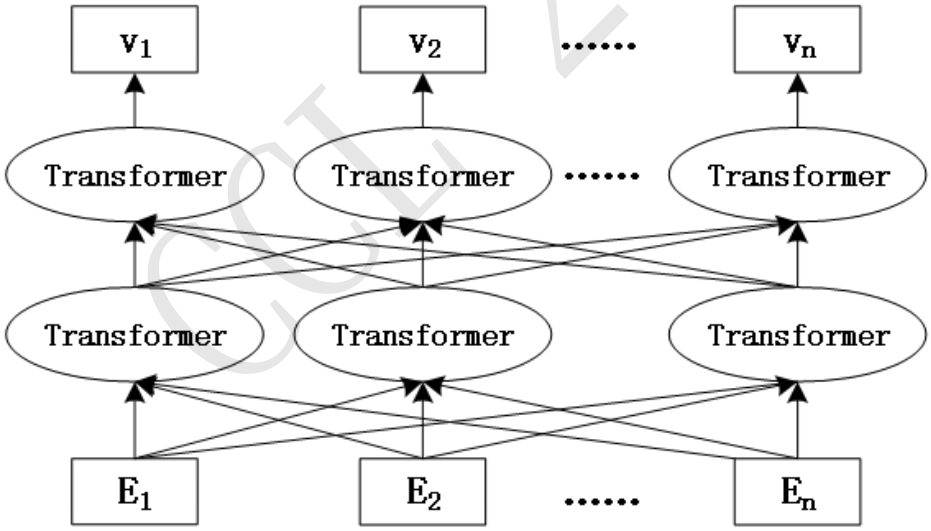


Figure 3: RoBERTa-WWM模型结构

3.2 语义信息获取层

对比词嵌入层得到的句子级表示，语义信息获取层用于获取所输入文本的更深层次的语义信息。将上一层得到句子级表示分别送入TextCNN(Kim (2014))和BiGRU(Cho et al. (2014))，中来学习文本中不同层次的局部信息和正反双向句子间的依赖关系，并将二者结果进行融合来对仇恨言论文本进行深层次的特征提取。

3.2.1 TextCNN层

TextCNN利用多个不同大小的卷积核来提取文本中的关键信息，自动对$n-gram$特征进行组合和筛选，从而能够充分捕捉文本中的局部相关信息，获得不同抽象层次的语义信息。TextCNN模型结构如图4所示。

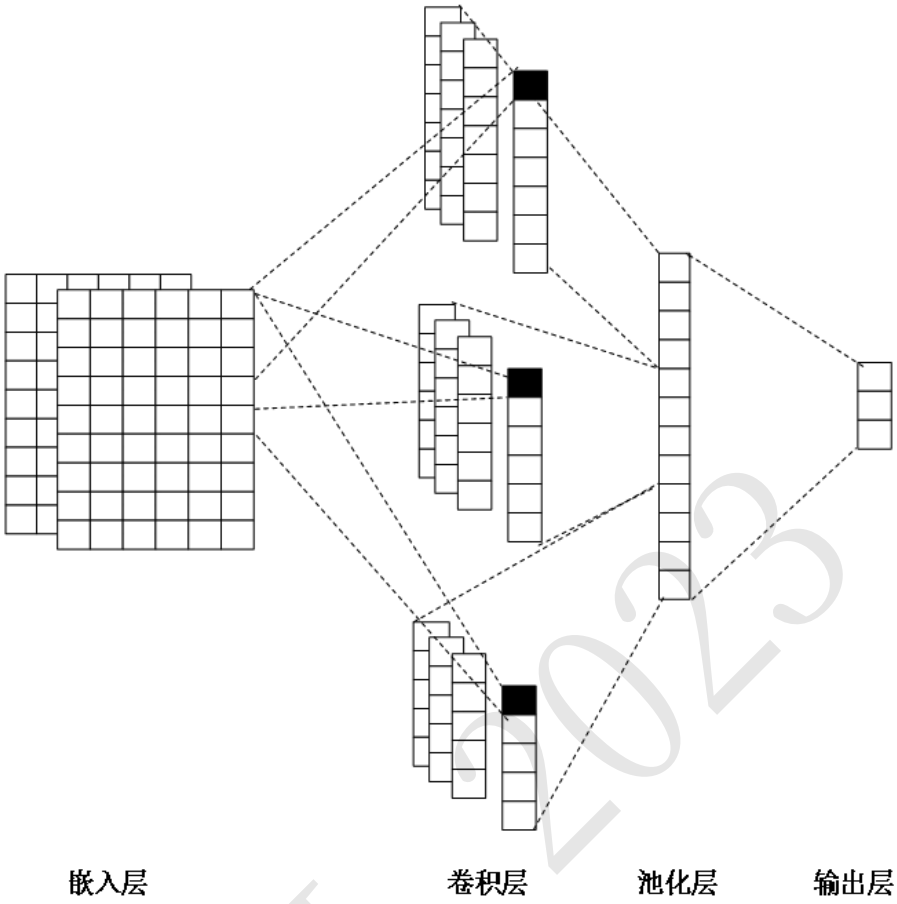


Figure 4: TextCNN模型结构图

令卷积核为$w \in R^{h \times d}$，d为卷积核的宽度，与RoBERTa-WWM层的输出维度一致，h为卷积核高度。设置三种不同的卷积核，h分别设置为3、4、5，输入通道为1，输出通道为256，保证卷积输出向量的维度一致。将卷积核w与词向量矩阵V中的第i个窗口$V_{i:i+h-1}$内的词向量进行卷积操作，得到特征c_i，计算如公式1所示。

$$c_i = f(w \cdot V_{i:i+h-1} + b) \tag{1}$$

其中，f为激活函数，b为偏置。

卷积核w与词向量矩阵V中所有窗口内词向量进行卷积操作后，得到特征图$c \in R^{n-h+1}$，如公式2所示。

$$c = [c_1, c_2, ..., c_{n-h+1}] \tag{2}$$

再进入池化层进行最大池化运算，将卷积层输出的特征进行池化操作来提取更显著的特征，卷积核w对应生成的特征图c经过池化操作得到$c^{'} = \max\{c\}$，并将池化后的特征向量进行拼接操作。通过联结所有卷积核的池化结果，得到新特征cnn_outs，计算如公式3所示。

$$cnn\_outs = [c_1^{'}, c_2^{'}, ..., c_k^{'}] \tag{3}$$

其中，k为TextCNN输出维度256。

第二十二届中国计算语言学大会论文集，第501页-第511页，哈尔滨，中国，2023年8月3日至5日。

3.2.2 BiGRU层

BiGRU层负责学习句子间双向依赖信息，其模型结构如图5所示，它是由双向的门控循环单元（Gated Recurrent Unit，GRU）组成的。可以对仇恨言论的上下文进行双向特征提取，捕获句子间双向依赖信息，以便更准确地获取上下文全局特征信息。

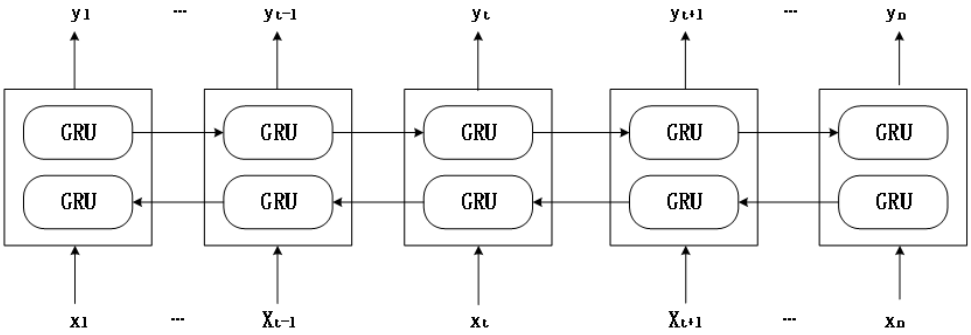


Figure 5: BiGRU模型结构

在t时刻BiGRU的双向隐状态输出分别如公式4和5所示,在t时刻的隐藏状态是双向结果地拼接，如公式6所示。

$$\overrightarrow{h_t} = \overrightarrow{GRU}(w_t, \overrightarrow{h_{t-1}}) \tag{4}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(w_t, \overleftarrow{h_{t-1}}) \tag{5}$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{6}$$

其中，w_t为t时刻单向GRU的权重矩阵。

整段文本的隐藏状态集合gru_outs为所有时刻h_i的拼接，如公式(7)所示。

$$gru\_outs = [h_1, h_2, \ldots\ldots, h_n] \tag{7}$$

最后将两个模型得到的结果进行拼接，得到该段文本的融合语义特征信息cat_outs。计算如公式8所示。

$$cat\_outs = cnn\_outs \oplus gru\_outs \tag{8}$$

3.3 分类输出层

将融合特征表示接入全连接层，将其映射到实例标签空间，对仇恨言论进行分类。语义信息获取层的输出cat_outs与全连接层权重矩阵计算后输出M，计算如公式9所示。

$$M = \tanh(W_d \cdot cat\_outs + b_d) \tag{9}$$

其中，W_d是全连接层权重矩阵，b_d是全连接层偏置。

输出层采用$Sigmoid$函数对全连接层的输出信息M进行归一化处理，得到每个倾向类别的概率值。计算如公式10所示。

$$y = Sigmoid(W_s \cdot M + b) \tag{10}$$

其中，W_s为输出层权重矩阵，b为输出层偏置。y表示模型对每个倾向类别的概率值。

4 实验与结果分析

4.1 数据集

本节主要是对中文冒犯言论数据集COLD进行二次处理，根据仇恨言论的定义对于数据进行重新标注，分别标注仇恨言论和非仇恨言论两种类别。为了更高效地标注数据，对于训练集，采用Model-in-the-loop(Sun et al. (2021))策略来对训练数据进行标注；对于测试集，为了保证结果可靠，测试数据集依旧采用手工标注的方式。对COLD数据集进行处理后，发现仇恨言论数据只有4629条，仅占整个COLD数据集的18.23%，其中性别主题有931条，地域主题有1289条，种族主题有2409条，话题分布明显不均衡。因此为了扩充数据集规模和实现仇恨言论数据集的话题分布类型均衡，考虑从性别主义数据集SWSR（Jiang et al. (2022)）中抽取1600条性别主题相关仇恨言论，从对话偏见数据集CDIAL-BIASDATASET（Zhou et al. (2022)）中的地域主题对话中抽取1250条地域主题相关仇恨言论来对数据集做扩充，最终提出提出一个包含17430条句子，覆盖种族，性别，地域主题的仇恨言论数据集CHSD。

（1）数据收集：对于训练集的构建，基于Model-in-the-loop的设置对数据进行半自动标记。首先，通过手动标记500个样本(仇恨或非仇恨)作为训练数据。其次，对一堆未标记的数据采用分类器，并预测它们的仇恨性。然后，根据预测分数对数据进行排序，并将数据分成多个集合进行样本检查。从每个集合中抽取大约10%的数据，并使用以下策略手动标记它们:1)如果预测标签的准确率达到90%，则直接将集合中的数据加入到训练集中;否则，2)将集合完全手动重新标记，然后添加到训练集中。通过这种方法，迭代更新分类器和训练集6轮。对于测试集的构建，完全采用人工标注的策略。

（2）标注策略：邀请了3位语言安全研究者来对数据集进行手工标注。他们的性别分布为两男一女，分别来自河南，东北，北京三个地区。根据仇恨言论的定义对于数据进行标注，最终结果选取票数更高的标签作为仇恨言论标签。

（3）数据分析：基于以上流程，构建了仇恨言论数据集CHSD，数据集的详细信息如下表2、表3所示。

Table 2: CHSD数据集分布

	仇恨	非仇恨	整体
训练集	5644	7944	13588
测试集	1840	2002	3842
整体	7484	9946	17430
平均长度	53.69	44.20	48.77

Table 3: CHSD数据集样例

标签	文本
仇恨	男生就活该被女性打
非仇恨	朋友，醒醒，重男轻女时代过去了
仇恨	河南人都偷井盖
非仇恨	人家没说山东人，并不是地域黑。扣帽子可不好
仇恨	黑人肤色和五官都不好看
非仇恨	印度人数学都很好

为了保证所构建仇恨言论数据集的可靠性和可用性，采用了Kappa值来对测试集中三位标注者的结果和最终数据集结果（可以视为一个新的标注者）进行一致性检验，计算如公式11所示。其具体结果如表4所示。可以看出，3个Kappa值都高于0.6，这说明数据具有较高的一致性，构建的仇恨言论数据集质量合格。

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{11}$$

其中，p_0是每一类正确分类的样本数量之和除以总样本数。

假设每位标注者标注的仇恨和非仇恨类别的样本个数分别为a_1, a_2,模型预测到的每类样本个数分别为b_1, b_2，总体样本数为n，p_e计算如公式12所示。

$$p_e = \frac{a_1 \cdot b_1 + a_2 \cdot b_2}{n \cdot n} \tag{12}$$

Table 4: CHSD数据集的一致性检验结果

数据集	A-Result	B-Result	C-Result
CHSD	0.813	0.792	0.832

4.2 实验参数设置

为了在仇恨言论数据集上取得最优的分类结果，通过设置不同的超参数，来做多次对比试验，最后选取参数的最优值。实验的超参数设置如表5所示。

Table 5: 实验参数设置

参数	说明	最优值
batch_size	一次训练选取的样本数	32
learning rate	学习率	2e-5
epochs	训练次数	30
GRU_units	GRU输出结果的维度	128
Dropout	随机舍弃的神经元比例，防止过拟合	0.5
max_seq_len	单个语句最大长度	64

其中，batch_size与实验所使用的计算平台算力相关，综合考虑实际情况将batch_size的值设置为32，BiGRU层的隐藏单元数设置为128。

4.3 对比实验与结果分析

为了验证RoBERTa-CHSD模型在仇恨言论侦测任务中的整体性能，将其与以下几种分类基线模型进行对比分析，测试了它们的精确率、召回率和F1 值，实验结果如表6所示。

1)BERT(bert-base-chinese)：使用基于BERT的中文预训练模型做仇恨言论侦测任务。

2)ALBERT(albert-chinese-tiny)：使用基于ALBERT的中文预训练模型做仇恨言论的文本分类任务。ALBERT模型在BERT模型的基础上进行改进，使用了自监督损失函数关注构建句子中的内在连贯性。它设计了参数减少的方法，用来降低内存消耗，同时加快了BERT的训练速度。

3)RoBERTa-WWM：使用RoBERTa-WWM预训练模型做仇恨言论侦测任务。

Table 6: 各模型实验结果

模型	精确率	召回率	F1值
BERT(bert-base-chinese)	85.07%	85.23%	85.15%
ALBERT(albert-chinese-tiny)	77.47%	77.47%	77.47%
RoBERTa-WWM	87.29%	87.43%	87.36%
RoBERTa-CHSD	**89.12%**	**89.13%**	**89.12%**

从表中可以看出，RoBERTa-CHSD模型与BERT,ALBERT,RoBERTa-WWM三种基线模型中性能最好的RoBERTa-WWM对比，在精确率，召回率和F1值方面分别高出1.83%，1.7%，1.76%。值得注意的是，ALBERT的参数共享策略会导致一些特征被过度

压缩，从而导致模型性能有一定的下降；BERT模型和RoBERTa-WWM模型没有考虑向前向后的信息，因此对上下文信息的理解不够充分。综上所述，相较于典型的基线模型，提出的RoBERTa-CHSD融合模型在仇恨言论侦测任务上具有更好的性能。

以F1值为评价指标,将RoBERTa-CHSD融合模型与RoBERTa-WWM模型在单个类别的分类性能上进行对比,实验结果如图6所示。

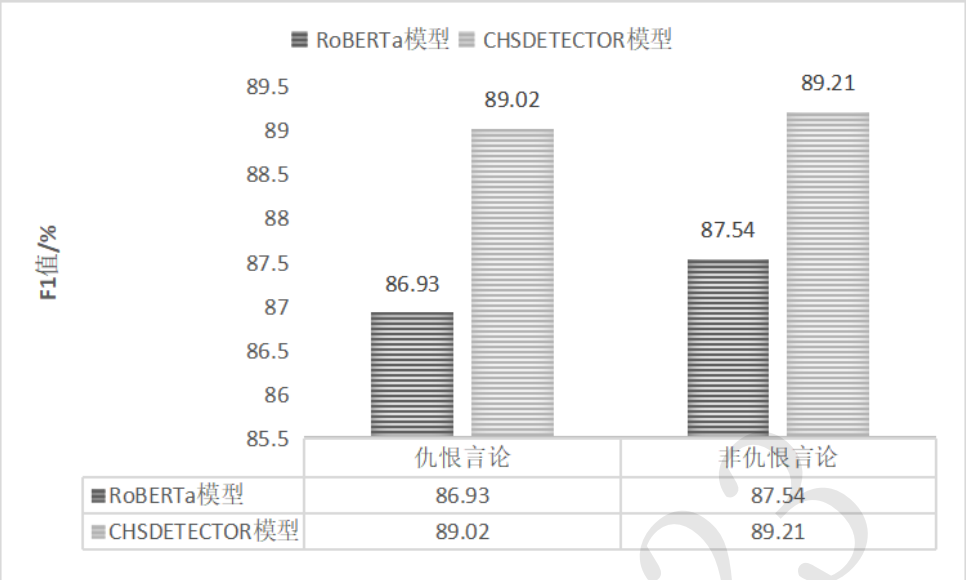


Figure 6: 不同模型的单个类别分类性能对比

相比RoBERTa-WWM模型，RoBERTa-CHSD融合模型在仇恨言论类型和非仇恨言论类型上的F1值分别提高了2.09%和1.67%，这说明融合TextCNN和BiGRU模型能够更充分地抽取文本语义特征，在各类别分类上都具有更优的性能。而且，二者都在非仇恨言论类型识别上更有优势，这是由于非仇恨言论的样本数相对而言比较多，因此模型对其语义特征学习得更充分。

4.4 消融实验设计与结果

为了进一步验证模型结构的有效性，进行消融实验，研究影响实验结果的独立因素。实验结果如表7所示。

Table 7: 不同模型实验结果对比

模型	精确率	召回率	F1值
RoBERTa-WWM	87.29%	87.43%	87.36%
RoBERTa-WWM+BiGRU	88.37%	88.06%	88.21%
RoBERTa-WWM+TextCNN	88.61%	88.50%	88.55%
RoBERTa-CHSD	**89.12%**	**89.13%**	**89.12%**

对比表中四个模型在CHSD数据集上的结果，可以发现普通基线模型RoBERTa-WWM效果最差，原因在于其只是简单地做了词嵌入的工作，未充分考虑文本的内部语义特征信息。模型RoBERTa-WWM+BiGRU利用BiGRU模型进一步提取句子间全局双向依赖信息，其F1值达到了88.21%,比单一的RoBERTa-WWM模型提升了0.85%。RoBERTa-WWM+TextCNN利用TextCNN提取多层次局部特征信息，其F1值达到了88.55%,比单一的RoBERTa-WWM模型提升了1.19%。综合来看，BiGRU和TextCNN都对文本的深层次仇恨特征提取有积极意义，因此RoBERTa-CHSD考虑融合BiGRU模型和TextCNN模型，同时提取全局信息和局部特征信息，实验结果表明，RoBERTa-CHSD模型的F1值达到了89.12%，相较于以上三种实验模型，分别提升了1.76%，0.91%，0.57%。该组对比实验体现了RoBERTa-CHSD的有效性。

5 结语

针对中文领域仇恨言论急需自动侦测的问题，提出了一个中文仇恨言论数据集CHSD，并在此基础上提出一种RoBERTa-CHSD中文仇恨言论侦测模型。该模型首先使用RoBERTa-WWM预训练语言模型捕获仇恨言论文本的语义特征，再分别使用TextCNN和BiGRU来学习文本中不同层次的局部信息和正反双向句子间的依赖关系，将二者结果进行融合来对仇恨言论文本进行深层次的特征提取，从而进一步提升了中文仇恨言论侦测模型的性能。通过实验结果可以看出，与现有的几种典型的文本分类模型相比，提出的RoBERTa-CHSD模型在仇恨言论侦测任务的整体性能上得到了有效提升。

参考文献

Ehab A Abozinadah, Alex V Mbaziira, and J Jones. 2015. Detection of abusive accounts with arabic tweets. *Int. J. Knowl. Eng.-IACSIT*, 1(2):113–119.

Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *Distributed Computing and Internet Technology: 11th International Conference, ICDCIT 2015, Bhubaneswar, India, February 5-8, 2015. Proceedings 11*, pages 431–442. Springer.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jiawen Deng, Jingyan Zhou, Hao Sun, Fei Mi, and Minlie Huang. 2022. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

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

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2021. On the safety of conversational models: Taxonomy, dataset, and benchmark. *arXiv preprint arXiv:2110.08466*.

Xiangru Tang, Xianjun Shen, Yujie Wang, and Yujuan Yang. 2020. Categorizing offensive language in social networks: A chinese corpus, systems and an explanation tool. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*, pages 300–315. Springer.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks. *arXiv preprint arXiv:2202.08011*.

王素格卢欣. 2019. 融合语言特征的卷积神经网络的反讽识别方法. 中文信息学报, 33(5):31–38.

联合国. 2019. 《联合国关于仇恨言论的战略和行动计划》. https://www.un.org/zh/hate-speech/understanding-hate-speech/what-is-hate-speech.

汉语被动结构解析及其在CAMR中的应用研究

胡康[1] ， 曲维光[1,2,4*] ， 魏庭新[3] ， 周俊生[1] ， 李斌[2] ， 顾彦慧[1]
(1.南京师范大学 计算机与电子信息学院/人工智能学院，江苏 南京 210023;
2.南京师范大学 文学院，江苏 南京 210097;
3.南京师范大学 国际文化教育学院，江苏 南京 210097;
4.南京师范大学 中北学院，江苏 丹阳 212334;
*通讯作者，Email：wgqu_nj@163.com)

摘要

汉语被动句是一种重要的语言现象。本文采用BIO结合索引的标注方法，对被动句中的被动结构进行了细粒度标注，提出了一种基于BERT-wwm-ext预训练模型和双仿射注意力机制的CRF序列标注模型，实现对汉语被动句中内部结构的自动解析，F1值达到97.31%。本文提出的模型具有良好的泛化性，实验证明，利用本文模型的被动结构解析结果对CAMR图后处理，能有效提高CAMR被动句解析任务的性能。

关键词： 被动结构解析 ； 双仿射注意力 ； CRF ； CAMR ； 后处理

Parsing of Passive Structure in Chinese and Its Application in CAMR

HU Kang[1], QU Weiguang[1,2,4*], WEI Tingxin[3], ZHOU Junsheng[1], LI Bin[2], GU Yanhui[1]
(1.School of Computer and Electronic Information/School of Artificial Intelligence,
Nanjing Normal University, Nanjing, Jiangsu 210023, China;
2.School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
3.International College for Chinese Studies, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
4.ZhongbeiColleg, Nanjing Normal University, Danyang, Jiangsu 212334, China;
*Corresponding, Email: wgqu_nj@163.com)

Abstract

Chinese passive sentences is an important linguistic phenomenon. In this paper, we use the BIO combined with indexing annotation method to annotate the passive structures in passive sentences at a fine-grained level. We propose a CRF sequence labeling model for passive structure parsing in Chinese based on the BERT-wwm-ext pre-training model and the biaffine attention mechanism, and the model achieves a significant F1 value of 97.31%. The proposed model exhibits excellent generalization capabilities. The experimental results have demonstrated that incorporating the parsing results of passive structures obtained from our model for post-processing the CAMR graph, can effectively improve the performance of passive sentence parsing in CAMR.

Keywords: passive structure parsing , biaffine attention , CRF , CAMR , post-processing

1 引言

被动句是一种常见的语法现象，它强调动作的承受者，将动作执行者放在句子中的其他位置或省略不表达，而一般的主动句则强调动作执行者。被动句的使用可以改变句子的重心，

基金项目：国家社会科学基金重大项目(21&ZD288); 国家自然科学基金(62277031)。

使得动作的承受者成为句子的核心，更加突出其在事件中的地位。被动句的使用十分广泛，表 1展示了被动句在不同语体中的出现频次(宋文辉等, 2007)。人们对被动句的广泛运用丰富了语言的多样性，但被动句构式的复杂多样性也给句法分析、语义解析等任务带来较大困难。

例1 这个问题已经被他解决了。
例2 检查情况已在当地曝光。
例3 这些建议虽然未被采纳，但其价值不可低估。

语体	会话	小说	新闻	学术	均值
有标记被动句	3.90	6.69	9.30	4.50	6.10
无标记被动句	3.55	9.19	3.65	2.56	4.70
总计	7.45	15.88	12.95	7.06	10.8

Table 1: 汉语被动句每万字出现次数

被动句自动解析任务的难点主要在于被动结构的复杂多样。被动结构即被动句中表达被动语态的内部结构，符号表示为A1+[M]+[A0]+V，其中A1表示广义受事，A0表示广义施事，M表示有标记被动结构中的标记词，V表示被动行为的谓语动词，[]表示可省略的成分。一个句子可能包含一个或多个被动结构，如例1中存在一个有标记被动结构“问题$_{A1}$+被$_{M}$+他$_{A0}$+解决$_{V}$”，例2存在一个无标记被动结构“检查情况$_{A1}$+曝光$_{V}$”，而例3存在两种被动结构，分别是“建议$_{A1}$+被$_{M}$+采纳$_{V}$”和”价值$_{A1}$+低估$_{V}$”。因此，能准确解析出句子中的被动结构可以为自然语言处理技术的发展提供技术支持，尤其是机器翻译、自动问答和自动摘要等领域。

本文把被动结构成分识别视为一种特殊的语义角色标注（Semantic Role Labeling, SRL）任务。首先对被动句进行细粒度标注，构建了一个用于模型训练的数据集。然后提出了一个针对汉语被动结构的自动解析模型PS-CRF(Passive Sentence CRF)，F1值达到97.31%。该模型利用融合整词掩码技术的预训练模型BERT-wwm-ext获取被动结构上下文语义信息，使用依存分析方法结合双仿射注意力(Biaffine Attention)评分机制进行特征学习，并通过TreeCRF模型预测输出。最后，为验证本文模型的泛化性和实用性，利用该模型对小学语文语料中被动句的解析结果对CAMR图进行后处理操作，实验证明在Smatch和Align-Smatch两个性能指标下，性能均有提升，尤其在Align-Smatch评价指标下的提升更为显著。

2 相关工作

被动句研究一直是语言学领域的重要课题，学者们除了对句子层面的句法、语义研究，还对被动句内具体成分尤其是谓语动词和标记词进行了研究。赵元任 (1979)提出能用于”被”字句的主要是处置动词，且动词必须前带或后带成分。兰宾汉 (2002)提出被字句的谓语中心语必须是表动作行为的及物动词、不及物动词、能愿动词、趋向动词、判断动词，而表示肯定或否定的”有、没有”等不能用来构成被字句。王振来 (2004)分析了自主动词比非自主动词更容易进入被动表述的原因，阐述被动表述式对动词选择具有制约作用。蚁坤 (2000)通过考察1000个被动句，验证了最常用于被动句中的动词是及物动词，不仅具有及物性特征，还具有高级及物性，其中高级及物性(王惠, 1997)指的是谓语动词具有[动作]、[完成]、[瞬时]、[自主]和[肯定]等特征。王一平 (1994)提出当”遭””挨””受”等遭受类动词后面紧跟一个及物动词，这时遭受类动词可以用”被”来替换，即在某些情况下，遭受类动词可以视为一种特殊的被动标记词。

在自然语言处理领域，有研究针对被动句的句式特点，提出了一种融合词性信息和动词论元框架信息的被动句自动识别模型(Hu et al., 2022)，可以实现从大规模语料中快速筛选被动句。但该模型只能从句子层面判断一个汉语句子是否是被动句，而无法对句中具体的成分进行解析。本文旨在实现对被动句内部结构成分的识别，并且把这个任务视为一种特殊的SRL任务，SRL是自然语言处理中的一项重要任务，其目的是识别一个句子中的每个单词或短语在句子中所扮演的语义角色，例如主语、宾语、谓语、时间状语等，SRL相关的前沿技术和方法对本文模型设计有重要参考意义。Li等人 (2021)分析了句法信息对基于序列、树和图的三种SRL基线模型的影响，提出句法信息能在一定程度上有助于模型学习，但这种帮助随着预训练模型的引入而受到限制，且句法信息的作用大小取决于模型集成句法信息的方式。Zhang等人 (2021)分析了词嵌入方法和不同的标注方法对SRL模型性能的影响，验证了预训练模型带来的性能提升优于静态词嵌入模型。Li等人 (2019)提出了一种可以同时解决span-based和dependency-based的端到端的SRL模型，该模型引入了双仿射注意力机制，能够对SRL两种表示方式进行统一有效的处理，有助于探索二者之间的联系。Zhang等

人 (2022)提出一种用依存句法分析解决SRL任务的方法，该方法的主要思想是：先通过一定的规则将SRL 结构转化为依存句法树，然后基于给定的依存句法树学习一个解析器，最后通过CRF模型将预测出的依存句法树恢复为SRL结构。

3 数据集构建

本文数据集的基础语料来源于被动句语料库(Hu et al., 2022)，该语料库包含4495条有标记被动句、4570条无标记被动句以及4465条非被动句。本文将在此基础上，针对该语料库中的被动句展开细粒度的标注，即对被动句中的被动结构进行成分标注。

3.1 数据集标注方法

被动结构可形式化表示为A1+[M]+[A0]+V，实现被动结构成分的识别，即定位句中包含的所有被动结构，并提取出每个被动结构的具体成分，可以用一个四元组(V, A1, A0, M)表示。该任务与语义角色标注任务类似，都是识别出句中谓语动词及其相关的论元角色，因此本文借鉴SRL任务数据集的标注方法对被动结构进行标注。首先对语料进行分词处理，然后使用词语级别的BIO序列标注方法结合索引的方式对语料中的汉语被动结构进行标注。图 1给出了一个被动结构标注示例，其中“B”表示被动结构中某成分的起始边界词语，“I”表示该成分的后续词语，“-”用于连接成分所属类别，“O”表示非被动结构的其他成分，“：”前的数字表示该成分指向的所属被动结构中动词在句中的索引，若当前成分是动词V，为便于在模型训练过程中对数据进行处理，冒号前的数字用不含实际意义的数字“0”代替。

索引	1	2	3	4	5	**6**	7	8
词语	这个	问题	已经	被	他	解决	了	。
标注	**6**:B-A1	**6**:I-A1	O	**6**:B-M	**6**:B-A0	0:V	O	O

Figure 1: 被动结构标注示例

3.2 数据集统计分析

本文从被动句语料库中随机抽取了3839个被动句进行细粒度标注，共标注被动结构4814个，平均每个句子含1.25个被动结构。数据集中的被动结构主要有以下四种类型：

一、普通有标记被动结构：A1+M+[A0]+V。其中A1指的是广义受事，包括受事、与事、感事、主事、材料、工具等论元角色，标记词M可由”被、由、为、给、让”等介词充当，也可是”受到、遭到”等遭受类动词。如例4中的标记词”被”是介词，而例5中”遭到”则是遭受类动词作为标记词。此外，二者省略了广义施事A0。

例4 裁判员$_{A1}$ 被$_{M}$ 袭击$_{V}$。

例5 裁判员$_{A1}$ 遭到$_{M}$ 袭击$_{V}$。

二、特殊有标记被动结构：M+[A0]+V+的+A1。当一个有标记被动结构作为一个定中结构时，由于不是句子主要成分，与普通有标记被动结构存在一定的差异，我们将之标注为特殊有标记被动结构。如例6中”全乡被洪水吞没的土地”是一个定中结构，其中隐含了一个被动表述”土地+被+洪水+吞没”。

例6 仅一冬一春，全乡被$_{M}$ 洪水$_{A0}$ 吞没$_{V}$ 的土地$_{A1}$ 全部修复。

三、普通无标记被动结构：A1+V。无标记被动结构与有标记被动结构的区别主要在于其不含标记词M和施事A0，前者直接强调了动作的完成，而后者因为标记词的存在更加强调动作的被动性，但二者均表示被动语态，如例7和例8所示。

例7 只有精通国家战略，君主的愿望$_{A1}$ 才可能实现$_{V}$。

例8 一座6000多平方米的晒谷场$_{A1}$ 也已建成$_{V}$。

四、特殊无标记被动结构：V+的+A1。与特殊有标记被动结构类似，也有少数无标记被动结构会出现在定中结构中，如例9所示。

例9 本次集中行动中列为$_{V}$ 全国重点的6大案件$_{A1}$ ，现已审结4件。

表 2列出了数据集中各种被动结构的数目以及占比，其中有标记的被动结构共2711个，无标记的被动结构共2103个。在有标记被动结构中，普通有标记被动结构有2655个，其中介词作

标记词的有2465个，遭受类动词作标记词的有190个。定中结构中的被动结构总体来说占比较少，有标记和无标记的被动结构分别占比1.16%和0.27%。

被动结构类别	细分类别	数量（个）	占比（%）	小计（个）
有标记被动结构	介词	2465	51.20	2711
	遭受类动词	190	3.95	
	定中结构	56	1.16	
无标记被动结构	普通无标记	2090	43.42	2103
	定中结构	13	0.27	
		总计	100	4814

Table 2: 被动结构数据集

4 模型设计

本文将被动结构成分识别任务建模为一个语义角色标注任务，提出一个使用高效的中文预训练模型并结合双仿射注意力机制的CRF序列标注模型，实现对汉语被动句中内部结构的自动解析，将该模型记为PS-CRF。对一个输入样本句子$S = w_1, w_2, \ldots, w_n$，其中$w_i$表示第$i$个词语，$i \in 1, 2, \ldots, n$，n为句子词语总数。模型输出为对应词语个数的标签序列$Y = o_1, o_2, \ldots, o_n$，其中$o_i$表示第$i$个词语对应的成分预测标签。模型由三大模块组成，分别是基于BERT-wwm-ext的词嵌入模块、基于双仿射注意力机制的评分模块以及基于TreeCRF的模型预测输出模块。模型整体架构如图 2 所示。

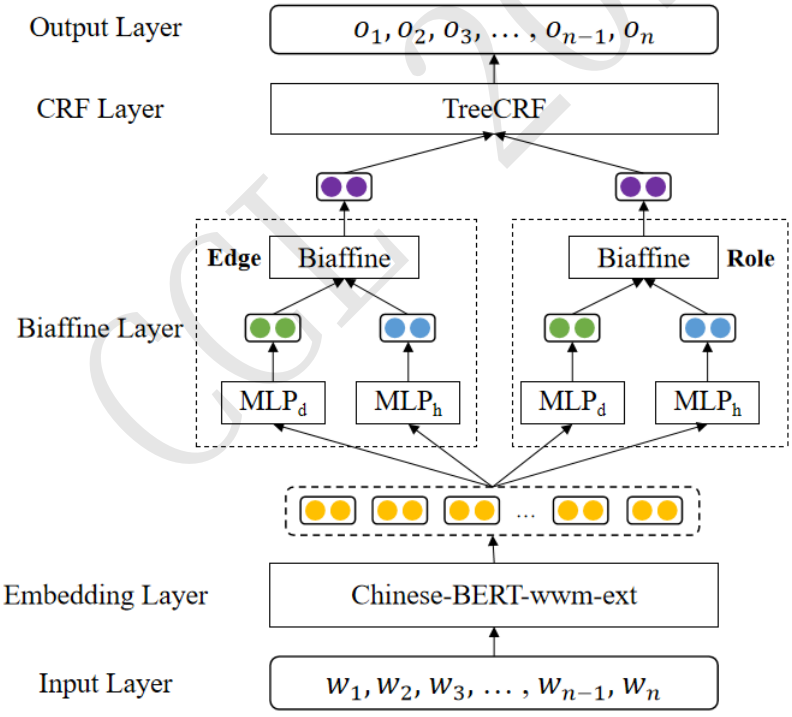


Figure 2: 被动结构成分识别模型图

4.1 基于BERT-wwm-ext的词嵌入层

本研究使用的BERT-wwm-ext预训练语言模型，是基于BERT模型和整词掩码技术(Whole Word Masking, WWM)技术扩展而来的。被动结构的成分大多数是多音节词语，普通BERT模

型在MLM(Masked Language Model)阶段是针对单字进行掩码操作，这容易导致词语被分割，影响句子的语义表示。Cui Y等人 (2021)把整词掩码(Whole Word Masking, WWM)技术引入BERT模型，在训练过程中，更加关注整个词语的语义，从而更好地捕捉语言的上下文信息。为进一步提升中文自然语言处理任务性能，该团队又往BERT模型训练语料增加了大量维基百科、新闻、问答等通用数据，同时对语料进行了严格的数据清洗和预处理，以确保训练得到的模型更加准确和有效，提出了BERT-wwm-ext模型。该模型一方面对中文词语的语义表示性能更好，另一方面丰富了对新闻领域文本的解析能力，因此它相比其他预训练模型更利于被动结构各成分的识别。

4.2 基于Biaffine Attention的评分模块

本文将基于依存分析的SRL方法应用到被动结构成分识别上，把一个句子中的所有被动结构转化为依存树结构，如图3所示，句中的一个被动结构可以转化为一棵多叉树，第一层为树的根节点，第二层为动词，第三层为被动结构除动词外的其他成分。被动结构转化为树结构后，模型的训练目标是从句子中解析出最佳子树，实现这一过程的关键步骤是对依存树进行评分，本文使用双仿射注意力(Bi-affine Attention)来实现这一过程。

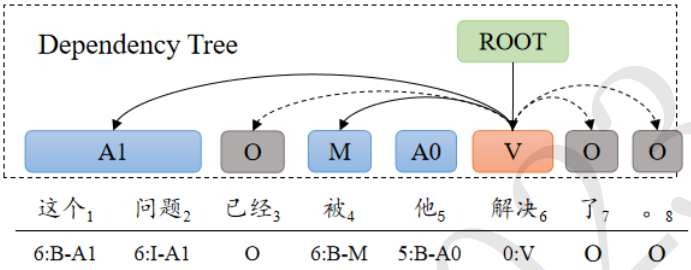


Figure 3: 被动结构的依存树形式

双仿射注意力机制 (Timothy et al., 2017)是一种用于解决依存句法分析问题的神经网络结构，具体来说，双仿射机制首先将每个词的词向量作为输入，经过一些线性变换和激活函数处理之后，得到一个隐向量表示。然后，使用双仿射函数来计算每对词之间的相关性得分，这个得分可以表示两个词之间的依存关系强度。对于两个词i和j，双仿射函数可以表示为公式(1):

$$Biaffine(i,j) = h_i^T W_1 U W_2^T h_j \tag{1}$$

其中，h_i和h_j分别表示词i和j的隐向量表示，W_1和W_2是可学习的权重矩阵，U是一个对称的得分矩阵，可以表示两个词之间的依存关系强度。

依存树的评分过程由结点预测和标签预测两个子任务构成。结点预测子任务是预测两个结点之间是否存在依存关系，标签预测子任务则是预测两个结点之间存在的被动关系属于何种具体关系。在本任务中有三种关系，第一种是动词与根节点之间的关系，记为ROOT-V；第二种是被动结构除动词外的其他成分与该动词之间的关系，记为V-A1、V-A0和V-M；第三种是对句中非被动结构的成分，记为V-O。

对于输入样本$S = w_1, w_2, \ldots, w_n$，经过预训练模型得到词嵌入$X = x_1, x_2, \ldots, x_n$。对句中任意两个结点$a$和$b$，判断二者之间是否存在依存关系，即依存树中是否存在$a \rightarrow b$的依存弧，先定义两个多层感知机MLP^h和MLP^t分别计算依存弧首尾两个结点的隐向量，然后把两个结点的隐向量代入双仿射函数中计算依存强度得分，如公式(2)-(4)。

$$r_a = MLP^h(x_a) \tag{2}$$

$$r_b = MLP^t(x_b) \tag{3}$$

$$Score(a \rightarrow b) = Biaffine(r_a, r_b) \tag{4}$$

标签预测子任务的算法与结点预测子任务类似，也是通过两个MLP结合一个Biaffine函数计算得分，最终遍历完所有可能的结点对，就可得到一棵依存树的得分。

4.3 基于TreeCRF的预测输出层

TreeCRF模型(McDonald and Pereira, 2006)是一种条件随机场模型，用于处理树形结构的数据。在TreeCRF模型中，每个节点表示一个观测变量，每条边表示一个潜在变量，即节点之间的关系。给定一个树和观测变量序列，TreeCRF模型的概率分布可以表示为一组特征函数的乘积，其中特征函数描述了每条边标注的条件概率。模型的学习是通过训练特征函数的权重来实现的。在预测时，使用维特比算法解码，得到最可能的标注序列。

给定一个树T和观测变量序列$x = (x_1, x_2, \ldots, x_n)$，TreeCRF模型的概率分布可以表示为：

$$P(y|x,T) = \frac{1}{Z(x,T)} \prod_{(i,j)\in E} \psi(y_{i,j}, x_i, x_j) \tag{5}$$

其中，E表示任意两个点之间构成边的集合，$y = (y_{1,2}, y_{1,3}, \ldots, y_{n-1,n})$ 表示给定观测变量x 对应的树T 的边上的潜在变量，$\psi(y_{i,j}, x_i, x_j)$表示边(i,j)的特征函数，$Z(x,T)$是归一化常数，用于保证模型的概率分布性质成立，即

$$P(x,T) = \sum_{y} \prod_{(i,j)\in E} \psi(y_{i,j}, x_i, x_j) \tag{6}$$

通过学习特征函数$\psi(y_{i,j}, x_i, x_j)$的权重，可以得到TreeCRF模型。在预测时，可以使用维特比算法进行解码，得到最可能的边的标注序列$\hat{y}$，即

$$\hat{y} = argmax_y P(y|x,T) \tag{7}$$

5 被动结构成分识别实验

5.1 参数设置及评价指标

本文的实验数据集按照6:2:2划分为训练集、验证集和测试集并随机打乱，实验使用的预训练模型基本参数为L-12_H-768_A-12，具体数据集划分和模型超参数设置如表 3和表 4所示。

数据集	句子数量（条）
训练集	2303
验证集	768
测试集	768

Table 3: 数据集划分

超参数	含义	值
epochs	数据集迭代次数	10
batch_size	单批次样本数量	128
pad_size	每个样本最大token数量	128
learning_rate	学习率	5e-5
dropout	丢弃概率	0.1

Table 4: 超参数设置

本文实验按准确率P、召回率R和F1得分进行评价，公式如(8)-(10)所示。

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{8}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{10}$$

其中，TP 表示模型正确预测的语义角色的数量，FP 表示模型错误预测的语义角色的数量，FN 表示模型未能正确预测的语义角色的数量。

5.2 实验结果与分析

本文采用CRF模型作为基线模型，在此基础上分别采用多种方法进行模型的构建并进行实验对比。首先对比了在相同预训练模型(GloVe)情况下使用句法依存分析方法（Biaffine +TreeCRF）和传统CRF序列预测的方法解决被动句解析的性能差异；然后对比了在相同序列预测模型下不同预训练模型对性能的影响；最后将这些模型与本文提出的模型，即BERT-wwm-ext+Biaffine+TreeCRF进行实验结果对比和分析。实验结果如表 5所示。 由实验结果

模型	P(%)	R(%)	F1(%)
CRF	78.53	76.68	77.59
Biaffine+TreeCRF	81.34	79.63	80.47
BERT+CRF	94.34	92.63	93.48
BERT+Biaffine+TreeCRF	97.18	95.72	96.44
本文模型	**98.46**	**96.18**	**97.31**

Table 5: 被动结构成分识别实验结果

可知，首先，Biaffine+TreeCRF模型比CRF的性能提高近3个百分点，说明将被动结构建模为一种依存句法树结构更有利于解析。这是由于一个句子中可能含有多个被动结构，且多个被动结构之间又可能存在嵌套，但CRF这种传统的序列标注方法无法应用于嵌套识别任务，而Biaffine+TreeCRF采用句法依存分析的方法，将每个被动结构解析为独立的依存树，不会相互影响，所以性能得到了明显的提升。

其次，相比于静态词向量GloVe，不论是CRF还是Biaffine +TreeCRF模型，在使用动态词向量预训练模型BERT之后，模型的性能都得到了较大的提升。这是因为静态词向量模型是基于全局统计信息，无法很好地处理不同语境中的上下文信息。而BERT采用了双向Transformer结构，可以同时考虑前后文信息，使得生成的词向量更具有上下文的代表性。

最后，相比上述四种模型，本文提出的模型取得了最好的解析性能，其F1值达到了97.31%。一方面是由于BERT-wwm-ext预训练模型采用了整词掩码技术，在MASK操作时能更好地学习到被动结构中每个词语完整的语义，从而提高模型的泛化能力和语义表示能力。另一方面该预训练模型是专门针对中文训练的模型，在预训练过程中学习了额外的新闻领域的文本知识，而本文数据基础也来源于新闻语料，因此能更好地捕捉中文句子的语义特征，进而提升了被动结构成分识别的性能。

通过对测试集中预测错误的被动结构进行分析，发现错误主要有两种情况：一是定中结构中的被动结构识别较差。如“.....群众互助互济活动的广泛开展。”中的“开展”一词是具有制作义的自主动词，能进入被动语态，但由于数据集中涉及到状中结构的句子占比只有不到2%，因此模型在训练时难以学习这种特殊结构的语义和句法特征。二是含多种义项的词语被错误识别为被动结构中的动词。例如“车辆乱停放等问题十分突出。”中的“突出”是动词兼形容词，而在该句中是作形容词，造成这一问题可能是由于句中的主语是“问题”，而这个词语在数据集中充当被动结构中的A1频次较高，使得模型对这个词语较为敏感，进而导致模型识别错误。

6 CAMR后处理实验

AMR是一种领域无关的句子语义表示方法，它将一个句子的语义抽象为一个单根有向无环图，其中句子中的实词抽象为概念节点，实词之间的关系抽象为带有语义关系标签的有向弧(曲维光等, 2017)。中文AMR也称作CAMR，它在AMR的基础上对汉语中常见的和特殊的语言现象作了细致的定义。但CAMR现有的自动解析模型对被动句的解析还存在一定的不足。本节实验通过利用被动结构成分的解析结果对CAMR解析图进行后处理，以期提升CAMR被动句解析任务的性能。

6.1 SPRING模型对被动句的自动解析

SPRING模型是BevilacquaM等人(2017)提出的一种AMR自动解析架构，此架构可以完成文本到AMR的解析和AMR到文本的生成两种任务，即利用BART的迁移学习能力完成这两个任务。本文首先利用被动句自动识别模型(Hu et al., 2022)从CAMR小学语文语料中识别被动句，经过自动识别和人工校对，筛选出有标记和无标记被动句各80条，共计160条被动句，对每个句子中的被动结构进行人工标注。然后利用SPRING模型中的Text-to-AMR任务框架进行AMR自动解析，考察了SPRING模型对被动句的解析性能。

Figure 4: CAMR黄金标注与自动解析示例

图4是句子“一天早晨，巨人被喧闹声吵醒了。”的黄金标注和自动解析图，句中包含一个被动结构“巨人$_{A1}$+被$_{M}$+喧闹声$_{A0}$+吵醒$_{V}$”。观察图(a)，句子的谓语动词解析为“x7_吵醒”，其中“x7”是一个概念对齐信息，它表示“吵醒”一词是句子中的第7个词语。同理，受事是“x4_巨人”，作为动词的arg1子结点；施事是“x6_喧闹声”，作为动词的arg0子结点。标记词“被”作为关系对齐包含在关系有向弧arg0中。图(c)是SPRING自动解析图，一方面，该CAMR图中概念节点“巨人”和“喧闹声”对应的语义角色标签解析错误了，正确的SPRING解析图应如(b)所示；另一方面，SPRING模型的解析结果不含概念对齐和关系对齐信息，即缺乏概念节点与句中词语间的索引信息和语义关系有向弧上的虚词信息，导致CAMR解析图中丢失了有标记被动结构中十分重要的标记词信息。

因此，为更加全面地了解SPRING模型对被动句的解析情况，本文基于CAMR图的结构设计了一组被动结构解析正确与否的判定规则，用于实现该模型对被动句解析正确率的统计分析。具体规则如下：

（1）**动词是否正确解析。**在以该动词为中心的被动关系对应的CAMR图中的某个子树，动词抽象而来的概念节点应当是这个子树的根节点。

（2）**受事主语是否解析为动词的arg1。**在当前子树中，受事对应的结点应当是动词概念结点的孩子节点，且关系标签是arg1。

（3）**施事是否解析为动词的arg0。**与arg1同理，若当前被动关系中出现了施事，则其对应的概念结点也应当是动词概念结点的孩子结点，关系标签是arg0。

（4）**不考虑概念和关系对齐信息。**由于SPRING模型的解析结果不具有概念和关系对齐信息，因此在判定的时候仅关注前三条规则。

利用上述判定规则对SPRING模型的CAMR解析图进行判定，如图4(b)可判定为解析正确，而图4(c)解析错误。同时利用本文提出的PS-CRF模型对160条被动句进行解析，统计两类被动句的自动解析正确率，实验结果如表 6所示。

类别	模型	黄金标注(条)	正确解析(条)	正确率(%)
有标记被动句	SPRING	80	36	45.00
	PS-CRF	80	76	**95.00**
无标记被动句	SPRING	80	42	52.50
	PS-CRF	80	73	**91.75**

Table 6: 两种模型解析被动句的正确率

可以看出，SPRING模型在被动句解析上性能较差，而本文提出的PS-CRF模型取得了良好的性能，两种被动句的解析正确率均达到90%以上。为提升CAMR对被动句的解析性能，本文尝试利用PS-CRF模型的识别结果对CAMR图进行后处理。

6.2 后处理算法设计

针对CAMR对被动句解析存在不足的问题，本节设计了一个CAMR后处理算法，来纠正CAMR图中错误解析的被动关系。CAMR后处理算法分为三个步骤：第一，把AMR图和被动结构成分都转化为多元组形式；其次，补充或修改被动结构中的概念节点和关系；最后，把AMR多元组还原成AMR解析图。算法流程图如图 5所示。

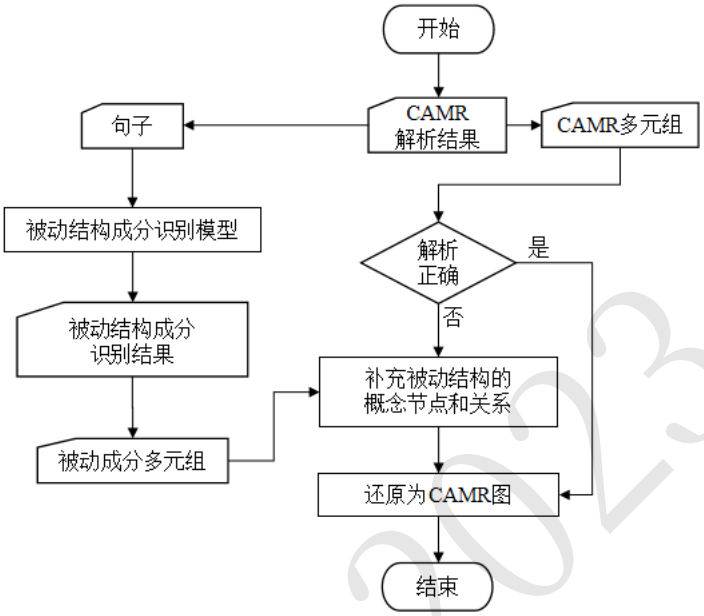


Figure 5: CAMR后处理算法流程图

由于SPRING解析结果不含概念对齐和关系对齐信息，而被动结构中的标记词对应CAMR中的关系对齐，且概念对齐信息有利于从CAMR解析图中定位相关的词语，便于后处理操作。因此本文还设置了两组对照实验，即先通过人工补充SPRING解析图中的概念对齐信息，再进行后处理操作。图 6是句子“许多人被火围困在山顶上。”的SPRING解析图在人工添加概念对齐信息前后的对比图，如(a)和(b)。在此基础上，分别利用被动解析结果对其进行后处理，得到图(c)和(d)。

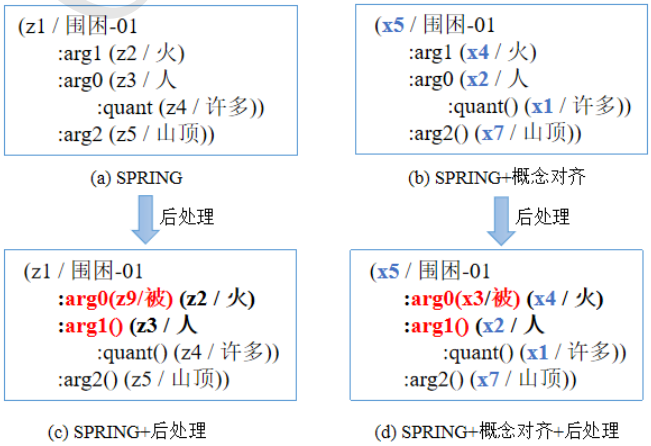


Figure 6: CAMR后处理的不同实验设置

6.3 实验结果与分析

本文采用了Align-Smatch和Smatch两种指标对CAMR解析图进行评价，其中Smatch是基于英文AMR设计的评测指标，Align-Smatch则是一种基于CAMR设计的评价指标，它兼容了中文AMR所特有的概念和关系对齐信息，还可以对有向弧上的虚词进行评估(肖力铭等, 2022)。实验结果如表 7所示。

模型	Align-Smatch(%)	Smatch(%)
SPRING	45.2	69.8
SPRING+后处理	47.6↑	70.7↑
SPRING+人工对齐	62.3	—
SPRING+人工对齐+后处理	65.8↑	—

Table 7: CAMR后处理实验结果

由实验结果可知，在不添加概念对齐信息时，CAMR解析图在利用被动结构成分进行后处理之后，两种评价指标得分均有所提高，其中Align-Smatch提升较大而Smatch值提升略低。这是因为Smatch指标是针对英文设计的，它在对两个AMR图进行匹配评分时，只关注概念节点和边的标签，且标签不含附加成分，实验使用的后处理数据是被动结构，其中施受事的修改、补充是概念节点层面的后处理，而有标记被动结构中的标记词作为论元关系标签中的附加成分，不作为独立的概念节点，因此对于Smatch评价指标而言，后处理带来的性能提升仅仅是由于动词及其施受事成分而不包含标记词，而Align-Smatch把所有的被动结构成分都利用了，所以提升效果明显。此外，后处理之前的Align-Smatch值比Smatch值低，这是由于Align-Smatch在计算得分的时候，不仅仅关注概念节点的匹配程度，更重要的是概念对齐信息和关系对齐信息，而现有的CARM解析器包括SPRING，生成的解析图都不包含这两种对齐信息，所以导致Align-Smatch得分较低。

而在对SPRING解析图人工添加概念对齐信息后，Align-Smatch得分由45.2提升到62.3，这验证了对齐信息对于Align-Smatch指标的重要性，在此基础上再利用被动结构成分进行后处理，分值达到了65.8，提升了3.5个百分点。而在人工添加概念对齐信息之前，后处理操作带来的性能提升为2.4个百分点，由此可见本文提出的被动结构成分识别模型对中文AMR的解析性能有一定的提升效果，尤其是对于包含对齐信息的CAMR图。

7 结语

本文把被动结构成分识别任务建模为一种语义角色标注任务。首先对被动句中的具体结构成分进行了细粒度标注；然后提出了一种BERT-wwm-ext预训练模型结合双仿射注意力机制的CRF序列标注模型，该模型取得了较好的解析性能，F1值达到了97.31%；最后基于CAMR小学语文语料，将本文模型应用到CAMR解析后处理任务中，提升了CAMR对被动结构的解析性能。

在后续工作当中，一方面我们将进一步完善标注规范，尤其是针对特殊被动结构和动词的标注，提升数据标注的一致性、平衡性。另一方面，考虑对被动句自动解析模型进一步优化，尝试融入更多语言学知识，以增强模型的可解释性。

参考文献

Bevilacqua M, Blloshmi R, and Navigli R. 2021. *One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. Proceedings of the AAAI*, 12564-12573.

Cui Y, Che W, Liu T, Qin B, and Yang Z. 2021. *Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 3504-3514.

Hu K, Qu W, Wei T, Zhou J, Gu Y, and Li B. 2022. *Automatic Recognition of Chinese Passive Sentences Based on Feature Fusion. Proceedings of the CCL*, 384-394.

Li Z, Zhao H, He S, Zhang Y, Zhang Z, Zhou X, and Zhou X. 2019. *Dependency or span, end-to-end uniform semantic role labeling. Proceedings of the AAAI*, 33(01): 6730-6737.

Li Z, Zhao H, He S, and Cai J. 2021. *Syntax role for neural semantic role labeling. Computational Linguistics*, 47(3): 529–574.

McDonald R, and Pereira F. 2006. *Online learning of approximate dependency parsing algorithms. Proceedings of the EACL*, 81-88.

Timothy D, and Christopher D. M. 2017. *Deep biaffine attention for neural dependency parsing. Proceedings of ICLR*, 2017.

Zhang Y, Xia Q, Zhou S, Jiang Y, Li Z, Fu G, and Zhang M. 2022. *Semantic Role Labeling as Dependency Parsing: Exploring Latent Tree Structures inside Arguments. Proceedings of the COLING*, 4212-4227.

Zhang Z, Emma S, and Eduard H. 2021. *Comparing span extraction methods for semantic role labeling. Proceedings of the SPNLP*, 67-77.

兰宾汉. 2002. 汉语语法知识与应用. 北京：石油工业出版社, 2002:123.

李斌, 闻媛, 宋丽, 卜丽君, 曲维光, 薛念文. 2017. 融合概念对齐信息的中文*AMR*语料库的构建. 中文信息学报, 31(06):93-102.

马庆株. 1988. 自主动词与非自主动词. 中国语言学报, 1998(03):157-180.

曲维光, 周俊生, 吴晓东, 戴茹冰, 顾敏, 顾彦慧. 2017. 自然语言句子抽象语义表示*AMR*研究综述. 数据采集与处理, 32(01):26-36.

宋文辉, 罗政静, 于景超. 2007. 现代汉语被动句施事隐现的计量分析. 中国语文, 2007(02):113-124.

王惠. 1997. 从及物性系统看现代汉语的句式. 语言学论丛, 1997:19.

王一平. 1994. 从遭受类动词所带宾语的情况看遭受类动词的特点. 语文研究, 1994(04):28-34.

王振来. 2004. 被动表述对自主动词和非自主动词的选择. 汉语学习, 2004(06):17-22.

肖力铭, 李斌, 许智星, 霍凯蕊, 冯敏萱, 周俊生, 曲维光. 2022. 基于概念关系对齐的中文抽象语义表示解析评测方法. 中文信息学报, 36(1): 21-30.

蚁坤. 2000. 汉语被动句的句法语义特征和使用条件. 北京语言文化大学.

赵元任. 1979. 汉语口语语法. 北京：商务印书馆, 1979:134-176.

人工智能生成语言与人类语言对比研究 ——以ChatGPT为例

朱君辉[1]，王梦焰[1]，杨尔弘[1]*，聂锦燃[1]，王誉杰[2]，岳岩[1]，杨麟儿[1]
北京语言大学[1]
北京交通大学[2]
nysyzxzjh@163.com

摘要

基于自然语言生成技术的聊天机器人ChatGPT能够快速生成回答，但目前尚未对机器作答所使用的语言与人类真实语言在哪些方面存在差异进行充分研究。本研究提取并计算159个语言特征在人类和ChatGPT对中文开放域问题作答文本中的分布，使用随机森林、逻辑回归和支持向量机（SVM）三种机器学习算法训练人工智能探测器，并评估模型性能。实验结果表明，随机森林和SVM均能达到较高的分类准确率。通过对比分析，研究揭示了两种文本在描述性特征、字词常用度、字词多样性、句法复杂性、语篇凝聚力五个维度上语言表现的优势和不足。结果显示，两种文本之间的差异主要集中在描述性特征、字词常用度、字词多样性三个维度。

关键词： ChatGPT ； 人类语言 ； 语言特征 ； 对比 ； 机器学习

A Comparative Study of Language between Artificial Intelligence and Human: A Case Study of ChatGPT

Junhui Zhu[1], Mengyan Wang[1], Erhong Yang[1] , Jinran Nie[1], Yujie Wang[2], Yan Yue[1], Liner Yang[1]
Beijing Language and Culture University[1]
Beijing Jiaotong University[2]
nysyzxzjh@163.com

Abstract

This paper aims to explore the differences between the language used in human-generated responses and responses generated by ChatGPT, a chatbot based on natural language generation technology. The study extracts and computes the distribution of 159 language features in real human text and ChatGPT-generated text. To evaluate the performance of these features, the study employs three machine learning algorithms: Random Forest, Logistic Regression, and Support Vector Machine (SVM).The experimental results demonstrate that both Random Forest and SVM can achieve high classification accuracy. The result reveals that the two texts differ significantly in three dimensions: descriptive features, word commonness, and word diversity.

Keywords: ChatGPT , Human language , Linguistic features , Comparison , Machine learning

* 通讯作者

第二十二届中国计算语言学大会论文集，第523页-第534页，哈尔滨，中国，2023年8月3日至5日。

1 引言

近年来，随着大数据的支持和计算能力的不断增强，人工智能（AI）在自然语言生成领域取得了长足的进展，特别是在机器翻译、对话生成和文章摘要等任务中，机器生成的语言已经达到了一定的准确性和自然度，并且具备了自己的语言风格。其中，基于神经网络的自然语言生成模型——如GPT系列（Generative Pre-trained Transformer）已成为当今最流行的自然语言处理技术之一。2022年11月30日，OpenAI发布了ChatGPT (Ouyang et al., 2022)，该模型以一问一答的对话形式设计，在理解用户查询和生成类人文本方面表现出色。在中文上，它也能够生成流畅、符合语法的回答，适用于来自各个领域不同类型的问题，引起了极大的关注。

虽然机器生成的语言在语法与逻辑性方面越来越接近于真实语言，但与人类真实语言相比，机器生成的文本在词汇、句法结构、衔接关系等具体语言特征的使用方面仍存在着一些明显的差异。分析这些语言特征的差异对于提高语言模型生成自然语言的准确性和真实性，以及认识人类智能与人工智能的区别至关重要。已有研究发现，机器生成的文本中高级的句法、语义特征占比较低(Pu et al., 2022)，缺乏情感和人情味，难以表达真实人类的情感和感受(Ma et al., 2023)等。然而，目前仍未出现从语言特征的角度深入挖掘二者在语言使用上的差异研究，另一方面，ChatGPT在中文使用上的表现如何也仍未得到探讨。

语言特征的提取与分析能够有效揭示文本中存在的语言规律，广泛应用于体裁分析和语言习得研究。例如，研究者使用一系列特征来识别口语和书面文本之间的语言差异(Louwerse et al., 2004)，正式和非正式类型(Dempsey et al., 2007)，不同文本的年代和作者差异，以及通过词汇丰富度、词汇密度、句法复杂性和句法相似性等特征的测量，探究语言学习者的词汇和句法知识水平(Crossley and McNamara, 2010; Nasseri and Thompson, 2021)等。目前，对ChatGPT生成语言的研究尤其是对比分析人类语言和ChatGPT语言差异的研究尚不多见。

随着计算机技术和自然语言处理技术的发展，研究者们主要聚焦于使用预训练语言模型探测人工智能生成的文本(Dou et al., 2021; Guo et al., 2023; Mitchell et al., 2023; Mitrović et al., 2023)。然而，采用经典特征工程的研究范式建立机器学习模型，操作简便、易于落地，并且能够直观地解释语言特征在其中的作用，仍具有其独特的价值和作用。

有鉴于此，本文从多维度语言特征的视角，深入探究机器生成文本与人类真实语言之间的差异所在，尝试挖掘影响二者语言风格的关键语言因素。具体来说，对平行问答语料进行自然语言处理，通过文本分析工具提取文本中不同维度的语言特征，基于机器学习方法构建分类模型，并对比分析各维度语言特征在人类与机器回答中的分布，以探索各自的语言风格。

2 研究方法

本研究基于ChatGPT与人类对中文开放域问题给出的回答，借助中文CTAP工具(Cui et al., 2022)与Python编程语言对二者语言特征进行量化，训练分类模型并选出预测力较强的特征，从各个维度研究ChatGPT生成文本与人类语言的差异。

具体过程如下：1）选取ChatGPT与人类在开放域问答中的6586篇语料作为研究样本；2）对语料进行分词、词性标注、短语结构标注等预处理，分别计算机器生成文本与人类真实文本五个维度下159项语言特征值；3）训练机器学习模型作为分类器，进而找出区分机器语言与人类语言的最具预测能力的特征；4）基于样本均值比较所选语言特征值在两种文本的分布，观察对模型贡献度强的语言特征，分析两者在不同维度上语言的表现。

2.1 研究问题

本研究主要讨论以下两个问题：1）使用特征工程结合机器学习的方法是否能够有效地区分人类的回答文本与ChatGPT的生成文本？哪些特征是有效预测变量？

2）ChatGPT生成语言与人类语言在不同维度特征上的表现有何具体差别？分别存在哪些优势与不足？

2.2 语料处理

本研究使用的语料来自于(Guo et al., 2023)发布的人机问答语料，选取开放域（不区分专

基金项目：高级汉语水平学习者多模态口语语料库的构建及应用(ZDI145-24)

业领域）下6586篇分别由人类与ChatGPT作答的平行语料构建人机问答语料库。为了更精确地提取语言特征，调用自然语言处理工具Stanford CoreNLP(Manning et al., 2014)工具依次对语料进行分词、词性、短语结构、依存句法等自动标注。

2.3 语言特征测量指标选取

本研究使用中文CTAP与Python编程语言提取两种文本中的语言特征。中文CTAP是一个全面的文本特征自动分析平台，能够分析文本的表层和深层语言特征，包括字、词、句、篇四个维度下的196个特征，可用于母语或二语等多种类型的文本测量指标提取。本研究选择了涵盖描述性特征、字词常用度、字词多样性、句法复杂性、篇章凝聚力五个方面的160个语言特征作为提取对象。在对文本语言特征进行量化时，我们以单个回答作为测量单位，主要使用总数、均值、方差、比例、比值和型例比（TTR）六种通用计算方法。接下来，分别计算159个指标在6586篇人类和ChatGPT回答文本中的测量值，对二者的平均值进行统计分析，初步探索其中的语言差异。在考察的160个语言特征中，删除测量值过小的“拟声词密度”特征后（$\bar{x} < 0.001$），最终确定159个语言特征作为分析对象。

3 实验

对于研究问题一，本节基于前文提出的159 种语言特征，通过传统机器学习算法构建人工智能探测器。本节将评估三种分类算法的预测能力，同时筛选出贡献度较高的有效预测变量。

3.1 模型构建

本研究属于文本分类任务。文本分类常用的机器学习算法有决策树（Decision Tree，DT）、随机森林（Random Forest，RF）、逻辑回归（Logit Regression，LR）、最近邻（K Nearest Neighbor，KNN）、支持向量机（Support Vector Machine，SVM）等，本文选取研究者使用较多的逻辑回归、SVM、随机森林三种经典文本分类算法构建分类模型。

1）逻辑回归（LR）是一种广泛使用的二分类模型，逻辑回归的目标是学习一个权重向量（或模型参数），使得模型能够最大程度地准确地预测二元输出变量。

2）支持向量机（SVM）能够有效地解决分类问题，特别是高维数据和非线性分类问题。SVM模型的主要思想是寻找一个最优的超平面，将不同类别的数据点分隔开来。其目标为最大化支持向量与分类边界之间的间隔，具有很好的泛化性能和较高的精度。

3）随机森林（RF）分类模型是一种集成算法，通过组合多个CART 决策树作为弱分类器，最终结果通过投票或取均值，具有较高的精确度和泛化性能。

在构建分类模型之前，对于每个特征值，我们进行了标准化处理，以避免不同特征值之间的比较出现偏差。我们将数据集随机分为训练集和测试集，比例为8:2，使用测试集对模型进行评估。通过准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1值（F1-score）四种常用指标来评估模型的分类性能。

3.2 有效预测变量及其预测力

实验结果表明，三种分类器都表现良好，如表1所示。其中，SVM的准确率与精确率最高，分别达到了97.27%与97.04%。随机森林的召回率最高，为98.07%。综合来看，SVM在分类性能最为优异。在F1值评估指标上，随机森林与支持向量机（SVM）均展示出了良好的性能。

分类模型	准确率（%）	精确率（%）	召回率（%）	F1值（%）
逻辑回归	96.36	96.99	95.83	96.41
随机森林	97.19	96.49	98.07	97.27
SVM	97.27	97.04	97.62	97.33

表 1: 三种机器学习模型性能对比

在三种模型中，随机森林和SVM均能达到较高的分类准确率，我们基于随机森林和SVM模型在159个语言特征中筛选贡献度较高的有效预测变量。随机森林算法默认采用基尼系数（Gini index）作为特征重要性的量化指标。具体来说，基尼系数越大，那么该语言特征所包含的信息

量就越大，对文本分类的影响力也越大。线性SVM通过找到一个能够最大化类别间隔的超平面来对数据进行分类，超平面由一个权重向量决定，每个分量对应一个特征的权重，特征权重的绝对值越大，说明特征对分类器的预测能力越重要。在本研究中，我们采用随机森林模型中各个语言特征的基尼系数进行评估，并根据这些指标建立特征影响力的排序序列，选取前31个具有较高影响力的特征（基尼系数>0.10）。接下来，我们从支持向量机（SVM）分类器中提取特征权重，并计算特征权重绝对值的均值作为筛选阈值。经过筛选，共有59个特征满足条件。通过综合运用这两种机器学习模型，我们共确定了77个关键特征，在这些特征中，有12个特征在两种算法的结果中均有所体现。这些结果为我们在分析和解释语言特征的差异上提供了重要的参考依据。

4 不同维度语言特征差异分析

参考贡献较大的77项特征，我们对描述性特征、字词常用度、字词多样性、句法复杂性、篇章凝聚力五个维度159项特征依次展开分析。

4.1 描述性特征

描述性特征指对文本中字、词、句、篇四个层面的基本描述性统计，用于表征文本中各个层面语言单位的数量、长度等，属于文本的视觉属性。我们将37项语言特征作为描述性特征维度的测量指标，具体如表2所示。

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
笔画	**少笔画字数**	73.033	134.343	词汇	三音词占比	0.044	0.042
	少笔画字比例(1-8)	0.668	0.651		**四音节及以上词占比**	0.027	0.047
	中笔画字数(9-16)	35.039	61.347		四音节及以上词数	1.653	4.636
	中笔画字比例(9-16)	0.325	0.297		平均词长	1.704	1.861
	高笔画字数(16以上)	0.364	0.441	句子	句子数	4.207	7.343
	高笔画字比例(16以上)	0.003	0.002		平均句长(以字为单位)	40.893	42.396
	字例平均笔画数	7.330	7.102		平均句长(以词为单位)	25.067	21.823
	字形平均笔画数	7.435	7.168		句长标准差(基于词例)	9.248	6.729
部件	**字形平均部件数**	1.754	1.692		**句长标准差(基于词形)**	6.838	5.042
	字例平均部件数	1.744	1.684		句长标准差(基于字例)	15.150	12.842
字数	**字例数**	134.617	262.117		句长标准差(基于字形)	10.034	7.654
	字形数	79.719	104.134		最长句字数	63.129	61.623
	词例数	84.040	146.615		最长句词数	38.447	31.858
	词形数	56.705	73.460	段落	**篇章段落数**	1.442	3.681
词汇	单音节词数	35.666	47.878		**最长段落长度(基于词)**	127.121	113.591
	单音节词占比	0.483	0.379		**最长段落长度(基于字)**	80.371	63.917
	双音节词数*	32.360	68.705		**平均段落长度(基于字)**	123.907	92.747
	双音节词占比	0.445	0.532		**平均段落长度(基于词)**	78.308	51.882
	三音节词数	3.005	5.226				

注：带*表示在两种算法中均贡献度突出，加粗表示在一种算法中贡献度突出，以下各表同理。

表 2: 描述性特征维度人类与ChatGPT文本特征均值对比

在37项描述性特征中，在两种算法中均贡献度突出的是双音节词数，在任一种算法中贡献度突出的有少笔画字数、中笔画字比例(9-16)、字形平均部件数等14项，占所有重要特征总数的19.5%。观察表2可知，ChatGPT语言特征指标高于人类语言的有17个，集中在笔画、字数、词数三个层面；低于人类语言的有20个，集中在部件、句长、段落三个层面。

汉字包含的笔画数与部件数一定程度上体现了汉字在书写方面的复杂程度。理论上说，在文本材料中，笔画数或部件数较多的汉字占比越大，汉字的字形复杂度越高，文本阅读难度越大(张倩倩, 2022)。去除文本长度因素的影响，在人类回答文本与ChatGPT生成文本中均为少笔画字占比最大，中笔画字次之，高笔画字占极低。即无论对于人类还是机器，少笔画字及中笔画字即可基本满足回答大多数开放域问题所使用语言的需求。不同的是，在人类回答文本

特征类别	汉语特征(频数)	人类	GPT	特征类别	汉语特征(频数)	人类	GPT
字形	**平均对数字形1**	6.256	6.123	实词词形	**平均对数实词词形1***	5.209	5.156
	平均对数字形2	4.021	2.923		**平均对数实词词形2**	3.441	3.239
	平均对数字形3	2.646	2.442		**平均对数实词词形3**	3.965	3.941
字例	**平均对数字例1**	6.424	6.202	实词词例	**平均对数实词词例1**	5.291	5.197
	平均对数字例2	2.945	2.862		**平均对数实词词例2**	3.517	3.302
	平均对数字例3	2.747	2.560		平均对数实词词例3	4.044	4.013
词形	**平均对数词形1***	5.250	5.288	虚词词形	**平均对数虚词词形1**	6.295	6.283
	平均对数词形2	2.579	2.273		**平均对数虚词词形2**	4.413	4.279
	平均对数词形3	2.923	2.722		平均对数虚词词形3	4.244	4.135
词例	平均对数词例1	5.528	5.475	虚词词例	**平均对数虚词词例1**	6.643	6.731
	平均对数词例2	2.753	2.590		**平均对数虚词词例2**	4.754	4.758
	平均对数词例3	2.122	2.970		平均对数虚词词例3	4.474	4.469

注：1 Gigaword字词频表　2 现代汉语语料库　3 汉语二语教材语料库

表 3: 字词常用度维度人类与ChatGPT文本特征均值对比

中，少、中、高笔画字所占比例与字形平均笔画数、字例平均笔画数均高于ChatGPT。数据表明人类拥有较广的中高笔画汉字储备量，在一定程度上更具备使用较多笔画数汉字的能力。

文本中各个语言单位的长度和数量能够用来衡量文本难度与文本质量(熊兵, 2016; 李绍山, 2000; 邢诗吟, 2022)。一般来说，文本中词数、句子数、段落数越多，平均词长、平均句长越长，文本的质量越高，难度越高。本研究测量了词长、句长、段落长度、语篇长度以及字例数、词例数、句子数及段落数等指标。在词汇层面，相较于人类的用词，ChatGPT生成语言倾向于使用大词（词长较长的词），表现在平均词长、双音节词、四音节词、四音节及以上词占比相对较高。与之照应，在句长上，ChatGPT语言以字为单位的平均句长大于人类语言，以词为单位的平均句长小于人类语言，即ChatGPT生成的句子字数更多，词数却更少。值得关注的是，在人类所给出的回答中，最长句子所包含的字数和词汇量均超过了ChatGPT，这表明人类具有生成更为复杂和详尽句子的能力。在段落层面，ChatGPT生成文本的平均段落长度和最长段落长度均低于人类文本。在语篇层面，语篇长度可由字例数、词例数、句子数及段落数体现，对于这四项指标，ChatGPT生成语言的测量值均高于人类语言。反映出人类的回答更加简短，倾向于将较长的回答浓缩在较少的自然段中；而ChatGPT则擅长生成较长的答案，并进行分段阐述。从某种角度来说，ChatGPT生成的文本在难度、质量上高于人类的回答。

此外，句长变化度通过计算文本中所有句子的句长标准差得到，用来评估文本中句子长短变化的情况。句长变化度较高的文本中出现的句子大多长短不一，反之，则说明文本中句子的长度大致相同(张倩倩, 2022)。无论是基于字还是基于词测量的句长标准差特征指标，人类回答的值均高于ChatGPT回答，且最长句子字数和最长句子词数大于ChatGPT回答。可知，相比ChatGPT语言，人类回答中的句子长度之间差异更大，长短句的使用更加灵活多变。

4.2 字词常用度

字词常用度由字词使用的频率信息测量，字词在书面文本中出现的频次反映了读者的实际接触频率和熟悉程度。本文引入《Gigaword字/词频表》、《汉语二语教材语料库字/词频表》和《现代汉语语料库字/词频表》三种字/词频表，通过汉字和词汇（实词/虚词）的平均对数频数来测量字词的常用度，共计24项测量指标，如表3所示。其中，在两种算法中均贡献度突出的是平均对数词形频数（Gigaword词频表）和平均对数实词词形频数（Gigaword词频表），在任一种算法中贡献度突出的有15个，占所有重要特征总数的22.1%。

字频（字形频数、字例频数）和词频（词形频数、词例频数）可以反映汉字熟悉度与词汇熟悉度，通常被作为衡量文本难度的重要指标。已有研究表明，词频取对数后的数值与词汇识别时间之间呈现线性负相关，频率效应显著(Balota and Chumbley, 1984; Haberlandt and Graesser, 1985; 蔡建永, 2020)。具体而言，如果一些词语在已有词频表中频次较高，即在大多数文本中出现和运用的次数较为频繁，表示这类词经常被使用，读者在阅读时遇到该类词便更加迅速地从记忆中提取出来并唤醒。反之，如果一些词语在文中显示和运用的次数极少，那么该词提取和唤醒需要的时长就会更多。汉字同理。数据显示，在字词常用度的24个特征中，人

类语言的各指标值均大于ChatGPT生成语言，这说明人类在回答中使用频次高的汉字与词汇占比均多于ChatGPT，文本的阅读难度相对较低。

4.3 字词多样性

字词多样性反映的是文本中汉字与词汇的使用是否丰富多样(蔡建永, 2020)，通过文本中字词被重复使用的程度进行衡量。表4呈现了衡量字词多样性的汉字多样性、词汇多样性、实词丰富度、词汇密度四个维度50项测量指标。

其中，在两种算法中均贡献度突出的是字型例比、出现一次的字占比、词形例比、仅出现一次的词占比、实词丰富度、连词密度，在任一种算法中贡献度突出的有字Log型例比、词Log形例比、词Uber形例比等21项，占所有重要特征总数的36.4%。

我们通过计算型例比（TTR）与仅出现一次的字/词占比来测量汉字多样性与词汇多样性。型例比值越高，仅出现一次的字/词占比越高，说明字词使用越丰富。为了缓解文本长度的影响，我们还采用了研究者改良后的Log TTR、Root TTR、Uber TTR、Corrected TTR等计算方法。从上表中可以观察到，在汉字多样性和词汇多样性的两个层面上，人类语言的字、词型例比以及出现一次的字、词占比都高于ChatGPT语言。数据表明，人类回答中所使用的字词种类丰富，词汇使用具有灵活性和创造性；ChatGPT生成的文本篇幅更长，但词汇选择范围较窄，重复性强，语言使用上趋于保守。

同时，字词多样性还可以由实词丰富度体现。实词丰富度反映的是名词、动词、形容词、副词四种实词类型在所有实词中的多样性。实词用于传递信息和表达意义，文本中的实词越多，概念密度也越大，包含的信息量越大(Johansson, 2008)。观察上表可知，人类语言的实词丰富度指标几乎均大于ChatGPT。在同样篇幅的文本中，人类提供的信息量更大。

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
汉字多样性	**字型例比***	0.648	0.470	词汇密度	“被”字结构密度	0.001	0.002
	字Log型例比	0.905	0.850		**标点密度**	0.135	0.136
	字Root形例比	6.753	6.739		**代词密度**	0.052	0.069
	字Uber形例比	55.612	40.810		**动词密度**	0.207	0.186
	字Corrected形例比	4.775	4.765		方位词密度	0.012	0.013
	仅出现一次的字数	53.656	57.133		**副词密度**	0.111	0.081
	出现一次的字占比*	0.520	0.308		基数词密度	0.034	0.025
词汇多样性	**词形例比***	0.725	0.543		**介词密度**	0.029	0.043
	词Log形例比	0.923	0.872		句均词性数量	5.507	3.151
	词Root形例比	6.022	6.095		量词密度	0.027	0.019
	词Uber形例比	58.070	40.659		**名词密度**	0.267	0.290
	词Corrected形例比	4.258	4.310		**能愿动词密度**	0.024	0.031
	仅出现一次的词数	42.977	46.554		**人称代词密度**	0.031	0.042
	仅出现一次的词占比*	0.588	0.365		**实词密度**	0.745	0.713
实词丰富度	**实词丰富度***	0.822	0.647		**数词密度**	0.036	0.025
	名词丰富度	0.309	0.269		叹词密度	0.001	0.000
	Squared动词丰富度1	10.949	11.346		形容词密度	0.023	0.016
	Corrected动词丰富度1	2.215	2.313		形式动词密度	0.001	0.002
	副词丰富度	0.127	0.089		**虚词密度**	0.118	0.150
	形容词丰富度	0.030	0.019		序数词密度	0.002	0.001
	修饰语丰富度	0.157	0.108		**疑问代词密度**	0.008	0.005
	动词丰富度	0.236	0.182		**语气词密度**	0.016	0.003
	动词丰富度1	0.823	0.672		**指示代词密度**	0.010	0.009
词汇密度	**连词密度***	0.013	0.036		助词密度	0.060	0.068
	“把”字结构密度	0.002	0.001		专有名词密度	0.028	0.020

表 4: 字词多样性维度人类与ChatGPT文本特征均值对比

从另一角度来讲，实词丰富性在某种程度上也反映了文本理解的难度。张必隐(1992)在其研究中指出，在阅读过程中，实词往往能够协助读者更快速地理解文本的含义。黄伯荣、廖序

东(黄伯荣and 廖序东, 2017)也在总结以往的实验研究后发现，文章中实词和虚词的数量及其比例对文章的易读性有一定的影响。因此，人类在其回答中使用更多的实词，这一事实验证了4.2一节所得出的结论，即人类文本具有更高的可读性。

词汇密度反映不同词汇在文本中的使用倾向，经常被用在语言类型、语言风格与文体类研究中。例如，在语言类型研究中，研究者发现英语常用介词来表达动态或动作的意思，而汉语使用者倾向于用动词表达，叙事性较强，善用副词等；在语言风格研究中，杨彬(2023)指出副词的使用能够灵活巧妙地调节叙事的节奏，从而建构出形态纷繁的文本。本研究参照现代汉语(黄伯荣and 廖序东, 2017)的词性体系，计算人类回答文本与ChatGPT生成本文中所有词类的密度与句均词性数量。人类回答中，大部分实词的密度与句均词性密度均大于ChatGPT语言，如形容词密度、动词密度、副词密度等；虚词中对叹词与语气词的使用倾向明显，而机器语言中几乎未出现叹词。这一事实表明人类语言更加生动，善于灵活处理变换词性，情感表达丰富，描写能力和表达能力远高于ChatGPT语言。相较于人类语言，ChatGPT生成语言更倾向于使用连词、介词、名词、能愿动词、人称代词、形式动词、助词等，虚词成分较多。这种差异可能源自于汉语训练语料数量不足，也在一定程度上表明，机器语言的连词和人称代词显化现象比人工语言更加突出，且使用助词频率较高，语法标记明显(蒋跃and 董贺, 2015)。

此外，虚词的分布也在一定程度上代表着文本的语体色彩。虚词的使用是写作者无意识的产物，能够反映不同作家的风格风貌(Stamatatos, 2009)。两者差异最大的是连词密度，在不同文体不同领域中，连词的使用及风格色彩也有所偏向。以最具代表性的“和、与、跟、同”为例，“和、与”具有书面语色彩，“跟”具有北方口语色彩，“同”具有南方口语色彩。我们分别计算了人类回答和ChatGPT回答中这四个连词使用频率的平均值，如表5所示，ChatGPT的回答更倾向于使用“和、与”作为句子成分的连接词，相比之下，人类回答中使用“跟、同”的频率更高。数据表明，ChatGPT生成的语言更倾向于书面语的表达方式。进一步支持这一结论的证据可以在第4.1节中各音节词数和占比中找到，具体来说，ChatGPT生成的语言中双音节词的数量和比例最高，这符合现代汉语双音化用词的习惯，而人类语言中单音节词和双音节词的比例则更为接近，更富有口语特色。

	“和”	“与”	“同”	“跟”
人类	4.13	0.92	0.73	0.18
ChatGPT	11.76	1.57	0.07	0.03

表 5: 人类与ChatGPT在“和、与、同、跟”上使用的差异

4.4 句法复杂性

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
并列短语	**并列短语数**	0.813	4.600	名词短语	名词短语数	28.895	54.184
	单句平均并列短语数	0.095	0.331		单句平均名词短语数	3.364	4.020
	句均并列短语数*	0.251	0.729		句均名词短语数	8.702	8.265
形容词短语	形容词修饰语数	1.838	4.114		名词短语平均长度/字token	4.054	4.816
短语结构	句均单句数	3.862	2.511	动词短语	动词短语数	28.727	44.556
	句法树高大于14的句子数量	0.484	0.759		单句平均动词短语数	3.222	2.958
	最大句法树高	13.999	14.919		句均动词短语数(en,de)	8.687	6.566
	平均句法树高	10.899	11.419		**动词短语平均长度/字token**	9.353	14.136
	句法树高大于14的句子占比	0.160	0.129	介词短语	介词短语数	2.197	5.422
	主要动词前平均词数	3.440	4.172		单句平均介词短语数	0.222	0.365
	主要动词前最大词数	7.814	11.432		句均介词短语数	0.663	0.900
依存句法	平均句子依存距离	3.900	3.659		介词短语平均长度/字token	6.274	9.886
	最大句子依存距离	29.452	23.991				

表 6: 句法复杂性维度人类与ChatGPT文本特征均值对比

	问题一：华尔街的课有效果吗？能提高英语水平吗？	**问题二**：北京有像上海七浦路一样的批发市场吗？
ChatGPT	...华尔街英语使用了多种教学方法，包括讲课、角色扮演、小组讨论和个人辅导等。...	...这些市场都提供各种服装产品，包括男装、女装、童装等。...
人类回答	...华尔街的话，其实价格蛮贵的，网上的叫骂声也蛮高的，但是我觉得培训方面还是非常不错的。...	...在南三环木犀园到南四环大红门一带，有很多服装批发大楼，其中的天雅是专门的品牌批发，购物环境不错，...

表 7: 并列短语在ChatGPT与人类回答文本中的使用示例

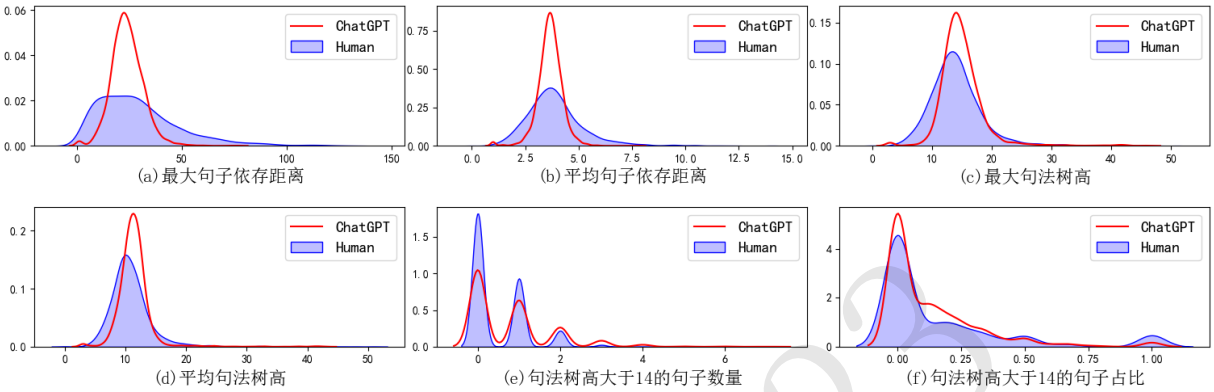


图1: 人类语言与ChatGPT句法树高与依存距离相关特征差异对比

句法复杂性包括名词短语、动词短语、介词短语等七个类别25项测量指标，如表6所示。其中，在两种算法中均贡献度突出的是句均并列短语数，在任一种算法中贡献度突出的有动词短语平均长度(以字token为单位）、并列短语数、单句平均并列短语数、句法树高大于14的句子占比4项，占所有重要特征总数的6.5%。句法复杂度主要通过句子中不同类型短语的数量和短语长度体现。人类回答中动词短语的使用频率较高，而对于名词短语、介词短语以及并列短语，ChatGPT回答的使用率较高。动词性成分具有较为突出的叙事性特征，人类回答中动词短语的大量使用意味着较强的交互性。相对而言，ChatGPT回答倾向于运用修饰性和概念性较强的表达方式。在各种类型的短语中，ChatGPT生成的文本所包含的短语平均长度普遍超过人类水平。

并列短语相关的3个特征均是分类模型中贡献度高的重要特征，经观察发现，ChatGPT在这三种指标上都远高于人类。通过观察数据我们发现，ChatGPT回答中经常使用多个并列成分，这些并列成分处于同一语义场之中，表7提供了两个问答示例，其中“教学方法、服装产品”是上位词，“包括”后是他们各自对应的下位词，诸如此类同一义场中下位词的并列使用，使得要表达的意思更加全面、具体，起到强调的作用(陈绍新, 2017)。

此外，主要动词前的平均词数、句法树高和依存距离（dependency distance）也是衡量句法复杂度的重要指标。主要动词前的平均词数越多，句子的句法树越高，依存距离越长，说明句子的句法关系越丰富，句法复杂度越高(McNamara et al., 2014; 吴思远, 2020)。数据显示，ChatGPT生成文本中主要动词前的平均词数与最大词数都多于人类。随后，我们统计了句法树高大于14的句子数量、占比以及平均句法树高和最大句法树高。数据表明，在ChatGPT生成的文本中，平均每篇文本中有75.9%的句子的句法树高度超过14，相较之下，人类编写的文本中仅有48.4%的句子表现出相同特征。这意味着ChatGPT生成的文本在句法结构复杂性方面往往高于人类撰写的文本。另一方面，我们采用平均句子依存距离、最大句子依存距离2 个常用的特征衡量依存距离。依存距离指句中两个有句法关系的词之间的线性距离，即支配词和被支配词之间的线性距离(Hudson, 1995)。然而，对于这两项特征，人类回答测量值的平均值均高于机器文本，与已有结论相悖。为了进一步探究原因，我们就句法树高与依存距离对3000篇文本绘制了核密度图，如图1所示。

观察图1可知，对于句法树各个相关指标的测量，人类与机器呈现出的取值范围总体相近；

而在依存距离上，人类与机器生成语言中测量值的范围相差较大。也就是说，虽然人类回答的平均句子依存距离与最大句子依存距离均高于机器文本，但人类的回答中最大句子依存距离介于0~100之间，在此区间内分布较为均匀，密集的区间为10~30；而ChatGPT生成的文本大多聚集在25~30之间，且密度高峰远超人类。这一事实揭示出，出于“省力”的考虑，在语言运用中，人类会尽量避免使用可能导致认知成本增加的长距离依存关系(陆前and 刘海涛, 2016)，倾向于使用简单的句子结构和句法成分。但面对难以简短回复的问题，也具备使用句法结构较为复杂的长句的语言能力。

4.5 语篇凝聚力

语篇凝聚力由语篇的衔接程度体现，包括指称、重复、衔接三个类别23项测量指标，如表8所示。其中，指代用代词比例来衡量，重复通过相邻句和全文中实词、名词、动词的重复性来衡量，代词和重复词语的使用可以从语义上让上下文的联系更加紧密。衔接则用各类连词比例来衡量，使用象征不同逻辑关系的关联词是衔接上下文的有效方法。

特征类别	汉语特征	人类	GPT	特征类别	汉语特征	人类	GPT
指称	**第一人称代词比例**	0.012	0.011	重复	**相邻句中动词的重复性**	0.241	0.437
	第三人称代词比例	0.005	0.007		全文中动词的重复性	0.181	0.267
	第二人称代词比例	0.010	0.021	衔接	转折连词比例	0.008	0.007
	人称代词比例	0.031	0.042		因果连词比例	0.011	0.007
	疑问代词比例	0.008	0.005		**选择连词比例**	0.003	0.015
	指示代词比例	0.010	0.009		条件连词比例	0.008	0.002
重复	**全文中词语的重复性***	0.380	0.519		顺承连词比例	0.013	0.003
	全文中实词的重复性*	0.335	0.491		目的连词比例	0.003	0.003
	全文中名词的重复性	0.192	0.368		假设连词比例	0.018	0.009
	相邻句中词语的重复性	0.545	0.875		递进连词比例	0.005	0.006
	相邻句中实词的重复性	0.481	0.831		并列连词比例	0.014	0.012
	相邻句中名词的重复性	0.263	0.631				

表 8: 语篇凝聚力维度人类与ChatGPT文本特征均值对比

语篇凝聚力维度中，在两种算法中均贡献度突出的是全文中词语的重复性、全文中实词的重复性，在任一种算法中贡献度突出的有人称代词比例、选择连词比例、全文中名词的重复性等10项特征，占所有重要特征总数的15.6%。

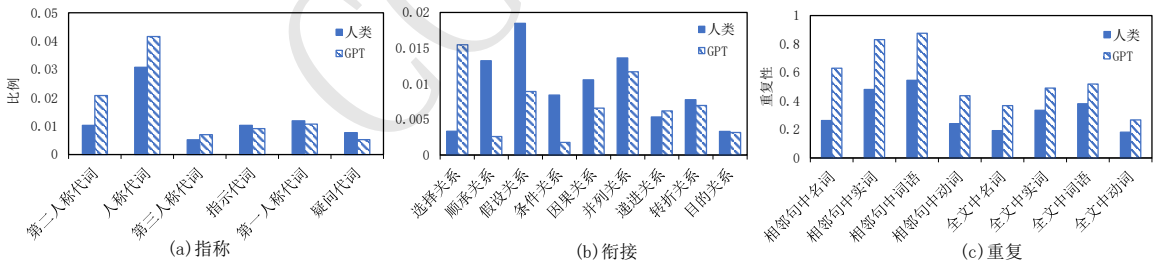


图2: 语篇凝聚力维度特征差异图

1.指称。指称关系指篇章中一个成分与另一成分之间所具有的相互解释的关系(洪秋月and 熊智伟, 2023)，指称衔接多用代词来体现，包括人称代词、指示代词、疑问代词。彭宣维(2005)在探讨代词的语篇语法属性时提出，代词在文章中的出现主要是代替别的成分发挥相应功能，从而使语句及整篇文章有很好的衔接关系。

如下图2（a）所示，在人称代词、指示代词、疑问代词这三项指标中，ChatGPT语言使用人称代词的比例要多于人类回答，且使用第二第三人称代词的比例相对较多，ChatGPT语言用多尊称“您”，而人类语言倾向于使用第一人称代词和指示代词、疑问代词。已有研究表示，第一人称代词、指示代词这两种形式在非正式语体中占主体部分，第一人称代词的使用显示了

研究结果的主观性，多是引导读者赞同所述观点和研究结果，传递出作者构建主体身份的特性(Seidel, 1975)。

指示代词和疑问代词的使用，显示出作者根据语境回指上文出现的事物，既有利于文本衔接上下连贯，又可强调观点表达(贾宇丹, 2022)。由此可见，ChatGPT语言多是以较为客观的态度进行分析并给出建议，遵循会话的礼貌原则，较少发表主观性强的意见，而人类回答拥有话语权，善于表达自己的观点和看法。

2.衔接。衔接以语篇序列为前提，在建立句子之间的衔接与联系方面占有非常重要的作用(Cain and Nash, 2011)，韩礼德和哈桑(Halliday and Hasan, 1976)提出，关联词不是直接通过它们自己来衔接，而是通过它们的特殊意义来间接衔接。本研究通过关联词的分布考察了选择关系、顺承关系、假设关系等9种衔接关系，以此分析ChatGPT语言与人类语言的文本衔接特征上的差异。观察数据可以发现无论是人类回答还是ChatGPT回答，所运用的关联词类别都比较全面，不存在某种衔接关系缺失的情况。但整体来看，ChatGPT语言使用上述9种衔接关系的比例低于人类语言，表明人类在回答问题时所运用的语言具有衔接显化的特点(邢诗吟, 2022)。

图2（b）对9种衔接关系的使用比例进行了差异排序，可以看到ChatGPT和人类回答中差距较大的是假设关系、顺承关系、条件关系、选择关系。ChatGPT回答中存在选择关系的比例高于人类语言，顺承关系、条件关系、选择关系的比例低于人类语言。ChatGPT最常使用陈述式关联词“或...或...”、“或者...或者...”，而人类回答中使用最多的是疑问式关联词“还是”。在人类语言中，使用比例最高的是假设关系，多是表示和结果一致的假设，如使用表示一致关系的关联词“就”。而ChatGPT回答在表达假设关系时使用最多的是表示相背关系（假设和结果不一致）的“...，也...”、“...，还...”。(黄伯荣and 廖序东, 2017)。

3.重复。重复指在同一语篇中反复出现具有相同含义和形式的词，对实现前后文的连贯有显著的作用。如图2（c）所示，ChatGPT语言中相邻句和全文中实词、名词、动词的重复性都高于人类语言，说明ChatGPT的篇章衔接紧密，文本的表达紧紧围绕同一主题，而人类文本的篇章重复性较低，词干、论元重叠度低(何清强et al., 2019)，行文发散。

5 总结与讨论

本研究旨在考察人类与ChatGPT回答文本中语言特征的差异，以及基于特征结合机器学习方法得到的ChatGPT探测器预测的能力。结果表明，第一，在描述性特征、字词常用度、字词多样性、句法复杂性、语篇凝聚力五个维度中，对模型分类贡献度较高的特征集中在描述性统计、字词常用度、字词多样性三个维度。第二，SVM与随机森林都表现出较好的性能，最优模型达到了97.27%的准确率与97.33%的F1值。

对于五个维度下两种回答文本的语言差异，本文得出以下结论：

1.ChatGPT生成语言倾向于使用大词，往往分段进行阐述生成长文本，人类的回答更加简短，自然段少。在一定程度上，ChatGPT生成的文本难度与质量高于人类的回答。人类具有生成更为复杂和详尽句子的能力与使用较多的笔画数汉字的能力，长短句的使用更加灵活多变。

2.人类的用词偏好有助于丰富语言表达（词汇多样性高）并降低文本理解难度（高频词和实词使用频率高），富有口语色彩。ChatGPT生成的语言中语法标记明显，倾向于书面语的表达方式。在同样篇幅的文本中，人类提供的信息量更大。整体来看，ChatGPT所体现出的语言特征更具英文偏好，比如和英文一样，ChatGPT倾向于使用介词、助词等修饰性较强的成分，这可能与训练语料大多是英语有关。

3.ChatGPT倾向于运用修饰性和概念性较强的表达方式，在句法结构复杂性方面往往高于人类撰写的文本。人类回答具有较强的交互性，倾向于使用简单的句子结构和句法成分。但面对难以简短回复的问题，也具备使用句法结构较为复杂的长句的语言能力。

4.ChatGPT在指称、重复上优于要优于人类文本，词语的重复性较多，语义重叠度高，生成的回答围绕同一主题展开。人类思维活跃，容易给出发散式的回答。

根据本文的研究结论可推断出，ChatGPT在中文使用上的表现与人类具有较大差异。为了使人工智能生成语言更加真实，高质量的中文数据集建设与大语言模型研究迫在眉睫。我们的研究详细分析了ChatGPT生成语言与人类语言在多个维度上的差异，但也存在一些局限。首先，本研究中仅选择了ChatGPT生成语言和人类语言在开放域的问答语料，样本量相对较小，未来的研究中可以使用更多包含不同语域与不同语体的数据集。其次，本研究中使

用GPT-3.5作为底层模型的ChatGPT，若使用更先进的模型（如GPT-4），这些语言特征的表现可能会有所不同。

参考文献

D. A. Balota and J. I. Chumbley. 1984. Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology Human Perception & Performance*, 10(3):340–357.

K. Cain and H. M. Nash. 2011. The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology*, 103(2):429.

Scott A Crossley and Danielle S McNamara. 2010. Interlanguage talk: What can breadth of knowledge features tell us about input and output differences? In *23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23*, pages 229–234.

Yue Cui, Junhui Zhu, Liner Yang, Xuezhi Fang, Xiaobin Chen, Yujie Wang, and Erhong Yang. 2022. Ctap for chinese: a linguistic complexity feature automatic calculation platform. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5525–5538.

Kyle B Dempsey, Philip M McCarthy, and Danielle S McNamara. 2007. Using phrasal verbs as an index to distinguish text genres. In *FLAIRS Conference*, pages 217–222.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Annual Meeting of the Association for Computational Linguistics*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

K. Haberlandt and A.C. Graesser. 1985. Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, 114:357–374.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Routledge, London.

Richard A. Hudson. 1995. English word grammar.

Victoria Johansson. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53:61–79.

Max M Louwerse, Philip M McCarthy, Danielle S McNamara, and Arthur C Graesser. 2004. Variation in language and cohesion across written and spoken registers. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human – differentiation analysis of scientific content generation.

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

D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, Cambridge.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Maryam Nasseri and Paul Thompson. 2021. Lexical density and diversity in dissertation abstracts: Revisiting english l1 vs. l2 text differences. *Assessing Writing*, 47:100511.

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

Jiashu Pu, Ziyi Huang, Yadong Xi, Guandan Chen, Weijie Chen, and Rongsheng Zhang. 2022. Unraveling the mystery of artifacts in machine generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6889–6898.

G. Seidel. 1975. Ambiguity in political discourse. In M. Bloch, editor, *Political Language and Oratory in Traditional Society*, pages 205–228. Academic Press, London.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

何清强, 王文斌, and 吕煜芳. 2019. 汉语叙述体篇内句的特点及其二语习得研究——基于汉英篇章结构的对比分析. 语言教学与研究, (06):1–11.

吴思远. 2020. 基于多层面语言特征的汉语文本可读性自动评估研究. 硕士学位论文, 北京语言大学.

张倩倩. 2022. 基于小学语文教材的文本易读性公式研究. Ph.D. thesis, 江南大学.

张必隐. 1992. 阅读心理学. 北京师范大学出版社, 北京.

彭宣维. 2005. 代词的语篇语法属性、范围及其语义功能分类. 语言教学与研究, (01):56–65.

李绍山. 2000. 易读性研究概述. 解放军外国语学院学报, 2000(04):1–5.

杨彬. 2023. 篇章动态视角下副词性成分的叙事价值分析. 当代修辞学, 2023(01):42–50.

洪秋月and 熊智伟. 2023. 语言经济原则下热搜词条的语篇衔接研究. 今古文创, (05):129–132.

熊兵. 2016. 基于语料库的旅游文本英译文词汇特征及翻译研究. 华中师范大学学报*(*人文社会科学版*)*, 55(05):94–103.

蒋跃and 董贺. 2015. 计量特征在人机译文语言风格对比中的应用. 语言教育, 3(03):69–74+81.

蔡建永. 2020. 汉语二语文本可读性研究. Ph.D. thesis, 北京语言大学, 北京.

贾宇丹. 2022. 中国外应专业研究生学术语篇非正式语体特征研究. 名家名作, (21):85–87.

邢诗吟. 2022. 基于语料库的初中英语记叙文写作语言特征研究. Master's thesis, 集美大学.

陆前and 刘海涛. 2016. 依存距离分布有规律吗? 浙江大学学报*(*人文社会科学版*)*, 46(4):63–76.

陈绍新. 2017. 元功能理论视角下的英语商务合同汉译研究. 湖南第一师范学院学报, 17(6):92–97.

黄伯荣and 廖序东. 2017. 现代汉语. 高等教育出版社.

古汉语通假字资源库的构建及应用研究

王兆基♠ 张诗睿♣ 张学涛♡ 胡韧奋♡,✉*
北京师范大学国际中文教育学院
♠zhaoji.wang@mail.bnu.edu.cn ♣1169882881@qq.com
♡{11112011118, irishu}@bnu.edu.cn

摘要

古籍文本中的文字通假现象较为常见，这不仅为人理解文意造成了困难，也是古汉语信息处理面临的一项重要挑战。为了服务于通假字的人工判别和机器处理，本文构建并开源了一个多维度的通假字资源库，包括语料库、知识库和评测数据集三个子库。其中，语料库收录11000余条包含通假现象详细标注的语料；知识库以汉字为节点，通假和形声关系为边，从字音、字形、字义多个角度对通假字与正字的属性进行加工，共包含4185个字节点和8350对关联信息；评测数据集面向古汉语信息处理需求，支持通假字检测和正字识别两个子任务的评测，收录评测数据19678条。在此基础上，本文搭建了通假字自动识别的系列基线模型，并结合试验结果分析了影响通假字自动识别的因素与改进方法。进一步地，本文探讨了该资源库在古籍整理、人文研究和文言文教学中的应用。

关键词： 古代汉语 ； 资源库 ； 通假字 ； 自动识别

The Construction and Application of an Ancient Chinese Language Resource on Tongjiazi

Zhaoji Wang Shirui Zhang Xuetao Zhang Renfen Hu✉
School of International Chinese Language Education, Beijing Normal University

Abstract

In ancient Chinese texts, it is common to use characters with the same sound or similar sounds instead of the original characters, that is, to use Tongjiazi. This not only creates difficulties for people to understand the meaning of the text, but also an important challenge for ancient Chinese information processing. In order to assist the manual analysis and machine processing of Tongjiazi, this paper builds a multi-dimensional language database, including three sub-databases, i.e. corpus, knowledge base and evaluation data set. Among them, the corpus contains more than 11,000 sentences with detailed annotations of Tongjia usages. The knowledge base is presented in graph data with 4185 characters as the nodes and 8350 relations between them as the edges. The attributes of the nodes and the edges are labeled from the perspectives of pronunciation, glyph and meaning. The evaluation data set is designed for automatic recognition of Tongjia usages, including training and testing data for two subtasks: Tongjiazi detection and the recognition of the original characters. Now the evaluation data covers 19678 entries. On this basis, this paper builds a series of baseline models for the automatic recognition of Tongjia usages, and analyzes the factors affecting the recognition results and the improvement methods. Further, this paper discusses the application of these resources in different fields, e.g. the collation of ancient books, humanities research and classical Chinese learning and teaching.

Keywords: ancient chinese , resource , Tongjiazi , automatic recognition

*Corresponding author.

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

1 引言

与现代汉语及其他语种不同的是，古籍文本中的文字通假较为常见，这为准确理解文意造成了困难。具体来说，通假指的是古人本有其字而不用，反而借用一个音同或音近字的现象，其中，被借用的字称作通假字，被代替的字称为正字或本字 (孔德明, 1993; 王宁, 2012)。例如，在“庄公寤生，惊姜氏。”（出自《左传》）中，“寤”为通假字，所通正字为“牾”，“牾生”即逆生，表示难产。

通假现象不仅常见于传世古籍，在出土文献中也有较高频率。据钱玄 (1980)统计，现存《老子》(据唐傅奕校《道德经古本篇》) 约5500余字，其中用通假字30多个，而马王堆帛书《老子》(乙本) 使用通假字320个，占全书的6%。整理古籍时，通假字识别对于准确理解文意来说十分重要，如王引之在《经义述闻·经文假借》中所述：“学者改本字读之，则怡然理顺；依借字解之，则以文害辞。”除了专业学者整理古籍时需要释读通假字，在中学文言文教学中，通假字也是一项重点和难点，掌握文言文常见通假字的用法是文言文阅读的基本功 (由明智, 2013)。值得一提的是，对于汉语史研究来说，通假字与被通假字之间的音同或音近关系可以为汉语古音和语音史研究提供宝贵的参考资料 (张儒, 1988; 党怀兴, 1998)；同时，字与字之间的通假关系亦有助于厘清词汇形式和词义演变的脉络，从而服务于词汇发展变化和词汇语义研究 (孙建伟, 2015)。可以说，无论是服务于通假字识别，还是汉语史研究，高质量的通假字资源库都必不可少。柳建钰和周晓文 (2017)从辅助校勘需求出发提出了构建通假字资源库的设想，拟基于各类通假字字典搜集整理通假字表，预计收录字头6000个左右，涵盖传世文献和出土文献的通假字。然而，目前除了字典和辞书以外，通假字相关数据资源的建设仍十分罕见。

近年来，伴随古籍整理利用的转型升级，古汉语信息处理资源建设和算法研究受到了越来越多的关注，涉及任务包括句读标点、分词和词性标注、词义消歧、命名实体识别、自动校勘、文本生成等 (邓三鸿等, 2021; 胡韧奋等, 2021; Liu 等, 2019; 苏祺等, 2021; 舒蕾等, 2022; Yang 等, 2021; Yi 等, 2018)。需要指出的是，在古汉语信息处理研究中，通假字识别是词义注释、文本校勘和文白翻译等技术的基础，目前学界尚无针对通假字自动识别的研究。当下，即使ChatGPT、GPT4等大模型具备极强的自然语言理解能力，其处理包含通假字的文本时仍会“以文害辞”，一例如表 1 所示。因此，对于古汉语信息处理来说，通假字资源库的建设及自动识别技术研究具有其现实价值和迫切性。

领域知识	输入	输出
无领域知识	请将下面的文言文翻译为现代汉语: 廉將軍雖老，尚善飯，然與臣坐，頃之，三遺矢矣。	尽管廉将军年事已高，但仍然精于弓箭。然而在与臣下一起坐着的过程中，不久，他三次失手放箭。
有领域知识	请将下面的文言文翻译为现代汉语: 廉將軍雖老，尚善飯，然與臣坐，頃之，三遺矢矣。 下面是可能的通假字用法，可结合语义判断是否使用: “矢”共有3种通假用法：①“矢”通“誓”。发誓永远要(做某事)。②“矢”通“屎”。粪便。③“矢”通“弛”。毁坏。	廉将军虽然年纪大了，但仍然能吃得下饭，然后与臣子坐下来。不久之后，他已经三次去上厕所了。 在这里，通假字“矢”应该使用第二种用法，表示“屎”。

表 1. 基于GPT-4的文白翻译实例

为了更好地辅助通假字的人工判别和机器处理，本文首先构建了一个涵盖标注语料库、知识库和评测数据集的通假字资源库。其中，标注语料库收录了11000余条包含通假字用例标注的语料，知识库以汉字为节点，通假关系为边，从字音、字形、字义多个角度对通假字与正字的属性进行加工，共包含4185个字节点、7700对通假字关联信息、650对通假字之间的形声关系信息；评测数据集分为基础版和拓展版，支持通假字检测和正字识别两个子任务的评测，收录评

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

测数据19678条。在此基础上，本文搭建基线模型开展了通假字检测和正字识别实验，并探讨了资源库在古籍整理、人文研究和文言文教学中的应用。

2 通假字资源库构建

为了让资源库更好地服务于与通假字有关的文史研究和自动识别算法研发，我们设计并构建了三个开源资源库，均以JSON格式发布[1]，包括：通假字标注语料库、通假字知识库与通假字识别评测集。

2.1 通假字标注语料库

目前，学界尚无专门标注通假字的文言文语料库，包含通假字的句篇信息主要见于各类辞书，其中也包括专门的通假字字典，如高亨《会典》收录了传世文献材料中的通假字用法，《简帛古书通假字大系》侧重于依据战国秦汉出土简帛文献。考虑到与通假字相关的辞书存在应用场景区别，为兼顾古汉语信息处理、文史研究与文言文教学的一般性需求，本文选择以《汉语大词典》为数据源，构建通假字标注语料库。该库可为通假字相关研究和应用提供高质量的基础性数据，亦可结合具体需求进行筛选、优化和扩充。

《汉语大词典》所收条目分为单字条目与多字条目。多字条目按“以字带词”的原则,列于单字条目之下。一个单字有两个以上字头的，在字头旁以阿拉伯数字标注序号。字头下依次标注现代音与古音，其中，现代音用汉语拼音字母标注，古音用反切标注。释义时，通假义用“通‘×’”和“‘×’的被通假字”表示。据此，我们以《汉语大词典》的标注为准，采集通假现象涉及的释义及例句，例如，在《汉语大词典》中，字头“耗3”的内容如表 2 所示，该字可通“眊1”，表示“昏乱不明”，词典收录了来自《荀子·修身》与《汉书·景帝纪》的两则包含通假字的例句。

耗3〔mào ㄇㄠˋ〕 〔《字彙》莫報切〕 通“眊1”。 昏乱不明。《荀子·修身》：“少而理曰治，多而亂曰耗。”《汉书·景帝纪》：“其令二千石各脩其職，不事官職耗亂者，丞相以聞，請其罪。”颜师古注：“耗，不明也，讀與眊同。”

表 2. 通假字标注语料库语料原文示例

经自动提取和人工校对，我们从《汉语大词典》中采集了较大规模的通假字属性及用例数据，在此基础上构建了高质量的通假字标注语料库，共包含11000余句繁体中文语料，覆盖2479个通假字。其中，用例最多的为“辨”，存在通“辯”、“變”、“班”、“般”等字的126条用例，同时，也有不少通假字的例句数量较少，例如，有833个通假字仅包含1条用例语料。如表 3 所示，语料库中的每一条语料包含11个属性：语料ID、语料文本、标注位置、通假字字头、正字[2]字头、出处、时代、释义、拼音、注音和古音。

语料ID	1
语料文本	少而理曰治，多而亂曰耗。
标注位置	10
通假字字头	耗3
正字字头	眊1
出处	《荀子·修身》
时代	战国
释义	昏乱不明。
拼音	mào
注音	ㄇㄠˋ
古音	《字彙》莫報切

表 3. 通假字标注语料库语料示例

[1]数据下载地址：https://github.com/frederick-wang/tongjiazi-resources
[2]资源库中采用“正字”标识被通假字。

2.2 通假字知识库

本文采用图数据结构设计通假字知识库，其中，汉字为节点(node)，通假关系和形声关系为边(edge)。在汉语史研究中，对通假这种字用现象的分析往往会从字音、字形、字义等多个角度展开，因此，我们在构建知识库时也分别针对音、形、义进行了属性标注。在**字音**方面，我们在通假关系边中标记了与其对应的注音和古音属性，同时参考胡韧奋等(2013)加工的形声字数据添加了字和字之间的形声关系边。在**字形**方面，我们在字节点上标记了部首、部件和结构信息，其中汉字部件信息参考了Yan et al. (2013)构建的数据集。在**字义**方面，字节点之间的通假关系考虑了义项的差别，即A字通B字时根据语境差别有多种含义，因此一对字节点之间允许有多个通假关系边相连。通假字知识库的详细规范参见附录 A。

为了更好地服务于汉语史及通假字自动识别研究，本文在《汉语大词典》的基础上，进一步从康熙字典[3]、汉典[4]、国学大师网汉语字典[5]中采集了与通假用法相关的多源异构数据，共计315万字，融合构建了通假字知识库。在融合数据时，对字而言，以字形为标准与其他来源的数据合并。对通假关系而言，以通假字、正字、释义为标准与其他来源的数据合并。最终，通假字知识库收录了4185个字节点、7700对通假关系、650对形声关系。

图 1 以“辟”为例，展示了知识库中节点和边的属性。其中，字节点属性标注在蓝框内；红色的有向边表示通假关系，通假关系的详细属性参见红框，与通假关系相关联的语料以紫框标注；绿色的有向边表示形声关系，对应的绿框为形声关系的具体属性。由图中内容可见，“辟”与“譬”之间存在3条通假关系连边，对应三种释义，同时，二者之间还包括一条形声关系连边，标识“辟”是“譬”的声符。

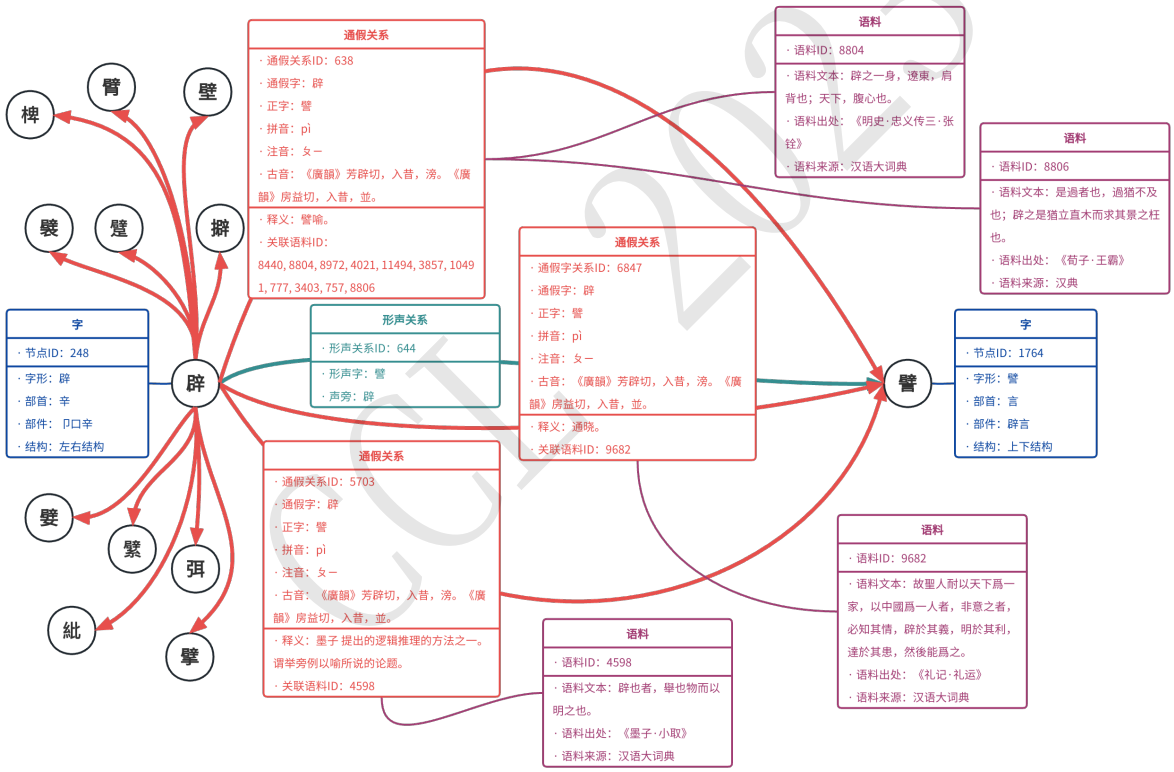


图 1. 通假字知识库示例“辟”：限于空间，图中仅展示了“辟”与“譬”之间的详细关系及部分关联语料，与“辟”有通假或形声关系的其他字仅在图中呈现节点，如“擗”、“繁”等。

通假字知识库对通假字与正字之间的关系进行了详细加工，相关信息可以从数据和特征两个维度为通假字的自动识别提供支持（参见本文第 3 节）。同时，通假字与正字之间的音、形、义关联也可为汉语史领域的相关研究提供参考。首先，它可以帮助相关研究者更高效地开展传统研究，常见的应用场景包括：

[3]使用“汉典”版的《康熙字典》数据。
[4]汉典：https://www.zdic.net/
[5]国学大师网汉语字典：http://www.guoxuedashi.net/zidian/

1. **字词考证**：通假字知识库可以帮助我们迅速辨别出通假字，识别出这是常用的通假还是在特定语境中出现的借字。例如，在图 1 中，“辟”字与“譬”字之间的通假关系，可以帮助我们了解到“辟”字在某些语境下可以作为“譬”的通假字使用。

2. **词汇语义研究**：通假字知识库可以帮助我们将某些和本义无关的假借义从词义引申中剔除，例如图 1 中“辟”通“譬”对应的三种释义。此外，通假字关联网络还能帮助系联同义、近义或词义相关的词，从而辅助词汇语义研究。

3. **形声字研究**：知识库中的字节点之间除了通假关系边，还有形声关系边，例如，在图 1 中，“譬”字是一个形声字，其声旁为“辟”。通假与形声关联数据可以辅助我们进一步研究形声字及语音的发展规律。

值得一提的是，通假字知识库能够提供传统辞书**无法呈现**的大规模通假字关联网络信息，这也为汉语史研究提供了新的视角，潜在的应用场景包括：

1. **量化通假强度**：在传统研究中，字与字之间的通假关系仅分为“有”和“无”，但这种粗粒度的判断方式并不精确。事实上，有些通假关系应用广泛，而有些仅为辞书中的孤例。基于通假字图知识库，我们可以通过字与字之间不同义项的通假关系数量（边数）以及相关联的语料数量来量化“通假关系的强度”，为后续研究提供更多可能性。

2. **利用子图探究通假规律**：传统研究范式下，由于人的时间和精力有限，研究通常仅针对一个字的通假关系及其相关的几个被通假字进行，相当于仅能研究图中的几个节点及其边。借助图数据库，我们可以根据分割条件迅速将所有数据划分为多个子图，研究子图中所有通假字节点与通假关系边的内在规律，并探讨子图间的联系。这将有助于我们发现更多的通假规律。例如，研究一个通假字的所有通假变化轨迹，实际上就是寻找该节点所在的子图并获得一个子图的生成树。

3. **辅助古汉语语音演变研究**：通假关系存在的前提是字之间的音同或音近，而不少汉字的读音在历史上经历了较大变化。利用通假字图知识库，我们可以为相关语音研究提供支持。例如，我们可以根据通假关系边关联语料的“出处”数据，获取不同时期的字与字之间的通假关系并生成关联子图，进而量化估计在某一特定时代，两个字的发音可能相同；而在另一时代，这两个字的发音可能不同。如此一来，我们便能从历时角度利用图知识库为语音演变研究提供支持。

2.3 通假字识别评测集

为了推动通假字自动识别算法研究，我们基于高质量的通假字标注语料库构建了通假字识别评测集。评测集分为两个子任务：通假字检测与正字识别。为了更好地评估模型的泛化能力，每个子任务均分为基础版与拓展版，其中，基础版任务的训练集与测试集覆盖的目标字范围一致，而拓展版任务的测试集中则包含训练集未出现的通假用法，其自动探测和识别的难度更高。接下来，本节将介绍两个评测子任务的形式及评测数据集的构建方法。

2.3.1 评测任务设计

表 4 给出了两个子任务的示例，其中，通假字检测任务旨在识别古汉语文本中的通假字位置，即给定一段输入文本，需输出文本中所有通假字的位置（从0开始计数）。如果该文本中没有通假字，则输出`[]`。计算精确率和召回率时，使用（句子，位置）二元组作为计算单位。

正字识别任务的目标是识别出古汉语文本中通假字所对应的正字，输入一段文本和通假字位置，需输出该位置的通假字所对应的正字。计算精确率和召回率时，使用（句子，位置，正字）三元组作为计算单位。

子任务	输入	输出
通假字检测	{"sentence": "北庭使劉涣躬行勃逆，委公斬之。"}	[7]
正字识别	{"sentence": "不韋毁身燋慮，出于百死。", "pos": 4}	"焦"

表 4. 评测任务示例

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

2.3.2 评测数据集的构建

考虑到通假字标注语料库主要收录目标字作为通假字使用的数据，为了评测模型的判别能力，兼使其适应真实应用情境，我们从词典中的其他义项例句中补充了目标字非通假用法的数据，构成正负例，如下面两例所示。

例1. **惠**心燭千仞，雄風扇八區。（正例，通“慧”，表“明慧”含义。）

例2. 必也君亂之，君終之，君之**惠**也。（负例，表“恩惠”含义）

考虑到通假字的常用度存在差异，且有必要对模型的泛化能力进行评估，我们构建了基础版和拓展版两类评测数据集。基础版评测旨在识别常用通假用法，其中，每个通假字收录至少10条正例，最多不超过20条[6]。同时，尽量补充与正例数量相等的负例，即目标字非通假用法的例句。进一步地，将每个通假字的正例和负例均按照8:2的比例拆分，分别划入训练集和测试集，从而保证训练集与测试集的数据分布相同。最终，基础版数据集覆盖了279个常见通假字，包含7962条语料，其中，训练集6190条，测试集1772条。

针对用例并不充足的通假字，我们又额外构建了拓展版评测数据。拓展版训练集与基础版训练集保持一致，拓展版测试集则在基础版测试集的基础上，额外补充了2200个通假字的正例和负例数据，其中，每个通假字的正例少于10句，负例与其数量相当，共计增补了11716条语料，因此拓展版测试集共收录13488条语料。由于拓展版测试集中收录了大量训练集未覆盖的通假用法，这便要求模型结合语境识别出训练时未见过的通假字，无疑挑战性极高，也更加接近真实的应用情境。

3 通假字自动识别评测

基于上节介绍的评测任务和数据，我们就通假字的自动识别开展了初步探索，以期为未来学界的相关研究提供基线（Baseline）结果[7]。接下来，将首先介绍本文引入的基线方法，然后，将分别报告通假字检测（基础版）、通假字检测（拓展版）、正字识别（基础版）和正字识别（拓展版）任务的评测结果，并展开分析和讨论。

3.1 实验方法

为了服务于通假字探测和正字识别，我们首先参考文本纠错的实验设定构建了一个“通假字-正字”混淆集。混淆集数据采集自通假字知识库和评测训练集中的“通假字-正字”字对。由于测试集中的语料来自《汉语大词典》，为避免测试数据的字对信息泄露，我们在使用通假字知识库中的字对数据时，排除了来自《汉语大词典》的数据。

3.1.1 通假字检测任务

在通假字检测任务中，我们采用了四类基线模型：Ngram语言模型、GPT2语言模型、BERT MLM语言模型和基于BERT的通假字检测微调模型。

对于Ngram语言模型来说，我们利用KenLM计算句子的概率得分p，并代入公式$p^{-\frac{1}{N}}$得到句子困惑度P。困惑度越低，句子表达越合理。对于句中每个位置的汉字，如果它位于混淆集中，将分别计算用它的混淆字替换原字后的句子困惑度，并与句子的初始困惑度进行比较。如果替换后的句子困惑度比原句低，则将该字所在的位置标记为通假字位置。实验中，我们分别基于殆知阁古代文献藏书2.0版和文渊阁版四库全书（繁体）语料库训练bigram和trigram模型，得到了四组结果，后文分别用DaizhigeBigram、DaizhigeTrigram、SikuBigram和SikuTrigram指代。

基于GPT2语言模型的通假字检测方法与Ngram模型类似，即利用困惑度和混淆集信息标记通假字位置。实验采用了Huggingface中两个开源的古汉语GPT2模型，分别基于殆知阁和四库全书语料训练，后文用DaizhigeGPT2[8]和SikuGPT2[9]指代。

利用Bert MLM语言模型进行实验时，我们依次判断句中的每一个字是否位于混淆集中。若在，则将该位置用[MASK]遮罩，并输出Mask LM的预测结果，从而得以比较原字与混淆集中对应字的预测概率，如果存在混淆字的预测概率高于原字，则将该字所处位置标记为通假字位置。

[6]为确保数据分布的均衡性，如果通假字在标注语料库中的例句大于20条，则随机抽取20句。

[7]https://github.com/frederick-wang/tongjiazi-evaluation

[8]https://huggingface.co/uer/gpt2-chinese-ancient

[9]https://huggingface.co/JeffreyLau/SikuGPT2

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

除了BERT MLM模型外，我们还引入了BERT微调方法。具体来说，通假字检测可建模为token序列标注任务，句中非通假字对应的标签为0，通假字的标签为1。微调阶段，采用BERT+全连接层的结构进行token标签学习[10]。推理阶段，如果句中某字既被模型标记为1，又是混淆集中收录的通假字，则将该字所在的位置标记为通假字位置。同时，我们也引入了一个无需混淆集的版本，只要该字被模型标记为1，就将对应位置标记为通假字位置。实验中，为了与前面三种方法对应，我们采用了基于殆知阁语料库训练的古汉语BERT模型和Huggingface上开源的SikuBERT模型[11]，经微调，得到了TongjiaziDetectionDaizhigeBert模型与TongjiaziDetectionSikuBert模型。

3.1.2 正字识别任务

与检测任务类似，正字识别任务也可基于Ngram语言模型、GPT2语言模型、BERT MLM语言模型和BERT微调模型实现。

对于Ngram、GPT2模型来说，我们将判断句中给定位置的字符是否在混淆集中，如果不在，将该字符直接作为识别的正字；如果在，则依次计算混淆字替换该字符后的句子困惑度，并与句子的初始困惑度进行比较，取句子困惑度最小的字作为识别的正字。BERT MLM模型的识别方法与之类似，如果给定位置的字符不在混淆集中，则将该字符作为识别的正字；如果在，则将该字符用[MASK]遮罩，利用Mask LM获取原字符与所有混淆字的预测概率，取预测概率最大的字作为识别的正字。

关于BERT微调方法，我们借鉴Mask LM任务的形式，要求模型预测出句中通假字所对应的正字，其余位置的字符不参与训练[12]。经过微调，模型加强了正字和上下文语境信息之间的关联，在推理阶段，采用与上述BERT MLM模型一致的方法获取正字识别结果。后文用ZhengziRecognition指代经微调训练的识别模型。

3.2 实验结果

3.2.1 通假字检测任务

表 5 和表 6 分别列出了通假字检测任务在基础版和拓展版数据集上的评测结果。在基础版测试集上，模型检测最优F1值达到66.94%，拓展版测试集的最优检测F1值为21.63%，可见通假字检测是一个极有挑战性的任务，在处理模型训练未见过的通假用法时尤为困难。通过对比不同模型，我们发现以下几点要素或对模型的检测性能产生影响。

序号	模型	精确率	召回率	F1
1	DaizhigeBigram	7.55%	22.26%	11.27%
2	SikuBigram	9.21%	18.52%	12.30%
3	DaizhigeTrigram	7.56%	18.62%	10.75%
4	SikuTrigram	10.00%	11.80%	10.83%
5	DaizhigeGPT2	8.59%	22.74%	12.47%
6	SikuGPT2	10.84%	17.47%	13.38%
7	DaizhigeBert	20.24%	55.28%	29.63%
8	SikuBert	29.09%	57.49%	38.63%
9	TongjiaziDetectionDaizhigeBert	65.02%	64.40%	64.71%
10	TongjiaziDetectionSikuBert	64.25%	69.87%	66.94%
11	TongjiaziDetectionDaizhigeBert（无混淆集）	62.10%	64.78%	63.41%
12	TongjiaziDetectionSikuBert（无混淆集）	61.96%	70.35%	65.89%

表 5. 通假字检测任务（基础版）实验结果

(1) 模型结构与复杂度

[10]训练模型时，Torch、Numpy和random模块的随机数种子为42，Batch大小设为8，Epoch数设为5，采用AdamW优化器，学习率设为5×10^{-5}。按照9：1的比例将训练数据划分为训练集与验证集，*RandomState*同样设为42。

[11]https://huggingface.co/SIKU-BERT/sikubert

[12]在微调模型时，Torch、Numpy和random模块的随机数种子、Batch大小、Epoch数、优化器，学习率、训练数据划分方法均与前文的TongjiaziDetectionBert模型相同。

序号	模型	精确率	召回率	F1
1	DaizhigeBigram	4.63%	10.81%	6.48%
2	SikuBigram	5.62%	8.74%	6.84%
3	DaizhigeTrigram	4.50%	8.64%	5.92%
4	SikuTrigram	6.21%	5.35%	5.75%
5	DaizhigeGPT2	5.25%	11.23%	7.16%
6	SikuGPT2	7.39%	8.60%	7.95%
7	DaizhigeBert	9.80%	22.73%	13.69%
8	SikuBert	15.54%	23.50%	18.71%
9	TongjiaziDetectionDaizhigeBert	31.54%	11.68%	17.05%
10	TongjiaziDetectionSikuBert	27.78%	12.12%	16.88%
11	TongjiaziDetectionDaizhigeBert（无混淆集）	32.94%	16.10%	21.63%
12	TongjiaziDetectionSikuBert（无混淆集）	29.48%	16.53%	21.18%

表 6. 通假字检测任务（拓展版）实验结果

实验结果显示，预训练语言模型具有较好的语境信息编码能力，在一定程度上能够辅助探测通假字，其中，基于BERT模型的方法效果普遍最优，GPT2模型次之，Ngram模型最弱。推测一方面与模型的复杂程度有关，Ngram模型最为简单，对上下文信息的捕捉能力最弱，另一方面也和模型结构有关，与GPT2单向自回归训练机制不同，BERT在预训练阶段的双向编码机制使其更擅长利用上下文语境信息进行字符判断。

(2) 预训练数据

在不同类型的模型上，基于文渊阁版繁体四库全书数据训练的模型表现普遍优于基于殆知阁数据训练的模型。殆知阁语料库规模更大，繁简混合，而文渊阁版四库全书（繁体）数据规模偏小，全部为繁体。考虑到我们的评测数据均为繁体中文，这与四库版预训练模型更为匹配。

(3) 微调机制的引入

在基础版评测数据集上，无论是DaizhigeBert还是SikuBert，微调后精确率和召回率均有显著提升，相较之下，精确率提升幅度更为突出，这意味着微调前，模型倾向于将非通假用法识别为通假字，而经过训练数据上的微调，模型熟悉了常见通假字用法，探测精确率得到显著改善。

在拓展版评测数据集上，微调同样提升了BERT模型的精确率，但也使其召回率出现了明显下降，推测这主要是由于拓展版测试集中收录了大量训练集未覆盖的通假用法，在训练集上微调使得模型聚焦于用例较多的常见通假字，对训练中未见过的通假用法不再关注，从而降低了识别的召回率。

(4) 混淆集的使用

在通假字检测任务（基础）中，使用混淆集的TongjiaziDetectionBert精确率略高于无混淆集版，召回率二者几乎一致。但是，在拓展版任务中，无混淆集的TongjiaziDetectionBert不论是精确率还是召回率都优于带混淆集版，这主要是由于拓展版数据集中存在不少混淆集未覆盖的通假用法，使用混淆集反而在一定程度上限制了模型的识别效果。

3.2.2 正字识别任务

表 7 示出了正字识别的实验结果，在基础版测试集上，模型最优准确率为65.64%，在拓展版评测集上，模型最优准确率为19.88%。与通假字检测任务类似，BERT系列模型普遍表现最优，同时，引入微调机制能够进一步提升识别效果，微调给基础版测试集带来的提升比拓展版更为显著。对于未经微调的模型来说，基于四库全书训练的模型效果普遍优于基于殆知阁语料训练的模型。

3.3 实验分析

由前文实验结果可见，对现有基线模型来说，通假字检测和正字识别均为十分具有挑战性的任务，拓展版评测集的难度大大高于基础版。为了进一步探析模型的识别和泛化能力，我们

序号	模型	准确率（基础版）	准确率（拓展版）
1	DaizhigeBigram	34.55%	13.18%
2	SikuBigram	40.69%	14.31%
3	DaizhigeTrigram	33.11%	11.38%
4	SikuTrigram	30.71%	9.93%
5	DaizhigeGPT2	40.79%	13.99%
6	SikuGPT2	43.38%	14.78%
7	DaizhigeBert	35.22%	12.97%
8	SikuBert	42.32%	14.90%
9	ZhengziRecognitionDaizhigeBert	65.64%	19.88%
10	ZhengziRecognitionSikuBert	61.61%	18.96%

表 7. 正字识别任务实验结果

将拓展版测试集按照目标字是否在训练集中收录分为两部分，分别计算了通假字检测和正字识别的实验结果，分别如表 8 和表 9 所示。

通假字分类	字数	精确率	召回率	F1
常见通假字（有训练数据）	279	26.22%	68.10%	37.86%
拓展通假字（无训练数据）	2200	37.60%	7.14%	12.00%
全部通假字	2479	29.48%	16.53%	21.18%

表 8. TongjiaziDetectionSikuBert（无混淆集）模型的通假字检测任务（拓展版）实验结果

通假字分类	字数	准确率
常见通假字（有训练数据）	279	58.97%
拓展通假字（无训练数据）	2200	11.68%
全部通假字	2479	18.96%

表 9. ZhengziRecognitionSikuBert模型的正字识别任务（扩展版）实验结果

如对于通假字检测任务来说，据表 8 可以发现：首先，对于训练数据中未出现的通假字，模型也可以检测出来一部分，并且具有较高精确率，这说明模型具有一定的泛化能力，能够探测出少量训练阶段未见过的通假用法，如例3中的“考”字。第二，对于训练数据收录的常见通假字，模型探测的召回率较高，但精确率却不理想，经过进一步地分析，发现主要有两点原因：(1) 模型倾向于将在训练数据中见过的通假字的非通假用法也判定为通假字，如例4中的“皇”字；(2) 模型实际预测正确，《汉语大词典》中的例句仅针对字头标注通假用法，句中还可能包括其他通假字，数据标注存在少量缺失情况，如例5中的“皇皇”。

例3. 陳登者，善術，夜過吉甫家，即捕登掠**考**，上言吉甫陰事。（“考”通“拷”，“考”字通假用法在训练集中未出现，模型正确预测其为通假字）

例4. 真宗**皇**帝之嘉嘆，面可其奏。（训练集中收录了“皇”的通假用法，但此处“皇”字并非通假，模型错误预测其为通假字）

例5. 孔子三月無君，則**皇皇**如也，出疆必載質。（此处“皇”通“惶”，模型正确预测其为通假字，但由于该句取自《汉语大词典中》“質”通“贄”的例句，其中“皇”的通假用法未被标注，导致评测时误将此例计为误探测条目。）

在正字识别任务中，如表 9 所示，ZhengziRecognitionSikuBert模型同样具有一定的泛化能力。对于训练数据中未覆盖的通假字，来自通假字知识库的混淆集发挥了作用，帮助模型将它们识别了出来。对于未识别出的正字，经分析，发现主要包括两种错误类型：第一，模型认为该位置填通假字比填正字更合适，如表 10 所示，在识别句中“台”的正字时，只有常见的通“鮐”被成功识别，而相对罕见的通“嗣”之用法则未被识别；第二，一个通假字对应着多个正

评测语料	训练数据覆盖	通假字	正字	识别结果
黃耇台背，以引以翼。	否	台	鮐	鮐（正确）
黃髮台背，壽胥與試。	否	台	鮐	鮐（正确）
有于德不台淵穆之讓，靡號師矢敦奮撝之容。	否	台	嗣	台（错误）
聖人共手，時幾將矣。	是	共	拱	拱（正确）
非吾所以共承宗廟意也。	是	共	恭	恭（正确）
唯是桃弧、棘矢以共禦王事。	是	共	恭	供（错误）

表 10. 拓展评测集上的正字识别结果示例

字，模型错误地识别为其他正字，例如，在识别“共”的正字时，存在通“恭”和通“供”两种通假用法，模型将部分通“恭”用法识别为了通“供”，如表 10 中的最后一例:“唯是桃弧、棘矢以共王事。”进一步查阅文献发现，不同学者对通假释读方式存在差异：唐代陆德明《经典释文》注此句中“共”音“恭”，成为清代之前学者共识，《汉语大词典》亦用此说。而以清代俞樾《群经平议》为代表的清人观点认为该字通“供”，并为现代人所继承，如杨伯峻《春秋左传注》、中华书局版《左传》(郭丹等译注)皆同此观点。可见，模型判定虽不同于“标准答案”，但有其合理之处。

总之，通假字的检测和识别是一个复杂的的问题，本文搭建的基线模型能够识别部分通假用法，但泛化能力尚显不足，对微调训练时未能覆盖的通假字，往往无法检测到或准确识别出本字。在识别本字时，对于不常见的通假关系，模型也往往无法正确识别。未来我们仍需要在设计模型时充分集合上下文语义信息与通假字、正字的释义信息，提升模型泛化能力，加强其对不常见通假关系的识别能力。

4 总结

通假是古汉语中的一种常见用字现象，为了服务于通假字的人工判别和机器处理，本文构建了一个涵盖标注语料库、知识库和评测数据集的多维度通假字资源库。在此基础上，本文基于Ngram、BERT、GPT2等主流语言模型开展了通假字自动检测和正字识别实验，为通假字检测和正字识别任务提供了基线结果：在收录常见通假字用法的基础版测试集上，通假字检测的F1值达到66.94%，正字识别的准确率达到65.64%；在拓展版测试集上，模型具备一定泛化能力，能够识别出少量在训练集中未见过的通假字及其正字，但识别效果远远低于基础版评测集。通过对比不同的基线模型，本文发现，模型结构、预训练数据、微调机制和混淆集的使用均会对两个子任务产生不同程度的影响。进一步地，本文对模型的预测误例及原因进行了初步分析。

需要指出的是，本文所开展的通假字资源库建设和通假字识别算法的研究只是该领域的初步探索性工作，研究还存在不少待改进之处。例如，在资源库的建设上，本研究基于《汉语大词典》采集基础性标注语料，词典仅针对字头给出通假例句，例句中仍可能存在其他通假字，有待在后续工作中通过人工标注进行补充；同时，《汉语大词典》所收录的通假用法旨在覆盖基础性、一般性需求，未来还有必要基于面向出土文献和传世文献的通假字辞书资源引入更大范围的通假用例数据，对现有的语料库和知识库进行扩充，从而更好地辅助汉语史领域的相关研究。在自动识别技术上，本研究搭建了通假字检测和正字识别的基线方法，由实验结果可见，通假字检测和正字识别是极具挑战性的自然语言处理任务，目前模型具有一定识别能力，但其准确性和泛化能力还有待进一步提升。此外，基于ChatGPT、GPT4等大模型开展通假字识别是一个值得探索的方向。

最后，在资源库和识别技术的应用上，仍有不少可以开展的工作。例如，通假字资源库及识别算法可以接入古籍整理或古文献检索平台，为该领域研究者提供可能的通假字用例及相关语料信息，辅助专家释读文献，提升古籍整理效率。如前文所述，基于图结构的知识库能够提供传统辞书无法呈现的大规模汉字通假关系网络信息，从而可为古汉语字用现象、词汇发展、词义关联、语音演变等研究提供新视角、新方法。此外，资源库中的高频常用通假字数据可以为文言文教学材料编写、考试命题提供参考，基于该库和其他古汉语领域现有语言资源（如词性标注语料库、词义标注语料库、文白翻译平行语料库等）还可进一步研发辅助文言文学习的工具应用，提升学生的文言文阅读理解能力。

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

致谢

本研究得到国家语委重大项目“古籍整理智能化关键技术研究”(ZDA145-9)、国家自然科学基金青年项目“面向古籍整理智能化的知识表示与加工研究”(62006021)、北京市社科重点项目“古典文献的智能化分析与关联技术研究”(21DTR037)资助。北京师范大学李隽琪、陈青、孟琢等师友为资源库设计提出了宝贵的建议，在此表示感谢。

参考文献

党怀兴. 1998. 通假成因说略. 陕西师范大学学报：哲学社会科学版, (1):61–65.

邓三鸿, 胡昊天, 王昊, and 王东波. 2021. 古文自动处理研究现状与新时代发展趋势展望. 科技情报研究, 3(1):1–20.

胡韧奋, 李绅, and 诸雨辰. 2021. 基于深层语言模型的古汉语知识表示及自动断句研究. 中文信息学报, 35(4):8–15.

胡韧奋, 曹冰, and 杜健一. 2013. 现代汉字形声字声符在普通话中的表音度测查. 中文信息学报, 27(3):41–48.

孔德明. 1993. 通假字概说. 北京广播学院出版社.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern chinese translation with a new large training dataset. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(1):1–13.

柳建钰and 周晓文. 2017. 计算机辅助古籍版本校勘资源库建设浅议. 图书馆理论与实践, (3):54–58.

钱玄. 1980. 秦汉帛书简牍中的通借字. 南京师大学报(社会科学版), (3):44–48.

舒蕾, 郭懿鸾, 王慧萍, 张学涛, and 胡韧奋. 2022. 古汉语词义标注语料库的构建及应用研究. 中文信息学报, 36(5):21–30.

孙建伟. 2015. 假借和通假研究综论. 宁夏大学学报(人文社会科学版), (2):29–33.

苏祺, 胡韧奋, 诸雨辰, 严承希, and 王军. 2021. 古籍数字化关键技术评述. 数字人文研究, 1(3):83.

王宁. 2012. 古代汉语. 高等教育出版社.

Zinong Yang, Ke-jia Chen, and Jingqiang Chen. 2021. Guwen-unilm: Machine translation between ancient and modern chinese based on pre-trained models. In Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10, pages 116–128. Springer.

Xiaoyong Yan, Ying Fan, Zengru Di, Shlomo Havlin, and Jinshan Wu. 2013. Efficient learning strategy of chinese characters based on network approach. PloS one, 8(8):e69745.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. Automatic poetry generation with mutual reinforcement learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3143–3153.

由明智. 2013. 谈人教版中学语文教材的通假字注释. 课程．教材．教法, 33(9):46–50.

张儒. 1988. 关于竹书、帛书通假字的考察. 山西大学学报：哲学社会科学版, (2):37–43+113–114.

附录A. 通假字知识库体例

通假字知识库采用图数据结构，以汉字为节点(node)，字节点之间有通假关系和形声关系两类连边(edge)，节点、边及其属性均以JSON Object形式存储。通假关系边属性会引用语料信息，这些语料没有像“通假字标注语料库”中的语料那样经过详细的标注与校对，只是将不同来源的语料去重后，解析为简单的结构化对象并存储。

字节点具有以下五个属性：

1. 节点ID：用于唯一标识字对象的编号，如“248”、“1764”。
2. 字形：字的书写形态，如“辟”、“譬”。
3. 部首：汉字的构造部分，用于分类和检索字，如“辛”、“言”。
4. 部件：汉字的基本构成单元，包括部首和其他部分，如“卩口辛”、“辟言”。
5. 结构：汉字的构造方式，如“左右结构”、“上下结构”等。

通假关系边具有以下八个属性：

1. 通假字关系ID：用于唯一标识通假关系对象的编号，如“638”。
2. 通假字：在该通假关系中通其他字的字，是有向边的起点，如“辟通譬”通假关系中的“辟”。
3. 正字：被通假的字，是有向边的终点，如“辟通譬”通假关系中的“譬”。
4. 拼音：该通假关系中字音的拼音表示，如“pì”。
5. 注音：该通假关系中字音的注音表示，如“ㄆㄧˋ”。
6. 古音：该通假关系中字音的古代发音，如“《廣韻》芳辟切，入昔，滂。《廣韻》房益切，入昔，並。”。
7. 释义：该通假关系中字的意义或用法解释，如“墨子提出的逻辑推理的方法之一。谓举旁例以喻所说的论题。”。
8. 关联语料ID：与通假关系对象相关的语料对象的编号列表，用逗号分隔，如“8440, 8804”。

形声关系边具有以下三个属性：

1. 形声关系ID：用于唯一标识形声关系对象的编号，如“644”。
2. 形声字：具有特定形声构造的汉字，是有向边的起点，如“譬”。
3. 声旁：形声字的声旁，是有向边的终点，如“辟”。

关联语料具有以下四个属性：

1. 语料ID：用于唯一标识语料对象的编号，如“8806”。
2. 语料文本：包含通假字与通假关系的文本内容，如“是過者也，過猶不及也；辟之是猶立直木而求其景之枉也。”。
3. 语料出处：语料的来源文献，如“《荀子·王霸》”。
4. 语料来源：语料的来源，为“汉语大词典”、“汉典”或“国学大师网汉语字典”。

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

SpaCE2022中文空间语义理解评测任务数据集分析报告

肖力铭[1,2,‡] 孙春晖[1,2,†] 詹卫东[1,2,3,†,*] 邢丹[1,2,‡] 李楠[1,2,†] 王诚文[3,†] 祝方韦[3,‡]
[1]北京大学 中文系
[2]北京大学 中国语言学研究中心
[3]北京大学 计算语言学教育部重点实验室
[†]{sch,zwd,linan2017,wangcw}@pku.edu.cn
[‡]{lmxiao,xingdan,zhufangwei2022}@stu.pku.edu.cn

摘要

第二届中文空间语义理解评测任务（SpaCE2022）旨在测试机器的空间语义理解能力，包括三个子任务：（1）中文空间语义正误判断任务；（2）中文空间语义异常归因与异常文本识别任务；（3）中文空间实体识别与空间方位关系标注任务。本文围绕SpaCE2022数据集介绍了标注规范和数据集制作流程，总结了改善数据集质量的方法，包括构建STEP标注体系，规范描述空间语义信息；基于语言学知识生成空间异常句子，提高数据多样性；采取双人标注、基于规则的实时质检、人工抽样审核等方式加强数据质量控制；分级管理标注数据，优选高质量数据进入数据集。通过考察数据集分布情况以及机器表现和人类表现，本文发现SpaCE2022数据集的标签分布存在明显偏差，而且正误判断任务和异常归因任务的主观性强，一致性低，这些问题有待在将来的SpaCE任务设计中做进一步优化。

关键词： 中文空间语义理解；评测基准数据集；质量控制；STEP标注规范

A Quality Assessment Report of the Chinese Spatial Cognition Evaluation Benchmark

Liming Xiao[1,2,‡] Chunhui Sun[1,2,†] Weidong Zhan[1,2,3,†,*] Dan Xing[1,2,‡]

Nan Li[1,2,†] Chengwen Wang[3,†] Fangwei Zhu[3,‡]
[1]Department of Chinese Language and Literature, Peking University
[2]Center for Chinese Linguistics, Peking University
[3]MOE Key Laboratory of Computational Linguistics, Peking University
[†]{sch,zwd,linan2017,wangcw}@pku.edu.cn
[‡]{lmxiao,xingdan,zhufangwei2022}@stu.pku.edu.cn

Abstract

The Second Chinese Spatial Cognition Evaluation Task (SpaCE2022) aims to test the machine's spatial semantic understanding capabilities, including three subtasks: (1) to judge whether the spatial information in a sentence is correct or incorrect; (2) to determine what causes the abnormal spatial information in a sentence, and locate text fragments with wrong information in a sentence; (3) to label the spatial roles and relations for spatial entities in a sentence. This paper introduces the annotation specifications and the development of the SpaCE2022 dataset, summarizing four quality control methods used in the project. A STEP annotation specification is proposed to standardize the annotation of spatial information of a sentence. Under the guidance of linguistic knowledge, we used the method of replacing spatial semantic words in

*通讯作者
基金项目：国家科技创新2030"新一代人工智能"重大项目（2020AAA0106701）；国家自然科学基金项目（62076008、61936012）

sentences to generate a large number of sentences that may contain spatial semantic anomalies. The sentences generated in this way are diverse in type. Quality control measures in corpus annotation include double-person annotation, rule-based real-time inspection, manual sampling review. According to different sources of annotated corpus, quality classification is carried out, and high-quality annotated data is selected into the final evaluation data set. By examining the dataset distribution as well as machine performance and human performance, this paper finds that the label distribution of the SpaCE2022 dataset exhibits significant bias. Both the correctness judgment task and spatial anomaly attribution task in SpaCE2022 are highly subjective and have low consistency, requiring further optimization in future version of SpaCE benchmark.

Keywords: Chinese Spatial Cognition Evaluation , Benchmark , Quality control , STEP annotation specification

1 引言

空间范畴是人类认知中重要的基础范畴，大量空间信息存在于自然语言文本中。在通往类人智能的道路上，空间语义理解是不可绕开的一环。为了评测机器的空间语义理解能力，自然语言处理领域发布了多项评测任务，具体可以分为以下三类：（1）**空间信息标注任务**，要求机器根据给定的语义角色标注文本中的空间实体和空间关系，形式上与语义角色标注任务以及事件抽取任务相当，代表性的工作有SpRL任务(Kordjamshidi et al., 2012)；(Kolomiyets et al., 2013)和SpaceEval任务(Pustejovsky et al., 2015)；（2）**空间关系推理任务**，要求机器根据文本中已有的空间信息回答涉及空间关系推理的问题，代表性的工作有bAbI任务集(Weston et al., 2015)中的位置推理任务和路径推理任务，以及SpartQA任务(Mirzaee et al., 2021)；（3）**空间语义异常判断任务**，要求机器判断文本是否存在空间信息异常以及异常的归因类型，CCL 2021发布的中文空间语义理解评测任务（Spatial Cognition Evaluation 2021，简称SpaCE2021）(詹卫东et al., 2022)首次提出此类任务，认为如果机器能甄别错误的空间信息并进行正确的归因，就说明机器具有一定的空间语义理解能力[0]。

三类任务提供了评测机器空间语义理解能力的不同角度，但单从一个角度出发并不能全面评估机器的空间语义理解能力。标注类任务只要求对空间相关元素加以识别，因而无法进一步验证机器是否真正理解了这些元素的语义。在已有的推理类任务中，语料都是围绕人为设计的几何图形场景构造出的非真实文本，其空间表达方面的自然度与人类真实对话仍存在一定差距。判断类任务不要求机器定位造成空间语义异常的文本片段，不能测试机器提取空间信息的能力。

在这样的背景下，第二十一届中国计算语言学大会（CCL 2022）发布了SpaCE2022技术评测任务[1]。SpaCE2022包含三个子任务：（1）**中文空间语义正误判断任务**，要求机器对文本的空间语义给出正常或异常的判断；（2）**中文空间语义异常归因与异常文本识别任务**，要求机器识别给定中文文本中空间信息异常的片段，并给出归因类型；（3）**中文空间实体识别与空间方位关系标注任务**，要求机器基于空间关系标注规范，对给定中文文本进行空间实体的识别与空间方位关系标注。

本文介绍SpaCE2022数据集的总体情况，并对参赛系统的表现和数据集质量进行评估和分析，指出数据集和机器模型存在的问题，以及这一任务相关的数据资源潜在的研究价值，为语言认知类评测任务的设计提供借鉴。下文第2节介绍数据集的标注规范、制作流程和数据分布情况，第3节展示基线模型和参赛队伍系统在各项任务上的表现，第4节通过人类表现来评估数据集质量，第5节总结数据集质量控制思路，展望语言认知类评测任务的发展。

[0]http://ccl.pku.edu.cn:8084/SpaCE2021/
[1]https://2030nlp.github.io/SpaCE2022/

第二十二届中国计算语言学大会论文集，第547页-第558页，哈尔滨，中国，2023年8月3日至5日。

2 SpaCE2022数据集总体情况

2.1 数据集标注规范

SpaCE2022数据集的标注面向三个子任务分为了三个层面[2]。一是面向空间语义正误判断任务，仅标注文本中是否存在异常空间信息，存在即标注“异常”，否则标注“正常”；二是面向空间语义异常归因任务，针对上一层标注中标签为“异常”的语料，进一步标注异常的类型以及存在异常的具体文本片段；三是面向空间实体识别与空间方位关系标注任务，精选了部分语料，基于STEP空间语义标注体系（S表示空间实体，T表示时间，E表示事件，P表示空间方位信息）进行精细的空间信息标注。

子任务一采用区间标注法，构造了一个空间语义完全正常到显然异常的连续区间，然后切分为四段子区间，分别对应四个标注选项，依次是：完全正常、尚能说通、比较牵强、显然异常。**完全正常**指句中实体的空间方位信息表达正确，毫无争议。**尚能说通**指实体的空间方位信息表达大致能成立，但表达的准确性和自然程度上存在一些问题，如“他听到一段钢琴声从茅屋边传出来”，用“茅屋里”表达感觉更自然，但“茅屋边”的表达也有出现的可能。**比较牵强**指实体的空间方位信息表达不大能成立，虽然这个句子的空间语义还没有到完全无法理解的程度，但空间语义信息罕见，如“他带着狗到森林旁去打猎”，这种情况基本不会出现，但从语义上看，“森林旁”所表示的空间关系仍然成立。**显然异常**指句中实体成分的空间方位信息表达错误，如“每辆自行车下座上都用一根扁担绑着两只很大的箩筐”，自行车只有“后座”，没有“下座”，与常识相悖，所以标注这个句子的空间信息为“显然异常”。

子任务二的标注需要选择造成空间信息异常的原因，并标注异常信息对应的文本片段。三个归因类型选项分别是：搭配不当、语义冲突、不符合常识或背景信息（以下简称不合常识），分别对应A、B、C三类标签。表1 是各选项的含义和标注示例。

选项	含义	示例
搭配不当	句中两个表示空间信息的语言成分，受到语法、韵律、语言习惯等因素的影响，不能组合。而且，这两个成分在其他语境中也都不能组合。	弯曲双膝，弯曲双肘，把头放在[**text1 地面**][**text2 边**]。
语义冲突	句中两处空间信息存在矛盾、冲突。这两处空间信息对应着两个事件。	池水照见了她的面容和身影；她笑，[**P1 池水里**]的[**S1 影子**]也向着她[**E1 笑**]；她假装生气，[**P2 池水外**]的[**S2 影子**]也向着她[**E2 生气**]。
不符合常识或背景信息	句中有一处文本片段的空间信息违反常识或者违反句子的背景信息。	那个苏联人孤零零地躺在那毫无遮掩的方场上，[**S 一只手臂**][**E 枕**][**P 在脑袋上面**]。

表 1: 子任务二的标注选项说明

在描述异常方面，搭配不当涉及语言使用中的搭配习惯问题，即两个相邻成分text1和text2是错误的组合。语义冲突和不合常识这两类的情况则更为复杂。为了更加规范地描述空间信息异常，子任务二引入空间语义三要素来描述句子的空间语义：空间实体（S）、空间方位（P）、空间义相关事件（E）。S指句中描述了空间方位信息的实体，P指空间实体在句中出现的空间方位信息，可能涉及处所、起点、终点等信息，E是动词性单位，表达了S位于P的方式、目的或原因。不合常识选择一组S-P-E三要素即可说明空间信息异常，而语义冲突需要两组S-P-E三要素来说明问题，见表1 中的示例。

子任务三在S-P-E标注法的基础上增加了时间信息的标注，形成了STEP空间语义四要素标注体系，描述信息可概括为：“某空间实体在某时，经由某事件，处于某种空间方位关系，这一

[2]了解标注规范详情请访问规范文档（https://2030nlp.github.io/Sp22AnnoOL/menu）和标注样例（https://2030nlp.github.io/Sp22AnnoOL/examples）。

第二十二届中国计算语言学大会论文集，第547页-第558页，哈尔滨，中国，2023年8月3日至5日。

命题的事实性为真/假”。事实性为假的空间信息用F要素来表示。按照该体系标注一段文本的空间信息，一条信息对应的数据为一个包含18项元素的元组，表2 是每个元素的含义。

序号	元素名称	所属要素	含义
1	空间实体	S	对应于被描述空间方位的空间实体。
2	参照实体	S	对应于与1号元素形成距离关系的另一个空间实体。
3	事件	E	与空间实体的空间方位直接关联的事件。
4	原文时间	T	文中写明的与空间方位相关联事件的时间表述。
5	参照事件	T	如果空间实体处于某种空间方位关系的时间在文中并未写明但可以判断，则可通过此元素和6号元素共同描述。此元素描述了6号元素所参照的事件。
6	参照时间	T	当文中未出现描述空间方位关系的具体时间，且该时间可以判断时，通过此元素描述空间方位关系的时间。值有”说话时”/”过去”/”将来”/”之时”/”之前”/”之后”/”之间”
7	事实性	F	如果空间方位命题是假的，则该字段为”假”。
8	处所	P	描述静态空间实体相对某外部参照物的位置。
9	起点	P	描述动态空间实体的方位发生变化的场景下，变化开始时实体的处所。
10	终点	P	描述动态空间实体的方位发生变化的场景下，变化结束时实体的处所。
11	方向	P	描述动态空间实体的位移方向。（空间实体在动态中才有方向特征）
12	朝向	P	描述空间实体某一侧面所朝向的位置。
13	部件处所	P	描述空间实体作为一个部件在整体中的位置。
14	部位	P	描述了空间实体的某个部位。
15	形状	P	描述了空间实体的形状。
16	路径	P	描述了空间实体位移时经过的轨迹。
17	显式距离	P	文中写明的描述了空间实体间距离关系的表述，与1号元素和2号元素共同描述。
18	隐式距离	P	文中并未写明空间实体间的距离关系但可以推断，值为”远”/”近”/”变远”/”变近”。

表 2: 子任务三的元素含义

2.2 数据集制作流程概要

数据集的构建流程包括：筛选数据、生成替换句、开展标注工作、划分数据集。下面分别介绍:

1. 筛选合适的原始句子。SpaCE2022注重语料类型的丰富性，收集了报刊、文学作品、中小学语文课文、交通事故判决书、体育动作训练手册、地理百科全书等多领域的数百万字生语料，使用pkuseg(Luo et al., 2019)对语料进行分词和词性标注，通过定义空间语义表达加权值筛选得到待进入标注流程的原始文本，排除错别字和词性标注错误等问题后，得到6,643条语料（称为“原句”），进入后续标注流程。

2. 基于词语替换得到替换句语料。制作空间语义表达词表，并从可替换角度，对词表进行分组。根据这些分组，由程序遍历每个原始句子中的空间方位意义词语，将之替换为同组的其他词。6,643个“原句”经过替换，得到了47,920个“替换句”。每一个替换句都有1或2个词语与原始句不同。替换句的空间信息既可能异常，也可能正常。

与SpaCE2021相比，SpaCE2022的替换词表更具系统性，并且在进行替换操作时加入了过滤规则，替换效果更好。SpaCE2021替换词表不记录词性，在进行替换操作时直接根据字符匹配定位可替换词；SpaCE2022替换词表区分了方位词、处所词、趋向动词、介词、副词等词类，能够在替换时区分兼类词，如“上”字一词既有可能是方位词，也有可能是趋向动词，前

者对应替换词“下、中、前、后”等，后者对应替换词“下”。考虑到任务主要关注物理空间的空间语义，SpaCE2022不对抽象名词后面的方位词进行替换。此外，SpaCE2022在替换单个词的基础上还新增了同时替换两个方位词的替换模式。数据集的替换情况显示，平均每个原句生成的替换句数量从SpaCE2021的40.52下降到SpaCE2022的8.41，词表中每个词的平均替换频次从SpaCE2021的6.03下降到SpaCE2022的4.39，说明SpaCE2022替换句的多样性得到充分提升。

3. 就正误判断任务进行人工标注，并作筛选。共招募229名标注人员参与此项工作，标注了44,921个句子（皆为替换句）。标注时需判断句子的异常程度，标注界面如图1所示。每条语料由2名标注员标注，并有质检员进行抽查，以控制标注质量。标注完成后，根据双人标注的一致程度，对语料进行分流，其中2人一致标注为“完全正常”或一致标注为“显然异常”的句子被认为是最可靠的标注语料，共计15,747条，进入最终评测数据集。

4. 就异常归因任务进行人工标注，并作筛选。共178名标注人员参与，标注了10,614个句子（皆为替换句）。此步骤以正误判断任务中被标注为异常的句子为原材料，要求标注者判断异常的类型，并选出存在异常的具体文本片段。异常类型参见上文表1 说明。标注界面如图2所示。每个句子由1-2名人员标注，并定期审核标注内容。最终根据双人标注的一致性、审核意见以及标注员标注水平分级情况，选出其中7,068条相对可靠的标注句进入评测数据集。

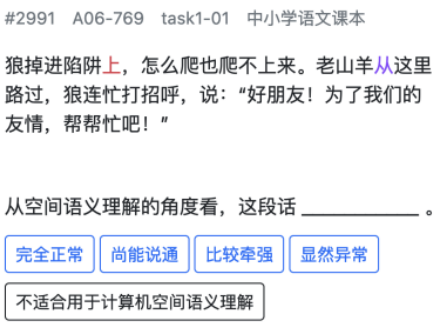


图 1: SpaCE2022正误判断标注系统工作界面

图 2: SpaCE2022异常归因标注系统工作界面

5. 使用STEP标注体系进行细粒度空间信息标注。共71名标注人员参与，标注了3,223个句子。此步骤要求标注者遵照前文介绍的STEP标注体系，将句子中涉及的空间信息详尽地标注出来。标注界面如图3所示。每个句子仅由1名人员标注。为保证质量，审核员对每名标注员的标注进行定期抽检；同时标注工具也提供了自动检查功能，将可能存在的问题实时地通过标注界面反馈给标注员，如被标注为空间方位P的片段首尾一般不为动词。标注完成后，通过程序剔除了怀疑有不符合标注规范情况的部分语料，最终选取了2,152个句子进入评测数据集。

划分训练集、验证集和测试集时，为避免机器通过比对同一原句生成的不同替换句学习到替换规律，规定某一原句和它生成的所有替换句只能出现在同一个数据集合中，即训练集、验证集或测试集三者之一。经过上述步骤，最终得到SpaCE2022数据集，共24,947条标注语料，数据集构成如表3所示。

子任务	训练集	验证集	测试集	总计
1.中文空间语义正误判断任务	10993	1602	3152	15747
2.中文空间语义异常归因与异常文本识别任务	4966	700	1402	7068
3.中文空间实体识别与空间方位关系标注任务	1529	207	396	2132

表 3: SpaCE2022数据集的构成

图 3: SpaCE2022空间语义角色标注系统工作界面

2.3 数据分布情况

SpaCE2022语料共计285万字符，每段语料字符数均值为114.23，标准差为49.57。语料涉及多种不同类型和来源，各类语料比例为：报刊（37%）、文学作品（25%）、中小学课本（20%）、交通事故判决书（9%），体育动作（6%）、地理百科（2%）、其他（1%）。下面分别考察各子任务的标签分布情况。

子任务一有两个标签："正常"和"异常"，下面用"正例"指标签为"正常"的语料，用"负例"指标签为"异常"的语料，表4是各子集的标签分布情况。正负例比重是正例数和负例数的比值，可以衡量二元标签的平衡性，越靠近1，正例和负例越平衡。从表中可看出，子任务一数据集以负例为主，说明替换空间义词语更有可能引发空间信息异常。测试集的正负例比例接近1，制作子任务一数据集时优先考虑了测试集的平衡性。

数据集	正例数	负例数	正负例之比
训练集	2677	8316	0.32
验证集	705	897	0.79
测试集	1695	1457	1.16
合计	5077	10670	0.48

表 4: 子任务一各子集的标签分布情况

子任务一中，替换词和替换对的分布呈现出了标签偏向。双音节处所词所在的语料倾向于空间信息正常，而有两个替换词（下文称为替换组）的语料则倾向于空间信息异常。近义词构成的替换对使用前后，空间义基本未发生改变，所以偏向正常，而反义词构成的替换对偏向异常，如"下来→上来"、"进去→出来"。此外，中心义词和外围义词的替换也容易导致空间信息异常，如"中→旁"、"中→外"。

子任务二共有三个标签，分别对应三种归因类型，具体参见上文表1。表5是子任务二数据集归因标签的频次。"&"代表联合归因。整个数据集中，三类标签的分布很不平衡，约58%的语料被归因为不符合常识或背景信息，其次是搭配不当和语义冲突。联合归因中，搭配不当经常和不符合常识或背景信息共现，说明这两个类型的区分并不显著。三类归因共现最少见。

标签	A	B	C	A&B	A&C	B&C	A&B&C
频次	870	821	4102	77	848	317	33

表 5: 子任务二数据集的归因标签分布情况

第二十二届中国计算语言学大会论文集，第547页-第558页，哈尔滨，中国，2023年8月3日至5日。

替换词和替换组在归因类型标签的分布上没有明显偏向。在联合归因中，替换词和替换组之间也都表现出单因大于双因、大于三因的现象，未发现数据偏差。替换对中，原词和替换词的搭配能力有明显区别时，偏向**搭配不当**，比如“当地”一般修饰名词，而“原地”一般修饰动词。偏向**语义冲突**的替换对包含了绝对方位词，替换后的方向在上下文构建的空间场景不能成立，与其他绝对方位词相冲突。偏向**不合常识**的替换对以“两侧”作原词为主，替换成其他单侧的方位词后，不能满足空间实体的要求，违反了常识。

子任务三的STEP标注体系共使用了18个元素，训练集、验证集和测试集的元素总量分别为25224, 3848和5357。每个元素占子集元素总量的占比情况见附录A。每个元素在各子集中均有分布且占比接近，但元素之间的分布严重失衡。空间实体比事件高约10个百分点，说明不是所有标注都有事件信息出现。这个特点对于使用事件抽取范式的系统而言是一大挑战，因为缺少动词性单位意味着缺少事件抽取所需的触发词。

3 机器在SpaCE2022任务上的表现

3.1 评测指标

子任务一是正误判断任务，以准确率为指标。子任务二设计了3种评价指标，分别是①异常文本归因准确性，②异常文本识别准确性，③异常元素识别准确性。指标①以准确率的形式考察参赛系统进行异常归类的能力；指标②以F1值的形式考察参赛系统对异常文本进行定位的能力；指标③则以F1值的形式考察参赛系统对于异常信息所属的具体要素进行分类的能力。鉴于指标③实际上涵盖了前两项指标所考察的方面，我们使用指标③作为评测赛事的排名依据。

子任务三的数据在评测阶段组织为元组形式，每个元组对应一条空间信息标注，每个句子对应若干个元组。每个元组含有18个槽位。评分程序会对参考答案和机器答案中的元组进行两两比较，对于每个参考元组和机器答案元组，程序根据一定标准计算其中每个元素的得分，求和得到该元组的得分，以及该题的总分。最后，根据每题得分计算所有题目的F1 值，作为最终得分。

3.2 机器在各项任务上的表现

3.2.1 基线系统

SpaCE课题组为评测建立了一套基线系统。子任务一使用预训练模型BERT(Devlin et al., 2019)构建了一个二元分类器。子任务二设置了一个分类层预测归因类型，以及一个序列标注层判断每个词所属的元素，两个模块采用独立编码器。子任务三首先进行序列标注任务，寻找文本中能够触发事件抽取的触发词，然后根据触发词抽取其他元素。

3.2.2 参赛系统

SpaCE2022共有32支队伍报名，最终3支队伍提交了测试结果。参赛系统普遍使用不同的模型架构来分别完成三个子任务。子任务一中，参赛系统均使用了判别式预训练模型Electra(Clark et al., 2020)来完成二分类任务，队伍1和队伍3进一步使用集成模型的方法提升准确率，队伍2则构建了方位词表，通过计算方位词的最大替换概率来判断句子是否存在空间信息异常。

子任务二中，队伍1使用阅读理解任务的范式，针对三个归因类型分别训练了三个模型，模型会预测每个异常文本片段的开头和结尾。队伍2训练了一个序列标注器，给每个词打上S、P、E或O（表示非目标词）的标签，从而找到异常文本片段。队伍3利用w2ner(Li et al., 2022)架构可以抽取不连续实体的特性，同时抽取所有归因类型的异常文本片段。

子任务三中，队伍1采用抽取和生成两阶段的方法，抽出18元组的主语（空间实体）后，使用生成模型生成其他部分。队伍2使用信息抽取预训练模型UIE抽取出每个空间元素。队伍3将18元组拆分为多个3元组，使用关系抽取模型gplinker(苏剑林, 2022)抽出3元组后，再合并为18元组。

表6是参赛系统和基线系统在三个子任务上的表现。所有参赛系统在子任务一的表现都超过了基线模型，因为基线模型不擅长同时捕捉多种错误模式，而参赛系统通过集成模型等方式能改善此问题。子任务二中，所有系统在文本准确性指标上的表现都优于归因准确率指标，一方面说明所有系统都更擅长识别异常文本片段，另一方面可能说明三种归因类型的区分度不明显，归因难度较大。所有系统的元素准确性指标得分也都显著低于文本准确性指标，说明系

统在为异常文本片段标注SPE要素时遇到了困难。子任务三中，所有参赛系统的得分均低于基线系统，且所有系统的得分均低于0.6，主要的原因可能是元组和元素的抽取数量较多，而且约30%的元组没有触发词，这对事件抽取模型而言有较大的难度。

系统	子任务一准确率	子任务二			子任务三F1值
		归因准确率	文本准确性F1值	元素准确性F1值	
队伍1	0.7865	0.0036	**0.8075**	**0.6748**	0.4950
队伍2	**0.7992**	**0.5827**	0.6432	0.4877	0.3870
队伍3	0.7985	0.2268	0.3324	0.2822	0.4387
基线系统	0.5864	0.5599	0.5812	0.4403	**0.5069**

表 6: 参赛系统和基线系统在任务上的表现

4 人类在SpaCE2022任务上的表现

课题组在数据集评测阶段进行了人类测试，包括用相同指标计算人类得分，以及进行一致性检验。人类得分反映了任务的难度，也为参赛系统提供人类基准。一致性检验关注不同人对同样的语料是否有一致的标注结果，反映人类对标注任务的理解是否趋同。一致性低，可能是评测任务的问题主观性比较强，不同人的看法不太容易趋同。如果存在上述问题，则有理由怀疑数据集的质量不可靠。下面分别介绍人类在三个子任务上的表现情况。

4.1 子任务一的人类表现

课题组招募了来自不同年级和专业的大学生共7名被试，进行子任务一的标注培训。培训合格后，他们需要独立完成100道测试题。这100道题目是根据语体占比随机抽取的，包括50条正例和50条负例。表7是所有被试的准确率，最高分为0.95，最低分为0.69，平均分为0.78，与参赛系统的得分相当。被试4和被试7属于异常值，最终取5名被试的结果进行了Kappa值的计算，衡量人类在子任务一的一致性，如表8所示。Kappa均值为0.53，在Kappa值分级标准下属于中等水平[3]。

人类	准确率
被试1	0.74
被试2	0.78
被试3	0.83
被试4	0.69
被试5	**0.95**
被试6	0.80
被试7	0.69
均值	0.78

表 7: 子任务一人类被试准确率

人类	被试1	被试2	被试3	被试5	被试6
被试1		0.53	0.57	0.49	0.42
被试2	0.53		0.51	0.56	**0.37**
被试3	0.57	0.51		**0.76**	0.48
被试5	0.49	0.56	0.76		0.59
被试6	0.42	0.37	0.48	0.59	
均值	0.43	0.47	0.52	0.54	0.41

表 8: 子任务一人类被试的Kappa值

进一步观察每一道题的答题情况，有76道题是至少4名被试做出了相同的判断，其中有38道题为5名被试的判断都相同，这说明这76道题的客观性比较强，被试有较为趋同的理解。包含近义词替换对的题目，被试倾向于做出空间信息正常的判断，因为可以构建出相似的空间场景。如“今晚在院子里坐着乘凉”，如果用“中”替换“里”，这个句子的空间信息依然正常。包含反义词替换对的题目，被试则倾向于做出空间信息异常的判断。如“口袋全挂在外面像是被抢劫了一样”，如果用“里面”替换“外面”，空间义与下文的“抢劫”相冲突。如果空间实体的维度形态在人类的认知中趋于一致，那么被试对实体可以或不可以搭配的方位词也有比较一致的认识，从而得出相同的答案。如“森林”是三维实体，当搭配强调二维平面的方位词“上”时，会出现搭配不当的情况，被试很容易发现异常。

剩余24道题的主观性较强，不同人的看法有所不同。一种情况是上下文缺少助于判断的线索，被试需要调用个人认知经验。比如“在宇航中心的食堂前吃了早饭”，有的被试认为可以构

[3]Kappa值在[0.41, 0.60]区间为中等的一致性。

建“食堂前”的场景并且在这里吃早饭，空间信息是正常的，而有的被试则认为基于常识，“食堂里”才是供应早饭并提供吃早饭的地方，所以空间信息异常。另一种情况是替换词为包含绝对方向词“东、南、西、北”等的交通文本，被试需要有较好的方位感和空间想象能力，否则容易判断错误。

4.2 子任务二的人类表现

子任务二共有9名被试参与人类一致性检验，经过培训后，他们需要独立标注50条随机抽取自子任务二测试集的语料。表9是所有被试的得分，文本准确率平均分为0.82，归因准确率均分为0.65。每名被试的文本准确性都比归因准确率高，被试2甚至高出约42个百分点，这说明被试能够较好地定位异常文本片段，但对归因类型的理解不到位，也反映出归因类型的划分主观性较强，有较大的改进空间。与参赛系统相比，人类在文本准确性上的表现显著优于队伍2和队伍3，类型准确率的表现显著优于队伍1和队伍3，但与表现最好的系统相当。

人类	文本准确性（F1）	归因准确率（Acc）
被试1	0.84	0.70
被试2	**0.78**	**0.36**
被试3	0.81	0.74
被试4	0.82	0.62
被试5	**0.87**	**0.76**
被试6	0.82	0.64
被试7	0.83	0.70
被试8	0.83	0.70
被试9	0.81	0.66
平均值	**0.82**	**0.65**

表 9: 子任务二人类被试的文本准确性和归因准确率

4.3 子任务三的人类表现

课题组对子任务三开展了数据集抽检质量评估工作。评估人员需要检查题目的标准答案，对认为有误的答案进行增加、删除或改动等操作。课题组招募了4名审核员完成这项工作。根据语体占比和标签占比，课题组从测试集中随机抽取了70条数据分发给审核员。最后，以每一位审核员的结果为标准答案，使用子任务三的评测脚本计算其他审核员的F1值。如果F1值高，说明他们有相同的增删改操作，对任务有较为一致的理解和认识。表10是审核员两两比对的F1值，总平均值达0.88，具有较高的一致性。其中，审核1和审核4的一致性最高。

人类	**审核1**	**审核2**	**审核3**	**审核4**
审核1		0.89	0.89	**0.92**
审核2	0.89		**0.86**	0.87
审核3	0.89	0.86		0.88
审核4	0.92	0.87	0.88	
均值	0.90	0.87	0.88	0.89

表 10: 子任务三人类审核的一致性得分

进一步考察每一条标注数据的一致性，约24%的数据完全一致，一致率大于80%的数据占八成。通过分析13条增删改次数差别较大的语料，发现造成标注差异的因素有：

（1）空间隐喻可能造成对空间实体的标注有主观认识上的差异。没有实体的抽象概念在空间隐喻的作用下可能会被视为空间实体。如“那三个老者的讥笑一句也没听进耳中”，使用了空间隐喻来描绘话语的传递过程，3号审核员认为“讥笑”是空间实体，标注了终点“耳中”、方向“进”和事实性“假”；

（2）空间推理可能造成对隐性空间信息的标注有主观认识上的差异。对于文本中通过简单推理可以得出的隐性空间语义，任务三要求进行标注，但标注规范本身对此缺少明确规定，

标注员往往难以把控尺度。简单的推理能够增加标注的丰富度，不合适的推理可能会让标注变得过于复杂。如“两个孩子往里面张望，一辆汽车都没有”，可以推理出“汽车不在里面”的空间义，简单且合理。但对于“一个巨大的小行星或彗星撞击地球”，通过标注距离角色来表示彗星与地球越来越近，不同的标注员处理就可能不一致，因为这个事件并不强调二者之间的距离，推理性较弱；

（3）对可能性与事实性的区分有主观认识上的差异。“可能”、“也许”等词可能影响事实性的判断，如“垃圾可能掉入河道内”，三名审核员认为“垃圾”位于“河道内”的事实为假，有一人认为是真。另外，带虚拟语气的句子和假设关系的句子也常常出现事实性标注不一致的问题；

（4）空间方位描述中存在一词充任多个角色的情况，容易漏标。如“手里的瓜子和蚕豆越掉越多”，在“手里的瓜子和蚕豆”中，“手里”是“处所”。在整个句子中，而“掉”蕴含了位移信息，“手里”又可以作为“起点”理解。4名审核员都只标注了其中的一种情况，存在信息标注不完备的问题。

通过考察人类在SpaCE2022任务上的表现，子任务一，对文本中空间信息是正常还是异常的判断，一致性较差，作为评测任务，评价难度较大；子任务二，识别异常文本片段的可操作性较强，一致性较高，但异常归因则界限模糊，一致性较低；子任务三，人类标注员在标注时也存在一定的主观差异，且达到标注信息完备的难度较高，反映出该任务的信息丰富度和复杂性，这是参赛系统普遍表现不佳的原因之一。

5 结语

SpaCE2022在SpaCE2021的基础上扩充语料规模和语料类型，提出了覆盖面更广的任务设计，包括：考察机器对空间信息异常的判断能力、异常片段的识别能力和归因能力，以及提取结构化文本空间信息的能力。为了构建高质量的数据集，SpaCE2022课题组在语料准备阶段，利用词类、约束规则等语言学知识，提高了空间信息异常句的生成多样性；在任务设计阶段，提出了S-P-E标注法和STEP标注体系，对空间信息异常和结构化空间语义信息的表征做了明确而系统的规范；在标注阶段，采用规则自动查错、双人标注和抽样审核等方式控制标注质量；在构建数据集阶段，对标注数据的价值进行合理分级，优先选用高质量数据，并充分考虑了语料在训练、验证和测试数据集上分布的合理性。

尽管如此，机器和人类的表现仍反映出数据集的质量控制问题。人类在子任务一和子任务二上的表现反映出正误判断和异常归因的任务设计存在一定的主观性，不同人对空间异常和归因类型的理解不一定相同，这可能降低评测结论的信度（Reliability）。机器和人类在子任务二的文本准确性指标上的表现都显著优于归因准确率指标，说明S-P-E标注法既有助于人类选取趋同的文本片段，也有利于机器学习描绘异常的方法。另一方面，也说明归因分类任务的设计还需要进一步优化，增加约束条件，提高标注的一致性。

数据集的质量控制还体现在标签的分布情况。标签分布失衡会降低数据集的效度（Validity），比如子任务一的替换组在标签分布上明显偏向负例，这可能使机器捕捉到替换词数量与标签分布的关系，无法有效反映机器真实的空间语义理解能力。再如，子任务三的标注规范虽然覆盖到了所有标签，但实际语料中某些标签数量过少，训练不充分，机器可能无法学习到相应的空间知识。标签分布失衡还在一定程度上反映了质量分级策略的局限性，分级规则可能让具有某一特征的可用数据被错误排除在外，如子任务三的标注规范要求P信息不以时体助词结尾，但分级规则没有区分时体助词“过”和表经过义的“过”，导致以经过义“过”为结尾的路径标签没有进入数据集。

为了更为全面和准确地探究机器空间语义理解能力，并推动语言认知类评测任务的发展，数据集的质量控制仍然是下一步待改进的重要问题。其中影响标注一致性的认知因素应研究更有效的规范和约束方法，而如果标注员经过多轮培训后，标注一致性仍不够理想，则需要考虑修改标注规范的体系设计。构建数据集时还应确保覆盖到尽可能多的评测对象，避免标签分布存在明显的偏差。

参考文献

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens, and Steven Bethard. 2013. Semeval-2013 task 3: Spatial role labeling. In *Second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 255–262.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: Spatial role labeling. In {** SEM 2012*}*: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* {*(SemEval 2012)*}, volume 2, pages 365–373. ACL.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

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

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

苏剑林. 2022. Gplinker：基于globalpointer的实体关系联合抽取. https://kexue.fm/archives/8888.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 2022. 空间语义理解能力评测任务设计的新思路—space2021数据集的研制. 语言文字应用, (02):99–110.

A 附录

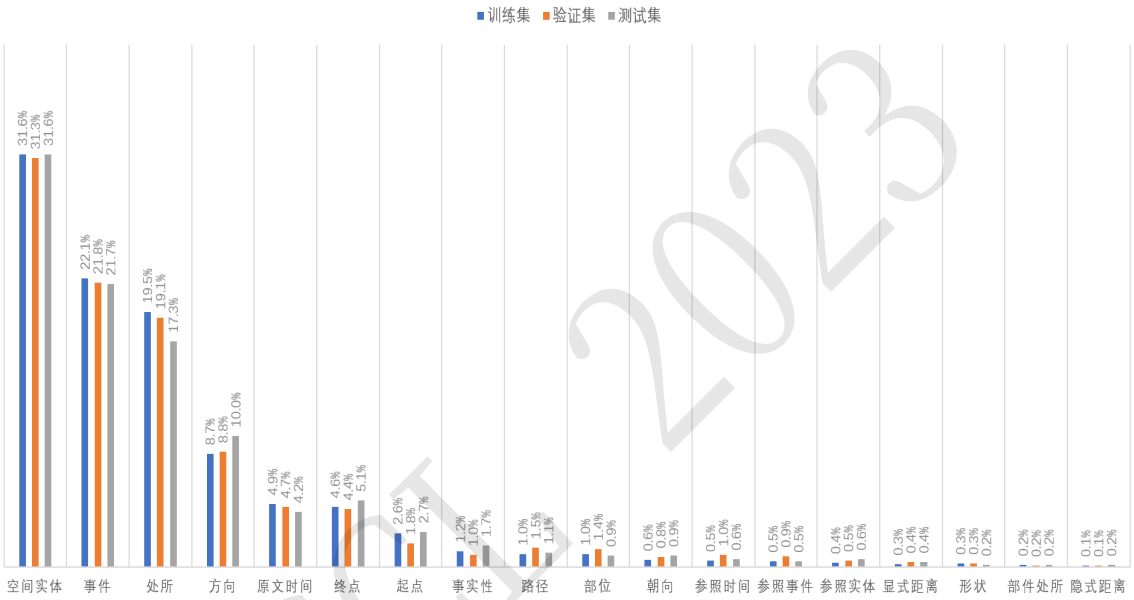


图 4: 子任务三数据集的标签分布图

基于预训练语言模型的端到端概念体系构建方法*

王思懿[1,2]，何世柱[1,2]，刘康[1,2]，赵军[1,2]
[1]中国科学院自动化研究所，复杂系统认知与决策实验室
[2]中国科学院大学，人工智能学院
wangsiyi2021@ia.ac.cn, {shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

摘要

概念体系是一种重要的知识资源，描述了概念之间的上下位关系并以层次结构进行组织。本文致力于研究概念体系的自动构建技术，旨在将给定的概念集合按照上下位关系组织成树状的概念体系（概念树）。传统方法通常将概念体系构建任务分解成两个独立的子任务：概念间上下位语义关系的判断和概念层次结构的生成任务。然而，这两个子任务缺乏信息反馈，容易导致错误累积等问题。近年来，使用预训练语言模型获取词语的语义特征、判断词语之间的语义关系已经成为一种流行的方法，在概念体系构建任务中取得了一定的效果，但是这种方法只能对第一个子任务进行建模，并且仍然存在错误累积的问题。为了解决这个问题并有效地获取词语及其关系的语义特征，本文提出了一种基于预训练语言模型的端到端概念体系构建方法。该方法一方面利用预训练语言模型获取概念及其上下位关系的语义信息和部分概念体系的结构信息，另一方面，利用强化学习端到端地建模概念关系的判断和完整体系结构的生成。在WordNet数据集上进行的实验表明，本文提出的方法取得了良好的效果。在相同条件下，本文方法比最好模型在F1值上能取得7.3%的相对性能提升。

关键词： 概念体系构建；强化学习；预训练语言模型

End to End Taxonomy Construction Method with Pretrained Language Model

Siyi Wang [1,2], Shizhu He [1,2], Kang Liu[1,2], Jun Zhao[1,2]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
wangsiyi2021@ia.ac.cn, {shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Concept system is an important knowledge resource, which describes the Hyponymy and hypernymy between concepts and organizes them in a hierarchical structure. This paper focuses on the automatic construction technology of concept system, aiming to organize the given concept set into a tree like concept system (generalization tree) according to the Hyponymy and hypernymy. Traditional methods usually decompose the task of building a conceptual system into two independent subtasks: determining the semantic relationships between concepts, and generating conceptual hierarchical structures. However, the lack of information feedback in these two subtasks can easily lead

*本论文受到国家重点研发计划项目（2022ZD0160503），国家自然科学基金项目（U1936207，61976211），中国科学院青年创新促进会和云南省重大科技专项（No.20220AD080004）资助。

to issues such as error accumulation. In recent years, using pre trained language models to obtain semantic features of words and determine semantic relationships between words has become a popular method, which has achieved certain results in concept system construction tasks. However, this method can only model the first subtask and still has the problem of error accumulation. In order to solve the problem of error accumulation in step-by-step methods and effectively obtain the semantic features of words and their relationships, this paper proposes an end-to-end conceptual architecture construction method based on pre-trained language models. On the one hand, this method uses the pre- trained language model to obtain the semantic information of concepts and their hyponymy and hypernymy and the structural information of some conceptual systems; on the other hand, it uses Reinforcement learning to model the judgment of conceptual relationships end-to-end and the generation of complete architectures. The experiments conducted on the WordNet dataset show that the proposed method has achieved good results. Under the same conditions, the proposed method can achieve a relative performance improvement of 7.3% on F1 values compared to the best model.

Keywords: Taxonomy Induction , Reinforcement Learning , Pretrained Language Model

1 引言

概念体系使用层次结构描述概念之间的上下位语义关系，是一类重要的知识体系，属于有向图的一种 (Yang and Ni, 2022)，被广泛的用于问答系统、信息检索 (Demeester et al., 2016; Yang et al., 2017)等任务中，例如，在医疗概念体系用于组织和管理疾病。同时，大量神经网络模型也常利用概念体系的知识内容增强语义表达能力。但是目前各类应用系统中的概念体系主要由人工完成，例如，语言概念体系WordNet (Miller, 1995)和常识概念体Cyc (Lenat, 1995)等完全由人工构建，DBpedia (Auer et al., 2007)中的概念体系是工程师通过观察Wikipedia中的标签（Tag）体系信息的组织方式及命名后总结得到，Schema.org[1]中类型、属性和关系的定义及规范由众多科技公司协商得到。这种方式不仅费时费力，构建的概念体系还常常存在概念遗漏的情况，使得更新和维护成本高昂，因此亟需设计和实现概念体系的自动构建方法。

传统方法将自动构建概念体系的任务分为两个独立的子任务，即上下位关系判断和概念体系构建。上下位关系判断任务通过基于模式的方法从大型语料库中提取候选词语对，并形成带有噪声的概念体系图，然后采用不同的修剪算法来提取树状概念体系 (Granot and Huberman, 1981)。然而，这种方法存在一些缺陷。首先，词语的上下位关系判断任务仅仅预测词语对的上下位关系概率，忽略了概念体系的结构信息。其次，在构建概念体系的第二个子任务中，使用固定的词语对特征表示，该特征表示是由第一个子任务产生的，缺乏跨任务的反馈修正，因而存在错误传播问题 (Sun et al., 2022)。

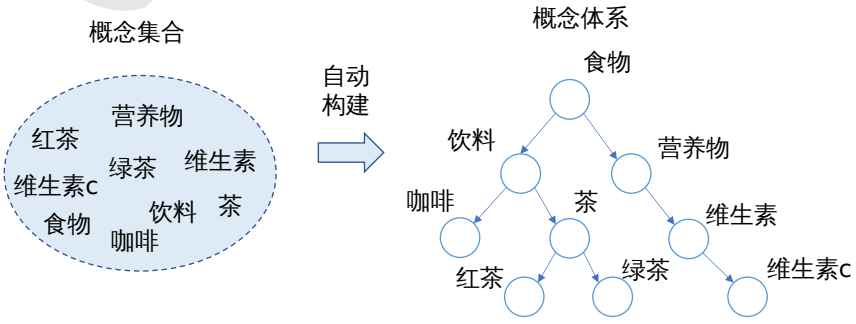


图 1: 概念体系构建示意图

强化学习在自然语言处理领域 (Sun et al., 2022)应用十分广泛 (Kwan et al., 2023)，为了解决这些问题，有研究提出了一种基于强化学习的联合建模方法 (Mao et al., 2018)，即利用强化

[1]https://schema.org

学习的策略网络逐步构建概念体系。然而，该方法中的词语选择过程缺乏已构建概念体系的信息，并且词语的特征向量表示是固定的，因此所取得的效果受到一定限制。在预训练语言模型被提出后 (Kenton and Toutanova, 2019)，其被广泛应用于分类和语言表示等任务。为了获得更好的效果，CTP (Constructing Taxonomies from Pretrained Language Models) (Chen et al., 2021)将上下位关系判断任务转化为二分类任务，取得了WordNet数据集上最佳的成果。然而，由于CTP是分步式的，仍然存在错误累积问题。

我们提出了一种名为基于预训练语言模型的端到端概念体系自动构建方法（E2TCM，End-to-End Taxonomy Construction with Pretrained Language Model），旨在解决现有方法的局限性。E2TCM利用预训练语言模型学习词语对的特征表示，并通过特定的动作（如概念选择、概念剔除、上位词节点选择等）来实现从集合到树结构-“Set to Tree”的映射，这一过程通过强化学习完成。为了更好地利用概念体系的信息，我们将语义特征和结构特征进行融合，得到概念体系的特征向量。同时，我们将该特征向量与待预测上下位关系的词语对特征向量进行联合表示，以提高预测的准确性。借助策略网络，我们可以精确地预测应选择哪个词语以及它在概念体系中的位置，并将反馈传递回特征学习模块来调整预训练语言模型对词语对特征的学习。所有模块都以端到端的方式进行训练，并累积每一步的奖励，直到形成完整的概念体系，并用于更新模型的所有参数。我们在公开数据集上进行了实验验证，结果表明所提出的方法是有效的。

我们的贡献可以总结为以下几点：

1.首次将预训练语言模型的端到端方法应用于概念体系自动构建任务，并且提出一种通用型框架，可有效提升模型的精度。

2.在强化学习动作预测中充分利用了已构建的部分概念体系的特征信息。我们使用可见矩阵将二维的树结构转换为自然语言形式，将概念体系的语义特征和结构信息结合，提取到概念体系的特征向量并有力支撑端到端概念体系自动构建。

3.我们在公开的WordNet数据集进行了实验，结果表明，我们的方法相比经典方法TaxoRL (End-to-End Reinforcement Learning for Automatic Taxonomy Induction) (Mao et al., 2018)有13.4%的绝对提升，相比基于预训练模型的CTP方法 (Chen et al., 2021)也有7.3%的性能提升（F1）。

2 相关工作

在图1中，我们展示了概念体系自动构建任务的流程图。任务包含两部分，首先是从给出的语料库中提取概念词语对，然后将提取到的概念词语集合按照上下位关系，组织成树状的概念体系。概念体系自动构建的方法可以大致分为基于模式的方法和基于分布的方法。

基于模式的方法主要用于从大型语料库中提取可能具有上下位关系的词语对 (Hearst, 1992; Snow et al., 2004; Kozareva and Hovy, 2010; Panchenko et al., 2016; Nakashole et al., 2012)，Hearst模式是指词语对的句法模式，也就是说如果一个句子中的一对词语符合某种形式，就可以认为他们是上下位关系，例如，“是一种”、“是一部分”等表示。有学者 (Snow et al., 2004)提出了一种系统，基于上述模式预测词语之间的关系。分布式方法则考虑词语的上下文，此时不需要词语对在句子中共存，有一些无监督方法 (Chang et al., 2017)和有监督方法 (Luu et al., 2016)。一些研究 (Aldine et al., 2021)使用基于顺序模式挖掘的新型三段式方法，改进了基于模式的方法的低召回率问题。我们主要研究如何将一组概念词语组织成概念体系形式。有学者将任务看作结构化的学习问题 (Bansal et al., 2014)，使用概率模型公式组织概念体系，使用循环信念传播整体考虑节点在树中的兄弟关系。有人将图神经网络应用于概念层次构建任务 (Shang et al., 2020)，用迁移学习的方式，提高了在SemEval-2016 Task 13数据集上构造大型概念体系的性能。

在基于分布的方法中，为了改进词语的特征表示，还有一些方法将词语的表示转换到双曲空间进行 (Aly et al., 2019; Torregrossa et al., 2021; Aly et al., 2019)使用双曲空间对树结构进行编码，得到词语在双曲空间中的表示，改进了 (Panchenko et al., 2016; Tan et al., 2016)等的分类结果，证明了双曲空间更适合描述树结构。此外，CTP模型 (Chen et al., 2021)使用预训练语言模型判断词语对的上下位关系，并且使用计算词语相似度的方法解决一词多义问题，取得了很好的效果；TaxoRL模型 (Mao et al., 2018)使用强化学习实现了端到端的分类体系构建法。

还有一些相关的工作则通过考虑使用概念体系节点的兄弟和祖先关系进行分类扩展 (Wang

et al., 2022; Zhang et al., 2021; Wang et al., 2021; Jiang et al., 2022)。分类扩展任务是在已有概念体系的基础上添加节点，使得概念体系更加完善，以前的分类扩展任务只将新添加的节点作为叶节点 (Wang et al., 2021)，这与我们的做法类似，在训练过程中，我们同样将带添加的节点只作为叶节点添加到已有分类书中，有研究 (Zhang et al., 2021)提出，在扩展概念体系时，新节点不应只作为叶节点，也可以作为父节点插入到已存在的边中。QEN算法 (Wang et al., 2022)在TMN算法的基础上额外考虑了兄弟节点信息。

这些方法都存在一些问题，大部分都是将任务分为两个子任务的做法，而这种做法存在错误累积问题；TaxoRL (Mao et al., 2018)忽略了概念体系的信息，且过度依赖于外部信息，这些都限制了模型的效果。

3 基于预训练语言模型的概念体系自动构建方法

任务定义 我们定义一个概念体系$T = (V, R)$,其中，T表示概念树节点集合，R表示概念树的边的集合。对于树中任意一个节点$v \in V$，v可以是一个单个词语或者是多个词语组成的词组。我们的任务是，将给定的概念集合按照上下位关系组织成树状的概念体系。

为了实现端到端概念体系构建，我们以强化学习技术为核心，联合建模概念上下文关系判断和完整概念树构建两个任务，本小节将具体介绍网络的实现。我们使用预训练语言模型对动作矩阵和概念树进行编码，将二者特征结合后通过两层神经网络得到模型选择每个词语的概率，图2所示的动作过程根据每个词语的概率大小做出选择，最终得到类似图5中的概念树。我们将详细介绍我们如何设计的动作、状态和奖励函数。

3.1 概念树构建动作空间

概念体系构建过程是一系列动作的组合。在每个时间步t中，T_t表示此时已构建的概念树，V_t表示剩余的概念集合，也就是还未加入T_t中的词语，我们定义的动作为：（1）从词汇表V_t中选择一个词语x_1；（2）将x_1从词汇表V_t中移除；（3）将x_1添加到T_t中，作为节点x_2的下位词。动作空间的大小为$|V_t| \times |T_t|$，其中$|V_t|$表示剩余词汇表V_t中词语的数量大小，$|T_t|$表示现有概念体系中节点的大小。回合开始时，概念集合V_0大小等于输入词汇表，T_0被初始化为空集合。在每个时间步t，$|V_t| = |V_{t-1}| - 1, |T_t| = |T_{t-1}| + 1$都始终成立，且$|V_t| + |T_t| = |V_0|$。当一个回合完成后，词汇表中的所有词语都被添加到了概念体系中。

在图2中，我们使用一个例子介绍了上述的动作过程，初始状态为某时刻词汇表与概念体系的状态，词汇表中灰色划线的词语表示已经从词汇表中被删除，我们可以看到这些被删除的词语已经被组织成了右图的概念树；动作1为从词汇表中选择词语“咖啡”，动作2将“咖啡”删去，动作3将“咖啡”连接到“饮料”，作为其子节点，经过一个完整的动作，模型的状态由初始状态转变为一个时间步之后的状态。圆圈中的数字表示词语被选择的顺序。

对于如何向概念体系中添加第一个节点，我们选择在一个回合开始时，随机选择一个词语作为根节点，并且在每一步选择词语时，允许新的词语作为当前根节点的上位词，也就是作为新的根节点。这样缓解了选择第一个节点时缺乏先验知识的问题，动作空间大小变为$|A_t| = |V_t| \times |T_t| + |V_t|$。由于每个一步词汇表的大小都会发生改变，因为在选择添加新的词语时，并非单独选择一个词语，而是列出概念体系T_t 和词汇表V_t中所有词语的可能组合(x_v, x_t)，选择词语对时，就已经确定了新选择的词语在概念体系T_t中的位置。

3.2 概念树构建状态建模

我们将每个时刻的已构建概念树T_t，概念集合V_t及其特征定义为状态。状态作为指导模型的重要因素，其特征的好坏对实验结果有极大影响。下面我们将介绍如何得到概念树的特征及动作矩阵的特征。如图4中所示，模型通过预训练语言模型获得树状概念体系的特征和词语的特征，然后将其输入策略网络中，策略网络通过处理得到的特征选择词语构建概念体系。

动作矩阵 动作矩阵也可以叫做词语对矩阵，其中的每一行都是一个词语对组成的句子“Term1 IS A Term2”（例如，“A cat is a mamal”，表示cat是mamal的下位词），其中Term1一般从词汇表中提取，Term2则从已构建概念树的节点集合中提取。为了得到词语对的特征，我们将动作矩阵输入到预训练语言模型中（如BERT），提取其中的隐藏层向量作为特征。实验证明，不同的句子形式不影响模型的特征提取结果，比如“Term1 a type of Term2”，“Term1 an example of Term2”等。

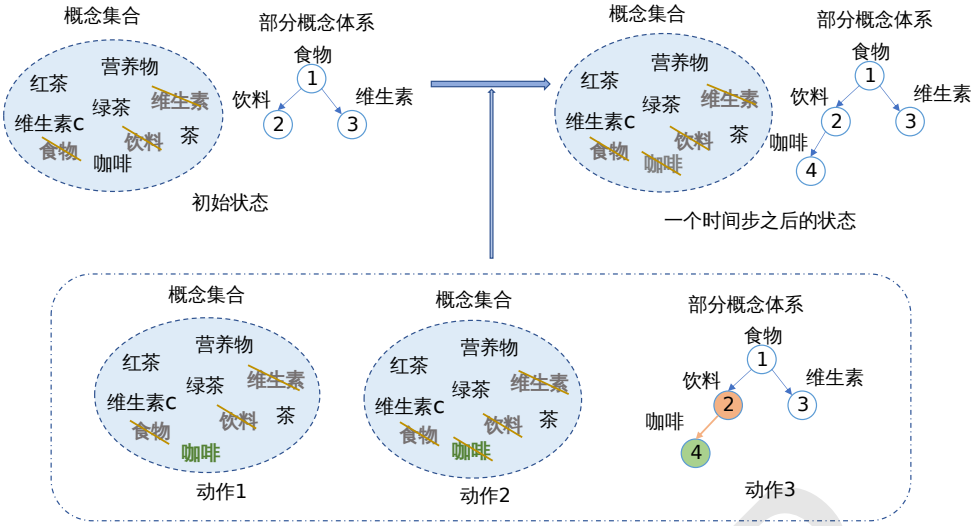


图 2: 每个时间步的概念树状态转变过程

树的特征 使用预训练语言模型可以方便地提取句子的特征，因为句子是一维的。然而，对于树状结构的概念体系，我们很难使用提取词语对特征的方法来提取树的特征。我们甚至很难定义树的特征，因为树状概念体系的特征不仅包括二维结构特征，还包括每个节点本身的语义特征。因此，如何将这两者融合起来是一个棘手的问题。

我们使用可视矩阵的方法 (Liu et al., 2020)，将概念体系的树状结构转换为词语之间能否相互“看见”的状态。我们认为，如果一对词语之间存在上下位关系，那么它们在语义上也一定存在很强的相关性，因此使这样的词语在可视矩阵中相互可见。在图3中展示了一个树状概念体系与其可视矩阵的示意图，我们使得有上下位关系的词语相互可见，因此可视矩阵是对称矩阵，矩阵的横轴和纵轴均表示树状概念体系的所有节点，矩阵(i, j)处的值为1时表示第i个节点可以“看见”第j个节点，值为0时表示“看不见”。通过将可视矩阵与树状概念体系同时输入预训练语言模型中，我们得到了概念体系的语义和结构相融合的特征。

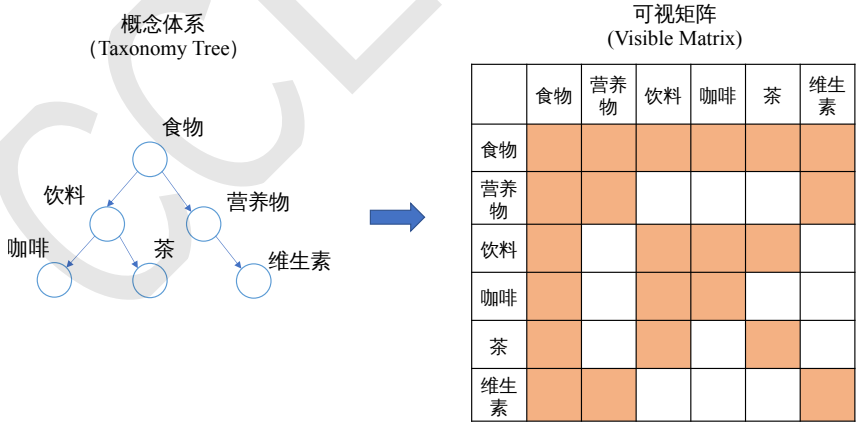


图 3: 概念树状态构建中的可视矩阵

3.3 概念树构建奖励函数

模型为了学习如何选择更好的动作，每进行一次动作都需要给予标量的奖励。一般会将结果的衡量指标作为奖励，也就是最终生成概念体系后，比较预测结果和标签的F1值，将其作为奖励。但是这种方式不能表现每一步动作选择的好坏，具有延迟性。因此我们计算每个动作执行前后概念树的F1值，将此差值作为奖励。也就是$r_t = F1_{e_t} - F1_{e_{t-1}}$。如果当前的F1值比上一次的时间好，奖励就为正，否则就为负。REINFORCE (Williams, 1992)算法的特点就是当完成一个回合后才会更新网络，即中间步骤的奖励不会即刻影响后续动作，而是在完成树的构建

后用所有奖励的集合更新策略网络。这样做的好处是，模型可以整体的考虑生成效果，而不是只考虑当前动作的好坏，某些动作在当下的收益可能一般，但在长远角度看来会带来更大的收益。

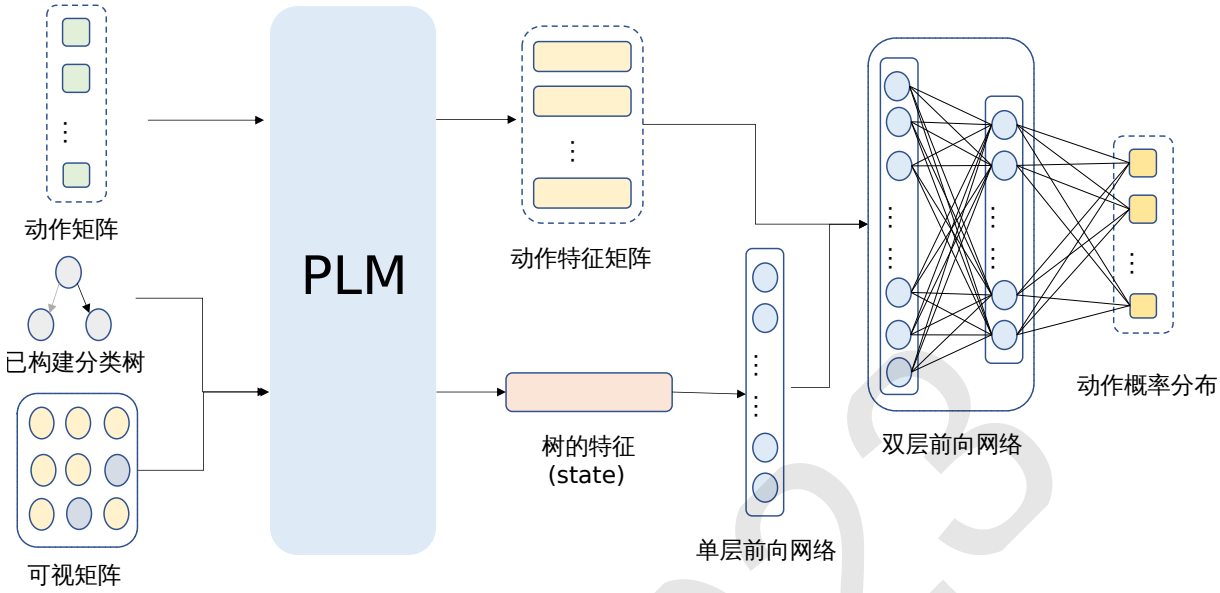


图 4: 概念树构建的策略网络结构图

3.4 概念树构建策略网络及模型训练

定义了动作、状态和奖励之后，我们将介绍如何设计模型的策略网络来选择每一步的动作。网络$\pi(a|S,U)$的作用是根据输入给出选择每一个动作的概率，动作空间大小随着状态的改变而改变，网络输入为动作特征矩阵，其结构如图4所示。结合所有词语对的向量后得到动作矩阵A_t，其尺寸为$||V_t| \times |T_t| + |V_t|| \times dim(P)$，其中$||V_t| \times |T_t| + |V_t||$表示动作维度，也就是所有词语对的数目，$dim(P)$表示词语对的特征向量维度。对于每个时刻得到的概念体系T_t，可视矩阵记为Vm_t，大小为$|T_t| \times |T_t|$。将概念体系T_t的节点与对应的可见矩阵输入到预训练语言模型中，就可以得到概念体系的特征向量S_t，将A_t与S_t一起输入到策略网络中。

策略网络由两部分组成，一个单层的前向传播网络和一个两层的前向传播网络。S_t由单层前向传播网络后，与A_t结合，再经由两层前向传播网络，最后经过softmax层，得到动作的概率分布P_t，大小为$|V_t| \times |T_t|$。模型将根据P_t选择动作也就是词语对，模型具体计算过程如下：

$$S_t = Relu(M_s + b_s)$$

$$\pi(a|S,U) = softmax(M_U^2(Relu(M_U^1(A_t;S_t) + b_U^1)) + b_U^2)$$

$$a_t \sim \pi(a|S,U)$$

其中，M_s和M_U均表示前向传播的网络的参数，U表示策略梯度网络的参数。在训练阶段，我们根据概率分布对动作进行采样，但是在测试阶段则是选择概率最高的动作。

我们选择REINFORCE算法作为优化算法，是策略梯度算法的一种。策略网络参数的更新如下：

$$U = U + \alpha \sum_{t=1}^{T} \bigtriangledown \pi(a_t|S,U) \cdot v_t$$

其中，$v_k = \sum_{t=1}^{T} \gamma^{t-k}$是完成一个回合后的累计奖励，$\gamma \in [0,1]$是折现因子，表示未来对现在的影响。

同时为了减小方差且使得模型更充分的学习每个样本，我们将每个样本训练十次后，将所得奖励的平均值作为基线。

4 实验

4.1 数据集

我们使用从WordNet中提取的数据集，该数据集是由 (Bansal et al., 2014)创建的中等大小的数据集，数据集涵盖各种领域，比如动物，食物，日常用品等等。数据集全部是由高度为4的中等大小的子树组成（将根节点的高度记为1，从根节点出发到达叶节点的最长路径为4），子树大小（子树拥有的所有节点个数）为均在10到50之间，子树之间互相没有重叠。数据集包含761棵树，我们将其按照训练集、验证集和测试集三类划分成533/114/114大小。

4.2 评价指标

Ancestor-F1. 我们沿用之前工作的评价指标 (Bansal et al., 2014)，该指标将预测概念体系(prediction tree)的祖先对与标准概念体系(gold tree)的祖先对进行比较，我们使用P_a, R_a和$F1_a$表示准确率、召回率和F1值:

$$P_a = \frac{|IS - A_{prediction} \cap IS - A_{gold}|}{|IS - A_{prediction}|}$$

$$R_a = \frac{|IS - A_{prediction} \cap IS - A_{gold}|}{|IS - A_{gold}|}$$

则可以得到$F1_a = \frac{2P_aR_a}{P_a+R_a}$，其中$IS - A_{prediction}$和$IS - A_{gold}$分别表示预测的上下位关系和正确的上下位关系。

Edge-F1. 这个评价指标只比较预测概念体系的边（也就是上下位关系的边）和标准概念体系的边。我们使用$P_e, R_e, F1_e$表示，其具体计算方式与上式类似。根据指标的公式可以看出，当预测结果与标准概念体系大小一致时，$P_e = R_e = F1_e$。

我们用$F1_e$计算强化学习的奖励值，使用$F1_a$作为评价指标衡量模型结果。

4.3 实验结果与分析

我们与两个典型模型进行对比：1）TaxoRL，基于端到端的概念体系自动构建方法；2）CTP，基于预训练语言模型的概念体系自动构建方法。表1中展示整体实验结果，对比了本文方法和对比方法。（我们在实验过程中使用的预训练语言模型是bert-base-uncased计算资源有限不能运行更大规模的模型），为了更加客观的比较我们的方法与CTP方法，在这里我们对比的CTP也是使用bert-base-uncased模型。从表中可以看出，在使用同样的预训练语言模型的条件下，我们的方法有最高的F1值，相比CTP有7.2%的相对提升。在CTP中，预训练语言模型用于估计词语之间的上下位关系的概率，并且使用二分类的方式对其进行微调。为了得到每一对词语之间上下位关系的概率，需要先得到所有可能的词语对，对于一个大小为n的集合，其可能的词语对有$\frac{n!}{2}$，而其中正确的仅有$n-1$对，其余均为负样本，存在严重的样本不平衡问题，将会影响预训练语言模型的判断能力，且由于分步式的缺陷，判断的错误后续无法改正。而我们将预训练网络嵌入到端到端模型中，如图4中所示，使用整体的损失微调预训练语言模型，使得预训练网络的错误可以用后续结果的反馈进行修正。为了更合理的比较模型，我们采用TaxoRL未使用额外信息时的结果，我们的方法相比之下有12.4%的绝对提升，

Model	P	R	F1
TaxoRL	41.3	49.2	44.9
CTP(bert-base-uncased)	**57.9**	51.8	53.4
Ours	56.6	**58.0**	**57.3**

表 1: 我们的方法在英文数据集上与以往方法的比较结果

4.4 消融实验

这一小节我们探究概念体系特征对于模型的作用。表2展示了我们在WordNet数据集上进行的消融实验。Glove+RL是使用Glove提取词语对的特征向量（实际上这就是TaxoRL的框架

Model	P	R	F1
Glove+RL	41.3	49.2	44.9
PLM(CTP)	57.9	51.8	53.4
PLM+RL	**58.7**	53.2	55.8
PLM+RL+State	56.6	**58.0**	**57.3**

表 2: 不同模型的结果比较

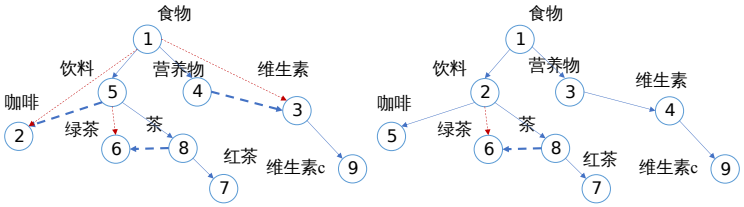
结构），当我们使用PLM作为提取特征向量的网络时，评价结果有了10.9%的极大提升，证明了预训练语言模型对词语特征表达更合理。相比仅使用预训练语言模型，我们将其与端到端方式结合后，评价结果提升2.4%，证明端到端方式可以更好的微调预训练语言模型的对词语的特征表示。当我们使用概念体系的特征时，结果又有1.5%的提升，说明我们定义的概念体系特征中包含有利于分类的结构信息。另外，我们还注意到加入状态特征（state）后，与CTP结果相比，模型的召回率增加但精确率降低。我们认为这可能是因为在使用PLM判断词语之间的上下位关系时，缺乏上下文信息，即概念树的特征信息，导致一些不明显的上下位关系被错误地判断为非上下位关系。然而，当增加状态特征（state）后，模型可以利用词语对的上下文信息来判断它们之间的关系，从而正确地判断一些不明显的上下位关系，因此召回率增加。但同时精确率也下降。这可能是导致该现象发生的原因。

State	P	R	F1
0	**58.7**	53.2	55.8
50	54.9	57.1	56.0
100	56.6	**58.0**	**57.3**
500	54.6	57.3	55.9
768	51.8	54.7	53.2

表 3: 状态维度维度不同时的结果

状态维度分析 我们使用BERT模型提取概念树的原始特征为768维，在图4中可以看到，在融合概念树特征和动作特征前，我们先将其经过了一个非线性的单层前向神经网络，为了探究这一步的效果，我们进行了如表3所示的对比试验。表3的左侧列表指的是前向神经网络输出的维度，0表示不使用该特征，768表示不经过神经网络直接与动作矩阵结合。

可以看出，模型的效果随着网络输出维度的增大呈现先上升后下降的趋势。使用$A_t = [a_1; a_2; ...; a_{|T_t|}]$表示动作特征矩阵，其中$|a_i| = 768$，用$S_t$表示概念树特征。直接将二者结合的结果为$G_t = [a_1, S_t; a_2, S_t; ...; a_i, S_t]$。策略网络决策的主要参考依据是$A_t$，$S_t$仅起到补充作用，维度过大时将严重影响$A_t$的特征表示，因此当其维度为768时，反而降低了模型的效果，而当维度适中时，既可以起到补充信息的作用，又不会影响A_t的原有特征，因此可以有效地提升效果。



(a) 训练初始阶段的构建过程　　(b) 训练后期的构建过程

图 5: 训练过程实例

训练实例 我们在图5中展示了训练过程的实例。图5(a)中，红色虚线为连接错误的边，蓝色虚线为正确的缺失的边，蓝色的实线为预测正确的边。图5(b)中节点的数字表示该节点加入概念树的先后顺序。在图5(a)中，策略网络错误的将“咖啡”节点先于“饮料”节点添加到概念树

中，造成了错误的边，“维生素”和“绿茶”等也同理。这是由于策略网络根据给出的特征认为“咖啡-食物”的概率比“饮料-食物”更高，然而在受到反馈后策略网络和特征提取网络同时调整参数，使得“饮料-食物”的概率更高，在图5(b)中，策略网络调整了“咖啡”与“饮料”的选择顺序，更改了之前的错误，其他的错误节点也同样会在后续得到修改。这就是我们的模型相比分步式方法的优势。

5 总结与展望

在本文中，我们提出了基于预训练语言模型的端到端概念体系构建方法，相比同等条件下的方法取得了更好的结果，且我们的模型为通用模型框架，可以任意更改预训练语言模型。但是由于实际设备限制，目前难以使用大模型。在未来我们将进一步探索更合理的动作设计（比如新的节点不仅可以作为叶节点，也可以插入已有的边中作为其他节点的父节点），如何更加充分的利用已构建概念体系的信息（例如，将节点与兄弟节点和祖先节点的关系纳入考虑）以及设置合理的奖励函数。

参考文献

Aldine, A. I. A., Harzallah, M., Berio, G., Béchet, N., and Faour, A. (2021). A 3-phase approach based on sequential mining and dependency parsing for enhancing hypernym patterns performance. *The Knowledge Engineering Review*, 36:e13.

Aly, R., Acharya, S., Ossa, A., Köhn, A., Biemann, C., and Panchenko, A. (2019). Every child should have parents: A taxonomy refinement algorithm based on hyperbolic term embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4811–4817.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 722–735. Springer.

Bansal, M., Burkett, D., De Melo, G., and Klein, D. (2014). Structured learning for taxonomy induction with belief propagation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1051.

Chang, H.-S., Wang, Z., Vilnis, L., and McCallum, A. (2017). Unsupervised hypernym detection by distributional inclusion vector embedding. *arXiv preprint arXiv:1710.00880*.

Chen, C., Lin, K., and Klein, D. (2021). Constructing taxonomies from pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4687–4700.

Demeester, T., Rocktäschel, T., and Riedel, S. (2016). Lifted rule injection for relation embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1389–1399.

Granot, D. and Huberman, G. (1981). Minimum cost spanning tree games. *Mathematical programming*, 21:1–18.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Jiang, M., Song, X., Zhang, J., and Han, J. (2022). Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations. In *Proceedings of the ACM Web Conference 2022*, pages 925–934.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kozareva, Z. and Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1110–1118.

Kwan, W.-C., Wang, H.-R., Wang, H.-M., and Wong, K.-F. (2023). A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20(3):318–334.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., and Wang, P. (2020). K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Luu, A. T., Tay, Y., Hui, S. C., and Ng, S. K. (2016). Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 403–413.

Mao, Y., Ren, X., Shen, J., Gu, X., and Han, J. (2018). End-to-end reinforcement learning for automatic taxonomy induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2472.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nakashole, N., Weikum, G., and Suchanek, F. (2012). Patty: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1135–1145.

Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, C., Ponzetto, S. P., and Biemann, C. (2016). Taxi at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1320–1327.

Shang, C., Dash, S., Chowdhury, M. F. M., Mihindukulasooriya, N., and Gliozzo, A. (2020). Taxonomy construction of unseen domains via graph-based cross-domain knowledge transfer. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 2198–2208.

Snow, R., Jurafsky, D., and Ng, A. (2004). Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems*, 17.

Sun, T.-X., Liu, X.-Y., Qiu, X.-P., and Huang, X.-J. (2022). Paradigm shift in natural language processing. *Machine Intelligence Research*, 19(3):169–183.

Tan, L., Bond, F., and van Genabith, J. (2016). Usaar at semeval-2016 task 13: Hyponym endocentricity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1303–1309.

Torregrossa, F., Allesiardo, R., Claveau, V., and Gravier, G. (2021). Unsupervised tree extraction in embedding spaces for taxonomy induction. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 302–309.

Wang, S., Zhao, R., Chen, X., Zheng, Y., and Liu, B. (2021). Enquire one's parent and child before decision: Fully exploit hierarchical structure for self-supervised taxonomy expansion. In *Proceedings of the Web Conference 2021*, pages 3291–3304.

Wang, S., Zhao, R., Zheng, Y., and Liu, B. (2022). Qen: Applicable taxonomy completion via evaluating full taxonomic relations. In *Proceedings of the ACM Web Conference 2022*, pages 1008–1017.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32.

Yang, H.-R. and Ni, W. (2022). Continuous-time distributed heavy-ball algorithm for distributed convex optimization over undirected and directed graphs. *Machine Intelligence Research*, 19(1):75–88.

Yang, S., Zou, L., Wang, Z., Yan, J., and Wen, J.-R. (2017). Efficiently answering technical questions—a knowledge graph approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Zhang, J., Song, X., Zeng, Y., Chen, J., Shen, J., Mao, Y., and Li, L. (2021). Taxonomy completion via triplet matching network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4662–4670.

Ask to Understand: Question Generation for Multi-hop Question Answering

Jiawei Li, Mucheng Ren, Yang Gao, Yizhe Yang
School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China
Beijing Engineering Research Center of High Volume Language Information
Processing and Cloud Computing Applications, Beijing, China
{jwli,renm,gyang,yizheyang}@bit.edu.cn

Abstract

Multi-hop Question Answering (QA) requires the machine to answer complex questions by finding scattering clues and reasoning from multiple documents. Graph Network (GN) and Question Decomposition (QD) are two common approaches at present. The former uses the "black-box" reasoning process to capture the potential relationship between entities and sentences, thus achieving good performance. At the same time, the latter provides a clear reasoning logical route by decomposing multi-hop questions into simple single-hop sub-questions. In this paper, we propose a novel method to complete multi-hop QA from the perspective of Question Generation (QG). Specifically, we carefully design an end-to-end QG module on the basis of a classical QA module, which could help the model understand the context by asking inherently logical sub-questions, thus inheriting interpretability from the QD-based method and showing superior performance. Experiments on the HotpotQA dataset demonstrate that the effectiveness of our proposed QG module, human evaluation further clarifies its interpretability quantitatively, and thorough analysis shows that the QG module could generate better sub-questions than QD methods in terms of fluency, consistency, and diversity.

1 Intorduction

Unlike single-hop QA (Rajpurkar et al., 2016; Trischler et al., 2017; Lai et al., 2017) where the answers could usually be derived from a single paragraph or sentence, multi-hop QA (Welbl et al., 2018; Yang et al., 2018) is a challenging task that requires soliciting hidden information from scattered documents on different granularity levels and reasoning over it in an explainable way.

The HotpotQA (Yang et al., 2018) was published to leverage the research attentions on reasoning processing and explainable predictions. Figure 1 shows an example from HotpotQA, where the question requires first finding the name of the company (Tata Consultancy Services) and then the address of the company (Mumbai). While, a popular stream of Graph Network-based (GN) approaches (De Cao et al., 2019; Tu et al., 2019; Ding et al., 2019; Fang et al., 2020) was proposed due to the structures of scattered evidence could be captured by the graphs and reflected in the representing vectors. However, the reasoning process of the GN-based method is entirely different from human thoughts. Specifically, GN tries to figure out the underlying relations between the key entities or sentences from the context. However, the process is a "black-box"; we do not know which nodes in the network are involved in reasoning for the final answer, thus showing relatively poor interpretability.

Inspired by that human solves such questions by following a transparent and explainable logical route, another popular stream of Question Decomposition-based (QD) approaches became favored in recent years (Fu et al., 2021; Nishida et al., 2019; Min et al., 2019; Jiang and Bansal, 2019b). The method mimics human reasoning to decompose complex questions into simpler, single-hop sub-questions; thus, the interpretability is greatly improved by exposing intermediate evidence generated by each sub-question. Nevertheless, the general performance is usually much worse than GN-based ones due to error accumulation that arose by aggregating answers from each single-hop reasoning process. Furthermore, the sub-questions are generated mainly by extracting text spans from the original question to fill the template.

Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

Hence the sub-questions are challenging to guarantee in terms of quality, such as uency, diversity, and consistency with the original question intention, especially when the original questions are linguistically complex.

In this work, we believe that asking the question is an effective way to elicit intrinsic information in the text and is an inherent step towards understanding it (Pyatkin et al., 2021). Thus, we propose resolving these dif culties by introducing an additional QG task to teach the model to ask questions. Speci cally, we carefully design and add one end-to-end QG module based on the classical GN-based module. Unlike the traditional QD-based methods that only rely on information brought by the question, our proposed QG module could generate uent and inherently logical sub-questions based on the understanding of the original context and the question simultaneously.

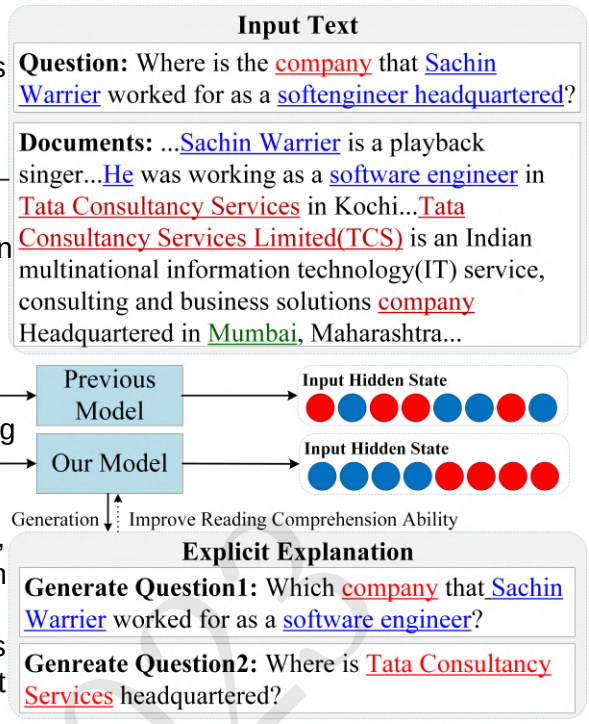


Figure 1: An example from HotpotQA dataset. Text in blue is the rst-hop information and text in red is the second-hop information. The mixed encoding of the rst-hop information ( ) and the second-hop information ( ) will confuse models with weaker reading comprehension.

Our method enjoys three advantages: First, it achieves better performance. Our approach preserves the GN module, which could collect information scattered throughout the documents and allows the model to understand the context in depth by asking questions. Moreover, the end-to-end training avoids the error accumulation issue; Second, it brings better interpretability because explainable evidence for its decision making could be provided in the form of sub-questions; Thirdly, the proposed QG module has better generalization capability. Theoretically, it can be plugged and played on most traditional QA models.

Experimental results on the HotpotQA dataset demonstrate the effectiveness of our proposed approach. It surpasses the GN-based model and QD-based model by a large margin. Furthermore, robust performance on the noisy version of HotpotQA proves that the QG module could alleviate the shortcut issue, and visualization on sentence-level attention indicates a clear improvement in natural language understanding capability. Moreover, a human evaluation is innovatively introduced to quantify improvements in interpretability. Finally, exploration on generated sub-questions clari es diversity, uency, and consistency.

2 Related Work

Multi-hop QA In multi-hop QA, the evidence for reasoning answers is scattered across multiple sentences. Initially, researchers still adopted the ideas of single-hop QA to solve multi-hop QA (Dhingra et al., 2018; Zhong et al., 2019). Then the graph neural network that builds graphs based on entities was introduced to multi-hop QA tasks and achieved astonishing performance (De Cao et al., 2019; Tu et al., 2019; Ding et al., 2019). While, some researchers paid much attention to the interpretability of the coreference reasoning chains (Fu et al., 2021; Nishida et al., 2019; Min et al., 2019; Jiang and Bansal, 2019b). By providing decomposed single-hop sub-questions, the QD-based method makes the model decisions explainable.

Interpretability Analysis in NLP An increasing body of work has been devoted to interpreting neural network models in NLP in recent years. These efforts could be roughly divided into structural analyses,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569–582, Harbin, China, August 3 – 5, 2023.

behavioral studies, and interactive visualization (Belinkov and Glass, 2019).

Firstly, the typical way of structural analysis is to design probe classi ers to analyze model characteristics, such as syntactic structural features (Elazar et al., 2021) and semantic features (Wu et al., 2021). Secondly, the main idea of behavioral studies is that design experiments that allow researchers to make inferences about computed representations based on the model's behavior, such as proposing various challenge sets that aim to cover speci c, diverse phenomena, like systematicity exhaustivity (Gardner et al., 2020; Ravichander et al., 2021). Thirdly, for interactive visualization, neuron activation (Durrani et al., 2020), attention mechanisms (Hao et al., 2020) and saliency measures (Janizek et al., 2021) are three main standard visualization methods.

Question Generation QG is the task of generating a series of questions related to the given contextual information. Previous works on QG focus on rule-based approaches. Fabbri et al. (2020) used a template-based approach to complete sentence extraction and QG in an unsupervised manner. Dhole and Manning (2021) developed Syn-QG using a rule-based approach. The system consists of serialized rule modules that transform input documents into QA pairs and use reverse translation counting, resulting in highly uent and relevant results. One of the essential applications of QG is to construct pseudo-datasets for QA tasks, thereby assisting in improving their performance (Zhang and Bansal, 2019; Alberti et al., 2019; Lee et al., 2020).

Our work is most related to Pyatkin et al. (2021), which produces a set of questions asking about all possible semantic roles to bring the bene ts of QA-based representations to traditional SRL and information extraction tasks. However, we innovatively leverage QG into complicated multi-hop QA tasks and enrich representations by asking questions at each reasoning step.

3 Methods

Multi-hop QA is challenging because it requires a model to aggregate scattered evidence across multiple documents to predict the correct answer. Probably, the nal answer is obtained conditioned on the rst sub-question is correctly answered. Inspired by humans who always decompose complex questions into single-hop questions, our task is to automatically produce naturally-phrased sub-questions asking about every reasoning step given the original question and a passage. Following the reasoning processing, the generated sub-questions further explain why the answer is predicted. For instance, in Figure 1, the answer 'Mumbai' is predicted to answer Question 2 which is conditioned on Question 1's answer. More importantly, we believe that the better questions the model asks, the better it understands the reading passage and boosts the performance of the QA model in return.

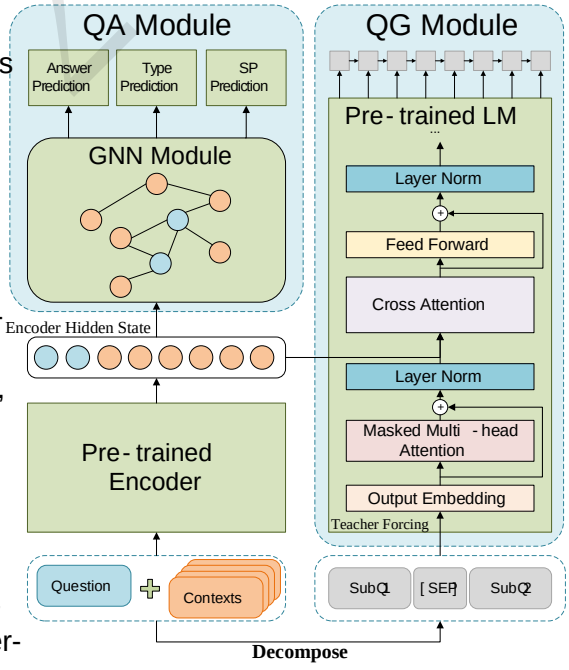


Figure 2: Overall model architecture.

Figure 2 illustrates the overall framework of our proposed model. It consists of two modules: QA module (Section §3.1) and QG module (Section §3.2). The QA module could help model to solve multi-hop QA in a traditional way, and the QG module allows the model to solve the question in an interpretable manner by asking questions. These two modules share the same encoder and are trained end-to-end with multi-task strategy.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

3.1 Question Answering Module

Encoder A key point of the GN-based approach to solving QA problems is the initial encoding of entity nodes. Prior studies have shown that pre-trained models are bene cial for increasing the comprehension of the model (Yang et al., 2019; Qiu et al., 2020), which enables better encoding of the input text. In Section 3.2 we will mention that encoder will be shared to the QG module to further increase the model's reading comprehension of the input text through the QG task. Here we chose BERT as the encoder considering its strong performance and simplicity.

GNN Encode Module The representation ability of the model will directly affect the performance of QA. Recent works leverage graphs to represent the relationship between entities or sentences, which have strong representation ability (Xiao et al., 2019; Tu et al., 2020; Fang et al., 2020). We believe that the advantage of graph neural networks is essential for solving multi-hop questions. Thus, we adopt the GN-based model DFGN[1] (Xiao et al., 2019) that has been proven to be effective in HotpotQA.

Xiao et al. (2019) build graph edges between two entities if they co-exist in one single sentence. After encoding the question Q and context C by the pre-trained encoder, DFGN extracts the entities' representation from the encoder output by their location information. Both mean-pooling and max-pooling are used to represent the entities' embeddings. Then, a graph neural network propagates node information to its neighbors. A soft mask mechanism is used to calculate the relevance score between each entity and the question in this process. The soft mask score is used as the weight value of each entity to indicate its importance in the graph neural network computation. At each step, the query embedding should be updated by the entities embedding of the current step by a bi-attention network (Seo et al., 2018). The entities embeddings in the t-th reasoning step:

$$E^t = \mathrm{GAT}([m_1^{t-1}e_1^{t-1}; m_2^{t-1}e_2^{t-1}; \ldots; m_n^{t-1}e_n^{t-1}]); \quad (1)$$

where e_i^{t-1} is the i-th entity's embedding at the $(t-1)$-th step and e_i^0 is the i-th entity's embedding produced both mean-pooling and max-pooling results from encoder output according to its position. m_i^{t-1} is the relevance score, which is also called soft mask score in previous, between i-th entity and the question at the $(t-1)$-th step calculated by an attention network. GAT is graph attention networks proposed by Veličković et al. (2017).

In each reasoning step, every entity node gains some information from its neighbors. An LSTM layer is then used to produce the context representation:

$$C^t = \mathrm{LSTM}([C^{t-1}; ME^t]); \quad (2)$$

where M is the adjacency matrix which records the location information of the entities.

The updated context representations are used for different sub-tasks: (i) answer type prediction; (ii) answer start position and answer end position; (iii) extract support facts prediction. All three tasks are jointly performed through multitasking learning.

$$L_{qa} = \lambda_1 L_{start} + \lambda_2 L_{end} + \lambda_3 L_{type} + \lambda_4 L_{para}; \quad (3)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyper-parameters.[2]

3.2 Question Generation Module

Question Generation Training Dataset A key challenge of training the QG module is that it is challenging to obtain the annotated sub-questions dataset. To achieve this, we take the following steps to generate sub-question dataset automatically:

First of all, according to the annotations provided by the HotpotQA dataset, the questions in the training set could be classi ed into the following two types: *Bridge* (70%) and *Comparison* (30%), where

[1] QA module is not the main focus of this work, and DFGN is one of the representative off-the-shelf QA models. In fact, any QA model could be adopted to replace it.

[2] In our experiments, we set $\lambda_1 = \lambda_2 = \lambda_3 = 1, \lambda_4 = 5$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569–582, Harbin, China, August 3 – 5, 2023.

the former one requires finding evidence from first hop reasoning then use it to find second-hop evidence, while the latter requires comparing the property of two different entities mentioned in the question.

Then we leverage the methods proposed by Min et al. (2019) to process these two types respectively. Specifically, we adopt an off-the-shelf span prediction tool Pointer to map the question into several points, which could be for segmenting the question into various text spans.

Finally, we generated sub-questions by considering the type of questions and index points provided by Pointer. Concretely, for Bridge questions like Kristine Moore Gebbie is a professor at a university founded in what year?, Pointer could divided the question into two parts: Kristin Moore Gebbie be a professor at a university and founded in what year?. Then some question words are inserted into the first part as the first-hop evidence like Kristin Moore Gebbie be a professor at which university?, denoted as S^A. Afterward, an off-the-shelf single QA model is used to find the answer for the first sub-question, and the answer would be used to form the second sub-question like Flinders University founded in what year?, denoted as S^B. On the other hand, for Comparison questions like Do The Importance of Being Icelandic and The Five Obstructions belong to different film genres?, Pointer would divide it into three parts: first entity (The Importance of Being Icelandic), second entity (The Five Obstructions), and target property (film genre). Then two sub-questions could be further generated by inserting question words to these parts like S^A : Do The Importance of Being Icelandic belong to which film genres? S^B : Do The Five Obstructions belong to which film genres?

Pre-trained Language Model (LM) as Generator After automatically creating the sub-question dataset, the next step is to train the QG module from scratch. Specifically, the structure of whole QG module is designed as seq2seq, where it shares the encoder with QA module and adopts GPT-2 (Radford et al., 2019) as the decoder. During training stage, the input of decoder is formed as: $[bos; y^A_1; y^A_2; :::; y^A_n; [SEP]; y^B_1; y^B_2; :::; y^B_n; eos]$, where [SEP] is the separator token, bos is the start token and eos is the end token. y^A_i and y^B_i are the i-th token in constructed sub-questions S^A and S^B respectively.

Then the training objective of the QG module is to maximize the conditional probability of the target sub-questions sequence as follows:

$$L_{qg} = \sum_{i=1}^{n} \log P(y_t | y_{<t}; h); \quad (4)$$

where h is encoder hidden state. Finally, QG module and QA module are trained together in end-to-end multi-task manner, and the overall loss is defined as:

$$L_{multitask} = L_{qa} + L_{qg}. \quad (5)$$

4 Experiments

4.1 Dataset

We evaluate our approach on HotpotQA (Yang et al., 2018) under the distraction setting, a popular multi-hop QA dataset taking the explanation ability of models into accounts. Expressly, for each question, two gold paragraphs with ground-truth answers and supporting facts are provided, along with 8 `distractor' paragraphs that were collected via bi-gram TF-IDF retriever (i.e., 10 paragraphs in total). Furthermore, HotpotQA contains two types of subtasks: a) Answer prediction; and b) Supporting facts prediction; both subtasks adopt the same evaluation metrics: Exact Match (EM) and Partial Match (F1).

4.2 Implementation Details

We implement the model via HuggingFace library (Wolf et al., 2020). In detail, DFGN is selected as a QA module by following the details provided by (Xiao et al., 2019). While, for the QG module, the pre-trained decoder language model is initialized with GPT2 (Radford et al., 2019). The number of shared encoder layers is set as 12, the number of decoder layers is 6, the maximum sequence length is 512. We train the model on four TITAN RTX GPUs for 30 epochs at a batch size of 8, where each epoch tasks for

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

Model	Answer		Sup Fact		Joint	
	EM	F1	EM	F1	EM	F1
Baseline Model	44.44	58.28	21.95	66.66	11.56	40.86
DecompRC	55.20	69.63	-	-	-	-
DFGN* (Bridge)	53.38	69.14	47.72	84.44	29.79	58.67
DFGN* (Comparison)	63.75	69.48	70.68	89.98	46.74	63.56
DFGN* (Total)	55.46	69.21	52.33	82.12	33.19	59.66
DFGN (Total)	55.66	69.34	53.10	82.24	33.68	59.86
Ours (Bridge)	56.24	71.67	51.06	81.16	33.61	61.75
Ours (Comparison)	63.08	69.59	73.03	90.36	49.23	64.45
Ours (Total)	57.79	71.36	55.77	83.33	36.99	62.52

Table 1: Performance comparison on the development set of HotpotQA in the distractor setting. * indicates the results implemented by us.

around 2 hours. We select Adam (Kingma and Ba, 2017) as our optimizer with a learning rate of 5e-5 and a warm-up ratio of 10%. In general, we determine the hyperparameters by comparing the nal EM and F1 scores.

4.3 Comparison Models

Baseline Model A neural paragraph-level QA model introduced in Yang et al. (2018) and original proposed by Clark and Gardner (2018).

DFGN The classic GN-based model (Xiao et al., 2019) , which is trained in an end-to-end fashion for multi-hop QA task. We select this as the primary QA module in our approach, and reproduce the DFGN model by using the BERT-base pre-trained model under the hyperparameter settings released by Yang et al. (2018).

DecompRC The classic QD-based model that decomposes each question into several sub-questions (Min et al., 2019). We reproduce the DecompRC model by following the same QD instruction illustrated in Min et al. (2019).

5 Analysis

Table 1 shows the performance of various models on the development set of HotpotQA. In general, our method attains substantial improvement across all tasks when compared to either the GN-based method or the QD-based approach. This demonstrates that the integration of the QG task can effectively augment the model's textual understanding capabilities. Additionally, our method exhibits consistent enhancement in performance for both types of questions. Notably, the performance on bridge-type questions, which necessitate linear reasoning chains, experiences a marked improvement, underscoring the ef cacy of posing questions at each reasoning stage. In subsequent sections, we will further explore the functionality, interpretability, and quality of the sub-questions generated by the QG module, providing a comprehensive analysis of our proposed method's strengths and potential applications.

5.1 Does it alleviate shortcut problem by adding question generation module?

In order to validate the capacity of the QG module to concentrate on uncovering the authentic reasoning process, as opposed to exploiting shortcuts for predicting answers, we further undertake QA tasks using baselines and our model on Adversarial MultiHopQA. This dataset was initially introduced by Jiang and Bansal (2019a) and is designed to challenge the model's comprehension capabilities. Speci cally, multiple noisy facts, constructed by substituting entities within the reasoning chain, are incorporated into the original HotpotQA dataset with the intent to confound the model. For instance, in the example provided in Figure 3, the noisy facts are formulated by replacing key entities present in Support Fact2.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

Figure 3: An example of the noisy dataset. The red text indicates a reasoning path with complete reasoning logic. The blue text indicates some other entities which have a similar structure with the red texts, but they can be inferred from the logical relationships.

These noisy facts retain the same sentence structure as the support facts but convey disparate meanings, thereby compelling the model to thoroughly comprehend the context. This additional layer of complexity serves to rigorously test our proposed QG module, ensuring it remains focused on elucidating the genuine reasoning process.

Table 2 shows the performance between the DFGN model and our model on the Adversarial-MultiHopQA dataset. In general, DFGN experiences a signi cant decline in performance, indicating that the existing QA model has poor robustness and is vulnerable to adversarial attacks. This further indicates that the model solves questions by mostly remembering patterns. On the other hand, by adding a QG module, the performance degradation of our method is signi cantly reduced. We think this is mainly because asking questions is an important strategy for guiding the model to understand the text.

Model	Answer	
	EM	F1
DFGN	55.66	69.34
DFGN*	48.08(-13.62%)	61.28(-11.62%)
Ours	57.79	71.36
Ours*	52.34(-9.43%)	65.12(-8.74%)

Table 2: Performance of DFGN model and ours on HotpotQA dataset and its noisy version Adversarial-MultiHopQA (marked with *).

We further prove this point through a case study shown in Figure 3. To answer the original question, the correct reasoning chain is: 2014 S/S → WINNER → YG Entertainment. However, when there exists an overlap in the context between facts (i.e. South Korean boy group), the current main-stream method, which strengthens representation by solely capturing internal relationships over entities or documents, usually regards the incorrect entity (i.e. YG Arthur or YG Republic) as a key node of reasoning chain, where so-called shortcut issue. It does not understand the reasoning process but remembers certain context patterns. However, our method mitigates such issues by reinforcing representations by asking a question at each reasoning step. As such, it could remain robust despite these disturbances.

5.2 Does generated sub-question provide better interpretability?

Past works have proved that interpretability can be improved by exposing evidence from decomposed sub-questions. However, few quantitative analyses have been carried out on interpretability due to its subjective nature. In this paper, we use human evaluation to quantify the improvement of interpretability brought by our QG module.

Speci cally, we design human evaluation by following steps: First, we assemble 16 well-educated vol-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

Indicators	Methods	Win	Tie	Loss
Diversity	QG vs. QD	57.64%	26.70%	15.66%
LM Score	QG vs. QD	60.22%	-	39.78%
Attention weight	QG vs. w/o QG	79.51%	-	20.49%

Table 4: Comparison between sub-questions generated by QG and template on diversity, LM score and Attention weights.

unteers and divide them into two groups, A and B; Second, we randomly sample 8 Bridge type questions from the dev set and manually write out the correct two-hop reasoning chain for solving each question. Afterward, we replace the entity that appeared in each correct reasoning chain with other confusing entities selected from context to generate three more wrong reasoning chains (i.e., each question has 4 reasoning chains.). Then shuf e them and combine them with the original question to form a four-way multi-choice QA; Third, for each group, we ask them to gure out the correct reasoning chain and record the time elapsed for nishing all questions. To be noticed, besides original questions and reasoning chains, we provide different additional information for each group to facilitate them, all supporting facts for Group A, and all sub-questions generated by our QG for Group B. For more details, please refer to Appendix.

Table 3 presents the results of the two groups. Remarkably, Group B has higher accuracy and takes less time. Therefore, we could argue that sub-questions generated by our QG contain more concise and precise explanations for problem-solving and further proves that the QG module can indeed improve interpretability.

Group	Accuracy	Time(s)
A (Support Facts)	65.63%	981
B (Sub-questions)	85.94%	543

Table 3: Average results for accuracy and time elapsed of human evaluation.

5.3 Does asking questions enhance the natural language understanding capability?

In this work, we believe that the ability to exhaustively generate a set of logical questions according to a complex scenario allows for a comprehensive, interpretable, and exible way of excavating the information hidden in natural language text, thereby enhancing the natural language understanding ability.

The self-attention mechanism in the pre-trained model is crucial for the model to understand the input information. Generally, the more critical a sentence is in its context, the greater attention weights it deserves. Thus, to verify whether the QG module could edify the model to carry out deep understanding intrinsically, we compare the sentence-level attention weight of our model with and without the QG module. In particular, we account for the number of increases in attention weight of support facts after adding the QG module. As shown in the last row of Table 4, the attention weight of around 80% of support facts is increased, which proves that the model is more prone to focus on meaningful information with the aid of the QG task.

Furthermore, Figure 4 visualizes the changes in attention weights over supporting facts between DFGN and our method. In this case, sentences $S_{2,5,6,7}$ are considered as supporting facts. DFGN fails to predict all supporting facts and focuses on the wrong ones while our method works properly.

5.4 Characteristics of Generated Questions

QG can indeed promote an in-depth understanding of the model, but what are the characteristics of the generated questions that contribute to this? Speci cally, what are the distinctive features of the sub-questions we generate using the QG module compared to the previous QD-based methods, which generate sub-questions using templates. Through case and statistical analysis, we nd that the sub-questions generated by the QG module exhibit the following characteristics:

Consistency As mentioned in Section 3.2, prior QD-based methods necessitate the implementation of a span predictor to dissect questions into constituent text spans. During the segmentation process, errors are predisposed to accumulate, rendering the generated sub-questions susceptible to inconsisten-

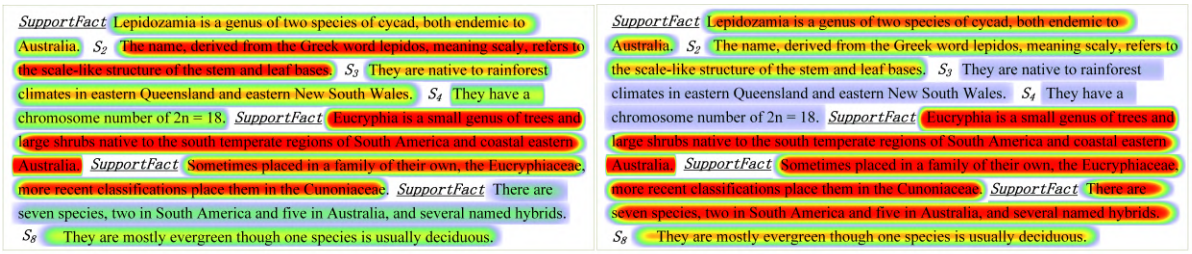


Figure 4: Visualization of attention weights at sentence-level between DFGN and our method. The depth of the color corresponds to the higher attention weights of the sentence.

ID			Question / Sub-question	Fluency	Diversity
1	Question		In 1991 Euromarche was bought by a chain that operated how any hypermarkets at the end of 2016?		
	QD	Q1	Which chain that operated how any hypermarkets?	×	×
		Q2	In 1991 Euromarche was bought by Euromarche at the end of 2016?		
	QG	Q1	In 1991 Euromarche was bought by which chain?	X	X
		Q2	Carrefour's oprated how many hypermarkets at the end of 2016?		
2	Question		Do The Importance of Being Icelandic and The Five Obstructions belong to different lm genres?		
	QD	Q1	Do the Importance of Being Icelandic and The Five Obstructions belong to different lm genres?	×	×
		Q2	Do the importance of?		
	QG	Q1	Does the Importance of Being Icelandic and The Five Obstructions belong to which lm genres?	X	X
		Q2	Does The Five Obstructions belong to which lm genres?		
.....					
7404	Question		Who was known by his stage name Aladin and helped organizations improve their performance as a consultant?		
	QD	Q1	Who was known by his stage name Aladin?	X	×
		Q2	Who helped organizations improve their performance as a consultant?		
	QG	Q1	His stage name Aladdin?	×	X
		Q2	Who was known by his stage name Aladdin and helped organizations improve their performance as a consultant?		
7405	Question		Which American lm actor and dancer starred in the 1945 lm Johnny Angel?		
	QD	Q1	Which 1945 le Johnny Angel?	×	-
		Q2	Which American lm actor and dancer starred in noir?		
	QG	Q1	Which American le actor and dancer?	X	-
		Q2	Which starred in the 1945 lm Johnny Angel?		

Table 5: Results on linguistic uency and diversity of sub-questions generated by QG compared to those generated by template. X indicate the method performs better, × indicate performs worse, and - indicate performs competitively.

cies with the original question. This issue becomes increasingly prevalent when the original question exhibits linguistic complexity. As illustrated by the second example in Table 5, the pair of sub-questions generated by template-based approaches erroneously deconstruct the original question, culminating in a question intention that deviates signi cantly from the original intent. Consequently, such sub-questions characterized by incongruent intent can mislead the model. In contrast, our proposed QG module is designed to facilitate a comprehensive understanding of the original question, utilizing abundant contextual information to generate logically ordered sub-questions. Ultimately, this approach ensures that the intentions of the combined sub-questions remain consistent with the original question, mitigating the risk of misinterpretation by the model.

Fluency The uidity and grammatical integrity of a sentence play a crucial role in accurately conveying meaning, particularly in the case of questions. When a question is plagued by grammatical inaccuracies or incoherence, it becomes challenging for individuals or computational models to comprehend, potentially leading to misinterpretation of the intended inquiry. This issue is widespread and inescapable in numerous datasets, primarily due to the manual construction of questions, as exempli ed by the rst instance in Table 5. In the original question, a typographical error (how many! how any) causes a shift in the intended meaning. Nonetheless, it remains feasible to discern the correct response from the additional information offered by the original question and general knowledge. Regrettably, the sub-question produced by the QD-based technique incorporates the typographical error, and the model fails to ascertain the accurate intention due to the limited information available within the sub-question. Moreover, syntactic errors are prone to accrue since determining the boundaries and attributes of text spans proves

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

to be a challenging task, leading to subpar readability.

Contrastingly, our QG module is capable of leveraging contextual information and the embedded knowledge within the language model to rectify typographical errors. Simultaneously, it can employ the capabilities of the pre-trained language model to generate coherent sentences, thus alleviating the impact of syntactic errors. To assess uency, we utilize the Language Model Score[3] (LMS) metric. As demonstrated in Table 4, over 60% of the questions generated by QG modules exhibit higher scores compared to those produced by the QD method.

Diversity Sultan et al. (2020) highlight that the diversity of generated questions can directly impact QA performance. However, sub-questions produced by QD methods tend to be monotonous and laborious due to constraints on vocabulary and templates. In contrast, our proposed QG module can gently mitigate these challenges and enhance question diversity. Relying on the pretrained LM, the QG module is capable of incorporating contextually appropriate words into sub-questions, adapting to various situations. This is exempli ed by the inclusion of Carrefour in the rst example provided in Table 5, which results in more diverse and rational sub-questions. In our analysis, we consider the number of words in sub-questions that did not appear in the original question as a measure of diversity. As demonstrated in Table 4, approximately 57% of sub-questions generated by our method exhibit greater diversity, underlining the advantages of our proposed QG module.

6 Conclusion

In this paper, drawing inspiration from human cognitive behavior, we posit that the act of asking questions serves as a crucial indicator for determining whether a model genuinely comprehends the input text. Consequently, we introduce a QG module designed to tackle multi-hop QA tasks in an interpretable manner. Building upon traditional QA modules, the incorporation of the QG module effectively enhances natural language understanding capabilities, delivering superior and robust performance through the process of asking questions. Furthermore, we conduct a quantitative analysis of interpretability, as provided by sub-questions, utilizing human evaluation and elucidating interpretability through attention visualization. Ultimately, we substantiate that the sub-questions derived from the QG method surpass those obtained via the QD method in terms of linguistic uency, consistency, and diversity, underscoring the bene ts of our proposed approach.

7 Limitations

Although our research presents numerous advantages, certain limitations persist. The lack of comparison with extant SOTA methods and validation on alternative datasets constitute two principal shortcomings. Despite these issues, we maintain our advocacy for the "ask to understand" concept, positing that the integration of a QG task can bolster a model's interpretability and comprehension capabilities.

Primarily, the rationale behind our decision not to utilize top SOTA models as baselines is that these approaches often entail the application of meticulously designed, task-speci c, and labor-intensive GNN to the encoder segment. Conversely, we posit that our method operates in a plug-and-play manner; validating its ef cacy on two rudimentary baselines suggests that it may also be applicable to other models. Consequently, outperforming SOTA methods in terms of performance is not the central contribution of this paper.Additionally, question decomposition serves as a vital component of our work, and we employ DecompRC to parse multi-hop questions into single-hop queries. Since DecompRC is tailored specifically for HotpotQA, adapting it to other multi-hop QA datasets may not yield the anticipated results; thus, we solely verify our methods on HotpotQA.

Finally, grounded in our core concept of "asking to understand," the applicability and reliability of the QA model in industrial contexts are signi cantly enhanced. Our model delivers answers accompanied by comprehensive multi-hop questions, enabling agents to evaluate the accuracy of the response. Furthermore, our model aids agents in "understanding by asking," delineating the steps involved in obtaining the answer and facilitating a more profound comprehension of the information's origin.

[3]https://github.com/simonepri/lm-scorer

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. arXiv preprint arXiv:1906.05416.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7:49–72.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In Proceedings of the 56th ACL (Volume 1: Long Papers), pages 845–855.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In NAACL, pages 2306–2317.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 42–48.

Kaustubh D. Dhole and Christopher D. Manning. 2021. Syn-qg: Syntactic and shallow semantic rules for question generation.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In ACL, pages 2694–2703.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. arXiv preprint arXiv:2010.02695.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics, 9:160–175.

Alexander R. Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing complex questions makes multi-hop qa easier and more interpretable. In EMNLP, pages 169–180.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. arXiv preprint arXiv:2004.02709.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Self-attention attribution: Interpreting information interactions inside transformer. arXiv preprint arXiv:2004.11207.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. Journal of Machine Learning Research, 22(104):1–54.

Yichen Jiang and Mohit Bansal. 2019a. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. In ACL.

Yichen Jiang and Mohit Bansal. 2019b. Self-assembling modular networks for interpretable multi-hop reasoning. In EMNLP-IJCNLP, pages 4474–4484.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. ACL, pages 785–794.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent qa pairs from contexts with information-maximizing hierarchical conditional vaes. arXiv preprint arXiv:2005.13837.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. ACL.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Multi-task learning for multi-hop qa with evidence extraction.

Valentina Pyatkin, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. In Proceedings of the 2021 Conference on EMNLP, pages 1429–1441.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63:1872–1897, Sep.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In EMNLP, pages 2383–2392.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. Noiseqa: Challenge set evaluation for user-centric question answering. arXiv preprint arXiv:2102.08345.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Bidirectional attention ow for machine comprehension.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In ACL, July.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 191–200, August.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In ACL, pages 2704–2713.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Interpretable multi-hop reading comprehension over multiple documents.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. arXiv preprint arXiv:1710.10903.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. TACL, 6:287–302.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Zhaofeng Wu, Hao Peng, and Noah A Smith. 2021. Infusing netuning with semantic dependencies. Transactions of the Association for Computational Linguistics, 9:226–242.

Yunxuan Xiao, Yanru Qu, Lin Qiu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In EMNLP, pages 2369–2380.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2346–2357, Florence, Italy, July. Association for Computational Linguistics.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. arXiv preprint arXiv:1909.06356.

Victor Zhong, Caiming Xiong, Nitish Shirish Keskar, and Richard Socher. 2019. Coarse-grain ne-grain coattention network for multi-evidence question answering. arXiv preprint arXiv:1901.00603.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

A Appedix: Human Evaluation Instruction

Speci cally, we design human evaluation by following steps:

1. We assemble 16 well-educated volunteers and randomly divide them into two groups, A and B. Each group contains 8 volunteers and evenly gender.
2. We randomly sample 8 Bridge type[4] questions from the dev set, and manually write out the correct two-hop reasoning chain for solving each question.
3. We replace the entity that appeared in each correct reasoning chain with other confusing entities selected from context to generate three more wrong reasoning chains (i.e., each question has 4 reasoning chains.), then shuf e them and combine them with the original question to form a four-way multi-choice QA.
4. For group A, except the original question, nal answer and four reasoning chains, we also provide supporting facts. Then volunteers are asked to nd the correct reasoning chain.
5. For group B, except the original question, nal answer and four reasoning chains, we also provide the sub-questions generated by our QG module. Then volunteers are asked to nd the correct reasoning chain.
6. We count the accuracy and time elapsed for solving problem.

Beyond that, some details are worth noting:

The volunteers participated in the human evaluation test are all well-educated graduate students with skilled English.

We use the online questionnaire platform to design the electronic questionnaire.

The questionnaire system can automatically score according to the pre-set reference answers, and count the time spent on answering the questions.

The timer starts when the volunteer clicks “accept" button on the questionnaire, and ends when the volunteer clicks “submit" button.

Volunteers are required to answer the questionnaire without any interruption, ensuring that all time spent is for answering questions.

Before starting lling the questionnaire, we provide a sample example as instruction to teach the volunteers how to nd the answer.

The interface of human evaluation for each group could be found in Figure 5 and Figure 6.

[4] Because Bride type questions always has deterministic linear reasoning chains.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 569-582, Harbin, China, August 3 - 5, 2023.

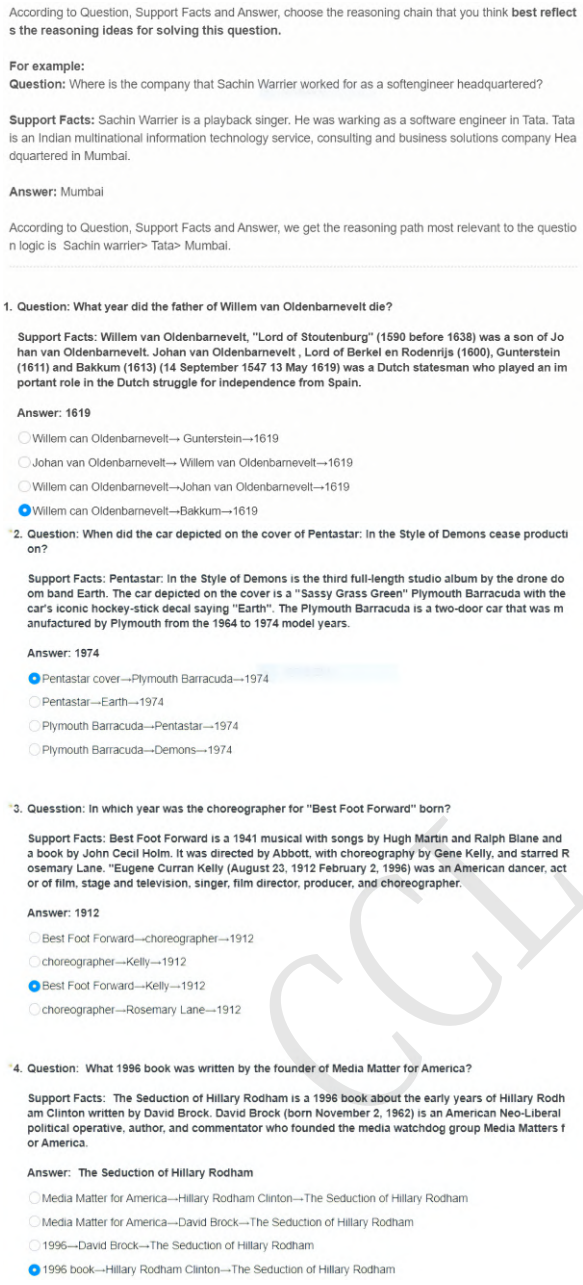


Figure 5: Interface for human evaluation of choosing reasoning chain based on support facts.

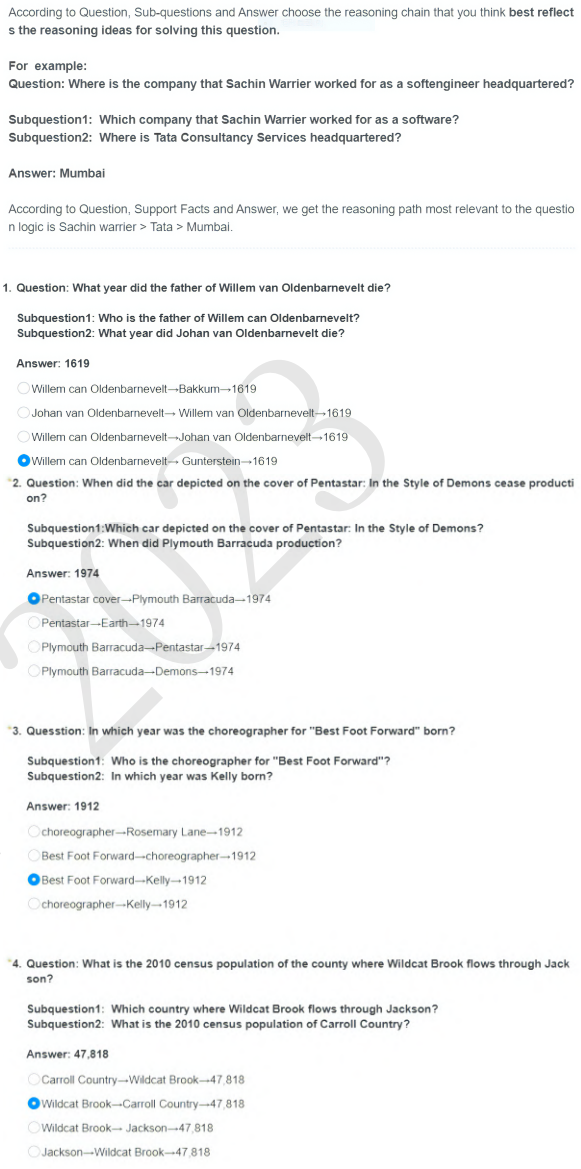


Figure 6: Interface for human evaluation of choosing reasoning chain based on sub-questions.s

Learning on Structured Documents for Conditional Question Answering

Zihan Wang, Hongjin Qian, Zhicheng Dou*
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
{wangzihan0527, ian, dou}@ruc.edu.cn

Abstract

Conditional question answering (CQA) is an important task in natural language processing that involves answering questions that depend on specific conditions. CQA is crucial for domains that require the provision of personalized advice or making context-dependent analyses, such as legal consulting and medical diagnosis. However, existing CQA models struggle with generating multiple conditional answers due to two main challenges: (1) the lack of supervised training data with diverse conditions and corresponding answers, and (2) the difficulty to output in a complex format that involves multiple conditions and answers. To address the challenge of limited supervision, we propose LSD (Learning on Structured Documents), a self-supervised learning method on structured documents for CQA. LSD involves a conditional problem generation method and a contrastive learning objective. The model is trained with LSD on massive unlabeled structured documents and is fine-tuned on labeled CQA dataset afterwards. To overcome the limitation of outputting answers with complex formats in CQA, we propose a pipeline that enables the generation of multiple answers and conditions. Experimental results on the ConditionalQA dataset demonstrate that LSD outperforms previous CQA models in terms of accuracy both in providing answers and conditions.

1 Introduction

Recently, question answering (QA) has gained increasing interest in the field of Natural Language Processing. Various types of question answering tasks, such as knowledge-based QA (Cui et al., 2017), open domain QA (Kwiatkowski et al., 2019), and multi-hop QA (Yang et al., 2018), have been extensively studied. Among them, conditional question answering (CQA) (Sun et al., 2022a) is becoming increasingly important in various contexts, such as medical diagnosis, legal consultation, financial analysis, and more. Unlike the traditional question answering problem that only accepts a question and returns an answer, CQA involves understanding a complex and lengthy document, finding all possible answers under different **conditions**, and determining under what **condition** the answer is applicable. Figure 1 shows an example for CQA, where the answer could be different when the questioner is under different conditions. The CQA task includes providing potential answers "yes" and "no" and their corresponding conditions based on the given question and scenario.

Previous studies on CQA can be broadly categorized into two groups: extractive and generative methods. Extractive methods (Ainslie et al., 2020) (Sun et al., 2021) extract the most relevant span from a document as answers and conditions. In contrast, generative methods (Izacard and Grave, 2021) (Sun et al., 2022b) use a generative model to generate answers along with their corresponding conditions directly. However, current CQA models face two common challenges. Firstly, the supervised data for CQA is limited and expensive to obtain. Unlike traditional QA datasets, CQA requires specific annotations that include scenarios, answers, and conditions, making the data collection process more extensive and time-consuming. Secondly, current CQA models are unable to provide multiple conditional answers in a coherent and controlled format. Extractive methods for CQA are mostly only able to provide a single answer or condition for a question, limiting their ability to produce multiple conditional answers.

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

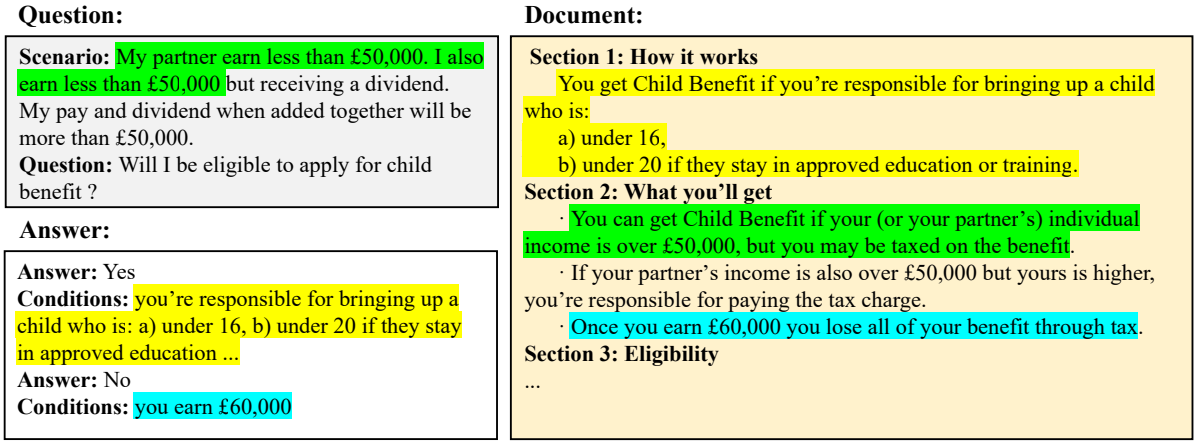


Figure 1: An example for CQA. The right side is a snapshot of a document discussing the policy of claiming Child Benefits. The green text span is the condition that has been satisfied. The yellow and blue text spans are the conditions for "Yes" and "No" respectively.

Conversely, generative methods may generate inconsistent and incoherent answers and conditions due to their inherent randomness, especially when generating multiple conditional answers. These challenges underscore the need for improved approaches to effectively handle the generation of multiple conditional answers in CQA.

In order to solve the problem of limited supervision, we propose a self-supervised learning method called LSD (Learning on Structured Documents). LSD consists of two main components: conditional question generation and contrastive learning. For conditional question generation, our intuition is that if a more precise context that contains sufficient information to answer a conditional question can be passed to the QA model, then the conditional answers given through this context will have high accuracy and can be used for subsequent training. To achieve this goal, we propose a selective extraction process that extracts parts of a structured document that are likely to be able to answer a conditional question. For a certain selected part of the document, we use a state generator to generate a conditional question and user scenario, and use a label generator to generate highly believed answers. For contrastive learning, we use four methods of document perturbation to perturb the structure of the document, including node reordering, repetition, masking, and deletion. These methods will change the content of the document but have little impact on its semantics. We design a contrastive learning objective that encourages the model to give similar representations of document corresponding sentences before and after perturbation, enabling the model to learn effective semantic representations from complex documents and helping with conditional question answering.

To solve the problem of complex output formats, we propose a pipeline that can generate multiple answers and their corresponding conditions. Our pipeline extracts answer spans from sentences, generating query vectors for each answer and key vectors for each candidate condition. Afterward, we calculate the query-key matching score for each answer and condition, and choose the best matches as the final output. Unlike existing methods, our pipeline utilizes the structure of documents to generate questions and conditions, and can generate controllable multiple conditional answers.

To verify the effectiveness of our method, we conduct experiments on the conditionalQA dataset (Sun et al., 2022a). The experimental results showed that our method outperformed all baseline models in terms of answer and condition accuracy, indicating that our method can provide accurate answers and corresponding conditions to effectively answer conditional questions.

In summary, our contributions are three-fold:

(1) We propose LSD, a self-supervised learning method for structured documents based on problem generation and contrastive learning, which effectively solves the problem of insufficient supervision for conditional question answering;

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

(2) We propose a pipeline that generates a query and key vectors for candidate answers and conditions and matching similarity for them, which can provide controllable conditional answers;

(3) The experimental results indicate that our method can answer conditional questions more effectively compared to previous conditional question answering methods.

2 Related Work

2.1 Conditional Question Answering

Conditional question answering (CQA) has been studied using extractive and generative methods. Extractive methods, such as ETC (Ainslie et al., 2020) and DocHopper (Sun et al., 2021), use two separate models to extract answers and conditions. ETC pipeline uses two separate encoders to extract answers from supporting documents and identify conditions. DocHopper, on the other hand, iteratively attends to different sentences to predict evidences, answers and conditions step-by-step. Generative methods such as FiD (Izacard and Grave, 2021) use a single generative model to generate answers with conditions. FiD splits documents into sentences, encodes the sentences separately, and jointly decodes all encoded representations to generate answers with conditions. TReasoner (Sun et al., 2022b) is a discriminative-generative model that first checks whether each sentence could be a condition and then generates the answer with the context. However, these models suffer from several limitations, including a lack of sufficient supervised data, which can lead to overfitting and poor reasoning capabilities. Furthermore, pipeline designs have a limited ability to generate multiple hybrid-type answers and conditions. Therefore, improving the performance of CQA through a suitable pipeline is crucial, and our work aims to address these challenges.

2.2 Self-Supervised Learning

Self-supervised learning methods have gained significant traction in recent years, as they allow models to learn powerful representations without relying on large amounts of labeled data. Various language models, such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020), have leveraged unsupervised pre-training to achieve remarkable results on extensive natural language tasks. There have also been multilingual approaches like XLM (Conneau et al., 2020), unsupervised machine translation (Lample et al., 2018), question generation techniques such as QA-based multiple-choice question generation (Le Berre et al., 2022), Web-pretraining (Guo et al., 2022), and deep reinforcement learning (Chen et al., 2019). On the other hand, contrastive learning has emerged as a powerful method for representation learning, with models like SimCSE (Gao et al., 2021), ELECTRA (Clark et al., 2020), DPR-QA (Karpukhin et al., 2020) and XMOCO (Yang et al., 2021) achieving state-of-the-art results across various natural language understanding and generation tasks by learning to distinguish between semantically similar and dissimilar inputs.

3 Preliminaries: Structured Documents

Structured documents contain complex and rich structural information, which is beneficial for learning conditional question answering. In this work, our model is trained on HTML documents, a widely used type of structured document. HTML documents are easily accessible and often contain rich semantic information, including tables, lists, and more. The underlying structure of an HTML document is represented by the Document Object Model (DOM) tree, wherein the entire document constitutes the root node, and individual elements are organized as child nodes within the hierarchy.

A diagram of a DOM tree is shown in Figure 2. Since HTML does not always demonstrate a clear hierarchy among elements, we adopt a tag precedence order to convert HTML documents into trees, thus making the relationships between elements explicit. We order commonly used tags as: ⟨title⟩ - ⟨h⟩ - ⟨p⟩ - ⟨li⟩ / ⟨tr⟩. Each node's parent is the closest preceding higher-level node. For example, the ⟨h1⟩ tag is a section title and is the parent of ⟨h2⟩ subsection titles. The ⟨h2⟩ tag is a subheading and is the parent of ⟨p⟩ text elements. We omit tags that do not contain important information, such as ⟨b⟩ (bold),⟨i⟩ (italic), and ⟨a⟩ (hyperlink) tags. With our approach, each sentence within the HTML document can be clearly represented as a node in the document tree.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

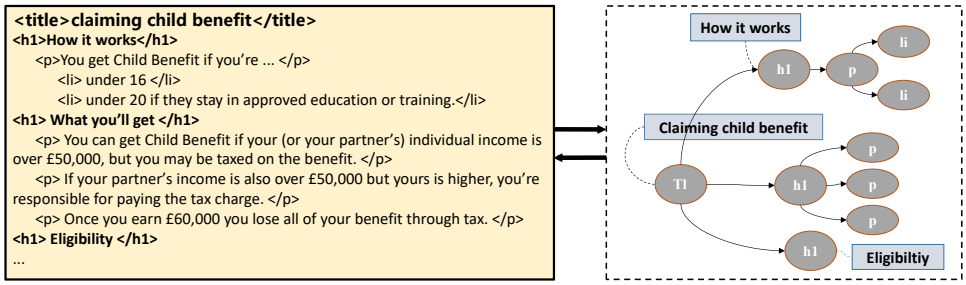


Figure 2: An example of the schematic diagram of a DOM tree. HTML tags can be used to create a hierarchy of sentences in a document, with some tags considered more senior than others. The nearest former superior tag of a node is its parent node.

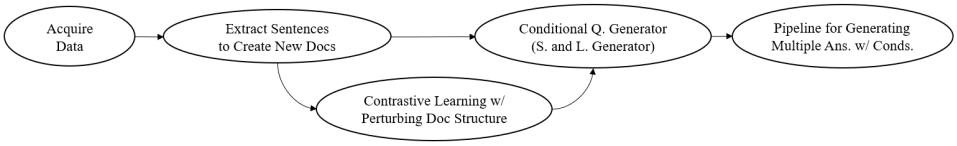


Figure 3: An overall illustration of our approach.

To compile a corpus of structured documents for the CQA task, we consider the following criteria:

- Logical Structure: Documents should possess clear logical structures, including specific conditions and provisions, to facilitate conditional reasoning in the CQA task
- Standardized Format: Documents should adhere to a standardized HTML format with minimal noise, such as advertisements.
- Data Quality: The corpus should comprise formal, authoritative, and reviewed documents to ensure data reliability and accuracy.

Based on these criteria, we propose to train our model to learn on **national government websites**, which are known for their formal and authoritative nature. We conduct web scraping to gather documents, filtering for policy documents, laws and regulations, and administrative guidelines, as they tend to exhibit clear logical structures and contain specific conditions relevant to the CQA task. For additional details regarding the construction of the corpus, which is referred to as DATASET, please refer to Appendix A.

4 Our Approach

In this section, we will introduce our proposed method LSD, which includes a conditional question generation module and a contrastive learning method for self-supervised learning on structured documents. After that, we will illustrate our pipeline that generates multiple conditional answers by calculating the matching score of answers and candidate conditions with query and key vectors. The overall process of our method is shown in Figure 3.

4.1 Decomposed Conditional Question Generation with Document Extraction

Let the conditional question generator be G and the conditional question answering model be M. Recall that the intuition of our approach is that if we can provide G with a more precise context with sufficient information for a conditional question, then G can answer the question correctly, and the obtained question-answer data can be used to train M. To achieve this, we adopt a two-step method: selective extraction and question generation. A specific overview of conditional question generation is in Algorithm 1.

4.1.1 Selective Extraction

Selective extraction aims for precise context to generate conditional questions. The main requirement for the selected context is to contain sufficient information to answer a conditional question. To guide our

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

Algorithm 1 Conditional Question Generation

Require: Structured doc set DATASET
Ensure: Cond. question q, scenario sc, answer a, condition c
1: **procedure** QUESTIONGEN(DATASET)
2: **Init:** state gen. G_S, label gen. G_L
3: **Sample** doc D from DATASET
4: **Select** non-leaf text node $s \in D$ as potential answer
5: **Construct** extracted $\overline{D}$ by selecting anc., child., sibl., and sibl. child. of s
6: **Gen.** question q, scenario sc using $G_S(\overline{D})$
7: **Gen.** cond. answers $A = (a_i, c_i)$ using $G_L(q, sc, \overline{D})$
8: **end procedure**

	answers	conditions
leaf node	86.93%	92.53%
text node	92.49%	98.33%

(a) Features of answers and condition nodes: whether they are leaf nodes or text nodes.

	a-a pairs	c-c pairs	a-c pairs
sibling-sibling	66.55%	53.67%	-
parent-child	-	-	39.59%

(b) Features of answer and condition pairs: answer pairs (a-a), condition pairs (c-c), and answer-condition pairs (a-c).

Table 1: Statistics of the ConditionalQA train dataset for guiding selective extraction.

extraction strategy, we analyzed the ConditionalQA dataset, which also leverages structural documents for the CQA task. (Table **??**). We analyzed the occurrence and correlations between answers and conditions, and observed several features: (1) answers and conditions are mainly located in leaf text nodes, such as ⟨p⟩ and ⟨li⟩ nodes; (2) answers are usually siblings; (3) conditions for extractive answers may be their child nodes; (4) sibling nodes with the same parent node can serve as parallel answers.

Guided by these insights, our extraction method involves the following steps. Firstly, we randomly select a non-leaf text node as a potential answer, because conditional answers are most likely to be such nodes. Then, we then extract its ancestors, children, siblings, and their children from the document tree, because: (1) ancestor nodes provide the macro context of higher-level topics; (2) child nodes offer potential conditions; (3) siblings, along with their children, provide parallel answers. Afterward, we obtain an extracted document that enables generating conditional questions aligned with the original text and answerable with accuracy.

4.1.2 Question Generation

The question generation approach are decomposed into two tasks: state generation and label generation. The first task is to generate question q and scenario sc given the extracted structured document $\overline{D}$, and the second task is to generate highly accurate conditional answers $A = \{(a_1, c_1), (a_2, c_2), \ldots\}$, where a_i is an answer and c_i denotes the corresponding conditions. We leverage a state generator G_S, a sequence-to-sequence (Sutskever et al., 2014) generative model to provide diverse content, and a conditional answer extraction model G_L, an extractive model to provide accurate answers. More information on the network structure and training process of G can be found in Appendix C.

In general, by leveraging structured documents for precise document extraction and supervised generator training, we ensure that we can identify the locations of potential answers and conditions within structured documents, thereby achieving the generation of high-quality conditional questions and ensuring the correct answering of questions for subsequent training.

4.2 Purturbation-based Document Contrastive Learning

Our contrastive learning approach on structured documents involves the following steps: document perturbation, positive sample generation, and contrastive loss computation. At the training stage, the computed loss is added to the total training loss for optimization.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

Operation	Description	Advantages
Node masking	Mask node with [MASK] of same length	Focus on structure & context
Node deletion	Delete non-root node & descendants	Learn node dependencies & importance
Node cloning	Clone node & descendants as another child	Identify semantically similar elements
Node shuffling	Shuffle child nodes within parent	Understand impact of node order

Table 2: Basic operations for Contrastive Learning.

4.3 Document Perturbation

To perturb the original document D and obtain a perturbed document $\hat{D}$, we introduce a set of basic operations T that can be applied to the document structure. These operations, detailed in Table 2, include node masking, node deletion, node cloning, and node shuffling. Assume the original document D has a title s_0 and m nodes $(n_1, n_2, ..., n_m)$. Starting with the original document D_0, we apply k random operations from the set T to generate the perturbed document $\hat{D} = D_k$. Each operation T_i is applied as $T_i(D_j) = D_{j+1}$ for any T_i selected from T.

4.3.1 Positive Pair Generation

We get positive pairs from D and $\hat{D}$ for loss calculation. For the i^{th} node n'_i in the perturbed document $\hat{D}$, there is a corresponding source node n_{k_i} in the original document D. We form positive pairs using tags t'_i and t_{k_i} that serves as global tokens of the nodes, which effectively convey node type and semantics despite structural changes during document perturbation.

4.3.2 Contrastive Loss Computation

We use the InfoNCE loss $\mathcal{L}_{CL}(D, \hat{D})$ for contrastive learning, defined as:

$$\mathcal{L}_{CL}(D, \hat{D}) = \sum_{i=1}^{m'} \frac{e^{\text{sim}(t'_i, t_{k_i})}}{e^{\text{sim}(t'_i, t_{k_i})} + \sum_{t^-_{k_i}} e^{\text{sim}(t'_i, t^-_{k_i})}}, \tag{1}$$

where m' is the total number of nodes in $\hat{D}$, t'_i and t_{k_i} represents a positive pair, and $t^-_{k_i}$ represents tags of any nodes other than n_{k_i} in D. sim computes the similarity between tags using the dot product of their hidden states from a neural document encoder, detailed in 4.3. The loss encourages high similarity between each t'_i and t_{k_i} while minimizing similarity with negative tags $t^-_{k_i}$.

In general, our contrastive learning approach enables self-supervised training by perturbing structured documents to construct contrastive pairs. By reinforcing node correspondence in structured documents, the method supports conditional question answering models in accurately capturing semantic connections between conditions and answers in complex contexts.

4.4 Pipeline for Answering Conditional Questions

Our proposed pipeline, illustrated in Figure 4, comprises three steps: (1) document encoding, (2) multiple answer extraction, (3) condition determination. An auxiliary task Evidence Node Finding is added when necessary (Appendix D).

4.4.1 Document Encoding

In the document encoding process, we first construct the input sequence, which consists of special tokens "[yes]" and "[no]" document content, question, and scenario. The special tokens are used to represent affirmative/negative answers. We represent the input sequence as follows:

$$\begin{aligned}\text{Input} = &\text{“[yes][no]document : ”} + D \\ &+ \text{“question : ”} + q + \text{“Scenario : ”} + sc,\end{aligned}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

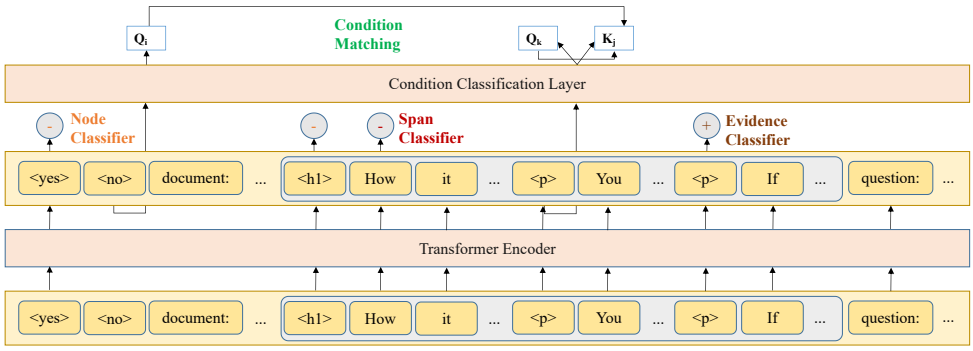


Figure 4: Our pipeline to answer conditional questions.

where [yes] and [no] are special tokens for yes / no answers. It is passed to E returning hidden states:

$$\begin{aligned}\text{Output} &= \text{Transformer}(\text{Input}) \\ &= h_{[\text{yes}]}, h_{[\text{no}]}, ...h_{t_i}, h_{a_{ij}}, ...,\end{aligned}$$

where $h_{[\text{yes}]}, h_{[\text{no}]}$ are hidden states of special tokens, h_{t_i} represents hidden state of the tag of the i^{th} node in the document, and $h_{a_{ij}}$ represents hidden state of the i^{th} node's j^{th} token. These hidden states are used by the multi-layer perceptron (MLP) classifiers P_S, P_N, P_V to calculate probabilities for answer extraction and condition determination.

4.4.2 Multiple Answer Extraction

To simplify the answer extraction process, we assume that a node has no more than one answer, and we retain only one answer if multiple exist. Since it's rare that a node has multiple answers, this process simplifies extraction by identifying potential answer nodes and determining the answer's start and end positions within the node.

We use two classifiers: a node classifier P_N to identify answer-containing nodes (or yes/no tokens) and an answer span classifier P_S to locate the answer's position within selected nodes.

For node classification, we set:

$$\begin{aligned}p^N_{\text{yes/no}} &= P_N(h_{[\text{yes}]/[\text{no}]}), \\ p^N_i &= P_N(h_{t_i}),\end{aligned} \tag{2}$$

where p represents probabilities given by these classifiers. From the above, we can obtain yes/no answers and sentences containing extractive answers from node classification results. At training, We set a Binary Cross Entropy (BCE) loss for node classification:

$$\mathcal{L}_{\text{bool}} = \frac{\text{BCE}(p^N_{\text{yes}}, \mathbb{I}^N_{\text{yes}}) + \text{BCE}(p^N_{\text{no}}, \mathbb{I}^N_{\text{no}})}{2}, \tag{3}$$

$$\mathcal{L}_{\text{extractive}} = \frac{1}{m}\sum_{i=1}^{m}\text{BCE}(p^N_i, \mathbb{I}^N_i), \tag{4}$$

$$\mathcal{L}_{\text{N}} = \mathcal{L}_{\text{bool}} + \mathcal{L}_{\text{extractive}}, \tag{5}$$

where $\mathbb{I}$ represents boolean labels to indicate whether the given element satisfies some requirements, e.g., $\mathbb{I}^N_i$ represents whether the i^{th} node is a potential answer node, assuming totally m nodes.

For answer span localization, we adopt a span locator P_S for any positive nodes of the above process by:

$$\begin{aligned}p^{S_i}_{j1}, p^{S_i}_{j2}, ... = P_{S_i}(a^A_{j1}), P_{S_i}(a^A_{j2}), ..., \\ (i \in (1,2), j \in (1,2,...,k)),\end{aligned} \tag{6}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583–599, Harbin, China, August 3 – 5, 2023.

where P_{S_1}, P_{S_2} predict start / end of the answer, a^A_{ju} denotes the u^{th} token of the j^{th} predicted node n^A_j to have an answer, and $p^{S_i}_{ju}$ are the predicted probabilities. At training, we adopt a span loss:

$$\mathcal{L}_{\text{S}} = \frac{1}{2k_r} \sum_{i=1}^{2} \sum_{j=1}^{k_r} \sum_{u=1}^{l_{n^A_j}} \frac{1}{l_{n^A_j}} \text{BCE}(p^{S_i}_{ju}, \mathbb{I}^{S_i}_{ju}), \tag{7}$$

where k_r represents the real count of answers and $l_{n^A_j}$ represents the number of tokens of n^A_j.

4.4.3 Condition Determination

To align with the document structure, we define that a potential condition must be a node in the document. Therefore, the condition determination process is to predict the probability of a node being the condition of an answer. To model this, we assign query vectors to answers, and key vectors to nodes:

$$\begin{aligned} h^Q_i &= W^Q \, \text{ReLU}(W^H h_i), \\ h^K_j &= W^K \, \text{ReLU}(W^H h_j), \end{aligned} \tag{8}$$

where h_i, h_j denotes the hidden state of i^{th} answer and j^{th} sentence. W^H, W^Q, W^K are transformation matrices, h^Q_i, h^K_j denotes the query vector of i^{th} answer and the key vector of j^{th} sentence.

Then, we calculate on conditions:

$$p^C_{ij} = \text{sigmoid}(h^Q_i \cdot h^K_j), \tag{9}$$

where p^C_{ij} denotes the probability of j^{th} node to be the condition of the i^{th} answer. We adopt the following loss for training:

$$\mathcal{L}_{\text{C}} = \frac{1}{k_r m} \sum_{i=1}^{k_r} \sum_{j=1}^{m} \text{BCE}(p^C_{ij}, \mathbb{I}^C_{ij}). \tag{10}$$

From the above process, we can fuse the representations of answers and conditions to model the condition determination process. Therefore, our pipeline has resolved the conditional question answering task. At training, we linearly mix up all losses mentioned:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{N}} + \mathcal{L}_{\text{S}} + \mathcal{L}_{\text{C}} + \mathcal{L}_{\text{CL}}. \tag{11}$$

5 Experiments

5.1 Datasets and Evaluation Metrics

To construct a dataset of structured documents, we scrape web pages from English websites. Our data collection process is detailed in Appendix A. To evaluate LSD's effectiveness on CQA, we conduct experiments on ConditionalQA (Sun et al., 2022a) dataset. It consists of extractive questions, yes / no questions, and not-answerable questions. The task is to find all answers with corresponding conditions on a structured document based on the given questions and scenarios.

Evaluation To evaluate model performance, we adopt the metrics of EM / F1 and EM / F1 with conditions, which are introduced in the ConditionalQA (Sun et al., 2022a) dataset. EM / F1 are conventional metrics, and EM / F1 with conditions jointly measures the correctness of the answer and the predicted conditions. For not answerable questions, EM and F1 are 1.0 if and only if no answer is predicted.

5.2 Results

We compared the LSD model with all of the baseline models for CQA. To evaluate the model's performance in both answering questions and providing conditions, we present results for the entire ConditionalQA dataset and its subset of conditional questions.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

	Yes / No		Extractive		Conditional		Overall	
	EM / F1	w/ conds	EM / F1	w/ conds	EM / F1	w/ conds	EM / F1	w/ conds
ETC-pipeline	63.1 / 63.1	47.5 / 47.5	8.9 / 17.3	6.9 / 14.6	39.4 / 41.8	2.5 / 3.4	35.6 / 39.8	26.9 / 30.8
DocHopper	64.9 / 64.9	49.1 / 49.1	17.8 / 26.7	15.5 / 23.6	42.0 / 46.4	3.1 / 3.8	40.6 / 45.2	31.9 / 36.0
FiD	64.2 / 64.2	48.0 / 48.0	25.2 / 37.8	22.5 / 33.4	45.2 / 49.7	4.7 / 5.8	44.4 / 50.8	35.0 / 40.6
TReasoner	**73.2 / 73.2**	**54.7 / 54.7**	34.4 / 48.6	30.3 / 43.1	51.6 / 56.0	12.5 / 14.4	57.2 / 63.5	**46.1 / 51.9**
LSD (ours)	71.6 / 71.6	51.6 / 51.6	**39.9 / 56.4**	**31.6 / 43.8**	**57.3 / 61.8**	**21.4 / 25.1**	**58.7 / 66.2**	45.0 / 50.5

Table 3: The results of our experiments on the ConditionalQA dataset. "EM / F1" shows the standard EM / F1 metrics based on the answer span only. "w/ conds" shows the conditional EM / F1 metrics introduced in (Sun et al., 2022a). The results for the baseline models are taken from (Sun et al., 2022a) (Sun et al., 2022b)

	Answer (w / conds)	Conditions (P / R / F1)
ETC-pipeline	/	/
DocHopper	/	/
FiD	3.2 / 4.6	98.3 / 2.6 / 2.7
FiD (cond)	6.8 / 7.4	12.8 / 63.0 / 21.3
TReasoner	10.6 / 12.2	34.4 / 40.4 / 37.8
LSD (ours)	**21.4 / 25.1**	**69.3 / 39.4 / 50.2**

Table 4: Experimental results on the subset of questions in ConditionalQA (dev) with conditional answers. Results of the baseline models are obtained from (Sun et al., 2022a) (Sun et al., 2022b). The first two models "do not provide any conditions when they achieved the best performance on the overall dataset".

5.2.1 Main Result

Table 3 shows the results on the entire conditionalQA dataset. The result indicates that:

(1) LSD outperforms all baselines in EM / F1 and conditional EM / F1 for extractive and conditional questions, demonstrating the effectiveness of our conditional question generation and contrastive learning.

(2) LSD performs not as well as TReasoner in Yes / No questions. We speculate that its attributed to LSD inclination to provide conditional answers due to training with our question generation system (Appendix B), which is penalized by the evaluation metric in (Sun et al., 2022a).

(3) In "w/ conds" overall results, LSD performs less well than TReasoner, potentially due to TReasoner's specialized multi-hop reasoning for condition determination, which may warrant further enhancement in LSD.

5.2.2 Conditional Accuracy

To further evaluate our model's ability to provide conditions for answers, we additionally report results on the subset of conditional questions in Table 4. We evaluate the results using the "w/ conds" metric, as well as precision, recall, and F1 of retrieved conditions for conditional answers. The result shows that our method significantly outperforms the current model in providing conditions.

6 Analysis

In this section, we conduct an ablation study to investigate the impact of our document modeling designs and contrastive learning. We further analyze the question generation process by evaluating the quality of generated questions and the accuracy of generated labels.

6.1 Ablation Study

We conduct an ablation study on the dataset to investigate the impact of conditional question generation and contrastive learning. Results on the dev set of ConditionalQA in Table 5 show that both conditional

	Yes / No		Extractive		Conditional		Overall	
	EM / F1	w/ conds	EM / F1	w/ conds	EM / F1	w/ conds	EM / F1	w/ conds
LSD (ours)	**71.6 / 71.6**	**51.6 / 51.6**	**39.9 / 56.4**	**31.6 / 43.8**	**57.3 / 61.8**	**21.4 / 25.1**	**58.7 / 66.2**	**45.0 / 50.5**
w/o CL	69.6 / 69.6	49.9 / 49.9	38.0 / 55.7	29.8 / 43.2	54.6 / 59.1	19.4 / 23.2	56.9 / 64.8	43.3 / 49.4
w/o QG	67.9 / 67.9	47.1 / 47.1	37.2 / 54.9	29.0 / 42.5	54.0 / 58.6	17.8 / 21.6	55.7 / 63.7	41.6 / 47.6

Table 5: Ablation study of our model on the dev set of ConditionalQA.

	ROUGE (%)	BLEU (%)
question	42.07	38.19
scenario	39.57	41.65

(a) Evaluation on state generator's output quality.

	Yes / No	Extractive	Conditional	Overall
EM / F1 (%)	79.6 / 79.6	51.2 / 67.2	69.9 / 73.8	67.8 / 75.0
w / conds (%)	50.8 / 50.8	38.9 / 51.3	33.4 / 35.5	47.9 / 53.4

(b) Evaluation on accuracy of generated labels.

Table 6: Evaluation on our question generation method.

question generation and contrastive learning are of importance, as removing either of them causes a significant performance drop in the final results.

6.2 State Generator's Output Quality

We use BLEU and ROUGE-L to measure the state generator's generated questions and scenarios' similarity to questions and scenarios from the evaluation dataset for question generation, QG-dev (detailed in Appendix C). The results are shown in Table 6a. Some examples are shown in Appendix E.

6.3 Label Generator's Output Accuracy

We evaluate our label generator's capability in providing accurate answers for questions given the extracted documents from QG-dev, shown in Table 6b. The result shows that the label generator can provide accurate answers given a selected context from the document.

7 Conclusion and Limitations

In this paper, we present Learning on Structured Documents (LSD), a self-supervised learning method for conditional question answering. LSD uses a conditional question generation method to leverage massive structured documents while improving conciseness, and applies contrastive learning to learn effective semantic representations from complex documents. We further propose a pipeline that could generate multiple answers and conditions to better handle the CQA task. We verify the effectiveness of the proposed method on the ConditionalQA dataset. For future work, we plan to investigate how to better generate conditional questions and improve our model's performance in providing correct answers.

Despite the effectiveness of LSD in utilizing the structure of massive unsupervised data, there are still some potential points for improvement. One issue is that the state generator is only trained on answerable questions, leading to a distribution bias that there might be unanswerable questions. In addition, our pipeline can still not handle the position where a sentence has more than one answer, which limits our model's performance for broader scenarios. We will resolve these issues in future work.

Acknowledgements

This work was supported by National Natural Science Foundation of China No. 62272467, Beijing Outstanding Young Scientist Program No. BJJWZYJH012019100020098, and Public Computing Cloud, Renmin University of China. The work was partially done at Beijing Key Laboratory of Big Data Management and Analysis Methods.

References

Joshua Ainslie, Santiago Ontañón, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: encoding long and structured inputs in transformers.

In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 268–284. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2019. Natural question generation with reinforcement learning based graph-to-sequence model. *CoRR*, abs/1910.08832.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. KBQA: learning question answering over QA corpora and knowledge bases. *Proc. VLDB Endow.*, 10(5):565–576.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Yu Guo, Zhengyi Ma, Jiaxin Mao, Hongjin Qian, Xinyu Zhang, Hao Jiang, Zhao Cao, and Zhicheng Dou. 2022. Webformer: Pre-training with web pages for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 15021512, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Guillaume Le Berre, Christophe Cerisara, Philippe Langlais, and Guy Lapalme. 2022. Unsupervised multiple-choice question generation for out-of-domain Q&A fine-tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 732–738, Dublin, Ireland, May. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

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

Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2021. End-to-end multihop retrieval for compositional question answering over long documents. *CoRR*, abs/2106.00200.

Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2022a. Conditionalqa: A complex reading comprehension dataset with conditional answers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3627–3637. Association for Computational Linguistics.

Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2022b. Reasoning over logically interacted conditions for question answering. *CoRR*, abs/2205.12898.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Nan Yang, Furu Wei, Binxing Jiao, Daxing Jiang, and Linjun Yang. 2021. xMoCo: Cross momentum contrastive learning for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6120–6129, Online, August. Association for Computational Linguistics.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

Appendix

A DATASET curation details

	UK	US	CA	Overall
count	17,881	577	12,115	30,573
Avg. w	709	179	2,538	1423
Avg. s	54	26	128	83
Avg. w/s	12.9	6.9	19.8	17.0
Tag dist.	14:45:41	38:40:22	10:40:50	12:41:57

Table 7: Statistics of our scraped dataset. We present document count, average document length measured by word (Avg. w) and sentences (Avg. s), average sentence length (Avg w/s) and tag distribution (h:p:li/tr).

DATASET contains a total of 30,573 documents, approximately 362MB in size (1×10^8 tokens). The statistics of our scraped dataset are shown in Table 7. The data curation process are detailed below.

A.1 Data Acquisition

To build DATASET, we scrape web pages from government websites: https://www.gov.uk, https://www.ca.gov, and https://www.usa.gov, as they have professional English material and have a massive number of well-structured documents, such as policies, regulations, and proposals.

A.2 Data Filtering

Page Category Filtering We use automated web scraping to categorize pages on the selected government websites based on URL. We retaine only pages related to policy documents, regulatory provisions, administrative guidelines, etc.

Content Validity Check We further examined the retained pages to exclude invalid, redundant, or duplicate documents.

A.3 Data Cleaning

Tag Normalization We use automated cleaning and standardization tools to fix irregular HTML tags and attributes in documents, close unclosed tags, and standardize attribute values.

Irrelevant Content Removal We remove nodes without text, advertisements, hyperlinks, images, videos, and other irrelevant information, retaining textual content for better model understanding of document structure and content.

Node Filtering We filtere nodes containing document content, i.e., <h1> to <h6> (headings), <p> (paragraphs), <li> (list items), <tr> (table rows), etc.

DOM Tree Construction We use an HTML parser to parse the filtered nodes and construct the Document Object Model (DOM) tree following the method proposed in section 3.

A.4 Dataset Splitting

We split the processed dataset into training and validation sets for model training and performance evaluation with a ratio of 4:1.

B Question Generation details

We present the statistics to show our question generation module's behavior on the scraped augmentation corpus. We randomly generate 1,000 samples with the QG module and present results in Table 8.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

Our Dataset	Yes / No	Extractive	Conditional
Percentage	52.4	47.5	45.1
Avg. answer	1.36	1.46	1.86
Avg. condition	0.89	1.04	2.14
Avg. context	292	350	413
Avg. document	1,467	1,260	1,525
ConditionalQA	Yes / No	Extractive	Conditional
Percentage	51.1	44.6	23.4
Avg. document		1358	

Table 8: Statistics of our generated dataset and ConditionalQA dataset in comparison. We present the percentage of every type of questions, average answer count, condition count, condition count, context length and document length (by word) if applicable.

C Implementation Details

C.1 Network Structure and Setup

For the conditional question generator G: we adopt BART [1] (Lewis et al., 2020), a seq-to-seq transformer for state generator G_S; for label generator G_L, we adopt the same setting of M, as detailed below.

For conditional question answering model M: We adopt Longformer [2] (Beltagy et al., 2020), a Transformer designed for long complex context, for the neural document encoder E; for MLP classifiers P_N, P_S, P_V, we set num_layers=2 and dim_hidden_states=768; for transformation matrices, we set $\dim(W_H) = (3072, 768)$ and $\dim(W^Q) = \dim(W^K) = (768, 3072)$.

To setup Longformer, we set the HTML tags as its global tokens. For extremely long documents beyond length limit, we chunk them into pieces with overlap and aggregate predicted answers from these pieces.

C.2 Training Conditional Question Generator

To train conditional question generator G, we use 80% data of the ConditionalQA train set, named QG-train, and the rest for evaluation, named QG-dev. We take the descendants and ancestors of all given evidence sentences from the document for extraction. We train G on QG-train for 10 epochs, adopting the Adam (Kingma and Ba, 2015) optimizer, setting learning rate to 3e-5 and batch size to 32.

C.3 Training Conditional Question Answerer

Training conditional question answering model M consists of two stages. In the self-supervised stage, we train M on our scraped dataset for 20 epochs, with a newly generated question and answer data for every epoch. We use the LAMB (You et al., 2020) optimizer for this stage, with the learning rate set to 1e-4 and the batch size set to 256. In the supervised stage, we adopt the Adam (Kingma and Ba, 2015) optimizer, setting the learning rate to 3e-5 and batch size to 32, and trained on ConditionalQA train set for 50 epochs. For both stages of training, we adopt a warm-up episode of 10% proportion with linear learning rate decay. For document chunking, We set the maximum of document length to 2000 to fit into the GPU memory, with an overlap of 100 tokens. For contrastive learning, we adopt k=5.

D Auxiliary Task: Evidence Node Finding

To improve model reasoning for yes / no questions, we introduce an auxiliary task to identify evidence nodes supporting the answer. The task is jointly trained with others and is active when datasets provide evidence information. We use an evidence classifier P_V for this task and define:

$$p_i^V = P_V(\mathrm{h}_{\mathrm{t_i}}), \tag{12}$$

[1] https://huggingface.co/facebook/bart-large

[2] https://huggingface.co/allenai/longformer-large-4096

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

$$\mathcal{L}_{\mathrm{E}} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{BCE}(p_i^V, \mathbb{I}_i^V), \tag{13}$$

When the evidence node finding task is activated, the training loss turns to:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\mathrm{N}} + \mathcal{L}_{\mathrm{S}} + \mathcal{L}_{\mathrm{C}} + \mathcal{L}_{\mathrm{CL}} + \mathcal{L}_{\mathrm{E}}. \tag{14}$$

E Case Studies for Question Generation

In this section, we provide 3 generated questions and the model-generated questions and conditional answers.

Example 1

Context:
⟨title⟩Funding Opportunities⟨/title⟩
⟨h1⟩Funding Opportunities⟨/h1⟩
⟨h2⟩Current Opportunities⟨/h2⟩
⟨p⟩Winter Collaborative Networks⟨/p⟩
⟨li⟩GC Key access⟨/li⟩
⟨li⟩SecureKey Concierge (Banking Credential) access⟨/li⟩
⟨li⟩Personal Access Code (PAC) problems or EI Access Code (AC) problems⟨/li⟩
⟨li⟩Social Insurance Number (SIN) validation problems⟨/li⟩
⟨li⟩Other login error not in this list⟨/li⟩

Generated Contents:
Question:
What type of access is required for the Winter Collaborative Networks funding opportunity?

Scenario:
I am interested in applying for the Winter Collaborative Networks funding opportunity, but I am unsure of what type of access is required.

Answers and Conditions:
[GC Key access, None],
[SecureKey Concierge (Banking Credential) access, None],
[Personal Access Code (PAC) problems or EI Access Code (AC) problems, None],
[Social Insurance Number (SIN) validation problems, None],
[Other login error not in this list, None]

Example 2, 3

Context:
⟨title⟩Claim Capital Allownces⟨/title⟩
⟨h1⟩What you can claim on⟨/h1⟩
⟨p⟩You can claim capital allowances on items that you keep to use in your business - these are known as plant and machinery.⟨/p⟩
⟨p⟩In most cases you can deduct the full cost of these items from your profits before tax using annual investment allowance (AIA).⟨/p⟩
⟨p⟩If youre a sole trader or partnership and have an income of č150,000 or less a year, you may be able to use a simpler system called cash basis instead.⟨/p⟩
⟨h2⟩What does not count as plant and machinery⟨/h2⟩
⟨p⟩You cannot claim plant and machinery allowances on:⟨/p⟩
⟨li⟩things you lease (unless you have a hire purchase contract or long funding lease) - you must own them⟨/li⟩
⟨li⟩items used only for business entertainment, for example a yacht or karaoke machine⟨/li⟩

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

⟨li⟩land⟨/li⟩
⟨li⟩structures, for example bridges, roads, docks⟨/li⟩
⟨li⟩buildings, including doors, gates, shutters, mains water and gas systems⟨/li⟩
⟨p⟩You may be able to claim structures and buildings allowance on structures and buildings.⟨/p⟩
⟨h2⟩What counts as plant and machinery⟨/h2⟩
⟨p⟩Plant and machinery includes:⟨/p⟩
⟨li⟩items that you keep to use in your business, including cars⟨/li⟩
⟨li⟩costs of demolishing plant and machinery⟨/li⟩
⟨li⟩parts of a building considered integral, known as integral features⟨/li⟩
⟨li⟩some fixtures, for example, fitted kitchens or bathroom suites⟨/li⟩
⟨li⟩alterations to a building to install plant and machinery - this does not include repairs⟨/li⟩
⟨p⟩Claim repairs as business expenses if youre a sole trader or partnership - deduct from your profits as a business cost if youre a limited company.⟨/p⟩
⟨h3⟩Integral features⟨/h3⟩
⟨p⟩Integral features are:⟨/p⟩
⟨li⟩lifts, escalators and moving walkways⟨/li⟩
⟨li⟩space and water heating systems⟨/li⟩
⟨li⟩air-conditioning and air cooling systems⟨/li⟩
⟨li⟩hot and cold water systems (but not toilet and kitchen facilities)⟨/li⟩
⟨li⟩electrical systems, including lighting systems⟨/li⟩
⟨li⟩external solar shading⟨/li⟩
⟨h3⟩Fixtures⟨/h3⟩
⟨p⟩You can claim for fixtures, for example:⟨/p⟩
⟨li⟩fitted kitchens⟨/li⟩
⟨li⟩bathroom suites⟨/li⟩
⟨li⟩fire alarm and CCTV systems⟨/li⟩
⟨p⟩You can claim if you rent or own the building, but only the person who bought the item can claim.⟨/p⟩
⟨p⟩When you buy a building from a previous business owner you can only claim for integral features and fixtures that they claimed for.⟨/p⟩
⟨p⟩You must agree on the value of the fixtures with the seller. If you do not you cannot claim for them. Agreeing the value also means the person selling the assets can account correctly for them.⟨/p⟩
⟨h2⟩If you let residential property⟨/h2⟩
⟨p⟩You can only claim for items to be used in a residential property if either:⟨/p⟩
⟨li⟩you run a furnished holiday lettings business⟨/li⟩
⟨li⟩the item is to be used in the common parts of a residential building, for example, a table in the hallway of a block of flats⟨/li⟩
⟨h2⟩Care workers⟨/h2⟩
⟨p⟩There are special rules if you run a care business.⟨/p⟩

Generated Contents:
Question:
Can I claim capital allowances on a yacht used for entertaining clients?

Scenario:
I am a business owner and I want to purchase a yacht to entertain clients. I want to know if they can claim capital allowances on it.

Answers and Conditions:
[No, None]

Question:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

Can I claim capital allowances on the commercial building I purchased for my business?

Scenario:

I am a business owner and I recently purchased a commercial building to use as a location for my business.

Answers and Conditions:

[No, [⟨p⟩You cannot claim plant and machinery allowances on:⟨/p⟩, ⟨li⟩buildings, including doors, gates, shutters, mains water, and gas systems⟨/li⟩]]

[Yes, ⟨p⟩You may be able to claim structures and buildings allowance on structures and buildings.⟨/p⟩]

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 583-599, Harbin, China, August 3 - 5, 2023.

Overcoming Language Priors with Counterfactual Inference for Visual Question Answering

Zhibo Ren, Huizhen Wang*, Muhua Zhu, Yichao Wang, Tong Xiao, Jingbo Zhu
NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
rzb1998@qq.com
wanghuizhen@mail.neu.edu.cn

Abstract

Recent years have seen a lot of efforts in attacking the issue of language priors in the field of Visual Question Answering (VQA). Among the extensive efforts, causal inference is regarded as a promising direction to mitigate language bias by weakening the direct causal effect of questions on answers. In this paper, we follow the same direction and attack the issue of language priors by incorporating counterfactual data. Moreover, we propose a two-stage training strategy which is deemed to make better use of counterfactual data. Experiments on the widely used benchmark VQA-CP v2 demonstrate the effectiveness of the proposed approach, which improves the baseline by 21.21% and outperforms most of the previous systems.

1 Introduction

As an AI-complete task to answer questions about visual content, Visual Question Answering (VQA) has seen surging interest in recent years. The task is thought to be extremely challenging since a VQA system requires the capability of visual and language understanding and the capability of multi-modal reasoning. Recent researches in this field have paid increasing attention to the issue of language priors, aka language bias (Agrawal et al., 2018). The issue of language priors is caused by spurious correlation between the question pattern and the answer. See the example in Figure 1, "yellow" is the most likely answer to the question "what color are the bananas" in the training data. So a simple solution to answering the question is to give the answer "yellow" with no reference to visual content. Such a short cut can achieve an accuracy of 54.5% for the question.

To overcome language priors in VQA, previous works generally resort to data augmentation. In this direction, visual and textual explanations can be used as the data for augmentation (Das et al., 2017; Park et al., 2018). Besides, counterfactual training samples are also regarded as a valuable source for the purpose (Chen et al., 2020; Zhu et al., 2020; Gokhale et al., 2020; Liang et al., 2020). In the direction of causal effect for VQA, more recent work is counterfactual VQA that focuses on the inference instead of training phase (Niu et al., 2021), though, we still think of counterfactual data augmentation as an efficient and effective way to solve the issue of language priors. So in this paper we first design novel causal graphs specifically for the task of VQA, and then use the causal graphs to guide the generation of counterfactual data. Finally, to make better use of counterfactual data, we propose a two-stage training strategy. We evaluate the proposed approach on the widely used benchmark VQA-CP v2. Extensive experiments demonstrate the effectiveness of the approach, which improves over the baseline by 21.21% and outperforms most of previous systems. Moreover, to evaluate the generalization ability of the approach, we also experiment with VQA v2 and find that our approach achieves the best performance on the dataset.

The contributions of the paper are as follows.

- For the task of counterfactual VQA, we design a novel causal graph and methods to construct counterfactual data.

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

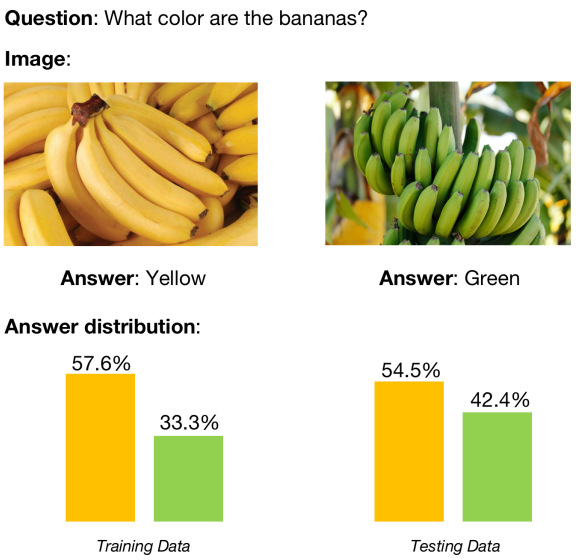


Figure 1: An example from VQA v2 which is used to illustrate 1) the task of visual question answering, and 2) the issue of language priors.

- Our approach achieves significant improvements over the baseline and is one of the best-performing systems on the benchmarks.

2 Methodology

In this section, we first describe the implementation of our baseline system. Then we introduce the design of VQA causal graphs which inspire us to come up with the proposed methods. Finally we describe the methods in detail. The system framework is presented in Figure 2.

2.1 The Baseline System

Following the conventional paradigm of VQA systems, we formalize the task as a multi-class classification problem. In general, a VQA dataset consists of N instances which are tuples of an image, a textual question, and the corresponding answer, denoted as $D = \{I_i, Q_i, A_i\}_{i=1}^{N}$. VQA models take an image-question pair (I, Q) as input, and predict an answer A by following

$$A^* = \arg\max_{A \in \mathcal{A}} P(A|I_i, Q_i), \tag{1}$$

where $P(A|I_i, Q_i)$ can be any model-based functions that map (I, Q) to produce a distribution over the answer space $\mathcal{A}$. Conventional VQA systems are generally composed of three components:

- **Feature Extraction**, which extracts the features of images and question as visual representation and text representation, respectively.
- **Multimodal Feature Fusion**, which fuses image and text features into the same vector space.
- **Answer Prediction**, which produces the answer prediction through a classifier.

We follow (Anderson et al., 2018) to implement our baseline system. The baseline system pays special attention to feature extraction by integrating a combined bottom-up and top-down attention mechanism to enable attention calculation at the fine-grained level of objects. Within the approach, the bottom-up attention proposes image regions while the top-down mechanism determines feature weightings.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

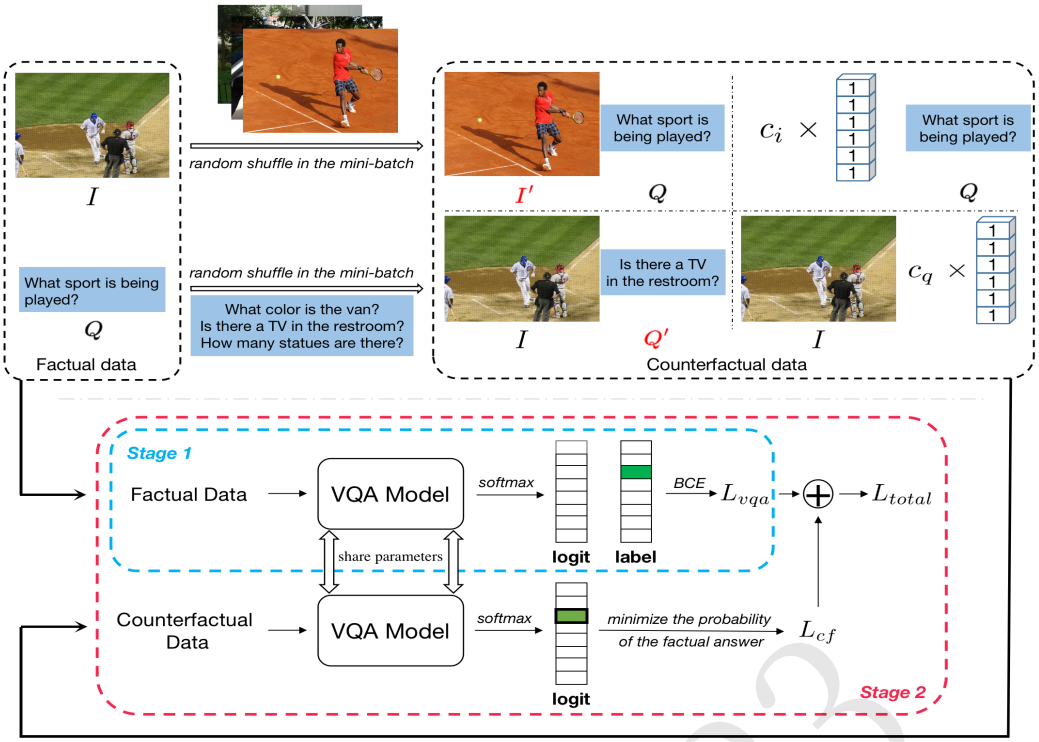


Figure 2: Illustration of our approach, where the upper half presents the process of counterfactual data generation and the bottom half represents the process of two-stage training.

2.2 Causal Graph for VQA

To better understand the casual graphs we propose for the VQA task, we need to revisit the procedure of VQA data annotation. Specifically, when curating a dataset, annotators are required to produce a question regarding visual content of a presented image and give a correct answer. Therefore, we can construct a casual graph to exhibit the relationship between three variables: the image I, the question Q, and the answer A. Figure 3(a) illustrates the casual graph, where I indirectly and directly affects A through $I \rightarrow Q \rightarrow A$ and $I \rightarrow A$, respectively. In the chain of $I \rightarrow Q \rightarrow A$, the question Q acts as a mediator to influence A. If we control the mediator Q, the causal association between I and A in the chain $I \rightarrow Q \rightarrow A$ will be blocked, that is, when the association between I and A is not well learned through $I \rightarrow A$ (the middle and right graph in Figure 3(a)), the model will give the answer based on the question only but ignore the content of the image. This phenomenon corresponds exactly to the language prior problem in VQA. Therefore, we propose to introduce counterfactual data to weaken the effect that comes from the chain $I \rightarrow Q \rightarrow A$, which is shown in Figure 3(b).

2.3 Automatic Generation of Counterfactual Data

We propose two methods to construct counterfactual data, corresponding to multimodal counterfactual data and unimodal counterfactual data, respectively.

Multimodal Counterfactual Data. First of all, we realize that the issue of language priors is caused by the chain $I \rightarrow Q \rightarrow A$, so we need to mitigate the influence of this branch on the selection of the answer. Inspired by (Zhu et al., 2020), for each pair(I_i, Q_i) in factual data, we construct counterfactual data (I_i', Q_i) by shuffling image I_i in the same mini-batch, such that the image and the question in counterfactual data are mismatched. The causal graph of counterfactual image data is shown in Figure 4(a). Following the same idea, we also propose to construct counterfactual question data by shuffling questions in the same mini-batch. The corresponding causal graph is illustrated in Figure 4(b). Subsequent experiments show that incorporation of multimodal counterfactual question data is also beneficial to the performance, which demonstrates the presence of vision bias in the VQA task, a phenomenon not often

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

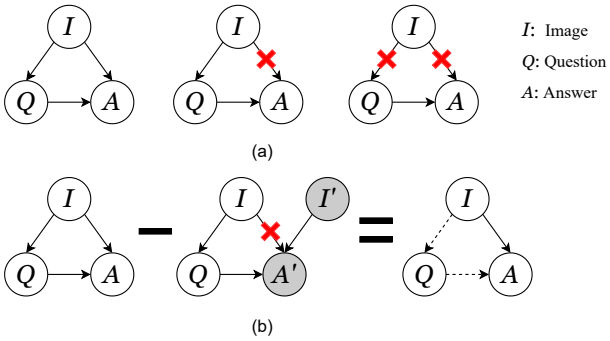


Figure 3: (a) Casual graph for VQA. (b) Overcome language priors with counterfactual data.

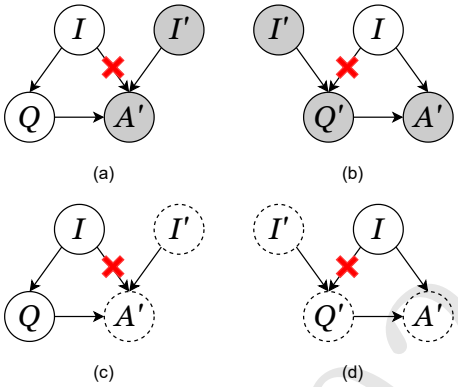


Figure 4: Causal graph demonstrating the methods for generating counterfactual data.

mentioned before.

It is worth noting that we do not resort to any extra human annotations during the construction of the multimodal counterfactual data, but simply make use of the factual data itself. The underlying idea is quite different from the methods proposed in previous works for the construction of counterfactual data (Chen et al., 2020; Liang et al., 2020; Gokhale et al., 2020).

Unimodal Counterfactual Data. We further consider to construct unimodal counterfactual data. We hope the model to accept information from only one modality as input. Concretely, we construct unimodal counterfactual data by passing only images$(I_i, \emptyset)$ or questions$(\emptyset, Q_i)$ into the model, which the causal graph is illustrated in Figure 4(c)(d). However, the model cannot handle the case where the input is empty during implementation, so we choose to use a learnable parameter c multiplied by a matrix whose elements are all ones and the shape is same as image representation or text representation as the null modal information. Finally, the unimodal counterfactual data can be represented as (I_i, c_q) and (c_i, Q_i).

2.4 Two-stage Training Strategy

In the real world, we can only give the right answer when we see the right factual image-question pair. Conversely, we often cannot give the correct answer when we see a counterfactual image-question pair. But usually in this case the correct answer will change and the previously correct answer will often become the wrong answer, which is the only thing we know for sure. We hope to solve language prior problems by using counterfactual image data in the manner shown in Figure 3(b). Specifically, when the VQA model takes the counterfactual image data as input, we construct the loss function by minimizing the probability of the ground truth answer:

$$\begin{aligned} P(A'|I'_i, Q_i) &= softmax(F(I'_i, Q_i)) \\ L_{mm\_cf\_i} &= P(A'|I'_i, Q_i)[k] \end{aligned} \tag{2}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

Systems	VQA-CP v2 test(%)				VQA v2 val(%)			
	All	Y/N	Num	Other	All	Y/N	Num	Other
UpDn	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66
GVQA	31.3	57.99	13.68	22.14	48.24	72.03	31.17	34.65
SAN	24.96	38.35	11.14	21.74	52.41	70.06	39.28	47.84
Systems without counterfactual inference								
DLR	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
VGQE	48.75	-	-	-	64.04	-	-	-
AdvReg	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
RUBi	44.23	67.05	17.48	39.61	-	-	-	-
LMH	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
CVL	42.12	45.72	12.45	48.34	-	-	-	-
Unshuffling	42.39	47.72	14.43	47.24	61.08	78.32	42.16	52.81
RandImg	55.37	83.89	41.6	44.2	57.24	76.53	33.87	48.57
SSL	57.59	86.53	29.87	50.03	63.73	-	-	-
Systems with counterfactual inference								
CSS	58.95	84.37	49.42	48.21	59.91	73.25	39.77	55.11
CSS+CL	59.18	86.99	49.89	47.16	57.29	67.29	38.40	54.71
CF-VQA	53.55	**91.15**	13.03	44.97	63.54	**82.51**	43.96	54.30
MUTANT	**61.72**	88.90	49.68	**50.78**	62.56	82.07	42.52	53.28
This Paper	60.95	87.95	**50.41**	49.70	**64.11**	82.23	**44.09**	**56.75**

Table 1: Comparison with the state-of-the-art methods on the VQA-CP v2 test set and VQA v2 validation set. The evaluation metric is accuracy, and the backbone of all models is UpDn. Overall best scores are **bold** and the second best of scores are underlined.

where k denotes the index of the ground truth in the answer set A. For the counterfactual question data, the corresponding loss function is similar to equation(2): , which can be defined as:

$$\begin{aligned} P(A'|I_i, Q_i') &= softmax(F(I_i, Q_i')) \\ L_{mm\_cf\_q} &= P(A'|I_i, Q_i')[k] \end{aligned} \tag{3}$$

Finally, the loss of the multimodal counterfactual data is defined as:

$$L_{mm\_cf} = \lambda_i^{mm} L_{mm\_cf\_i} + \lambda_q^{mm} L_{mm\_cf\_q}, \tag{4}$$

where λ_i and λ_q are hyperparameters.
Similar to multimodal counterfactual data, the unimodal counterfactual loss function can be defined as:

$$\begin{aligned} P(A'|c_i, Q_i) &= softmax(F(c_i, Q_i)) \\ L_{um\_cf\_i} &= P(A'|c_i, Q_i)[k] \end{aligned} \tag{5}$$

$$\begin{aligned} P(A'|I_i, c_q) &= softmax(F(I_i, c_q)) \\ L_{um\_cf\_q} &= P(A'|I_i, c_q)[k] \end{aligned} \tag{6}$$

The total loss of unimodal counterfactual data is defined as:

$$L_{um\_cf} = \lambda_i^{um} L_{um\_cf\_i} + \lambda_q^{um} L_{um\_cf\_q} \tag{7}$$

Simply combining counterfactual and factual data together as training data may render these two types of data interfere with each other. For this reason, we adopt a two-stage training strategy, which utilize factual data and the normal VQA loss function for training in the first stage and utilize counterfactual data and counterfactual loss functions in the second stage. are introduced on top of the first stage to alleviate the problem of the language priors of the VQA model:

$$L_{total} = L_{vqa} + \lambda^{mm} L_{mm\_cf} + \lambda^{um} L_{um\_cf} \tag{8}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

3 Experiments

3.1 Datasets and Comparative Systems

Datasets. We conducted extensive experiments on the most widely used benchmark VQA-CP v2 (Agrawal et al., 2018) adopting the standard evaluation metric. Because the dataset of VQA v2 (Goyal et al., 2017) has the language prior problem, (Agrawal et al., 2018) reorganized the data splitting of VQA v2 to construct VQA-CP v2 where answers have different distributions in the training and validation set. Thus, VQA-CP v2 is an appropriate benchmark for evaluating the generalization ability of VQA models. Briefly, the training set of VQA-CP v2 contains approximately 121k images and 245k questions, and the test set consists of approximately 98k images and 220k questions.

Comparative Systems. System participating in the comparison against our approach can be categorized into two groups: 1) systems without counterfactual inference, including **DLR** (Jing et al., 2020), **VGQE** (KV and Mittal, 2020), **AdvReg** (Ramakrishnan et al., 2018), **RUBi** (Cadène et al., 2019), **LMH** (Clark et al., 2019), **Unshuffling** (Teney et al., 2021), **RandImg** (Teney et al., 2020), **SSL** (Zhu et al., 2020), and 2) systems with counterfactual inference, including **CF-VQA** (Niu et al., 2021), **CSS** (Chen et al., 2020), **CL** (Liang et al., 2020), and **MUTANT** (Gokhale et al., 2020).

3.2 Implementation Details

As mentioned above, our VQA system builds on the base of UpDn (Anderson et al., 2018). Following previous researches, we use the Faster-RCNN (Ren et al., 2015) model previously trained by (Anderson et al., 2018) to extract image features. We extract 36 region features for each image and the dimension of each region feature is set to 2048. Moreover, each question is padded so as to have the same length of 14 tokens, and each token in questions is encoded by the pretrained language model BERT (Devlin et al., 2019) with a dimension of 768. Then word embeddings are fed into GRUs to obtain the question representation with a dimension of 1280. Inspired by SSL (Zhu et al., 2020), we also add a BatchNorm layer before the MLP classifier of UpDn. We train our model for 25 epochs every time. We adopt the Adam optimizer to update model parameters, whose learning rate is set to 0.001 and the learning rate decreases by half every 5 epochs after 10 epochs. The batch size is set to 256. We implement our system using PyTorch, and we train our model with one Nvidia 2080Ti card.

3.3 Main Experimental Results

Table 1 presents the comparison results between our approach and previous systems on both VQA-CP v2. From the results, we can see that our approach significantly improves the baseline UpDn by $+21.21\%$ on VQA-CP v2. The improvement demonstrates the effectiveness of our approach on mitigating the issue of language prior. Moreover, our approach outperforms all the comparative systems on VQA-CP v2 except for MUTANT which requires additional human annotations of key objects in images. Moreover, we can see our approach achieves stable performance on VQA v2 with the best performance over all the previous systems. To demonstrate the generality of our approach, we also experiment with VQA v2, and the results show that our approach achieves the best performance among all the participating systems.

3.4 Experiment Analysis

Impact of Counterfactual Data Combination

We proposed several types of counterfactual data, so we conducted a study on the effect of each type of counterfactual data and the effect of their combinations. From the results shown in Table 2, we have the following observations:

- Both counterfactual image data (I_i', Q_i) and counterfactual question data (I_i, Q_i') are able to improve the performance. The use of counterfactual image data achieves significant improvements, while the counterfactual question data achieves relatively limited improvements. This suggests that the main cause of the language prior problem is the superficial correlation between questions and answers, but there are also some vision bias that cannot be ignored.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

	Counterfactual Data (I_i', Q_i)	(I_i, Q_i')	(c_i, Q_i)	(I_i, c_q)	Acc.
			-		41.52
MM	✔	-	-	-	57.59
	-	✔	-	-	41.87
	✔	✔	-	-	**59.05**
UM	-	-	✔	-	41.83
	-	-	-	✔	41.70
	-	-	✔	✔	**41.88**
Total	✔	✔	✔	✔	**60.95**

Table 2: Impact of different types of counterfactual data, evaluated on VQA-CP v2 test set. MM refers to multimodal counterfactual data and UM refers to unimodal counterfactual data, respectively

λ	Ratio	VQA-CP v2 test(%)
λ_i^{mm}:λ_q^{mm}	1:0.5	58.06
	1:0.7	**59.46**
	1:1	59.32
	1:2	59.15
	1:3	58.76
λ_i^{um}:λ_q^{um}	1:0.5	60.03
	1:0.7	60.29
	1:1	**60.34**
	1:2	59.51
	1:3	58.07
λ^{mm}:λ^{um}	1:0.5	60.17
	1:0.7	60.53
	1:1	**60.95**
	1:2	58.21
	1:3	60.29

Table 3: Impact of different ratio between λ. We divide λ into three groups($\lambda_i^{mm} : \lambda_q^{mm}$), ($\lambda_i^{um} : \lambda_q^{um}$),($\lambda^{mm} : \lambda^{um}$) according to the counterfactual data used, with the latter group realized on the best results of the previous group's experiment. The evaluation metric is accuracy(%).

- Both multimodal counterfactual data and unimodal counterfactual data can improve the model performance, which demonstrates that these data can prompt the generalization ability of model.

In summary, the above experimental results verify the validity of the counterfactual data.

Impact of Varying Settings of λ

As we can see from the results in Table 2, different types of counterfactual data have diverse effect on the performance. So we need to evaluate the effect of varying settings of the hyperparameters λ in the loss functions. We divide λ into three groups for comparison and conducted extensive experiments with different λ values. From results in Table 3, we can observe that the model gets the best performance when $\lambda_i^{mm} : \lambda_q^{mm}$ is 1:0.7, $\lambda_i^{um} : \lambda_q^{um}$ is 1:1, and $\lambda^{mm} : \lambda^{um}$ is 1:1.

Impact of Varying Starting Points of the Second Stage Training

In the process of two-stage training, different starting points of the second stage tend to achieve different results. So we conducted an experiment to show the effect of varying starting points. As can be seen in Figure 5, starting the training on counterfactual data too early or too late will bring negative effect on

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

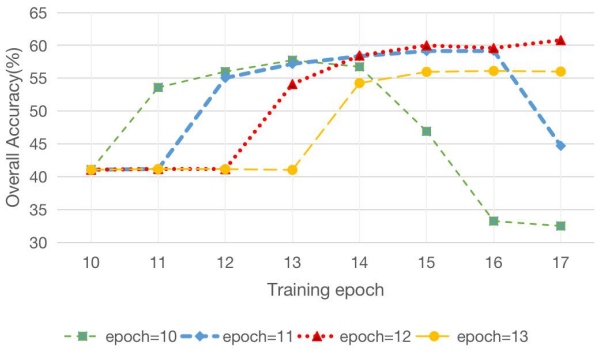


Figure 5: Impact of different starting points of the second stage training, evaluated on the VQA-CP v2 test set.

Methods	Overall(%)	Gap Δ ↑
UpDn	39.74	**+21.21**
UpDn + counterfactual data	**60.95**	
SAN	24.96	**+27.46**
SAN + counterfactual data	**52.42**	

Table 4: Performance of different backbones on VQA-CP v2 test set.

the performance. Empirically, we find the second stage can start its training at the 12th epoch.

Impact of Different Backbones

We also conducted experiments on another backbones SAN (Yang et al., 2016) to verify that our approach is model agnostic. From the results in Table 4, we can observe that our approach can achieve significant improvements no matter what backbone is used.

4 Related Work

Visual Question Answering

Visual Question Answering aims to answer the question according to the given image, which involves both natural language processing and computer vision techniques. At present, the dominant methods are attention-based models. (Anderson et al., 2018; Yu et al., 2019; Yang et al., 2016) use attentions mechanisms to capture the alignment between images and natural language in order to learn the intrinsic interactions between image regions and words. (Antol et al., 2015) maps two modal features(visual and textual features) into a common feature space and then passes the joint embedding into the classifier to obtain the answer of the question. Another methods including that compositional models that (Andreas et al., 2016) applies neural module network to the VQA task, which is a combination of several modular networks. The neural module network is dynamically generated according to the linguistic structure of the question. (Wu et al., 2016) introduces external knowledge to help model with answering the questions.

Attacking Language Priors in VQA

Despite the progress made in the field of VQA, recent researches have found that VQA systems tend to exploit superficial correlations between question patterns and answers to achieve state-of-the-art performance (Agrawal et al., 2016; Kafle and Kanan, 2017). To help build a robust VQA system, (Agrawal et al., 2018) propose a new benchmark named VQA-CP whose training and testing data have vast distributions. Recent solutions to overcome the language priors can be grouped into two categories as without counterfactual inference (Clark et al., 2019; Zhu et al., 2020; Teney et al., 2021) and with counterfactual inference (Agrawal et al., 2019; Pan et al., 2019; Chen et al., 2020; Liang et al., 2020;

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

Gokhale et al., 2020).
For the methods that without counterfactual inference, RUBi (Cadène et al., 2019) proposes to dynamically adjust the weights of samples according to the effect of the bias, LMH (Clark et al., 2019) ensembles a question-only branch to discriminates which questions can be answered without utilizing image and then penalizes these questions. Unshuffling (Teney et al., 2021) describes a training procedure to capture the patterns that are stable across environments while discarding spurious ones. SSL (Zhu et al., 2020) proposes a self-supervised framework that generates labeled data to balance the biased data.
For the methods that with counterfactual inference, One solution is is to modify model architecture that implement counterfactual inference to reduce the language bias (Niu et al., 2021). The other one is to synthesize counterfactual samples to improve the robustness of VQA systems (Agrawal et al., 2019; Pan et al., 2019; Chen et al., 2020; Liang et al., 2020; Gokhale et al., 2020). CSS (Chen et al., 2020) generates the counterfactual samples by masking objects in the image or some keywords in the question. Based on CSS, CL (Liang et al., 2020) introduces a contrastive learning mechanism to force the model to learn the relationship between original samples, factual samples and counterfactual samples. MUTANT (Gokhale et al., 2020) utilizes the extra object-name annotations to locates critical objects in the image and critical words in the question and then mutates these critical elements to generate counterfactual samples.

5 Conclusion and Future Work

To mitigate the effect of language priors in the VQA task, we proposed a causal inference approach that automatically generates counterfactual data and utilize the data in a two-stage training strategy. We also designed several causal graphs to guide the generation of counterfactual data. Extensive experiments on the benchmark VQA-CP v2 shows that our system achieves significant improvements over the baselines and outperforms most of previous works. Moreover, our system achieves the best performance on VQA v2 which demonstrates the capability of generalization.

The starting point of the the second stage training is critical to the performance, in our future work, we would like to determine the starting point in an automatic way. Moreover, it is interesting to evaluate the performance when other networks such as SAN are used as the backbone. We will also study this problem in our future work.

Acknowledgements

This work was supported in part by the National Science Foundation of China (No. 62276056), the National Key RD Program of China, the China HTRD Center Project (No. 2020AAA0107904), the Natural Science Foundation of Liaoning Province of China (2022-KF-16-01), the Yunnan Provincial Major Science and Technology Special Plan Projects (No. 202103AA080015), the Fundamental Research Funds for the Central Universities (Nos. N2216016, N2216001, and N2216002), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No. B16009).

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of EMNLP*, pages 1955–1960.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of CVPR*, pages 4971–4980.

Vedika Agrawal, Rakshith Shetty, and Mario Fritz. 2019. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of CVPR*, pages 9690–9698.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*, pages 6077–6086.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of CVPR*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *Proceedings of ICCV*, pages 2425–2433.

Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Proceedings of NeurIPS*, pages 839–850.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of CVPR*, pages 10797–10806.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of EMNLP-IJCNLP*, pages 4067–4080.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of EMNLP*, pages 878–892.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*, pages 6325–6334.

Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in VQA via decomposed linguistic representations. In *Proceedings of AAAI*, pages 11181–11188.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of CVPR*, pages 1983–1991.

Gouthaman KV and Anurag Mittal. 2020. Reducing language biases in visual question answering with visually-grounded question encoder. In *Proceedings of ECCV*, volume 12358, pages 18–34.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of EMNLP*, pages 3285–3292.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of CVPR*, pages 12700–12710.

Jingjing Pan, Yash Goyal, and Stefan Lee. 2019. Question-conditional counterfactual image generation for VQA. In *arXiv, preprint arXiv:1911.06352*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of CVPR*, pages 8779—-8788.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of NeurIPS*, pages 1548–1558.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of NeurIPS*, pages 91–99.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. 2020. On the value of out-of-distribution testing: An example of goodhart's law. In *Proceedings of NeurIPS*.

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. In *Proceedings of ICCV*, pages 1397–1407.

Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of CVPR*, pages 4622–4630.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of CVPR*, pages 21–29.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 600-610, Harbin, China, August 3 - 5, 2023.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of CVPR*, pages 6281–6290.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of IJCAI*, pages 1083–1089.

Rethinking Label Smoothing on Multi-hop Question Answering

Zhangyue Yin[◇*] **Yuxin Wang**[◇*] **Xiannian Hu**[◇] **Yiguang Wu**[◇] **Hang Yan**[◇]
Xinyu Zhang[♠] **Zhao Cao**[♠] **Xuanjing Huang**[◇] **Xipeng Qiu**[◇†]
[◇]School of Computer Science, Fudan University
[♠]Huawei Poisson Lab
{yinzy21,wangyuxin21,xnhu21}@m.fudan.edu.cn
{ygwu20,hyan19,xjhuang,xpqiu}@fudan.edu.cn
{zhangxinyu35,caozhao1}@huawei.com

Abstract

Multi-Hop Question Answering (MHQA) is a significant area in question answering, requiring multiple reasoning components, including document retrieval, supporting sentence prediction, and answer span extraction. In this work, we present the first application of label smoothing to the MHQA task, aiming to enhance generalization capabilities in MHQA systems while mitigating overfitting of answer spans and reasoning paths in the training set. We introduce a novel label smoothing technique, F1 Smoothing, which incorporates uncertainty into the learning process and is specifically tailored for Machine Reading Comprehension (MRC) tasks. Moreover, we employ a Linear Decay Label Smoothing Algorithm (LDLA) in conjunction with curriculum learning to progressively reduce uncertainty throughout the training process. Experiment on the HotpotQA dataset confirms the effectiveness of our approach in improving generalization and achieving significant improvements, leading to new state-of-the-art performance on the HotpotQA leaderboard.

1 Introduction

Multi-Hop Question Answering (MHQA) is a rapidly evolving research area within question answering that involves answering complex questions by gathering information from multiple sources. This requires a model capable of performing several reasoning steps and handling diverse information structures. In recent years, MHQA has attracted significant interest from researchers due to its potential for addressing real-world problems. The mainstream approach to MHQA typically incorporates several components, including a document retriever, a supporting sentence selector, and a reading comprehension module (Tu et al., 2020; Wu et al., 2021; Li et al., 2022). These components collaborate to accurately retrieve and integrate relevant information from multiple sources, ultimately providing a precise answer to the given question.

Despite the remarkable performance of modern MHQA models in multi-hop reasoning, they continue to face challenges with answer span errors and multi-hop reasoning errors. A study by S2G (Wu et al., 2021) reveals that the primary error source is answer span errors, constituting 74.55%, followed by multi-hop reasoning errors. This issue arises from discrepancies in answer span annotations between the training and validation sets. As illustrated in Figure 1(a), the training set answer includes the quantifier "times", while the validation set answer does not. Upon examining 200 samples, we found that around 13.7% of answer spans in the HotpotQA validation set deviate from those in the training set.

Concerning multi-hop reasoning, we identified the presence of unannotated, viable multi-hop reasoning paths in the training set. As depicted in Figure 1(b), the non-gold document contains the necessary information to answer the question, similar to gold doc1, yet is labeled as an irrelevant document. During training, the model can only discard this reasoning path and adhere to the annotated reasoning path. Given that current MHQA approaches primarily use cross-entropy loss for training multiple components,

[*]Equal contribution.
[†] Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Training set:
Gold Doc1: ***Love or Leave***
"Love or Leave" was the Lithuanian entry in the Eurovision Song Contest 2007, performed in English by 4FUN.
Gold Doc2: ***Lithuania in the Eurovision Song Contest***
Lithuania has participated in the Eurovision Song Contest (known in Lithuania as "Eurovizija") 18 times since its debut in 1994, where Ovidijus Vyšniauskas finished last, receiving nul points.
Question: How many times does the song writer of "Love or Leave" have participated in the Eurovision Song Contest?
Answer: 18 times

Validation set:
Gold Doc1: ***Binocular (horse)***
"Love or Leave" was the Lithuanian entry in the Eurovision Song Contest 2007, performed in English by 4FUN.
Gold Doc2: ***Tony McCoy***
Based in Ireland and the UK, McCoy rode a record 4,358 winners, and was Champion Jockey a record 20 consecutive times, every year he was a professional.
Question: The primary jockey of Binocular was Champion Jockey how many consecutive times?
Answer: 20

(a) Different Answer Span

Non-Gold Doc: ***F. W. Woolworth Building (Watertown, New York)***
(1) The Woolworth Building is an historic building in Watertown, New York.
(2) It is a contributing building in the Public Square Historic District.
(3) Plans for the Woolworth Building were begun in 1916 by Frank W. Woolworth, the founder of the Woolworth's chain of department stores.

Gold Doc1: ***Woolworth Building***
(1) The Woolworth Building, at 233 Broadway, Manhattan, New York City, designed by architect Cass Gilbert and constructed between 1910 and 1912, is an early US skyscraper.

Gold Doc2: ***1 New York Plaza***
(1) 1 New York Plaza is an office building in New York City's Financial District, built in 1969 at the intersection of South and Whitehall Streets.
(2) It is the southernmost of all Manhattan skyscrapers.

Question: Which was built first Woolworth Building or 1 New York Plaza?
Answer: Woolworth Building
Evidence Sentences: ["Woolworth Building", 0], ["1 New York Plaza",0]

(b) Multiple Feasible Reasoning Paths

Figure 1: Causes of errors in answer span and multi-hop reasoning within the HotpotQA dataset. In Figure (a), the answer from the training set contains a quantifier, while the answer from the validation set does not. Figure (b) demonstrates that the correct answer can be inferred using a non-gold document without requiring information from gold doc1.

they tend to overfit annotated answer spans and multi-hop reasoning paths in the training set. Consequently, we naturally pose the research question for this paper: *How can we prevent MHQA models from overfitting answer spans and reasoning paths in the training set?*

Label smoothing is an effective method for preventing overfitting, widely utilized in computer vision (Szegedy et al., 2016). In this study, we introduce label smoothing to multi-hop reasoning tasks for the first time to mitigate overfitting. We propose a simple yet efficient MHQA model, denoted as $\mathbf{R}^3$, comprising **R**etrieval, **R**efinement, and **R**eading Comprehension modules. Inspired by the F1 score, a commonly used metric for evaluating MRC task performance, we develop F1 Smoothing, a novel technique that calculates the significance of each token within the smooth distribution. Moreover, we incorporate curriculum learning (Bengio et al., 2009) and devise the **L**inear **D**ecay **L**abel Smoothing **A**lgorithm (LDLA), which gradually reduces the smoothing weight, allowing the model to focus on more challenging samples during training. Experimental results on the HotpotQA dataset (Yang et al., 2018) demonstrate that incorporating F1 smoothing and LDLA into the $\mathbf{R}^3$ model significantly enhances performance in document retrieval, supporting sentence prediction, and answer span selection, achieving state-of-the-art results among all published works.

Our main contributions are summarized as follows:

- We introduce label smoothing to multi-hop reasoning tasks and propose a baseline model, $\mathbf{R}^3$, with retrieval, refinement, and reading comprehension modules.
- We present F1 smoothing, a novel label smoothing method tailored for MRC tasks, which alleviates errors caused by answer span discrepancies.
- We propose LDLA, a progressive label smoothing algorithm integrating curriculum learning.
- Our experiments on the HotpotQA dataset demonstrate that label smoothing effectively enhances the MHQA model's performance, with our proposed LDLA and F1 smoothing achieving state-of-the-art results.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

2 Related Work

Label Smoothing Label smoothing is a regularization technique first introduced in computer vision to improve classification accuracy on ImageNet (Szegedy et al., 2016). The basic idea of label smoothing is to soften the distribution of true labels by replacing their one-hot encoding with a smoother version. This approach encourages the model to be less confident in its predictions and consider a broader range of possibilities, reducing overfitting and enhancing generalization (Pereyra et al., 2017; Müller et al., 2019; Lukasik et al., 2020a). Label smoothing has been widely adopted across various natural language processing tasks, including speech recognition (Chorowski and Jaitly, 2017), document retrieval (Penha and Hauff, 2021), dialogue generation (Saha et al., 2021), and neural machine translation (Gao et al., 2020; Lukasik et al., 2020b; Graça et al., 2019).

In addition to traditional label smoothing, several alternative techniques have been proposed in recent research. For example, Xu et al. (2020) suggested the **T**wo-**S**tage **LA**bel smoothing (TSLA) algorithm, which employs a smoothing distribution in the first stage and the original distribution in the second stage. Experimental results demonstrated that TSLA effectively promotes model convergence and enhances performance. Penha and Hauff (2021) introduced label smoothing for retrieval tasks and proposed using BM25 to compute the label smoothing distribution, which outperforms the uniform distribution. Zhao et al. (2020) proposed Word Overlapping, which uses maximum likelihood estimation (Su et al., 2020) to optimally estimate the model's training distribution.

Multi-hop Question Answering Multi-hop reading comprehension (MHRC) is a demanding task in the field of machine reading comprehension (MRC) that closely resembles the human thought process in real-world scenarios. Consequently, it has gained significant attention in the field of natural language understanding in recent years. Several datasets have been developed to foster research in this area, including HotpotQA (Yang et al., 2018), WikiHop (Welbl et al., 2018), and NarrativeQA (Kočiský et al., 2018). Among these, HotpotQA (Yang et al., 2018) is particularly representative and challenging, as it requires the model to not only extract the correct answer span from the context but also identify a series of supporting sentences as evidence for MHRC.

Recent advances in MHRC have led to the development of several graph-free models, such as QUARK (Groeneveld et al., 2020), C2FReader (Shao et al., 2020), and S2G (Wu et al., 2021), which have challenged the dominance of previous graph-based approaches like DFGN (Qiu et al., 2019), SAE (Tu et al., 2020), and HGN (Fang et al., 2020). C2FReader (Shao et al., 2020) suggests that the performance difference between graph attention and self-attention is minimal, while S2G's (Wu et al., 2021) strong performance demonstrates the potential of graph-free modeling in MHRC. FE2H (Li et al., 2022), which uses a two-stage selector and a multi-task reader, currently achieves the best performance on HotpotQA, indicating that pre-trained language models alone may be sufficient for modeling multi-hop reasoning. Motivated by the design of S2G (Wu et al., 2021) and FE2H (Li et al., 2022), we introduce a our model $\mathbf{R}^3$.

3 Framework

Figure 2 depicts the overall architecture of $\mathbf{R}^3$. The retrieval module serves as the first step, where our system selects the most relevant documents, which is essential for filtering out irrelevant information. In this example, document1, document3, and document4 are chosen due to their higher relevance scores, while other documents are filtered out. Once the question and related documents are given, the refinement module further selects documents based on their combined relevance. In this instance, the refinement module opts for document1 and document4. Following this, the question and document1, document4 are concatenated and used as input for the reading comprehension module. Within the reading comprehension module, we concurrently train supporting sentence prediction, answer span extraction, and answer type selection using a multi-task approach.

3.1 Retrieval Module

In the retrieval module, each question Q is typically accompanied by a set of M documents $D_1, D_2 \ldots, D_M$, but only $C, |C| << M$ (two in HotpotQA) are genuinely relevant to question Q.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

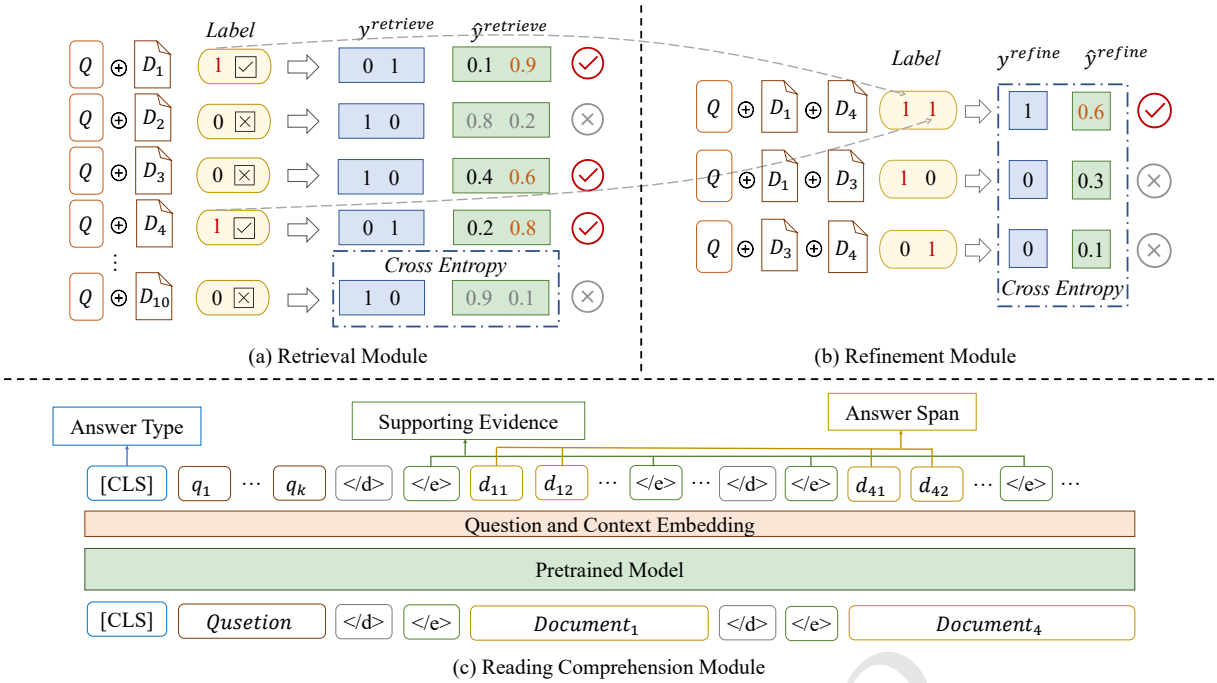


Figure 2: Overview of our $\mathbf{R}^3$ model, which consists of three main modules: **R**etrieval, **R**efinement, and **R**eading Comprehension.

We model the retrieval process as a binary classification task. Specifically, for each question-document pair, we generate an input by concatenating [CLS], question, [SEP], document, and [SEP] in sequence. We then feed the [CLS] token output from the model into a linear classifier. $\mathcal{L}_{\text{retrieve}}$ represents the cross-entropy between the predicted probability and the gold label. In contrast to S2G (Wu et al., 2021), which employs a complex pairwise learning-to-rank loss, we opt for a simple binary cross-entropy loss, as it maintains high performance while being significantly more efficient.

$$\begin{aligned}\mathcal{L}_{\text{retrieve}} = \mathbb{E}[-\frac{1}{M}\sum_{i=1}^{M}(y_i^{\text{retrieve}} \cdot log(\hat{y}_i^{\text{retrieve}}) \\ +(1-y_i^{\text{retrieve}}) \cdot log(1-\hat{y}_i^{\text{retrieve}}))],\end{aligned} \tag{1}$$

where $\hat{y}_i^{\text{retrieve}}$ is the probability predicted by the model and y_i^{retrieve} is the ground-truth label. M is the number of provided documents. $\mathbb{E}$ means the expectation of all samples.

$$y_i^{\text{retrieve}} = \begin{cases} 1 & D_i \text{ is a golden document.} \\ 0 & D_i \text{ is a non-golden document.} \end{cases} \tag{2}$$

3.2 Refinement Module

In the refinement module, we select the top K relevant documents from the previous step and form pairs, resulting in C_K^2 combinations. Emphasizing inter-document interactions crucial for multi-hop reasoning, we concatenate the following sequence: [CLS], question, [SEP], document1, [SEP], document2, [SEP]. Similar to the retrieval module, we extract the [CLS] token output from the model and pass it through a classifier. Pairs containing two gold-standard documents are labeled as 1, while others are labeled as 0. The refinement module thus filters out irrelevant documents, producing a more concise set for further processing.

$$\mathcal{L}_{\text{refine}} = \mathbb{E}[-\sum_{i=1}^{C_K^2} y_i^{\text{refine}} log(\hat{y}_i^{\text{refine}})], \tag{3}$$

where $\hat{y}_i^{\text{refine}}$ is predicted document pair probability and y_i^{refine} is the ground-truth label, C_K^2 is number of all combination.

$$y_i^{\text{refine}} = \begin{cases} 1 & \mathcal{C}_i \text{ consists of two gold documents.} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

We use a single pretrained language model as the encoder for both the retrieval and refinement module, and the final loss is a weighted sum of $\mathcal{L}_{\text{retrieve}}$ and $\mathcal{L}_{\text{refine}}$. λ_1 and λ_2 are accordingly coefficients of $\mathcal{L}_{\text{retrieve}}$ and $\mathcal{L}_{\text{refine}}$.

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{retrieve}} + \lambda_2 \mathcal{L}_{\text{refine}}. \tag{5}$$

3.3 Reading Comprehension Module

In the reading comprehension module, we use multi-task learning to simultaneously predict supporting sentences and extract answer span. HotpotQA (Yang et al., 2018) contains samples labeled as "yes" or "no". The practice of splicing "yes" and "no" tokens at the beginning of the sequence (Li et al., 2022) could corrupt the original text's semantic information. To avoid the impact of irrelevant information, we introduce an answer type selection header trained with a cross-entropy loss function.

$$\mathcal{L}_{\text{type}} = \mathbb{E}[-\sum_{i=1}^{3} y_i^{\text{type}} log(\hat{y}_i^{\text{type}})], \tag{6}$$

where $\hat{y}_i^{\text{fine}}$ denotes the predicted probability of answer type generated by our model, and y_i^{fine} represents the ground-truth label. answer type includes "yes", "no" and "span".

$$y_i^{\text{type}} = \begin{cases} 0 & \text{Answer is no.} \\ 1 & \text{Answer is yes.} \\ 2 & \text{Answer is a span.} \end{cases} \tag{7}$$

To extract the span of answers, we use a linear layer on the contextual representation to identify the start and end positions of answers, and adopts cross-entropy as the loss function. The corresponding loss terms are denoted as $\mathcal{L}_{\text{start}}$ and $\mathcal{L}_{\text{end}}$ respectively. Similar to previous work S2G (Wu et al., 2021) and FE2H (Li et al., 2022), we also inject a special placeholder token $< /e >$ and use a linear binary classifier on the output of $< /e >$ to determine whether a sentence is a supporting fact. The classification loss of the supporting facts is denoted as $\mathcal{L}_{\text{sup}}$, and we jointly optimize all of these objectives in our model.

$$\mathcal{L}_{\text{reading}} = \lambda_3 \mathcal{L}_{\text{type}} + \lambda_4 (\mathcal{L}_{\text{start}} + \mathcal{L}_{\text{end}}) + \lambda_5 \mathcal{L}_{\text{sup}}. \tag{8}$$

4 Label Smoothing

Label smoothing is a regularization technique that aims to improve generalization in a classifier by modifying the ground truth labels of the training data. In the one-hot setting, the probability of the correct category $q(y|x)$ for a training sample (x, y) is typically defined as 1, while the probabilities of all other categories $q(\neg y|x)$ are defined as 0. The cross-entropy loss function used in this setting is typically defined as follows:

$$\mathcal{L} = -\sum_{k=1}^{K} q(k|x) \log(p(k|x)), \tag{9}$$

where $p(k|x)$ is the probability of the model's prediction for the k-th class. Specifically, label smoothing mixes $q(k|x)$ with a uniform distribution $u(k)$, independent of the training samples, to produce a new distribution $q'(k|x)$.

$$q'(k|x) = (1 - \epsilon) q(k|x) + \epsilon u(k), \tag{10}$$

where ϵ is the weight controls the importance of $q(k|x)$ and $u(k)$ in the resulting distribution. $u(k)$ is construed as $\frac{1}{K}$ of the uniform distribution, where K is the total number of categories. Next, we introduce two novel label smoothing methods.

Algorithm 1 Linear Decay Label Smoothing.

Require: training epochs $n > 0$; smoothing weight $\epsilon \in [0, 1]$; decay rate $\tau \in [0, 1]$; uniform distribution u

Initialize: Model parameter $w_0 \in \mathcal{W}$;
Input: Optimization algorithm $\mathcal{A}$
for $i = 0, 1, \ldots, n$ **do**
  $\epsilon_i \leftarrow \epsilon - i\tau$
  if $\epsilon_i < 0$ **then**
    $\epsilon_i \leftarrow 0$
  end if
  sample(x_t, y_t)
  $y_t^{LS} \leftarrow (1 - \epsilon_i)y_i + \epsilon u$
  $w_{i+1} \leftarrow \mathcal{A}{-}step(w_i; x_i, y_i^{LS})$
end for

4.1 Linear Decay Label Smoothing

Our proposed Linear Decay Label Smoothing Algorithm (LDLA) addresses the abrupt changes in training distribution caused by the two-stage approach of TSLA, which can negatively impact the training process. In contrast to TSLA, LDLA decays the smoothing weight at a constant rate per epoch, promoting a more gradual learning process.

Given a total of n epochs in the training process and a decay size of τ, the smoothing weight ϵ for the i-th epoch can be calculated as follows:

$$\epsilon_i = \begin{cases} \epsilon - i\tau & \epsilon - i\tau \geq 0 \\ 0 & \epsilon - i\tau < 0 \end{cases} \tag{11}$$

Algorithm 1 outlines the specific steps of the LDLA algorithm. LDLA employs the concept of curriculum learning by gradually transitioning the model's learning target from a smoothed distribution to the original distribution throughout the training process. This approach incrementally reduces uncertainty during training, enabling the model to progressively concentrate on more challenging samples and transition from learning with uncertainty to certainty. Consequently, LDLA fosters more robust and effective learning.

4.2 F1 Smoothing

Unlike traditional classification tasks, MRC requires identifying both the start and end positions of a span. To address the specific nature of this task, a specialized smoothing method is required to achieve optimal results. In this section, we introduce F1 Smoothing, a technique that calculates the significance of a span based on its F1 score.

Consider a sample x that contains a context S and an answer a_{gold}. The total length of the context is denoted by L. We use $q_s(t|x)$ to denote the F1 score between a span of arbitrary length starting at position t in S and the ground truth answer a_{gold}. Similarly, $q_e(t|x)$ denotes the F1 score between a_{gold} and a span of arbitrary length ending at position t in S .

$$q_s(t|x) = \sum_{\xi=t}^{L-1} F1\left((t, \xi), a_{\text{gold}}\right). \tag{12}$$

$$q_e(t|x) = \sum_{\xi=0}^{t} F1\left((\xi, t), a_{\text{gold}}\right). \tag{13}$$

The normalized distributions are noted as $q_s^{'}(t|x)$ and $q_e^{'}(t|x)$, respectively.

$$q_s^{'}(t|x) = \frac{exp(q_s(t|x))}{\sum_{i=0}^{L-1} exp(q_s(i|x))}. \tag{14}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

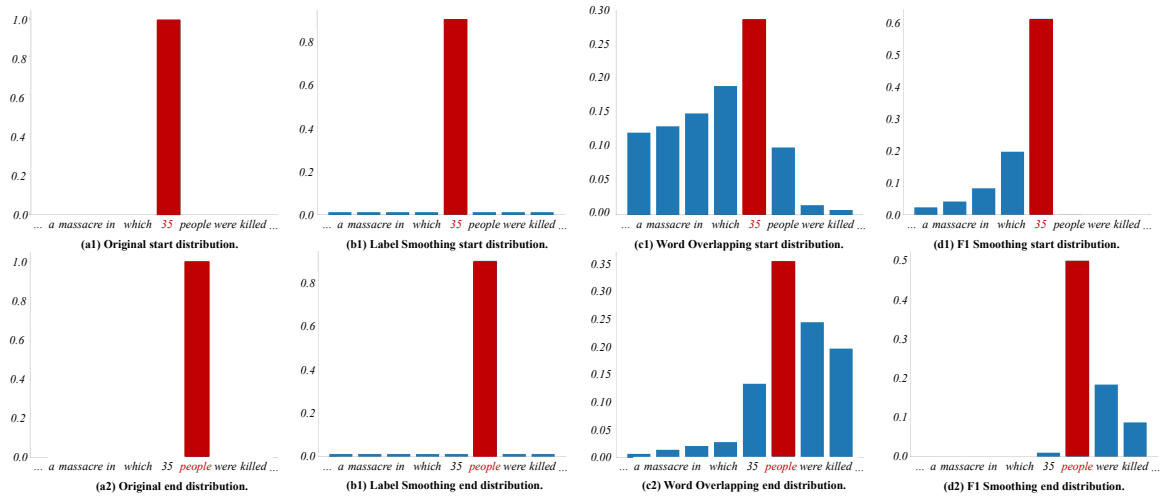


Figure 3: Visualization of original distribution and different label smoothing distributions, including Label Smoothing, Word Overlapping, and F1 Smoothing. The first row shows the distribution of the start token, and the second row shows the distribution of the end token. The gold start and end tokens are highlighted in red.

$$q_e^{'}(t|x) = \frac{exp(q_e(t|x))}{\sum_{i=0}^{L-1} exp(q_e(i|x))}. \tag{15}$$

To decrease the computational complexity of F1 Smoothing, we present a computationally efficient version in Appendix 7. Previous research (Zhao et al., 2020) has investigated various label smoothing methods for MRC, encompassing traditional label smoothing and word overlap smoothing. As illustrated in Figure 3, F1 Smoothing offers a more accurate distribution of token importance in comparison to Word Overlap Smoothing. This method reduces the probability of irrelevant tokens and prevents the model from being misled during training.

5 Experiment

5.1 Dataset

We evaluate our approach on the distractor setting of HotpotQA (Yang et al., 2018), a multi-hop question-answer dataset with 90k training samples, 7.4k validation samples, and 7.4k test samples. Each question in this dataset is provided with several candidate documents, two of which are labeled as gold. In addition to this, HotpotQA also provides supporting sentences for each question, encouraging the model to explain the inference path of the multi-hop question-answer. We use the Exact Match (EM) and F1 score (F1) to evaluate the performance of our approach in terms of document retrieval, supporting sentence prediction, and answer extraction.

5.2 Implementation Details

Our model is built using the Pre-trained language models (PLMs) provided by HuggingFace's Transformers library (Wolf et al., 2020).

Retrieval and Refinement Module We used RoBERTa-large (Liu et al., 2019) and ELECTRA-large (Clark et al., 2020) as our PLMs and conducted an ablation study on RoBERTa-large (Liu et al., 2019). Training on a single RTX3090 GPU, we set the number of epochs to 8 and the batch size to 16. We employed the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 5e-6 and a weight decay of 1e-2.

Reading Comprehension Module We utilized RoBERTa-large (Liu et al., 2019) and DeBERTa-v2-xxlarge (He et al., 2021) as our PLMs, performing ablation studies on RoBERTa-large (Liu et al., 2019). To train RoBERTa-large, we used an RTX3090 GPU, setting the number of epochs to 16 and the batch

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

Model	Answer		Supporting	
	EM	F1	EM	F1
Baseline Model (Yang et al., 2018)	45.60	59.02	20.32	64.49
QFE (Nishida et al., 2019)	53.86	68.06	57.75	84.49
DFGN (Qiu et al., 2019)	56.31	69.69	51.50	81.62
SAE-large (Tu et al., 2020)	66.92	79.62	61.53	86.86
C2F Reader (Shao et al., 2020)	67.98	81.24	60.81	87.63
HGN-large (Fang et al., 2020)	69.22	82.19	62.76	88.47
FE2H on ELECTRA (Li et al., 2022)	69.54	82.69	64.78	88.71
AMGN+ (Li et al., 2021)	70.53	83.37	63.57	88.83
S2G+EGA (Wu et al., 2021)	70.92	83.44	63.86	88.68
FE2H on ALBERT (Li et al., 2022)	71.89	**84.44**	64.98	89.14
$\mathbf{R}^3$ (ours)	71.27	83.57	65.25	88.98
Smoothing $\mathbf{R}^3$ (ours)	**72.07**	84.34	**65.44**	**89.55**

Table 1: In the distractor setting of the HotpotQA test set, our proposed F1 Smoothing and LDLA has led to significant improvements in the performance of the Smoothing $\mathbf{R}^3$ model compared to the $\mathbf{R}^3$ model. Furthermore, the Smoothing $\mathbf{R}^3$ model has outperformed a number of strong baselines and has achieved the highest results.

Model	EM	F1
SAE_{large} (Tu et al., 2020)	91.98	95.76
S2G_{large} (Wu et al., 2021)	95.77	97.82
FE2H_{large} (Li et al., 2022)	96.32	98.02
$\mathbf{R}^3$ (ours)	96.50	98.10
Smoothing $\mathbf{R}^3$	**96.85**	**98.32**

Table 2: Comparison of our retrieval and refinement module with previous baselines on HotpotQA dev set. Label smoothing can further enhance model performance.

size to 16. For the larger DeBERTa-v2-xxlarge model, we employed an A100 GPU, setting the number of epochs to 8 and the batch size to 16. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 4e-6 for RoBERTa-large and 2e-6 for DeBERTa-v2-xxlarge, along with a weight decay of 1e-2 for optimization.

5.3 Experimental Results

We utilize ELECTRA-large (Clark et al., 2020) as the PLM for the retrieval and refinement modules, and DeBERTa-v2-xxlarge for the reading comprehension module. The $\mathbf{R}^3$ model incorporating F1 Smoothing and LDLA methods is referred to as Smoothing $\mathbf{R}^3$. LDLA is employed for document retrieval and supporting sentence prediction, while F1 Smoothing is applied for answer span extraction. As shown in Table 1, Smoothing $\mathbf{R}^3$ achieves improvements of 0.8% and 0.77% in EM and F1 for answers, and 0.19% and 0.57% in EM and F1 for supporting sentences compared to the $\mathbf{R}^3$ model. Among the tested label smoothing techniques, F1 smoothing and LDLA yield the most significant performance improvement.

We compare the performance of our retrieval and refinement module, which uses ELECTRA-large as a backbone, to three advanced works: SAE (Tu et al., 2020), S2G (Wu et al., 2021), and FE2H (Li et al., 2022). These methods also employ sophisticated selectors for retrieving relevant documents. We evaluate the performance of document retrieval using the EM and F1 metrics. Table 2 demonstrates that our $\mathbf{R}^3$ method outperforms these three strong baselines, with Smoothing $\mathbf{R}^3$ further enhancing performance.

In Table 3, we evaluate the performance of the reading comprehension module, which employs DeBERTa-v2-xxlarge (He et al., 2021) as the backbone, on documents retrieved by the retrieval and

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

Model	Answer		Supporting	
	EM	F1	EM	F1
SAE	67.70	80.75	63.30	87.38
S2G	70.80	-	65.70	-
$\mathbf{R}^3$	71.39	83.84	66.32	89.54
Smoothing $\mathbf{R}^3$	**71.89**	**84.65**	**66.75**	**90.08**

Table 3: Performances of cascade results on the dev set of HotpotQA in the distractor setting.

Setting	EM	F1
Baseline	95.93±.05	97.91±.09
LS	96.06±.11	97.94±.04
TSLA	96.21±.01	98.05±.05
LDLA	**96.57**±.05	**98.18**±.04

Table 4: Various label smoothing methods applied to retrieval modules.

Setting	EM	F1
Baseline	66.94±.05	90.50±.02
LS	66.88±.02	90.53±.02
TSLA	67.42±.05	90.72±.05
LDLA	**67.63**±.04	**90.85**±.03

Table 5: Various label smoothing methods applied to supporting sentence prediction.

refinement module. Our $\mathbf{R}^3$ model outperforms strong baselines SAE and S2G, and further improvements are achieved by incorporating F1 Smoothing and LDLA. These results emphasize the potential for enhancing performance through the application of label smoothing techniques.

5.4 Label Smoothing Analysis

In our study of the importance of label smoothing, we used RoBERTa-large (Liu et al., 2019) as the backbone for our model. To ensure the reliability of our experimental results, we conducted multiple runs with different random number seeds (41, 42, 43, and 44) to ensure stability.

In our experiments, we compared three label smoothing strategies: Label Smoothing (LS), Two-Stage Label smoothing (TSLA), and Linear Decay Label smoothing (LDLA). The initial value of ϵ in our experiments was 0.1, and in the first stage of TSLA, the number of epochs was set to 4. For each epoch in LDLA, ϵ was decreased by 0.01.

Retrieval Module As shown in Table 4, label smoothing effectively enhances the generalization performance of the retrieval module. LDLA outperforms TSLA with a higher EM (0.36%) and F1 score (0.13%), demonstrating superior generalization capabilities.

Supporting Sentence Prediction We assess the impact of label smoothing on the supporting sentence prediction task. The results presented in Table 5 indicate that TSLA exhibits an increase of 0.48% in EM and 0.22% in F1 compared to the baseline. Additionally, LDLA further enhances the performance by 0.21% in EM and 0.13% in F1 when compared to TSLA.

Answer Span Extraction Table 6 highlights the impact of label smoothing methods on answer span extraction in the reading comprehension module. LS, TSLA, and LDLA exhibit slight improvements compared to the baseline. The advanced Word Overlapping technique demonstrates an average improvement of 0.49% in EM and 0.47% in F1, respectively, compared to the baseline. In contrast, our proposed F1 Smoothing technique achieves an average EM improvement of 0.82% and an average F1 score improvement of 0.84%. These results suggest that F1 Smoothing can enhance performance on MRC tasks more effectively than other smoothing techniques.

5.5 Error Analysis

To gain a deeper understanding of how label smoothing effectively enhances model performance, we examined the model's output on the validation set, focusing on answer span errors and multi-hop reasoning errors. First, we define these two types of errors as follows:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

Methods	EM	F1
Baseline	69.11±.02	82.21±.03
LS	69.30±.02	82.56±.09
TSLA	69.32±.10	82.66±.09
LDLA	69.39±.12	82.69±.03
Word Overlapping	69.60±.09	82.68±.13
F1 Smoothing	**69.93**±.07	**83.05**±.10

Table 6: Analysis of different label smoothing methods for Answer Span Extraction.

Model	Answer Span Errors	Multi-Hop Reasoning Errors
S2G	1612	550
$\mathbf{R}^3$	1556	562
Smoothing $\mathbf{R}^3$	1536 ($\downarrow$ 1.3%)	545($\downarrow$ 3.0%)

Table 7: Error analysis on Answer Span Errors and Multi-hop Reasoning Errors.

- Answer Span Errors: The predicted answer and the annotated answer have a partial overlap after removing stop words, but are not identical.
- Multi-hop Reasoning Errors: Due to reasoning errors, the predicted answer and the annotated answer are entirely different.

By implementing label smoothing, as shown in Table 7, Smoothing $\mathbf{R}^3$ experienced a 1.3% reduction in answer span errors, decreasing from 1556 to 1536, and a 3.0% decrease in multi-hop reasoning errors, dropping from 562 to 545. Smoothing $\mathbf{R}^3$ shows a significant reduction in both types of errors compared to the S2G model. This finding suggests that incorporating label smoothing during training can effectively prevent the model from overfitting the answer span and reasoning paths in the training set, thereby improving the model's generalization capabilities and overall performance.

6 Conclusion

In this study, we first identify the primary challenges hindering the performance of MHQA systems and propose using label smoothing to mitigate overfitting issues during MHQA training. We introduce F1 smoothing, a novel smoothing method inspired by the widely-used F1 score in MRC tasks. Additionally, we present LDLA, a progressive label smoothing algorithm that incorporates the concept of curriculum learning. Comprehensive experiments on the HotpotQA dataset demonstrate that our proposed model, Smoothing $\mathbf{R}^3$, achieves significant performance improvement when using F1 smoothing and LDLA. Our findings indicate that label smoothing is a valuable technique for MHQA, effectively improving the model's generalization while minimizing overfitting to particular patterns in the training set.

Acknowledgement

We would like to express our heartfelt thanks to the students and teachers of Fudan Natural Language Processing Lab. Their thoughtful suggestions, viewpoints, and enlightening discussions have made significant contributions to this work. We also greatly appreciate the strong support from Huawei Poisson Lab for our work, and their invaluable advice. We are sincerely grateful to the anonymous reviewers and the domain chairs, whose constructive feedback played a crucial role in enhancing the quality of our research. This work was supported by the National Key Research and Development Program of China (No.2022CSJGG0801), National Natural Science Foundation of China (No.62022027) and CAAI-Huawei MindSpore Open Fund.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Jan Chorowski and Navdeep Jaitly. 2017. Towards better decoding and language model integration in sequence to sequence models. In *INTERSPEECH*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Yingbo Gao, Weiyue Wang, Christian Herold, Zijian Yang, and Hermann Ney. 2020. Towards a better understanding of label smoothing in neural machine translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 212–223, Suzhou, China. Association for Computational Linguistics.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy. Association for Computational Linguistics.

Dirk Groeneveld, Tushar Khot, Mausam, and Ashish Sabharwal. 2020. A simple yet strong pipeline for HotpotQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8839–8845, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv preprint*, abs/2111.09543.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Ronghan Li, Lifang Wang, Shengli Wang, and Zejun Jiang. 2021. Asynchronous multi-grained graph network for interpretable multi-hop reading comprehension. In *IJCAI*, pages 3857–3863.

Xin-Yi Li, Wei-Jun Lei, and Yu-Bin Yang. 2022. From easy to hard: Two-stage selector and reader for multi-hop question answering. *ArXiv preprint*, abs/2205.11729.

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

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

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. 2020a. Does label smoothing mitigate label noise? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6448–6458. PMLR.

Michal Lukasik, Himanshu Jain, Aditya Menon, Seungyeon Kim, Srinadh Bhojanapalli, Felix Yu, and Sanjiv Kumar. 2020b. Semantic label smoothing for sequence to sequence problems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4992–4998, Online. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2335–2345, Florence, Italy. Association for Computational Linguistics.

Gustavo Penha and Claudia Hauff. 2021. Weakly supervised label smoothing. In *European Conference on Information Retrieval*, pages 334–341. Springer.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150, Florence, Italy. Association for Computational Linguistics.

Sougata Saha, Souvik Das, and Rohini Srihari. 2021. Similarity based label smoothing for dialogue generation. *ArXiv preprint*, abs/2107.11481.

Nan Shao, Yiming Cui, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Is Graph Structure Necessary for Multi-hop Question Answering? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7187–7192, Online. Association for Computational Linguistics.

Lixin Su, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Label distribution augmented maximum likelihood estimation for reading comprehension. In James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang, editors, *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 564–572. ACM.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bohong Wu, Zhuosheng Zhang, and Hai Zhao. 2021. Graph-free multi-hop reading comprehension: A select-to-guide strategy. *ArXiv preprint*, abs/2107.11823.

Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, and Rong Jin. 2020. Towards understanding label smoothing. *ArXiv preprint*, abs/2006.11653.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhenyu Zhao, Shuangzhi Wu, Muyun Yang, Kehai Chen, and Tiejun Zhao. 2020. Robust machine reading comprehension by learning soft labels. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2754–2759, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 611-623, Harbin, China, August 3 - 5, 2023.

7 Appendix A

In order to alleviate the complexity introduced by multiple for loops in the F1 Smoothing method, we have optimized Eq. (12) and Eq. (13). We use $L_a = e^* - s^* + 1$ and $L_p = e - s + 1$ to denote respectively the length of gold answer and predicted answer.

$$q_s(t|x) = \sum_{\xi=t}^{L-1} \mathrm{F1}\left((t,\xi), a_{\text{gold}}\right). \tag{16}$$

If $t < s^*$, the distribution is

$$q_s(t|x) = \sum_{\xi=s^*}^{e^*} \frac{2(\xi - s^* + 1)}{L_p + L_a} + \sum_{\xi=e^*+1}^{L-1} \frac{2L_a}{L_p + L_a}, \tag{17}$$

else if $s^* \leq t \leq e^*$, we have the following distribution

$$q_s(t|x) = \sum_{\xi=s}^{e^*} \frac{2L_p}{L_p + L_a} + \sum_{\xi=e^*+1}^{L-1} \frac{2(e^* - s + 1)}{L_p + L_a}. \tag{18}$$

In equation 17 and 18, $L_p = e - i + 1$.

We can get $q_e(t|x)$ similarly. If $t > e^*$,

$$q_e(t|x) = \sum_{\xi=s^*}^{e^*} \frac{2(e^* - \xi + 1)}{L_p + L_a} + \sum_{\xi=0}^{s^*-1} \frac{2L_a}{L_p + L_a}, \tag{19}$$

else if $s^* \leq t \leq e^*$,

$$q_e(t|x) = \sum_{\xi=s^*}^{e} \frac{2L_p}{L_p + L_a} + \sum_{\xi=0}^{s^*-1} \frac{2(e - s^* + 1)}{L_p + L_a}. \tag{20}$$

In equation 19 and 20, $L_p = i - s + 1$.

Improving Zero-shot Cross-lingual Dialogue State Tracking via Contrastive Learning

Yu Xiang[1], **Ting Zhang**[2], **Hui Di**[3], **Hui Huang**[4], **Chunyou Li**[1], **Kazushige Ouchi**[3], **Yufeng Chen**[1], **Jinan Xu**[1*]

[1] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China

[2] Global Tone Communication Technology Co., Ltd.

[3] Toshiba (China) Co., Ltd. [4] Harbin Institute of Technology

{21120422,21120368,chenyf,jaxu}@bjtu.edu.cn;
zhangting01@gtcom.com.cn;dihui@toshiba.com.cn;
22b903058@stu.hit.edu.cn;kazushige.ouchi@toshiba.co.jp

Abstract

Recent works in dialogue state tracking (DST) focus on a handful of languages, as collecting large-scale manually annotated data in different languages is expensive. Existing models address this issue by code-switched data augmentation or intermediate fine-tuning of multilingual pre-trained models. However, these models can only perform implicit alignment across languages. In this paper, we propose a novel model named **C**ontrastive **L**earning for **C**ross-**L**ingual DST (CLCL-DST) to enhance zero-shot cross-lingual adaptation. Specifically, we use a self-built bilingual dictionary for lexical substitution to construct multilingual views of the same utterance. Then our approach leverages fine-grained contrastive learning to encourage representations of specific slot tokens in different views to be more similar than negative example pairs. By this means, CLCL-DST aligns similar words across languages into a more refined language-invariant space. In addition, CLCL-DST uses a significance-based keyword extraction approach to select task-related words to build the bilingual dictionary for better cross-lingual positive examples. Experiment results on Multilingual WoZ 2.0 and parallel MultiWoZ 2.1 datasets show that our proposed CLCL-DST outperforms existing state-of-the-art methods by a large margin, demonstrating the effectiveness of CLCL-DST.

1 Introduction

Dialogue state tracking is an essential part of task-oriented dialogue systems (Zhong et al., 2018), which aims to extract user goals or intentions throughout a dialogue process and encode them into a compact set of dialogue states, i.e., a set of slot-value pairs. In recent years, DST models have achieved impressive success with adequate training data. However, most models are restricted to monolingual scenarios since collecting and annotating task-oriented dialogue data in different languages is time-consuming and costly (Chen et al., 2018). It is necessary to investigate how to migrate a high-performance dialogue state tracker to different languages when no annotated target language dialogue data are available.

Previous approaches are generally divided into the following three categories: (1) Data augmentation methods with neural machine translation system (Schuster et al., 2019). Although translating dialogue corpora using machine translation is straightforward, it has inherent limitation of heavily depending on performance of machine translation. (2) Pre-trained cross-lingual representation (Lin and Chen, 2021). The approach applies a cross-lingual pre-trained model, such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) as one of the components of the DST architecture and then is trained with task data directly. However, the approach does not introduce cross-lingual information during the training process. (3) Code-switched data augmentation (Liu et al., 2020a; Liu et al., 2020b; Qin et al., 2021). The method replaces words randomly from the source language to the target language with a bilingual dictionary as a way to achieve data

* Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

augmentation. Nevertheless, a synonym substitution with some meaningless words may introduce noise that impairs the semantic coherence of the sentence. Besides, the model only use the code-switched corpus as the training data, ignoring the interaction between the original and code-switched sentences. Consequently, these models can not sufficiently learn the semantic representation of the corpus.

To address the above-mentioned issues, we propose a novel model named **C**ontrastive **L**earning for **C**ross-**L**ingual DST (CLCL-DST), which utilizes contrastive learning (CL) for cross-lingual adaptation. CLCL-DST first captures comprehensive cross-lingual information from different perspectives and explores the consistency of multiple views through contrastive learning (Lai et al., 2021). Simultaneously, as dialogue state tracking is to predict the state of slots in each turn of the dialogue, we consider it as a token-level task and then employ the same fine-grained CL. Specifically, we obtain the encoded feature representation of each slot in the original sentence and the corresponding code-switched sentence from the multilingual pre-trained model, respectively. We then employ fine-grained CL to align the representations of slot tokens in different views. By introducing CL, Our model is able to distinguish between the code-switched utterance and a set of negative samples, thus encouraging representations of similar words in different languages to align into a language-invariant feature space (Subsection 3.1).

Furthermore, CLCL-DST introduces a significance-based keyword extraction approach to obtain task-related keywords with high significance scores in different domains. For example, in the price range domain, some words like “cheap”, “moderate” and “expensive” are more likely to have higher significance scores than background words, such as “a”, “is” and “do”. Specifically, Our approach obtains the semantic representation of sentences and corresponding subwords by encoder. Then the approach gets the significance scores of the words by calculating the cosine similarity and get the keywords of the dataset based on the scores. We then replace these keywords with the corresponding words in the target language to generate multilingual code-switched data pairs. These code-switched keywords can be considered as cross-lingual views sharing the same meaning, allowing the shared encoder to learn some direct bundles of meaning in different languages. Thus, our keyword extraction approach facilitates the transfer of cross-lingual information and strengthens the ties across different languages (Subsection 3.2).

We evaluate our model on two benchmark datasets. For the Multilingual WoZ 2.0 dataset (Mrkšić et al., 2017) which is single-domain, our model outperforms the existing state-of-the-art model by 4.1% and 4.8% slot accuracy for German (De) and Italian (It) under the zero-shot setting, respectively. For the parallel MultiWoZ 2.1 dataset (Gunasekara, 2021) which is multi-domain, our method outperforms the current state-of-the-art by 22% and 38.7% in joint goal accuracy and slot f1 for Chinese (Zh), respectively. Moreover, further experiments show that introducing fine-grained CL performs better than coarse-grained CL. We also investigate the impact of different keyword extraction approaches on the model to demonstrate the superiority of our extraction approach.

Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first work on DST that leverages fine-grained contrastive learning to explicitly align representations across languages.
- We propose to utilize a significance-based keyword selection approach to select task-related keywords for code-switching. By constructing cross-lingual views through these keywords makes the model more effective in transferring cross-lingual signals.
- Our CLCL-DST model achieves state-of-the-art results on single-domain cross-lingual DST tasks, and it boasts the unique advantage of performing effective zero-shot transfer under the multi-domain cross-lingual setting, demonstrating the effectiveness of CLCL-DST.

2 Related Work

2.1 Dialogue State Tracking

Methods of dialogue state tracking can be divided into two categories, ontology-based and open-vocabulary DST. The first method selects the possible values for each slot directly from a pre-defined ontology and the task can be seen as a value classification task for each slot (Lee et al., 2019; Goel et

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

al., 2019; Lin et al., 2021; Wang et al., 2022). However, in practical applications, it is difficult to define all possible values of slots in advance, and the computational complexity increases significantly with the size of the ontology.

The open-vocabulary approach attempts to solve the above problems by extracting or generating slot values directly from the dialogue history (Ren et al., 2019). (Wu et al., 2019) generates slot values directly for each slot at every dialogue turn. The model uses GRU to encode the dialogue history and decode the value with a copy mechanism. Some recent works (Kim et al., 2020; Zeng and Nie, 2020b) adopt a more efficient approach by decomposing DST into two tasks: state operation prediction and value generation. SOM-DST (Kim et al., 2020) firstly predicts state operation on each slot and then generates the value of the slot that needs updating. (Zeng and Nie, 2020a) proposes a framework based on the architecture of SOM-DST, with a single BERT as both the encoder and the decoder.

2.2 Zero-shot Cross-Lingual Dialogue State Tracking

There is a growing demand for dialogue systems supporting different languages, which requires large-scale training data with high quality. However, these data are only available within a few languages. It remains a challenge to migrate dialogue state tracker from the source language to the target language.

Cross-lingual dialogue state tracking can be divided into two categories: single-domain and multi-domain. In single-domain, XL-NBT (Chen et al., 2018) first implements cross-lingual learning under the zero-shot setting by pre-training a dialogue state tracker for the source language using a teacher network. MLT (Liu et al., 2020a) adopts a code-mixed data augmentation framework, leveraging attention mechanism to obtain the code-mixed training data for learning the interlingual semantics across different languages. CLCSA (Qin et al., 2021) further explores the dynamic replacement of words from source language to target language during training. Based on CLCSA architecture, XLIFT-DST (Moghe et al., 2021) improves the performance by intermediate fine-tuning of pre-trained multilingual models using parallel and conversational movie subtitles datasets.

In multi-domain, the primary benchmark is the Parallel MultiWoZ 2.1 dataset (Gunasekara, 2021) originating from the Ninth Dialogue Systems and Technologies Challenge (DSTC-9) (Gunasekara, 2021). This challenge is designed to build a dialogue state tracker to evaluate a low-resource target language dataset using the learned knowledge of the source language. All the submissions in this challenge use the translated version of the dataset, transforming the problem into a monolingual dialogue state tracking task. XLIFT-DST employs SUMBT (Lee et al., 2019) as the base architecture and achieves competitive results on the parallel MultiWoZ 2.1 dataset through intermediate fine-tuning. Unlike these works, we leverage code-switched data with CL to further align multiple language representations under the zero-shot setting.

2.3 Contrastive Learning

Contrastive learning aims at pulling close semantically similar examples (positive samples) and pushing apart dissimilar examples (negative samples) in the representation space. SimCSE (Gao et al., 2021) proposes a simple dropout approach to construct positive samples and achieves state-of-the-art results in semantic textual similarity tasks. Cline (Wang et al., 2021) constructs semantically negative instances without supervision to improve the robustness of the model against semantically adversarial attacks. GL-CLEF (Qin et al., 2022) leverages bilingual dictionaries to generate code-switched data as positive samples, and incorporates different grained contrastive learning to achieve cross-lingual transfer. Our model incorporates fine-grained CL to align similar representations between the source and target languages.

3 Methodology

In this section, we set up the notations that run throughout the paper first, before describing our CLCL-DST model which explicitly uses contrastive learning to achieve cross-lingual alignment in dialogue state tracking. Then, we introduce a significance-based code-switching approach on how to select task-related keywords in the utterance and code-switch the input sentence dynamically in detail. The main

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

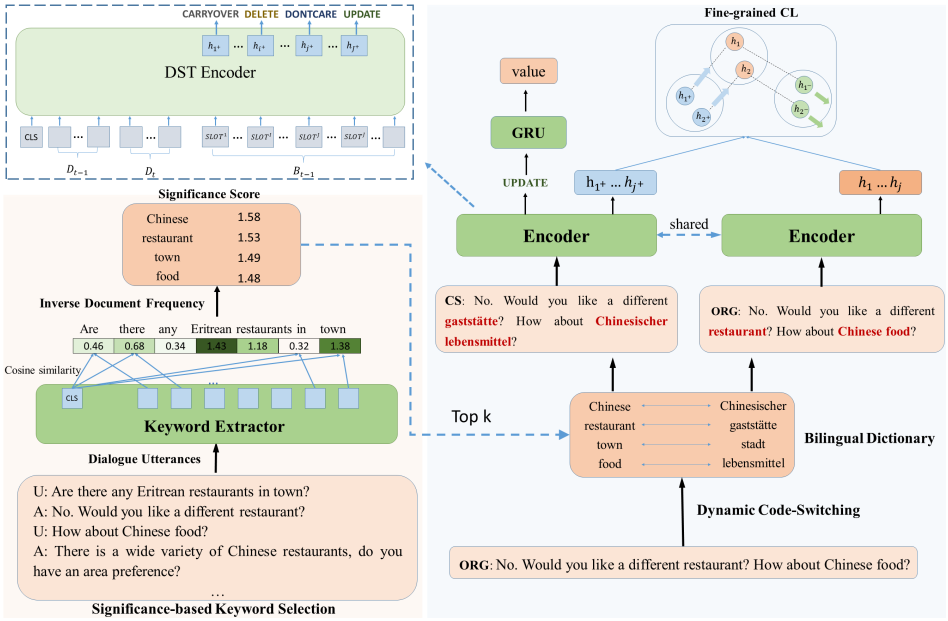


Figure 1: The overview of the proposed CLCL-DST. The input of our model consists of previous turn dialogue utterances D_{t-1}, current turn dialogue utterances D_t and previous dialogue state B_{t-1}. For simplicity, we only put one turn of dialogue on the picture. The model constructs a bilingual dictionary by obtaining keywords from the significance-based code-switching approach, and then generates code-switched data. The data are fed to the encoder to obtain a feature representation of each slot subsequently. **ORG** denotes the original sentence and **CS** denotes the corresponding code-switched sentence. In the part of **Fine-grained CL**, different color denotes different representation spaces for origin utterance, positive and negative samples. The decoder generates the value for the slot whose state operation is predicted to **UPDATE**.

architecture of our model is illustrated in Figure 1.

Notation. Suppose the dialogue has T turns. We define the dialogue utterance at turn t as $D_t = R_t \oplus ; \oplus U_t \oplus [\text{SEP}]$, where R_t and $U_t (1 \leq t \leq T)$ are the system response and the user utterance respectively. $\oplus$ denotes token concatenation, and the semicolon ; is a separation symbol, while [SEP] marks the end boundary of the dialogue. Besides, we represent the dialogue states as $B = \{B_1, ..., B_T\}$, where $B_t = [\text{SLOT}]^1 \oplus b_t^1 \oplus ... \oplus [\text{SLOT}]^{\text{I}} \oplus b_t^I$ denotes I states combination at the t-th turn. I is the total number of slots. The i-th slot-value pair b_t^i is defined as:

$$b_t^i = S^i \oplus - \oplus V_t^i, \tag{1}$$

where S^i is a slot and V_t^i is the corresponding slot value. $[\text{SLOT}]^{\text{i}}$ and $-$ are separation symbols. The representations at $[\text{SLOT}]^{\text{i}}$ position are used for state operation prediction and contrastive learning. We use the same special token [SLOT] for all $[\text{SLOT}]^{\text{i}}$. The input tokens in CLCL-DST are spliced by previous turn dialogue utterance D_{t-1}, current turn dialogue utterance D_t and previous turn dialogue states B_{t-1} (Kim et al., 2020):

$$X_t = [\text{CLS}] \oplus D_{t-1} \oplus D_t \oplus B_{t-1}, \tag{2}$$

where [CLS] is a special token to mark the start of the context. Next, we will elaborate each part in detail.

3.1 Fine-grained Contrastive Learning Framework

We introduce our fine-grained contrastive learning framework (CLCL-DST) with an encoder-decoder architecture consisting of two modules: state operation prediction and value generation. The encoder, i.e., state operation predictor, uses a multilingual pre-trained model to predict the type of the operations

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

to be performed on each slot. The decoder, i.e., slot value generator, generates values for those selected slots.

Encoder The encoder of CLCL-DST is based on mBERT architecture. We feed the code-switched sentence $X_{t,cs}$ into the encoder and obtain the output representation $H_{t,cs} \in \mathbb{R}^{|X_t| \times d}$, where $h_{t,cs}^{[\mathrm{CLS}]}, h_{t,cs}^{[\mathrm{SLOT}]^i} \in \mathbb{R}^d$ are the outputs corresponding to [CLS] and $[\mathrm{SLOT}]^i$. $h_{t,cs}^{[\mathrm{SLOT}]^i}$ is passed into a four-way classification layer to calculate the probability $P_{enc,t}^i \in \mathbb{R}^{|\mathcal{O}|}$ of operations in the i-th slot at the t-th turn:

$$P_{enc,t}^i = \mathrm{softmax}\left(W_{enc} h_{t,cs}^{[\mathrm{SLOT}]^i} + b\right), \tag{3}$$

where W_{enc} and b are learnable parameters. $\mathcal{O} = \{\mathrm{CARRYOVER}, \mathrm{DELETE}, \mathrm{DONTCARE}, \mathrm{UPDATE}\}$ denotes four state operations of each slot (Kim et al., 2020). Specifically, CARRYOVER indicates that the slot value remains unchanged; DELETE changes the value to NULL; and DONTCARE means that the slot is not important at this turn and does not need to be tracked (Wu et al., 2019). Only when the UPDATE is predicted does the decoder generate a value for the corresponding slot.

Our main learning objective is to train the encoder to match predicted state operation with the ground truth operation. So the loss for state operation is formulated as:

$$\mathcal{L}_{enc,t} = -\frac{1}{I} \sum_{i=1}^{I} \left(Y_{enc,t}^i\right)^\top \log\left(P_{enc,t}^i\right), \tag{4}$$

where $Y_{enc,t}^i \in \mathbb{R}^{|\mathcal{O}|}$ is the ground truth operation for the j-th slot.

Decoder We employ GRU as decoder to generate the value of dialogue state for each domain-slot pair whose operation is UPDATE. GRU is initialized with $g_t^{i,0} = W_t$ and $e_t^{i,0} = h_t^{[\mathrm{SLOT}]^i}$. The probability distribution of the vocabulary is calculated as:

$$P_{dec,t}^{i,k} = \mathrm{softmax}\left(\mathrm{GRU}\left(g_t^{i,k-1}, e_t^{i,k}\right) \times E\right) \in \mathbb{R}^{|V|}, \tag{5}$$

where k is decoding step, $E \in \mathbb{R}^{|V| \times d}$ is the word embedding space shared with the encoder, and $|V|$ is the size of multilingual vocabulary. The overall loss for generating slot value is the average of the negative log-likelihood loss:

$$\mathcal{L}_{dec,t} = -\frac{1}{|\mathbb{U}_t|} \sum_{i \in \mathbb{U}_t} \left[\frac{1}{K_t^i} \sum_{k=1}^{K_t^i} \left(Y^{i,k}\right)^\top \log\left(P_{dec,t}^{i,k}\right)\right], \tag{6}$$

where $|\mathbb{U}_t|$ is the number of slots which require value generation, K_t^i indicates the number of ground truth value to be generated for the i-th slot. $Y^{i,k} \in \mathbb{R}^{|V|}$ represents the one-hot vector of the ground truth token generated for the i-th slot at the k-th decoding step.

Fine-grained Contrastive Learning In order to better capture the common features between the source language and the target language, our model utilizes fine-grained CL to pull closer the representation of similar sentences across different languages. The key to CL is to find high-quality positive and negative pairs corresponding to the original utterance. The positive sample should be semantically consistent with the original utterance and provides cross-lingual view as well. In our scenario, we choose code-switched input $X_{t,cs}$ as the positive sample of X_t, while other inputs in the same batch are treated as negative samples.

As state operation of each slot is a token-level task, we utilize a fine-grained CL loss to facilitate token alignment. To achieve fine-grained cross-lingual transfer, our method selects the output representation $h_t^{[\mathrm{SLOT}]^i}$ of the special token $[\mathrm{SLOT}]^i$ for contrastive learning, as these I tokens are able to convey the semantics of the slots in the query. The i-th slot token loss is defined as:

$$\mathcal{L}_{cl,t}^i = -\frac{1}{I} \sum_{j=1}^{I} \log \frac{cos\left(h_t^i, h_t^{j^+}\right)}{cos\left(h_t^i, h_t^{j^+}\right) + \sum_{k=0,k\neq j}^{\mathrm{I}-1} cos\left(h_t^i, h_t^{k^-}\right)}, \tag{7}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

where h_t^i is the abbreviation of $h_t^{[\text{SLOT}]^i}$, $h_t^{j^+}$ and $h_t^{k^-}$ are positive and negative samples of $h_t^{[\text{SLOT}]^i}$ respectively. The total loss $\mathcal{L}_{cl,t}$ is calculated by adding up all tokens CL loss.

The overall objective in CLCL-DST at dialogue turn t is the sum of individual losses above:

$$\mathcal{L}_t = \mathcal{L}_{enc,t} + \mathcal{L}_{cl,t} + \mathcal{L}_{dec,t}. \tag{8}$$

3.2 Significance-based Code-switching

The importance of different words in a dialogue utterance varies. For example, in the price range domain, "cheap" and "expensive" are more likely to be keywords, while in the area domain, keyword set might include orientation terms such as "center", "north" and "east". Assuming that a dataset contains v words constituting a vocabulary $\mathcal{V}$, we construct a subset of keywords $\mathcal{K} \subseteq \mathcal{V}$ for code-switching. Subsequently, the encoder of CLCL-DST serves to extract keywords in the training data.

Given the input token $X_t = (w_t^1, w_t^2, ..., w_t^n)$ at the t-th turn, n denotes the number of words. We feed X_t into encoder, and obtain the representation $h_t^{[\text{CLS}]} \in \mathbb{R}^d$ of the special token [CLS]. Then the sentence embedding vector W_t is calculated as:

$$W_t = \tanh(W_{pool} h_t^{[\text{CLS}]} + b), \tag{9}$$

where W_{pool} and b are learnable parameters. Then the cosine similarity between each token $w_t \in X_t$ and the sentence embedding vector W_t is computed as:

$$Sim(w_t) = \cos(w_t, W_t). \tag{10}$$

$Sim(w_t)$ reflects the degree of associations between w_t and sentence embedding W_t. A higher value of the significance score $Sim(w_t)$ indicates a higher probability of w_t to be a keyword. For words that are tokenized into subwords, we average the significance scores of each subword to obtain the word score.

Equation 10 calculates the significance score of words in a sentence. To get the keyword set $\mathcal{K}$ in training set, we add all significance scores for token w in training set and multiply them by the inverse document frequency (IDF) (Yuan et al., 2020) of w:

$$S(w) = \log \frac{N}{|\{x \in X : w \in x\}|} \cdot \sum_{x \in X : w \in x} Sim(w), \tag{11}$$

where N denotes the number of the input in the training dataset, $|\{x \in X : w \in x\}|$ indicates the number of the input containing w. The IDF term can reduce the weight of words which appear frequently in the dataset, assigning meaningless words (e.g., "for" and "an") with a lower score.

We select top-k words according to the significance scores to get a keyword set K, and use the bilingual dictionary MUSE (Lample et al.,) to construct the code-switched dictionary $Dic = ((s_1, t_1), ...(s_k, t_k))$, where s and t refer to the source and target language words respectively. k is the number of keywords. In addition, we translate the whole words in ontology and add them to Dic due to their important role in the sentence.

Inspired by (Qin et al., 2021), we randomly replace some words in source language sentence with corresponding target words with a fixed probability if they appear in Dic. Since words from the source language may have multiple translations in Dic, we randomly select one of them for substitution. Notably, the input token X in our model includes dialogue utterance D and dialogue states B, we just replace source words in D as B shares the same slots across languages. Finally, we can get the code-switched input tokens $X_{t,cs}$ from X_t as:

$$X_{t,cs} = [\text{CLS}] \oplus D_{t-1,cs} \oplus D_{t,cs} \oplus B_{t-1}, \tag{12}$$

4 Experiments

4.1 Datasets

We evaluate our model on two datasets as follows:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

- **Multilingual WoZ 2.0 dataset** (Mrkšić et al., 2017): A restaurant domain dialogue dataset expanded from WoZ 2.0 (Wen et al., 2017), which contains three languages (English, German, Italian) and 1200 dialogues for each language. The corpus consists of three goal-tracking slot types: food, price range and area. The task is to learn a dialogue state tracker only in English and evaluate it on the German and Italian datasets, respectively.
- **Parallel MultiWoZ dataset** (Gunasekara, 2021): A seven domains dialogue dataset expanded from MultiWoZ 2.1 (Eric et al., 2020). Parallel MultiWoZ contains two languages (English, Chinese) and 10K dialogues. The Chinese corpus is obtained through Google Translate and manually corrected by experts.

4.2 Compared Methods

We compare our approach with the following methods:

- **XL-NBT** (Chen et al., 2018) utilizes bilingual corpus and bilingual dictionaries to transfer the teacher's knowledge of the source language to a student tracker in the target languages.
- **MLT** (Liu et al., 2020a) constructs code-switched data through the attention layer for training.
- **CLCSA** (Qin et al., 2021) dynamically constructs multilingual code-switched data by randomly replacing words, so as to better fine-tune mBERT and achieve outstanding results in multiple languages.
- **SUMBT** (Lee et al., 2019) uses a non-parametric distance measure to score each candidate slot-value pair. We replace BERT with mBERT on the cross-lingual setup.
- **SOM-DST** (Kim et al., 2020) employs BERT as the encoder and uses a copy-based RNN to decode upon BERT outputs.
- **DST-as-PROMPTING** (Lee et al., 2021) introduces an approach that uses schema-driven prompting to provide history encoding and then utilizes T5 to generate slot values directly. Here, we use the multilingual version of T5 - mT5 (Xue et al., 2021).
- **XLIFT-DST** (Moghe et al., 2021) leverages task-related parallel data to enhance transfer learning by intermediate fine-tuning of pre-trained multilingual models. For parallel MultiWoZ, XLIFT-DST uses the architecture of SUMBT, while uses the state tracker in CLCSA for Multilingual WoZ 2.0.

4.3 Implementation Details

Our method leverages the pre-trained mBERT-base[0] implemented by HuggingFace as the encoder, with 12 Transformer blocks and 12 self-attention heads. One layer GRU is used as the decoder. The encoder shares the same hidden size s with the decoder, which is 768. Adam optimizer (Kingma and Ba, 2014) is applied to optimize all parameters with a warmup strategy for the 10% of the total training steps. The peak learning rate is set to 4e-5 for encoder and 1e-4 for decoder, respectively. Besides, we use greedy decoding for generating slot values.

For Multilingual WoZ dataset, the batch size is set to 64 and the maximum sequence length to 200. For parallel MultiWoZ dataset, the batch size and the maximum sequence length are 16 and 350 respectively. We replace the word for each dialogue with a fixed probability of 0.6. The training is performed for 100 epochs as default, and we choose the best checkpoint on the validation set to test our model.

4.4 Evaluation Metrics

The metrics in dialogue state tracking are turn-level which include Slot Accuracy, Joint Goal Accuracy and Slot F1. Slot Accuracy is the proportion of the correct slots predicted in all utterances. Joint Goal Accuracy is the proportion of dialogue turns where all slot values predicted at a turn exactly match the ground truth values, while Slot F1 is the Macro-average of F1 score computed over the slot values at each turn.

[0]https://huggingface.co/bert-base-multilingual-uncased

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

Model	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
XL-NBT (Chen et al., 2018)	55.0	30.8	72.0	41.2
MLT (Liu et al., 2020a)	69.5	32.2	69.5	31.4
Transformer based				
mBERT	57.6	15.0	54.6	12.6
CLCSA (Qin et al., 2021)	83.0	63.2	82.2	61.3
XLIFT-DST (Moghe et al., 2021)	85.2	**65.8**	84.3	66.9
CLCL-DST (ours)	**89.3**	63.2	**89.1**	**67.0**

Table 1: Slot accuracy and joint goal accuracy on Multilingual WoZ 2.0 dataset under zero-shot setting when trained with English task data. Please see text for more details. **Bold** indicates the best score in that column. CLCL-DST denotes our approach.

Model	joint acc.	slot f1.
SUMBT (Lee et al., 2019) †	1.9	14.8
SOM-DST (Kim et al., 2020) ‡	1.7	10.6
DST-as-PROMPTING (Lee et al., 2021) ‡	2.5	17.6
XLIFT-DST †	5.1	40.7
CLCL-DST (ours)	**27.1**	**79.4**
In-language training †	15.8	70.2
Translate-Train †	11.1	54.2
Translate-Test †	26.5	77.0

Table 2: Joint goal accuracy and slot F1 on parallel MultiWoZ dataset under zero-shot learning setting when trained with English task data and tested on Zh language. '†' denotes results from (Moghe et al., 2021). '‡' denotes our re-implemented results for the models based on corresponding multilingual pretrained models.

4.5 Main Results

Results for the Multilingual WoZ dataset are illustrated in Table 1. We can see that CLCL-DST outperforms the state-of-the-art model (XLIFT-DST) by 4.1% and 4.8% in slot accuracy for De and It respectively. This demonstrates that our model is able to explicitly bring similar representations of different languages closer together through contrastive learning than augmenting transfer learning process with intermediate fine-tuning of pre-trained multilingual models.

To further study the effectiveness of our model under the zero-shot setting, We also test CLCL-DST on parallel MultiWoZ in Table 2. As there are only a few baselines available for this dataset, we re-implement some monolingual models such as SUMBT, SOM-DST, DST-as-PROMPTING into multilingual scenarios. We find that our model has 22% and 38.7% improvement over XLIFT-DST in joint goal accuracy and slot f1 for target language Zh under the zero-shot setting. It is worth noting that the joint goal accuracy of all these baseline models is relatively low. The possible reason is that these models do not learn considerable cross-lingual representations in the multi-domain cases, making it difficult to migrate for complex slots. Specifically, In the SOM-DST model, its decoder utilizes the soft-gated copy mechanism (See et al., 2017) in addition to GRU, which introduces additional noise from the source language and is not applicable to multilingual settings. In DST-as-PROMPTING, the model only leverages mT5 to generate slot values directly without learning deeply cross-lingual interaction information. Besides, we also refer to the results of translation-based methods from (Moghe et al., 2021) in Table 2. Our model still outperforms all of them. These results further indicate that our proposed CLCL-DST leveraging code-switched data with contrastive learning boosts the performance of dialogue state tracker.

5 Ablation Studies

We conduct ablation experiments to explore the effect of fine-grained contrastive learning and the significance-based keyword extraction approach on the overall performance for the Multilingual WoZ

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

2.0 dataset.

5.1 The Effect of Fine-grained Contrastive Learning

In addition to fine-grained CL, we also introduce coarse-grained CL for aligning similar sentences across different languages. To be specific, we align the sentence embedding W_t from equation 9 with its corresponding code-switched positive representations W_t^+. The objective for coarse-grained CL is written as follows:

$$\mathcal{L}_{sl,t} = -\log \frac{\cos\left(W_t, W_t^+\right)}{\cos\left(W_t, W_t^+\right) + \sum_{k=0,k\neq j}^{\mathrm{I}-1} \cos\left(W_t, W_t^{k-}\right)}, \tag{13}$$

where $W_t^{k^-}$ is the negative sample for W_t at the t-th turn.

Method	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
w/o CL	82.5	52.0	86.8	60.0
Coarse-grained CL	87.9	57.7	79.8	41.0
Fine-grained CL	**89.3**	**63.2**	**89.1**	**67.0**

Table 3: Slot accuracy and joint goal accuracy for different grained contrastive learning under zero-shot setting. "CL" denotes the abbreviation of contrastive learning.

As results shown in Table 3, we can conclude that different granularities of contrastive learning are effective for our model, especially fine-grained CL since it can bring more improvement to CLCL-DST. Using fine-grained CL improves 1.4% and 5.5% in slot accuracy and joint goal accuracy for De, and 9.3% and 26% for It, respectively, compared to coarse-grained CL. Since the goal of dialogue state tracking is to predict the state of slots in each turn of the dialogue, it can be considered as a token-level task, so fine-grained CL is better suited for this task compared to coarse-grained CL. Also, our approach selects specific tokens representing slots instead of all tokens in the dialogue for contrastive learning, which can reduce the noise caused by other semantically irrelevant tokens.

5.2 The Effect of significance-based code-switching

In this section we further explore the impact of keyword extraction algorithm on CLCL-DST. Table 4 shows the performance of different keyword extraction strategies. We try other four approaches to obtain the mapping dictionaries and compare them with the significance-based code-switching approach: (1) choosing words based on their frequency in our training set and converting them to target languages by MUSE; (2) using the whole ontology, which contains 90 words approximately; (3) combining the dictionaries obtained from (1) and (2) to form a new dictionary; (4) extracting keywords using only TF-IDF algorithm.

Method	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
MUSE	86.4	59.4	84.0	54.5
Onto	86.2	56.0	81.8	46.8
MUSE+Onto	88.0	57.8	88.4	66.3
TF-IDF+Onto	86.5	55.3	87.9	66.0
Significance-based	87.9	60.4	**89.1**	63.5
Significance-based+Onto	**89.3**	**63.2**	**89.1**	**67.0**

Table 4: Slot accuracy and joint goal accuracy on Multilingual WoZ 2.0 dataset for different keywords extraction approaches under zero-shot setting. The Method column represents the strategy for extracting keywords. "Onto" is the abbreviation of ontology. "+" denotes the merging of dictionaries obtained by the two methods.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

Number of keywords	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
200	86.5	60.4	85.1	61.5
500	88.2	62.3	86.8	64.4
1000	**89.3**	63.2	**89.1**	**67.0**
2000	88.6	**63.3**	86.9	66.3
5000	88.9	62.9	87.4	66.5

Table 5: Slot accuracy and joint goal accuracy on Multilingual WoZ 2.0 dataset for different number of keywords under zero-shot setting.

Compared with only considering the frequency of words in the corpus, our significance-based code-switching approach can also make use of the numerous linguistic information carried in the multilingual pretrained model, so that the selected words are more representative of the utterance. This approach enables the selected words to better express the main idea of the text. At the same time, words in ontology such as place names, food names, etc. are originally special words in the dataset, which occupy an important position in the text. Adding these words to our dictionary can further improve the performance of the model.

Table 5 shows the influence of different number of keywords on our model. We can see that the model has the best or second-best performance when k is 1000. As k continues to increase, the additional keywords are less indicative, so they even have a negative impact on model performance.

6 Conclusion

In this paper, we propose a novel zero-shot adaptation method CLCL-DST for cross-lingual dialogue state tracking. Our approach leverages fine-grained contrastive learning to explicitly align representations across languages. Besides, we introduce the significance-based code-switching approach to replace task-relevant words with target language for generating code-switched sentences on downstream tasks. Our method obtains new state-of-the-art results on Multilingual WoZ dataset and parallel MultiWoZ dataset, which demonstrates its effectiveness. In the future, we would investigate better training objectives for cross-lingual DST task, especially on multi-domain area, to further boost the dialogue system on multi-lingual scenarios. We would also explore better positive and negative samples when applying contrastive learning on DST task.

Acknowledgement

The research work descried in this paper has been supported by the National Key RD Program of China (2020AAA0108005), the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130) and Toshiba (China) Co., Ltd. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics (ACL).

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. *Proc. Interspeech 2019*, pages 1458–1462.

Chulaka Gunasekara. 2021. Overview of the ninth dialog system technology challenge: Dstc9. In *DSTC9 Workshop at AAAI 2021*.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Siyu Lai, Hui Huang, Dong Jing, Yufeng Chen, Jinan Xu, and Jian Liu. 2021. Saliency-based multi-view mixed language training for zero-shot cross-lingual classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 599–610.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. Sumbt: Slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949.

Yen-Ting Lin and Yun-Nung Chen. 2021. An empirical study of cross-lingual transferability in generative dialogue state tracker. *arXiv preprint arXiv:2101.11360*.

Weizhe Lin, Bo-Hsiang Tseng, and Bill Byrne. 2021. Knowledge-aware graph-enhanced gpt-2 for dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7871–7881.

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

Zihan Liu, Genta Indra Winata, Peng Xu, Zhaojiang Lin, and Pascale Fung. 2020b. Cross-lingual spoken language understanding with regularized representation alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7241–7251.

Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual intermediate fine-tuning improves dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150.

Nikola Mrkšić, Ivan Vulić, Diarmuid O Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

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

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. Gl-clef: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686.

Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and accurate dialogue state tracking via hierarchical sequence generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1876–1885.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. Cline: Contrastive learning with semantic negative examples for natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2332–2342.

Yifan Wang, Jing Zhao, Junwei Bao, Chaoqun Duan, Youzheng Wu, and Xiaodong He. 2022. Luna: Learning slot-turn alignment for dialogue state tracking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3319–3328.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996.

Yan Zeng and Jian-Yun Nie. 2020a. Jointly optimizing state operation prediction and value generation for dialogue state tracking. *arXiv preprint arXiv:2010.14061*.

Yan Zeng and Jian-Yun Nie. 2020b. Multi-domain dialogue state tracking based on state graph. *arXiv preprint arXiv:2010.11137*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

Unsupervised Style Transfer in News Headlines via Discrete Style Space

Qianhui Liu, Yang Gao*, Yizhe Yang
School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China
{3120201048,gyang,yizheyang}@bit.edu.cn

Abstract

The goal of headline style transfer in this paper is to make a headline more attractive while maintaining its meaning. The absence of parallel training data is one of the main problems in this field. In this work, we design a **d**iscrete style space for unsupervised **h**eadline **s**tyle **t**ransfer, short for **D-HST**. This model decomposes the style-dependent text generation into content-feature extraction and style modelling. Then, generation decoder receives input from content, style, and their mixing components. In particular, it is considered that textual style signal is more abstract than the text itself. Therefore, we propose to model the style representation space as a discrete space, and each discrete point corresponds to a particular category of the styles that can be elicited by syntactic structure. Finally, we provide a new style-transfer dataset, named as **TechST**, which focuses on transferring news headline into those that are more eye-catching in technical social media. In the experiments, we develop two automatic evaluation metrics — style transfer rate (STR) and style-content trade-off (SCT) — along with a few traditional criteria to assess the overall effectiveness of the style transfer. In addition, the human evaluation is thoroughly conducted in terms of assessing the generation quality and creatively mimicking a scenario in which a user clicks on appealing headlines to determine the click-through rate. Our results indicate the D-HST achieves state-of-the-art results in these comprehensive evaluations.

1 Introduction

A style makes sense under pragmatic use and becomes a protocol to regularize the manner of communication [Jin et al.2022, Khalid and Srinivasan2020]. So, the task of text style transfer is to paraphrase the source text in a desired style-relevant application [Toshevska and Gievska2021]. In practical use, the style is data-driven and task-oriented in different area [Jin et al.2022].

The absence of parallel training data for a certain style is one of the difficult problems. Continuous latent space mapping is a typical method for unsupervised style transfer to address the issue. Guo et al. (2021; Liu et al. (2020) model the latent space to a Gaussian distribution. Points in latent space are moved to the target representation with some style guidance. Nangi et al. (2021; John et al. (2018; Romanov et al. (2018) disentangle the continuous latent representation purely according to its content, and replace the source attribute to the target one. However, there are two problems of the continuous space approach. Firstly, the style is highly abstract so that it is unstable and too sparse to accurately represent the style in the continuous space. Second, the continuous vector-based representation is difficult to manipulate and cannot be examined at a finer level. To control the style transfer and enhance its explainability, several kinds of discrete signals are used to represent the style. For instance, Reid and Zhong (2021; Tran et al. (2020; Li et al. (2018) employ Mask-Retrieve-Generate strategy to decompose style attributes by word-level editing actions. But, these methods express styles in a highly discrete way which fail to capture the relationships between words or sentences.

To more effectively describe the style in a highly abstract and discrete manner while also capturing the semantic relations in the texts, we propose a latent and **d**iscrete style space for **h**eadline **s**tyle **t**ransfer,

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

abbreviated as **D-HST**. This model decomposes style-dependent text generation into content-feature extraction and style modeling. Therefore, we design a dual-encoder and a shared single-decoder framework to accomplish the overall generation. Due to the lack of parallel training data, we have to synthesize adequate training pairs to accommodate the content extraction and the style modeling. Given a target stylistic headline, we first automatically generate a content-similar input as well as style-consistent input for feeding the dual encoders. As the textual style signal is expected to be rather abstract and limited compared to the text itself, we propose to model the style representation space as a discrete space, with each discrete point denoting a particular category of the styles that can be elicited by syntactic structure.

Also, we provide a new style-transfer dataset derived from the real scenarios, named as **TechST**, which transfers news headlines into the ones that are more attractive to readers. Although several datasets are currently available for this purpose [Jin et al. (2020], but the appealing styles—such as humor and romance—are taken from fictional works of literature, which we believe makes them unsuitable for usage as an *attractive* style for headlines. In the experiments, we design two automatic evaluation metrics, including style transfer rate (STR) and style-content trade-off (SCT) - along with a few traditional criteria to assess the overall of the style transfer. Additionally, the quality of the generation is thoroughly evaluated, and the click-through rate is calculated by creatively simulating a scenario in which a user clicks on attractive headlines. Our findings show that the D-HST performs at the cutting edge in these thorough assessments. In conclusion, our article mainly has the following contributions.

- We propose an unsupervised style transfer method with discrete style space, which is capable of disentangling content and style.
- We propose new metrics in automatic evaluation and human evaluation, and achieves state-of-the-art results in these comprehensive evaluations.
- We provide a novel dataset derived from actual events to convert news headlines into catchy social media headlines.

2 Related Work

Attractive Headline Generation It is crucial to generate eye-catching headlines for an article. Gan et al. (2017) proposes to generate attractive captions for images and videos with different styles. Jin et al. (2020) introduces a parameter sharing scheme to generate eye-catchy headlines with three different styles, humorous, romantic, and clickbait. Li et al. (2021) proposes a disentanglement-based model to generate attractive headlines for Chinese news. We build upon this task by rewriting source headlines to attractive ones.

Text Style Transfer There are mainly three kinds of methods used in TST task. 1) **Modeling in the Latent space** Mueller et al. (2017; Liu et al. (2020) use continuous space revision to search for target space. Shen et al. (2017; Sun and Zhu () learn a mapper function in source and target space. John et al. (2018; Romanov et al. (2018; Hu et al. (2017) explicitly disentangle content and style in latent space. However, the style is highly abstract so that it is unstable and too sparse to accurately represent the style in the continuous space. 2) **ProtoType Editing** It is a word replacement method. Li et al. (2018; Tran et al. (2020) propose three-stage methods to replace stylist words with retrieved words in the target corpus. Reid and Zhong (2021) uses Levenshtein editing to search target stylist words. These methods work well on Content-Preferences dataset, like sentiment, debias. 3) **Control Code Index** Keskar et al. (2019; Dai et al. (2019) use a control sign embedding to controls the attribute of generated text. Yi et al. (2021) controls style using a style encoder. These methods don't learn style in a fine-grained way and the style space is a block-box. We combine the first and third methods, using a control code to control style and modeling a style space with appropriate distribution.

To model the discrete style in an unsupervised fashion, we propose to inherit the third and fourth methods. Specifically, we construct pseudo data to enrich the content-based parallel data and style-based para. Further, different from the previously styled latent space, we model it as a discrete one based on the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

claim that style is highly abstract and more sparse compared to content. We will describe this in detail in the next section.

3 Methodology

We are given samples $Y = \{y_1, y_2, \dots y_m\}$ from the style dataset S. The objective of our task is to transfer a headline sentence to a new headline equipped with the style of the target data S, while maintaining its originally semantic content.

3.1 Model Overview

Our proposed **D-HST** model consists of a duel-encoder and a single shared decoder in an unsupervised setting. It begins by constructing a pseudo-parallel dataset which comprises of two pairs of inputs-and-outputs. One of the inputs is X^Y_{cont}, which is generated by using a pre-trained paraphrasing model and has input that is content-similar to output Y. The other input is X^Y_{style}, which is collected in style dataset S and uses inputs of sentences with the same style as output Y based on the defined style (Section 3.2).

The model structure is described in Section 3.3. One of the inputs is **content input** X^Y_{cont} encoded by a content encoder, then fed into a content pooling to extract its sentence-level feature, denoted as $Z^Y_{cont} = pool_{cont}(enc_{cont}(X^Y_{cont}))$. Similarly, the other input is **style input** X^Y_{style} encoded by a style encoder, then fed into a style pooling to get style representation Z^Y_{style}. The hypothesis is that the pooling serves as a bottleneck which can disentangle the representation of content and style with help of proper loss function (Section 3.4). The overall model architecture is shown in Figure 1.

3.2 Pseudo Parallel Data Construction

Content Input Prior work has demonstrated that paraphrasing techniques can translate source sentences into standard written sentences while maintaining their substance [Mitamura and Nyberg (2001]. In our approach, we assume that a special style (such as attractiveness, informality in the experiment) of a sentence can be removed after paraphrasing. We use a pretrained paraphrasing model[0] to remove stylist attribute, and construct the content inputs X^Y_{cont}.

As the paraphrasing model often produces multiple outcomes, in the experiment, we select top 5 generations as a candidate set for the content input. Then, we calculate bertscore to estimate the similarity between the generated candidates and the output Y. Only candidates with similarity between 0.75 to 0.95 are kept to preserve as much content information as possible and prevent significantly overlapping generation.

Style Input We suppose that a certain syntactic structure can reflect the style. For example, attractive headlines often employ interrogative questions; informal conversations frequently use ellipsis; and impolite language often employ imperative sentences. To collect more parallel headlines to train the style-based modules, we construct the style input X^Y_{style} that shares the same syntactic structure yet different content with target Y, from the data in style dataset S. In order to filter out the content information in the style input, we use a set of sentences C^Y_{style} that share the same syntactic structure for X^Y_{style}, then average these sentences with a learnable parameter.

Specifically, we use a chunk parser FlairNLP[1] to get the syntactic structure of these headlines. We first get the chunk label for each word using the chunk parser. Then, we merge the spans having the same label. Based on the assumption that words such as "who", "whether" and "how" are function words that guide special sentence patterns, we set a separate label QP to mark the leading words of interrogative sentences. We get some distinct syntactic structures, each of which has some corresponding headlines. We assume that if one syntactic structure occurs in less than 10 headlines, it is not representative. Then, we filter the syntactic structure and its corresponding sentences if its syntactic structure occurs in less than 10 headlines. Table 1 shows examples of processed syntactic structures and their corresponding sentences.

[0]https://huggingface.co/tuner007/pegasus_paraphrase
[1]https://github.com/flairNLP/flair

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

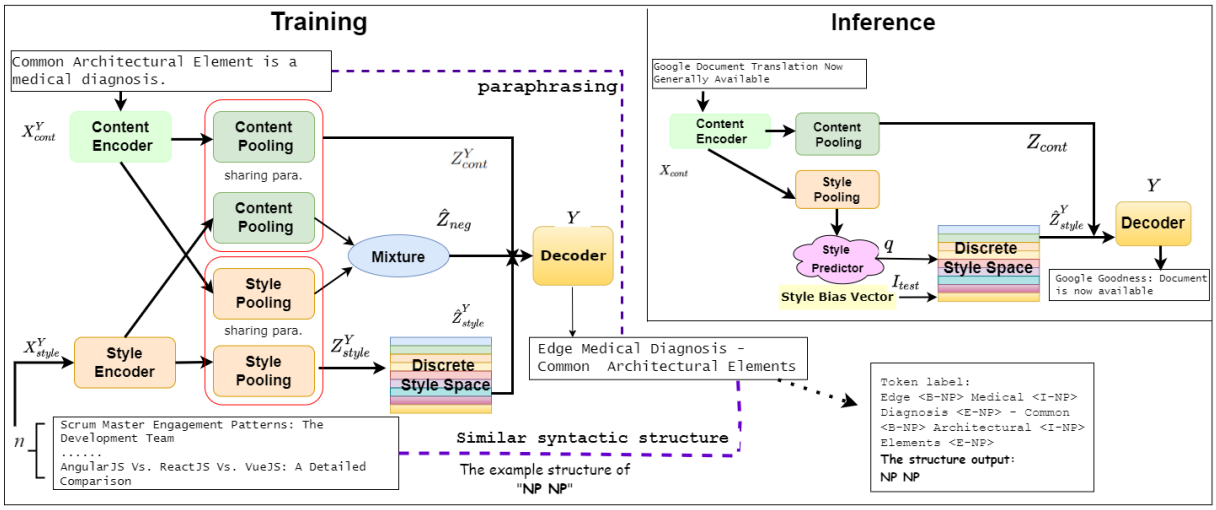


Figure 1: The framework of the D-HST model. The training phase and inference phase are depicted in the figure.

Syntactic Structure	QP VP NP	CD NP VP
Sentences	How to Become a DevOps Engineer How to Scale Your SaaS Business How to Do API Testing	3 Tech Debt Metrics Every Engineer Should Know 7 Top Kubernetes Health Metrics You Must Monitor 10 Software Testing Interview Questions You Haven't Heard Before

Table 1: Examples of syntactic structures and their corresponding headline sentences. These examples indicate that some sentences can express the same style representation by syntactic structure.

3.3 Model Architecture

The duel-encoder and the shared decoder are both based on standard Transformer model [Vaswani et al. (2017]. The content inputs and style inputs are both encoded by their separate encoders, that are content encoder and style encoder, respectively. Each token is fed to the encoder and obtains embeddings $\{e_1, e_2, ..., e_{|X|}\} = enc(X)$, where $|X|$ is the length of the input sentence, $e_t \in R^H$, H is the dimension of transformer.

Feature Extractor To facilitate the disentanglement between the content semantics and the stylistic attributes, we elicit their distinct features by pooling the multi-dimensional representation in accordance with the method used in Liu and Lapata (2019). Specifically, a multi-head pooling is adopted to extract features. We employed attention a_t, where t represents a token, to calculate its importance score for the whole sentence. The equation is :

$$\alpha_t = \frac{\exp\ a_t}{\sum_{t \in |X|} \exp\ a_i} \tag{1}$$

$$a_t = k_t e_t \tag{2}$$

where $k_t \in R^H$ is a learnable parameter. The value of each token V_t is also computed using a linear projection of e_t. Finally, we take a weighted average to get the pooling output Z.

$$Z = \sum_{t \in |X|} \alpha_t V_t \tag{3}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

Discrete Style Space Inspired by Hosking and Lapata (2021) who claim style is limited and sparse, we therefore propose to extract a specific style from a discrete style space. The space maintains a discrete table $C \in R^{K \times D}$, K is the number of style categories[2], equal to the number of distinct syntactic structure in style dataset S. We use q to represent the category distribution and $\tilde{q} \in [0, K]$ to represent the sampled category. The category distribution q is mapped from the style pooling Z^Y_{style}, and it can be formulated as $p(q|Z^Y_{style})$.

Finally, we draw $\tilde{q}$ from the Gumbel-Softmax distribution of q. The equation can be written as:

$$\tilde{q} \sim \text{Gumbel-Softmax}(q) \tag{4}$$

The style representation, $\hat{Z}^y_{style} = C(\tilde{q})$, maps from the discrete code $\tilde{q}$. $\hat{Z}^Y_{style}$ ought to be as near as the input Z^Y_{style}. So we get a loss term:

$$\mathcal{L}_q = \| Z^Y_{style} - sg(C(\tilde{q})) \|_2 \tag{5}$$

Because the gradient could be broken at stop gradient sg, the loss is not derivable. We employ a reparameterization trick [Kingma and Welling (2013] to update parameters and exponential moving average [Roy et al. (2018] strategy to update the discrete table.

Style Bias We assume that each sentence has its own style score. For example "You Can't Reset Your Fingerprint" is more obviously attractive than "AI-Assisted Coding with Tabnine" in terms of its expressing style, although they are both in style dataset. Therefore, we manually rank each sentence in the style dataset S based on external knowledge I_{test}. Details of the external knowledge are shown in Section 5.

We believe that syntactic structure can be used to define the style category and that sentences with the same structure may score similarly in terms of style. Each style is expected to be encoded into a specific category, and categories with higher style scores are more likely to be selected in inference. $I \in R^K$ is a one-hot vector and serves as a pre-labeled supervisory signal, representing the correspondence between styles and categories. For example, $I_m \in R^K$ encodes the style category to which the sentence m belongs. In training phase, we expect each style is encoded into a specific category, so we let the output of the category distribution q fit supervisory signal I. The equation can be written as:

$$\mathcal{L}_r = \| I - \text{softmax}(q) \|_2 \tag{6}$$

In inference phase, for all sentences, we use the fixed discrete style bias distribution $I_{test} \in R^K$ to increase the probability of choosing a high-scoring style. And we set the probability for each category in I_{test} to be the normalized style score.

Mixture Module We also design a mixture module to serve as negative knowledge to guide decoder to leave away from the content of X^Y_{style} and the style of X^Y_{cont}. We use a small full connect network with the concatenation of $Z_{cs} = pool_{cont}(enc_{style}(X^Y_{style}))$ and $Z_{sc} = pool_{style}(enc_{cont}(X^Y_{cont}))$ as input, written as $Z_{neg} = MLP(Z_{cs}, Z_{sc})$

Finally, the overall hidden representation Z can be written as $Z = Z^Y_{cont} + \hat{Z}^Y_{style} + Z_{neg}$. And the target distribution $p(Y|Z) = dec(Z)$.

3.4 Model Training

We first describe the training process that makes the model to capture its local independence information separately.

We set triples $((X^Y_{cont}, X^Y_{style}), Y)$ as input and output, respectively. To produce strong style signals, we use a set of style sentences C^Y_{style} in the same style as Y. The selection strategy has been described in Section 3.2 and the style representation is weighted with a learnable parameter κ, such as $Z^Y_{style} = \sum_{c_i \in C^Y_{style}} pool_{style}(enc_{style}(c_i))\kappa_{ci}$. Then, $\hat{Z}^Y_{style}$ is sampled from the style space. It is trained

[2] K=324 in our TechST dataset

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

to generate target Y with the overall hidden representation Z, which is the sum of content encoding Z^Y_{cont}, style encoding $\hat{Z}^Y_{style}$ and negative knowledge encoding Z_{neg}. The factorised reconstruction loss term can be written as:

$$\mathcal{L}_Y = \sum_t \log p(w_t|w_1, w_2...w_{t-1}; Z) \tag{7}$$

The final objective function is:

$$\mathcal{L} = \mathcal{L}_Y + \delta\mathcal{L}_q + \epsilon\mathcal{L}_r \tag{8}$$

3.5 Inference

Since we don't have any style input X^Y_{style} for inference, only X_{cont} in source dataset is available and transferred to the defined target style. As such, the well-trained style encoder and mixture module can not be directly adopted in the inference. To fill this gap, we further train a style predictor module to alternatively select a sample to represent the most stylistic category for the following decoder. This predictor is formulated as $p(q|X^Y_{cont}) = MLP(pool_{style}(enc_{cont}(X^Y_{cont})))$. The additional predictor is trained to predict the well-trained style category distribution q through X^Y_{cont}. q is mapped from Z^Y_{style} and represents as $p(q|Z^Y_{style})$. So we distill the distribution $p(q|X^Y_{cont})$ to the well-trained distribution $p(q|Z^Y_{style})$. The loss term is:

$$\mathcal{L}_{KL} = -KL(sm(p(q|X^Y_{cont}))||sm(p(q|Z^Y_{style}))) \tag{9}$$

where sm is short for softmax function. In inference phase, we sample $\tilde{q} \sim ((1-\gamma)\text{sm}(q) + \gamma I_{test})$ 3 times and generate 3 candidate outputs. Finally, we select the one with highest content preservation with the input, calculated by bertscore.

4 Tasks and Datasets

For the headline style transfer task, we focus on attractive news headline transfer on technology topics. Technology news headlines are always formal. For example, "Google Document Translation Now Generally Available." is a common style for an event headline. On the contrary, technology blog headlines in social media tend to be special and catch readers' eyes. In this paper, we define this kind of headline as "**Attractive**" style. To highlight the characteristics of style, the previous example can be transferred as "Google Goodness: Document Now Available". The goal of this task is to transfer the formal news headlines to more attractive blog headlines in technology domain.

Datasets Our attractive technology dataset **TechST** was crawled from Dzone[3], including stylistic technology blog headlines and users' pageviews. This data was used to train the style transfer model. We also crawled technology news headlines from InfoQ[4] as non-stylistic headlines for testing. The task is to transfer the headlines in InfoQ to a new style that is modelled with the Dzone dataset. Both of them were crawled from the beginning to November 2011. We filtered out the blog headlines with pageviews less than 500 and the ones more than 22 words as we believe shorter headlines are attractive. Finally, we get 60,000 samples for training and 2,000 samples for testing.

We also use a cornerstone dataset Grammarly's Yahoo Answers Formality Corpus (GYAFC) [Rao and Tetreault (2018] for formality transfer. It contains 53,000 paired formal and informal sentences in two domains. To meet our requirement of unsupervised style transfer setting, the task is to transfer the formal sentences to informal ones. Only informal sentences in the Family and Relationships categories were used for training and validation.

5 Experiments and Results

External Knowledge As mentioned in Section 3.3, external knowledge is used to estimate the style strength. To some extents, users' pageviews reflect attractiveness of the style. We first parsed all the

[3]https://dzone.com/

[4]https://www.infoq.com/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

syntactic structures of sentences in the style dataset. Then, we calculate average-pageviews for each syntactic structure. The more average-pageviews the structure receives, the higher style score it has. We acknowledge that style isn't the only factor that affects pageviews, content also contributes to it. For example, headlines with syntactic structure like "NP VP" are common, but some headlines with such structure may have high pageviews. To eliminate the impact of content, we add a pageview variance term. Specifically, if sentences with same syntactic structure show little pageview variance, it is speculated that pageviews are determined by the syntactic structure. On the contrary, if the variance is significant, it suggests that other elements, such as content, are influencing pageviews. As such, the style score must be penalized. Finally, we define our style score as:

$$I^i_{test} = \frac{mean(a)^\omega}{var(a)^\nu} \tag{10}$$

I^i_{test} represent the style score of category i, a is the collection of the sentences having style i. ω and ν are hyperparameters.

For GYAFC dataset, no such corresponding information is provided, so we set all syntactic structures the same style score.

Experiment Setup We use 6-layers transformers to train our model. Each transformer has 8 attention heads and 768 dimensional hidden state. Dropout with 0.1 was added to each layer in the transformer. Encoder and decoder initialized from BART base. Hyperparameters δ and ϵ in loss function are set to 0.5. In external knowledge building, we set $\omega = 2$, $\mu = 0.05$.

We trained our model on a 3090 GPU for 20 epochs taking about 5 hours with gradient accumulation every 2 steps. We chose the best checkpoint for the testing through a validation process.

Baselines We compared the proposed model against the following three strong baseline approaches in text style transfer: **BART+R** [Lai et al. (2021] is trained by fine-tune BART model with an extra BLEU reward and style classification reward. This model uses parallel dataset. In order to meet our requirement of unsupervised style transfer setting, we used pseudo-parallel data X^Y_{cont} and Y as input and target in the following experiment. **StyIns** [Yi et al. (2021] leverages the generative flow technique to extract stylistic properties from multiple instances to form a latent style space, and style representations are then sampled from this space. **TSST** [Yi et al. (2021] proposes a retrieval-based context-aware style representation that uses an extra retriever module to alleviate the domain inconsistency in content and style.

5.1 Automatic Metrics

To quantitatively evaluate the effectiveness of style transfer task which calls for both the transfer of styles as well as the preservation of content semantics, we newly designed two metrics of Style Transfer Rate (STR) and Style-Content Trade-off (SCT), respectively.

Content Preservation (CP) It is calculated by the similarity between the input and the transferred output leveraged by standard metric Bertscore [Zhang et al. (2019].

Style Transfer Rate The traditional style transfer methods [Lai et al. (2021] use a well-trained style classifier to testify if a sentence has been successfully transferred into a targeted style. But, this method is more suitable for polar word replacement, such as sentiment transfer in review generation. For the cases of eye-catching or written formality transfer, we propose a rule-based yet easy-to-use transfer metric, named as STR. We calculate the STR according to the percentage of syntactic structures changed between the generated output and its input as follows:

$$STR = \frac{\sum_{i \in C_{test}} \text{structure}(X^i_{cont}) \neq \text{structure}(O^i)}{|C_{test}|} \tag{11}$$

where $|C_{test}|$ is the number of testing data, X^i_{cont} and O^i represent content input and generated output, respectively.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

Dataset	Model	CP	STR	SCT	PPL
TechST	StyIns	0.773	0.377	0.253	48.39
	BART+R	**0.962**	0.394	0.280	92.48
	TSST	0.874	0.488	0.313	104.68
	D-HST	0.665	**0.846**	**0.372**	**15.48**
GYAFC	StyIns	0.811	0.666	0.366	26.51
	BART+R	**0.896**	0.663	0.381	12.61
	TSST	0.829	0.625	0.356	23.19
	D-HST	0.641	**0.944**	**0.382**	**10.11**

Table 2: The automatic evaluation results on our model and all baselines on both TechST and GYAFC datasets.

Models	Interestedness	Fluency
D-HST	1.711	1.763
StyIns	1.05	1.413
BART+R	1.219	1.906
TSST	1.181	1.463

Table 3: Human evaluation.

Style-Content Trade-off In order to integrate the STR and CP into a single measure, we take their harmonic means as follows:

$$SCT = \frac{2}{\frac{1}{\text{STR}} + \frac{1}{\text{CP}}} \tag{12}$$

Language Fluency We fine-tuned the GPT-2 model (Radford et al., 2019) on our stylistic dataset S and use it to measure the perplexity (PPL) on the generated outputs.

5.2 Overall Performance

We compared the performance of our model against with the baselines in Table 2. D-HST performs the best across all the metrics except for the CP metric. From the results we can find that, firstly, our model achieves very obvious advantage in STR metric (nearly 50% margin) indicate the thorough and outstanding performance on style transfer; Secondly, our D-HST identifies the most harmonious balance point between content preservation and style transfer revealed by the SCT metric; Thirdly, our language model GPT-2 was fine-tuned in stylistic data, therefore, the PPL metric favors fluent sentences adhere more closely to the given style format. Although the BART+R model receives best fluency in human evaluation (Table 3), it mostly fails in our automatic fluency metric. When evaluating the content preservation, we discourage the CP metric from being as high as possible since the extremely high similarity (like close to 1) implies the exactly same words are used in sentences. However, what is required is a change in style that involves a particular number of words. Therefore, we argue that CP is acceptable around 0.64-0.66[5], which can preserve the source content while transferring the style.

To gain further insight on the performance of the style transfer, we sampled real examples from our model and baselines on TechST dataset, as shown in Table 4. StyIns and BART+R nearly copy the content of input; TSST has difficulty in generating fluent sentences. D-HST can transfer the style on the premise of basically preserving the content.

5.3 Human Evaluation

To assess the quality of text generated using D-HST from human perspective, we designed two human evaluations based on the performance in TechST dataset. First, we randomly sampled 20 groups head-

[5]we randomly sample 60 headlines from the baseline model and our model evenly, and ask the annotators to select the ones that transfer style and preserve content, and the bertscore of the selected headlines mostly falls between 0.63-071.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

	Example #1	Example #2	Example #3
Input	IBM to Acquire Red Hat for $34 Billion	Microsoft Releases Azure Open AI Service Including Access to Powerful GPT-3 Models	EF Core Database Providers
StyIns	IBM to Acquire Red Hat for $34 Billion	AWS and Cloudflare Add Bot Management Features to Their Firewalls	A Core Database Providers
BART+R	IBM to Acquire Red Hat for $ 34 Billion	Microsoft Releases Azure Open AI Service with Powerful GPT-3 Models	EF Core Database Providers
TSST	IBM to Acquire Red Hat for $ 34 Billion	Microsoft Releases Azure Open AI Service Including Access to Powerful QR Models	Using Core Database Providers
D-HST	Why IBM Acquires Red Hat for $34M	Microsoft Azure: Accessing Open-Source Microsoft Machine Models	Going Into Core Database Providers

Table 4: Example outputs generated by different models. Red parts represent stylistic attributes D-HST captures.

	Example #1	Example #2
Input	The New Microsoft Edge - Microsoft Build 2020	Qwik, a Resumable Javascript Framework
Category	Category2: VP NP	
	Introducing Microsoft's New Microsoft Edge	Using a Resumable JavaScript Framework
	Category95: QP VP NP	
	How to Build Microsoft's New Microsoft Edge	How to Develop a Resumable Javascript Framework

Table 5: Examples of generated headlines given specific style category.

lines generated from baselines and D-HST, respectively. 10 postgraduates annotators were asked to score the candidates according to the following attributes from 0 to 2. *Fluency*: how fluent and readable the headline is? *Interestedness*: is the generated headline interesting? The final score of each model is calculated by averaging all judged scores. The results in Table 3 show that headlines generated by our proposed D-HST model receives most popularity compared to other models, indicated by the *Interestedness* metric. Additionally, both BART+R and D-HST generate fluent headlines.

The second human evaluation was designed to compute the click-through rate based on users' real click behavior. It is the most straightforward method of testifying **attractiveness**. When giving many headlines to real readers, we will examine which model receives the most clicks in this evaluation. Specifically, we selected 11 postgraduate annotators, each of whom was given a list of news headlines. The annotators were asked to click on those headlines that are most attractive to them. To make the selection as fair as possible, we carefully design to let the headlines generated by each model distribute evenly across the list, and the headline order was randomly shuffled to eliminate the effect of position on the probability of being clicked. Finally, each list contained 36 headlines (Each model generates 9 headlines, D-HST and three baselines models compared in this experiment) and the annotators were asked to click on 5 most attractive ones. As shown in Figure 2, the largest rate (reaches 58%) obtained by

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

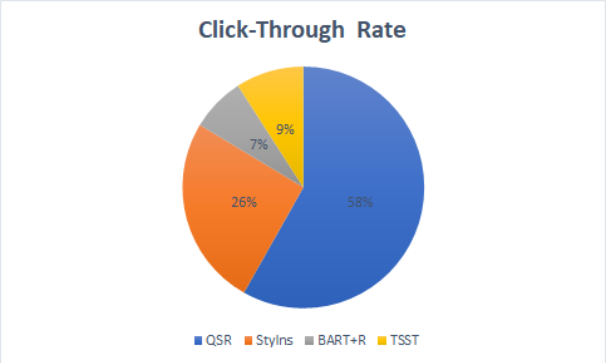


Figure 2: Human evaluation of click-through rate.

Dataset	Strength	CP	STR	SCT	PPL
TechST	γ=0	0.669	0.817	0.368	14.97
	γ=0.1	0.668	0.824	0.369	15.17
	γ=0.3	0.665	0.844	0.372	15.48
	γ=0.5	0.661	0.857	0.373	15.81

Table 6: Evaluation of the style bias strength γ

the D-HST mainly conform to the previously quantitative results. We can conclude that D-HST generates the most appealing and acceptably fluent headlines.

5.4 Discrete Style Space and Controllability

To investigate whether style information is encoded in categories of discrete style space, we inspect to select two kinds of structures to control the generated headlines' styles in the inference stage. The outcomes are shown in Table 5. As it clearly demonstrates, category 2 and category 95 contain two distinct syntactic structures which are "VP NP" and "QP VP NP", respectively. Based on them, given the same input, our D-HST model is capable of generating different attractive headlines match the chosen structures. The results again indicate that the stylistic features are well disentangled and it is easy to control the style of generated results.

5.5 Style Bias Strength

As mentioned in Section 3.5, external knowledge I_{test} is inserted as the style bias in the inference. The style category $K = 324$ in TechST dataset. To investigate how the style bias strength γ affects the final generation, we chose different values on γ and evaluate the performance in a series of automatic metrics, presented in Table 6. Through the experiment, we find that adding a style bias is effective for style transfer, and the scores of STR and SCT increase. The generation quality of the model has no significant fluctuation as the style strength increase, indicating that the model has strong generalization and is insensitive to the parameter.

6 Conclusion

This paper presents an unsupervised model for headline style transfer. It consists of content, style and their mixing components, which are together fed to decoder for headline generation. In particular, we propose to extract the style features in a discrete style space, and each discrete point corresponds to a particular category of the styles. Our system is comprehensively evaluated by both quantitative and qualitative metrics, and it produces cutting-edge outcomes in two typical datasets. Our work can be applied in the scenarios of formality machine translation, politeness transfer in intelligent customer service, spoken language transfer in live broadcast delivery. It can also be followed by the task of paraphrase and data augmentation.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

References

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Qipeng Guo, Zhijing Jin, Ziyu Wang, Xipeng Qiu, Weinan Zhang, Jun Zhu, Zheng Zhang, and Wipf David. 2021. Fork or fail: Cycle-consistent training with many-to-one mappings. In *International Conference on Artificial Intelligence and Statistics*, pages 1828–1836. PMLR.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. *arXiv preprint arXiv:2105.15053*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. *arXiv preprint arXiv:2004.01980*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for non-parallel text style transfer. *arXiv preprint arXiv:1808.04339*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Osama Khalid and Padmini Srinivasan. 2020. Style matters! investigating linguistic style in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 360–369.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you bart! rewarding pre-trained models improves formality style transfer. *arXiv preprint arXiv:2105.06947*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Mingzhe Li, Xiuying Chen, Min Yang, Shen Gao, Dongyan Zhao, and Rui Yan. 2021. The style-content duality of attractiveness: Learning to write eye-catching headlines via disentanglement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13252–13260.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8376–8383.

Teruko Mitamura and Eric Nyberg. 2001. Automatic rewriting for controlled language translation. In *The Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001) Post-Conference Workshop, Automatic Paraphrasing: Theories and Applications*.

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544. PMLR.

Sharmila Reddy Nangi, Niyati Chhaya, Sopan Khosla, Nikhil Kaushik, and Harshit Nyati. 2021. Counterfactuals to control latent disentangled text representations for style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 40–48.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. *arXiv preprint arXiv:2105.08206*.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2018. Adversarial decomposition of text representation. *arXiv preprint arXiv:1808.09042*.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. 2018. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Shangquan Sun and Jian Zhu. Plug-and-play textual style transfer.

Martina Toshevska and Sonja Gievska. 2021. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*.

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. *arXiv preprint arXiv:2011.00403*.

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

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2021. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3801–3807.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 636-647, Harbin, China, August 3 - 5, 2023.

Lexical Complexity Controlled Sentence Generation for Language Learning

Jinran Nie[1], Liner Yang[1]*, Yun Chen[2], Cunliang Kong[1], Junhui Zhu[1], Erhong Yang[1]
[1]Beijing Language and Culture University
[2]Shanghai University of Finance and Economics
njrbarry@gmail.com

Abstract

Language teachers spend a lot of time developing good examples for language learners. For this reason, we define a new task for language learning, lexical complexity controlled sentence generation, which requires precise control over the lexical complexity in the keywords to examples generation and better fluency and semantic consistency. The challenge of this task is to generate fluent sentences only using words of given complexity levels. We propose a simple but effective approach for this task based on complexity embedding while controlling sentence length and syntactic complexity at the decoding stage. Compared with potential solutions, our approach fuses the representations of the word complexity levels into the model to get better control of lexical complexity. And we demonstrate the feasibility of the approach for both training models from scratch and fine-tuning the pre-trained models. To facilitate the research, we develop two datasets in English and Chinese respectively, on which extensive experiments are conducted. Experimental results show that our approach provides more precise control over lexical complexity, as well as better fluency and diversity.

1 Introduction

In the fields of language teaching and acquisition, language instructors and textbook compilers need to make teaching materials with example sentences, either synthetically designed or from authentic resources (Caro and Mendinueta, 2017; Lu et al., 2019). In most cases, they are required to create appropriate example sentences that only use the words at particular complexity for language learners passing through different learning levels (Nordlund and Norberg, 2020; Laufer, 2021), which is very time-consuming and exhausting. Automatically generating good examples can support educators and language learners in obtaining, analyzing, and selecting proper example sentences. Besides, it can also assist in the development of graded reading materials (Ryu and Jeon, 2020; Al-Jarf, 2021; Amer, 2021).

For language learners, good examples are not only required to be fluent and diverse but also match the level of the learners, especially the level of vocabulary. Therefore, it is necessary to effectively control the lexical complexity in good examples generation, which is a task of controllable text generation.

Controllable text generation (CTG), a significant area of natural language generation, contains a series of tasks that aim to generate text according to the given controlled requirements (Prabhumoye et al., 2020; Zhang et al., 2022). CTG systems usually focus on controlling text attributions such as sentiment (Hu et al., 2017; Zhang et al., 2019; Samanta et al., 2020), topic (Dathathri et al., 2019; Tang et al., 2019; Khalifa et al., 2020) or keywords (He, 2021; Zhang et al., 2020; He and Li, 2021), generating poems or couplets with specific formats (Chen et al.,

* Corresponding author:Liner Yang

This work was supported by the funds of Research Project of the National Language Commission No. ZDI145-24.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

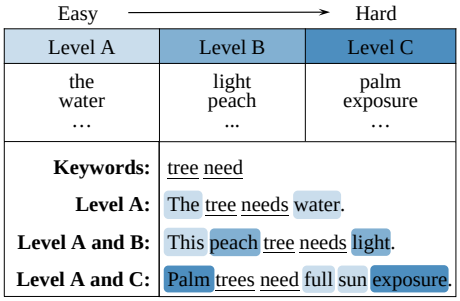


Figure 1: An example for lexical complexity controlled sentence generation. There are three complexity levels (A, B, and C) from easy to hard. Given the keywords "tree" and "need", we will generate "The tree needs water." if required to use all words from level A and generate "This peach tree needs light." if required to use words from both level A and B as both "peach" and "light" are in level B.

2019; Shao et al., 2021; Sheng et al., 2021), and even predicting descriptions from structured data (Zhao et al., 2020; Su et al., 2021; Ribeiro et al., 2021). However, few works have been devoted to strict control over the lexical complexity for text generation. Although lexical simplification has been paid attention to the text simplification task through substitution (Kriz et al., 2018), it cannot strictly control the lexical complexity levels of the generated sentence.

To this end, we propose a new task of lexical complexity controlled sentence generation, which requires that keywords and complexity levels be given to generate a sentence including the keywords and consisting of the words in the given complexity levels. For example, as shown in Figure 1, we assume that there are three complexity levels (A, B, and C) from easy to hard. Given the keywords, we can generate sentences consisted with words of different complexity according to the given levels.

It is challenging to generate fluent sentences for given keywords while using the words only at specific complexity levels. This can be regarded as an extension and a particular case of lexical CTG task (He and Li, 2021; Miao et al., 2019; Zhang et al., 2020). Differently, it combines two aspects of constraints during generation: keywords constraint the semantics, and lexical complexity levels constraint the surface form. It is difficult for the model to select suitable words from a specific subspace satisfying the above two constraints in each generation process. We formulate this problem in Section 2.1.

Some previous works can be customized as solutions to this problem, which are divided into three branches: controlled decoding, prompting, and reranking. The first method forces to change the probability distribution during the decoding phase to ensure that only words of the specified levels are used in the generation (Dathathri et al., 2019; Post and Vilar, 2018). But the hard constraint may lead to poor quality generation quality. The second one considers lexical complexity through prompting (Brown et al., 2020; Raffel et al., 2020; Li and Liang, 2021) in the input of the model, which introduce coarse grained information of training and inference. The method of reranking is to select the sentence that best meets the lexical complexity requirements from the candidates (Ravaut et al., 2022; Pandramish and Sharma, 2020), which executes after decoding and does not consider lexical complexity in the training time.

The complexity constraint requires models to aware of lexical complexity and respond to complexity control signals. Therefore, we use two mechanisms as enhancements to the transformer-based models. *For the complexity awareness*, we propose the Complexity Embedding (**CE**) method, which represents the complexity levels with trainable embeddings. We incorporate the CEs into both training and prediction processes by fusing the CEs and word embeddings as token representations, which is simple but effective. *For responding to complexity control signals*, we concatenate special tokens corresponding to specific complexity levels with the keywords as the input sequence. To combine the awareness and response, we use CEs to represent these

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

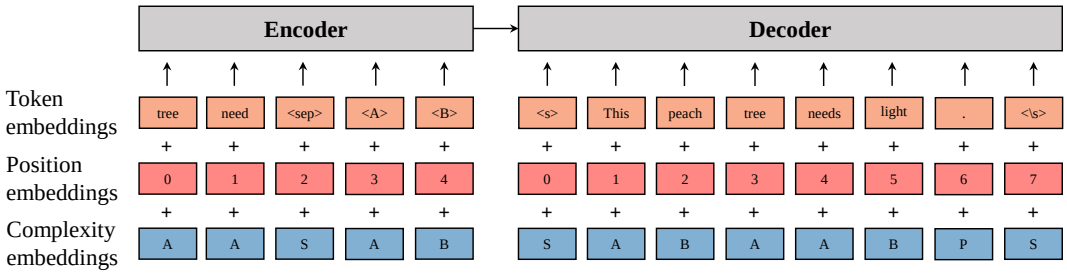


Figure 2: Encoder-Decoder model with our proposed CE method. The representation of each input token is a summary of three embeddings, which are token embedding, position embedding, and complexity embedding. And we concatenate the keywords and complexity level tokens as the input sequence of the encoder. Note that the special tokens correspond to the complexity level of "S", and the punctuation correspond to "P".

special tokens. The experiments show that our proposed method is effective for both training from scratch and fine-tuning the pre-trained language models. And compared to the baseline methods, our method achieves significant improvement in the restriction of lexical complexity levels and generation quality. Our main contributions include:

- We propose a new task of lexical complexity controlled sentence generation and two datasets in English and Chinese for this task. To evaluate the satisfaction of the lexical complexity constraint, we develop four metrics.
- We propose a new method for this task based on complexity embedding.
- The experimental results show that the complexity embedding method we proposed significantly outperforms the baseline methods which are implemented for this task.

2 Method

2.1 Problem Definition

Lexical Complexity Controlled Sentence Generation aims at keywords to sentence generation with desired complexity levels. First, we give the keywords set $K = \{k_1, k_2, ..., k_m\}$ and the complexity levels $L = \{l_1, l_2, ..., l_n\}$ which correspond to a subset $D = \{W_1 \cup W_2 \cup ... \cup W_n\}$ of the whole vocabulary V and W_i is the word set of complexity level l_i. The control elements in this task include three parts:

First, we define a predicate $F(K, Y)$ to be a boolean function indicating the occurrence of keyword k_i in a generated sequence $Y = y_1, y_2, ..., y_t$, and t is the sequence length.

$$C_1 = F(K, Y) \tag{1}$$

$$F(K, Y) \equiv \forall\ i, k_i \in Y \tag{2}$$

where C_1 is the keywords constraint which means the keywords are required to be included in the generated sentence.

Second, we define a predicate $G(Y, D)$ to be a boolean function indicating the occurrence of a word y_i which is a word of the sentence Y in a word set D.

$$C_2 = G(Y, D) \tag{3}$$

$$G(Y, D) \equiv \forall\ i, y_i \in D \tag{4}$$

where C_2 is the complexity constraint on word which means the words in the generated sentence are required to be the words of the given complexity levels.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Then, we define a predicate $H(Y, W_i)$ to be a boolean function indicating that there exist at least one word in the generated sentence in the W_i.

$$C_3 = H(Y, W_1) \wedge H(Y, W_2) ... \wedge H(Y, W_n) \tag{5}$$

$$H(Y, W_i) \equiv \exists\ j, y_j \in W_i \tag{6}$$

where C_3 is the constraint on the species of complexity level which means the lexical levels of the generated sentence need cover all the given levels.

The task requires to seek optimal sequences in which all constraints are satisfied as much as possible. The formula is as follows:

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \log P_\theta\big(Y|K, L\big) \quad \text{where} \quad \sum_{i=1}^{N} C_i = N \tag{7}$$

where N is the number of constraints and $N = 3$.

2.2 Complexity Embedding

As illustrated in Figure 2, our model is based on the encoder-decoder architecture. To make the model aware of the complexity levels, we fuse the complexity into the task by designing a lexical complexity embedding for each token. To make the model respond to specific complexity levels, we insert special tokens corresponding to complexity levels into the input sequence as controllable elements. This section introduces these two key components as well as the training and inference strategy.

We initialize a learnable matrix $\mathbf{M} \in \mathbb{R}^{U \times dim}$ as representations of complexity levels, where U is the total number of complexity levels, and dim is the dimensions of each embedding. For each token input to the encoder and decoder, we retrieve a predefined hash-table to obtain its complexity level l_i. Then we get the corresponding complexity embedding by $com_i = \mathbf{M}_i$. The final embedding of this token emb_i is as following:

$$emb_i = tok_i + pos_i + com_i \tag{8}$$

where tok_i and pos_i are token and positional embeddings, which are obtained according to Transformer model(Vaswani et al., 2017).

For example, as shown in Figure 2, when two keywords "tree" and "need" along with two complexity levels A and B are required, the sentence "This peach tree needs light." is generated which satisfies both constraints. We use different complexity representations (mapping into a complexity embedding) for words of different complexity levels. And the complexity representations of special tokens and punctuation are also different.

In practice, we apply the BPE (byte pair encoding) (Sennrich et al., 2015) algorithm to split words into sub-word tokens to mitigate the OOV (out-of-vocabulary) problem. We mark each sub-word with the same complexity level as the original word. More details about the complexity levels can be found in the Appendix A.

2.3 Controllable Elements

As illustrated in Equation 4, each word in the sentence Y is constrained to the word set D. To achieve this, we design a set of special tokens $Z = \{z_1, z_2, \ldots, z_n\}$, where each token corresponds to a complexity level in L.

We concatenate the keywords and the special tokens as the input sequence $X = [K; \langle sep \rangle; Z]$. And we refer the special tokens Z as controllable elements, as they control the complexity of the generated sentence. Note that the complexity embedding of z_i is that of the level l_i.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

2.4 Training Complexity Embedding

We train the complexity embedding in the Transformer model from scratch or fine-tune the pre-trained model discriminatively as there is no complexity embedding layer in the pre-trained process. If a model is trained from scratch, the parameters of complexity embedding will be trained the same as other parameters in the model. If the complexity embedding is added to a pre-trained model for fine-tuning, we first train the complexity embedding layer by fixing the original parameters of the pre-trained model and then fine-tune the whole model.

During the training process, in fact, both the word embedding and the complexity embedding are in a teach-forcing pattern through the ground truth. At the time of inference, the next word embedding at each step will be predicted by the probability distribution of the vocabulary of the model. Since the complexity level of the next word is unknown at each step of the inference stage, we utilize a look-up table method to map the predicted token id to complexity id. The table is a mapping relation between the token id and its complexity id on the whole vocabulary. At each step, the token id will be predicted by the model. We get its complexity id through its token id and the table. The complexity id and token id will then be given as the input for the next step of inference.

2.5 Length and Syntactic Complexity Control

The length of the generated text is also a factor that language learners may consider, and there is a correlation between text length and syntactic complexity. From a statistical view, text length and syntactic complexity are generally positively correlated. Thus, we design a method to dynamically control text length and syntactic complexity, which is used in the decoding stage. We set three sentence length modes: short, normal, and long, and the sentence length mode also corresponds to the syntactic complexity. We introduce length penalties to beam search in the decoding time in different modes. The formula for calculating the penalty coefficient is as follows:

$$Penalty = N^{pen} \tag{9}$$

where N is the counts of keywords, $pen = -1, 0, 1$ if the mode is short, normal or long respectively. We have observed from statistics that the larger the number of given keywords leads the longer the generated sentences. Therefore, we set the relationship between the length penalty and the number of keywords. In the mode of short or long, if the number of keywords is larger, the greater the penalty required.

3 Datasets and Evaluation Metrics

3.1 Dataset Construction

We present two datasets for lexical complexity controlled sentence generation in English and Chinese. The English raw corpus is collected from the monolingual English News dataset in ACL2019 WMT. The Chinese raw corpus is collected from 500 textbooks for Chinese L2 learners. We adopt the English word complexity levels in the Common European Framework of Reference for Languages (CEFR) [0] which is divided into six complexity levels (A1, A2, B1, B2, C1, and C2). The word complexity levels in Chinese Proficiency Grading Standards for International Chinese Language Education (CPGS) [1] is divided into seven complexity levels (1 to 7). The process for cleaning data is divided into three steps: split the raw data into sentences and choose the proper sentences; obtain the keywords from the sentences; get the lexical complexity levels from the sentences. More details of the two datasets are in the Appendix B.

3.2 Evaluation Metrics

Generated Quality To evaluate the quality of generated text, we employ some automatic evaluate metrics in three aspects. 1) N-gram Similarity with References: we use **BLEU** (Papineni et

[0] https://www.englishprofile.org/wordlists/evp
[1] http://www.chinesetest.cn

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

al., 2002), **METEOR** (Lavie and Agarwal, 2007), and **NIST** (Doddington, 2002) evaluate the difference between generated texts and reference texts, which are commonly utilized in machine translation and text generation. 2) Diversity: We use 2-gram and 4-gram of **Entropy** (Zhang et al., 2018) and 1-gram and 2-gram of **Distinct** (Li et al., 2015) to evaluate lexical diversity. 3) Fluency: Following previous works (Zhang et al., 2020; He and Li, 2021), to assess the fluency of generated sentences, we report the perplexity (**PPL**) over the test set using the pre-trained GPT-2 (Radford et al., 2019) large model.

Satisfaction of Lexically Controlling The control elements of lexical complexity controlled sentence generation have introduced in the Section 2.1. Our metrics are corresponding to the three constraints.

- **Keywords Constraint**. For this aspect, we introduce Keywords Constraint (**K-C**) satisfaction metric on word-level, which is computed using the percentage of the keywords contained in the generated sentences. The formular describe is as below:

$$K - C = \frac{1}{N} \sum_{i=1}^{N} \text{count}_i^{C_1} / m_i \tag{10}$$

  where N is the total number of samples in the test dataset, $\text{count}_i^{C_1}$ is the number of keywords included in the generated sentence of the i-th sample, which satisfy the constraint of C_1, and m_i is the number of the keywords of the input on the i-th sample.

- **Word Complexity Constraint**. The purpose of this metric is to calculate the Accuracy (**ACC**) of the words that meet the lexical complexity levels requirement in the generated sentence. As shown in the following formula:

$$ACC = \frac{1}{N} \sum_{i=1}^{N} \text{count}_i^{C_2} / t_i \tag{11}$$

  where $\text{count}_i^{C_2}$ is the number of the words that satisfy the constraint C_2 of the i-th sample, and t_i is the length of the generated sentence of the i-th sample.

- **Complexity Levels Constraint**. We propose three metrics to evaluate the satisfaction of the species of the required complexity levels. It is unreasonable that the ACC is still 100% if given two complexity levels but the words of generated sentence only covers one of the levels. Thus we design the metrics of Precision (**P**), Recall (**R**), and **F1** to calcuate the saticfaction of complexity level constraint. The formular discribes are as follows:

$$P = \frac{1}{N} \sum_{i=1}^{N} \text{count}_i^{C_3} / g_i \tag{12}$$

$$R = \frac{1}{N} \sum_{i=1}^{N} \text{count}_i^{C_3} / n_i \tag{13}$$

$$F1 = \frac{2}{N} \sum_{i=1}^{N} \text{count}_i^{C_3} / (n_i + g_i) \tag{14}$$

  where $\text{count}_i^{C_3}$ is the number of the complexity levels satisfy the constraint C_3 of the i-th sample, n_i is the number of the complexity levels given in the source of the i-th sample, and g_i is the number of the complexity levels of the generated sentence of the i-th sample.

4 Experiments

Our experiments are based on the two datasets introduced in Section 3. Besides the strong baselines of controlled decoding, prompting and reranking mentioned in Section 4.2, we generate the sentence by setting the keys as the input directly as the basic baseline (K2S). This baseline does not require complexity levels, which are just learnt from the data. Our evaluations include automatic evaluation and human evaluation. The automatic metrics have been introduced in the Section 3.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Metrics	BLEU(%)		NIST(%)		METEOR(%)	Entropy(%)		Distinct(%)		PPL
	B-2	B-4	N-2	N-4		E-2	E-4	D-1	D-2	
Training Transformer from scratch										
K2S	16.58	4.57	3.14	3.27	15.23	8.20	10.23	**5.93**	24.76	74.91
Ctrl-decoding	12.12	3.16	2.45	2.61	11.72	7.28	9.22	5.27	20.14	286.50
Prompting	18.19	5.73	3.57	3.64	15.93	8.30	10.36	6.10	25.55	52.10
Reranking	**18.47**	6.27	3.52	3.60	15.99	7.87	9.79	5.93	22.70	47.81
CE (ours)	18.37	**6.66**	**3.64**	**3.69**	**16.06**	**8.43**	**10.47**	5.80	**25.75**	**42.06**
Fine-tuning BART										
K2S	17.40	5.96	3.20	3.26	15.60	8.60	10.52	6.36	28.53	33.11
Ctrl-decoding	14.17	3.55	2.73	2.48	13.15	8.03	9.87	5.96	21.96	223.43
Prompting	19.36	6.88	3.59	3.67	16.09	**8.93**	**10.81**	**7.22**	**33.84**	39.65
Reranking	18.95	6.54	3.54	3.58	16.03	8.72	10.67	6.60	30.09	34.24
CE (ours)	**19.80**	**7.22**	**3.61**	**3.69**	**16.34**	8.50	10.48	6.41	27.56	**28.48**

Table 1: Generation quality evaluation results on English dataset.

Metrics (%)	K-C	ACC	P	R	F1
Training Transformer from scratch					
K2S	96.93	95.68	89.03	83.27	84.93
Ctrl-decoding	85.56	99.02	97.84	83.51	89.19
Prompting	96.85	98.91	97.35	90.86	93.46
Reranking	97.33	96.80	91.81	87.97	88.98
CE (ours)	**98.00**	**99.10**	**98.09**	**92.84**	**94.96**
Fine-tuning BART					
K2S	97.51	95.26	88.79	84.63	85.58
Ctrl-decoding	89.73	**99.34**	**98.57**	84.19	90.33
Prompting	96.57	97.79	95.77	90.17	92.25
Reranking	98.52	96.10	92.36	88.96	91.87
CE (ours)	**98.68**	99.13	98.54	**93.72**	**95.77**

Table 2: Satisfaction of controlling evaluation results on English dataset.

4.1 Experimental Setup

Our experimental setup contains two aspects:training from scratch and fine-tuning. From scratch training experiments are on the Transformer model (Vaswani et al., 2017), which is the most widely used model in text generation. The fine-tuning experiments are on the pre-trained model of BART (Lewis et al., 2019), which has superior generation ability. During inference, we run greedy decoding on all models for a fair comparison. We implement all models with the Fairseq library [2] and the BART pre-trained model is from HuggingFace Transformers library [3] (Wolf et al., 2019). All models are trained and tested on NVIDIA TITAN Xp GPU.

From Scratch Training Setup We adopt the typical Transformer (Vaswani et al., 2017) as the model trained from scratch. We utilize a learning rate of 3e-4 and set the warming-up schedule with 4000 steps for training. We train our model for around 100 epochs. The optimization algorithm is Adam (Kingma and Ba, 2014). We set the maximum number of input tokens as 8192, which is the same as transformer-based baselines. **Fine-tuning Setup** We initialize our model with BART-base (Lewis et al., 2019), which has comparable parameters to generation baselines. For generation baselines and our models, we use Adam (Kingma and Ba, 2014) with an initial learning rate of 1e-5 to update parameters for four epochs and choose the checkpoints with the lowest validation loss. We train our model for around 30 epochs. We set the maximum number of input tokens as 2048.

[2]https://github.com/pytorch/fairseq
[3]https://github.com/huggingface/transformers

Metrics	BLEU(%)		NIST(%)		METEOR(%)	Entropy(%)		Distinct(%)		PPL
	B-2	B-4	N-2	N-4		E-2	E-4	D-1	D-2	
Training Transformer from scratch										
K2S	13.92	4.17	2.73	2.76	15.00	8.83	10.20	8.60	37.70	48.32
Ctrl-decoding	12.84	3.57	2.48	2.50	13.70	8.70	10.30	6.08	34.90	224.59
Prompting	13.90	3.81	2.70	2.73	14.35	8.53	10.05	7.47	33.35	45.61
Reranking	15.46	5.37	**2.98**	**3.02**	15.34	8.84	10.15	9.13	37.88	38.56
CE (ours)	**15.69**	**6.27**	2.91	2.94	**16.04**	**9.28**	**10.58**	**10.68**	**47.71**	**34.53**
Fine-tuning BART										
K2S	14.97	4.39	3.08	3.10	16.56	8.60	10.06	9.91	37.13	**21.76**
Ctrl-decoding	12.54	3.71	2.38	2.55	14.04	8.73	10.25	9.96	37.85	129.86
Prompting	16.81	5.47	3.15	3.17	16.24	8.69	10.13	10.04	38.33	31.75
Reranking	16.53	6.42	**3.29**	**3.36**	16.61	8.81	10.08	10.15	38.96	53.47
CE (ours)	**17.07**	**6.46**	3.18	3.26	**16.73**	**9.34**	**10.27**	**10.55**	**48.76**	26.52

Table 3: Generation quality evaluation results on Chinese dataset.

Metrics (%)	K-C	ACC	P	R	F1
Training Transformer from scratch					
K2S	87.36	92.74	85.40	68.40	73.75
Ctrl-decoding	71.83	**99.96**	**99.96**	61.79	74.73
Prompting	85.54	98.88	97.79	80.23	86.88
Reranking	88.22	96.70	93.05	75.74	81.59
CE (ours)	**89.61**	98.87	97.49	**88.80**	**92.17**
Fine-tuning BART					
K2S	92.12	93.73	86.88	68.87	74.37
Ctrl-decoding	82.52	**99.18**	**98.65**	65.26	76.41
Prompting	86.94	98.73	97.98	81.78	88.02
Reranking	90.14	97.21	95.44	76.78	83.95
CE (ours)	**92.58**	99.07	97.91	**89.34**	**92.85**

Table 4: Satisfaction of controlling evaluation results on Chinese dataset.

4.2 Baseline

Controlled decoding We consider a strategy of controlled decoding (Dathathri et al., 2019) to realize the generated sentence consists of the words belonging to the given complexity levels. Since we know the words of the complexity level to be used in the sentence, we can restrict the words of the subset of the vocabulary to only be used in the decoding stage. The specific method is to set the probability of words outside the subset to zero so that they can meet the requirements of the word complexity level.

Prompting Prompting is another feasible method for controlled text generation (Zou et al., 2021). Inspired by the prefix-tuning (Li and Liang, 2021), which uses continuous vectors as prompts, we add the required complexity levels as the prefix for controlling in the input of the generation model.

Reranking Inspired by previous works (Ravaut et al., 2022; Pandramish and Sharma, 2020), we select the sentence that best meets the lexical complexity requirements from the N-best candidates. We take the score that is the sum of ACC score and $F1$ score on the test reference hypothesis from this N-best list and choose the candidate that has the largest score. The detail of the re-ranking method is shown as the Algorithm 1 in Appendix C.

4.3 Experimental Results

The experimental results on English dataset are shown in Table 1 and Table 2. From the evaluation of generation quality in Table 1, it can be seen that the method of complexity embedding has competitive results in different aspects, especially on fluency. In general, the CE method

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Metrics (%)	Semantics	Fluency	Diversity
English dataset			
Ctrl-decoding	2.68	2.40	2.92
Prompting	**4.63**	3.25	3.45
Reranking	4.60	3.39	3.40
CE (ours)	4.62	**3.82**	**3.54**
Chinese dataset			
Ctrl-decoding	3.89	2.82	3.27
Prompting	4.23	3.08	3.02
Reranking	4.37	3.29	3.16
CE (ours)	**4.57**	**3.80**	**3.71**

Table 5: Human evaluations for fine-tuning BART model on two datasets.

has better performance in the control of lexical complexity, especially on the metrics of R and F1. The method of controlled decoding has poor performance on PPL because it forces the distribution of the logits to concentrate on the words of given complexity levels in the decoding stage. This hard constraint pattern will impact the fluency of the generated sentences. But its performances on the metrics of ACC and P are better than other methods from Table 2. The methods of prompting and reranking are two competitive baselines. The prompting method has better performance in the control of the word complexity because it has considered the word complexity levels in training. But the reranking method has better generation quality on the whole metrics of Table 1.

The experimental results on Chinese dataset are shown in Table 3 and Table 4. We can draw similar conclusions from these two tables. Our approach performs well in terms of both text generation quality and lexical complexity control. The rerank approach outperforms prompt in all aspects of generation quality, both in terms of similarity to ground truth and in diversity and fluency, and even achieves the best NIST metrics for the Chinese dataset.

4.4 More Analyses and Discussion

The CE method we proposed has an excellent performance in controlling lexical complexity. The reason is that the CE method not only keeps the consistency of training and prediction but also considers the information of the complexity at the token level. Thus, it has more precise control of lexical complexity. And it also has competitive generation quality in the aspect of fluency and similarity with the reference. From the metrics of Entropy and Distinct, its diversity has a little poor performance in terms of the fine-tuning pattern on the English dataset. We think the main reason is that the vocabulary of the English word complexity levels is less than which of the Chinese, so the token level restrictions of complexity embedding will impact the diversity of the sentences. The Chinese dataset, on the other hand, has a much larger coverage of voabulary with complexity and the dataset comes from the field of second language teaching, so the diversity of our model is better. It is worth noting that our CE method performs best in terms of lexical complexity control, especially the metrics of K-C, R, and F1, compared to the baseline model. This indicates that the CE method has higher coverage on complexity levels due to it takes into account the complexity of each word.

4.5 Length and Syntactic Complexity Control

We evaluate the length and the depth of the syntactic tree of generated text in the modes of short, normal and long, which can reflect the complexity of the generated text. As shown in the table 6, the experiment of controlling sentence length and syntactic complexity is on the English dataset. In the long mode, the generated sentences are longer, and the syntactic tree is deeper. In the short mode, the generated sentences are shorter, and the syntactic tree depth is smaller. The length penalty in the decoding stage can effectively control the sentence length while affecting the complexity of the syntax.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Metric/Mode	Short	Normal	Long
Length	15.3	24.6	36.8
Syn-Depth	9.3	11.1	13.5

Table 6: The length and depth of syntactic tree of generated sentences in different modes.

4.6 Human Evaluation

We conduct a human evaluation to further compare our model with the three baselines with fine-tuning the BART model on two datasets. For each model, we randomly select 200 generated sentences from the test set for each dataset and invite three annotators to label the sentences, who are postgraduates of the major in linguistics. To evaluate the quality of the sentences, annotators rate the sentences on three dimensions: semantic consistency between the keywords and sentence; the fluency of the sentence; the diversity of the sentence (Zhang et al., 2020). The score is range from 0 to 5. As shown in Table 5, our method has better performance at the three aspects of human evaluation, especially the fluency and diversity. We give some real cases of two datasets in the Appendix D. From the cases study we can find that the CE method can cover more lexical complexity levels than the baseline methods. This also confirms the reason why the CE method that we proposed has a better performance on R and F1 metrics of the automatic evaluation.

5 Related Work

Lexical constraint text generation is to generate a complete text sequence, given a set of keywords as constraints (Zhang et al., 2020). Previous works involve enhanced beam search (Post and Vilar, 2018; Hu et al., 2019) and the stochastic search methods (Zhang et al., 2020; Sha, 2020). Currently, Seq2Seq-based models such as Transformer and pre-trained models have been increased in generation with lexical constraint (Wang et al., 2021b; Liu et al., 2020; Wang et al., 2021a; Fan et al., 2020; Liu et al., 2021). But lexically constrained text generation is not able to control the complexity of words used in the generation, which is different from our work.

Text readability assess research has shown that lexical complexity is also a crucial aspect of evaluating the complexity of a text for text readability assess task (Chakraborty et al., 2021). In the relevant study of sentence-level readability, it is generally accepted that apart from sentence length, the most predictive indicator is the number of difficult words in the sentence (Weiss and Meurers, 2022). In our work, we follow the definition and vocabulary of lexical complexity of text readability assess.

Text simplification In text simplification field, lexical substitution, the replacement of complex words with simpler alternatives, is an integral part of sentence simplification and has been the subject of previous work (Alonzo et al., 2020; Nishihara et al., 2019). Differently, our work can strictly control the lexical complexity levels of the generated sentence, not only simplify the lexical complexity.

6 Conclusions

To summarize, we introduce a new task of lexical complexity controlled sentence generation, where word complexity must be strictly controlled in generating. To promote the development of this task, we develop two datasets and four metrics for the controlled element. In this paper, we also develop a series of alternate solutions for this task and propose a novel method based on complexity embedding to obtain better control of lexical complexity in a generation. Our results indicate that the complexity embedding method has better performance in controlling the lexical complexity and competitive generation quality.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

References

Reima Al-Jarf. 2021. Efl students' difficulties with lexical and syntactic features of news headlines and news stories. *Technium Soc. Sci. J.*, 17:524.

Oliver Alonzo, Matthew Seita, Abraham Glasser, and Matt Huenerfauth. 2020. Automatic text simplification tools for deaf and hard of hearing adults: Benefits of lexical simplification and providing users with autonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Mohammad Ahmad Bani Amer. 2021. Lexical density and readability of secondary stage english textbooks in jordan. *International Journal for Management and Modern Education*, 2(2):11–20.

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

Keiby Caro and Nayibe Rosado Mendinueta. 2017. Lexis, lexical competence and lexical knowledge: a review. *Journal of Language Teaching & Research*, 8(2).

Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. 2021. Simple or complex? learning to predict readability of bengali texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12621–12629.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *IJCAI*, pages 4925–4931.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Zhihao Fan, Yeyun Gong, Zhongyu Wei, Siyuan Wang, Yameng Huang, Jian Jiao, Xuanjing Huang, Nan Duan, and Ruofei Zhang. 2020. An enhanced knowledge injection model for commonsense generation. *arXiv preprint arXiv:2012.00366*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Xingwei He and Victor OK Li. 2021. Show me how to revise: Improving lexically constrained sentence generation with xlnet. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12989–12997.

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with bart. *arXiv preprint arXiv:2109.12487*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850.

Kenji Imamura and Eiichiro Sumita. 2017. Ensemble and reranking: Using multiple models in the nict-2 neural machine translation system at wat2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 127–134.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2020. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Reno Kriz, Eleni Miltsakaki, Marianna Apidianaki, and Chris Callison-Burch. 2018. Simplification using paraphrases and context-based lexical substitution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 207–217.

Batia Laufer. 2021. Lexical thresholds and alleged threats to validity: A storm in a teacup? *Reading in a Foreign Language*, 33(2):238–246.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. 2020. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. *arXiv preprint arXiv:2009.12677*.

Yixian Liu, Liwen Zhang, Wenjuan Han, Yue Zhang, and Kewei Tu. 2021. Constrained text generation with global guidance–case study on commongen. *arXiv preprint arXiv:2103.07170*.

Dawei Lu, Xinying Qiu, and Yi Cai. 2019. Sentence-level readability assessment for l2 chinese learning. In *Workshop on Chinese Lexical Semantics*, pages 381–392. Springer.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pages 260–266.

Marie Nordlund and Cathrine Norberg. 2020. Vocabulary in efl teaching materials for young learners. *International Journal of Language Studies*, 14(1):89–116.

Vinay Pandramish and Dipti Misra Sharma. 2020. Checkpoint reranking: An approach to select better hypothesis for neural machine translation systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 286–291.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *arXiv preprint arXiv:1804.06609*.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Mathieu Ravaut, Shafiq Joty, and Nancy F Chen. 2022. Summareranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. *arXiv preprint arXiv:2203.06569*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Leonardo FR Ribeiro, Yue Zhang, and Iryna Gurevych. 2021. Structural adapters in pretrained language models for amr-to-text generation. *arXiv preprint arXiv:2103.09120*.

Jisu Ryu and Moongee Jeon. 2020. An analysis of text difficulty across grades in korean middle school english textbooks using coh-metrix. *Journal of Asia TEFL*, 17(3):921.

Bidisha Samanta, Mohit Agarwal, and Niloy Ganguly. 2020. Fine-grained sentiment controlled text generation. *arXiv preprint arXiv:2006.09891*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Lei Sha. 2020. Gradient-guided unsupervised lexically constrained text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8692–8703.

Yizhan Shao, Tong Shao, Minghao Wang, Peng Wang, and Jie Gao. 2021. A sentiment and style controllable approach for chinese poetry generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4784–4788.

Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13798–13805.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. *arXiv preprint arXiv:2108.13740*.

Hongyin Tang, Miao Li, and Beihong Jin. 2019. A topic augmented text generation model: Joint learning of semantics and structural features. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5090–5099.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021a. Retrieval enhanced model for commonsense generation. *arXiv preprint arXiv:2105.11174*.

Yufei Wang, Ian Wood, Stephen Wan, Mark Dras, and Mark Johnson. 2021b. Mention flags (mf): Constraining transformer-based text generators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 103–113.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for german language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *arXiv preprint arXiv:1809.05972*.

Rui Zhang, Zhenyu Wang, Kai Yin, and Zhenhua Huang. 2019. Emotional text generation based on cross-domain sentiment transfer. *IEEE Access*, 7:100081–100089.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020. Pointer: Constrained progressive text generation via insertion-based generative pre-training. *arXiv preprint arXiv:2005.00558*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491.

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2450–2460.

A Complexity Embedding Id

The English words have six levels. And the Chinese words have seven levels (Diff 1-7). We give the design of the complexity embedding id for this two language in the table 7. Note that, if a word is out of the complexity level vocabulary, its complexity is "$\langle out \rangle$" which is mapping into id 7 in English corpus and 8 in Chinese corpus. In addition, the special tokens such as "$\langle s \rangle$" "$\langle pad \rangle$" "$\langle \backslash s \rangle$" "$\langle unk \rangle$" are the common meaning in data preprocessing for model training.

English		Chinese	
Token	Id	Token	Id
Punctuation	0	Punctuation	0
A1-C2	1-6	Diff 1-7	1-7
$\langle out \rangle$	7	$\langle out \rangle$	8
$\langle sep \rangle$	8	$\langle sep \rangle$	9
$\langle s \rangle$	8	$\langle s \rangle$	9
$\langle pad \rangle$	8	$\langle pad \rangle$	9
$\langle \backslash s \rangle$	8	$\langle \backslash s \rangle$	9
$\langle unk \rangle$	8	$\langle unk \rangle$	9

Table 7: Complexity Embedding Id.

B Details of Datasets Construction

B.1 English Dataset

We adopt the English word complexity levels in the Common European Framework of Reference for Languages (CEFR) [4] which is divided into six complexity levels (A1, A2, B1, B2, C1, and C2). First, we need to restrict the words in the corpus to ensure most of the words are in the complexity level vocabulary. Then, we need to extract keywords from the sentences. In this process, we command the number of keywords is related to the length of the sentence, and the number of keywords is between 1 to 5. Finally, we obtain the complexity information of each sentence through the complexity level vocabulary. The English raw corpus is collected from the monolingual English News dataset in ACL2019 WMT. We select those sentences which have 90% words in the complexity level vocabulary of CEFR. After the processes mentioned above, we get 199k samples in the English corpus, and we split the train, validation and test dataset as shown in the Table 8.

B.2 Chinese Dataset

The word complexity levels in Chinese Proficiency Grading Standards for International Chinese Language Education (CPGS) [5] is divided into six complexity levels (1 to 7). The Chinese raw corpus is collected from 500 textbooks for Chinese learners. These textbooks contain two types of text: essay and dialogue. We split these texts into sentences and throw away those short sentences. If the raw text is a dialogue, after splitting, we need to remove the speaker's name to guarantee it is a proper sentence. Then, we command the number of keywords is related to the length of the sentence, and the number of keywords is between 1 to 5. After the processes mentioned above, we get 156k samples in the Chinese corpus, as shown in the Table 8.

[4] https://www.englishprofile.org/wordlists/evp
[5] http://www.chinesetest.cn

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

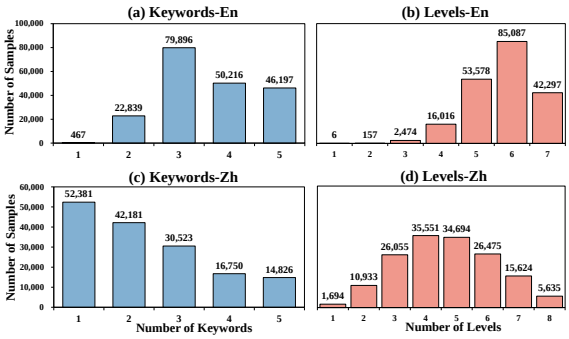


Figure 3: Distributions of the number of keywords and complexity levels.

Dataset	Train	Valid	Test	Total
English	180,000	16,000	3,615	199,615
Chinese	140,000	14,000	2,661	156,661

Table 8: Statistics of the two datasets.

B.3 Analysis of the Datasets

B.3.1 Coverage of Words with Levels

We first analyze the two datasets from the coverage rate of complexity level vocabulary. Due to the requirement of complexity level, the target text is proper to cover most of the vocabulary of complexity level. Both of the two datasets have covered over 93% of the vocabulary of complexity levels.

B.3.2 Distributions of the Number of Keywords and Complexity Levels

One or multiple complexity levels and keywords are given as the input to generate sentences. We give the distribution of the number of keywords and the complexity levels in Figure 3. From the statistics of (a) and (c) in Figure 3, the number of keywords in all samples has covered the range of 1 to 5 both in the English and Chinese datasets, but the distributions are quite different. On account of the average sentence length of English news data is longer than the Chinese corpus, the number of keywords in English is larger. From the statistics in (b) and (d) of Figure 3, the number of complexity levels distribution of the Chinese dataset is close to a standard normal distribution, and the English dataset concentrates on a wider range of complexity levels. This indicates that in the English dataset it tends to use more words of different complexity levels in the same sentence.

C Algorithm of Reranking

The algorithm is the detail of reranking method. We select the sentence that best meets the lexical complexity requirements from the N-best candidates, and $N = 10$. On the test set, We take the sum of ACC score and $F1$ score. The, we choose the candidate that has the largest score.

D Case Study

We choose some cases of the fine-tuning pattern from two datasets. The English cases are in the Table 9, and the Chinese cases are in the Table 10. In both tables, the required keywords as well as appearing in the sentences are shown in blue font, and certain given grades as well as words actually appearing in the sentences for the corresponding grade are shown in red font.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Algorithm 1 Reranking Method

Input: Generated n best candidate sentences $H = (h_0, h_1, h_2, ..., h_{n-1})$ for given keywords and $n = 10$

Output: Sentence having highest score

Let $score = 0$
for each sentence h_j in H **do**
  $ACC = F_{acc}(h_j)$
  $F1 = F_{f1}(h_j)$
  $score_j = ACC + F1$
  if $score_j > score$ **then**
    $score = score_j$
    $ret = h_j$
  end if
end for
return ret

E Related Methods

E.1 Controlled Decoding

The gradients of an external discriminator is directly used to the generation of a pre-trained language model toward the target topic(Dathathri et al., 2019). The output probabilities of a language model is modified by using the output of a discriminator that determines whether the future text will contain the desired attribute. Different from the controlled decoding methods, our method considers the constraint of lexical complexity during both training and prediction.

E.2 Prompting

The prompting method has emerged as a new way to perform natural language processing by conditioning on extra information. Brown et al. propose to use a task description and a few examples to adapt the GPT-3 model to downstream tasks, which is referred to as in-context learning (Brown et al., 2020). Their prompts are manually designed. Gao et al. present LM-BFF for automatic prompts generation(Gao et al., 2020). Liang et al. propose prefix-tuning, which uses continuous vectors as prompts(Li and Liang, 2021). Compared to the prompting method, our method fuses more fine-grained information on lexical complexity in model training.

E.3 Reranking

The reranking approach has been proved to have excellent performance in machine translation (Pandramish and Sharma, 2020) and text generation (Ravaut et al., 2022). The reranking method rescores the n-best candidates through a model or a function and selects the highest scoring candidate as the final prediction (Imamura and Sumita, 2017). Unlike the reranking method, our method do not need to process the outputs after decoding.

F Limitation

Our proposed task has wide applications in the field of language teaching, and the proposed method has precise control over lexical difficulty. However, the task requires that the lexical complexity is known first. The vocabulary difficulty table is the experience summed up by the predecessors, and it is difficult to apply to all vocabulary. Therefore, we are actively exploring how to make the model automatically understand all vocabulary difficulties so that it can cover a wider vocabulary at generation.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Keywords: refuse, accept, country ; **Complexity Levels:** A1, A2, B1, C1	
Ctrl-decoding	I refuse to accept that this is a matter of time for the country, but I do not accept it.
Prompting	I refuse to accept that I am the only one who has been elected to the country, but I am not.
Reranking	I refuse to accept that this is the best thing I can do for the country, ” he said.
CE (ours)	I refuse to accept that this is a country that is going to be a place where people are going to be unsafe.
Keywords: public, maintain, lose ; **Complexity Levels:** A1, A2, B1, B2	
Ctrl-decoding	I do not think the public will maintain the power to lose, but I do not think it will.
Prompting	The public will maintain the public's ability to lose, and the public will not lose, and they will not lose.
Reranking	I don't want to be in public, but I don't want to maintain my weight and lose.
CE (ours)	The public must maintain their faith and not lose , and we will continue to do everything we can to protect them.
Keywords: football, Leicester, City, magical ; **Complexity Levels:** A1, A2, B1, B2, C2	
Ctrl-decoding	I think football is a great way to play for the game and to be able to play for the best of the game against the game against the game against the game and the way we play against the game against the game against the game against the game is not the same, but the way we are magical.
Prompting	I think football is going to be the best of Leicester City, but I think it's a magical moment for us to have to play with.
Reranking	This is a football game for Leicester City, and it's a magical moment for us to be part of the game and be a part of it.
CE (ours)	It's a football club that has been around for a long time and to see Leicester City win the title is magical.

Table 9: Generated examples from the English dataset.

Keywords: 生活 ; **Complexity Levels:** 1, 2, 3, 7	
Ctrl-decoding	生活中，我最喜欢的就是吃面。
Prompting	生活中的一切都是为了生活。
Reranking	我们的生活很美好，也很快乐。
CE (ours)	没有工作压力的生活多让人羡慕啊！
Keywords: 春天, 姑娘 ; **Complexity Levels:** 1, 2, 3	
Ctrl-decoding	春天来了，姑娘又来了，他还在那儿等着我呢。
Prompting	春天来了，姑娘又来了，他的眼睛也变得很美丽。
Reranking	春天到了，一个年轻的姑娘爱上了他。
CE (ours)	春天来了，我认识了一位美丽的姑娘，他非常喜欢这个姑娘。
Keywords: 喜欢 ; **Complexity Levels:** 1, 7	
Ctrl-decoding	我喜欢吃中国菜，我喜欢吃中国菜。
Prompting	我喜欢看中国电影。
Reranking	你喜欢什么样的女孩子呢?
CE (ours)	我最喜欢逗它们玩儿。
Keywords: 别人, 可能, 意思; **Complexity Levels:** 1, 2	
Ctrl-decoding	别人看我懂了，我觉得我有可能去做一个人，意思是要看他。
Prompting	别人可能不会说，如果你觉得自己可能有可能，你可能会觉得自己是个很难的意思。
Reranking	如果别人问你一个问题，你的意思是什么?
CE (ours)	别人可能不知道你的意思，你要做我喜欢的，要我愿意跟别人说。

Table 10: Generated examples from the Chinese dataset.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 648-664, Harbin, China, August 3 - 5, 2023.

Dynamic-FACT: A Dynamic Framework for Adaptive Context-Aware Translation

Linqing Chen, Weilei Wang*
(PatSnap Co., LTD. Suzhou, Jiangsu 215000)
{chenlinqing, wangweilei}@patsnap.com

Abstract

Document-level neural machine translation (NMT) has garnered considerable attention since the emergence of various context-aware NMT models. However, these static NMT models are trained on fixed parallel datasets, thus lacking awareness of the target document during inference. In order to alleviate this limitation, we propose a dynamic adapter-translator framework for context-aware NMT, which adapts the trained NMT model to the input document prior to translation. Specifically, the document adapter reconstructs the scrambled portion of the original document from a deliberately corrupted version, thereby reducing the performance disparity between training and inference. To achieve this, we employ an adaptation process in both the training and inference stages. Our experimental results on document-level translation benchmarks demonstrate significant enhancements in translation performance, underscoring the necessity of dynamic adaptation for context-aware translation and the efficacy of our methodologies.

1 Introduction

Numerous recent studies have introduced a variety of context-aware models aiming to effectively harness document-level context either from the source side (Maruf and Haffari, 2018; Zhang et al., 2018; Miculicich et al., 2018; Tan et al., 2019; Zheng et al., 2020; Kang et al., 2020), target side (Xiong et al., 2019; Yu et al., 2020; Sugiyama and Yoshinaga, 2021), or both (Kuang et al., 2018; Tu et al., 2018; Maruf et al., 2019; Chen et al., 2020; Chen et al., 2022). In the prevailing practice, a context-aware model remains fixed after training and is then employed for every testing document. Nonetheless, this approach presents a potential challenge, as the model is required to encapsulate all translation knowledge, particularly from diverse domains, within a predefined set of parameters. Accomplishing this task within the confines of reality poses a formidable undertaking.

The "one sentence one model" approach for sentence-level NMT, as proposed by (Li et al., 2018), aims to familiarize the model with each sentence in the test dataset by fine-tuning the NMT model for every testing sentence. However, acquiring suitable fine-tuning sentences for a given testing sentence proves to be highly time-consuming, as they require meticulous extraction from the bilingual training data through similarity search. This presents a significant challenge when attempting to replicate their methodology by seeking similar documents from the bilingual document-level training data. Moreover, this approach assesses sentence similarity solely based on the Levenshtein distance, thereby disregarding the document-level context of these sentences extracted from distinct documents.

To address the potential challenge of employing a fixed, trained model for all testing documents, we propose the "one document one model" approach in this paper. This alternative approach aims to achieve the objective by introducing the *document adapter*. Unlike other methods, the adapter relies solely on the input document itself and does not require additional input forms. Its primary function is to reconstruct the original document from a deliberately corrupted version, thereby enabling the model to familiarize itself with the task of document-level translation. Notably, this approach differs from previous methods where the input and output are in different languages, as opposed to the same language. Following adaptation, this modified model is utilized to translate the document.

*Corresponding author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

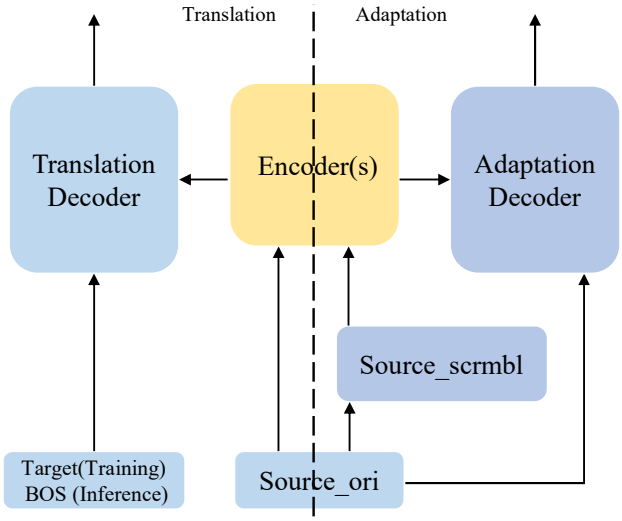


Figure 1: The figure presented in this section depicts the adapter-translator architecture designed for context-aware neural machine translation. In this architecture, the encoder(s) are shared between the adapter denoted as ϕ and the translator denoted as ψ. It is important to note that the translator and adapter constitute two distinct stages within the same model, rather than being treated as separate models.

Both the adapter model and the NMT model employed in our study are context-aware and utilize shared encoder(s), while each having its dedicated decoder. In this paper, we present a training methodology that aims to adapt a pre-trained NMT model to a specific document through a process of alternating document reconstruction and document translation for each document batch. This approach is employed during both the training and inference stages. To evaluate the effectiveness of our proposed approach, we conducted experiments on three English-to-German document-level translation tasks. The results reveal significant enhancements in translation performance, providing strong evidence for the necessity of employing a one document one model approach and the efficacy of our proposed methodology.

Overall, we make the following contributions.

- We present an enhanced context-aware document-level auto-encoder task to facilitate dynamic adaptation of translation models.
- We propose an adapter-translator framework for context-aware NMT. To the best of our knowledge, this is the first study that investigates the one-document-one-model approach specifically for document-level NMT.

2 Adapter-Translator Architecture

The Adapter-Translator architecture entails an iterative procedure involving an adaptation process denoted as ϕ and a translation process denoted as ψ. Figure 1 presents a visual representation of the proposed architecture. The translator ψ, which is a context-aware NMT model, comprises context-aware encoder(s) and a decoder specific to translation.[1] The adapter shares the encoder(s) with the translator while possessing a decoder specifically designed for adaptation. Given a source document $\mathcal{X}$, the corpus processing script generates a deliberately corrupted version $\widehat{\mathcal{X}}$ of the document. This corrupted version is then utilized to optimize the adapter in order to reconstruct the scrambled segments of the original document $\mathcal{X}$. As the encoder(s) are shared between the adapter and the translator, the capability to capture context during document adaptation can also be harnessed during the document translation process. The translation component of this architecture resembles that of other document-level translation models.

[1]It is worth noting that while not all context-aware NMT models possess an additional context encoder (Ma et al., 2020), the adapter-translator architecture can still be adapted to accommodate these models.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

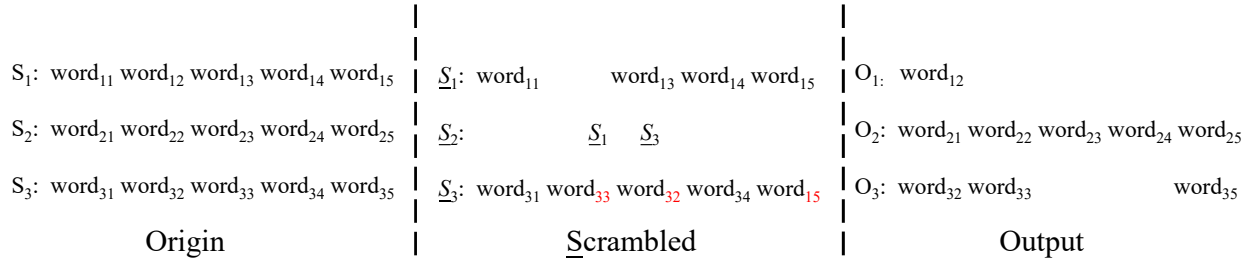


Figure 2: Illustration of document reconstruction task.

Due to its straightforward yet impactful architecture, the proposed method can be employed with diverse document-level translation models.

2.1 Document Adapter

Motivated by the work of (He et al., 2022), we present an adapter-based methodology to restore the scrambled segments of an input document. To be more precise, we adopt a strategy where sentences or words are randomly omitted from the original document, and the adapter is trained to reconstruct these scrambled portions by minimizing the cross-entropy reconstruction loss between the output of its decoder and the corresponding correct part of the original document.

Given a document $\mathcal{X} = (X_i)\,|_{i=1}^{N}$ consisting of N sentences, we apply token substitution, insertion, and deletion operations to each sentence X_i. Following the approach of BERT (Devlin et al., 2019), we randomly select 15% of the tokens. However, unlike BERT, we do not replace these tokens with [MASK] tokens. Instead, the adapter is responsible for identifying the positions that require correct inputs. Furthermore, we do not preserve 10% of the selected tokens unchanged, as our method does not rely on the [MASK] token. In our experiments, we observed that compared to generating the entire original document, generating only the corrected scrambled part significantly reduced the computational time. Nevertheless, this modification did not significantly compromise the model's ability to capture context and become familiar with the document to be translated.

Figure 2 depicts an example involving 3 sentences in the original document. In this example, the first sentence undergoes a word deletion operation, while the third sentence experiences word scrambling and replacement operations. The scrambled preceding and succeeding sentences serve as context for the second sentence. The adapter produces the missing words in the first sentence, the corrected words in the third sentence, and the complete second sentence. While our document reconstruction task draws inspiration from the similar proposal in (Devlin et al., 2019), there exist two significant distinctions. Firstly, instead of substituting selected words with `[MASK]` tokens, we introduce contextual document corruption by allowing token substitution, insertion, and deletion. Secondly, in contrast to BERT, our training objective simultaneously considers the utilization of both sentence-level and document-level context.

To summarize, we define the document adaptation task by employing the following two sub-tasks:

- Sentence-level: The adapter generates corrected words based on a deliberately scrambled version of the original sentence.
- Document-level: The adapter utilizes the concatenated context sentences to generate the original sentences.

2.2 Context-Aware Translator

The context-aware translation model in our framework is designed as a relatively independent model, which shares the encoder(s) with the adaptation model while having a dedicated decoder for translation. This design ensures the flexibility of the framework, allowing it to be easily integrated with different translation models by simply incorporating the adapter model's encoder.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

#	Model	TED		News		Europarl		Average	
		BLEU	Meteor	BLEU	Meteor	BLEU	Meteor	BLEU	Meteor
1	DocT (Zhang et al., 2018)	24.00	44.69	23.08	42.40	29.32	46.72	25.47	44.60
2	+ Adapter	24.70	45.20	23.68	43.01	29.84	47.15	26.07	45.12
3	HAN (Miculicich et al., 2018)	24.58	45.48	25.03	44.02	28.60	46.09	26.07	45.20
4	+ Adapter	24.90	45.89	25.51	44.38	29.07	46.61	26.49	45.63
5	SAN (Maruf et al., 2019)	24.42	45.26	24.84	44.17	29.75	47.22	26.34	45.55
6	+ Adapter	24.80	45.69	25.24	44.63	30.11	48.20	26.72	46.17
7	QCN (Yang et al., 2020)	25.19	46.09	22.37	41.88	29.82	47.86	25.79	45.28
8	+ Adapter	25.83	46.80	22.89	42.40	30.32	48.35	26.35	45.85
9	GCNMT (Chen et al., 2022)	25.81	46.33	25.32	44.35	29.80	47.77	26.98	46.15
10	+ Adapter	26.50	46.96	25.71	44.83	30.43	48.46	27.55	46.75
11	Transformer (Vaswani et al., 2017)	23.02	43.66	22.03	41.37	28.65	45.83	24.57	43.62

Table 1: Performance on test sets. + Adapter indicates we use our proposed context-aware adapter to guidance the context-aware encoder. Significance test (Koehn, 2004) shows that the improvement achieved by our approach is significant at 0.05 on almost all of the above models.

From a structural perspective, this approach facilitates the applicability of the framework to a wide range of translation models. However, in terms of translation performance, there are significant differences between the output of the adaptation phase and the translation phase. Sharing the decoder between these two phases may introduce bias towards shorter output text during translation, given the relatively short length of the corrected scrambled part produced in the adaptation phase. Furthermore, sharing the decoders may increase the vocabulary size of the translation model decoding end and the dimension of the vector, thereby increase the computational cost of training and inference. Additionally, changes in the decoder's vocabulary may alter the semantic space of the translation model, necessitating retraining even if a well-trained translation model is available.

As discussed earlier, the adapter model's decoder only generates the corrected part of the original document. Therefore, employing two different decoders does not significantly impact the time required during the translation inference phase.

2.3 Training and Inference

During the model training phase, the framework follows different procedures based on whether it is built upon a pre-trained translation model or trained from scratch. When using a pre-trained model, the parameters of the translator are frozen, and only the decoder part of the adapter is trained. In the case of training from scratch, parallel corpora are employed as input and output for the translator, while the source corpus and its scrambled versions are used as input and output for the adapter. Training is performed iteratively, alternating between the translation and reconstruction tasks.

In the framework's inference phase, the decoders of both the translator and adapter are frozen for two primary reasons. Firstly, these decoders have undergone sufficient training during the training phase. Secondly, freezing them saves computational time during inference. Similarly, a certain percentage (P%) of the context-aware encoder parameters are also frozen for similar reasons. This not only reduces computational overhead but also facilitates multi-round learning by utilizing multiple scrambled versions of the same document, enabling the translation model to become familiar with the document to be translated. By freezing most of the encoder parameters and increasing the dropout rate, overfitting on a single document is mitigated, preventing potential performance degradation on other documents in the test set.

Specifically, the document restoration process consists of the following steps:

1. Expansion of $\mathcal{X}$: We expand the original document $\mathcal{X}$ by creating K copies, where K is the expansion ratio. Each copy is processed independently, forming instances for the document restoration task.

2. Freezing of Translator and Adapter Parameters: We freeze a portion of the parameters in both the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

#	Model	MT06	MT02	MT03	MT04	MT05	MT08	All		
		BLEU	BLEU	BLEU	BLEU	BLEU	BLEU	BLEU	Meteor	d-BLEU
1	DocT (Zhang et al., 2018)	37.08	43.40	43.83	41.51	41.79	32.47	40.35	27.45	42.91
2	+ Adapter	38.65	44.57	44.17	42.80	43.19	33.75	41.52	28.66	44.07
3	HAN (Miculicich et al., 2018)	37.20	42.96	44.53	41.89	42.31	32.57	40.83	28.00	43.28
4	+ Adapter	38.11	43.62	45.99	43.51	43.03	33.91	42.47	29.49	45.10
5	SAN (Maruf et al., 2019)	37.40	43.28	44.82	41.99	42.60	32.46	41.01	28.19	43.54
6	+ Adapter	**39.62**	**45.37**	46.72	**43.91**	43.59	**34.48**	**42.93**	**30.01**	**45.38**
7	GCNMT (Chen et al., 2022)	38.39	44.33	46.43	42.92	43.60	33.41	41.51	28.73	44.08
8	+ Adapter	39.51	45.28	**47.26**	43.70	**44.56**	34.27	42.43	29.50	44.96
9	Transformer (Vaswani et al., 2017)	36.27	42.71	43.51	41.25	41.07	31.54	39.64	26.70	42.16

Table 2: Performance on ZH-EN test sets with and without the context-aware adapter is presented in this Table. The "+Adapter" indicates that our proposed context-aware adapter was used to guide the context-aware encoder. Significance testing (Koehn, 2004) demonstrates that the improvements achieved by our approach are statistically significant at the 0.05 level for almost all of the aforementioned models.

translator and adapter. The dropout rate is set to 0.2, while P% (the percentage of frozen context-aware encoder parameters) is set to 99%.

3. Training the Context-Aware Model: We utilize the corrupted instances to train the context-aware model, which follows the adapter-translator architecture, with the aim of familiarizing it with the document. This involves updating part of the parameters in the context-aware encoder(s). During adaptation, the learning rate is set to 0.1.

4. Document Translation: We employ the adapted model to translate the original document $\mathcal{X}$. This entails utilizing the updated parameters in the context encoder and the sentence encoder to encode the source sentences, and employing the translator decoder to decode the target sentences.

3 Application to various Document-level NMT Model

To evaluate the effectiveness of our proposed framework in context-aware NMT, we select the following five representative NMT models:

- DocT (Zhang et al., 2018): This model considers two previous sentences as context. It employs a document-aware transformer that incorporates context representations into both the sentence encoder and decoder.

- HAN (Miculicich et al., 2018): HAN leverages all previous source and target sentences as context and introduces a hierarchical attention network to capture structured and dynamic context. The context representations are then fed into the decoder.

- SAN (Maruf et al., 2019): SAN extends the context coverage to the entire document. It adopts sparse attention to selectively attend to relevant sentences and focuses on key words within those sentences.

- MCN (Zheng et al., 2020): MCN employs an encoder to generate local and global contexts from the entire document, enabling the model to understand inter-sentential dependencies and maximize the utilization of contextual information.

- GCNMT (Chen et al., 2022): GCNMT comprises a global context encoder, a sentence encoder, and a sentence decoder. It incorporates two types of global context to enhance translation performance.

All of these models utilize a context encoder to encode global or local contexts, thereby improving document-level translation performance. To apply our proposed adapter-translator architecture to these models, we introduce an adapter decoder.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Set	TED		News	
	#SubDoc	#Sent	#SubDoc	#Sent
Training	7,491	206,126	10,552	236,287
Dev	326	8,967	112	2,169
Test	87	2,271	184	2,999
Set	**Europarl**			
	#SubDoc	#Sent		
Training	132,721	1,666,904		
Dev	273	3,587		
Test	415	5,134		

Table 3: Statistics of the training, development, and test sets of the three translation tasks.

K	**BLEU**	**Meteor**
0	26.98	46.15
1	27.33	46.50
5	27.55	46.75
10	27.41	46.50
15	27.13	46.32

Table 4: Averaged performance with respect to different data expansion ratio in inferring stage.

4 Experimentation

4.1 Settings

Datasets and Evaluation Metrics. We conduct experiments on English-to-German (EN→DE) translation tasks in three different domains: talks, news, and speeches. Additionally, we evaluate our proposed framework for the Chinese-to-English translation task.

- TED: This dataset is obtained from the IWSLT 2017 MT track (Cettolo et al., 2012). We combine test2016 and test2017 as our test set, while the remaining data is used as the development set.
- News: This dataset is derived from the News Commentary v11 corpus. We use news-test2015 and news-test2016 as the development set and test set, respectively.
- Europarl: This dataset is extracted from the Europarl v7 corpus. We randomly split the corpus to obtain the training, development, and test sets.
- For ZH-EN: The training set consists of 41K documents with 780K sentence pairs.[2] We use the NIST MT 2006 dataset as the development set and the NIST MT 02, 03, 04, 05, and 08 datasets as the test sets. The Chinese sentences are segmented using Jieba, while the English sentences are tokenized and converted to lowercase using Moses scripts.

We obtained the three document-level translation datasets from (Maruf et al., 2019). [3] For the source-side English sentences, we segmented them using the corresponding BPE model trained on the training data. Meanwhile, for the target-side German sentences, we used the BPE model with 25K operations trained on the corresponding target-side data. Table 3 provides a summary of the statistics for the three translation tasks. It should be noted that we divided long documents into sub-documents containing at most 30 sentences to enable efficient training.

Model Settings. For all translation models, we have set the hidden size to 512 and the filter size to 2048. The number of heads in the multi-head attention mechanism is 8, and the dropout rate is 0.1. During the training phase, we train the models for 100K steps using four A100 GPUs, with a batch size of 40960 tokens. We employ the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a learning rate of 1, incorporating a warm-up step of 16K. As for the fine-tuning stage, we fine-tune

[2]It consists of LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, LDC2010T03.
[3]https://github.com/sameenmaruf/selective-attn/tree/master/data

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

the models for 40K steps on a single A100 GPU, with a batch size of 40960 tokens, a learning rate of 0.3, and a warm-up step of 4K. During the inference phase, we set the beam size to 5.

4.2 Experimental Results

We utilize two evaluation metrics, BLEU (Papineni et al., 2002) and Meteor (Lavie and Agarwal, 2007), to assess the quality of translation. The results, presented in Table 1, demonstrate that our proposed approach consistently achieves state-of-the-art performance, outperforming previous context-aware NMT models on average. We observe significant improvements across all datasets by adapting the NMT model to the characteristics of each input document. Of particular note is the comparison between models #9 and #10, where our approach demonstrates a notable improvement with a gain of +0.57 in BLEU and +0.60 in Meteor.

Table 2 showcases the performance results for Chinese-English translation. The table presents the BLEU scores for each sub-test set and the average Meteor score across all sets. The results demonstrate that our proposed adapter-translator framework consistently achieves state-of-the-art performance when compared to the original versions of previous context-aware NMT models. Moreover, we consistently observed improvements across all datasets by adapting the trained NMT model to fit each input document. For instance, comparing models #8 and #7, our approach achieves an improvement with a gain of +0.92 in BLEU, +0.77 in Meteor, and +0.88 in d-BLEU.

Effect of Hyper-Parameter K in Dynamic Translation In the inference stage, the expansion ratio is an important hyperparameter for dynamic translation. A low ratio may restrict the effectiveness of adaptation in parameter optimization, whereas a high ratio may lead to overfitting of the model to the document restoration task. As indicated in Table 4, we observe that the optimal performance is attained with a ratio of 5 for the EN-DE translation task using the GCNMT model.

5 Analysis and Discussion

In this section, we employ the Chinese-to-English translation task as a representative to offer additional evidence for the efficacy of our proposed framework. In addition to reporting s-BLEU, we also present case-insensitive document-level BLEU (d-BLEU) scores.

5.1 Effect of Adapting Task

In a previous study (Li et al., 2020), it was suggested that context encoders not only utilize context to guide models but also encode noise. Therefore, the improvement in translation quality can sometimes be attributed to enhanced model robustness. The authors discovered that two context-aware models exhibited superior performance during inference even when the context input was replaced with noise. To ascertain whether our framework genuinely benefits from the document adaptation task, we compare the experimental results with and without an adapter in a Chinese-to-English translation task.

We conducted an investigation on the impact of adapting a document prior to translation. We define **Fake adapting** as the process wherein nonsensical words are employed as the target output during the model's adaptation phase, and **Noisy adapting** as the process wherein the model employs shuffled noisy sentences as input and corrects portions of these sentences as output. The results in Table 5 demonstrate that our proposed framework achieves improvements of +3.12 and +1.67 compared to Fake adapting and Noisy adapting, respectively. Furthermore, a notable performance disparity is observed between the results of Fake adapting and Noisy adapting. The adapter that employs shuffled documents as input achieves a gain of +1.45 compared to Fake adapting, indicating that document adaptation indeed has a positive effect on the translation model.

5.2 Architecture of the Adapter

As elaborated in Section 2 on the **Adapter-Translator Architecture**, our proposed framework employs shared encoder(s) for both the adaptation process and translation process. It is worth noting that some previous context-aware models have utilized multiple encoders. To determine whether this architecture

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Context	s-BLEU	d-BLEU	Meteor
HAN (Miculicich et al., 2018)	40.83	43.28	28.00
Fake adapting	39.35	42.00	26.29
Noisy adapting	40.80	43.55	28.33
ours	**42.47**	**45.10**	**29.49**

Table 5: Performance on ZH-EN test sets of effectiveness of adapting process.

Share	s-BLEU	d-BLEU	Meteor
Sentence encoder	41.59	44.19	28.40
Context encoder	41.55	44.13	28.42
Both	**42.47**	**45.10**	**29.49**

Table 6: Performance on ZH-EN test sets of sharing the sentence encoder, the context encoder, or both.

is the optimal choice for our research objectives, we investigated the impact of the adapter architecture on the translation model's performance.

In our framework, the encoder(s) are shared between the adapter and translator; however, the effectiveness of each encoder remains uncertain. To explore this, we conducted experiments and present the results in Table 6. The table demonstrates that sharing either the sentence encoder, the context decoder, or both leads to significant improvements in translation performance. These findings align with our intuition, and we observe that sharing both encoders yields the best performance, as indicated in the first row of the table. A possible explanation for these results is that sharing both encoders maximizes the preservation and exchange of information acquired during the reconstruction process in the adaptation phase, specifically concerning the test document.

5.3 Designing of Adapting Task

Masked sentence auto-encoding tasks have been extensively utilized in natural language processing and have consistently shown their effectiveness and generalizability in numerous previous studies. In Table 7, we present the performance of various document adaptation tasks on the Chinese-to-English translation task. Interestingly, we observe a decline in performance when using the translation process itself as a document adaptation task, which aligns with findings from prior research on double-translation. Similarly, the experiment employing the reconstruction of typical masked sentences as an adaptation task also exhibited a similar phenomenon. These findings indicate that our proposed approach effectively assists translation models in capturing valuable information from documents.

5.4 Pronoun Translation

To evaluate coreference and anaphora, we adopt the reference-based metric proposed by Werlen and Belis (2017), following the methodology of Miculicich et al.(2018) and Tan et al.(2019). This metric measures the accuracy of pronoun translation. Table 8 displays the performance results. We observe that our proposed approach significantly improves the translation of pronouns, indicating that pronoun translation benefits from leveraging global context. This finding is consistent with the results reported in related studies (Werlen and Popescu-Belis, 2017; Miculicich et al., 2018; Tan et al., 2019).

5.5 Adapting with Human Feedback

Adapting with human feedback has been widely employed in various natural language models, and its effectiveness and generalization have been demonstrated in numerous prior studies. We sought to investigate whether human feedback could enhance our translator-adapter framework.

Table 9 presents the performance of the adapting task augmented with human feedback on the Chinese-to-English translation task. The term "**Fake feedback**" refers to using the adapter's outputs as simulated human feedback, while "**Real feedback**" denotes the process of reviewing and correcting the adapter's outputs, and using the corrected sequences as target sentences. From the results, we observe that using

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Task	s-BLEU	d-BLEU	Meteor
Translation	41.30	44.00	28.01
Masked sentences	41.98	44.60	28.33
Ours	**42.47**	**45.10**	**29.49**

Table 7: Performance of different document adapting task on ZH-EN translation task.

Model	Pronoun
Transformer	68.68
GCNMT (Chen et al., 2022)	68.77
+ adapter	68.95
SAN (Maruf et al., 2019)	69.37
+ adapter	**69.84**

Table 8: Evaluation on pronoun translations of ZH-EN.

the adapter's output as simulated human feedback leads to a decrease in performance. Additionally, employing human-corrected sentences as feedback incurs a doubling of the adaptation task cost, but only yields marginal improvements in translation performance. One possible assumption is that significant positive impact on translation quality can be achieved only when a substantial amount of high-quality human feedback data is available. Therefore, we did not integrate this method into our adapter-translator framework.

5.6 The Impact of Frozen Encoder Parameters Proportion

We performed preliminary experiments to examine the optimal proportion of frozen encoder parameters during the inference phase of the translator. The results in Table 10 demonstrate that the translator's performance steadily improved as we increased the proportion of frozen encoder parameters, reaching its peak at 99%. However, when we further increased the proportion to 99.5%, the translator's performance started to decline. Consequently, in our experiments, we set the proportion of frozen encoder parameters to 99% during the inference phase of the translator.

6 Related Work

Local context has been extensively investigated in neural machine translation (NMT) models, including both RNN-based RNNSearch and Transformer-based models (Bahdanau et al., 2015; Vaswani et al., 2017). An early attempt in RNN-based NMT was the concatenation method proposed by (Tiedemann and Scherrer, 2017). Subsequently, the adoption of multiple encoders emerged as a promising direction in both RNNSearch and Transformer-based NMT models (Jean et al., 2017; Wang et al., 2017; Zhang et al., 2018; Bawden et al., 2018; Voita et al., 2018; Voita et al., 2019b; Yang et al., 2020). Cache/memory-based approaches (Tu et al., 2018; Kuang et al., 2018; Maruf and Haffari, 2018) also fall under this category, as they utilize a cache to store word/translation information from previous sentences.

An alternative approach in document-level NMT treats the entire document as a unified translation unit and dynamically extracts pertinent global knowledge for each sentence within the document. This global context can be derived either from the source side (Maruf and Haffari, 2018; Mace and Servan, 2019; Maruf et al., 2019; Tan et al., 2019) or the target side (Xiong et al., 2019).

Moreover, several endeavors have been undertaken to enhance the performance of document-level translation through the utilization of monolingual document data. For instance, in order to improve translation coherence within a document, Voita et al.(2019a) propose DocRepair, which is trained on monolingual target language document corpora to address inconsistencies in sentence-level translations. Similarly, Yu et al.(2020) train a document-level language model to re-evaluate sentence-level translations. In contrast, Dowmunt(2019) harness monolingual source language document corpora to investigate multi-task training using the BERT-objective on the encoder.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Task	s-BLEU	d-BLEU	Meteor
Real feedback	42.64	45.37	29.60
Fake feedback	42.03	44.71	28.50
Ours	**42.47**	**45.10**	**29.49**

Table 9: Performance of human feedback augmented adapting task on ZH-EN translation task.

K	BLEU	Meteor
97.0%	23.57	43.46
98.0%	26.69	45.83
99.0%	27.55	46.75
99.5%	27.50	46.68
99.7%	27.46	46.60

Table 10: The impact of frozen encoder parameters proportion.

7 Conclusion

To enhance the alignment between the trained context-aware NMT model and each input document, we present in this paper an adapter-translator framework, designed to facilitate the model's familiarity with a document prior to translation. Our modification to the NMT model involves incorporating an adapter encoder, which reconstructs the intentionally corrupted portions of the original document. Empirical findings from Chinese-to-English translation tasks and various English-to-German translation tasks demonstrate the considerable performance improvement achieved by our approach compared to several robust baseline models.

Limitations

Our experimental findings and analysis validate the effectiveness of the proposed adapter-translator framework in facilitating model familiarity with documents prior to translation, thereby yielding substantial enhancements across multiple evaluation benchmarks. However, it should be noted that the inclusion of the adapter module may introduce a certain degree of computational overhead to the framework's efficiency. Nevertheless, it is widely recognized that the time-consuming aspect of machine translation during the inference stage primarily stems from the serial decoding process of beam search. In contrast, our approach, as described in this paper, does not employ beam search during the adaptation stage; instead, it leverages parallel attention and mask mechanisms that align with the training stage. The main increase in computational time for this approach arises from the storage of checkpoints after the completion of parameter updates during the adaptation stage.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of NAACL*, page 1304–1313.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*, pages 261–268.

Linqing Chen, Junhui Li, and Zhengxian Gong. 2020. Hierarchical global context augmented document-level neural machine translation. In *Proceedings of CCL*, pages 434–445.

Linqing Chen, Junhui Li, Zhengxian Gong, Min Zhang, and Guodong Zhou. 2022. One type context is not enough: Global context-aware neural machine translation. In *TALLIP*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, page 4171–4186.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of CVPR*, pages 16000–16009.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machinetranslation benefit from larger context? *Computing Research Repository*, arXiv:1704.05135.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of WMT*, page 225–233.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of EMNLP*, page 2242–2254.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, page 388–395.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of COLING*, page 596–606.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of WMT*, pages 228–231, June.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2018. One sentence one model for neural machine translation. In *Proceedings of LREC*, pages 910–917.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of ACL*, page 3505–3511.

Valentin Mace and Christophe Servan. 2019. Using whole document context in neural machine translation. In *Proceedings of IWSLT*.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of ACL*, pages 1275–1284.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of NAACL*, pages 3092–3102.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of EMNLP*, pages 2947–2954.

Kishore Papineni, Salim Roukos, Ward Todd, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Amane Sugiyama and Naoki Yoshinaga. 2021. Context-aware decoder for neural machine translation using a target-side document-level language model. In *Proceedings of NAACL*, pages 5781–5791.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of EMNLP-IJCNLP*, page 1576–1585.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, pages 1264–1274.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of EMNLP-IJCNLP*, pages 877–886.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of ACL*, pages 1198–1212.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of EMNLP*, page 2826–2831.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *Proceedings of Workshop on Discourse in Machine Translation*, pages 17–25.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of AAAI*, pages 7338–7345.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2020. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of EMNLP*, pages 1527–1537.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of EMNLP*, pages 533–542.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of IJCAI*, pages 3983–3989.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 665-676, Harbin, China, August 3 - 5, 2023.

TERL: Transformer Enhanced Reinforcement Learning for Relation Extraction

Yashen Wang
China Academy of Electronics and Information Technology of CETC, Artificial Intelligence Institute of CETC
yswang.arthur@gmail.com

Tuo Shi
Beijing Police College
shituo@bjpc.edu.cn

Xiaoye Ouyang
National Engineering Laboratory for Risk Perception and Prevention (RPP), China Academy of Electronics and Information Technology
ouyangxiaoye@cetc.com.cn

Dayu Guo *
CETC Academy of Electronics and Information Technology Group Co.,Ltd.
guodayu1@cetc.com.cn

Abstract

Relation Extraction (RE) task aims to discover the semantic relation that holds between two entities and contributes to many applications such as knowledge graph construction and completion. Reinforcement Learning (RL) has been widely used for RE task and achieved SOTA results, which are mainly designed with rewards to choose the optimal actions during the training procedure, to improve RE's performance, especially for low-resource conditions. Recent work has shown that offline or online RL can be flexibly formulated as a sequence understanding problem and solved via approaches similar to large-scale pre-training language modeling. To strengthen the ability for understanding the semantic signals interactions among the given text sequence, this paper leverages Transformer architecture for RL-based RE methods, and proposes a generic framework called **T**ransformer **E**nhanced **RL** (**TERL**) towards RE task. Unlike prior RL-based RE approaches that usually fit value functions or compute policy gradients, TERL only outputs the best actions by utilizing a masked Transformer. Experimental results show that the proposed TERL framework can improve many state-of-the-art RL-based RE methods.

1 Introduction

Relation Extraction (RE) aims to discover the binary semantic relation between two entities in a sequence of words. E.g., given a sentence "$\cdots$ Carey will succeed Cathleen P. Black, who held the position for 15 years and will take on a new role as chairwoman of Hearst Magazines, the company said$\cdots$" (Xue et al., 2020), and we aim to predict the relation type between two entities "Cathleen P. Black" and "chairwoman" and the result is "per:title".

Deep neural network (DNN) driven methods have gained decent performance when labeled data is available (Hu et al., 2021b; Guo et al., 2020). While Reinforcement Learning (RL) based RE methods gain a lot of attention recently and show encouraging effects (Takanobu et al., 2018; Hu et al., 2021b; Wang and Zhang, 2021), especially in low-resource and few-shot conditions. Since this kinds of work requires *fewer* labeled data or could expand limited labeled data by exploiting information on unlabeled data to iteratively improve the performance (Hu et al., 2021b).

Recent works have shown Transformers (Vaswani et al., 2017) can model high-dimensional distributions of semantic concepts at scale, and several attempts have demonstrated the combination between transformers and RL architecture (Parisotto and Salakhutdinov, 2021; Parisotto et al., 2020; Zambaldi et al., 2019). These works have shown that the Transformer's efficiency for modeling beneficial semantic interactions in the given sequence (Chen et al., 2021a; Zheng et al., 2022), which is very enlightening for RE task. Given the diversity of successful applications of such models (Chen et al., 2021a), this paper seeks to investigate their application to sequential RE problems formalized as RL, because of the three main advantages of transformers: (i) Its ability to model long sequences has been demonstrated in many tasks; (ii) It could perform long-term *credit assignment* via self-attention strategy, contrary to Bellman backups (Lee et al., 2021) which slowly propagate rewards and are prone to distractor signals (Hung et al., 2019) in Q-learning, which could enable Transformer-based architecture to still work effectively in the presence of distracting rewards (Chen et al., 2021a); and (iii) It can model a wide distribution of behaviors, enabling better generalization (Ramesh et al., 2021). Hence, inspired by (Chen et al., 2021a;

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

Zheng et al., 2022), we try to view the RL-based RE as a conditional sequence understanding problem. Especially, we model the joint distribution of the sequence of states, actions and rewards, and discuss whether generative sequence understanding could can serve as a substitute for traditional RL algorithms in RE task. Overall, we propose **T**ransformer **E**nhanced **R**einforcement **L**earning (**TERL**), which abstracts RL paradigm as autoregressively sequence understanding and utilize Transformer architecture in BERT [1] to model text sequences with minimal modification to native transformer's architecture, and we investigate whether the sequence understanding paradigm can perform policy optimization by evaluating TERL on RL benchmarks in RE task. This enables us to leverage the scalability of the Transformer's architecture, as well as the related advancements in pre-training language modeling (such as the BERT's series).

Especially, following the backbone proposed in (Chen et al., 2021a), we train Transformer architecture on collected experience with a sequence understanding objective for RE task, instead of training a policy through conventional RL algorithms (Hu et al., 2021b; Wang and Zhang, 2021). This transformer is trained to predict next token in a sequence of rewards (forward-cumulative-rewards emphasized here), states, and actions. This paper shows that leveraging Transformers can open up another paradigm to solve RL-based RE problem. The main differences between this work and previous RL-based RE methods, can be concluded as follows: (i) RL is transformed into sequence understanding; (ii) We learn the natural projection from reward and state to action, instead of maximizing cumulative discount rewards or *only* modeling state and action in conventional behavior cloning paradigm (Chen et al., 2021b); (iii) Q/V-functions are *no* need to be learned, while we directly model it as a sequence problem, wherein as long as given the expected return, we can get the corresponding action; and (iv) Bellman backups or other temporal difference frameworks is *no* need; In RE tasks (even relation and entity joint extraction tasks) with our work, the expected target return is highly correlated with the actual observed return. Under certain conditions, the proposed TERL could successfully generate sequences that almost completely match the required returns. In addition, we can prompt TERL with a higher return than the maximum event available in the dataset, indicating that our TERL can sometimes be extrapolated. Moreover, the proposed framework can also be used as a plug-in unit for any RL-based RE architecture, and be extended to relation and entity joint extraction task (Zhou et al., 2019). Experimental results show that the proposed TERL framework can improve many state-of-the-art RL-based RE methods.

2 Related Work

Relation Extraction (RE) aims to predict the binary relation between two entities in a sequence of words. Recent work leverages deep neural network (DNN) for learning the features among two entities from sentences, and then classify these features into pre-defined relation types (Hu et al., 2021b). These methods have achieved satisfactory performance when labeled data is sufficient (Zeng et al., 2015; Guo et al., 2020), however, it's labor-intensive to obtain large amounts of manual annotations on corpus. Hence, few-shot (even zero-shot) RE methods gained a lot of attention recently, since these methods require *fewer* labeled data and could expand limited labeled information by exploiting information on unlabeled data to iteratively improve the performance. Wherein, Reinforcement Learning (RL) based methods have grown rapidly (Zeng et al., 2019; Wang and Zhang, 2021), which has been widely used in Nature Language Processing (NLP) (Narasimhan et al., 2016; Zhou et al., 2019; Li et al., 2021). These methods are all designed with rewards to force the correct actions to be chosen during the model's training procedure. For RE task, (Qin et al., 2018) proposes a RL strategy to generate the false-positive indicator, where it automatically recognizes false positives for each relation type without any supervised information. (Li et al., 2021) addresses the RE task by capturing rich contextual dependencies based on the attention mechanism, and using distributional RL to generate optimal relation information representation. (Hu et al., 2021b) proposes gradient imitation RL method to encourage pseudo label data to imitate the gradient descent direction on labeled data.For relation and entity joint extraction task, (Takanobu et al., 2018) proposes a hierarchical RL framework which decomposes the whole extraction process into a hierarchy of two-level RL policies for relation extraction and entity extraction, respectively. (Zeng et al., 2019) applies policy gradient method to model future reward in a joint entity and relation extraction task. (Wang

[1] Other transformer architecture is also applicable.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

and Zhang, 2021) jointly extracts entities and relations, and propose a novel bidirectional interaction RL model.

Recently, there exist many exciting works which formulate the Reinforcement Learning (RL) problem as a context-conditioned "sequence understanding" problem (Chen et al., 2021a; Zheng et al., 2022). For *offline* RL settings, (Chen et al., 2021a) trains a transformer (Vaswani et al., 2017) as a model-free context-conditioned policy, and (Janner et al., 2021) trains a transformer as both a policy and model and shows that beam search can be used to improve upon purely model-free performance. These works focus on exploring *fixed* datasets that transformers are traditionally trained with in NLP applications, which is similar to our focus. For *online* RL settings, (Zheng et al., 2022) proposes a RL algorithm based on sequence understanding that blends offline pre-training with online fine-tuning in a unified framework. To best of our knowledge, this work is the *first* test to leverage Transformer for enhancing RL-based RE task.

3 Methodology

This section presents the proposed TERL for RE task, as summarized in Fig. 1.

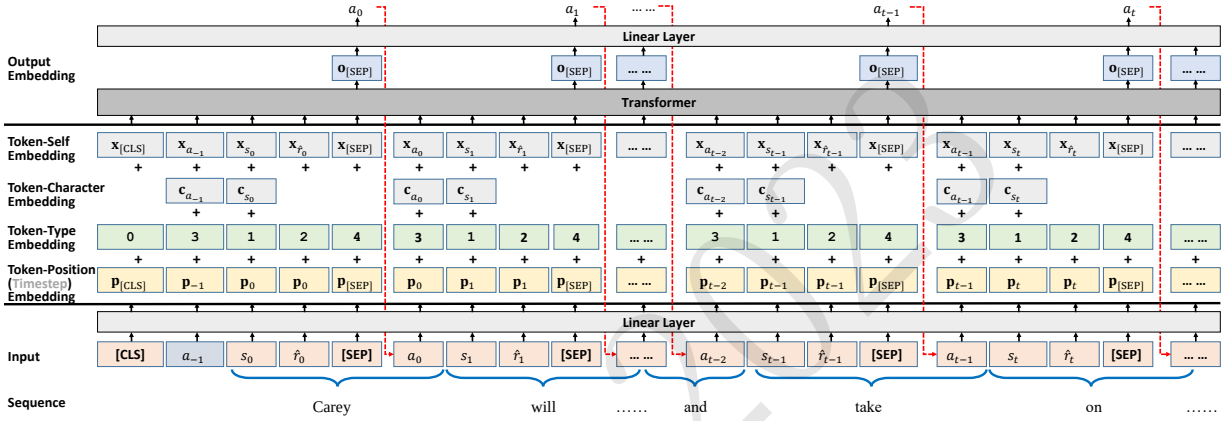


Figure 1: The architecture of TERL for RE task.

3.1 Relation Extraction with RL

The RL policy π for Relation Extraction (RE), usually aims to detect the relations in the given word sequence $\tau_1 = \{w_0, w_1, w_2, \cdots, w_T\}$, which can be regarded as a conventional RL policy over actions. A Markov Decision Process (MDP) described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ (Wang and Zhang, 2021), is usually used for learning procedure. Especially, the MDP tuple consists of states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$, transition probability $P(s'|s, a)$ and rewards $r \in \mathcal{R}$. At timestep t, s_t, a_t, and $r_t = \mathcal{R}(s_t, a_t)$ denote the state, action, and reward, respectively. The goal in RL is to learn a desired policy which maximizes the expected reward $\mathbb{E}(\sum_{t=1}^{T} r_i)$ in MDP (Chen et al., 2021a).

Action: The action a_t is selected from $\mathcal{A} = R \bigcup \texttt{None}$, wherein notation `None` indicates that *no* relation exists in the given context, and R is the pre-defined relation-type set.

State: The state $s_t \in \mathcal{S}$ of the relation extraction RL process at timestep t, can be represented by (Wang and Zhang, 2021; Takanobu et al., 2019): (i) the current hidden state vector $\mathbf{h}_t$, (ii) the relation-type vector $\mathbf{a}_{t-1}$ (the embedding of the latest action a_{t-1} that $a_{t-1} \neq \texttt{None}$, a learnable parameter), and (iii) the state from the last timestep $\mathbf{s}_{t-1}$, formally represented as follows:

$$\mathbf{s}_t = f(\mathbf{W}_{\mathcal{S}}[\mathbf{h}_t; \mathbf{a}_{t-1}; \mathbf{s}_{t-1}]) \tag{1}$$

where $f(\cdot)$ denotes a non-linear function implemented by MLP (Other encoder architecture is also applicable, which is not the focus of this paper). To obtain the current hidden state $\mathbf{h}_t$, sequence Bi-LSTM over the current input word embedding $\mathbf{x}_t$, character embedding $\mathbf{c}_t$, token-type embedding $\mathbf{v}_t$, and

token-position embedding $\mathbf{p}_t$, can be used here, as follows:

$$\begin{aligned}\overrightarrow{\mathbf{h}_t} &= \overrightarrow{\texttt{LSTM}}(\overrightarrow{\mathbf{h}_{t-1}}, \mathbf{x}_t, \mathbf{c}_t, \mathbf{v}_t, \mathbf{p}_t) \\ \overleftarrow{\mathbf{h}_t} &= \overleftarrow{\texttt{LSTM}}(\overleftarrow{\mathbf{h}_{t+1}}, \mathbf{x}_t, \mathbf{c}_t, \mathbf{v}_t, \mathbf{p}_t) \\ \mathbf{h}_t &= [\overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_t}]\end{aligned} \tag{2}$$

Policy: The stochastic policy for detecting relation-type can be defined as $\pi : \mathcal{S} \rightarrow \mathcal{A}$, which specifies a probability distribution over actions:

$$a_t \sim \pi(a_t|s_t) = \text{SoftMax}(\mathbf{W}_\pi \mathbf{s}_t) \tag{3}$$

Reward: The environment provides intermediate reward r_t to estimate the future return when chose action a_t. The reward is computed as follows:

$$r_t = \begin{cases} 1, & a_t \text{ conforms to } \tau_1, \\ 0, & a_t = \texttt{None}, \\ \text{-1}, & a_t \text{ not conforms to } \tau_1. \end{cases} \tag{4}$$

If a_t equals to `None` at certain timestep t, the agent transfers to a new relation extraction state at the next timestep $t+1$. Such a MDP procedure mentioned above continues until the *last* action about the *last* word w_T of current sequence is sampled. Finally, a final reward r_* is obtained to measure the RE's performance that the RL's policy π detects, which is obtained by the weighted harmonic mean of precision and recall in terms of the relations in given sentence sequence τ_1 (Wang and Zhang, 2021): $r_* = \frac{(1+\beta^2)\cdot Prec \cdot Rec}{\beta^2 \cdot Prec + Rec}$. Wherein, notation $Prec$ and Rec indicate the precision value and recall value respectively, computed over the current sequence τ_1.

3.2 Transformer

For simplicity, we take BERT as an example. BERT (Devlin et al., 2019) is the first bidirectional language model, which makes use of left and right word contexts simultaneously to predict word tokens. It is trained by optimizing Masked Language Model (MLM) objective etc.,. The the architecture of conventional BERT is a multi-layer bidirectional transformer encoder (Vaswani et al., 2017), and the inputs are a sequence of tokens $\{x_1, x_2, \cdots, x_n\}$. The tokens go through several layers of *transformers*. At each layer, a new contextualized embedding is generated for each token by weighted-summing all other tokens' embeddings. The weights are decided by several attention matrices (*multi-head attention*). Note that: (i) tokens with *stronger* attentions are considered *more* related to the target word; (ii) *Different* attention matrices capture *different* types of token relations, such as exact match and synonyms.

The entire BERT model is pre-trained on large scale text corpora and learns linguistic patterns in language. It can be viewed as an interaction-based neural ranking model (Guo et al., 2016). Given the widespread usage of BERT, we do not detail the architecture here. See (Devlin et al., 2019) for more details about the conventional architecture of BERT and its variants for various applications.

3.3 Input Generation

Given sequence under RL's paradigm $\{s_0, r_0, a_0, s_1, r_1, a_1, \cdots, s_T, r_T, a_T\}$, the reward of a sequence at step t, is defined as the forward-cumulative-rewards from the current timestep, similar to (Chen et al., 2021a): $\hat{r}_t = \sum_{i=t}^{T} r_i$, without discount. Wherein, r_i denotes the reward from environment at timestep i. Because we want to generate actions based on *future* (forward direction) expected returns rather than *past* (backward direction) rewards. Hence the input sequence towards our Transformer, is defined as follows, which consists of states, actions and rewards:

$$\tau = \{a_{-1}, s_0, \hat{r}_0, a_0, s_1, \hat{r}_1, a_1, s_2, \hat{r}_2, a_2, \cdots, s_T, \hat{r}_T, a_T\} \tag{5}$$

It represents the whole sequence from the beginning to the end, but in the actual training process, we often only intercept K timesteps (i.e., context length) as input (details in Sec. 3.4). Wherein, K is a hyper-parameter with different values towards different tasks, and a_{-1} in E.q. (5) is a padding indicator.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

3.4 Procedure

We feed the last K timesteps into TERL, for a total of $3 \times K$ tokens (one for each type: states, actions and rewards). As shown in Fig. 1, to obtain token embeddings: (i) for state and action, E.q. (1) and E.q. (2) are used to generate state embedding and action embedding, which consider word embeddings, character embeddings, type embeddings and position embeddings (Wang and Zhang, 2021; Zhou et al., 2019); (ii) for forward-cumulative-rewards, we learn a linear-layer, which projects inputs to the embedding dimension, followed by layer-normalization (Chen et al., 2021a; Xiong et al., 2020).

Moreover, a token-position (respect to timestep) embedding, a token-type embedding for each token as well as a token-character embedding respect to action token or state token, is learned and added to each token, as one timestep corresponds to 4 types of tokens in our framework. Wherein, we define the token-type projection as: $\{[\mathrm{CLS}], action, state, reward, [\mathrm{SEP}]\} \rightarrow \{0, 1, 2, 3, 4\}$. The tokens are then processed by a BERT (Devlin et al., 2019) or GPT (Radford and Narasimhan, 2018) model (as well as their variants), which predicts future (forward) action: $\{a_{t-1}, s_t, \hat{r}_t\} \rightarrow a_t$.

With efforts above, after executing the generated actions for the current state, we reduce the target return by the rewards we receive and repeat until the end of the episode. The output is action sequence $\{a_0, a_1, a_2, \cdots, a_T\}$, which is generated with a linear layer (on top of Fig. 1). Note that, the output can also includes sequence of states or rewards. For simplicity, we do not use them and leave for future discussion.

The details about training procedure and testing procedure, can be concluded as follows:

(i) In training procedure, we sample mini-batches of sequence length K (i.e., context length) from the training dataset, and mainly use the self-attention paradigm in Transformer. a_{-1} with zero-padding is added before the entire sequence. As shown in Fig. 1, predicting action at each timesetp a_t with cross-entropy loss, is used as the training objective.

(ii) At test time, we use the definition of E.q. (4) as the desired performance. At the beginning, given the desired performance (e.g., $\hat{r}_0 = 1$) as well as the initial state s_0, transformer generates action a_0. Let the agent perform actions a_0, the environment will give return r_0 and the next state s_0, and we can get $\hat{r}_1$. Then $\{a_0, s_1, \hat{r}_1\}$ can be added into the input sequence, and we can get a_1. The aforementioned testing procedure is *autoregressive*, because the output a_{t-1} in previous timestep will intuitively be the viewed as input in the following timestep: $\{a_{t-1}, s_t, \hat{r}_t\} \rightarrow a_t$.

4 Experiments

This paper constructs relation extraction task and relation and entity joint extraction task for evaluations.

4.1 Datasets and Metrics

For relation extraction (RE) task examination, we follow (Hu et al., 2021b) to leverage two public RE datasets for conducting experiments on, including SemEval 2010 Task 8 (SemEval) (Hendrickx et al., 2010), and TAC Relation Extraction Dataset (TACRED) (Zhang et al., 2017): (i) SemEval dataset is a standard benchmark dataset for testing RE models, which consists of training, validation and test set with 7,199, 800, 1,864 relation-mentions respectively, with totally 19 relations types (including `None`). (ii) TACRED dataset is a more large-scale crowd-sourced RE dataset, which is collected from all the prior TAC KBP relation schema. It consists of training, validation and test set with 75,049, 25,763, 18,659 relation-mentions respectively, with totally 42 relation types (including `None`).

We also test the extension of the proposed framework for relation and entity joint extraction task. For this task, we conduct experiments on two public datasets NYT (Riedel et al., 2010) and WebNLG (Gardent et al., 2017): (i) NYT dataset is originally produced by a distant supervision method, which consists of 1.18M sentences with 24 predefined relation types; (ii) WebNLG dataset is created by Natural Language Generation (NLG) tasks and adapted by (Zeng et al., 2018) for relational triple extraction task. It consists of 246 predefined relation types.

For both datasets, we follow the evaluation setting used in previous works. A triple (head entity, relation-type, tail entity) is regarded as *correct* if the relation-type (belongs to R) and the two corresponding entities (head entity and tail entity head entity) are all correct. Precision, Recall and F1-score are introduced here

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

as metrics for all the compared models. For each dataset, we randomly chose 0.5% data from the training set for validation (Wang and Zhang, 2021).

4.2 Baselines

For relation extraction task, the baselines include three categories:

(i) When comparing with supervised relation encoders with only labeled data, we choose **LSTM**(Hochreiter and Schmidhuber, 1997), **PCNN**(Zeng et al., 2015), **PRNN**(Zhang et al., 2017), and **BERT**(Devlin et al., 2019) as baselines.

(ii) When comparing with semi-supervised relation encoders with both labeled data and unlabeled (or pseudo labeled) data, we choose **Self-Training**(Rosenberg et al., 2005), **Mean-Teacher**(Tarvainen and Valpola, 2017), **DualRE**(Lin et al., 2019), and **MetaSRE**(Hu et al., 2021a) as baselines.

(iv) When comparing with the RL-based models, we choose **RDSRE**(Qin et al., 2018), **DAGCN**(Li et al., 2021) and **GradLRE**(Hu et al., 2021b) as baselines. **RDSRE** is a RL strategy to generate the false-positive indicator, where it automatically recognizes false positives for each relation type without any supervised information. **DAGCN** addresses the RE task by capturing rich contextual dependencies based on the attention mechanism, and using distributional RL to generate optimal relation information representation. **GradLRE** is gradient imitation RL method to encourage pseudo-label data to imitate the gradient descent direction on labeled data and bootstrap its optimization capability through trial and error. As our work can be viewed as a plug-in unit for this kind of RL-based model, the variant model with help of our work is named with suffix "**+TERL**".

Our framework can be easily extended to relation and entity joint extraction method based on RL. For evaluating joint extraction task, the baselines include four categories:

(i) The traditional pipeline models are **FCM** (Kim, 2014) and **LINE** (Gormley et al., 2015). Wherein, **FCM** is a conventional and compositional joint model by combining hand-crafted features with learned word embedding for relation extraction task. **LINE** is a network embedding method which embeds very large information networks into low-dimensional vectors. Note that, following (Wang and Zhang, 2021), both of them obtain the NER results by CoType (Ren et al., 2017), and then the results are fed into the two models to predict relations.

(ii) The joint learning baselines used here include feature-based methods (e.g., **DS-Joint** (Yu and Lam, 2010), **MultiR** (Hoffmann et al., 2011) and **CoType** (Ren et al., 2017)), and neural-based methods (e.g., **SPTree** (Li and Ji, 2014) and **CopyR** (Zeng et al., 2018)). Wherein, **DS-Joint** is an incremental joint framework extracting entities and relations based on structured perceptron and beam-search. **MultiR** is a joint extracting approach for multi-instance learning with overlapping relation types. **CoType** extracts entities and relations by jointly embedding entity mentions, relation mentions, text features, and type labels into two meaningful representations. **SPTree** is a joint learning model that represents both word sequence and dependency tree structures using bidirectional sequential and tree-structured LSTM-RNNs. **CopyR** is a sequence-to-sequence learning framework with a copy mechanism for relation and entity jointly extracting.

(iii) The tagging mechanism based models include **Tagging-BiLSTM** (Zheng et al., 2017) and **Tagging-Graph** (Wang et al., 2018). Wherein, **Tagging-BiLSTM** utilizes a Bi-LSTM-based architecture to capture the context representation of the input sentences through and then uses an LSTM network to decode the tag sequences. **Tagging-Graph** converts the joint extraction task into a directed graph by designing a novel graph scheme.

(iv) RL-based joint extraction models include **HRL**(Takanobu et al., 2018), **JRL**(Zhou et al., 2019), **Seq2SeqRL**(Zeng et al., 2019) and **BIRL**(Wang and Zhang, 2021). Wherein, **HRL** presents a hierarchical RL framework decomposing the whole joint extraction process into a hierarchy of two-level RL policies for relation extraction and entity extraction, respectively. **JRL** consists of two

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

components, including a joint network and a RL agent (which refines the training dataset for anti-noise). **Seq2SeqRL** applies RL strategy into a sequence-to-sequence model to take the extraction order into consideration. **BIRL** proposes a novel bidirectional interaction RL model for jointly extracting entities and relations with both inter-attention and intra-attention.

4.3 Experimental Settings

For a fair comparison, we build our **TERL** implementation for RE and joint extraction task with BERT (Devlin et al., 2019), as BERT-based work has achieved the state-of-the-art performance in RE task. Besides, we adopt BERT as the base encoder for both our **TERL** and other RL-based baselines for a fair comparison. Although GPT is also tested, the experimental trend is consistent. All hyper-parameters are tuned on the validation set. The word vectors are initialized using Word2Vec vectors and are updated during training. DQN encoder (Mnih et al., 2015) with an additional linear layer is introduced here for projecting to the embedding dimension. The main list of hyper-parameters is concluded as follows: Number of layers is 6; Number of attention heads is 8; Embedding dimensionality is 256; Batch size is 512; Context length $K = 30$; Max epochs is 5; Dropout is 0.1; Learning rate is 10^{-4}.

4.4 Performance Summary

F1 results with various labeled data on Relation Extraction (RE) task, are shown in Table 1. Average results over 20 runs are reported, and the best performance is bold-typed. As our work can be viewed as a plug-in unit for RL-based model, the variant model with help of our work is named with suffix "**+TERL**". RL-based methods outperforms all baseline models consistently. We could observe that **+TERL** improve all the RL-based methods. More specifically, compared with the previous SOTA model **GradLRE**, which defeats other models across various labeled data, **+TERL** is also more robust than all the baselines. Considering low-resource RE when labeled data is very scarce, e.g. 5% for SemEval and 3% for TACRED, the improvement from **+TERL** is significant: **+TERL** could achieve an average 3.15% F1 boost compared with **GradLRE**. Moreover, the improvement is still robust when more labeled data can be used (i.e., 30% for SemEval and 15% for TACRED), and the average F1 improvement is 1.15%. Especially, **RDSRE** fall behinds **DualRE** in most cases, while it outperforms **DualRE** when plugged with our **TERL** (i.e., **RDSRE+TERL**). This because the attention mechanism gives our **TERL** an excellent ability of long-term credit assignment, which can capture the effect of actions on rewards in a long sequence. We believe this phenomenon is meaningful and important for document-level RE task. Moreover, a key difference between our TERL and previous RL-based RE SOTA methods, can be concluded that this work dos *not* require policy regularization or conservatism to achieve optimal performance, which is consistent with the observation in (Chen et al., 2021a) and (Zheng et al., 2022). Especially, our speculation is that an algorithm based on time difference learning paradigm learns an approximation function and improves the strategy by optimizing the value function.

Relation and entity joint extraction is a more challenging task, and the proposed Transformer enhanced RL framework can be easily extend to this task. The experimental results on NYT and WebNLG datasets are shown in Table 2. It can be concluded that, the proposed model consistently outperforms all previous SOTA models in most cases, especially RL-base methods. Especially, RL-based methods usually defeats encoder-decoder based methods. E.g., RL-based **HRL** and **JRL** significantly surpass **Tagging-BiLSTM** and **CopyR**. Compared with **HRL**, **JRL** and **BIRL**, the their **+TERL**'s variants improve the F1 score by 3.94%, 3.66% and 4.22% on WebNLG dataset, respectively. This phenomenon shows that, our **TERL**-based variant matches or exceeds the performance of SOTA model-free RL algorithms, even without using dynamic programming. Note that, the behavior of optimizing the learning function in previous work, may unfortunately exacerbate and exploit any inaccuracies in the approximation of the value function, leading to the failure of policy improvement. Due to the fact that the proposed TERL does *not* require explicit optimization with learning functions as the objective, it *avoids* the need for regularization or conservatism, to a certain degree. Moreover, when we represent the distribution of policies, just like sequence understanding, context allows the converter to identify which policies generate actions, thereby achieving better learning and improving training dynamics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

Table 1: Performance comparisons on Relation Extraction (RE) task (F1).

Model	SemEval			TACRED		
	5%↑	10%↑	30%↑	3%↑	10%↑	15%↑
LSTM(Hochreiter and Schmidhuber, 1997)	0.226	0.329	0.639	0.287	0.468	0.494
PCNN(Zeng et al., 2015)	0.418	0.513	0.637	0.400	0.504	0.525
PRNN(Zhang et al., 2017)	0.553	0.626	0.690	0.391	0.522	0.546
BERT(Devlin et al., 2019)	0.707	0.719	0.786	0.401	0.532	0.556
Self-Training(Rosenberg et al., 2005)	0.713	0.743	0.817	0.421	0.542	0.565
Mean-Teacher(Tarvainen and Valpola, 2017)	0.701	0.734	0.806	0.443	0.531	0.538
DualRE(Lin et al., 2019)	0.744	0.771	0.829	0.431	0.560	0.580
MetaSRE(Hu et al., 2021a)	0.783	0.801	0.848	0.462	0.570	0.589
RDSRE(Qin et al., 2018)	0.729	0.756	0.812	0.422	0.549	0.568
RDSRE+TERL(Ours)	0.787	0.801	0.853	0.435	0.560	0.574
DAGCN(Li et al., 2021)	0.781	0.801	0.838	0.464	0.570	0.587
DAGCN+TERL(Ours)	0.804	0.817	0.846	0.478	0.582	0.593
GradLRE(Hu et al., 2021b)	0.797	0.817	0.855	0.474	0.582	0.599
GradLRE+TERL(Ours)	**0.820**	**0.833**	**0.864**	**0.488**	**0.594**	**0.605**

4.5 Analysis and Discussion

This section investigates whether our TERL variant can remain robust performance on metric of *imitation learning* (like **GradLRE** etc.,) on a *subset* of the dataset. Hence, we adopt baseline **GradLER** which is based on imitation learning, by following the experimental setting of Percentile Behavior Cloning strategy proposed by (Chen et al., 2021a), wherein we run behavior cloning on *only* the top $X\%$ of timesteps in the corresponding dataset, following (Chen et al., 2021a). The Percentile Behavior Cloning variant of **GradLER** is denoted as %**GradLER** here in Table 3. The percentile $X\%$ interpolates between standard behavior cloning ($X = 100\%$) that trains on the complete dataset and only cloning the best observed sequence ($X \approx 0\%$), which in a manner trades off between better generalization by training on more data with training a specialized model that focuses on a subset of the dataset. Table 3 shows experimental results comparing %**GradLRE** to **+TERL**, when the value of X are chosen in $\{10\%, 30\%, 50\%, 100\%\}$. From the experimental results, we conclude that, lower X reduces the performances of **GradLRE**, however **+TERL** successfully exceeds the performance and pulls F1 metric back. Especially, when X is 30, with enhancement from our **TERL**, 30%**GradLRE+TERL** could even defeats 50%**GradLRE**, while 30%**GradLRE** lags behind 50%**GradLRE** obviously. Moreover, 50%**GradLRE+TERL** nearly matches the performance of 100%**GradLRE**. This phenomenon indicates that, the improvement of our **TERL** can be made to the specific subset, after training the distribution of the complete dataset.

Then, to evaluate the importance of accessing previous states, actions, and returns, we discuss the context length K. This is interesting because when using frame stacking, it is usually assumed that the previous state is sufficient for the RL algorithm. Fig. 2 and Fig. 3 is evaluated on RE task (with TACRED dataset and 15% labeled data) and joint extraction task (with WebNLG dataset), respectively. TERL with different K is loaded into baselines **RDSRE**, **DAGCN** and **GradLRE**, as well as baselines **JRL**, **Seq2SeqRL** and **BIRL**. Experimental results show that performance of TERL is significantly worse when K is small (i.e., $K = 1$ or $K = 5$), indicating that past information (i.e., previous states s_t, actions a_t, and returns$\hat{r}_t$) is useful for RE task. Especially, when K becomes small, the performances have fallen off a cliff, even falling behind the original with side effect. Note that, the proposed framework still match the MDP properties when $K = 1$, while the results is worse, which demonstrates the sequence understanding is *highly* context dependent. When $K = 20$ and $K = 30$, **+TERL** defeats the corresponding original comparative baseline and the performances have changed little when K becomes larger. Besides, the context information (i.e., larger K) enables the transformer to figure out which actions are generated, which can lead to higher returns.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

Table 2: Performance comparisons on relation and entity joint extraction task (Precision, Recall, and F1).

Model	NYT			WebNLG		
	Precision↑	Recall↑	F1↑	Precision↑	Recall↑	F1↑
FCM(Kim, 2014)	0.561	0.118	0.193	0.472	0.072	0.124
LINE(Gormley et al., 2015)	0.340	0.251	0.277	0.286	0.153	0.193
MultiR(Hoffmann et al., 2011)	0.344	0.250	0.278	0.289	0.152	0.193
DS-Joint(Yu and Lam, 2010)	0.572	0.201	0.291	0.490	0.119	0.189
CoType(Ren et al., 2017)	0.521	0.196	0.278	0.423	0.175	0.241
SPTree(Li and Ji, 2014)	0.492	0.557	0.496	0.414	0.339	0.357
CopyR (Zeng et al., 2018)	0.569	0.452	0.483	0.479	0.275	0.338
Tagging-BiLSTM(Zheng et al., 2017)	0.624	0.317	0.408	0.525	0.193	0.276
Tagging-Graph(Wang et al., 2018)	0.628	1.632	0.844	0.528	0.194	0.277
HRL(Takanobu et al., 2018)	0.714	0.586	0.616	0.601	0.357	0.432
HRL+TERL(Ours)	0.750	0.604	0.641	0.631	0.368	0.449
JRL(Zhou et al., 2019)	0.691	0.549	0.612	0.581	0.334	0.410
JRL+TERL(Ours)	0.712	0.582	0.613	0.610	0.344	0.425
Seq2SeqRL(Zeng et al., 2019)	0.779	0.672	0.690	0.633	0.599	0.587
Seq2SeqRL+TERL(Ours)	**0.802**	0.692	0.711	0.665	0.617	0.611
BIRL(Wang and Zhang, 2021)	0.756	0.706	0.697	0.660	0.636	0.617
BIRL+TERL(Ours)	0.794	**0.727**	**0.725**	**0.693**	**0.655**	**0.643**

Table 3: Performance comparisons on Percentile Behavior Cloning (F1).

Model	TACRED		
	3%↑	10%↑	15%↑
10%**GradLRE**	0.190	0.233	0.240
10%**GradLRE+TERL**	0.342	0.416	0.424
30%**GradLRE**	0.356	0.437	0.449
30%**GradLRE+TERL**	**0.410**	**0.499**	**0.508**
50%**GradLRE**	0.379	0.466	0.479
50%**GradLRE+TERL**	0.464	0.564	0.575
100%**GradLRE**	0.474	0.582	0.599
100%**GradLRE+TERL**	0.488	0.594	0.605

5 Conclusion

In this work, we try to combine transformers and Reinforcement Learning (RL) based sequence relation extraction (RE), and extend Transformer paradigm to RL. We design a novel framework (TERL) that abstracts RL-based RE as a sequence understanding task, which could leverage the simplicity and scalability of the Transformer-based architecture for understanding textual sequence, as well as the advancements released by pre-training language modeling (such as the BERT/GPT series). Moreover, the proposed framework can also be used as a plug-in unit for any RL-based RE architecture, and be extended to relation and entity joint extraction task. Experimental results show that the proposed TERL framework can improve many state-of-the-art RL-based RE methods.

Acknowledgements

We thank anonymous reviewers for valuable comments. This work is funded by: the National Natural Science Foundation of China (No. U19B2026, 62106243, U22B2601).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

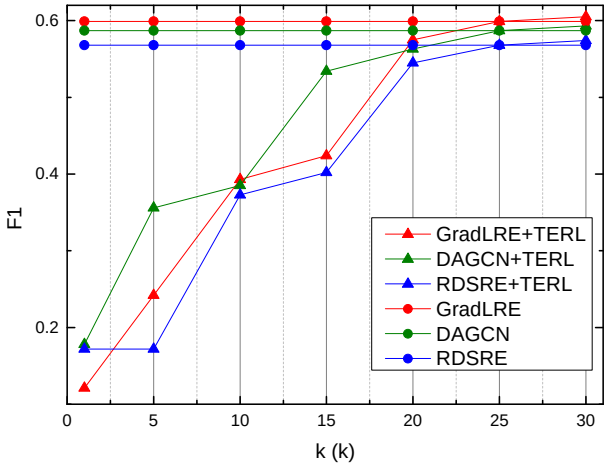


Figure 2: Effect of context length K on RE task.

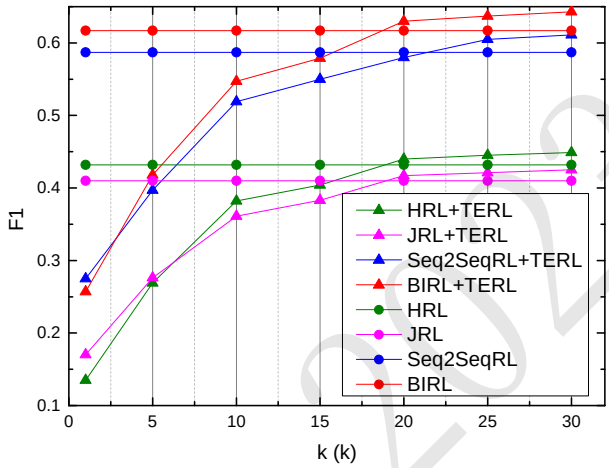


Figure 3: Effect of context length K on joint extraction task.

References

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, P. Abbeel, A. Srinivas, and Igor Mordatch. 2021a. Decision transformer: Reinforcement learning via sequence modeling. In *NeurIPS*.

Yangyi Chen, Jingtong Su, and Wei Wei. 2021b. Multi-granularity textual adversarial attack with behavior cloning. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planners. In *ACL*.

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *EMNLP*.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*.

Zhijiang Guo, Guoshun Nan, Wei Lu, and Shay B. Cohen. 2020. Learning latent forests for medical relation extraction. In *IJCAI*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In **SEMEVAL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*.

Xuming Hu, Fukun Ma, Chenyao Liu, Chenwei Zhang, Lijie Wen, and Philip S. Yu. 2021a. Semi-supervised relation extraction via incremental meta self-training. In *EMNLP*.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021b. Gradient imitation reinforcement learning for low resource relation extraction. In *EMNLP*.

Chia-Chun Hung, Timothy P. Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. 2019. Optimizing agent behavior over long time scales by transporting value. *Nature Communications*, 10.

Michael Janner, Qiyang Li, and Sergey Levine. 2021. Reinforcement learning as one big sequence modeling problem. *ArXiv*, abs/2106.02039.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Kimin Lee, Michael Laskin, A. Srinivas, and P. Abbeel. 2021. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *ICML*.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *ACL*.

Zhixin Li, Yaru Sun, Suqin Tang, Canlong Zhang, and Huifang Ma. 2021. Reinforcement learning with dual attention guided graph convolution for relation extraction. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 946–953.

Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning dual retrieval module for semi-supervised relation extraction. *The World Wide Web Conference*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature*, 518:529–533.

Karthik Narasimhan, Adam Yala, and Regina Barzilay. 2016. Improving information extraction by acquiring external evidence with reinforcement learning. In *EMNLP*.

Emilio Parisotto and Ruslan Salakhutdinov. 2021. Efficient transformers in reinforcement learning using actor-learner distillation. *ArXiv*, abs/2104.01655.

Emilio Parisotto, H. Francis Song, Jack W. Rae, Razvan Pascanu, aglar Gülehre, Siddhant M. Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, Matthew M. Botvinick, Nicolas Manfred Otto Heess, and Raia Hadsell. 2020. Stabilizing transformers for reinforcement learning. In *ICML*.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *ACL*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. *Proceedings of the 26th International Conference on World Wide Web*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML/PKDD*.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, 1:29–36.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2018. A hierarchical framework for relation extraction with reinforcement learning. In *AAAI*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. *ArXiv*, abs/1811.03925.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Yashen Wang and Huanhuan Zhang. 2021. Birl: Bidirectional-interaction reinforcement learning framework for joint relation and entity extraction. In *DASFAA*.

Shaolei Wang, Yue Zhang, Wanxiang Che, and Ting Liu. 2018. Joint extraction of entities and relations based on a novel graph scheme. In *IJCAI*.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *ICML*.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2020. Gdpnet: Refining latent multi-view graph for relation extraction. *ArXiv*, abs/2012.06780.

Xiaofeng Yu and Wai Lam. 2010. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *COLING*.

Vinícius Flores Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David P. Reichert, Timothy P. Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew M. Botvinick, Oriol Vinyals, and Peter W. Battaglia. 2019. Deep reinforcement learning with relational inductive biases. In *ICLR*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *ACL*.

Xiangrong Zeng, Shizhu He, Daojian Zeng, Kang Liu, Shengping Liu, and Jun Zhao. 2019. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning. In *EMNLP/IJCNLP*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *EMNLP*.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. *ArXiv*, abs/1706.05075.

Qinqing Zheng, Amy Zhang, and Aditya Grover. 2022. Online decision transformer. In *ICML*.

Xin Zhou, Luping Liu, Xiaodong Luo, Haiqiang Chen, Linbo Qing, and Xiaohai He. 2019. Joint entity and relation extraction based on reinforcement learning. *IEEE Access*, 7:125688–125699.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 677-688, Harbin, China, August 3 - 5, 2023.

P-MNER: Cross Modal Correction Fusion Network with Prompt Learning for Multimodal Named Entity Recognitiong

Zhuang Wang[1], Yijia Zhang[1*], Kang An[1], Xiaoying Zhou[1], Mingyu Lu[2*], Hongfei Lin[3]
[1]College of Information Science and Technology, Dalian Maritime University / Dalian, China
[2]College of Artificial Intelligence, Dalian Maritime University / Dalian, China
[3]College of Computer Science and Technology, Dalian University of Technology / Dalian, China
{wang_1120211498,zhangyijia,1120221416_ankang,zhouxiaoying}@dlmu.edu.cn
lumingyu@dlmu.edu.cn
hflin@dlut.edu.cn

Abstract

Multimodal Named Entity Recognition (MNER) is a challenging task in social media due to the combination of text and image features. Previous MNER work has focused on predicting entity information after fusing visual and text features. However, pre-training language models have already acquired vast amounts of knowledge during their pre-training process. To leverage this knowledge, we propose a prompt network for MNER tasks (P-MNER).To minimize the noise generated by irrelevant areas in the image, we design a visual feature extraction model (FRR) based on FasterRCNN and ResNet, which uses fine-grained visual features to assist MNER tasks. Moreover, we introduce a text correction fusion module (TCFM) into the model to address visual bias during modal fusion. We employ the idea of a residual network to modify the fused features using the original text features.Our experiments on two benchmark datasets demonstrate that our proposed model outperforms existing MNER methods. P-MNER's ability to leverage pre-training knowledge from language models, incorporate fine-grained visual features, and correct for visual bias, makes it a promising approach for multimodal named entity recognition in social media posts.

1 Introduction

With the rapid development of the Internet, social media platforms have experienced an exponential growth of content. These platforms offer a wealth of user-generated posts that provide valuable insights into the events, opinions, and preferences of both individuals and groups. Named Entity Recognition (NER) is a crucial task in which entities contained in textual data are detected and mapped to predefined entity types, such as location (LOC), person (PER), organization (ORG), and miscellaneous (MISC). Incorporating visual information from posts has been shown to significantly enhance the accuracy of entity prediction from social media content. For instance, as illustrated in Fig.1, the sentence "Alban got Rikard a snowball in the snow" can be easily resolved by leveraging the visual cues in the accompanying image, allowing us to identify "Rikard" as an animal. However, relying solely on textual data to predict entities may lead to erroneous predictions, such as identifying "Rikard" as a name.

With the continuous evolution of deep learning models, several multi-modal Named Entity Recognition (NER) models have been proposed to enhance the prediction performance of entities by incorporating visual information. These models employ techniques such as cross-attention(Wu et al., 2020; Zhang et al., 2018), adversarial learning(Goodfellow et al., 2014; Frankle and Carbin, 2018), and graph fusion(Xiao et al., 2021; Wu et al., 2020). However, previous methods fused text features with visual features and directly fed them into a neural network model for prediction. This approach overlooks the wealth of information embedded

This work is supported by the National Natural Science Foundation of China (No.61976124)

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

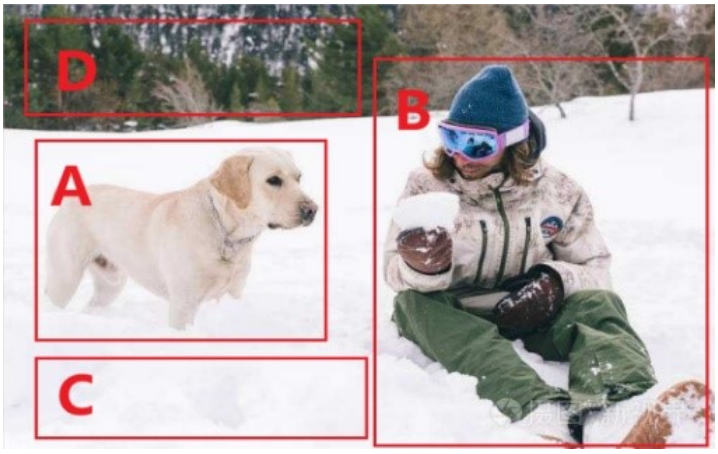


Figure 1: An example for MNER with (A and B) the useful visual clues and the (C and D) useless visual clues.

in the pre-training language model itself. To overcome this limitation, we propose the use of prompt learning(Liu et al., 2023) to process the fused features, followed by final training.

The presence of irrelevant content in an image may negatively impact the performance of Named Entity Recognition (NER) models. As illustrated in Fig.1, regions A and B in an image may aid in identifying entities in a sentence, while regions C and D may not contribute to model prediction. In previous Multi-modal NER (MNER) tasks, however, all visual regions were involved in cross-modal fusion. To address this issue, we propose a novel model(FRR), which utilizes visual objects in images for modal fusion. This approach effectively eliminates extraneous image features that are irrelevant to the corresponding text.

In this paper, we present a new Transformer-based(Vaswani et al., 2017) text correction fusion module (TCFM) to address the issue of cross-modal visual bias in the named entity recognition (NER) task. Inspired by the residual network, the TCFM continuously integrates the original text features with the fusion features to iteratively correct the fusion features. This approach effectively alleviates the problem of visual bias and enhances the performance of the NER task.

In order to showcase the effectiveness of our proposed approach, we conducted a comprehensive set of experiments on two publicly available datasets: Twitter-2015 and Twitter-2017. The obtained experimental results unequivocally demonstrate that our method outperforms the existing MNER algorithm in terms of performance.

The significant contributions of our work can be summarized as follows:

- We introduce a novel approach, the Prompt Network for Named Entity Recognition (P-MNER), which aims to leverage the abundant information present in pre-trained language models. To accommodate the specific requirements of our proposed prompt network, we further present a novel Text Correction Fusion Module (TCFM) that effectively minimizes the visual bias in the fusion process.

- To mitigate the impact of irrelevant visual regions on modal fusion, we propose a novel Feature Extraction Module (FRR) that leverages fine-grained visual objects for more precise feature extraction.

- Experimental results show that our proposed P-MNER network achieves SOTA performance on both datasets.

2 Related work

Named Entity Recognition (NER) has emerged as a crucial component in a plethora of downstream natural language processing (NLP) applications, including but not limited to affective

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.

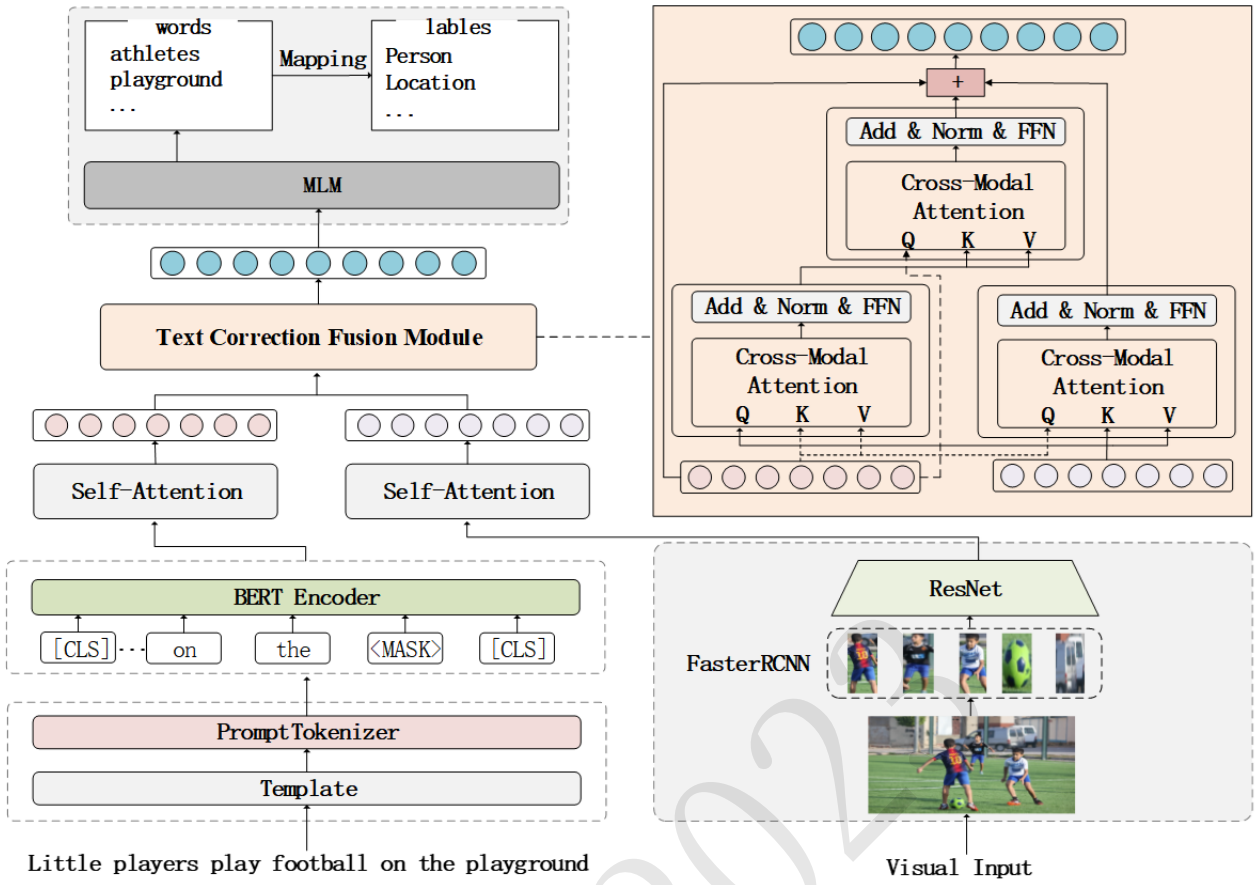


Figure 2: The overall architecture of our P-MNER model.

analysis(Chen et al., 2022), relationship extraction(Gupta et al., 2017), and knowledge graph(Cui et al., 2021) construction. With the advent of neural network models, the use of Bi-LSTM(Chiu and Nichols, 2016) or Convolutional Neural Network (CNN)(Zhao et al., 2017) as encoders, and Softmax(Joulin et al., 2017), RNN(Liu et al., 2016), or Conditional Random Field (CRF)(Zhuo et al., 2016) as decoders has gained popularity in the NER community. For instance, Huang et al.(Huang et al., 2015) utilized BiLSTM and CRF as the encoder and decoder, respectively, for NER tasks. Similarly, Chiu and Nichols et al.(Chiu and Nichols, 2016) proposed a CNN-based encoder with CRF as the decoder to accomplish the final prediction.

Social media posts are characterized by their brevity and high levels of noise, which often lead to suboptimal performance of conventional NER methods when applied to such data. To address this challenge, several recent studies have proposed novel approaches for cross-modal fusion in the context of NER. For instance, Sun et al.(Ritter et al., 2011; Sun et al., 2021) introduced a novel image-text fusion approach for NER tasks, while Zhang et al.(Zhang et al., 2018) employed BiLSTM to combine visual and textual features in multimodal social posts. Similarly, Zheng et al.(Zheng et al., 2020) proposed an adversarial learning approach to tackle the issue of semantic gap in multimodal NER tasks. Lu et al.(Lu et al., 2018) integrated an attention mechanism into the modal fusion process and introduced a visual gate to filter out noise in the image. Lastly, Yu et al.(Yu et al., 2020) designed a multimodal Transformer model for the MNER task and introduced an entity span module to facilitate the final prediction.

In order to utilize the vast amount of knowledge encoded in pre-trained language models, Wang et al.(Wang et al., 2022) proposed a prompt-based method, namely PromptMNER, which extracts visual features and subsequently fuses them with input text for enhanced performance.

Previous studies on named entity recognition (NER) in social media have yielded promising results. However, these methods have been unable to fully utilize the power of pre-training language models in capturing the contextual nuances of input text. This shortcoming has limited the overall effectiveness of NER in social media. To overcome this challenge, our paper proposes a novel prompt learning approach that directly leverages the knowledge embedded in pre-training language models to enrich input text and image features. By doing so, we are able to tap into the full potential of these models and achieve a higher level of accuracy in NER tasks.In essence, our approach represents a significant step forward in the field of NER for social media. It addresses a critical limitation of previous methods and provides a more effective way of leveraging pre-training language models. Our approach offers a promising new avenue for future research.

3 Methods

The objective of this study is to predict a tag sequence $L = (l_1, l_2, ..., l_n)$. given a sentence Y and an associated picture V as input. Here, belongs to a predefined set of tags for the BIO2 tagging pattern.

Fig.2 depicts the overall architecture of our proposed model, which comprises four modules: visual feature extraction, text feature extraction, modal fusion, and prompt learning. The visual feature extraction module takes the objects in the picture as input. In the text feature extraction module, we leverage BERT to extract features by processing the wrapped text. We also introduce a text correction fusion module to obtain more precise fusion features. In the prompt learning module, we utilize the Bert Masked Language Model for prompt learning.

3.1 Text Feature Extraction

In Fig.2, the input sentence $Y = (y_1, y_2, , y_n)$ is demonstrated on the left. To structure the sentences, a wrapping class is utilized, which adheres to a predetermined template. For MNER direction, a prompt template is introduced: "<sentence>, the word of <entity> is <mask>." Within the template, <sentence> represents the sentence Y, <entity>∈y.

In order to effectively process complex information from input text and templates, a new Tokenizer has been introduced to tokenize the input sentences that are wrapped by the wrapper class. Following the pre-training models, special tokens are added to the tokenized sequences to form a sequence, denoted as $S = (s_0, s_1, , s_{n+1})$, where s_0 and s_{n+1} represent the two special tokens at the beginning and end of the final sequence. The tokenized sequences are then sent to the embedding layer where BERT embedding is utilized to convert each word into a vector form that includes token embedding, segment embeddings, and position embeddings.

$$x_i = e^t(s_i) + e^s(s_i) + e^p(s_i) \tag{1}$$

where $\{e^t, e^s, e^p\}$ denotes the embeddings lookup table.$X = (x_0, x_1, \ldots x_{n+1})$ is the word representation of S, where x_i is the sum of word, segment, and position embeddings for token y_i.

In various social media posts, the same word may have different meanings depending on the context. To address this challenge, we adopt BERT as the sentence encoder. The resulting embedding representation is fed into the BERT encoder, producing a signature of encodings $R = \{r_0, r_1, \ldots, r_{n+1}\}$.

The Self-Attention mechanism establishes direct links between any two words in a sentence through a calculation step, significantly reducing the distance between distance-dependent features and enabling their efficient use. Consequently, we feed the hidden representation output by BERT Encoder into Self-Attention to capture long-distance dependencies in a sentence.

$$T = \text{softmax}\left(\frac{[W_{q_t}R]^T[W_{k_t}R}{\sqrt{d_t}}\right)[W_{v_t}R]^T \tag{2}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.

where $\{W_{k_t}, W_{v_t}, W_{q_t}\}$ is parameter matrices for the key, value and query.The final text feature $T = \{t_0, t_1, \ldots, t_{n+1}\}$, where t_i is the generated contextualized representation for y_i.

3.2 Visual Feature Extraction

We utilize object-level visual features in the visual feature extraction module to aid named entity recognition, and introduce a novel method for feature extraction.

To begin with, we feed the image into the Faster RCNN(Faster, 2015) detection module to extract the visual object area. Specifically, we input the image into a feature extraction network that includes convolutional layers, pooling layers, and rectified linear unit (ReLU) layers to obtain feature maps of the image. Next, we pass the feature maps to the Region Proposal Networks (RPN) to train them to extract Region Proposal regions from the original maps. Then, we use RoI Pooling to normalize candidate recognition areas of different sizes and shapes into fixed-size target recognition areas. RoI Pooling collects proposals (coordinates of each box) generated by RPN and extracts them from feature maps. Finally, we process the resulting proposals with Fully Connected and Softmax to determine the probability that each proposal corresponds to a particular category.

Typically, only a small number of visual entities are needed to emphasize the entities in a sentence. To accomplish this, we choose the first m visual objects with a probability exceeding 0.95. Then, we crop the original picture based on these proposals to obtain the final visual object set, denoted as $I = \{I_1, I_2, \ldots, I_m\}$, where I_i represents the $i-th$ visual object.

The residual network is among the most advanced CNN image recognition models, with the ability to extract meaningful features from input images. Thus, we feed the resulting visual objects into a pre-trained 152-layer ResNet and use the output of the last convolution layer as the visual characteristics of each object, denoted as $\tilde{V} = \left\{\widetilde{V_1}, \widetilde{V_2}, \ldots, \widetilde{V_m}\right\}$, where $\widetilde{V_i} \in R^{1024}$ represents the features of the $i-th$ object. We then employ Self-Attention to enable each visual block to fully comprehend the context of the visual features:

$$V = \text{softmax}\left(\frac{\left[W_{q_v}\widetilde{V}\right]^T\left[W_{k_v}\widetilde{V}\right]}{\sqrt{d_v}}\right)\left[W_{v_v}\widetilde{V}\right]^T \tag{3}$$

where $\{W_{q_v}, W_{k_v}, W_{v_v}\}$ denote the weight matrices for the query, key and value. The final visual features are: $V = \{v_1, v_2, \ldots, v_m\}$, v_i refer to the visual features processed by Self-Attention.

3.3 Text Correction Fusion Module

In the text feature extraction module, we have extracted text features through con-textual comprehension. However, the short length of social media posts and the presence of irrelevant information make it challenging to accurately identify entities using text information alone. To address this issue, we utilize visual objects in pic-tures to guide text-based word representations for improved accuracy. Nevertheless, the challenge of visual bias in modal fusion remains. Therefore, we propose a text correction fusion module to generate the final fusion features.

As shown in the right of Fig.2, we initially apply a k-head cross-modal attention mechanism. This involves using the visual features $V = \{v_1, v_2, \ldots, v_m\}$ as queries in the self-attention mechanism and utilizing the text features $T = \{t_0, t_1, \ldots, t_{n+1}\}$ as keys and values:

$$H_i(V,T) = \text{softmax}\left(\frac{[W_{q_i}V]^T[W_{k_i}T]}{\sqrt{d/k}}\right)[W_{v_i}T]^T \tag{4}$$

$$M\_H(V,T) = W'[H_i(V,T), \ldots, H_k(V,T)]^T \tag{5}$$

where H_i refers to the i-th head of cross-modal attention, $\{W_{q_i}, W_{k_i}, W_{v_i}\}$ and W' denote the weight matrices for the query, key, value, and multi-head attention, respectively. By utilizing

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.

this cross-attention approach, we can derive feature representations based on the correlation between words and visual objects in the text. We then process the fused features through two normalization layers and a feed-forward neural network(Vaswani et al., 2017):

$$\tilde{P} = LN\left(V + M\_H(V, T)\right) \tag{6}$$

$$P = LN(\tilde{P} + FFN(\tilde{P})) \tag{7}$$

where FFN is the feed-forward network, LN is the layer normalization. Get the text features based on visual objects, denoting as $P = \{p_0, p_1, \ldots, p_{n+1}\}$. Similar to the description above, We use the text feature $T = \{t_0, t_1, \ldots, t_{n+1}\}$ as queries in our own attention and the visual feature $V = \{v_1, v_2, \ldots, v_m\}$ as keys and values. The result is a text-based visual object, denoting as $q = \{q_1, q_2, \ldots, q_m\}$.

During the process of acquiring visual object-based text features, the resulting features may exhibit bias towards the visual mode, as the queries used are primarily based on visual features. In order to alleviate such bias, we propose the use of a cross-modal layer for the re-fusion of text features. In this approach, the original text features are employed as queries, while the visual-based text fusion features are utilized as keys and values. The final cross-modal text representation is obtained as $C = \{c_0, c_1, \ldots, c_{n+1}\}$.

Previous studies have simply connected cross-modal visual features and cross-modal text features, which may lead to biased final fusion features. In this paper, we propose an alternative approach for the final stitching process by connecting initial text features to both cross-modal visual features and cross-modal text features. This method aims to mitigate bias and enhance the quality of the fusion features.

$$H = T + V + C \tag{8}$$

where T is the initial text features, V is the cross-modal visual features, and C is the cross-modal text features.

By incorporating the original text features in the final fusion process, it is effectively reduce visual bias. The resulting fusion feature is denoted as $H = \{h_0, h_1, , h_{n+1}\}$.

3.4 Prompt-Learning Module

In this module, we employ the Bert model as our Pre-trained Language Model (PLM). Our approach involves inputting the resulting fusion feature H into the PLM and leveraging the masked language model (MLM) to reconstruct sequences with <MASK>. The predicted part of the text is replaced with <MASK> during packaging to optimize the pre-training language model stimulation. Our method follows the pre-training language model training process for processing fusion features.

In PLM, our aim is to predict a probability distribution for the <MASK> section that aligns with the objectives of MLM. Here, we are only predicting that part of <MASK> belongs to a certain vocabulary. The ultimate goal is to predict <MASK> as predefined tags in a sentence. To accomplish this, we introduce a verbalizer class to process the output of the MLM model. This class constructs a mapping from original tags to words. When PLM predicts a probability distribution for a masked location in the vocabulary, the verbalizer maps the word to the original label. The output layer can be defined as:

$$c_i = \text{plm}\left(h_i\right) \tag{9}$$

$$d_i = \text{ver}\left(c_i\right) \tag{10}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.

Table 1: Statistics of Twitter datasets.

Entity Type	Train-15	Dev-15	Test-15	Train-17	Dev-17	Test-17
PER	2217	552	1816	2943	626	621
LOC	2091	552	1697	731	173	178
ORG	928	247	839	1674	375	395
MISC	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
Tweets	4000	1000	3257	3373	723	723

where *plm* is masked language model (MLM), *ver* refers to the verbalizer.c_i is the probability distribution of predicted positions on the vocabulary, d_i is a label for prediction. Finally, the prediction tag distribution is $D = \{d_0, d_1, \ldots, d_{n+1}\}$.

During the training phase, we calculate the loss of verbalizer-mapped labels and real labels:

$$L = -\sum_{i=1}^{n} o_i \log(d_i) \tag{11}$$

where o_i is the true tag for d_i.

4 Experiments

We tested the model on two common datasets. Furthermore, we compare our model with the single-mode NER model and the existing multimodal methods.

4.1 Experiment Settings

Datasets: During the model training and evaluation phase, we employed a publicly available dataset from Twitter, comprising four distinct entity types, namely PER, LOC, ORG, and MISC, with non-entity words marked as O. Following the same protocol established by Zhang et al.(Zhang et al., 2018), the dataset was partitioned into training, development, and test sets. Table 1 provides an overview of the dataset, including the number of samples in each set and the count of each entity type.

Hyperparameter: Compared with other NER methods, our model is an experiment performed on a GUP. For visual object extraction, the first five objects with an accura-cy above 0.95 are selected for feature extraction using a pre-trained 152-layer Res-Net. The maximum sentence length is set to 128, and the batch size is 8. The input template has a maximum length of 20, while the encoded text length is set at 256. Cross-modal multi-head attention is applied to facilitate modal fusion, utilizing 12 attention heads. The learning rate and learning attenuation rate are set at 0.005 and 0.01, respectively. During the evaluation phase, standard precision, recall rate, and F1-score are employed as evaluation metrics. The model with the highest performance in the evaluation phase is selected, and its performance is reported on the test dataset.

4.2 Main Result

Table 2 presents the experimental results of our proposed model and the comparative approaches. During model evaluation, we calculated the precision (P), recall (R), and F1-score (F1) of our model.

In the upper section of Table 2, we initially conducted a series of experiments using a text-only model to extract features. Our findings revealed that employing BERT as the encoder for text feature extraction resulted in significantly superior results compared to other methods. We believe that the contextualized word representation and contextual understanding of the input

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.

Table 2: Performance comparison on two TWITTER datasets. Specifically, B-L+CRF and C+B-L+CRF refers to Bi-LSTM+CRF and CNN+Bi-LSTM+CRF, respectively.

Models	TWITTER-2015			TWITTER-2017		
	P	R	F1	P	R	F1
B-L+CRF	68.14	61.09	64.42	79.42	73.42	76.31
C+B-L+CRF	66.24	68.09	67.15	80.00	78.76	79.31
T-NER	69.54	68.65	69.09	-	-	-
BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
MNER-MA	72.33	63.51	67.63	-	-	-
AGBAN	74.13	72.39	73.25	-	-	-
UMT	71.67	75.23	73.41	85.28	85.34	85.31
UMGF	74.49	75.21	74.85	86.54	84.50	85.51
PromptMNER	78.03	79.17	78.60	89.93	90.10	90.27
Ours	79.18	79.55	79.43	90.11	91.23	91.31

text played a crucial role in enhancing the performance of the NER models. In order to achieve even deeper text representation, we leveraged BERT to extract hidden features of the text.

Moreover, we took our analysis a step further and experimented with some representative multimodal NER models to compare their performance with single-mode NER models. As shown in Table 2, the results demonstrate that MNER-MA outperforms the single-mode NER models, indicating the effectiveness of combining visual information in NER tasks. However, we noticed that when BERT was utilized to replace the encoder in the model, the observed improvement was relatively modest. Therefore, it is evident that novel methods need to be developed and employed to address the current limitations in this area.

Prompt learning, a novel paradigm, has demonstrated strong potential in the field of NLP. Wang et al. (Wang et al., 2022)propose utilizing prompt learning to aid in the extraction of visual features. Specifically, they suggest employing the CLIP model as a prompt language model (PLM) to leverage the learned information from the pre-training stage for visual feature extraction. During training, both visual and text information are processed and fed into the PLM to obtain visual features based on prompts. Finally, the extracted visual and text features are fused together.

The results presented in Table 2 unequivocally demonstrate the superior performance of our proposed approach over existing single-mode approaches in the task of named entity recognition (NER). Our method outperforms the current state-of-the-art MNER method as well, owing to our incorporation of prompt learning, which allows us to extract rich information from the pre-trained language model. The primary reason for our success is the utilization of visual context, which enables us to make full use of the available information and improve the overall accuracy of the model.Our approach outperforms the promptMNER method as well. The incorporation of prompt learning in our model allows us to effectively fuse the visual and text features, thereby making the most of the pre-trained model's knowledge during the training process. As a result, we are able to achieve a better overall performance in the NER task.In summary, our proposed approach offers a significant improvement over existing single-mode approaches in the NER task. Our method outperforms both the current state-of-the-art MNER method and the promptMNER method. By incorporating visual context and prompt learning, we are able to effectively extract and utilize the rich information contained in the pre-trained language model, resulting in superior performance.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.

Table 3: The effect of each module in our model.

Models	TWITTER-2015			TWITTER-2017		
T+V	72.76	72.53	72.31	83.74	83.24	84.33
T+V+TCFM	74.38	74.12	73.35	85.16	84.35	85.41
T+FRR+TCFM	75.32	75.54	75.14	85.47	84.89	86.02
OURS	79.18	79.55	79.43	90.11	91.23	91.31

4.3 Ablation Result

To evaluate the effectiveness of each component in our proposed P-MNER model, we conduct ablation experiments. Our results, presented in Table 3, indicate that all components in the P-MNER model have contributed significantly to the final predicted results.

T+V is the baseline of our MNER task, with BERT utilized as the encoder for text feature extraction and ResNet employed as the encoder for visual data. The experimental results presented in Table 3 demonstrate that our proposed baseline model achieves a higher F1-score than all single-mode models, thereby validating the effectiveness of incorporating visual information into our model.

T+V+TCFM replaced the modal splicing part with TCFM.Table 3 shows a significant increase in F1-score of 1.22% and 1.08%, respectively, upon implementation of the proposed text correction fusion module, which validates our proposed modal fusion mechanism. Our TCFM module improved accuracy by 1.62% and 1.42% on the two datasets, due to its ability to continuously utilize text information to correct feature bias during mode fusion. This effectively addresses the problem of feature alignment and improves model performance.

T+FRR+TCFM uses a new visual feature extraction method (FRR).Table 3 illustrates that our proposed visual feature extraction module achieved F1-scores of 75.14% and 86.02% on the two datasets, respectively, surpassing other NER methods.

Our proposed model, OURS, is a comprehensive approach that employs prompt learning throughout the entire system. The effectiveness of prompt learning is demonstrated in Table 3, where F1-scores of 79.43% and 91.31% were achieved on two different datasets, respectively, surpassing the current state-of-the-art methods in MNER. The superiority of OURS can be attributed to its ability to deeply explore latent knowledge within pre-trained language models, thanks to the prompt learning technique. Moreover, we achieved a 0.83% and 1.04% improvement in F1-score compared to promptMNER, due to our use of prompt learning for feature fusion processing. Our method is more effective in extracting hidden knowledge from pre-trained language models.

4.4 Case analysis

To further strengthen our argument regarding the effectiveness of our proposed method, we have conducted a comprehensive case study analysis. We present the results of this analysis in Fig.3, where we compare the performance of three models for entity prediction: BERT-CRF, UMGF, and P-MNER.

BERT-CRF is a text-only NER model, while UMGF and P-MNER are MNER models that incorporate both visual and textual information. In the first case of our analysis, BERT-CRF failed to accurately predict the entity "Susie". We attribute this to the model's lack of attention to visual information. This highlights the importance of incorporating visual data to improve entity prediction accuracy.

In the second case, all three models correctly predicted the entities. However, this case also revealed that not all image information is semantically consistent with the accompanying text. Hence, the incorporation of visual data should be done thoughtfully and with a proper

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.

Visual Modaliy							
Textual Modality	**[Susie MISC]** is playing football in the **[park LOC]**		**[Sonam Kapoor PER]** to walk on red carpet at **[cannes film festival MISC]**		**[NFT ORG]** star **[patrick willis PER]** is thriving in retirement as a **[silicon valley LOC]** tech worker		
BERT-CRF	(Susie MISC) (park LOC)	× √	(Sonam Kapoor PER) (cannes film festival MISC)	√ √	(NFT ORG) (patrick willis PER) (silicon valley LOC)	× √ ×	
UMGF	(Susie MISC) (park LOC)	√ √	(Sonam Kapoor PER) (cannes film festival MISC)	√ √	(NFT ORG) (patrick willis PER) (silicon valley LOC)	√ √ ×	
P-MNER	(Susie MISC) (park LOC)	√ √	(Sonam Kapoor PER) (cannes film festival MISC)	√ √	(NFT ORG) (patrick willis PER) (silicon valley LOC)	√ √ √	

Figure 3: Three cases of the predictions by BERT-CRF, UMGF and OUR MODE

understanding of the context.

Finally, in the third case, both BERT-CRF and UMGF failed to accurately predict the entity types. In contrast, our P-MNER model leverages the pre-trained language model to effectively acquire knowledge and make accurate entity type predictions. Our model outperformed the other models by a considerable margin, thereby highlighting the superiority of our proposed method.

In conclusion, the case study analysis provides strong evidence to support our claim that incorporating visual information enhances the accuracy of entity prediction. Additionally, our proposed P-MNER model outperforms the other models by leveraging the pre-trained language model to acquire knowledge and make accurate predictions.

5 Conclution

In this paper, we have introduced the P-MNER architecture, which has been specifically designed to tackle named entity recognition (MNER) tasks. Our proposed architecture leverages the power of prompt learning to process modal fusion features, thereby enabling the model to fully exploit the wealth of knowledge that pre-trained language models have to offer during training. We also proposed a fine-grained visual object feature extraction module (FRR) to address the issue of noise caused by irrelevant visual areas. This module aids in the MNER task by extracting only the relevant visual information, thus improving the accuracy of the model. To further address the issue of visual bias across modes, we proposed a new text correction fusion module. This module aligns the fusion features with text features to reduce visual bias and improve the model's performance. Experimental results on benchmark datasets demonstrate that our P-MNER model outperforms state-of-the-art approaches. Our model's superior performance is attributed to its ability to effectively utilize pre-trained language models and its innovative feature extraction and fusion modules. Overall, our proposed P-MNER architecture offers a promising solution for named entity recognition tasks, and we believe that our approach can be extended to other natural language processing tasks to improve their performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61976124)

References

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete opinion tree induction for aspect-based sentiment analysis. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2051–2064.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. Transactions of the association for computational linguistics, 4:357–370.

Zijun Cui, Pavan Kapanipathi, Kartik Talamadupula, Tian Gao, and Qiang Ji. 2021. Type-augmented relation prediction in knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 7151–7159.

RCNN Faster. 2015. Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 9199(10.5555):2969239–2969250.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2681–2690.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.

Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou, et al. 2017. Efficient softmax approximation for gpus. In International conference on machine learning, pages 1302–1310. PMLR.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9):1–35.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1990–1999.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In Proceedings of the 2011 conference on empirical methods in natural language processing, pages 1524–1534.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In Proceedings of the AAAI conference on artificial intelligence, volume 35, pages 13860–13868.

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

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022. Promptmner: Prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III, pages 297–305. Springer.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In Proceedings of the 28th ACM International Conference on Multimedia, pages 1038–1046.

Zeguan Xiao, Jiarun Wu, Qingliang Chen, and Congjian Deng. 2021. Bert4gcn: Using bert intermediate layers to augment gcn for aspect-based sentiment classification. arXiv preprint arXiv:2110.00171.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In Proceedings of the AAAI conference on artificial intelligence, volume 32.

Zhehuan Zhao, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2017. Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC medical genomics, 10:75–83.

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. IEEE Transactions on Multimedia, 23:2520–2532.

Jingwei Zhuo, Yong Cao, Jun Zhu, Bo Zhang, and Zaiqing Nie. 2016. Segment-level sequence modeling using gated recursive semi-markov conditional random fields. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1413–1423.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 689-700, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Self Question-answering: Aspect Sentiment Triplet Extraction via a Multi-MRC Framework based on Rethink Mechanism

Fuyao Zhang[1],Yijia Zhang[1,✉],Mengyi Wang[1],Hong Yang[1],Mingyu Lu[1],Liang Yang[2]
[1]Dalian Maritime University, Dalian
{zhangfuyao,zhangyijia,mengyiw,yanghong,lumingyu}@dlmu.edu.cn
[2]Dalian University of Technology, Dalian
liang@dlut.edu.cn

Abstract

The purpose of Aspect Sentiment Triplet Extraction (ASTE) is to extract a triplet, including the target or aspect, its associated sentiment, and related opinion terms that explain the underlying cause of the sentiment. Some recent studies fail to capture the strong interdependence between ATE and OTE, while others fail to effectively introduce the relationship between aspects and opinions into sentiment classification tasks. To solve these problems, we construct a multi-round machine reading comprehension framework based on a rethink mechanism to solve ASTE tasks efficiently. The rethink mechanism allows the framework to model complex relationships between entities, and exclusive classifiers and probability generation algorithms can reduce query conflicts and unilateral drops in probability. Besides, the multi-round structure can fuse explicit semantic information flow between aspect, opinion and sentiment. Extensive experiments show that the proposed model achieves the most advanced effect and can be effectively applied to ASTE tasks.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) is a fine-grained task (Zhang et al., 2022). Its purpose is to detect the sentiments of different entities rather than infer the overall sentiment of sentences. As shown in Figure 1, researchers proposed many subtasks of ABSA, such as Aspect Term Extraction (ATE) (Ma et al., 2019), Opinion Term Extraction (OTE) (Zhao et al., 2020), Aspect-based Sentiment Classification (ABSC) (Hazarika et al., 2018), Aspect-oriented Opinion Extraction (AOE) (Fan et al., 2019), etc. Aspect terms refer to words or phrases that describe the attributes or characteristics of an entity. Opinion terms refer to words or phrases that express the corresponding attitudes of the aspect terms. ATE and OTE aim to extract aspects and opinions from sentences, respectively. For ABSC, given a sentence and an aspect within the sentence, it is possible to predict the sentiment (positive, neutral, or negative) associated with that aspect. In the sentence "The service is good, but the food is not so great", ATE extracts "service" and "food", and OTE extracts "good" and "not so great". ABSC predicts the sentiment polarity of "service" and "food" as positive and negative, respectively. However, these studies focus on individual tasks respectively while neglecting their interdependencies.

Recent studies have focused on joint tasks to explore the interactions among different tasks. Figure 1 provides examples of Aspect Term Extraction and Sentiment Co-classification (AESC) as well as Aspect-Opinion Pair Extraction(pair). However, these subtasks still cannot tell a complete story. Hence Aspect Sentiment Triplet Extraction (ASTE) was introduced. The purpose of ASTE is to extract aspect terms, related opinion terms, and sentiment polarities for each aspect simultaneously. ASTE has two advantages: first, opinions can enhance the expressiveness of the model, helping to determine the sentiment of the aspects better; second, the sentiment dependency between aspects and opinions can narrow the gap of sentiment decision-making, further improving the interpretability of the model.

Peng (Peng et al., 2020) proposed the first solution for ASTE, which jointly extracts aspect-sentiment pairs and opinions using two sequence taggers. Sentiment is attached to aspects through a unified tagging

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

S: The service is good, but the food is not so great.

Substask			
Aspect Term Extraction(ATE)	S	⇨	{service, food}
Opinion Term Extraction(OTE)	S	⇨	{good, not so great}
Aspect-based Sentiment Classification(ABSC)	S+service	⇨	Positive
	S+food	⇨	Negative
Aspect-oriented Opinion Extraction(AOE)	S+service	⇨	good
	S+food	⇨	not so great
Aspect Term and Sentiment Co-extraction(AESC)	S	⇨	{service, Positive}
	S	⇨	{food, Negative}
Aspect-Opinion Pair Extraction(Pair)	S	⇨	{service, good}
	S	⇨	{food, not so great}
Aspect Sentiment Triplet Extraction(ASTE)	S	⇨	{service, good, Positive}
	S	⇨	{food, not so great, Negative}

Figure 1: Illustration of ABSA subtasks

process, and then an exclusive classifier is used to pair the extracted aspect-sentiment pairs with opinions. While this method achieved significant results, there are also some issues. **Firstly**, the model has low computational efficiency because its framework involves two stages and requires training three independent models. **Secondly**, the model does not fully recognize the relationship between ATE and OTE, and does not effectively utilize the correspondence between aspect terms and opinion terms. **Thirdly**, the correspondence between aspect and opinion expressions can be very complex, involving various relationships such as one-to-many, many-to-one, overlapping, and nesting, which makes it difficult for the model to flexibly and accurately detect these relationships. Therefore, we take the solution to the above problems as our challenge.

To address the **first** problem mentioned above, inspired by (Chen et al., 2021), this paper proposes an improved multi-round MRC framework (R-MMRC) with a rethink mechanism to elegantly identify ASTE within a unified framework. To address the **second** problem, we decompose the ASTE into multiple rounds and introduce prior knowledge from the previous round to the current round, which effectively learns the associations between different subtasks. In the first round, we design static queries to extract the first entity of each aspect-opinion pair. In the second round, we design dynamic queries to identify the second entity of each aspect-opinion pair based on the previously extracted entity. In the third round, we design a dynamic sentiment query to predict the corresponding sentiment polarity based on the aspect-opinion pairs obtained in the previous round. In each step, the manually designed static and dynamic queries fully utilize the sentence's explicit semantic information to improve the extraction or classification performance. Based on these steps, we can flexibly capture complex relationships between entities, effectively mine the connection between ATE and OTE, and use these relationships to guide sentiment classification. To address the **third** issue, inspired by human two-stage reading behaviour (Zheng et al., 2019), we introduce a rethink mechanism to validate candidate aspect-opinion pairs further, enhance the information flow between aspects and opinions, and improve overall performance. Our contributions are summarized as follows:

- We proposed an improved multi-round machine reading comprehension framework (R-MMRC) with a rethink mechanism to address the ASTE task effectively.
- The model introduced the rethink mechanism to enhance the information flow between aspects and opinions. The exclusive classifier was added to avoid interference and query conflicts between different Q&A steps. The probability generation algorithm was also introduced to improve the prediction performance further.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

- The proposed model conducts extensive experiments on four public datasets, and experimental results show that our framework is very competitive.

2 Related work

We present related work in two parts, including various subtasks of aspect-based sentiment analysis and machine reading comprehension.

2.1 Aspect-based Sentiment Analysis

ATE. Locating and extracting terms that are pertinent for sentiment analysis and opinion mining is the task of ATE (Xu et al., 2018). Recent studies use two ways to alleviate the noise from pseudo-labels generated by self-learning (Wang et al., 2021).

OTE. OTE is to extract opinion terms corresponding to aspect terms, hoping to find specific words or phrases that describe sentiment (Chen and Qian, 2020).

ABSC. The task's aim is to forecast sentiment polarity of specific aspects. The latest development of ABSC focuses on developing various types of deep learning models: CNN-based (Huang and Carley, 2019), memory-based methods (Majumder et al., 2018), etc. Dependencies and graph structures have also been used effectively for sentiment classification problems (Xu et al., 2020a; Zhang and Qian, 2020).

AOE. Fan (Fan et al., 2019) first proposed this subtask, which aims to extract corresponding opinion terms for each provided aspect term. The difference between AOE and OTE is that the input of AOE contains aspect terms.

AESC. AESC aims to simultaneously extract aspect terms and sentiment. Recent work removes the boundaries of these two subtasks using a unified approach. Chen (Chen and Qian, 2020) proposes a relational awareness framework that allows subtasks to coordinate their work by stacking multitask learning and association propagation mechanisms.

Pair. The Pair task usually uses the pipeline method or directly uses the unified model. Gao (Gao et al., 2021) proposed a machine reading comprehension task based on question answering and span annotation.

ASTE. Peng (Peng et al., 2020) defined a triplet extraction task intending to extract all possible aspect terms, their corresponding opinion terms, and sentiment polarities. Xu (Xu et al., 2021) propose a span-based method to learn the interaction between target words and opinion words and propose a two-channel span pruning strategy.

2.2 Solving NLP Tasks by MRC

The purpose of machine reading comprehension (MRC) is to enable machines to answer questions from a specific context based on queries. Xu (Xu et al., 2021) proposed a post-training method for BERT. Yu (Yu et al., 2021) introduced role replacement into the reading comprehension model and solved the coupling problem in different aspects. To sum up, MRC is an effective and flexible framework for natural language processing tasks.

2.3 Aspect Sentiment Triplet Extraction

ASTE is the latest subtask in the field of ABSA. Xu (Xu et al., 2020b) proposed a position-aware tagging scheme that efficiently captures interactions in triplets. However, they generally overlooked the relationship between words and language features. In a similar vein, Yan (Yan et al., 2021) converted the ASTE task into a generative formulation, but also tended to ignore the linguistic aspects of word features. Meanwhile, Chen (Chen et al., 2022) introduced an enhanced multi-channel GCN that incorporated various language features to enhance the model. However, they failed to consider the interaction between these different language features. In summary, there are still many issues waiting to be resolved in ASTE, and we will try our best to make breakthroughs in ASTE tasks.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

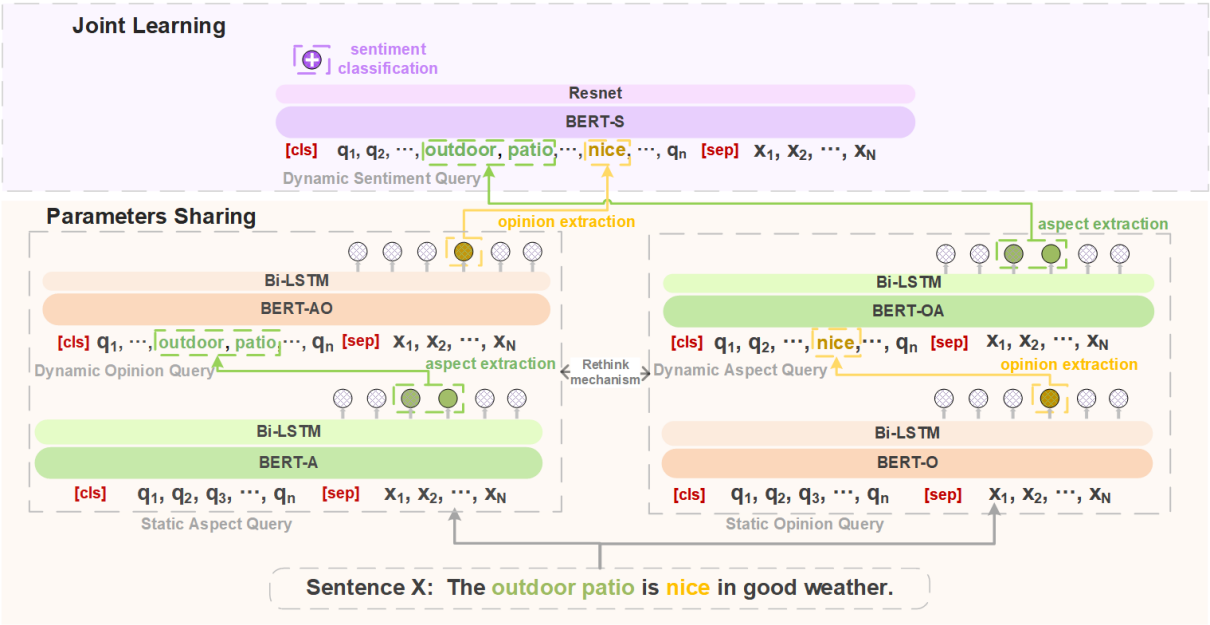


Figure 2: Overview of R-MMRC framework

3 Methodology

3.1 Model Framework

As shown in Figure 2, to address the ASTE task, we propose a multi-round machine reading comprehension framework based on a rethink mechanism. Specifically, we design two modules: parameter sharing and joint learning. First, for the parameter sharing module, we design a bidirectional structure to extract aspect-opinion pairs, consisting of two querying rounds. The first round is static queries aimed at extracting all aspect or opinion sets based on the given query statements. The second round is dynamic queries, aimed at identifying the corresponding opinion or aspect sets based on the results of the static queries and generating aspect-opinion pairs. Then, the rethink mechanism is used to filter out invalid aspect-opinion pairs in the parameter sharing stage. For the joint learning module, the framework employs dynamic sentiment queries to predict the sentiment polarity of the filtered aspect-opinion pairs. During the probability generation stage, the model combines the answers from different queries and forms triplets.

3.2 Query Template Construction

In R-MMRC, we build queries using a template-based method. Specifically, we designed static queries $Q^S = \{q_i^S\}_{i=1}^{|Q^S|}$ and dynamic queries $Q^D = \{q_i^D\}_{i=1}^{|Q^D|}$, where i represents the i-th token in the sentence. In particular, static queries do not carry any contextual information. Dynamic queries require the results of static queries as keywords to search for valid information in sentences. Static and dynamic queries are used to formalize the ASTE task as an MRC task:

Parameter Sharing.

Static Aspect Query q_A^S: We design the query ' Find the aspect in the text? ' to extract a set of aspects $A = \{a_i\}_{i=1}^{|A|}$ from a given review sentence X.

Dynamic Opinion Query q_O^D: We design the query ' Find the opinion of the aspect a_i ?' to extract the relevant opinions $O_{ai} = \{o_{ai,j}\}_{j=1}^{|O_{ai}|}$ for each aspect a_i.

Static Opinion Query q_O^S: We design the query ' Find the opinion in the text? ' to extract the collection of opinions $O = \{o_i\}_{i=1}^{|O|}$ from a given review sentence X.

Dynamic Aspect Query q_A^D: We design the query ' Find the aspect of the opinion o_i ' to extract the corresponding aspects $A_{oi} = \{a_{oi,j}\}_{j=1}^{|A_{oi}|}$ for each opinion O_i.

Through the above queries, dynamic queries elegantly learn the conclusions of static queries and

naturally integrates entity extraction and relationship detection. Although the entity results of these two queries are the same, the latter conveys the information of the former and searches for all entities described by the former, while the former does not carry any contextual information. Then, in the joint learning module, we classify the sentiment corresponding to the aspect-opinion pairs.

Joint Learning.

Dynamic Sentiment Query $q^{D'}$: We build the query ' Find the sentiment of the aspect a_i and the opinion o_i? ' to anticipate the sentiment polarity s_i of each aspect a_i.

Through the queries, we can fully consider the semantic relationship of aspect terms and corresponding opinion terms.

3.3 Input Representations

This section focuses on the triplet extraction task. Given a sentence $X = \{x_1, x_2, \ldots, x_N\}$ with max-length N as the input, and each query $q_i = \left\{q_1^i, q_2^i, \ldots, q_{|q_i|}^i\right\}$ with $|q_i|$ tokens. We use BERT as the model's encoder, and the encoding layer's role is to learn each token's context representation. First, we associate the query Q_i with the review sentence X and obtain the input $I = \left\{[CLS], q_1^i, q_2^i, \ldots, q_{|q_i|}^i, [SEP], x_1, x_2, \ldots, x_N\right\}$ after combination, where $[CLS]$ and $[SEP]$ are the start tag and the segment tag. Bert is used to encode an initial representation sequence $E = \left\{e_1, e_2, \ldots, e_{|q_i|+2+N}\right\}$,which is encoded as a hidden representation sequence $H_e = \left\{h_1, h_2, \ldots, h_{|q_i|+2+N}\right\}$ with stacked transformer blocks.

3.4 Query Answer Prediction

For the first two rounds of static and dynamic queries, the answer is to extract aspect terms or opinion terms from review sentence X. For instance, in Figure 2, the aspect term "outdoor patio" should be extracted as the answer to the Static Aspect Query.

In the original BMRC (Chen et al., 2021), all queries shared a single classifier, which could lead to interference between different types of queries and cause query conflicts. Since there are four different queries in the parameter sharing part, we set an exclusive BERT classifier for each query, which can effectively avoid interference of query conflict and answering step. Classifiers are BERT-A, BERT-AO, BERT-O, and BERT-OA, respectively. The context representation generated by BERT is used for Bi-LSTM to generate sentence hidden state vectors. Since H_e already contains information about aspect or opinion, we obtain specific context representation by aggregating the hidden states of two directions: $H = \left[\overrightarrow{H_{e_f}}; \overleftarrow{H_{e_b}}\right]$, where $\overrightarrow{H_{e_f}}$ is the hidden state of the forward LSTM and $\overleftarrow{H_{e_b}}$ is of the backward LSTM. We adopted the strategy of (Xu et al., 2019) and employ two binary classifiers to predict the answer spans based on the hidden representation sequence H. We utilize two classifiers to predict the possibility that the token x_i is the start or end of the answer. Then, we obtain the logits and probabilities for start and end positions:

$$p_{x_i,q}^{\text{start}} = \text{softmax}\left(W_s h_{|q|+2+i}\right) \tag{1}$$

$$p_{x_i,q}^{end} = \text{softmax}\left(W_e h_{|q|+2+i}\right) \tag{2}$$

where $W_s \in R^{d\times 2}$ and $W_e \in R^{d\times 2}$ are model parameters, d represents the dimension of hidden representations, and $|q|$ stands for the query length.

For dynamic sentiment queries, we utilize the hidden representation of [CLS] to predict the answer. We add a three-class classifier in BERT, called "BERT-S" for short, to predict the sentiment of aspect-opinion pairs. In addition, we add two layers of ResNet network to protect the integrity of information and reduce the loss of information.

$$h = \sigma F\left(h_1, \{W_{ri}\}\right) + h_1 \tag{3}$$

$$p_{X,q}^{D'} = \text{softmax}\left(W_c h\right) \tag{4}$$

where h_1 is the hidden representation of $[CLS]$, refers to ReLU activation function, $F()$ is the residual mapping of fitting, W_{ri} and $W_c = R^{d\times 3}$ is the model parameter.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

3.5 Rethink Mechanism

During the inference process, we combine the answers from different queries into tuples. As shown in Figure 2, the left-side static aspect query q_A^S first identifies all aspect items $A = \{a_1, a_2, \ldots, a_{|A|}\}$. For each aspect item a_i, the corresponding opinion expression set $O_i = \{o_{i,1}, o_{i,2}, \ldots, o_{i,|O_i|}\}$ is identified through the dynamic opinion query q_O^S, resulting in a set of aspect-opinion pairs $V_{AO} = \left[\left(a_i^k, o_{i,j}^k\right)\right]_{k=1}^{I}$, and ultimately obtaining the probability of each candidate pair $p\left(a_i, o_{i,j}\right) = p\left(a_i\right) p\left(o_{i,j} \mid a_i\right)$. Similarly, on the right side, the model first identifies all the opinion items and then queries all corresponding aspect items, and we finally obtain another set of aspect-opinion pairs $V_{OA} = \left[\left(a_{j,i}^k, o_j^k\right)\right]_{k=1}^{J}$, from which we obtain the probability of each candidate pair $p\left(a_{j,i}, o_j\right) = p\left(o_j\right) p\left(a_{j,i} \mid o_j\right)$.

However, the above approach may introduce incorrect aspect-opinion pairs. To better address this issue, we implement a rethink mechanism through a soft-selection strategy. If there exist identical candidate pairs in sets V_{AO} and V_{OA}, then the corresponding aspect-opinion pairs are added to the valid set V. If there are unmatched candidate pairs in V_{AO} and V_{OA}, it indicates that one side's output may be invalid. Therefore, in the soft selection strategy, we adjust the probabilities and introduce a probability threshold λ. If the probability $p(a, o)$ of a certain candidate pair in the difference set is greater than or equal to the probability threshold λ, then this candidate pair is added to the valid set V; otherwise, it is discarded. By using a rethink mechanism, invalid pairs can be better filtered out, reducing the interference of erroneous candidate pairs on the model.

3.6 Entity Pair Probability Generation

After filtering with the rethink mechanism, we obtained a set of valid aspect-opinion pairs, and the next step is to calculate the probability of each candidate pair. In BMRC, the probability of an entity is the product of the probabilities of its start and end positions, and the probability of a candidate pair is the product of the probabilities of the aspect item and opinion item. However, this can result in a product of high probabilities equaling a lower probability value, which does not well represent the model's prediction. As shown in the formula, we balance the probabilities of entities and candidate pairs by taking the square root, which keeps the probability within the range of two related probabilities. This approach can avoid unilateral decrease of probability and better meeting the expectation of the model.

$$p(e) = \sqrt{p\left(e_{\text{start}}\right) * p\left(e_{end}\right)} \tag{5}$$

$$p(a, o) = \begin{cases} \sqrt{p(a) * p(o \mid a)} \cdots \text{ if } (a, o) \in V_{AO} \\ \sqrt{p(o) * p(a \mid o)} \cdots \text{ if } (a, o) \in V_{OA} \end{cases} \tag{6}$$

where e represents the aspect or opinion entity, $start$ and end represent the start and end positions of the entity, and $p(a, o)$ represents the probability of the final candidate pair.

Finally, we employ the dynamic sentiment query $q_i^{D'}$ to predict the various aspects of emotion a_i. We obtain the output of labeled triplets for input sentence X_i, denoted as $T_i = \{(a, o, s)\}$, where $s \in$(positive, neutral, negative) and (a, o, s) refers to (aspect term, opinion term, sentiment polarity).

3.7 Loss Function Construction

In order to learn triplet subtasks jointly and make them promote each other, we integrate loss functions from various queries. For static queries in different directions, we minimize the loss of cross-entropy:

$$L_S = -\sum_{i=1}^{|Q^S|} \sum_{j=1}^{|S|} \left[p_{x_j,q_i}^{start} \cdot \log \hat{p}_{x_j,q_i}^{start} + p_{x_j,q_i}^{end} \cdot \log \hat{p}_{x_j,q_i}^{end}\right] \tag{7}$$

where $p()$ represents the distribution of gold, $\hat{p}()$ indicates the predicted distribution.

Similarly, the loss of dynamic queries in different directions is as follows:

$$L_D = -\sum_{i=1}^{|Q^D|} \sum_{j=1}^{|D|} \left[p_{x_j,q_i}^{start} \cdot \log \hat{p}_{x_j,q_i}^{start} + p_{x_j,q_i}^{end} \cdot \log \hat{p}_{x_j,q_i}^{end}\right] \tag{8}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

Datasets	Train		Dev		Test	
	#S	#T	#S	#T	#S	#T
14-Lap	920	1265	228	337	339	490
14-Res	1300	2145	323	524	496	862
15-Res	593	593	148	238	318	455
16-Res	842	1289	210	316	320	465

Table 1: Statistics of 4 datasets. # S and # T denotes number of sentences and triples.

For dynamic sentiment classification queries, we minimize the cross-entropy loss function:

$$L_{D'} = -\sum_{i=1}^{\left|Q^{D'}\right|} p_{X,q_i}^{D'} \cdot \log \hat{p}_{X,q_i}^{D'} \tag{9}$$

Then, we integrate the aforementioned loss functions to generate the overall model's losses. In this paper, we used the method of AdamW (Loshchilov and Hutter, 2017) to optimize:

$$L(\theta) = L_S + L_D + L_{D'} \tag{10}$$

4 Experiments

4.1 Datasets

To verify the validity of our proposed approach, we conducted experiments on four benchmark datasets from the SemEval ABSA challenge (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016) and listed the statistics for these datasets in Table 1.

4.2 Subtasks and Baselines

To demonstrate the validity of the proposed model, we compared the R-MMRC with the following baseline.

- CMLA+ (Peng et al., 2020) modifies CMLA (Yu et al., 2021), the attention mechanism is used by CMLA to detect the relationship between words and to extract aspects and opinions jointly. CMLA+ incorporates MLP to further determine whether the triplet is accurate during the matching phase.

- Two-Stage (Peng et al., 2020) is a two-stage pipeline model for ASTE. The task of the first stage is to mark all aspects and opinions. The goal of the second stage is to match all aspects with the corresponding opinion expression.

- RACL+ is improved by RACL framework (Chen and Qian, 2020), which uses mechanisms for relationship propagation and multi-task learning to enable subtasks to cooperate in a stacked multi-layer network. Then researchers (Chen et al., 2021) construct the query "Matching aspect a_i and opinion expression o_j?" to detect relationships.

- JET (Xu et al., 2020b) is a first end-to-end model with a novel position-aware tagging scheme that is capable of jointly extracting the triple.

- GTS-BERT (Wu et al., 2020) address the ASTE task in an end-to-end fashion with one unified grid tagging task.

- BMRC (Chen et al., 2021) transforms the ASTE task into a bi-directional MRC task and designs three types of queries to establish relationships between different subtasks.

	Models	14Lap			14Res			15Res			16Res		
		AESC	Pair	ASTE	AESC	Pair	ASTE	AESC	Pair	ASTE	AESC	Pair	ASTE
Precision	CMLA+	54.70	42.10	31.40	67.80	45.17	40.11	49.90	42.70	34.40	58.90	52.50	43.60
	TS	63.15	50.00	40.40	74.41	47.76	44.18	67.65	49.22	40.97	71.18	52.35	46.76
	RACL+	59.75	54.22	41.99	75.57	73.58	62.64	68.35	67.89	55.45	68.53	72.77	60.78
	JET	-	-	52.00	-	-	66.76	-	-	59.77	-	-	63.59
	GTS-BERT	-	66.41	57.52	-	76.23	70.92	-	66.40	59.29	-	71.70	68.58
	BMRC	**72.73**	74.11	**65.12**	77.74	76.91	71.32	72.41	**71.59**	63.71	**73.69**	76.08	67.74
	Ours	70.32	**74.60**	63.76	**78.95**	**78.36**	**72.69**	**72.95**	69.57	**63.96**	72.22	**78.04**	**68.64**
Recall	CMLA+	59.20	46.30	34.60	73.69	53.42	46.63	58.00	46.70	37.60	63.60	47.90	39.80
	TS	61.55	58.47	47.24	73.97	68.10	62.99	64.02	65.70	54.68	72.30	70.50	62.97
	RACL+	**68.90**	**66.94**	51.84	**82.23**	67.87	57.77	**70.72**	63.74	52.53	**78.52**	71.83	60.00
	JET	-	-	35.91	-	-	49.09	-	-	42.27	-	-	50.97
	GTS-BERT	-	64.95	51.92	-	74.84	69.49	-	68.71	58.07	-	**77.79**	66.60
	BMRC	62.59	61.92	54.41	75.10	75.59	70.09	62.63	65.89	58.63	72.69	76.99	**68.56**
	Ours	62.92	63.27	**54.69**	77.00	**78.54**	**72.85**	68.49	**70.33**	**62.64**	68.49	70.33	67.31
F1-score	CMLA+	56.90	44.10	32.90	70.62	48.95	43.12	53.60	44.60	35.90	61.20	50.00	41.60
	TS	62.34	53.85	43.50	74.19	56.10	51.89	65.79	56.23	46.79	71.73	60.04	53.62
	RACL+	64.00	59.90	46.39	**78.76**	70.61	60.11	69.51	65.46	53.95	**73.19**	72.29	60.39
	JET	-	-	42.48	-	-	56.58	-	-	49.52	-	-	56.59
	GTS-BERT	-	65.67	54.58	-	75.53	70.20	-	67.53	58.67	-	74.62	67.58
	BMRC	**67.27**	67.45	59.27	76.39	76.23	70.69	67.16	68.60	61.05	73.18	76.52	68.13
	Ours	66.41	**67.61**	**61.45**	77.96	**78.45**	**72.77**	**69.70**	**69.95**	**62.30**	72.41	**77.62**	**69.67**

Table 2: Statistics of 4 datasets. # S and # T denotes number of sentences and triples.

4.3 Model Settings and Evaluation Metrics

We adopted a Bert (Xu et al., 2019) model for the encoding layer with 12 attention heads, 12 hidden layers, and 768 hidden sizes. The fine-tuning rate of BERT and the learning rate of the training classifier are set to 1e-5 and 1e-3, respectively. We use AdamW optimizer with a weight decay of 0.01 and a warm-up rate of 0.1. At the same time, we set the batch size to 8 and the dropout rate to 0.3. The F1-score is extracted according to the triplet state on the development set. The threshold λ manually adjusted to 0.8, and the step size is set to 0.1.

We use precision, recall, and f1-score as measurement indicators to measure performance, including aspect term and sentiment co-extraction, aspect-opinion pair extraction, and aspect sentiment triplet extraction, respectively. Only when the prediction of aspects, opinions, and sentiments is correct, the triplet's prediction is correct.

4.4 Main Results

Table 2 shows the comparison results for all approaches, from which we derive the following conclusions. The proposed model R-MMRC achieves competitive performance on all datasets, which demonstrates the efficacy of our model. Under the F1 metric, the R-MMRC model is superior to the pipeline method in all datasets. Our model's F1-score on AESC exceeded the baseline average by 2.09%, on Pair by 3.66%, and on ASTE by 2.67%, respectively. The result shows that our method extracts more practical features. We observe that the method based on MRC achieves more significant improvement than the pipeline method, because it establishes the correlation between these subtasks by jointly training multiple subtasks, and alleviates the error propagation problem. It is worth noting that our model also has a significant improvement in precision, which indicates that the model's prediction ability is more reliable than those baselines.

The Pair and ASTE of our model achieve the best performance on all datasets, but the scores of two datasets in AESC are inferior to RACL+. We think that the idea that RACL+ first jointly trains the underlying shared features, then independently trains the advanced private features, and finally exchanges subtask information clues through the relationship propagation mechanism is very effective. TS performs better than CMLA+, since it uses a unified tagging schema to resolve sentiment conflicts. It is noteworthy that the improvement of precision contributes the most to the increase in F1 score. We believe that the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

Model	14Lap			14Res			15Res			16Res		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
R-MMRC	63.76	**54.69**	**61.45**	**72.69**	**72.85**	**72.77**	**63.96**	**62.64**	**62.30**	**68.64**	67.31	**69.67**
—rethink mechanism	**64.45**	53.21	58.30	71.76	65.42	68.34	60.21	59.26	59.57	67.61	65.02	67.32
—exclusive classifier	63.60	55.26	60.58	72.02	68.91	72.36	63.67	61.85	61.98	68.50	**68.39**	69.15
—probability generation	62.51	53.03	59.03	70.64	69.73	70.50	61.16	60.03	60.69	67.26	66.16	67.80
—dynamic query	60.12	50.41	53.27	65.32	67.63	61.16	55.71	56.63	54.05	62.74	60.56	60.13

Table 3: Ablation study results (%). P represents precision, R represents recall, F1 represents Macro-F1 score.

high precision score is due to the rethink mechanism filtering out some negative samples. Both JET and GTS-BERT used labeling schemes, but the latter yielded better results due to the use of more advanced grid labeling and the design of effective inference strategies. The sentiment classification task is more challenging than the previous extraction task because sentiment heavily relies on the extracted aspect-opinion pairs. However, with the help of dynamic sentiment queries constructed based on aspect-opinion information, compared to BMRC, an overall improvement has been achieved.

There is a certain performance gap between the baseline model and our proposed model, which confirms the rationality of the architecture we proposed. We believe that the design of static and dynamic queries can naturally integrate entity extraction and relation detection to enhance their dependency. The rethink mechanism validates each candidate aspect-opinion pair by modeling the information flow from aspect to opinion (or from opinion to aspect), effectively filtering out negative samples and improving the performance of the model. At the same time, the exclusive classifier we introduced, as well as the probability generation algorithm, further improve the performance of the model.

4.5 Ablation Test

We conduct further ablation studies to analyze the impact of different components of R-MMRC. We present the results of ASTE in Table 3, where the first row shows the reproduced results of R-MMRC. The next three rows show the results after removing the rethink mechanism, exclusive classifier, and probability generation, respectively. The last row shows the final results after removing these three parts of the R-MMRC model.

The results show that each component improves the performance of the model, demonstrating their advantages and effectiveness. We remove the dynamic query in the parameter sharing stage of R-MMRC and keep only static queries and the dynamic sentiment query, which is referred to as "-dynamic query". Obviously, removing the dynamic query resulted in a significant drop in model performance. We analyze that after removing the dynamic query, the model could not capture the dependency relationships between entities and separated entity extraction from relation detection. The results indicate that the dynamic query in the parameter sharing stage is highly effective in capturing dependencies.

The advantage of the rethink mechanism is quite significant. Specifically, compared with R-MMRC, the rethink mechanism achieved F1-score improvements of 3.15%, 3.43%, 2.73%, and 2.35% on the four datasets, demonstrating the effectiveness of the rethink mechanism. The probability generation also has a certain improvement effect, which proves that our model better avoids unilateral decline of probability and is more consistent with the model's expectation. For the exclusive classifier, the model's F1 score improvement is relatively smaller compared to the previous two components. Moreover, we find that it has a significant downside of slowing down the model's runtime.

4.6 Case Study

We conduct a case study to illustrate the effectiveness and perform an error analysis in Table 4. We select three cases from datasets and compare our results with RACL+. The reason for choosing RACL+ is that its performance is second only to our R-MMRC model.

The first case has two aspect terms: "exterior patio" and "ambiance". RACL+ cannot extract the triplets corresponding to "ambiance". We speculate that the model only considers the relationship be-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

Case	Ground Truth	RACL+	R-MMRC
The outdoor patio is really nice in good weather, but what ambience the indoors possesses is negated by the noise and the crowds.	(outdoor patio, nice, POS) (ambience, negated, NEG)	(outdoor patio, nice, POS)	(outdoor patio, nice, POS) (ambience, negated, NEG)
The food is pretty good, but after 2 or 3 bad experiences at the restaurant (consiste--ntly rude, late with RSVP'd seating).	(food, pretty good, POS) (seating, RSVP, NEU)	(food, pretty good, POS) (seating, rude, NEG) × (seating, late, NEG) ×	(food, pretty good, POS) (seating, RSVP, NEU)
Dinner is okay not many vegetarian options and the portions are small.	(Dinner, okay, NEU) (positions, small, NEG)	(Dinner, okay, POS) × (vegetarian options, not many, NEG) × (portions, small, NEG)	(Dinner, okay, POS) × (vegetarian options, not many, NEG) × (portions, small, NEG)

Table 4: Case study. Marker × indicates incorrect predictions. The table's abbreviations POS, NEU, and NEG represent positive, neutral, and negative sentiments, respectively.

tween sentence representations of subtasks, which weakens aspect terms in long and complicated sentences. Our proposed model considers all triplets in the sentence because it can guarantee that an aspect or an opinion can produce a pair, precisely like human reading behavior.

The second case is a long sentence with two triplets, and the corresponding sentiments are positive and neutral, respectively. Our R-MMRC correctly extracted aspect terms and opinion terms, and successfully predicted the corresponding polarity. However, RACL+ correctly extracts all aspect terms, but it misjudges the polarity of "seating". The reason is that RACL+ is good at making use of different semantic relationships between subtasks, so it may use irrelevant "rule" and "late" as keywords, and predict the sentiment of "seating" as "negative". On the contrary, R-MMRC can more accurately identify aspect terms and the corresponding opinion terms in complex sentences.

The third case is error analysis. Although the sentence is not long, both models predict the sentiment of "dinner" incorrectly. We analyze that "ok" is usually considered a positive opinion term, so the two models define "dinner" as positive. However, by carefully observing this sentence, we find that the seldom choices in "vegetarian options " are the reason why guests say "dinner" is just "okay" rather than "good". So, sentiment polarity should be "neutral" rather than "positive". We speculate that we are looking for the training loss of maximum likelihood cross entropy in the training set, which may be the reason for the wrong prediction in this case. More interestingly, RACL+ and our R-MMRC, as two excellent solutions, incorrectly consider (vegetarian options, not many, NEG) as a triplet. Therefore, we think that understanding sentence structure through logic and even causal reasoning may provide new ideas for the future research of sentiment analysis.

5 Conclusion

In this paper , we investigate ASTE task and propose an improved multi-round MRC framework with a rethink mechanism(R-MMRC). This framework sequentially extracts aspect-sentiment pairs and performs sentiment classification, which can handle complex correspondences between aspects, opinions, and sentiments. In each round, explicit semantic information can be effectively utilized. Additionally, the rethink mechanism models the bidirectional information flow to verify each candidate aspect-opinion pair, effectively utilizing the corresponding relationship between entities. Exclusive classifiers avoid interference between different queries, and probability generation algorithms further improve prediction performance. The experimental results demonstrate the effectiveness of the R-MMRC framework, further improving the overall performance of the system. More importantly, our model can serve as a general framework to address various tasks of ABSA. However, our model still suffers from the issue of high computational cost, and we hope to compress the model in the future to make it more lightweight.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

Acknowledgements

This work is supported by a grant from the Social and Science Foundation of Liaoning Province (No. L20BTQ008)

References

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3685–3694.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12666–12674.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985.

Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518.

Lei Gao, Yulong Wang, Tongcun Liu, Jingyu Wang, Lei Zhang, and Jianxin Liao. 2021. Question-driven span labeling model for aspect–opinion pair extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12875–12883.

Devamanyu Hazarika, Soujanya Poria, Prateek Vij, Gangeshwar Krishnamurthy, Erik Cambria, and Roger Zimmermann. 2018. Modeling inter-aspect dependencies for aspect-based sentiment analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 266–270.

Binxuan Huang and Kathleen M Carley. 2019. Parameterized convolutional neural networks for aspect level sentiment classification. *arXiv preprint arXiv:1909.06276*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3538–3547.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Md Shad Akhtar, Erik Cambria, and Asif Ekbal. 2018. Iarm: Inter-aspect relation modeling with memory networks in aspect-based sentiment analysis. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3402–3411.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.

M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, and S. Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of International Workshop on Semantic Evaluation at*.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021. Progressive self-training with discriminator for aspect term extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 257–268.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *arXiv preprint arXiv:2010.04640*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. *arXiv preprint arXiv:1805.04601*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Lu Xu, Lidong Bing, Wei Lu, and Fei Huang. 2020a. Aspect sentiment classification with aspect-specific opinion spans. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3561–3567.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020b. Position-aware tagging for aspect sentiment triplet extraction. In *Conference on Empirical Methods in Natural Language Processing*.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. *arXiv preprint arXiv:2107.12214*.

Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, et al. 2021. A unified generative framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2106.04300*.

Guoxin Yu, Jiwei Li, Ling Luo, Yuxian Meng, Xiang Ao, and Qing He. 2021. Self question-answering: Aspect-based sentiment analysis by role flipped machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1331–1342.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3540–3549.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3239–3248.

Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human behavior inspired machine reading comprehension. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 425–434.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 701-712, Harbin, China, August 3 - 5, 2023.

Enhancing Ontology Knowledge for Domain-Specific Joint Entity and Relation Extraction

Xiong Xiong[1,2], **Chen Wang**[1,2], **Yunfei Liu**[1,2], **Shengyang Li**[1,2]*
[1]Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{xiongxiong20,wangchen21,liuyunfei,shyli}@csu.ac.cn

Abstract

Pre-trained language models (PLMs) have been widely used in entity and relation extraction methods in recent years. However, due to the semantic gap between general-domain text used for pre-training and domain-specific text, these methods encounter semantic redundancy and domain semantics insufficiency when it comes to domain-specific tasks. To mitigate this issue, we propose a low-cost and effective knowledge-enhanced method to facilitate domain-specific semantics modeling in joint entity and relation extraction. Precisely, we use ontology and entity type descriptions as domain knowledge sources, which are encoded and incorporated into the downstream entity and relation extraction model to improve its understanding of domain-specific information. We construct a dataset called SSUIE-RE for Chinese entity and relation extraction in space science and utilization domain of China Manned Space Engineering, which contains a wealth of domain-specific knowledge. The experimental results on SSUIE-RE demonstrate the effectiveness of our method, achieving a 1.4% absolute improvement in relation F1 score over previous best approach.

1 Introduction

Extracting relational triples from plain text is a fundamental task in information extraction and it's an essential step in knowledge graph (KG) construction (Lin et al., 2015). Traditional methods perform Named Entity Recognition (NER) and Relation Extraction (RE) in a pipelined manner, that is, first extract entities, and then perform relation classification on entity pairs (Zhou et al., 2005; Chan and Roth, 2011; Gormley et al., 2015). However, since the entity model and relation model are modeled separately, pipelined methods suffer from the problem of error propagation. To address this issue, some joint methods have been proposed (Yu and Lam, 2010; Li and Ji, 2014; Zheng et al., 2017; Wang and Lu, 2020; Yan et al., 2021; Xiong et al., 2022). The task of joint entity and relation extraction aims to simultaneously conduct entity recognition and relation classification in an end-to-end manner.

In recent years, with the development of pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), many entity and relation extraction methods have adopted the paradigm of pre-training and fine-tuning. They utilize PLMs to encode the contextual representations of input text and design various downstream models for task-specific fine-tuning. However, when employed for domain-specific entity and relation extraction, this paradigm suffers from problems of semantic redundancy and insufficiency of domain-specific semantics, particularly in highly specialized domains. On the one hand, PLMs are usually trained on general-domain corpora, which results in a significant amount of redundant semantic information that may not be relevant to specific domains and a lack of sufficient domain-specific semantic information. On the other hand, modeling domain-specific information in this paradigm depends primarily on the role of downstream model and domain-specific labeled data in the fine-tuning stage. However, due to the significantly smaller parameter size of downstream model compared to PLMs and the limited availability of domain-specific labeled data, the effectiveness of domain-specific semantic information modeling is constrained.

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Consequently, some methods attempt to incorporate domain knowledge into entity and relation extraction models to enhance the their comprehension of domain-specific information . These methods can be broadly categorized into two groups according to how knowledge is introduced: pre-training domain-specific language models and integrating domain-specific knowledge graph information into models. Methods of domain-specific pre-training utilize large-scale domain corpora to facilitate continuous pre-training of existing general-domain language models (Araci, 2019; Peng et al., 2019; Lee et al., 2020) or, alternatively, to perform domain-specific pre-training from scratch (Chalkidis et al., 2020; Gu et al., 2021). However, in certain specialized domains, there may be a dearth of enough domain-specific corpora to support domain-specific pre-training. Another category of methods involve integrating domain-specific knowledge graph information into models, where entity mentions in input text are linked to the corresponding entities in knowledge graph, and then the relevant information of the linked entities in the knowledge graph is incorporated into models (Lai et al., 2021; Roy and Pan, 2021; Yang et al., 2021; Zhang et al., 2022). Some of these knowledge graph integration methods are designed simply for the task of relation extraction (RE) where the entities in the sentence are pre-specified, rather than the task of joint entity and relation extraction. In addition, a prerequisite for this kind of approaches is the availability of a well-constructed domain-specific knowledge graph, which is scarce and expensive for some highly specialized domains.

In this study, we explore how to incorporate domain knowledge for the task of joint entity and relation extraction in space science and utilization domain of China Manned Space Engineering. Due to the lack of sufficient domain-specific corpora to support the pre-training of large-scale language models and the absence of well-constructed domain-specific knowledge graphs, the aforementioned approaches cannot be directly used for domain knowledge enhancement. We propose an ontology-enhanced joint entity and relation extraction method (**OntoRE**) for space science and utilization domain. The predefined domain-specific ontology involves many highly specialized entity types that interconnected by different semantic relations, which frames the knowledge scope and defines the knowledge structure in this domain, so it is an appropriate source of domain knowledge. The ontology can be formalized as a graph structure containing nodes and edges, where nodes represent entity types and edges represent relation types. Furthermore, drawing inspiration from the manner in which humans comprehend specialized terminology, we add descriptions for each entity type in the ontology to enhance the semantic information of entity types. We serialize the ontology graph and then adopt an ontology encoder to learn the embeddings of ontology knowledge. The encoded ontology features are fused with input sentence features, and then the entity and relation extraction is carried out under the guidance of ontology knowledge. To evaluate our model, we construct a dataset called **SSUIE** (**S**pace **S**cience and **U**tilization **I**nformation **E**xtraction), which contains rich knowledge about space science and utilization in the aerospace field. This work exclusively pertains to the problem of entity and relation extraction, therefore our model was evaluated on the subset of SSUIE specifically designed for entity and relation extraction, namely **SSUIE-RE**.

The main contributions of this work are summarized below:

1. A dataset named SSUIE-RE is proposed for Chinese entity and relation extraction in space science and utilization domain of China Manned Space Engineering. The dataset is enriched with domain-specific knowledge, which contains 19 entity types and 36 relation types.
2. An ontology-enhanced method for domain-specific joint entity and relation extraction is proposed, which substantially enhances domain knowledge without the need of domain knowledge graphs or large-scale domain corpora. Experimental results show that our model outperforms previous state-of-the-art works in terms of relation F1 score.
3. The effect of domain ontology knowledge enhancement is carefully examined. Our supplementary experiments show that the ontology knowledge can improve the extraction of relations with varying degrees of domain specificity. Notably, the benefit of ontology knowledge augmentation is more evident for relations with higher domain specificity.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

2 Related Work

Among the representative entity and relation extraction approaches in recent years, some focus on solving the problem of triple overlapping (Zeng et al., 2018; Nayak and Ng, 2020; Yu et al., 2020; Wei et al., 2020; Wang et al., 2020) and some focus on the problem of task interaction between NER and RE (Wang et al., 2018; Yan et al., 2021; Xiong et al., 2022). However, the challenge of effectively integrating domain knowledge into entity and relation extraction models to improve their applicability in specific fields, has not been solved well by previous works. We survey the representative works on topics that are most relevant to this research: *domain-specific pre-training* and *integrating knowledge graph information*.

Domain-specific pre-training In order to enhance the domain-specific semantics in PLMs, this family of approaches uses domain corpora to either continue the pre-training of existing generic PLMs or pre-train domain-specific language models from scratch. FinBERT (Araci, 2019) is initialized with the standard BERT model (Devlin et al., 2019) and then further pre-trained using financial text. BioBERT (Lee et al., 2020) and BlueBERT (Peng et al., 2019) are further pre-trained from BERT model using biomedical text. Alsentzer et al. (2019) conduct continual pre-training on the basis of BioBERT, and PubMedBERT (Gu et al., 2021) is trained from scratch using purely biomedical text. Chalkidis et al. (2020) have explored both strategies of further pre-training and pre-training from scratch and release a family of BERT models for the legal domain.

Integrating knowledge graph information This category of methods infuse knowledge into the entity and relation extraction models with the help of external knowledge graph. Lai et al. (2021) adopt the biomedical knowledge base *Unified Medical Language System (UMLS)* (Bodenreider, 2004) as the source of knowledge. For each entity, they extract its semantic type, sentence description and relational information from UMLS with an entity mapping tool MetaMap (Aronson and Lang, 2010), and then integrate these information for joint entity and relation extraction from biomedical text. Roy and Pan (2021) fuse UMLS knowledge into BERT model for clinical relation extraction and explore the effect of different fusion stage, knowledge type, knowledge form and knowledge fusion methods. Zhang et al. (2022) integrate the knowledge from Wikidata [0] into a generative framework for relational fact extraction.

To the best of our knowledge, only a limited number of specialized domains can meet the conditions for applying the two aforementioned methods of enhancing domain knowledge, mainly including biomedical, financial, and legal fields. These fields are comparatively prevalent in human life, so there are more likely to be a considerable amount of domain corpora and data in these fields. However, in highly specialized fields like aerospace, both the large-scale domain-specific corpora and well-constructed domain-specific knowledge graph are scarce. Our proposed method only utilize the ontology and entity type descriptions to inject domain knowledge into entity and relation model without additional prerequisites.

3 Method

In this section, we introduce the architecture of OntoRE. As shown in Figure 1, the model mainly includes four parts: knowledge source, knowledge serialization, knowledge encoding and knowledge fusion. In the following subsections, we provide a detailed description of each component.

3.1 Knowledge Source

In the process of human learning professional knowledge, a common practice is to first understand the meaning of each specialized term and then establish the interrelationships between them. Following this pattern, we leverage ontology and entity type descriptions as domain knowledge sources to augment the capacity of entity and relation extraction models to comprehend domain-specific information. Ontology defines the semantic associations among specialized entity types in the domain, while entity type descriptions provide further explanations for each type of specialized terms. For the space science and utilization domain, the ontology is predefined in SSUIE-RE dataset (see Section 4.1). We collect the official descriptions of domain-specific entity types from the *China Manned Space* official website [1].

[0] https://www.wikidata.org
[1] http://www.cmse.gov.cn

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

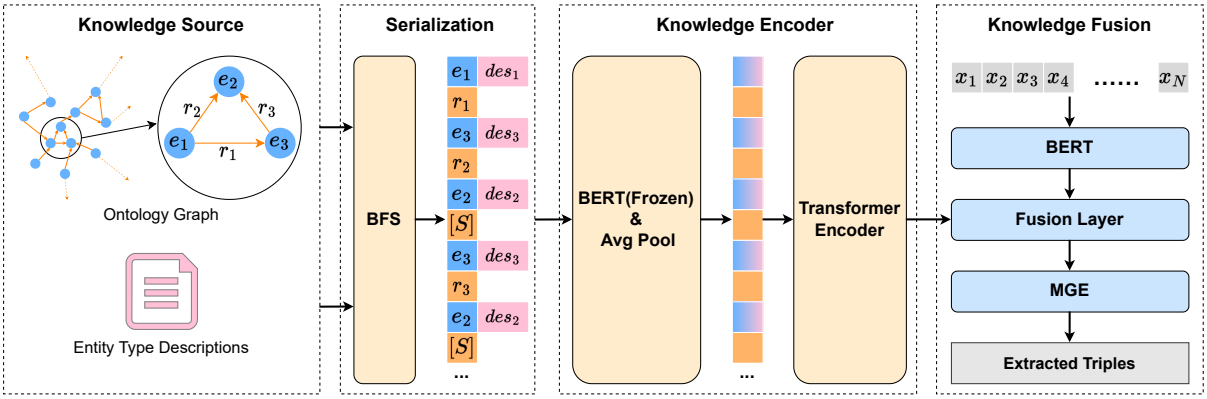


Figure 1: The architecture of the proposed OntoRE framework. We formalize the ontology as a directed graph, where the nodes (blue) represent the predefined entity types and the edges (orange) represent the predefined relation types. The ontology graph is serialized through Breadth First Search (BFS) algorithm. The special marker "$[S]$" represents the end of each level of BFS. des_i denotes the descriptions of entity type e_i. MGE (Xiong et al., 2022) is used as a baseline to verify the effect of our knowledge enhancement method.

Compared to large-scale pre-training corpora and domain-specific knowledge graphs, ontology and entity type descriptions are more accessible for highly specialized domains like space science and utilization.

3.2 Knowledge Serialization

The ontology is a graph structure, where nodes represent entity types and edges represent relation types. It can be formalized as a directed graph $G = (V, E)$, where $V = \{e_1, e_2, \ldots, e_M\}$ denotes the set of predefined entity types and M is the number of predefined entity types. E denotes a multiset of predefined relation types. Additionally, to enrich the semantic information of the entity type nodes in the graph, we append the corresponding entity type descriptions to each node:

$$V' = \{(e_1, des_1), (e_2, des_2), \ldots, (e_M, des_M)\}, \tag{1}$$

where des_m denotes the descriptions of entity type e_m. Then the resulting new graph with the added entity type descriptions can be represented as $G' = (V', E)$.

To facilitate the integration of ontology graph knowledge into entity and relation extraction models that are typically based on sequences, we serialize it using the Breadth First Search (BFS) algorithm while maintaining the structural and semantic properties of the original graph. The graph is represented as an adjacency list in BFS. Before performing the BFS traversal, we initially sort the nodes based on the frequency of their occurrence as entity types in the dataset. Subsequently, we sort the neighboring nodes and edges based on the sum of the node frequency and the edge frequency. This ensures that the nodes and edges with higher frequency in dataset will be traversed first. Then the sorted adjacency list of G is input into the BFS algorithm. During the BFS traversal, we insert a special marker "$[S]$" at the end of each layer of BFS traversal. Taking the nodes e_1, e_2, and e_3 shown in Figure 1 as an example, the first special marker denotes the end of traversing the triple types with e_1 as the head entity, while the second special marker denotes the end of traversing the triple types with e_3 as the head entity, which conveys the structural information among the nodes in the graph. Formally, the BFS serialization process is summarized in Algorithm 1. Then we get an ontology sequence s^π of nodes and edges in the order visited by BFS with level markers:

$$s^\pi = \{s_1^\pi, s_2^\pi, \ldots, s_L^\pi\}, \tag{2}$$

where L represents the length of the ontology sequence obtained by BFS traversal.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Algorithm 1 BFS traversal with level markers

Input: A sorted adjacency list of ontology graph $G' = (V', E)$
Output: A list s^{π} of nodes and edges in the order visited by BFS, with level markers

$s^{\pi} \leftarrow$ empty list
$q \leftarrow$ empty queue
Enqueue the first node of G' into q
Mark all the nodes as unvisited
while q is not empty **do**
  $size \leftarrow$ size of q
  for $i \leftarrow 1$ to $size$ **do**
    $v \leftarrow$ dequeue a node from q
    if v is visited **then**
      break
    end if
    Append v to s^{π}
    for each unvisited neighbor w of v **do**
      Enqueue w into q
      Append the edge (v, w) to s^{π}
      Append w to s^{π}
    end for
    Append a level marker "$[S]$" to s^{π}
    Mark v as visited // The triple types with v as the head entity type have all been traversed
  end for
end while
return s^{π}

3.3 Knowledge Encoding

The elements in s^{π} consist of texts with varying lengths, which encompass relation type words, special markers, and texts formed by concatenating entity type words with their corresponding entity type description words. To get a preliminary semantic representations of these texts, we initialize the representation of each element in s^{π} with a frozen BERT encoder (Devlin et al., 2019) and employ average pooling to unify the feature size. Then we can generate a representation h_k for each element in s^{π} as follows:

$$h_k = \mathrm{AvgPool}(\mathrm{BERT}_{\mathrm{frozen}}(s_k^{\pi})), k \in \{1, 2, \ldots, L\}, \tag{3}$$

where $h_k \in \mathbb{R}^d$ and d is the hidden size of BERT. $\mathrm{AvgPool}(\cdot)$ denotes the operation of average pooling. The representations of the whole ontology sequence s^{π} is concatenated by h_k:

$$H_{s^{\pi}} = [h_1, h_2, \ldots, h_L], \tag{4}$$

where $H_{s^{\pi}} \in \mathbb{R}^{L \times d}$. The feature information in $H_{s^{\pi}}$ are individually encoded from each element in s^{π}. To further capture the inherent information in the ontology sequence, we use a Transformer Encoder (Vaswani et al., 2017) to obtain the final ontology knowledge representations $H_{know} \in \mathbb{R}^{L \times d}$:

$$H_{know} = \mathrm{TransformerEncoder}(H_{s^{\pi}}), \tag{5}$$

3.4 Knowledge Fusion

Given the encoded ontology knowledge representations H_{know}, it can be integrated into different downstream entity and relation extraction models for knowledge enhancement. We select the state-of-the-art methods that have performed best on publicly available benchmark datasets in recent years, and then we evaluate these algorithms on the SSUIE-RE dataset (evaluation results are shown in Table 1). We select the MGE model (Xiong et al., 2022), which performs the best on SSUIE-RE, as our baseline for

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

comparison, and infuse ontology knowledge into it to verify the effectiveness of ontology knowledge enhancement. MGE model uses BERT to encode the contextual information of input sentences, and designs a multi-gate encoder (MGE) based on gating mechanism to filter out undesired information and retain desired information, then performs decoding with table-filling module (Miwa and Sasaki, 2014). We infuse ontology knowledge at the position between the BERT layer and MGE layer, as shown in Figure 1.

We have explored different fusion methods to integrate ontology knowledge representations with input sentence representations, including appending, concatenation and addition. Regarding the appending operation, we concatenate the ontology knowledge representations H_{know} with the input sentence representations along the sequence length dimension. We then apply a self-attention layer to model the guiding effect of ontology knowledge on the extraction of entities and relations from the sentence. The fused representations are calculated as follows:

$$H_{append} = \mathrm{SA}\left([H_b; H_{know}]\right), \tag{6}$$

where $\mathrm{SA}(\cdot)$ means the self-attention layer and H_b denotes the input sentence representations extracted by BERT. $[;]$ denotes the operation of appending, that is, concatenating along the sequence length dimension.

In the case of the concatenation and addition fusion methods, a linear transformation is initially employed to unify the feature dimensions. After this step, the input representations H_b and ontology knowledge representations H_{know} are concatenated along the hidden size dimension or added. The concatenation fusion method can be formalized as below:

$$H_{concat} = \mathrm{Concat}\left(H_b, \mathrm{Linear}(H_{know})\right), \tag{7}$$

where $\mathrm{Concat}(\cdot)$ means the operation of concatenation along the hidden size dimension and $\mathrm{Linear}(\cdot)$ denotes linear transformation. And the fusion method of addition can be formalized as below:

$$H_{add} = H_b + \mathrm{Linear}(H_{know}). \tag{8}$$

Then the representations fused with ontology knowledge is input into the downstream MGE model to obtain the final results of entity and relation extraction.

4 Experiments

4.1 SSUIE-RE Dataset

To evaluate our method, we construct a SSUIE-RE dataset for entity and relation extraction in the space science and utilization domain , which contains rich domain expertise about space science and utilization in the aerospace field. The process of creating SSUIE-RE can be divided into two steps:

Corpora collection and preprocessing The corpora is collected from published professional technical documents in the field, official websites related to China Manned Space Engineering, and Web pages returned by the Google search engine for in-domain professional terms. Before annotation, we preprocess the collected corpora using the following measures:

- We only select Chinese texts and discard texts that are in other languages.
- The invisible characters, spaces and tabs are removed, which are generally meaningless in Chinese.
- In order to eliminate excessively short sentences and incomplete sentences, we split the texts at the Chinese sentence-ending punctuation symbols (e.g., period, question mark, exclamation point), and only retain sentences with more than 10 characters.
- We deduplicate the segmented sentences.

Human annotation We invite annotators with related majors in aerospace field to annotate the processed corpora on the brat[2] platform. The brat platform is an online environment for collaborative text annotation. To ensure the annotation quality, pre-labeling is carried out prior to the formal labeling stage, which aims

[2] https://brat.nlplab.org/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

to ensure that all annotators reach a unified and accurate understanding of the labeling rules. During the annotation process, each sentence is annotated by at least two annotators. If there are inconsistent annotations, the annotation team will discuss the corresponding issue and reach a consensus.

Our final constructed dataset contains 19 entity types, 36 relation types, and 66 triple types. The dataset contains 6926 sentences, 58,771 labeled entities and 30,338 labeled relations. We randomly split the dataset into training (80%), development (10%) and test (10%) set.

4.2 Evaluation and Implementation Details

Following standard evaluation protocol, we use precision (Prec.), recall (Rec.), and micro F1 score (F1) to evaluate our model. The results of NER are considered as correct if the entity boundaries and entity types are both predicted correctly. The results of RE are considered correct if the relation types, entity boundaries and entity types are all predicted correctly.

We use the official implementation of the comparison models to evaluate them on the SSUIE-RE dataset. For fair comparison, we adopt *chinese-bert-wwm* (Cui et al., 2021) as the pre-trained language model for all the models. Our proposed OntoRE model is trained with Adam optimizer for 100 epochs, and the batch size and learning rate are set to be 4 and 2e-5 respectively. The max length of input sentence is set to 300 characters. All the models are trained with a single NVIDIA Titan RTX GPU. The models that achieves the best performance on the development set is selected, and its F1 score on the test set is reported.

4.3 Comparison Models

To ensure a rigorous evaluation, we carefully select state-of-the-art algorithms that have demonstrated superior performance on publicly available benchmark datasets in recent years, and then evaluate their performance on the SSUIE-RE dataset. We compare our model with the following models: (1) **TPLinker** (Wang et al., 2020): this method formulates the task of joint entity and relation extraction as a token pair linking problem, and introduces a handshaking tagging scheme that aligns the boundary tokens of entity pairs for each relation type. (2) **CasRel** (Wei et al., 2020): it models the relations as functions that map subjects to objects rather than discrete labels of entity pairs, allowing for the simultaneous extraction of multiple triples from sentences without the issue of overlapping. (3) **PFN** (Yan et al., 2021): this work utilizes a partition filter encoder to produce task-specific features, which enable effective modeling of inter-task interactions and improve the joint entity and relation extraction performance. (4) **PURE** (Zhong and Chen, 2021): this study constructs two distinct encoders for NER and RE, respectively, and conducts entity and relation extraction in a pipelined manner. (5) **PL-Marker** (Ye et al., 2022): this work is similar to PURE except that it adopts a neighborhood-oriented packing strategy to better model the entity boundary information and a subject-oriented packing strategy to model the interrelation between the same-subject entity pairs. (6) **MGE** (Xiong et al., 2022): this work designs interaction gates to build bidirectional task interaction and task gates to ensure the specificity of task features, based on gating mechanism.

4.4 Main Result

Table 1 shows the comparison of our model OntoRE with other comparison models on SSUIE-RE dataset. As is shown, OntoRE achieves the best results in terms of relation F1 scores. Although PURE achieves the best performance on NER, its relation F1 score is relatively low due to the pipelined architecture which may encounter error accumulation issues. Similarly, PLMarker, which is also a pipelined method, achieves mediocre results on the SSUIE-RE dataset. Among other compared joint methods, MGE achieves the best relation extraction F1 score, and is therefore selected as the baseline model for ontology knowledge injection. As we can see in the table, OntoRE achieves an absolute entity F1 improvement of +0.6% and absolute relation F1 improvement of +1.4% compared to MGE, which indicates that the ontology knowledge injection can enhance the performance of entity and relation extraction in highly specialized domain. Further observation reveals that the models with the best performance on general domain datasets may not perform well in specific professional domains, which reflects the necessity of introducing domain knowledge for entity and relation extraction in specialized fields.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Model	NER			RE		
	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
TPLinker (Wang et al., 2020)	77.0	56.0	64.8	65.3	40.8	50.2
CasRel (Wei et al., 2020)	-	-	-	57.8	53.5	55.6
PFN (Yan et al., 2021)	74.9	75.8	75.4	57.8	62.0	59.8
PURE (Zhong and Chen, 2021)	80.5	80.6	**80.6**	55.0	67.4	60.6
PL-Marker (Ye et al., 2022)	80.2	62.6	70.3	55.5	33.4	41.7
MGE (Xiong et al., 2022)	75.8	76.3	76.0	60.0	64.2	62.0
OntoRE (Ours)	75.0	78.3	76.6	62.4	64.5	**63.4**

Table 1: Overall results of different methods on SSUIE-RE Dataset. The results of all comparison models are implemented using official code. We use the same *chinese-bert-wwm* (Cui et al., 2021) pre-trained encoder for all these models. Results of PURE and PL-Marker are reported in single-sentence setting for fair comparison. Results of OntoRE are reported under the utilization of addition fusion method.

4.5 Effect of Domain Knowledge Enhancement

Although our proposed OntoRE achieves the best results on the SSUIE-RE dataset in terms of the overall relation F1 score, in this section, we take a deeper look and further investigate whether OntoRE's integration of domain knowledge essentially improves the model's ability to comprehend domain-specific information.

We observe that the SSUIE-RE dataset includes entity types with varying levels of specialization, ranging from highly specialized entity types (such as *Space Mission*, *Experimental Platform* and *Space Science Field*, etc.) to more general entity types (such as *Person*, *Location*, and *Organisation*, etc.). We refer to the former as in-domain entity types and the latter as general entity types. According to the degree of domain specificity, we categorize 15 out of the 19 entity types defined in the SSUIE-RE dataset as in-domain entity types, and the remaining 4 as general entity types, as shown in Table 2. Based on this categorization, in-domain entities account for 68% of the total entities in the SSUIE-RE dataset, while general entities account for 32%.

Domain Specificity	Entity Types
In-domain (68%)	*Space Mission, Space Station Segment, Space Science Field, Prize, Experimental Platform, Experimental Platform Support System, Experimental System, Experimental System Module, Patent, Criterion, Experimental Project, Experimental Data, Academic Paper, Technical Report, Research Team*
General (32%)	*Organisation, Person, Time, Location*

Table 2: We divide entity types into in-domain and general entity types according to the degree of domain specificity.

To more accurately evaluate OntoRE's ability to understand domain-specific information, we further differentiate the domain specificity of relation types. A triple type defined in the dataset is composed of a head entity type, a relation type, and a tail entity type in the form of (s, r, o). We assess the degree of domain specificity of a relation type by determining whether the head and tail entities it connects are classified as in-domain entity types, as listed in Table 2. Specifically, we consider a relation to be highly domain-specific when both the head and tail entity types are in-domain. If only one of the two entity types is in-domain and the other is general, the corresponding relation is considered to exhibit weaker domain specificity. Furthermore, relations with head and tail entity types are both general rather than in-domain

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

entity types, are considered to exhibit the weakest domain specificity.

We compare our model with the baseline MGE on the performance of recognizing in-domain and general entities, respectively. And for relation extraction, we compare the performance of our model and baseline in extracting relation types with varying degrees of domain specificity. The experimental results are shown in Table 3 and Table 4.

As shown in Table 3, OntoRE outperforms the baseline in recognizing in-domain and general entity types, with a respective improvement of +0.4% and +0.8% in terms of entity F1 score. Table 4 demonstrates that OntoRE obtains an absolute relation F1 score improvement of +0.5%, +1.5% and 1.7% respectively, as the domain specificity of the relation types increases. The experimental results show that OntoRE improves the performance of extracting relation types with varying degrees of domain specificity, and the benefit of ontology knowledge augmentation is more evident for relations with higher domain specificity. This indicates that the incorporation of ontology knowledge appears to be an effective approach for enhancing the model's ability to understand domain-specialized information, while not weakening its understanding of general information.

Entity Type	**Model**	**NER**		
		Prec.(%)	Rec.(%)	F1(%)
In-domain (68%)	Baseline	73.4	73.0	73.2
	OntoRE	72.0	75.3	**73.6 (+0.4)**
General (32%)	Baseline	79.6	81.6	80.6
	OntoRE	79.8	83.0	**81.4 (+0.8)**

Table 3: NER results of in-domain and general entity types on SSUIE-RE test set. In-domain entities account for 68% in the dataset, and general entities account for 32%.

Relation Type	**Model**	**RE**		
		Prec.(%)	Rec.(%)	F1(%)
IDE = 0 (26%)	Baseline	68.5	72.2	70.3
	OntoRE	69.7	71.8	**70.8 (+0.5)**
IDE = 1 (11%)	Baseline	55.5	42.7	48.2
	OntoRE	60.5	42.2	**49.7 (+1.5)**
IDE = 2 (63%)	Baseline	56.8	64.6	60.4
	OntoRE	59.3	65.3	**62.1 (+1.7)**

Table 4: RE results of relation types with varying degrees of domain specificity on SSUIE-RE test set. IDE (In-Domain Entity) represents the number of in-domain entity types contained in a triple type according to ontology definition. The proportions of relations with IDE=0, IDE=1, and IDE=2 in the SSUIE-RE dataset are 26%, 11%, and 63%, respectively.

4.6 Ablation Study

In this section, we conduct ablation study on the SSUIE-RE dataset to examine the effectiveness of our model, specifically with regard to three factors: knowledge source, knowledge fusion method, and the number of knowledge encoder layers.

4.6.1 Knowledge Source and Fusion Method

We put the two factors of knowledge source and knowledge fusion method together for experimental analysis. For the aspect of knowledge source, we remove the entity type descriptions (denoted as *Des* in Table 5) from the complete OntoRE framework to examine the role of entity type descriptions in

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

knowledge enhancement. For knowledge fusion method, we examine the effects of three fusion methods: appending, concatenation and addition.

Table 5 presents a comparison of the experimental results for different combinations of these factors on the SSUIE-RE dataset. The experimental results show that, under the fusion methods of appending and concatenation, the incorporation of entity type descriptions improves NER F1 scores by 1.4% and 1.4% respectively. However, under the addition fusion method, there is a slight decrease in NER F1 score. This can be attributed to the need for compressing the dimension of the entity type description tensor to match the input sentence tensor before addition, leading to information loss. Across all three fusion methods, the inclusion of entity type descriptions consistently improve the relation F1 scores. Additionally, when using the same combination of knowledge sources, the performance of the appending and concatenation fusion methods is comparable, while the addition fusion method achieves the best relation F1 score. This suggests that the optimal approach is to employ ontology and entity type descriptions as knowledge sources and use the addition fusion method to integrate knowledge representations into the model.

Knowledge Source	**Fusion Method**	**NER**			**RE**		
		Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
Ontology	Append	71.3	77.9	74.5	61.0	62.2	61.6
Ontology	Concat.	72.9	78.3	75.5	61.3	64.4	62.8
Ontology	Add	74.2	79.4	76.7	62.2	64.4	63.3
Ontology + *Des*	Append	73.3	78.8	75.9	60.6	65.4	62.9
Ontology + *Des*	Concat.	75.7	78.0	76.9	63.0	63.0	63.0
Ontology + *Des*	Add	75.0	78.3	76.6	62.4	64.5	**63.4**

Table 5: Ablation study on SSUIE-RE for knowledge source and knowledge fusion method. *Des* denotes entity type descriptions.

4.6.2 Number of Knowledge Encoder Layers

In the knowledge encoding stage, we utilize Transformer encoder to encode the serialized ontology knowledge, as described in Section 3.3. We conduct ablation study to investigate whether stacking multiple layers of encoders could improve the model performance. Considering the parameter size of Transformer encoder, we only experiment with encoder layers up to three. As shown in Table 6, using two layers of Transformer encoders achieved the best performance (which we employed in our final model), and further stacking of encoders does not result in additional performance improvements.

Knowledge Encoder Layers	**NER**			**RE**		
	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)
L = 1	70.6	80.8	75.3	58.4	64.9	61.5
L = 2	75.0	78.3	76.6	62.4	64.5	**63.4**
L = 3	75.0	77.7	76.4	61.1	62.0	61.5

Table 6: Ablation study on SSUIE-RE for the number of knowledge encoder layers.

5 Conclusion

In this work, we propose an ontology-enhanced method for joint entity and relation extraction in space science and utilization domain. Our model utilizes ontology and entity type descriptions as sources of domain knowledge, and incorporate them into downstream model to enhance model's comprehension of domain-specific information. We introduce a new dataset, SSUIE-RE, which contains rich domain-specialized knowledge. Experimental results on SSUIE-RE demonstrate that our approach outperforms previous state-of-the-art methods. Moreover, we conduct a detailed analysis of the extraction of entities

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

and relations with different degrees of domain specificity and validate the effectiveness of ontology knowledge enhancement. Overall, our proposed method provides a promising direction for improving the performance of entity and relation extraction in specialized domains with limited resources. In the future, we would like to further explore how to generalize the ontology knowledge enhancement idea to other domain-specific information extraction tasks.

Acknowledgements

This work was supported by the National Defense Science and Technology Key Laboratory Fund Project of the Chinese Academy of Sciences: Space Science and Application of Big Data Knowledge Graph Construction and Intelligent Application Research and Manned Space Engineering Project: Research on Technology and Method of Engineering Big Data Knowledge Mining.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November. Association for Computational Linguistics.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560, Portland, Oregon, USA, June. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE Transactions on Audio, Speech and Language Processing*.

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

Matthew R. Gormley, Mo Yu, and Mark Dredze. 2015. Improved relation extraction with feature-rich compositional embedding models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1774–1784, Lisbon, Portugal, September. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6248–6260, Online, August. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, Baltimore, Maryland, June. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar, October. Association for Computational Linguistics.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of AAAI*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

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

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November. Association for Computational Linguistics.

Shaolei Wang, Yue Zhang, Wanxiang Che, and Ting Liu. 2018. Joint extraction of entities and relations based on a novel graph scheme. In *IJCAI*, pages 4461–4467. Yokohama.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online, July. Association for Computational Linguistics.

Xiong Xiong, Liu Yunfei, Liu Anqi, Gong Shuai, and Li Shengyang. 2022. A multi-gate encoder for joint entity and relation extraction. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 848–860, Nanchang, China, October. Chinese Information Processing Society of China.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 185–197, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 987–991, Online, August. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland, May. Association for Computational Linguistics.

Xiaofeng Yu and Wai Lam. 2010. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *Coling 2010: Posters*, pages 1399–1407, Beijing, China, August. Coling 2010 Organizing Committee.

Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Yubin Wang, Tingwen Liu, Bin Wang, and Sujian Li. 2020. Joint extraction of entities and relations based on a novel decomposition strategy. In *Proceedings of ECAI*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia, July. Association for Computational Linguistics.

Sheng Zhang, Patrick Ng, Zhiguo Wang, and Bing Xiang. 2022. Reknow: Enhanced knowledge for joint entity and relation extraction. In *NAACL 2022 Workshop on SUKI*.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada, July. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June. Association for Computational Linguistics.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 713-725, Harbin, China, August 3 - 5, 2023.

Document Information Extraction via Global Tagging

Shaojie He[1,2], **Tianshu Wang**[2], **Yaojie Lu**[2], **Hongyu Lin**[2]*, **Xianpei Han**[2]*, **Yingfei Sun**[1], **Le Sun**[2]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Chinese Information Processing Laboratory,
Institute of Software, Chinese Academy of Sciences, Beijing, China
{heshaojie2020,tianshu2020,luyaojie,hongyu,xianpei}@iscas.ac.cn
yfsun@ucas.ac.cn, sunle@iscas.ac.cn

Abstract

Document Information Extraction (DIE) is a crucial task for extracting key information from visually-rich documents. The typical pipeline approach for this task involves Optical Character Recognition (OCR), serializer, Semantic Entity Recognition (SER), and Relation Extraction (RE) modules. However, this pipeline presents significant challenges in real-world scenarios due to issues such as unnatural text order and error propagation between different modules. To address these challenges, we propose a novel tagging-based method – Global TaggeR (GTR), which converts the original sequence labeling task into a token relation classification task. This approach globally links discontinuous semantic entities in complex layouts, and jointly extracts entities and relations from documents. In addition, we design a joint training loss and a joint decoding strategy for SER and RE tasks based on GTR. Our experiments on multiple datasets demonstrate that GTR not only mitigates the issue of text in the wrong order but also improves RE performance.

1 Introduction

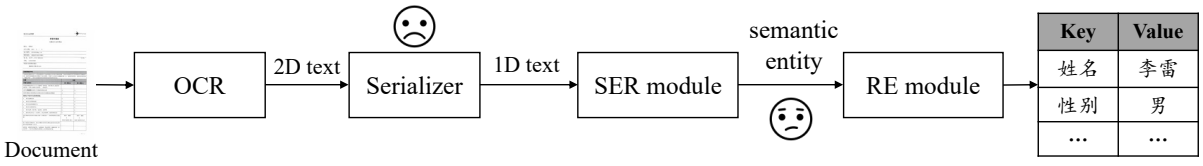
Document Information Extraction (DIE), which is to extract key information from document with complex layouts, has become increasingly important in recent years (Zhang et al., 2022; Hong et al., 2022). It not only enables us to efficiently compress document data, but also facilitates the retrieval of important information from documents. A typical pipeline approach for the DIE task is depicted in Figure 1(a) (Denk and Reisswig, 2019; Hwang et al., 2021a). First, the document with complex layout is converted into text blocks using Optical Character Recognition (OCR) tools. Next, the serializer module organizes these text blocks into a more appropriate order. Finally, the well-ordered text blocks are input sequentially into the Semantic Entity Recognition (SER) and Relation Extraction (RE) modules to extract key-value pairs.

However, the pipeline approach in Figure 1(a) presents significant challenges in real-world scenarios. (1) Mainstream models for the DIE task, such as LayoutLM (Xu et al., 2020), LayoutLMv2 (Xu et al., 2021) and LayoutXLM (Xu et al., 2022), usually use sequence labeling in the Beginning-Inside–Outside (BIO) tagging schema, which assume that tokens belonging to the same semantic entity are grouped together after serialization. If the serializer module fails to order the text blocks correctly, the final performance can be severely impacted. A potential solution is to train a strong and robust serializer module, but this is difficult due to the labor-intensive labeling process under rich and diverse styles of documents; (2) In addition to the issue of text order, this pipeline also suffers from error propagation when using a SER module and a RE module. In research settings, the results of the SER and RE tasks are generally tested separately, with the ground truth of the SER results being used as default auxiliary information for the RE task. However, in real-world scenarios, the SER module in the pipeline cannot provide 100% accurate results, which ineluctably leads to error propagation on RE performance.

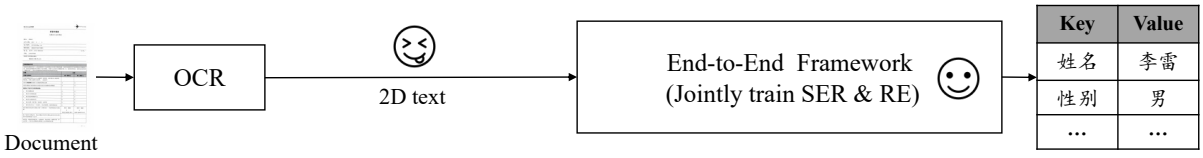
Researchers have explored alternative methods for modeling OCR results directly without serializer module to tackle the issue of text in the wrong order. Some have utilized graph convolution networks to

*Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China



(a) The pipeline approach in the DIE task.



(b) The end-to-end approach in this paper.

Figure 1: A comparison between (a) current pipeline approach and (b) our end-to-end approach.

model the relationships between tokens (Yu et al., 2020; Lee et al., 2022; Wei et al., 2020) . Others have converted the DIE task into a parsing problem, modeling tree structure for the document (Hwang et al., 2021b; Mathur et al., 2023). Besides, generative encoder-decoder frameworks are applied to avoid the weakness of the BIO tagging schema essentially (Kim et al., 2022). While these methods can mitigate the problem of text in the wrong order, they still face challenges. For example, graph-based methods require a more delicate model design, and generative models are usually difficult to train and require a great amount of document data for pre-training.

To address abovementioned two problems, we propose a simple yet effective method named Global TaggeR (GTR). Our approach is inspired by Wu et al. (2020), which converts the original sequence labeling task into a token relation classification task. For the SER task, we tag all token pairs and design a decoding strategy based on disjoint sets to decode the semantic entities. And we find GTR naturally resistant to wrong text order to a certain extent. For example, there is a document fragment "登记表姓名李雷性别男" and we tag the token pair {姓, 名} so that we know "姓名" is a semantic entity. Even if we shuffle this fragment to "登记表姓李性男名雷别", we still can know "姓名" is a semantic entity using the same tag {姓, 名}. In other words, GTR enables us to recognize discontinuous semantic entities, regardless of text in the wrong order. Additionally, for the RE task, we combine RE and SER tags for joint training and extend the decoding strategy for joint decoding. The pipeline of this study is depicted in Figure 1(b). We remove the serializer module from the original pipeline to make it easier and propose an end-to-end extraction framework for jointly training the SER and RE tasks to prevent error propagation problem. The contributions of this work are summarized as follows:

- We propose an end-to-end extraction framework for the document information extraction, which simplifies the traditional pipeline approach and alleviates error propagation issues.
- In this end-to-end extraction framework, we propose the Global TaggeR (GTR) method, which contains a global tagging schema and a joint decoding strategy for the SER and RE tasks.
- Our experiments on multiple datasets demonstrate that the GTR proposed not only mitigates the issue of text in the wrong order but also facilitates the interaction of entity and relation information, resulting in improvement of RE performance.

2 Background

2.1 Task Definition

Given a document image I and its OCR results that containing a sequence of tokens $S = \{t_1, ..., t_n\}$ paired with corresponding bounding boxes $L = \{b_1, ..., b_n\}$, the goal of the DIE task is to extract a set of entities $E = \{e_1, ..., e_m\}$ in the document and their corresponding relations $R = \{(e_i, e_j)\}$. We usually divide the DIE task into two sub-tasks named SER and RE. For the SER task, we try to recognize

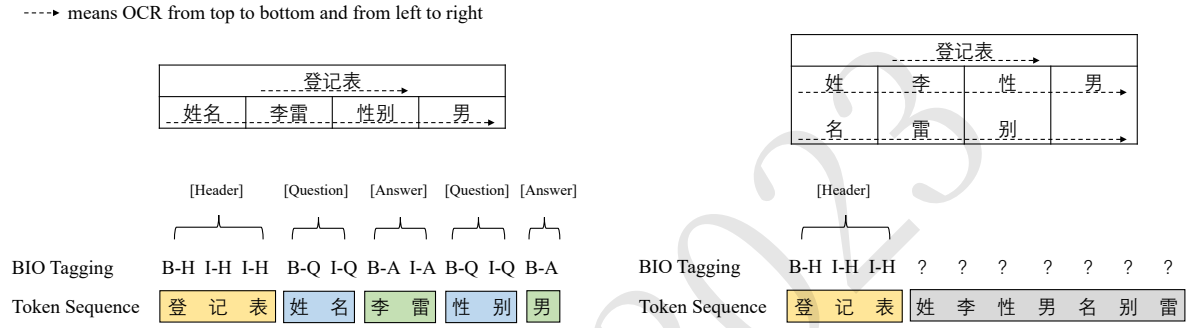
Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

all possible semantic entities in token sequence S and classify them with three entity types {[Header], [Question], [Answer]}. For the RE task, based on semantic entities that we have recognized, we match each two of them if they are question-answer pairs, or key-value pairs. The relations only have two types, paired or not.

2.2 LayoutXLM

We choose LayoutXLM (Xu et al., 2022) as our baseline model, which is a multilingual and multi-modal pre-trained language model designed with a single encoder architecture. The model first feeds token sequence S and bounding box sequence L, along with visual features extracted from document image I. Next, it adopts visual and text embedding, position embedding and layout embedding as the representation of tokens, and then employs multi-modal Transformer encoder layers to generate the representations of the given tokens $H = \{h_1, ..., h_n\}$. Finally, a simple classifier is connected to the encoder, enabling it to perform downstream SER and RE tasks.

2.3 BIO Tagging Schema



(a) The BIO tagging schema successfully tags when the input text is well-ordered.

(b) The BIO tagging schema fails to tag when the order of the input text is incorrect.

Figure 2: The illustration of the BIO tagging schema.

The BIO tagging schema, which is a popular sequence labeling technique, is widely used for the SER task. In this schema, each token in the document is labeled with a prefix that indicates whether it is the beginning (B), inside (I), or outside (O) of an entity span. Figure 2(a) provides a simple illustration. However, layout-rich documents often result in OCR text in the wrong order. Given text in the wrong order, the BIO tagging schema cannot express span boundaries correctly, illustrated in Figure 2(b). Therefore, it is necessary to find new approaches to tackle this issue.

3 Approach

In this section, we introduce our GTR approach in four parts. First, we propose the global tagging schema of the DIE task. Next, a token pair scoring layer added to baseline model is proposed. Then, we design a corresponding decoding strategy to decode entities and relations from the predicted tagging matrix. Finally, we introduce our training loss for jointly training SER and RE tasks.

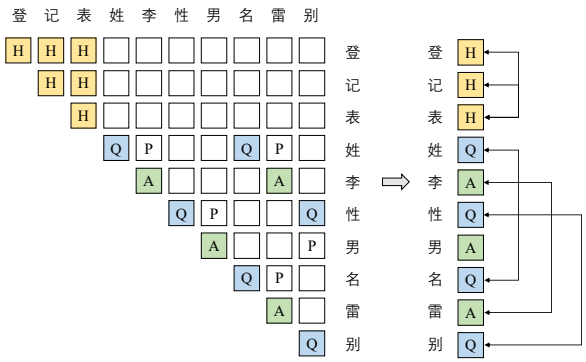
3.1 Global Tagging Schema

For the DIE task, we use five tags {O, H, Q, A, P} to represent relations between token t_i and t_j. Table 1 shows the meanings of these five tags.

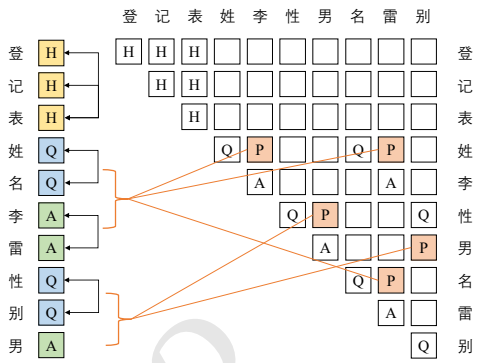
Figure 3(a) illustrates the global tagging schema tags entities that are difficult to tag using the BIO tagging schema in Figure 2(b). Tokens in the same semantic entity are tagged with the same label pairwise. The labels are {H, Q, A}, representing the entity types {[Header], [Question], [Answer]}, respectively. For example, in Figure 3(a), the token pair {姓, 名} = Q means that the tokens "姓" and "名" belong to the same entity span, and the entity type is [Question]. Similarly, {登, 记, 表} belongs

Tags	Meanings
H	Token t_i and t_j belong to the same entity span, and the entity type is [Header].
Q	Token t_i and t_j belong to the same entity span, and the entity type is [Question].
A	Token t_i and t_j belong to the same entity span, and the entity type is [Answer].
P	Token t_i and t_j belong to two paired entities, with the types of [Question] and [Answer].
O	No above four relations for token t_i and t_j.

Table 1: The meanings of tags for the DIE task.



(a) SER task. The left is the tagging matrix and the right is the corresponding semantic entities.



(b) RE task. The right is the tagging matrix and the left is the rearranged semantic entities.

Figure 3: The illustration of global tagging schema for jointly labeling (a) SER and (b) RE tasks. We only display the upper triangular of tagging matrix on account of its symmetry.

to the [Header] entity, {性, 别} belongs to the [Question] entity, and {李, 雷}, {男} belong to the [Answer] entity.

Figure 3(b) illustrates the global tagging schema tags relations after tagging entities. For each [Question]-[Answer] (QA) relation in the document, tokens from the two associated entities, are tagged with the same label P pairwise. For example, given the premise that {姓, 名} belongs to [Question] entity and {李, 雷} belongs to [Answer] entity, the token pairs {姓, 李}, {姓, 雷}, {名, 雷} = P, indicating that {姓, 名} and {李, 雷} are paired QA relation. Similarly, {性, 别} and {男} are paired QA relation.

The global tagging schema offers two primary advantages in the DIE task. (1) First, it allows for the tagging of discontinuous semantic entity spans. Due to the diversity of document layouts, the token sequence produced by OCR tools is usually in an incorrect order. Even if the tokens in the same semantic entity span are discontinuous in the token sequence, they can still be tagged using this global tagging schema. (2) Second, it supports joint training of the SER and RE tasks. Using the global tagging schema, the SER task can be expanded to token-to-token relationship classification task. This schema unifies task format and enables unified modeling and joint training for the SER and RE tasks.

3.2 Token Pair Scoring

For the representations $H = \{h_1, ..., h_n\}$ generated from given token sequence S, we employ simple linear transformation and multiplication operation to obtain the global score $s_{ij|c}$ of token pair t_i and t_j classified to class c:

$$q_{i,c} = W_{q,c}h_i + b_{q,c} \tag{1}$$

$$k_{j,c} = W_{k,c}h_j + b_{k,c} \tag{2}$$

$$s_{ij|c} = (\mathcal{R}_i q_{i,c})^T (\mathcal{R}_j k_{j,c}) \tag{3}$$

where $q_{i,c}$ and $k_{j,c}$ are intermediate representations created by linear transformation operation. $\mathcal{R}$ is a rotary position embedding (Su et al., 2021), which helps to embed relative position information and accelerate training process.

During the training stage, we directly use $s_{ij|c}$ to compute loss function. For the supervision signal of $s_{ij|c}$, we assign signal 1 to represent {H, Q, A, P} tags and signal -1 to represent the absence of any of the above four relations. Therefore, during the inference stage, we obtain the predicted tagging matrix by processing $s_{ij|c}$ with a threshold of 0, where values greater than 0 are regarded as tags.

3.3 Decoding Strategy

With the predicted tagging matrix, we design a decoding strategy to extract the semantic entities and relations, as shown in Algorithm 1. Following the proposed decoding strategy, we decode in two steps:

Algorithm 1 Decoding Strategy for DIE

Input: The predicted tagging matrix T. The predicted tag of token pair t_i and t_j is denoted as $T(t_i, t_j)$. The predicted tag of token pair t_i and t_i is abbreviated as $T(t_i)$. If all tokens in a set e share the same tag, abbreviated as $T(e)$.

Output: Entity set E and relation set R.

Initialize the entity set E and relation set R with $\varnothing$, and $n \leftarrow len(S)$.
while $i \leq n$ **do**
  if $T(t_i) \in \{\mathrm{H}, \mathrm{Q}, \mathrm{A}\}$ **then**
    $E \leftarrow E \cup \{t_i\}$
  end if
end while
while $i \leq n$ and $j \leq n$ **do**
  if $i \neq j$ **and** $T(t_i, t_j) \in \{\mathrm{H}, \mathrm{Q}, \mathrm{A}\}$ **and** $T(t_i, t_j) = T(t_i) = T(t_j)$ **then**
    $E \leftarrow$ Merge the set where t_i resides and the set where t_j resides in E.
  end if
end while
while $e_i \in E$ and $e_j \in E$ **do**
  if $T(e_i) = \mathrm{Q}$ and $T(e_j) = \mathrm{A}$ and any $T(t_k, t_l) = \mathrm{P}$ that $t_k \in e_i$ and $t_l \in e_j$ **then**
    $R \leftarrow R \cup \{(e_i, e_j)\}$
  end if
end while
return the set E and the set R

SER. Firstly, we recognize the diagonal tags, and use these tags to label the token sequence S. Then, we recognize the non-diagonal tags belonging to {H, Q, A}, and use these tags for merging tokens. Iterating through these tags, we use a disjoint set algorithm with additional judgement to merge semantic entity tokens. Therefore, we can extract the semantic entity set E.

RE. Using the semantic entity set E, we iterate through all possible `[Question]`-`[Answer]` entity pairs. If there exists any token pair t_k and t_l that t_k in an entity e_i with type `[Question]` and t_l in an entity e_j with type `[Answer]` and $T(t_k, t_l)$ is tagged with label P, we add (e_i, e_j) into relation set R. Finally, we can extract the semantic entity set E as well as the relation set R.

3.4 Training Loss

For token pair t_i and t_j, we denote y_{ij} as the ground truth tag and $P_{ij}(\hat{y} = k)$ as the predicted probability for class k. A cross entropy loss is applied:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in C}\mathbb{I}(y_{ij} = k)\log P_{ij}(\hat{y} = k), \quad P_{ij}(\hat{y} = k) = \frac{e^{s_{ij|k}}}{\sum\limits_{k'\in C} e^{s_{ij|k'}}} \tag{4}$$

where $\mathbb{I}$ is an indicator function and C is the label set {H, Q, A, P, O}. And $s_{ij|k}$ denotes the predicted score for token pair t_i and t_j classified to class k.

We attempt to train the baseline model using the above loss function but fail due to convergence issues. And the training results always output O tags. We suggest that our global tagging schema requires the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

prediction of a probability matrix of $n*n$, which results in very sparse supervised signals, facing a severe class imbalance problem, and making it challenging to train the model effectively. Inspired by Su et al. (2022), we improve $\log P_{ij}(\hat{y}=k)$ with a class imbalance likelihood:

$$\log P_{ij}(\hat{y}=k) = \log(1+e^{-s_{ij|k}}) + \log(1+\sum_{k' \in C, k' \neq k} e^{s_{ij|k'}}) \quad (5)$$

which turns loss into a pairwise comparison of target category scores and non-target category scores.

4 Experiments

4.1 Experimental Setup

Dataset. We use FUNSD (Jaume et al., 2019) and XFUN (Xu et al., 2022) datasets to evaluate our proposed approach. (1) FUNSD is an English dataset for document understanding, comprising 199 annotated documents. The dataset is split into a training set of 149 documents and a testing set of 50 documents; (2) XFUN is a multilingual dataset for document understanding that comprises seven languages [Chinese (ZH), Japanese (JA), Spanish (ES), French (FR), Italian (IT), German (DE), Portuguese (PT)], totaling 1,393 annotated documents. Each language's data has separate training and testing sets, with 199 and 50 documents respectively.

Parameter Settings. For training, we follow the hyper-parameter settings of Xu et al. (2022), setting the learning rate to 5e-5 and the warmup ratio to 0.1. The max length of input token sequence is set to 512, which means a split of chunk size 512 if the input token sequence is too long. For a fair comparison, we set the batch size to 64 and run the training for 2000 steps to ensure that the models have well converged.

Input Settings. Golden input and OCR input are two types of input text order for experiment input settings. (1) Golden input means that we concatenate the ground truth text blocks into a token sequence and feed it into the model, which implies that all semantic entity spans are continuous. (2) OCR input means that we concatenate all tokens following the recognition pattern of a common OCR from top to bottom and left to right before feeding them into the model. This implies that under complex layouts, the same semantic entity span may be discontinuous.

Evaluation Metrics. For evaluation, we use F1-score on two sub-tasks: (1) Semantic Entity Recognition (SER), where semantic entities are identified by tagging as either {`[Header]`, `[Question]`, `[Answer]`}. When the entity type and all entity tokens are correct, the entity is regarded as a correct entity. (2) Relation Extraction (RE), where paired relation of question and answer entities are identified. We use a strict evaluation metrics that only the paired two entities are exactly correct at the token-level, the relation is regarded as a correct relation.

Baseline Model. We use LayoutXLM$_{\text{BASE}}$ model as the baseline model. Its original RE results are tested based on the given ground truth semantic entities. To test the RE results in the pipeline for baseline model, we first reproduce the results of Xu et al. (2022) and then re-test the RE results using the semantic entities generated by its SER module.

4.2 Result

We evaluate the baseline model with the BIO tagging and the global tagging on language-specific fine-tuning settings (training on X, and testing on X).

Table 2 presents the results under Golden input settings. We compare our global tagger approach with the reproduced baseline. The results show that our global tagger method outperforms the baseline model on average F1-score of the 8 languages for the SER task. Moreover, when combining the SER and RE tasks in an end-to-end extraction framework, the RE performance of average F1-score significantly surpassed that of the baseline model pipelined, and is even higher on two languages compared with the baseline model using ground truth semantic entity information .

Table 3 presents the result under OCR input settings. We directly use the baseline model trained under Golden input settings to predict the SER and RE results for evaluating the BIO tagging schema. We observe that the SER performance on average F1-score of the 8 languages for the baseline model is

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	BIO♣	0.7940	0.8924	0.7921	0.7550	0.7902	0.8082	0.8222	0.7903	0.8056
	BIO	0.8013	**0.8944**	0.7864	0.7426	0.7852	0.8073	0.7951	0.7848	0.7996
	GTR	**0.8079**	0.8818	**0.7972**	**0.7631**	**0.8067**	**0.8210**	**0.8032**	**0.8071**	**0.8110**
gtSER+RE	BIO♣	0.5483	0.7073	0.6963	0.6896	0.6353	0.6415	0.6551	0.5718	0.6432
	BIO	0.5560	0.7047	0.6519	0.7041	0.6664	0.6725	0.6485	0.5893	0.6492
SER+RE	BIO	0.4340	0.5965	0.5082	0.498	0.5064	0.4861	0.4258	0.3765	0.4789
	GTR	**0.5910**	**0.7739**	**0.6470**	**0.5363**	**0.6063**	**0.6594**	**0.5531**	**0.5247**	**0.6115**

Table 2: Main result under Golden input settings. ♣: results reported in Xu et al. (2022). Best results are in **bold** comparing reproduced BIO tagging (abbreviated as BIO) with our global tagger (abbreviated as GTR). gtSER+RE denotes evaluating the RE results using ground truth SER results. And SER+RE denotes evaluating the RE results using SER results of the model.

	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	BIO	0.5735	0.3970	0.4017	0.6287	0.6916	0.7055	0.6823	0.6863	0.5958
	GTR	**0.7412**	**0.8444**	**0.7205**	**0.7165**	**0.7676**	**0.7772**	**0.7508**	**0.7811**	**0.7624**
SER+RE	BIO	0.2712	0.1441	0.1759	0.3665	0.4141	0.4206	0.3566	0.3086	0.3072
	GTR	**0.5828**	**0.6920**	**0.5427**	**0.5686**	**0.5712**	**0.5888**	**0.5933**	**0.5580**	**0.5872**

Table 3: Main result under OCR input settings. Best results are in **bold** comparing reproduced BIO tagging (abbreviated as BIO) with our global tagger (abbreviated as GTR).

significantly impacted, making it difficult to perform the RE process based on its SER results. However, with joint training and decoding using our global tagger approach, we are able to alleviate this issue.

4.3 Analysis

4.3.1 Golden Input vs. OCR Input

The BIO tagging schema requires well-ordered input, while the global tagging schema accepts unordered input. Comparing the average F1-score of SER performance in Table 2 and Table 3, we observe a significant drop from 0.7996 to 0.5958 when changing Golden input settings into OCR input, indicating a great impact by the order of input tokens using the BIO tagging schema. On the other hand, under the global tagging schema, we find the model's average SER performance only drops from 0.8110 to 0.7624 between Golden input settings and OCR input settings, demonstrating that the global tagging schema can effectively alleviate suboptimal input token order issue.

4.3.2 Pipeline Framework vs. End-to-End Framework

The pipeline framework with the BIO tagging schema suffers from error propagation, while the end-to-end GTR method can greatly mitigate it. In Table 2, we observe a drop of average F1-score on the RE results from 0.6492 to 0.4789 when combining the SER and RE modules in the pipeline, demonstrating that pipeline framework can greatly impact performance. Particularly, when both SER and RE modules have poor performance under OCR input settings, we observe a terrible performance, which is only 0.3072 average F1-score on the RE task. In such case, joint training and decoding in GTR method can significantly alleviate error propagation issue with the average F1-score of 0.5872 rather than 0.3072 of 8 language datasets on the RE task.

Besides, in Table 2, the SER+RE results using GTR approach are even higher than the baseline RE results using ground truth semantic entities on English(FUNSD) and Chinese(ZH) language datasets, indicating that the end-to-end extraction framework is potential for facilitating the interaction of entity and relation information, resulting in better RE performance.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

5 Related Work

In recent years, benefited from pre-training and fine-tuning paradigm, information extraction for documents has gained significant attention in both research and industry (Li et al., 2021b; Li et al., 2021a; Appalaraju et al., 2021; Wang et al., 2022; Li et al., 2022; Sun et al., 2023). However, there are still numerous challenges in the pipeline when applied in real-world scenarios. Addressing the issue of text order in the pipeline, related works are organized into three perspectives.

5.1 Sequence-based Perspective

Sequence-based models, such as LayoutLM (Xu et al., 2020) and LayoutLMv2 (Xu et al., 2021), aim to encode serialized token sequence from complex and diverse document, integrating layout, font, and other features. These models offer several advantages, such as simplicity, scalability, and suitability for Masked Language Modeling (MLM) pre-training. However, these models are constrained by the traditional BIO tagging mode and require a well-ordered token sequence as a basis.

5.2 Graph-based Perspective

Graph-based models usually treat tokens as nodes in a graph and allow interactions between tokens explicitly to enhance their representations (Yu et al., 2020; Lee et al., 2022; Wei et al., 2020). Even though these models leverage the graph structure to capture more complex relationships between entities, they still use the BIO paradigm for SER task. Alternatively, some works, like SPADE (Hwang et al., 2021b), take a different approach by converting DIE task into document parsing task. It models the document as a dependency tree to represent entities and relations.

Our work in this paper also lies in graph-based perspectives. Similar to the tack of SPADE that converting the DIE task to a different task, we view the DIE task as a token relation classification task. But unlike SPADE, we do not utilize a graph generator and graph decoder. Rather, we simply modify the tagging schema and do not change the encoding model.

5.3 End-to-End Perspective

End-to-End model typically combines the entire pipeline into one model. Dessurt (Davis et al., 2022), TRIE++ (Cheng et al., 2022), for example, unify OCR, reordering, and extraction into a single model. Meanwhile, models like Donut (Kim et al., 2022), GMN (Cao et al., 2022), use a generative encoder-decoder architecture to unify OCR and generation. In contrast to extraction-based works, they directly generate the structured output, making it more flexible for varying output formats.

6 Conclusion

In this paper, we propose an end-to-end approach named global tagger to solve the document information extraction task. Experiments on the FUNSD and XFUN datasets demonstrate its efficacy in effectively mitigating the gap between token order in OCR input and golden input. Furthermore, our experimental results indicate that joint training and decoding of semantic entity recognition and relation extraction tasks in this end-to-end extraction framework can alleviate the negative impact of error propagation and improve the performance of the relation extraction results.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. U1936207, 62122077 and 62106251.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 973–983. IEEE.

Haoyu Cao, Jiefeng Ma, Antai Guo, Yiqing Hu, Hao Liu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. GMN: generative multi-modal network for practical document information extraction. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3768–3778. Association for Computational Linguistics.

Zhanzhan Cheng, Peng Zhang, Can Li, Liang Qiao, Yunlu Xu, Pengfei Li, Shiliang Pu, Yi Niu, and Fei Wu. 2022. TRIE++: towards end-to-end information extraction from visually rich documents. *CoRR*, abs/2207.06744.

Brian L. Davis, Bryan S. Morse, Brian L. Price, Chris Tensmeyer, Curtis Wigington, and Vlad I. Morariu. 2022. End-to-end document recognition and understanding with dessurt. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13804 of *Lecture Notes in Computer Science*, pages 280–296. Springer.

Timo I. Denk and Christian Reisswig. 2019. BERTgrid: Contextualized embedding for 2d document representation and understanding. In *Workshop on Document Intelligence at NeurIPS 2019*.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10767–10775.

Wonseok Hwang, Hyunji Lee, Jinyeong Yim, Geewook Kim, and Minjoon Seo. 2021a. Cost-effective end-to-end information extraction for semi-structured document images. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3375–3383. Association for Computational Linguistics.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021b. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343, Online. Association for Computational Linguistics.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3735–3754. Association for Computational Linguistics.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. Structurallm: Structural pre-training for form understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6309–6318. Association for Computational Linguistics.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021b. Structext: Structured text understanding with multi-modal transformers. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1912–1920. ACM.

Qian Li, Hao Peng, Jianxin Li, Jia Wu, Yuanxing Ning, Lihong Wang, Philip S. Yu, and Zheng Wang. 2022. Reinforcement learning-based dialogue guided event extraction to exploit argument relations. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:520–533.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

Puneet Mathur, Rajiv Jain, Ashutosh Mehra, Jiuxiang Gu, Franck Dernoncourt, Anandhavelu Natarajan, Quan Hung Tran, Verena Kaynig-Fittkau, Ani Nenkova, Dinesh Manocha, and Vlad I. Morariu. 2023. Layerdoc: Layer-wise extraction of spatial hierarchical structure in visually-rich documents. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 3599–3609. IEEE.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *CoRR*, abs/2208.03054.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2023. Learning implicit and explicit multi-task interactions for information extraction. *ACM Trans. Inf. Syst.*, 41(2):27:1–27:29.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7747–7757. Association for Computational Linguistics.

Mengxi Wei, Yifan He, and Qiong Zhang. 2020. Robust layout-aware IE for visually rich documents with pre-trained language models. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2367–2376. ACM.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, Dublin, Ireland. Association for Computational Linguistics.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2020. PICK: processing key information extraction from documents using improved graph learning-convolutional networks. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 4363–4370. IEEE.

Zhenyu Zhang, Bowen Yu, Haiyang Yu, Tingwen Liu, Cheng Fu, Jingyang Li, Chengguang Tang, Jian Sun, and Yongbin Li. 2022. Layout-aware information extraction for document-grounded dialogue: Dataset, method and demonstration. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 7252–7260. ACM.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 726-735, Harbin, China, August 3 - 5, 2023.

A Distantly-Supervised Relation Extraction Method Based on Selective Gate and Noise Correction

Zhuowei Chen[1], **Yujia Tian**[1], **Lianxi Wang**[✉1,2], **Shengyi Jiang**[1,2]
[1]School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China, 510006
[2]Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangzhou, China, 510006
{20211003051,20211003065,wanglianxi}@gdufs.edu.cn
jiangshengyi@163.com

Abstract

Entity relation extraction, as a core task of information extraction, aims to predict the relation of entity pairs identified by text, and its research results are applied to various fields. To address the problem that current distantly supervised relation extraction (DSRE) methods based on large-scale corpus annotation generate a large amount of noisy data, a DSRE method that incorporates selective gate and noise correction framework is proposed. The selective gate is used to reasonably select the sentence features in the sentence bag, while the noise correction is used to correct the labels of small classes of samples that are misclassified into large classes during the model training process, to reduce the negative impact of noisy data on relation extraction. The results on the English datasets clearly demonstrate that our proposed method outperforms other baseline models. Moreover, the experimental results on the Chinese dataset indicate that our method surpasses other models, providing further evidence that our proposed method is both robust and effective.

1 Introduction

Entity Relation Extraction (RE) is a crucial task in information extraction that aims to identify the relation between entity pairs in text. The findings of RE have practical applications in several fields, such as the construction of knowledge graphs (KG), semantic web annotation, and the development and optimization of question-and-answer systems and search engines, which have a significant impact on daily life. However, the task of RE is challenging due to the limited availability of annotated data. To address this challenge, distant supervision has been proposed, which automatically annotates data, significantly increasing the number of annotated samples.

However, distant supervision suffers from a strong hypothesis, leading to a large number of noisy labels during data annotation. Training on a dataset with noisy labels can result in model overfitting to the noise, which adversely impacts the model's performance (Li et al., 2022b; Christou and Tsoumakas, 2021).

To mitigate these issues, this paper proposes a novel method for RE that incorporates selective gate and the end-to-end noise correction method. In our model, selective gate is utilized to rationally select sentence features in the sentence bag, while noise correction is used to correct the labels of small classes of samples that are misclassified into larger classes during model training. These techniques reduce the negative impact of noisy data on the distant DSRE model. Additionally, since common word embedding models, such as Word2Vec and Glove, produce static vectors that overlook contextual semantics and the flexible use of multiple-meaning words, this paper introduces a pre-trained language model (PLM) to encode and extract features from sentences. This approach provides richer sentence semantic features, effectively improving prediction accuracy and reducing training time. Experiment results demonstrate that this method significantly outperforms the baseline models, improving the DSRE model's performance.

The major contributions of this paper can be summarized as follows:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

- We propose a DSRE method, named PLMG-Pencil, which combines PLM and selective gate and introduces an end-to-end noise correction training framework called pencil. Selective gate prevents the propagation of noisy representations and pencil corrects noise labels during the training process, reducing the impact of noise on the dataset and improving the performance of the DSRE model.

- We present a novel algorithm for DSRE that combines selective gate mechanism and pencil framework within a three-stage training process. This process involves training the backbone model, gradually correcting noisy labels, and subsequently fine-tuning our model using the corrected data. Empirical experiments demonstrate the robustness and effectiveness of our proposed method.

- Our experiments on three different Chinese and English datasets demonstrate that effective sentence-level feature construction methods and training methods, combined with noise correction, are crucial for improving the performance of models on DSRE tasks.

2 Related Work

2.1 Distantly Supervised Relation Extraction Models

Numerous RE models have been proposed, with deep learning-based models like convolutional neural networks (CNNs) being the current mainstream. CNNs can automatically extract features from sentences, making them a fundamental model for future research(Zhang and Wallace, 2015). However, the maximum pooling operation used in this model ignores important structural and valid information about the sentence.

Socher (2012) was the first to propose a recurrent neural network (RNN) to train relational extractors by encoding sentences. In addition, Zeng (2018) proposed a piecewise convolutional neural network (PCNN) that uses maximum pooling processing based on CNN to effectively preserve the information features of long texts while also reducing the time complexity. Zhou (2016) introduced an attention mechanism based on the long short-term memory network (LSTM) to form the classical BiLSTM-ATT model. The model can reasonably assign weights to features to obtain a better representation of the sentence. Riedel (2010) proposed a multi-instance learning (MIL) framework with a basic annotation unit of a sentence bag containing a common entity pair, rather than a single individual sentence. For sentence bag level labeled data, the model can be made to implicitly focus on correctly labeled sentences through an attention mechanism, thus learning from noisy data to become a stable and robust model. Subsequently, Ye and Ling (2019) proposed a DSRE method based on the intra- and inter-sentence bag, combining sentence-level and bag-level attention for noise correction. Alt(2019) introduced a transformer-based PLM for DSRE. Chen (2021) proposed a new contrastive instance learning method (CIL) to further improve the performance of DSRE models. Further, Li (2022a) introduces a hierarchical contrast framework (HiCLRE) on top of Chen's CIL method to enhance cross-layer semantic interaction and reduce the impact of noisy data. These methods are generally neural network driven and use neural network models that have strong generalization capabilities compared to traditional methods.

2.2 Noise Correction Methods

There are three categories of noise correction methods for DSRE: rule-based statistical methods, multi-instance learning-based methods, and adversarial and reinforcement learning-based methods. Multi-instance learning-based approaches have received the most attention from scholars, due to their effectiveness in correcting noise labels as demonstrated by Yao (2018).

In deep neural networks, designing robust loss functions has also proven effective in coping with noise by making models robust during training. Several studies have examined the robustness of different loss functions such as mean square loss and cross-entropy loss. Zhang (2018) combined the advantages of mean absolute loss and cross-entropy loss to obtain a better function. Li (2020a) proposed DivideMix framework that separates noisy samples using a Gaussian mixture model before training the model. Tanaka (2018) proposed an optimization strategy while Jiang(2018) introduced MentorNet technique for regularizing deep CNNs on test data with weakly supervised labels.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

Moreover, Wu(2017) and Shi(2018) have investigated adversarial training based approaches where simulated noise is mixed with real samples during training in order to improve model's robustness against noisy datasets by distinguishing between real versus noisy samples. Although this type of approaches improves corpus quality up to some extent, it requires simultaneous training of two models which can lead to instability and difficulty when applied directly into production systems at scale.

3 Methodology

To mitigate the impact of noise on the DSRE model, this paper proposes a two-pronged approach, PLM-based selective gate pencil (PLMG-Pencil) method. As shown in Figure 1, first, we encode the text using PLM and employ the selective gate mechanism to select sentence-level features that contribute to the bag-level feature. Second, we replace all labels with soft labels and train the model in the pencil framework. This framework uses soft labels that are updated during training and can be corrected for noisy data. This approach reduces the chances of noise being selected in the selective gate, even if it cannot be corrected in the pencil framework. These two methods complement each other, reducing the degree of noise interference and improving the model's RE performance. In this section, we will describe our approach from the backbone model architecture, noise correction framework, and the RE algorithm.

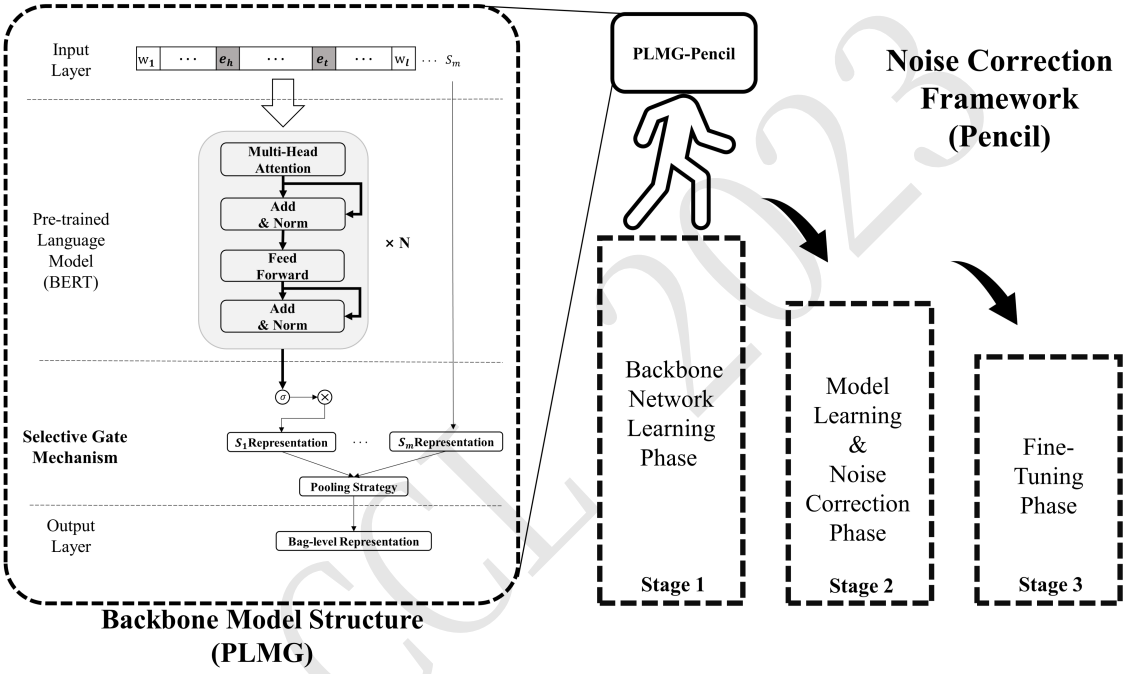


Figure 1: Overview of PLMG-Pencil

3.1 Backbone Model

This paper proposes the PLM-based selective gate as the backbone model, inspired by the Entity-aware Self-attention Enhanced selective gate (SeG) framework proposed by Li (2020b). The primary architecture of our model is presented in the backbone model structure part in Figure 1, and it comprises two main components: (1) **PLM**, structured to encode sentence, entity, and location features for semantic enhancement. (2) **Selective gate**, which enhances the representation of bag-level features by assigning weights to different sentences in the bag. The selective gate mechanism reduces the impact of noise on the model by weighing the contribution of each sentence in the bag.

3.1.1 Input Embeddings

To convert a sentence into a sequence of tokens, we use the BERT tokenizer, which results in a token sequence $S = \{t_1, t_2, \ldots, e_1, \ldots, e_2, \ldots, t_L\}$, where t_n denotes tokens, e_1 and e_2 denote the head and

tail entities, respectively, and L represents the maximum length of the input sentence. We add two special tokens, [CLS] and [SEP], to signify the beginning and end of the sentence, respectively.

However, the [CLS] token is not ideal for RE tasks as it only serves as a pooling token to represent the entire sentence. Therefore, to incorporate entity information into the input, we introduce four tokens: [unused1], [unused2], [unused3], and [unused4], which mark the start and end of each entity.

3.1.2 Selective Gate Enhanced Bag Representation

To obtain an effective bag representation, we introduce the selective gate mechanism, which dynamically calculates the weight of each sentence in the bag. We first represent each sentence using a PLM, such as BERT, which accepts structured sequences of tokens S that integrate entity information e_1 and e_2. The PLM's sentence encoder then sums the embeddings, including tokens, entities, and position, to generate context-aware sentence representations $H = \{h_{t_1}, h_{t_2}, \ldots, h_{e_1}, \ldots, h_{e_2}, \ldots, h_{t_L}\}$:

$$H = \mathrm{PLM}(S) \tag{1}$$

where h_{t_n} denotes the hidden features of the token t_n and PLM represents a pre-trained language model, such as BERT, that serves as the sentence encoder. We use special tokens to encode sentences to generate structural representations of sentences for RE task, including [CLS] for sentence-level pooling, its hidden features denoted as $h_{[CLS]}$. [unused1] and [unused2] mark the start and end of the head entity, [unused3] and [unused4] for the tail entity.

$$h_{e_h} = \mathrm{mean}(h_{t_{[unused1]}}, h_{t_{[unused2]}}) \tag{2}$$

$$h_{e_t} = \mathrm{mean}(h_{t_{[unused3]}}, h_{t_{[unused4]}}) \tag{3}$$

Representations of two entities, h_{e_h} and h_{e_t}, are generated by Equation (2) and Equation (3). The hidden features of these special tokens are denoted as $h_{t_{[unused1]}}$, $h_{t_{[unused2]}}$, $h_{t_{[unused3]}}$ and $h_{t_{[unused4]}}$. The sentence representations are generated using the following formulas:

$$h_{S_i} = \sigma([h_{e_h} \,||\, h_{e_t} \,||\, h_{[CLS]}] \cdot W_S) + b_S \tag{4}$$

where $||$ represents the concatenation operation, σ is the activation function, W_S is a weight matrix, and b_S is the bias.

Bag Representation The use of PLMs allows us to obtain sentence representations S_n, which can be stacked to form the initial bag representation $B = \{S_1, S_2, ..., S_n\}$. While selective attention modules are commonly used to aggregate sentence-level representations into bag-level representations, our proposed model leverages SeG's novel selective gate mechanism for this purpose. Specifically, when dealing with noisy data, the selective attention mechanism may be inefficient or ineffective if there is only one sentence in the bag, or if that sentence is mislabeled. Given that approximately 80% of the RE benchmark datasets contain single-sentence bags with mislabeled instances, our selective gate mechanism offers a more effective solution by dynamically reducing the alignment of gating values with instances of mislabeling, thereby preventing the propagation of noisy representations.

To generate gate values for each S_j, we employ a two-layer feed-forward network with the following formula:

$$g_j = \sigma(W^{(g_1)}\sigma(W^{(g_2)}S_j + b^{(g_2)}) + b^{(g_1)}), \forall j = 1, ..., m \tag{5}$$

We have $W^{(g2)} \in R^{3d_c \times d_h}$ and $W^{(g1)} \in R^{d_h \times d_h}$, $\sigma(\cdot)$denotes the activation function and $g_i \in (0, 1)$, after that, values of the gates are calculated and the mean pooling aggregation is performed in the bag to generate bag-level representation thus the further relation classification can be performed. The formula of this process is as follows, and m denotes the size of the sentence bag.

$$Q = \frac{1}{m}\sum\nolimits_{j=1}^{m} S_j g_j \tag{6}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

3.1.3 Classifier

We feed Q into a multi-layer perception (MLP) and apply the $|c|$-way softmax function to determine the relation between the head and tail entities, where $|c|$ represents the number of distinct relation classes. The formula for this process is as follows:

$$p = \text{Softmax}(\text{MLP}(Q)) \in R^{|c|} \tag{7}$$

3.1.4 Model Learning

To train the model, we minimize the negative log-likelihood loss plus an L2 regularization penalty, which is expressed by the following formula:

$$L_{NLL} = -\frac{1}{|D|}\sum_{k=1}^{|D|} \log p^k + \beta||\theta||_2^2 \tag{8}$$

where p^k represents the predicted distribution of the k-th example in the dataset D from Equation (8). The term $\beta||\theta||_2^2$ is the L2 regularization penalty, where θ is the set of model parameters, and β controls the strength of the regularization.

By minimizing this loss function using an optimization algorithm such as stochastic gradient descent, the model learns to predict the correct relation between the head and tail entities.

3.2 Noise Correction Framework

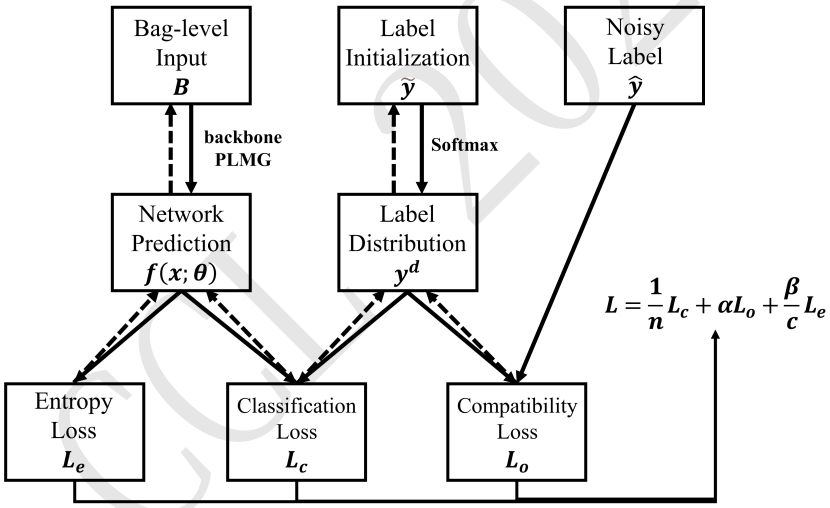


Figure 2: Pencil Framework

In this section, we introduce pencil, a noise correction framework based on the end-to-end noise-labeled learning correction framework proposed by Yi and Wu (2019). The framework is illustrated in Figure 2, with solid arrows representing forward computation and dashed arrows indicating backward propagation.

The pencil framework is designed to update both the network parameters and the data labels simultaneously using gradient descent and backpropagation. To accomplish this, the model generates a vector $\widetilde{y}$ to construct soft labels.

$$y^d = \text{Softmax}(\widetilde{y}) \tag{9}$$

With Equation (9), $\widetilde{y}$ can be updated by gradient descent and backpropagation. The following equation shows the initialized representation of the label with noise in the initial value.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

$$\widetilde{y} = K\hat{y} \tag{10}$$

where $\hat{y}$ is the original label with noise, and K is a large constant which ensures y^d and $\hat{y}$ has the most similar distribution in Equation (9), i.e., $y^d \approx \hat{y}$.

An intricately devised loss function is employed to correct the noise labels during the model training procedure, with L_e and L_o as penalty terms and L_c as the classification loss. This loss function incorporates two hyperparameters, denoted as α and β, which can be flexibly adjusted to accommodate diverse datasets with varying proportions of noisy data. Specifically, increasing the value of α and reducing the value of β will yield a diminished degree of label correction. In a c-class classification problem, the loss function is presented as follows.

$$L = \frac{1}{c}L_c + \alpha L_o + \frac{\beta}{c}L_e \tag{11}$$

where c denotes the number of classes.

The classification loss, which works as the main loss of the model guiding the model to learn, is measured using the dual form of the KL divergence between the predicted distribution and the soft labels. The formula for the classification loss L_c is given by:

$$L_c = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} f_j(x_j;\theta)\log\left(\frac{f_j(x_i;\theta)}{y_{ij}^d}\right) \tag{12}$$

where n denotes the batch size and y_{ij}^d denotes the corresponding soft label. In this equation, KL divergence is used in a symmetric form, which has been shown to perform better than using it directly in this framework in previous studies (Wu et al., 2017). Based on the gradient of the loss function L_c, it can be observed that a larger gap between the predicted value and the true label tends to correspond to a larger gradient. In this framework, the model parameter and noise labels can be updated together, which effectively serves to balance the disparity between the prediction and the true label, facilitating the gradual correction of noisy labels.

To avoid falling into a local optimum, the model sets the entropy loss L_e, using the predicted values of the network and its calculation of the cross-entropy loss. The formula is as follows.

$$L_e = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} f_j(x_i;\theta)\log f_j(x_i;\theta) \tag{13}$$

The compatibility loss function L_o is formulated as follows, which uses noise labels and soft labels to calculate the cross-entropy loss so as to avoid large deviations between the corrected label and the original noise label.

$$L_o = -\frac{1}{n}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{c}\hat{y}_{ij}\log y_{ij}^d \tag{14}$$

3.3 PLMG-Pencil Relation Extraction Method

This paper presents a RE algorithm that utilizes selective gate and noise correction, as shown in Algorithm 1. The complete training process of the algorithm is described below.

- **Stage 1 - Backbone Network Learning Phase**: Initially, the PLMG-Pencil network is trained from scratch with a larger fixed learning rate. The noise in the data is not processed in this stage, and the loss calculation formula only utilizes the classification loss. The network parameters obtained at this stage serve as the initialized network parameters for the next training step.

- **Stage 2 - Model Learning and Noise Correction Phase**: In this stage, the network parameters and label distributions are updated together using the model, thus, noisy labels can be corrected. To avoid overfitting the label noise, the label distribution is corrected for the noise in the original labels. We obtain a vector of label distributions for each sentence bag at the end of this stage. Due

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

to the dissimilarity of the learning rate used for soft labels update and the global model parameters update, a hyperparameter λ is set to adjust it.

- **Stage 3 - Final Fine-Tuning Phase**: The label distribution learned by the model in the previous stages are utilized to fine-tune the network in this stage. Sample labels in the training set are not updated, and the network parameters are updated using the classification loss as the loss function of the model. There are no additional adjustments to the learning rate, and the same decay rules are followed for general neural network training.

Algorithm 1: PLMG-Pencil Distantly Supervised Relation Extraction Algorithm

Input: Dataset $D = x_i, \widetilde{y}_i (1 < i < n)$, epoch of stages T_1, T_2.

Stage 1:
Initialization: $t \leftarrow 1$.
while $t \leq T_1$ **do**
  Train and update the model parameters θ, while calculating the loss in equation (14) with $\alpha = 0$ and $\beta = 0$. Hold off on using $\widetilde{y}_i$;
  $t \leftarrow t + 1$;

Stage 2:
Initialization: $\widetilde{y}_i = K\hat{y}_i$.
while $T_1 \leq t \leq T_2$ **do**
  Train and update the model parameters θ and y_i^d;
  $t \leftarrow t + 1$;

Stage 3:
while $T_2 \leq t$ **do**
  Train and update the model parameters θ and y_i^d;
  Train and update the model parameters θ, while calculating the loss in equation (14) with $\alpha = 0$ and $\beta = 0$. Do not update sample labels.
  $t \leftarrow t + 1$;

Output: θ, noise-corrected labels.

4 Experiments

4.1 Datasets

We evaluate our proposed model on three different datasets: the New York Times (NYT10) dataset and the GDS dataset in English, the SanWen dataset in Chinese. Datasets statistics are shown in Appendix A.

NYT10 (Riedel et al., 2010): This dataset is widely used in models based on DSRE, which is annotated with 58 different relations and the NA relation account for over 80% of the total. It has 522K and 172K sentence sets in the training and test sets respectively.

GDS (Jat et al., 2018): This dataset is created from the Google RE corpus, which contains 5 relations. It has 18K and 5K instances in the training and test sets, respectively.

SanWen (Xu et al., 2017): This dataset contains 9 relations from 837 Chinese documents. It has 10K, 1.1K, and 1.3K sentences in the training set, test set, and validation set respectively.

4.2 Baselines

To validate the effectiveness of the RE model proposed in this paper, we compare it with mainstream RE methods on the three datasets mentioned above. The following baseline methods are used.

Mintz (Mintz et al., 2009): It concatenates various features of sentences to train a multi-class logistic regression classifier.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

PCNN+ATT (Lin et al., 2016): It uses selective attention to multiple instances to alleviate the problem of mislabelling.
RESIDE (Vashishth et al., 2018): It exploits the information of entity type and relation alias to add a soft limitation for relation classification.
MTB-MIL (Baldini Soares et al., 2019): It proposes a method for matching gaps and learning sentence representations through entity-linked text.
DISTRE (Alt et al., 2019): It combines the selective attention with its PLM.
SeG (Li et al., 2020b): It uses an entity-aware embedding-based self-attentive enhancement selective gate based on PCNN+ATT to rationally select sentence features within sentence bags to reduce the interference of noise.
CIL (Chen et al., 2021): It proposes a comparative instance learning method in the MIL framework.
HiCLRE (Li et al., 2022a): It incorporates global structural information and local fine-grained interactions to reduce sentence noise.

4.3 Parameter Settings

Table 1 presents the hyperparameter settings used in our experiments. The English datasets are trained on the bert-base-uncased model from the Huggingface platform, while the Chinese dataset uses the bert-base-chinese model. To effectively train our model, we use the parameter settings from Yi and Wu (2019) as initialization settings for our experiments. The model's dropout rate, learning rate, α, β, batch size, and epoch settings are shown in the table.

Params	Dropout	LR	α	β	BatchSize	Epoch 1	Epoch 2
Value	0.5	0.035	0.1	0.4	64	15	20

Table 1: Parameter Settings. Epoch 1 and Epoch 2 mark the end of Stage 1 and Stage 2, respectively, and LR stands for the learning rate.

It is important to note that the optimal values for α and β may vary based on the level of noise in different datasets. Therefore, these values should be adjusted accordingly to improve the loss calculation and enhance the overall performance of the model.

4.4 Results

To evaluate the performance of our model in DSRE tasks, we use AUC and P@N values as evaluation metrics. AUC measures the area under the ROC curve, while P@N indicates the average accuracy of top N instances. Finally, P@M represents the average of these three P@N results.

4.4.1 Evaluation on English Dataset

Table 2 and Table 3 present a comparison of our proposed model with baseline models on dataset GDS and NYT10, respectively. Our model achieves promising results, as shown by the following observations: (1) Our proposed model shows competitive performance in terms of AUC values on both datasets. As shown in Table 2, on the GDS dataset, the AUC values of our model reach comparable levels with CIL and HiCLRE. Furthermore, as shown in Table 3, on the NYT10 dataset, our model outperforms CIL and DISTRE by 4.1% and 5.2% in AUC values respectively. (2) Our model demonstrates a clear advantage in terms of P@N values. On the NYT10 dataset, the P@100 value is 2.5% higher than CIL, which uses a contrast learning framework. The maximum difference in P@N values appears on the P@300 value, of which our method is 5.9% higher. In comparison to the DISTRE model, which also uses the PLM and MIL framework, our model outperforms it by 16%, 13.5%, and 12.7% on P@100, P@200, and P@300 values respectively.

We further conduct ablation experiments to highlight the benefits of the pencil framework. Specifically, we train our model using a conventional MIL training framework. When comparing the results of the PLMG model with the PLMG-Pencil model on the GDS dataset, we observe a 0.2% decrease in

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

the AUC value and a 0.1% decrease in the P@1K value for the PLMG model. These findings provide compelling evidence for the effectiveness of the pencil framework and our proposed algorithm. On the dataset NYT10, the proposed model shows a significant improvement compared to the model without pencil framework. Precisely, we observe a 6%, 2.5% and 2% improvement in P@100, P@200 and P@300 values respectively.

Dataset	Models	AUC	P@500	P@1K	P@2K	P@M
GDS	Mintz† (Mintz et al., 2009)	-	-	-	-	-
	PCNN-ATT† (Lin et al., 2016)	79.9	90.6	87.6	75.2	84.5
	MTB-MIL† (Baldini Soares et al., 2019)	88.5	94.8	92.2	87.0	91.3
	RESIDE† (Alt et al., 2019)	89.1	94.8	91.1	82.7	89.5
	REDSandT† (Christou and Tsoumakas, 2021)	86.1	95.6	92.6	84.6	91.0
	DISTRE† (Alt et al., 2019)	89.9	97.0	93.8	87.6	92.8
	CIL† (Chen et al., 2021)	90.8	**97.1**	94.0	87.8	93.0
	HiCLRE(Li et al., 2022a)	90.8	96.6	93.8	88.8	**93.1**
	PLMG-Pencil	**91.0**	95.4	**94.1**	88.8	92.8
	-without pencil (PLMG)	90.8	95.4	94.0	**89.0**	92.8

Table 2: Model Performances on GDS. (†) marks the results are reported in the previous research.

Dataset	Models	AUC	P@100	P@200	P@300	P@M
NYT10	Mintz† (Mintz et al., 2009)	10.7	52.3	50.2	45.0	49.2
	PCNN-ATT† (Lin et al., 2016)	34.1	73.0	68.0	67.3	69.4
	MTB-MIL† (Baldini Soares et al., 2019)	40.8	76.2	71.1	69.4	72.2
	RESIDE† (Alt et al., 2019)	41.5	81.8	75.4	74.3	77.2
	REDSandT† (Christou and Tsoumakas, 2021)	42.4	78.8	75.0	73.0	75.3
	DISTRE† (Alt et al., 2019)	42.2	68.0	67.0	65.3	66.8
	CIL† (Chen et al., 2021)	43.1	81.5	75.5	72.1	76.9
	HiCLRE(Li et al., 2022a)	45.3	82.0	78.5	74.0	78.2
	PLMG-Pencil	**47.0**	**84.0**	**80.5**	**78.0**	**80.8**
	-without pencil (PLMG)	47.0	78.0	78.0	76.0	77.3

Table 3: Model Performances on NYT10. (†) marks the results are reported in the previous research.

Figure 3 shows the PR curves for our proposed model and the baseline model. Our model clearly outperforms the baselines, particularly compared to the DISTRE model, which also uses PLM and MIL. Based on the ablation experiments conducted on the NYT10 dataset, it can be observed that the PLMG-Pencil method demonstrates a notable superiority in terms of precision at N (P@N) values. These results suggest that the selective gate has a positive impact on constructing sentence bag features and improving model performance. Furthermore, the pencil framework effectively corrects for noisy samples during training, leading to improved performance.

4.4.2 Evaluation on Chinese Dataset

We conduct additional experiments on the SanWen dataset to further validate the effectiveness of the pencil framework and selective gate mechanism. Figure 4 presents the model performances on this dataset.

Our model exhibits superior performance compared to HiCLRE, which utilizes the contrast learning framework, with a notable increase of 4.4% in AUC values. Furthermore, when compared to the SeG model that employs the selective gate mechanism, our PLMG-Pencil model, which incorporates the pencil approach, demonstrates a significant enhancement in AUC values. The ablation experiment further

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

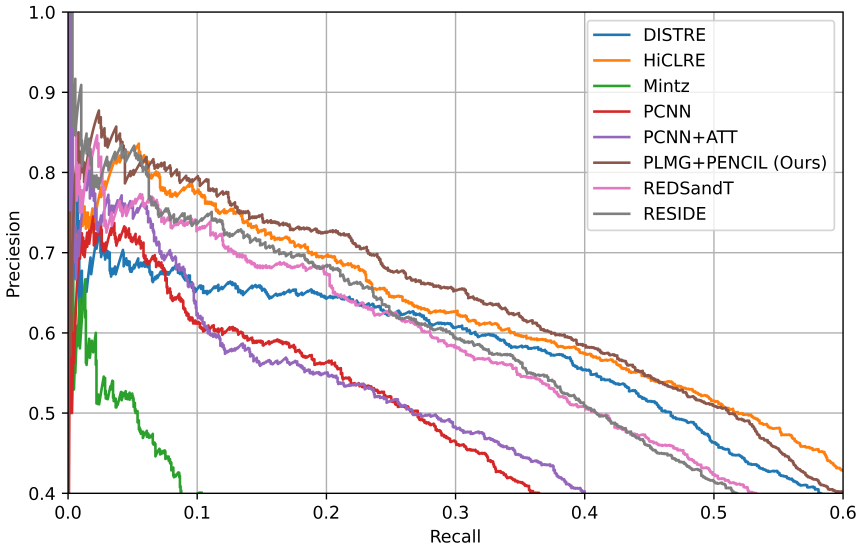


Figure 3: PR-Curve on NYT10

validates the effectiveness and robustness of our method. These results highlights the positive influence of the PLM and noise correction framework on the RE task.

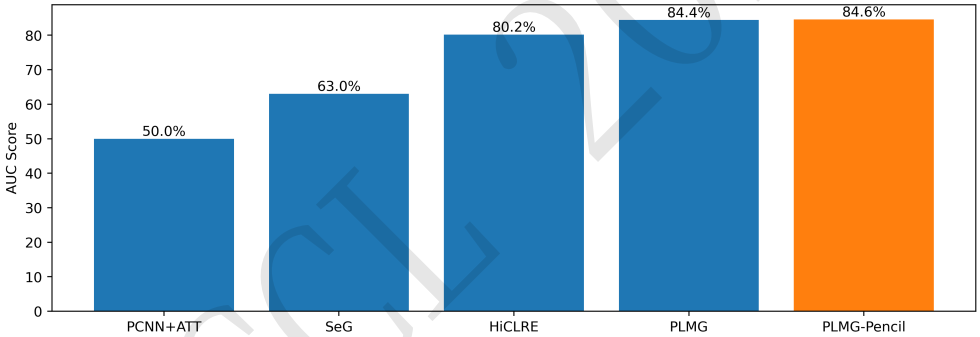


Figure 4: AUC Values of Models on SanWen

Based on the experimental results and the analysis of the dataset features described in Section 4.1, our model tends to perform better on datasets with more relations, such as NYT and SanWen. Compared with baselines, our model can achieve greater advantages on such datasets. In addition, the experimental results on the NYT10 dataset reveal that the pencil framework generates more significant performance enhancements compared to those obtained through experiments performed on the GDS dataset. The GDS dataset employs various methods to mitigate noise interferences and thus has higher quality annotations (Jat et al., 2018). Moreover, the pencil framework is designed to conduct a noise correction process for optimizing model performance, thus, it tends to bring larger improvements on datasets with greater amounts of noisy data.

5 Conclusion

In this paper, we propose the PLMG-Pencil method for DSRE. Our approach automatically learns the weights of different sentences in a sentence bag and selects the features that best represent the sentence bag through a gate mechanism. Additionally, we introduce a noise correction framework based on end-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

to-end probability with noise label learning for improved performance in RE. The experimental results clearly demonstrate that our proposed model outperforms baselines and achieves significant improvement in the RE task. Our approach shows great potential for practical application in the field of information extraction.

Acknowledgements

This work was supported by the National Social Science Fund of China (No. 22BTQ045).

A Datasets statistics

Dataset	#Relation	#Train	#Dev	#Test	Language
NYT	58	520K	-	172K	English
GDS	5	18K	-	5K	English
SanWen	9	10K	1.1K	1.3K	Chinese

Table 4: Datasets statistic.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy, July. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July. Association for Computational Linguistics.

Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021. CIL: Contrastive instance learning framework for distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6191–6200, Online, August. Association for Computational Linguistics.

Despina Christou and Grigorios Tsoumakas. 2021. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, 9:62574–62582.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020a. Dividemix: Learning with noisy labels as semi-supervised learning. *ArXiv*, abs/2002.07394.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020b. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8269–8276.

Dongyang Li, Taolin Zhang, Nan Hu, Chengyu Wang, and Xiaofeng He. 2022a. HiCLRE: A hierarchical contrastive learning framework for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2567–2578, Dublin, Ireland, May. Association for Computational Linguistics.

Rui Li, Cheng Yang, Tingwei Li, and Sen Su. 2022b. Midtd: A simple and effective distillation framework for distantly supervised relation extraction. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–32.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and He-Yan Huang. 2018. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1023.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium, October-November. Association for Computational Linguistics.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*.

Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. 2018. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2810–2819, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kun Yi and Jianxin Wu. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025.

Daojian Zeng, Yuan Dai, Feng Li, R Simon Sherratt, and Jin Wang. 2018. Adversarial learning for distant supervised relation extraction. *Computers, Materials & Continua*, 55(1):121–136.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.

Improving Cascade Decoding with Syntax-aware Aggregator and Contrastive Learning for Event Extraction

Zeyu Sheng , Yuanyuan Liang , Yunshi Lan*
School of Data Science & Engineering, East China Normal University, Shanghai, China
{51205903051,leonyuany}@stu.ecnu.edu.cn, yslan@dase.ecnu.edu.cn

Abstract

Cascade decoding framework has shown superior performance on event extraction tasks. However, it treats a sentence as a sequence and neglects the potential benefits of the syntactic structure of sentences. In this paper, we improve cascade decoding with a novel module and a self-supervised task. Specifically, we propose a syntax-aware aggregator module to model the syntax of a sentence based on cascade decoding framework such that it captures event dependencies as well as syntactic information. Moreover, we design a type discrimination task to learn better syntactic representations of different event types, which could further boost the performance of event extraction. Experimental results on two widely used event extraction datasets demonstrate that our method could improve the original cascade decoding framework by up to 2.2% percentage points of F1 score and outperform a number of competitive baseline methods.

1 Introduction

As an important yet challenging task in natural language processing, event extraction has attracted much attention for decades (Chen et al., 2015; Nguyen and Grishman, 2018; Zheng et al., 2019; Lai et al., 2021; Wang et al., 2021; Li et al., 2022; Ma et al., 2022; Zhou et al., 2022). This task aims at predicting event types, triggers and arguments from a given sentence. We display three examples in Figure 1. Given an example sentence (a) "*In 2018, Chuangwei Tech acquired equity of Qianhong Electronics for 1.5 billion* ...", an event extraction system is able to recognize the trigger "*acquired*", that corresponds to the event type "*invest*", and the argument "*Chuangwei Tech*", that plays the subject role of "*sub*" in the event.

A great number of methods have been developed for event extraction. Early methods formulate the event extraction as a sequence labeling task, where each token is considered as a candidate for labeling. They perform trigger extraction and argument extraction with joint learning (Li et al., 2013; Nguyen et al., 2016; Nguyen and Nguyen, 2019), which easily causes the label conflict issue. Considering the precedence relationship between the components in an event, pipeline methods are explored to perform trigger and argument extraction in separate stages (Chen et al., 2015; Du and Cardie, 2020; Liu et al., 2020; Ma et al., 2022). But the error is accumulated along with the pipeline. Recently, a cascade decoding framework (Xu et al., 2020; Sheng et al., 2021) is proposed to extract events with a cascade tagging strategy, which could not only handle the label conflict issue, but also avoid error propagation.

In above methods, a sentence is treated as a sequence, and methods suffer from the low efficiency problem in capturing long-range dependency. We take sentence (a) in Figure 1 as an example. The argument "*1.5 billion*" is far from the trigger "*acquired*" based on the sequential order while they are closely connected via the dependency arc. Therefore, it is necessary to take advantage of the syntactic structure to capture the relations between triggers and arguments. Some researches managed to include syntactic information of sentences in event extraction. Chen et al.(Chen et al., 2015) first employed dependency trees to conduct event extraction. Nguyen et al.(Nguyen and Grishman, 2018) and Yan et al.(Yan et al., 2019) treated each dependency tree as a graph and adopted Graph Convolution Network (GCN) (Kipf and Welling, 2017) to represent the sentence. More recent studies strengthened the graph representation via gate mechanism to filter out noisy syntactic information (Lai et al., 2020) or empowered the graph

*Corresponding author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

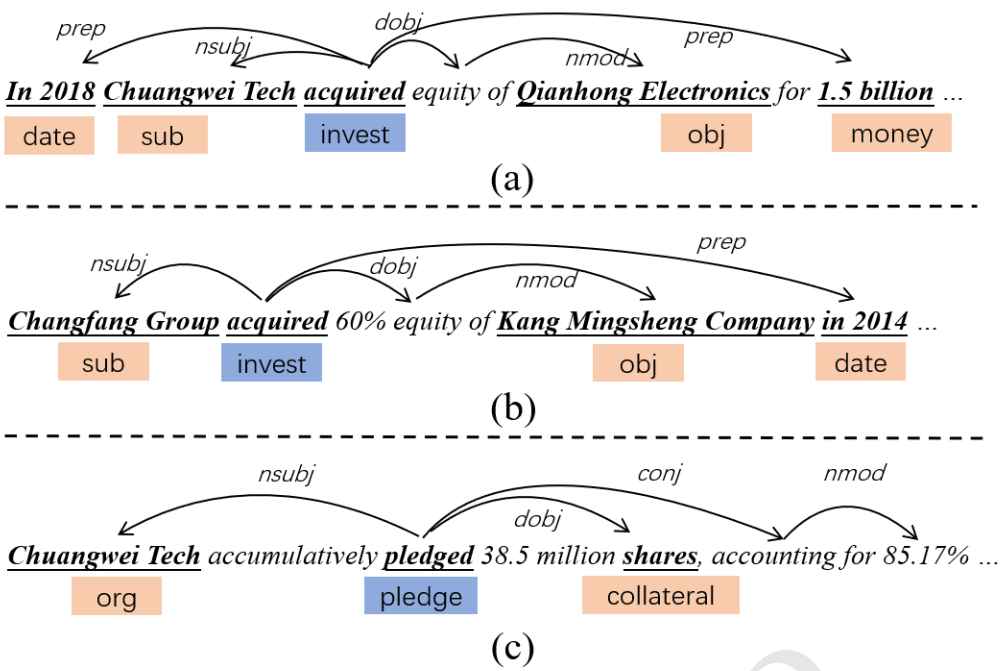


Figure 1: Three examples of event extraction. We annotate the event types with blue boxes under the triggers, and label the argument roles with orange boxes under the arguments.

encoder with more advanced Transformer (Ahmad et al., 2021). These methods could effectively solve the long-range dependency issue. However, they either follow the joint learning paradigm or pipeline paradigm thus still encounter the issue of label conflict or error propagation. In this paper, we develop our approach modeling syntactic information for event extraction based on cascade decoding framework. To achieve this, two main challenges should be addressed.

First, cascade decoding represents event types, triggers as well as arguments in the format of a triple. It sequentially predicts components in triples as subtasks and learns the implicit dependencies of the subtasks. It is not trivial to design a syntax encoder which is customized for the cascade decoders. In this paper, we propose a novel *Syntax-enhanced Aggregator* which could not only integrate the information from the precedent subtask with the current subtask but also model the syntactic structure of sentences. Moreover, this module could fuse the heterogeneous features together. In detail, our aggregator processes both subtask dependency and syntactic information via two channels. The final representation will be fused based on the alignment between tokens of a sentence and components in a dependency tree. Such aggregators are deployed in cascade decoders.

Second, existing methods involving syntactic structure rarely consider the interaction among event types. As examples (a) and (b) shown in Figure 1, the sentences of the same event type usually have similar syntactic structure despite different involved entities. In contrast, the sentences of the different event types usually have different syntactic structure despite similar involved entities, as examples (a) and (c). We design *Contrastive Learning* of syntactic representation to capture the interactions between sentences. Specifically, we define a type discrimination task to distinguish whether two sentences belong to the same event type based on their syntactic representations. This is jointly trained with event extraction task.

We conduct experiments on two event extraction datasets, FewFC (Zhou et al., 2021) and DuEE (Li et al., 2020b). The experiments show that compared with original cascade framework, our method can clearly perform better on both datasets. Our method also outperforms competitive baseline methods that represent the state-of-the-art on event extraction tasks. To reveal the working mechanism of our method, we also conduct ablation study and visualization that shed light on where the improvement comes from.

We summarize the contributions of this paper as follows: (1) We propose a novel syntax-enhanced aggregator to model the syntactic structure of sentences, which is a good fit for the cascade decoding framework. This aggregator is able to model syntax and fuse with dependencies of events. (2) To

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

further benefit from the syntax modeling, we design a type discrimination task to refine the syntactic representation via contrastive learning. (3) We empirically show the effectiveness of our method on two datasets. Our proposed method outperforms the baseline methods with remarkable margins based on F1 score of all measurement metrics.

2 Background

2.1 Problem Formulation

The task of event extraction aims at identifying event triggers with their types and event arguments with their roles. Specifically, a pre-defined event schema contains an event type set $\mathcal{C}$ and an argument role set $\mathcal{R}$. Given a sentence $x = \{w_1, w_2, ..., w_n\}$, the goal is to predict all events in gold set $\mathcal{E}_x$ of the sentence x, where the components of $\mathcal{E}_x$ are in the format of triples (c, t, a_r). Here, $c \in \mathcal{C}$ is an event type, t is a trigger word in sentence x, and a_r is an argument word corresponding to the role $r \in \mathcal{R}$. A dataset $\mathcal{D}$ consists of a set of $(x, \mathcal{E}_x)$.

2.2 A Cascade Decoding Framework

To solve the task, we follow the existing cascade decoding approach, CasEE method (Sheng et al., 2021), which is proposed to predict the events by maximizing the following joint likelihood:

$$\begin{aligned}&\prod_{(x,\mathcal{E}_x)\in\mathcal{D}}[\prod_{(c,t,a_r)\in\mathcal{E}_x} P((c,t,a_r)|x)]\\ =&\prod_{(x,\mathcal{E}_x)\in\mathcal{D}}[\prod_{c\in\mathcal{C}} P(c|x)\prod_{t\in\mathcal{T}_x} P(t|x,c)\prod_{a_r\in\mathcal{A}_{x,r}} P(a_r|x,c,t)],\end{aligned} \tag{1}$$

where $\mathcal{T}_x$ and $\mathcal{A}_{x,r}$ denote trigger and argument sets of x, respectively.

The joint likelihood explicates the dependencies among the type, trigger, and argument. The order of cascade decoding indicates that the framework first learns a *Type Decoder* $P(c|x)$ to identify the event types in the sentence. Then, it extracts the trigger words from the sentence via a *Trigger Decoder* $P(t|x, c)$ which corresponds to the detected type. After that, an *Argument Decoder* $P(a_r|x, c, t)$ is developed to extract role-specific arguments.

In the cascade decoding approach, the decoders are built upon a sharing BERT encoder:

$$\{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\} = \text{BERT}(x), \tag{2}$$

where $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\}$ is the hidden representation of x for downstream decoding. Next, an attention layer followed by a simple feed-forward neural network is leveraged as the type decoder to predict the event type. We denote it as:

$$P(c|x) = \text{TypeDecoder}(\mathbf{H}). \tag{3}$$

After that, the predicted type embedding $\mathbf{c}$ is concatenated with each token representation. This will be further processed via a conditional layer normalization (CLN) (Lee et al., 2021) layer and a self-attention layer to form the hidden representation $\mathbf{H}^c$. A pointer network takes charge of predicting the position of start and end indexes based on $\mathbf{H}^c$. We denote the above trigger extraction procedure as follows:

$$\begin{aligned}\mathbf{H}^c &= \text{Aggregator}(\mathbf{H}, \mathbf{c}),\\ P(t|x,c) &= \text{Pointer}(\mathbf{H}^c).\end{aligned} \tag{4}$$

For argument decoder, the trigger information is concatenated with $\mathbf{H}^c$ and processed with a CLN to form the hidden representation $\mathbf{H}^{ct}$. Given $\mathbf{H}^{ct}$, the start and end indexes of role-specific arguments are then predicted as follows:

$$\begin{aligned}\mathbf{H}^{ct} &= \text{Aggregator}(\mathbf{H}^c, \mathbf{t}),\\ P(a_r|x,c,t) &= \text{Pointer}(\mathbf{H}^{ct}).\end{aligned} \tag{5}$$

More details could be found in the original paper (Sheng et al., 2021).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

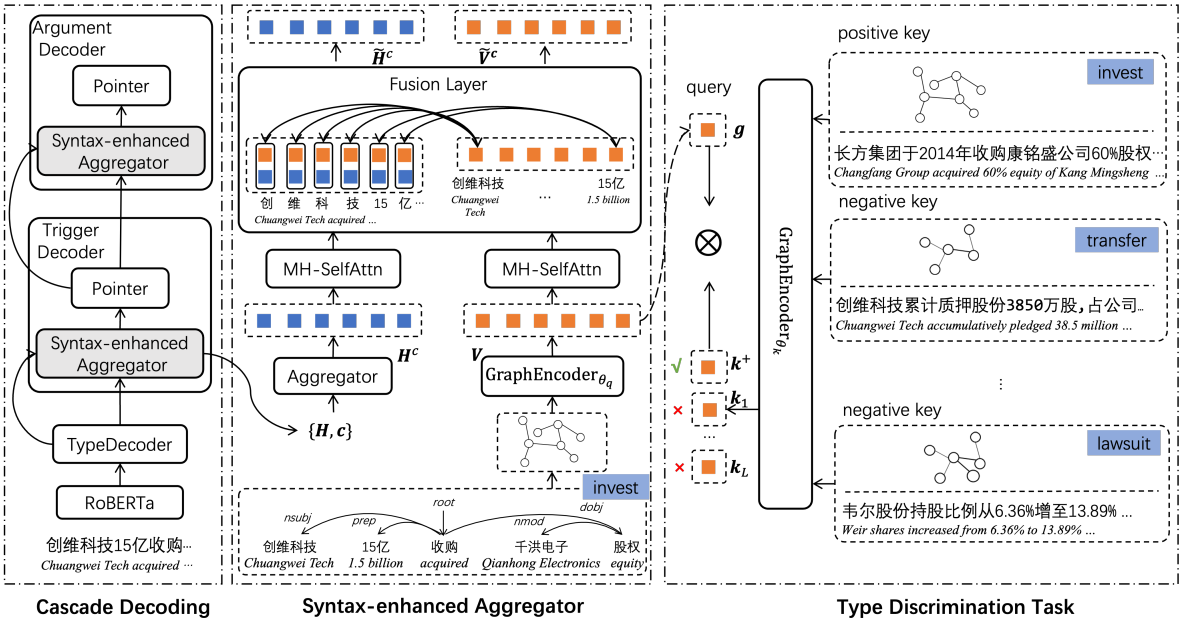


Figure 2: The overall architecture of our approach. The network modules are annotated with solid boxes and data is annotated with imaginary boxes. The left part is the cascade decoding framework. We modify the original aggregators to syntax-enhanced aggregators. The middle part shows the details of proposed syntax-enhanced aggregator in trigger decoder, where dependency and syntactic information are carried via two channels and eventually fuse together in fusion layer. The right part shows the details of type discrimination task, where syntactic representations belonging to the same event type are learned to be closer. Please note the imaginary line from the aggregator to the discriminator is meant to show the input of the discriminator rather than forward pass of the architecture.

3 Our Approach

The cascade decoding framework that we described in Section 2.2 decodes different components of events in a cascading manner, the inputs of which are hidden representations of tokens featured with subtask dependencies. Our approach follows the framework, but we improve it by introducing a module to fuse the syntactic information over the decoding process and a self-supervised task to further refine the syntactic representation. Specifically, we propose *Syntax-enhanced Aggregators* to take place of the original aggregators. The proposed aggregator elaborately models the syntactic structure of the sentence and fuses syntax with the original hidden representation, as we will explain in Section 3.1. To better capture the interactions among event types, we design a *Type Discrimination Task* to distinguish whether the representations belonging to the same type are syntactically close or not, which will be presented in Section 3.2. Eventually, event detection and type discrimination generate their training objectives and we join them together, as we will describe in Section 3.3. The overall architecture of our approach is displayed in Figure 2.

3.1 Syntax-enhanced Aggregator

Recall that we could prepare the hidden representations enriched with dependency information $\mathbf{H}^c$ and $\mathbf{H}^{ct}$ through the aggregators in trigger decoder and argument decoder, respectively. Now we describe, in our syntax-enhanced aggregator, how we obtain the syntactic representations and fuse these heterogeneous features to form new representations $\tilde{\mathbf{H}}^c$ and $\tilde{\mathbf{H}}^{ct}$. For simplicity, we take $\mathbf{H}^c$ in trigger decoder as the example. The similar procedure is conducted for $\mathbf{H}^{ct}$ in argument decoder.

We first extract the dependency tree of the sentence via existing parsing tools. To avoid one way message transition from the root to leaf nodes, we add reversed edges and distinguish them with different edge labels in the dependency tree. This results in a syntactic graph $\mathcal{G}(v, e)$, where v is the entity in a

dependency tree and e is the grammatical link between these entities. The representation of entities are updated along with the graph structure. Let us use $\mathbf{V} = \{\mathbf{v}_1, ...\mathbf{v}_m\}$ to denote the representations of m entities in $\mathcal{G}$. Each entity is initially represented via the average embeddings of their tokens.

To model the syntactic structure of sentences, we adopt the commonly used Relational Graph Convolutional Network (R-GCN) (Schlichtkrull et al., 2018) as our graph encoder to capture the message transition of the syntactic graph:

$$\{\mathbf{v}_1, ...\mathbf{v}_m\} = \text{GraphEncoder}(\{\mathbf{v}_1, ...\mathbf{v}_m)\}. \tag{6}$$

In this way, the updated entity representation is featured with sentence syntax. Next, we aggregate them with the original hidden representations $\mathbf{H}^c = \{\mathbf{h}_1^c, ..., \mathbf{h}_n^c\}$, which are arranged in token level, such that we can fuse these two types of information together.

We first utilize two individual multi-head self-attentions (MH-SelfAttns) to process both $\mathbf{H}^c$ and $\mathbf{V}$, respectively. Inspired by the Knowledgeable Encoder proposed in prior work (Zhang et al., 2019), where the language representation is enhanced with knowledge graphs, we align an entity with its corresponding tokens or characters if it is formed by multiple tokens or characters. As shown in Figure 2, the entity "创维科技 *(Chuangwei Tech)*" is aligned with "创", "维", "科", and "技". Thus there are explicit links between this entity and the four characters. We define the fusion layer as follows:

$$\mathbf{z}_j = \sigma(\mathbf{U}_1\mathbf{h}_j^c + \sum_{v_i \in \mathcal{A}lign(w_j)} \mathbf{W}_1\mathbf{v}_i + \mathbf{b}_1) \tag{7}$$

$$\tilde{\mathbf{h}}_j^c = \sigma(\mathbf{U}_2\mathbf{z}_j + \mathbf{b}_{21}) \tag{8}$$

$$\tilde{\mathbf{v}}_i^c = \sigma(\sum_{w_j \in \mathcal{A}lign(v_i)} \mathbf{W}_2\mathbf{z}_j + \mathbf{b}_{22}), \tag{9}$$

where σ is non-linear activation function GELU (Hendrycks and Gimpel, 2016) and $\mathcal{A}lign$ indicates the alignment between tokens and entities. The inputs are hidden representation $\mathbf{H}^c$ and entity representation $\mathbf{V}$. $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{b}$ with subscripts are parameters to learn. $\mathbf{z}_j$ indicates fused hidden representation of j-th token. As a result, $\tilde{\mathbf{H}}^c = \{\tilde{\mathbf{h}}_1^c, \tilde{\mathbf{h}}_2^c, ..., \tilde{\mathbf{h}}_n^c\}$ is the token representation with fusion of syntax information. It will be leveraged as the input of pointer network in Equation (4) for trigger extraction. $\tilde{\mathbf{V}}^c = \{\tilde{\mathbf{v}}_1^c, \tilde{\mathbf{v}}_2^c, ..., \tilde{\mathbf{v}}_m^c\}$ is the entity representation enriched with subtask dependencies. It will be utilized in downstream decoding.

When it comes to the argument decoder, $\tilde{\mathbf{V}}^c$ is used as the input entity representation to be continuously processed via the graph encoders and eventually fuse with the hidden representation $\mathbf{H}^{ct}$ to generate $\tilde{\mathbf{H}}^{ct}$. This will be fed into pointer network in Equation (5) for argument extraction. Compared with the original aggregator, besides capturing dependency information, our syntax-enhanced aggregators encode syntactic structure and fuse both subtask dependencies and syntactic information to generate a more expressive representation for decoding.

3.2 Type Discrimination Task

Type discrimination task aims at predicting whether two sentences are syntactically close or not. The intuition behind is that sentences describing the same event type usually have similar syntactic structure. To this end, we adopt the idea of contrastive learning and push the syntactic representations of positive pairs closer than negative pairs. The syntactic representations learned from type discrimination task can further boost the performance of cascade decoding.

We conduct dependency parsing for all sentences and obtain a collection of syntactic graphs denoting as $\mathcal{U}$, each $\mathcal{G} \in \mathcal{U}$ deriving from a sentence is labeled with their event type. Then, we train the representations of a pair of syntactic graphs that share the same event type to be closer in the space. We adopt Momentum Contrast (MoCo) (He et al., 2020) for self-supervised representation learning, which formulates contrastive learning as a dictionary look-up task and is effective in maintaining a large-scale dynamic dictionary.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

Specifically, given a syntactic graph $\mathcal{G}$ as a query, we represent it by the average of all entities encoded via the graph encoder of Equation (6) and obtain $\mathbf{g} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{v}_i$ to indicate the status of the syntactic graph. Meanwhile, we sample a set of syntactic graphs from $\mathcal{U}$ as keys of a dictionary and encode these key graphs via another graph encoder to obtain their representations. For clear presentation, we denote the query graph encoder and key graph encoder as $\text{GraphEncoder}_{\theta_q}$ and $\text{GraphEncoder}_{\theta_k}$, respectively. In the dictionary, the positive key (denoted as $\mathbf{k}^+$) is the only graph having the same type as the query. The others are negative keys $\{\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_L\}$, as depicted in Figure 2. We define the loss function of the type discrimination task as follows:

$$\mathcal{L}_{TD} = -\sum_{\mathcal{G}\in\mathcal{U}} log\frac{\exp(\mathbf{g}^\intercal\mathbf{k}^+/\tau)}{\sum_{i=0}^{L}\exp(\mathbf{g}^\intercal\mathbf{k}_i/\tau)}, \tag{10}$$

where τ is a temperature hyper-parameter. For each query, we construct one positive key and L negative keys.

Similar as MoCo, during training, the keys in the dictionary are progressively updated. For each new query graph $\mathcal{G}$, the old key graphs in the dictionary are removed and new key graphs are collected. Moreover, the parameters of the encoder of keys are driven by momentum update as follows:

$$\text{GraphEncoder}_{\theta_k} \leftarrow \gamma\text{GraphEncoder}_{\theta_k} + (1-\gamma)\text{GraphEncoder}_{\theta_q}, \tag{11}$$

where γ is the momentum coefficient. This results in a smooth evolution of $\text{GraphEncoder}_{\theta_k}$ as we can control the evolving progress.

3.3 Training Objective

During our training procedure, event extraction and type discrimination tasks are performed simultaneously. For each sampled data, a sentence and its corresponding syntactic graph are both collected for event extraction training. A dictionary of key graphs for a query graph is also prepared for contrastive learning.

The overall training objectives of our improved cascade decoding framework is shown as follows:

$$\mathcal{L} = \lambda\mathcal{L}_{EE} + (1-\lambda)\mathcal{L}_{TD}, \tag{12}$$

where $\mathcal{L}_{EE}$ is the negative logarithm of the joint likelihood of event extraction task in Equation (1), and λ is a hyper-parameter. All the parameters except for $\text{GraphEncoder}_{\theta_k}$ are updated by back-propagation.

4 Experiments

In this section, we conduct experiments to evaluate the proposed method. We first introduce our experiment settings including datasets and evaluation metrics, comparable methods, and implementation details in Section 4.1, Section 4.2, and Section 4.3. Next, we discuss our main results in Section 4.4. We show further analysis in Section 4.5

4.1 Datasets and Evaluation Metrics

We conduct experiments on two commonly used event extraction datasets:

- **FewFC** (Zhou et al., 2021)[1] is a public Chinese dataset for event extraction in the financial domain. It contains 10 event types and 19 role types. There are 12,890 sentences in the dataset. Following previous setting (Sheng et al., 2021), we split the dataset with the ratio 8 : 1 : 1 to form training, development, and test sets.
- **DuEE** (Li et al., 2020b)[2] is a relatively large Chinese event extraction dataset, which contains 19,640 sentences in total. The data is collected by crowdsourcing and contains 65 event types associated with 121 role types in real-world scenarios. We follow its default split setting to construct the data sets.

[1] https://github.com/TimeBurningFish/FewFC
[2] http://ai.baidu.com/broad/download

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

Methods	TI			TC			AI			AC		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
DMCNN⋆	82.0	79.4	80.7	69.4	68.2	68.8	70.2	66.3	68.2	66.8	65.7	66.2
GCN-ED⋆	84.4	83.7	84.0	71.7	68.9	70.3	69.1	69.6	69.4	71.2	65.7	68.3
GatedGCN⋆	88.9	85.0	86.9	76.2	73.4	74.8	72.3	70.1	71.2	71.4	68.8	70.1
BERT-CRF	88.4	84.1	86.2	74.2	70.5	72.3	69.4	68.1	68.7	70.8	68.2	69.5
MQAEE	88.7	86.2	87.4	77.2	76.4	76.8	**72.7**	69.7	71.2	70.2	66.5	68.3
CasEE	89.1	87.8	88.4	77.8	78.6	78.2	71.6	73.2	72.4	71.2	72.4	71.8
Ours⋆	**90.1**	**88.9**	**89.5**	**78.1**	**79.4**	**78.7**	71.9	**77.0**	**74.4**	**71.5**	**75.7**	**73.5**

Table 1: Event extraction results on test set of FewFc dataset. P(%), R(%) and F1(%) denote percentages of precision, recall and F1 score, respectively. The methods annotated with "⋆" are those enriched with syntactic features.

We utilize the standard evaluation metrics (Chen et al., 2015; Du and Cardie, 2020) to evaluate performance of trigger detection and argument detection: (1) Trigger Identification (**TI**): If a predicted trigger word matches the gold word, this trigger is identified correctly. (2) Trigger Classification (**TC**): If a trigger is correctly identified and assigned to the correct type, it is correctly classified. (3) Argument Identification (**AI**): If an event type is correctly recognized and the predicted argument word matches the gold word, it is correctly identified. (4) Argument Classification (**AC**): If an argument is correctly identified and the predicted role matches the gold role type, it is correctly classified. We measure Precision, Recall and F1 score based on the above four metrics.

4.2 Comparable Methods

We choose a range of advanced approaches for event extraction as our baselines:

- **DMCNN** (Chen et al., 2015) is a pipeline with dynamic multi-pooling convolutional neural network and enriched encoded syntactic features. It is the early attempt adopting syntactic information into event extraction.
- **GCN-ED** (Nguyen and Grishman, 2018) develops a GCN based on dependency trees to perform event detection, where each word is treated as a trigger candidate and joint learning is performed to label words with event types.
- **GatedGCN** (Lai et al., 2020) is GCN-based model for event detection which uses a gating mechanism to filter noisy information. It also follows a joint learning paradigm.
- **BERT+CRF** (Du and Cardie, 2020) is a sequence labeling model with advanced pre-trained language model BERT for encoding sentences and conditional random field (CRF) for tagging labels.
- **MQAEE** (Li et al., 2020a) is a pipeline method that formulates the extraction task as a multi-turn question answering without any syntactic information involved.
- **CasEE** (Sheng et al., 2021) is the representative cascade decoding approach for event extraction, which simply treats a sentence as a sequence.

We either utilize official source codes or follow their descriptions to re-implement the baseline methods.

4.3 Implementation Details

For implementation, we use Chinese BERT Model (Devlin et al., 2018) in Transformers library[3] as our basic textual encoder to convert words into vector representations. For other parameters, we randomly initialize them. To obtain syntactic graphs, we extract the syntactic dependency of sentences via StanfordNLP parsing tool[4] and convert dependency trees into graphs via DGL[5] library. In our syntax-

[3] https://huggingface.co/
[4] https://nlp.stanford.edu/software/lex-parser.shtml
[5] https://www.dgl.ai/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

Methods	TI			TC			AI			AC		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
DMCNN⋆	78.4	80.2	79.3	79.4	76.3	77.8	69.2	67.4	68.3	67.2	65.6	66.4
GCN-ED⋆	82.4	76.2	79.2	81.6	76.2	78.8	71.3	69.5	70.4	70.9	64.5	67.5
GatedGCN⋆	**88.6**	83.0	85.7	82.4	80.5	81.4	**73.8**	71.6	72.7	**72.5**	68.4	70.4
BERT-CRF	87.2	77.6	82.1	80.4	77.4	78.8	70.6	68.1	69.3	70.5	66.7	68.5
MQAEE	87.9	82.1	84.9	80.9	79.4	80.1	73.2	71.7	72.4	71.0	69.7	70.3
CasEE	85.5	88.2	86.8	83.6	83.9	83.7	70.3	75.4	72.8	68.6	75.7	72.0
Ours⋆	87.7	**89.0**	**88.3**	**83.7**	**86.8**	**85.2**	72.8	**76.9**	**74.8**	71.2	**77.4**	**74.2**

Table 2: Event extraction results on test set of DuEE dataset. P(%), R(%) and F1(%) denote percentages of precision, recall and F1 score, respectively. The methods annotated with "⋆" are those enriched with syntactic features.

enhanced aggregator, we use 8 heads for MH-SelfAttns layers and 2 stacked R-GCN layers to form a GraphEncoder. For hyper-parameters, we search via grid search through pre-defined spaces and decide the best configuration based on the best F1 score on the development set. The dimension of hidden representations in graph encoders or aggregators are all set to 768. We use an Adam optimizer (Kingma and Ba, 2015) to train all trainable parameters. The initial learning rate is set to $1e-5$ for BERT parameters and $1e-4$ for the other parameters. A warmup proportion for learning rate is set to 10%. The training batch is set to 16 and the maximum training epoch is 30. The size of dictionary L is set to 1000 for contrastive learning. We set $\tau = 0.07$, $\lambda = 0.5$ and $\gamma = 0.8$. To avoid overfitting, we apply dropout layers in syntax-enhanced aggregators with a dropout ratio as 0.3.

4.4 Main Results

The performance of all methods on FewFC and DeEE datasets is displayed in Table 1 and Table 2, respectively. Based on the two tables, we have the following observations:

(1) For both datasets, our method surpasses all baseline methods with a remarkable margin and obtains new state-of-the-art results on F1 score of all measurement metrics. This shows that our method incorporating syntactic information with cascade decoding framework indeed brings the largest benefit for event extraction task. Compared with CasEE, our method shows gains on TI as well as AI measurement. This may because that leveraging syntactic relation of sentences captures long-range dependency and enables the model to retrieve more accurate trigger and arguments. Also, the gains on TC and AC may comes from contrastive learning, which helps the model label events by discriminating the different syntactic structure of event types.

(2) In the perspective of framework, compared with the joint learning and pipeline paradigms, cascade decoding could achieve better performance. CasEE outperforms BERT-CRF as well as MQAEE with marginal improvement on both datasets. As discussed in Section 1, cascade decoding framework could avoid label conflicts and error propagation effectively (Sheng et al., 2021), which reveals the necessity of developing methods based on cascade decoding framework.

(3) For methods featured with syntactic information, different methods show different effects. Specifically, DMCNN and GCN-ED are methods involving syntactic information, their performance on both datasets are not ideal, this may because that these two methods are developed upon un-contextual word embeddings thus cannot fully capture the deep semantics of sentences. GatedGCN takes advantage of BERT encoder and encodes syntactic information via GCN model and it could outperform the BERT-CRF method. Our method is also built upon BERT encoder and featured with syntax-enhanced aggregator and type discrimination task, which is effective in solving the label conflict and modeling syntactic information of sentences.

4.5 Further Analysis

Ablation study. To explore details of our proposed method, we show the result of ablation study in Table 3. As we can see, both syntax-enhanced aggregators and contrastive learning contribute to the entire system. After we omit the contrastive learning, the performance decreases. This indicates that

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

	TI(%)	TC(%)	AI(%)	AC(%)
Our Model	**89.5**	**78.7**	**74.4**	**73.5**
w/o Contrastive learning	88.7	78.3	73.6	72.4
w/o Fusion Layer	89.0	78.4	73.6	72.8
w/o SA in Trigger Decoder	88.5	78.1	72.4	71.9
w/o SA in Argument Decoder	89.3	78.6	73.0	72.1

Table 3: Results of ablation study on FewFC dataset. We display the percentages of F1 score on all measurement metrics. **SA** denotes Syntax-enhanced Aggregator.

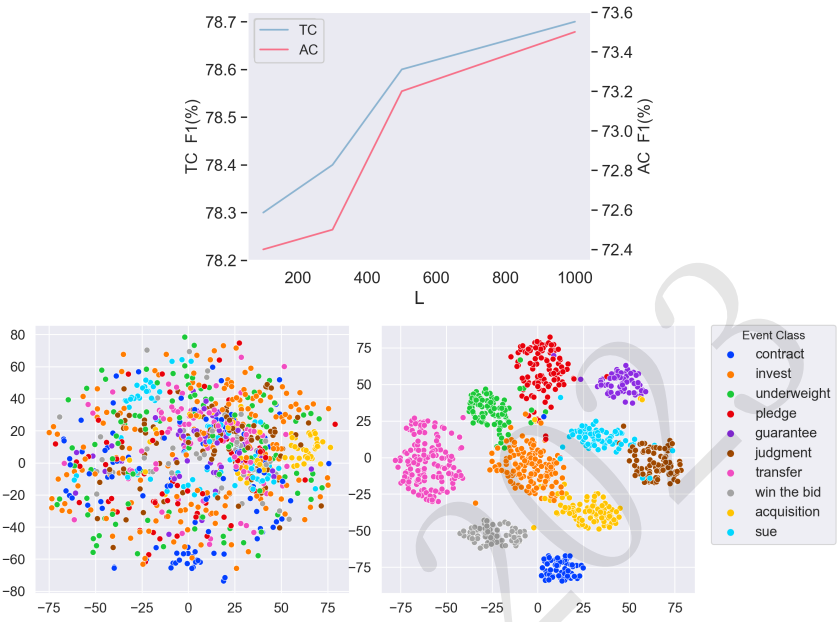


Figure 3: (a) shows the performance change of TC and AC on FewFC with increasing L value in contrastive learning. (b) shows the t-SNE plots of representations of query graphs of FewFC without and with contrastive learning.

capturing the syntactic structure of sentences is key for detecting the event types. Similarly, After we omit the fusion layer in syntax-enhanced aggregator and simply add the hidden representation of syntactic graph to $\mathbf{H}^c$, the performance drops. This indicates that the way to combine syntactic feature and subtask dependencies is critical. We remove the syntax-enhanced aggregators in trigger and argument decoders in turn. The performance decrease indicates that the proposed syntax-enhanced aggregators contribute to both trigger extraction and argument extraction.

Effect of L. In order to show the effect of L value in contrastive learning, we train our method on FewFC dataset with varying dictionary sizes and draw curves in Figure 3 (a). The figure shows that with the increase of L value in contrastive learning, the performance of trigger classification and argument classification increases. This is because seeing more interactions of different event types could help the model learn more distinct syntactic features.

Representation visualization. In Figure 3 (b), we display the learned query representations in FewFC dataset by mapping them into two dimensional space via t-distributed stochastic neighbor embedding (t-SNE) (Hinton and Roweis, 2002). The data points with different colors indicate query graphs of different categories of event types. As we can observe, the query representations of different event types without contrastive learning mix together and exhibit random distribution. In contrast, after including type discrimination task with contrative learning, the same event types clustered. This verifies that contrastive learning leads to a better syntactic representation for each sentence.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

5 Related Work

5.1 Frameworks of Event Extraction

The frameworks of event extraction can be roughly categorized into three groups. Joint learning framework solves event extraction in a sequence labeling manner (Li et al., 2013; Nguyen et al., 2016; Sha et al., 2018a; Liu et al., 2018; Nguyen and Nguyen, 2019; Shen et al., 2020; Huang et al., 2020). They treat each token as the candidate of a trigger or an argument and tag it with types. However, joint learning has the disadvantage of solving sentences where one token could have more than one event types. Pipeline framework performs trigger extraction and argument extraction in separate stages (Yang et al., 2019; Wadden et al., 2019; Li et al., 2020a; Du and Cardie, 2020; Liu et al., 2020; Chen et al., 2020; Ma et al., 2022; Zhou et al., 2022). This framework could avoid the label conflict issue but it ignores the potential label dependencies in modeling and suffers from error propagation. The cascade decoding framework formulates triples to represent event types, triggers and arguments (Xu et al., 2020; Sheng et al., 2021; Yang et al., 2021). It jointly performs predictions for event triggers as well as arguments based on shared feature representations and learns the implicit dependencies of the triples. It could avoid label conflict and error propagation. Empirical results show it is an effective solution for event extraction. The cascade decoding framework is also effective in jointly extracting relations and entities from text (Zheng et al., 2017; Wei et al., 2020).

5.2 Syntax Modeling for Event Extraction

There are a number of studies that incorporate the syntactic structure of sentences into event extraction tasks. The early work (Chen et al., 2015) collected syntactic features from the dependency tree and fed them into a dynamic multi-pooling convolutional neural network for extracting events. Li et al.(Li et al., 2018) also utilized dependency-based embeddings to represent words semantically and syntactically and proposed a PMCNN for biomedical event extraction. Some studies tried to enhance the basic network with syntactic dependency, Sha et al.(Sha et al., 2018b) proposed a novel dependency bridge recurrent neural network and Zhang et al. (Zhang et al., 2018) transformed dependency trees into target-dependent trees. The follow-up studies (Nguyen and Grishman, 2018; Liu et al., 2018; Yan et al., 2019) employed graph convolutional network to encode the dependency tree and utilized it for predicting event types. More advanced neural networks are leveraged to model syntax in event extraction tasks. The gate mechanism and Transformer (Lai et al., 2020; Ahmad et al., 2021; Xie et al., 2021) have shown to be effective in encoding the graph information of dependency tree. (Li et al., 2021) utilized the relationships of event arguments based on a reinforcement learning and incremental learning. (Lu et al., 2021) designed a sequence-to-structure framework to uniformly models different subtasks of event extraction. However, some of them focus on detecting event types with syntax modeling which can be treated as a joint learning framework of event extraction, the others follow a pipeline framework of event extraction to enhance syntactic information. To fully make use of the cascade decoding framework, we propose our method based on the cascade decoding architecture, which captures the subtask dependencies and syntactic structure simultaneously.

6 Conclusions

In this paper, we improved cascade decoding with syntax-aware aggregator and contrastive learning for event extraction. We demonstrated the effectiveness of our proposed method on two datasets. The results showed that our method outperforms all baseline methods based on F1 score. Considering that many scenes have relatively high requirements for real-time performance, we will explore to optimize the computational complexity of the model and improving the universality of the model in the future.

Acknowledgements

This work was supported in part by ECNU Research Fund on Cultural Inheritance and Innovation (Grant No. 2022ECNU—WHCCYJ-31) and Shanghai Pujiang Talent Program (Project No. 22PJ1403000). We sincerely thank the anonymous reviewers for their valuable comments and feedback.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

References

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: graph attention transformer encoder for cross-lingual relation and event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12462–12470.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, and Benjamin Van Durme. 2020. Reading the manual: Event extraction as definition comprehension. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*.

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

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415*.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15.

Peixin Huang, Xiang Zhao, Ryuichi Takanobu, Zhen Tan, and Weidong Xiao. 2020. Joint event extraction with hierarchical policy network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2653–2664.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5405–5411.

Viet Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L. Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 93–102.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

Lishuang Li, Yang Liu, and Meiyue Qin. 2018. Extracting biomedical events with parallel multi-pooling convolutional neural networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(2):599–607.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020b. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 534–545.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

Qian Li, Hao Peng, Jianxin Li, Jia Wu, Yuanxing Ning, Lihong Wang, S Yu Philip, and Zheng Wang. 2021. Reinforcement learning-based dialogue guided event extraction to exploit argument relations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:520–533.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2022. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 5900–5907.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6851–6858.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018a. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018b. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. 2020. Hierarchical Chinese legal event extraction via pedal attention mechanism. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 100–113, December.

Jiawei Sheng, Shu Guo, Bowen Yu, Qian Li, Yiming Hei, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2021. CasEE: A joint learning framework with cascade decoding for overlapping event extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 164–174.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, November.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6283–6297.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488.

Jianye Xie, Haotong Sun, Junsheng Zhou, Weiguang Qu, and Xinyu Dai. 2021. Event detection as graph parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, page 1630–1640.

Nuo Xu, Haihua Xie, and Dongyan Zhao. 2020. A novel joint framework for multiple Chinese events extraction. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 950–961.

Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event detection with multi-order graph convolution and aggregated attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5766–5770.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294.

Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. Document-level event extraction via parallel prediction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6298–6308.

Wenbo Zhang, Xiao Ding, and Ting Liu. 2018. Learning target-dependent sentence representations for chinese event detection. In *China Conference on Information Retrieval*, pages 251–262. Springer.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346.

Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, and Jinlong Li. 2021. What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14638–14646.

Jie Zhou, Qi Zhang, Qin Chen, Qi Zhang, Liang He, and Xuanjing Huang. 2022. A multi-format transfer learning model for event argument extraction via variational information bottleneck. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 748-760, Harbin, China, August 3 - 5, 2023.

Learnable Conjunction Enhanced Model for Chinese Sentiment Analysis

Bingfei Zhao, Hongying Zan *, Jiajia Wang, Yingjie Han
Zhengzhou University
zbf472670570@163.com, iehyzan@zzu.edu.cn, wjj4work@163.com
ieyjhan@zzu.edu.cn

Abstract

Sentiment analysis is a crucial text classification task that aims to extract, process, and analyze opinions, sentiments, and subjectivity within texts. In current research on Chinese text, sentence and aspect-based sentiment analysis is mainly tackled through well-designed models. However, despite the importance of word order and function words as essential means of semantic expression in Chinese, they are often underutilized. This paper presents a new Chinese sentiment analysis method that utilizes a Learnable Conjunctions Enhanced Model (LCEM). The LCEM adjusts the general structure of the pre-trained language model and incorporates conjunctions location information into the model's fine-tuning process. Additionally, we discuss a variant structure of residual connections to construct a residual structure that can learn critical information in the text and optimize it during training. We perform experiments on the public datasets and demonstrate that our approach enhances performance on both sentence and aspect-based sentiment analysis datasets compared to the baseline pre-trained language models. These results confirm the effectiveness of our proposed method.

1 Introduction

Sentiment analysis is a crucial area of research within the field of natural language processing. Before the advent of Transformer (Vaswani et al., 2017), Recurrent Neural Networks (RNNs) were the primary method used to model sequences in language modeling tasks (Tang et al., 2016a; Li et al., 2018; Li et al., 2019; Majumder et al., 2022). RNN, along with its variants LSTM (Long-Short Term Memory) and GRU (Gated Recurrent Unit), are powerful models for processing sequences of varying lengths and addressing long-term dependencies. However, the sequential nature of RNNs makes parallelization difficult. Transformer introduces the attention mechanism to encode the context information, which can well capture the internal correlation and ease the problem of long-term dependencies. This allows for greater parallelization and improved performance on certain tasks.

Nevertheless, since self-attention discards sequential operations when processing sequences, the position information in the sequence cannot be fully utilized. In languages such as Chinese, word order plays a crucial role in conveying grammatical meaning [0], making it important to consider the sequential nature of the language when developing natural language processing models. Word order refers to the sequence of words in a phrase or sentence, while Chinese word order is relatively fixed, and the change of word order can make the phrase or sentence express different meanings. "Speak well/说好话", "easy to speak with/好说话", and "easier said/话好说" are three Chinese phrases that demonstrate the importance of word order in conveying meaning. Although these phrases share similar characters, their meanings differ greatly depending on how those characters are arranged. "Speak well/说好话" means to speak positively or say good things about someone or something, while "easy to speak with/好说话" describes someone who is easy to communicate with. Lastly, "easier said/话好说" implies that something may sound simple or easy to do but can be more difficult in practice. It's essential to consider

*Corresponding author

[0]Higher Education Press.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

both the context and word order when interpreting or translating Chinese phrases. In addition, function words in Chinese play an important role in constructing the grammatical structure of a sentence and reflecting specific grammatical relationships. They are a crucial grammatical tool necessary for expressing meaning [1]. Among them, conjunctions connect grammatical units at different levels, and their positions in sentences are significantly different (Liu, 2016), which can be used as an essential aspect of studying syntactic distribution.

Therefore, in this paper, we propose LCEM, a learnable conjunctions augmentation model for Chinese sentiment analysis. By adjusting the structure of the pre-trained language model, LCEM introduces the conjunction position information into the fine-tuning process. The paper also explores variants of residual structure and constructs an enhanced model capable of learning critical information during training and optimization of the residual structure.

The main contributions of this paper can be summarized as follows:

- LCEM is a generic structure that can be easily integrated into a pre-trained language model based on Transformer using an adaptive update optimized network of learnable parameter factors.
- By incorporating the relative position of conjunctions in each layer of the pre-trained language model, LCEM enhances multi-head self-attention and effectively considers the sentiment range of sentences connected by conjunctions.
- Additionally, LCEM combines a learnable residual structure to better balance the network and optimize semantic representation more efficiently.
- LCEM is evaluated on benchmark datasets for sentence and aspect-based sentiment analysis. Experiments show that LCEM consistently achieves state-of-the-art performance across all test datasets.

2 Related Work

2.1 Chinese sentiment Analysis

Early Chinese sentiment analysis methods (Zhu et al., 2006; SHI Wei, 2021; Liu et al., 2015) primarily relied on sentiment lexicons, such as HowNet sentiment word dictionary and National Taiwan University Sentiment Dictionary (NTUSD), and classified sentiment polarity based on dictionaries and rules. However, these methods are limited by the quality and coverage of lexicons. The sentiment analysis in a specific field needs to construct a specific dictionary, which is time-consuming and laborious. When traditional machine learning algorithms are used in sentiment classification, different features enable different classifiers to obtain higher accuracy than dictionary methods (Xu et al., 2007; Yang and Lin, 2011; He et al., 2018). However, traditional machine learning methods rely on the quality of the annotated corpus and cannot fully use contextual semantic information.

With the rapid development of deep learning, neural network and attention mechanism have been widely concerned and applied in Chinese sentiment analysis (Cheng et al., 2019; Peng et al., 2018). Transformer with self-attention mechanism, which employs an encoder-decoder framework to better address long-term dependencies and allows for more robust scalability of parallel computations, is widely used in natural language processing. Based on the Transformer architecture, a series of landmark pre-trained language models have emerged, showing a strong ability to learn generic Chinese representations. Li (2021) fully extracted context information using improved attention to encode relative position between words based on ELMo (Peters et al., 2018). Xie(2020) used BERT to encode the set of sentiment words extracted from texts and used attention to obtain sentiment information. However, in the above studies, although the pre-trained language model has powerful modeling ability, it neglects the application of syntactic structure or semantic information in sentiment analysis and fails to use sentiment features effectively.

[1]The Commercial Press.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

2.2 Relative Position Feature

In order to leverage the sequential information contained within input text, Transformers incorporate position embeddings into the original input embedding. This process is calculated as follows:

$$\begin{aligned} PE_{(pos,2i)} &= sin(pos/10000^{(2i/d_{model})}) \\ PE_{(pos,2i+1)} &= cos(pos/10000^{(2i/d_{model})}) \end{aligned} \tag{1}$$

where pos represents position, i represents the number of dimensions, d_{model} is the input and output vector dimensions. The sines and cosines enable the model to learn the relative position and easily extend to longer sequences.

The BERT-based pre-trained language model adopts the encoder structure in Transformer and selects absolute position embedding to better adapt to downstream tasks. In the input layer, word embedding is combined with position embedding to ensure that identical words at different positions can learn representations that are appropriate for their respective contexts. Li (2021) improved attention by encoding relative positions between words. Shaw (2018) used relative encoding as an additional value in the self-attention to capture information about the relative position differences between input elements. According to different task characteristics, different position embeddings contain different meanings. For instance, in the named entity recognition task, entity term is often introduced by designing different position features (Li et al., 2020; Yan et al., 2019; Mengge et al., 2020). In the causality extraction task, position features can reflect the position of connectives and the distance between causal events and connectives (Zhao et al., 2016).

2.3 Residual Structure

Neural networks have a strong representation ability and can optimize and update the network structure through the back propagation algorithm. However, during backpropagation, gradients may either vanish or increase exponentially, resulting in ineffective updates to the underlying parameters, or gradient explosion. Furthermore, deeper networks are susceptible to degradation problems. He (2016) verified that adding more layers to a network model with a certain depth will lead to higher training errors.

Recently, residual learning has been widely used in natural language processing and computer vision as a technique for optimization of deep neural network to alleviate gradient vanishing or explosion problems (He et al., 2016; Srivastava et al., 2015; Liu et al., 2019a; Liu et al., 2021). Since each submodule of the Transformer encoder contains residual structures with layer normalization, BERT-based pre-trained variants can also make full use of residual connections to optimize the network.

This paper introduces the learnable residual structure based on enhanced self-attention by the position features of conjunctions. By assigning learnable parameters to each branch, the residual structure can be adjusted adaptively, and performance can be improved through simple model adjustment.

3 Methodology

3.1 Overview

LCEM is based on the basic architecture of the pre-trained language model. The overall structure of LCEM is described in Figure 1. LCEM uses the conjunction relative position enhanced multi-head attention to replace the multi-head attention module in each layer of the pre-trained language model. By combining the relative position feature with the attention mechanism, the model can learn global semantic information while still paying close attention to important local ranges. In addition, the residual structure of the pre-trained language model is improved to a more flexible structure to optimize the network and enable better internal information sharing. The learnable factors can adaptively control the residual structure, better integrating the semantic information learned by the relative position feature and further optimizing by assigning different importance to each residual branch.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

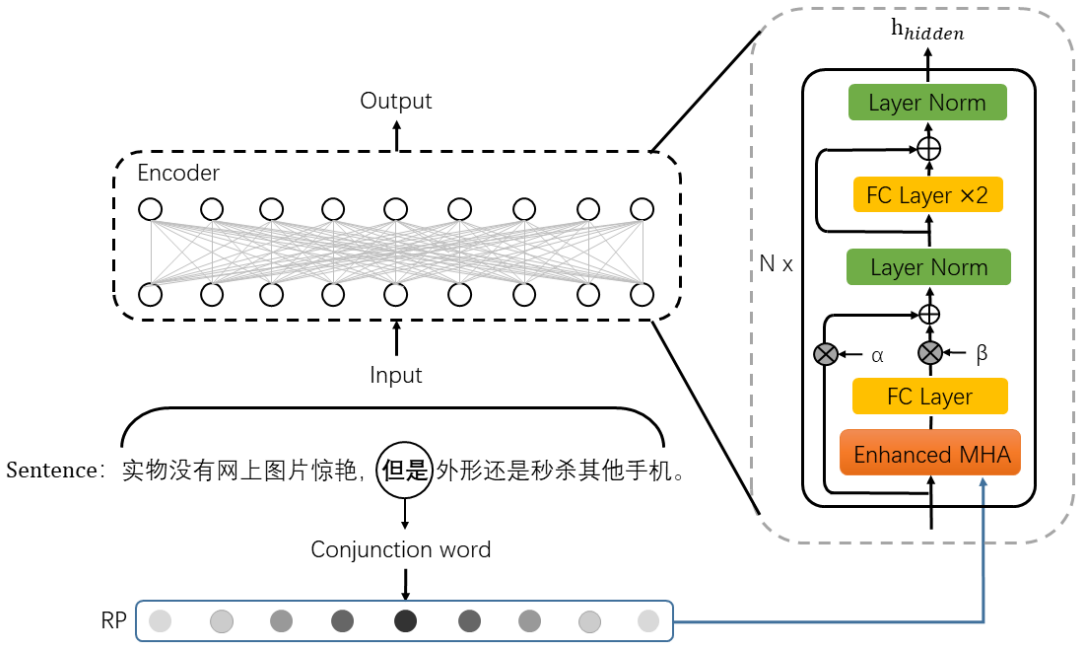


Figure 1: Overview of LCEM

3.2 Conjunction Relative Position Enhanced Multi-Head Attention

LCEM uses the relative position feature to enhance attention to learn the interaction between input text and the conjunctions representation. Conjunctions of transition, progression, selection, and coordinate are selected in the Chinese Function Word Usage Knowledge Base(CFKB) (Zan et al., 2011; Kunli et al., ; Zhang et al., 2015), and the distance $d(d \geq 0)$ between each character in a sentence and the first character of the conjunction is calculated. We map the relative position of conjunctions into the interval of $(0, 1)$ to obtain the relative position feature RP, and the calculation is as follows:

$$RP = 1 - Sigmoid(d) = 1 - \frac{1}{1 + e^{-d}} \tag{2}$$

If there is no conjunctions in the sentences, the d in the formula is the distance between each word in the sentences and the beginning of the sentences.

Then, as shown in Figure 2, RP increases the attention to the context near conjunctions. At the same time, the learnable parameter ω is introduced to reduce the noise caused by introducing the relative position feature to the original input representation H. The attention after adding the relative position feature is as follows:

$$\begin{aligned} Attetnion(Q, K, V) &= Softmax(\frac{QK^T}{\sqrt{d_K}} + \omega RP)V \\ where \quad Q &= HW^Q, K = HW^K, V = HW^V \end{aligned} \tag{3}$$

3.3 Learnable Residual Structure

Some studies (Liu et al., 2019a; Liu et al., 2021) divided the problems existing in residual connection into two types: the balance problem of each residual branch and the optimization problem. Liu (2019a) analyzed existing works and summarized the general residual structure as follows:

$$\mathcal{Y} = \alpha x + \beta \mathcal{F} + \gamma LN(x + \mathcal{F}) \tag{4}$$

Where x is the input branch, i.e., the skip connection, $\mathcal{F}$ is the residual branch, LN is layer normalization, $\mathcal{Y}$ is the output of the residual block, and α, β, γ are the weight factors. The residual block can be

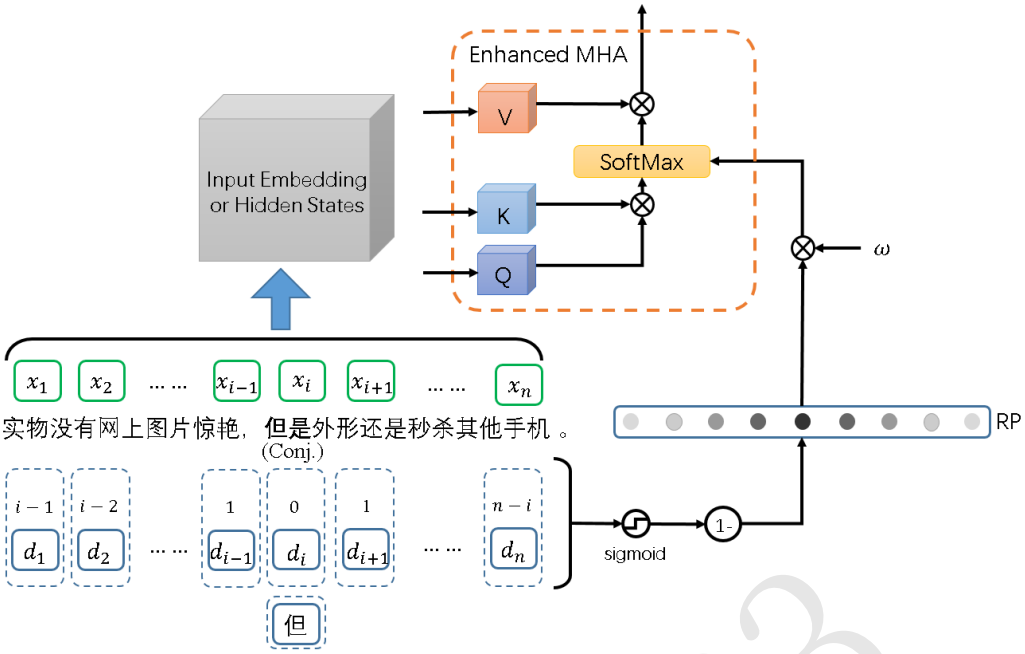


Figure 2: Details of Conjunction Relative Position Enhanced Multi-Head Attention

(a) Transformer (b) Learnable Residual Structure in LCEM

Figure 3: Residual Structure in Transformer and LCEM

adjusted and optimized adaptively by adjusting values for α, β, and γ. Liu (2021) proposed formula 5 to summarize the residual connection with normalization. Normalization $\mathcal{G}$ was placed outside the sum of input x and nonlinear transformation $\mathcal{F}(x, W)$, and λ was used to enhance the input branch.

$$\mathcal{Y} = \mathcal{G}(\lambda x + \mathcal{F}(x, W)) \tag{5}$$

Drawing inspiration from the residual structure present in every layer of the Transformer (Figure 3 (a)), layer normalization plays a crucial role in the model's overall performance. It can help the optimization of nonlinear transformation to a certain extent. And, in combination with the idea of adjusting each branch of residual in the neural network by the weight factor mentioned above, the residual structure is summarized as follows:

$$\mathcal{Y} = LN(\alpha x + \beta \mathcal{F}) \tag{6}$$

As shown in Figure 3 (b), the residual structure in Transformer can be regarded as a particular case $\mathcal{Y} = LN(x + \mathcal{F})$ when $\alpha = \beta = 1$. In Transformer, the residual branch $\mathcal{F}$ can be either multi-headed attention or feedforward networks. In this paper, we focus on the residual structure of multi-head attention. We propose to replace the residual branch with conjunctions relative position enhanced

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

Table 1: Statistical data of each category in the datasets.

Datasets	COAE2013		NLPCC2014		SemEval16_CAM		SemEval16_PHO	
	Train	Test	Train	Test	Train	Test	Train	Test
Positive	753	305	5000	1250	809	344	758	310
Negative	876	239	5000	1250	450	137	575	219

attention. Meanwhile, α and β are set as learnable parameters so that the model can self-learn appropriate scaling factors. The proportion of input branch x and residual branch $\mathcal{F}$ in the network is constantly modified to achieve optimization.

The semantic representation obtained by the enhanced attention will further learn the appropriate proportion in the propagation under the adjustment of scaling factor β, reducing the noise caused by the introduction of the relative position feature. Scaling factors α and β jointly determine the different distribution of x and $\mathcal{F}$. The layer normalization is used to make the distribution of each layer in the network relatively consistent to avoid gradient vanishing or explosion caused by the change of learnable parameters. Through multi-layer structure with learnable conjunctions enhanced attention, the final output is obtained by a linear classifier.

4 Experimental Settings

4.1 Datasets

In this paper, we study two granular subtasks in Chinese sentiment analysis. Statistical data of the above datasets are shown in Table 1.

For Chinese sentence-level sentiment analysis, COAE2013 and NLPCC2014 are selected. COAE2013 is a dataset of annotated data from The Fifth Chinese Opinion Analysis Evaluation, consisting of 1004 positive reviews and 834 negative reviews. The dataset was divided into train set and test set according to the ratio of 9:1. NLPCC2014 is from the 3rd CCF Conference on Natural Language Processing & Chinese Computing, including reviews of books, DVDs, electronic products, and other domains. The train set consisted of 5,000 positive and 5,000 negative texts, and the test set consisted of 2,500 texts.

For the Chinese aspect-based sentiment analysis task, this paper selects SemEval2016 (Pontiki et al., 2016). Task 5 of SemEval2016 provides a Chinese dataset of electronic product aspect-based reviews in two specific domains, including phone and camera, including 400 samples, a total of about 4100 sentences.

4.2 Baselines

We evaluate LCEM with typical sentiment analysis and text classification models as baselines for sentence-level sentiment analysis, including BiLSTM (Zhang et al., 2015), BiLSTM+Att (Zhang and Wang, 2015), TextCNN (Kim, 2014), DPCNN (Johnson and Zhang, 2017), and pre-trained language models like EBi-SAN (2021), BERT, BERT_wwm (Cui et al., 2021), RoBERTa (Liu et al., 2019b), ERNIE (Sun et al., 2019b). For aspect-based sentiment analysis, we compare our solution to several models that can be applied to Chinese text, including MemNet (Tang et al., 2016b), ATAE-LSTM (Wang et al., 2016), IAN (Ma et al., 2017), Ram (Chen et al., 2017), AOA (Huang et al., 2018), MGAN (Fan et al., 2018), Tnet (Li et al., 2018), and QA-B (Sun et al., 2019a) and NLI-B (Sun et al., 2019a), and also BERT and ERNIE.

The word vector pre-trained by the Sogou News corpus is selected as the initial embedding in the general baselines. The batch size is 128, the learning rate is 1E-5, and 30 epochs are trained by Adam optimization. Based on the pre-trained model, the baselines all follow the default 12 hidden layers with a size of 768, the batch size is 20, and the learning rate is 5E-5. Adam is used to optimize the cross-entropy loss function and fine-tunes the parameters.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

5 Experimental Results

5.1 Results on Sentence-level Sentiment Analysis

Table 2 shows the results of comparative experiments on the sentence-level datasets.Compared with the pre-trained model ERNIE and neural network models based on RNN and CNN, such as TextCNN and DPCNN, the results indicate that the fine-tuned pre-trained language model performs better on the datasets than the neural network models based on RNN and CNN, highlighting the huge advantage of pre-trained language models in sentiment analysis tasks. Additionally, compared to other pre-trained models, ERNIE performs better on two sentiment analysis datasets. By using relative positional encoding of conjunctions and learnable residual structures based on ERNIE, LCEM further optimized the model and improved its performance, demonstrating the effectiveness of the proposed method in this paper.

Table 2: Results on sentence-level sentiment analysis datasets.

Datasets	COAE2013		NLPCC2014	
	Acc(%)	F1(%)	Acc(%)	F1(%)
BiLSTM	85.74	85.39	60.48	60.48
BiLSTM+Att	86.91	86.76	69.60	69.56
TextCNN	89.65	89.46	69.04	68.85
DPCNN	87.30	87.07	62.48	58.88
EBi-SAN	-	-	79.08	78.48
BERT	93.57	93.53	79.61	79.61
BERT_wwm	94.88	94.83	80.21	80.20
RoBERTa	94.99	95.01	79.57	79.56
ERNIE	95.77	95.74	80.89	80.88
LCEM	**96.69**	**96.68**	**81.08**	**81.08**

5.2 Results on Aspect-based Sentiment Analysis

Experimental results are shown in Table 3 compared with aspect-based sentiment analysis baselines. Under the accuracy and F1, LCEM outperforms all baselines in SemEval16_CAM and SemEval16_PHO. The accuracy of LCEM on the SemEval16_CAM is 1.25% higher than that of ERNIE, and the F1 value is 0.72% higher than that of QA-B. Compared with IAN, MGCN, and other non-pre-trained language models, the fine-tuned results of the pre-trained model have great advantages. On the one hand, the pre-trained model has been trained on large text corpus and has learned rich language representation capabilities, which enables the pre-trained model to better understand the semantics and context of the text, which is very helpful for sentiment analysis tasks. On the other hand, pre-trained models can achieve better results on small datasets, while recurrent neural networks require large amounts of manually annotated training data, and the size of the training data will limit the performance of the model.

5.3 Ablation Study

Table 4 shows the results of LCEM ablation experiments on four datasets.

In which, $+RP$ and $+\omega RP$ respectively represent adding relative position encoding (RP) and weighted relative position encoding (Weighted RP) only in the self-attention module on top of the baseline model. Comparing $+RP$ and $+\omega RP$ with baseline ERNIE, we can see that $+\omega RP$ is better than $+RP$, improves performance on both sentence-level datasets and SemEval16_PHO. But on SemEval16_CAM, neither $+RP$ nor $+\omega RP$ can achieve effective performance enhancement, which may be because the relative position feature is added to each layer of the pre-trained language model. The output of each layer will serve as input to the next layer and participate in the residual structure. As the network depth increases, each addition of the relative position feature will introduce some noise into the original representation. Although the weighted relative position feature($+\omega RP$) introduces parameters that can learn relative positional shifts with the network structure, its effect varies on different datasets.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

Table 3: Results on aspect-based sentiment analysis datasets.

Datasets		SemEval16_CAM		SemEval16_PHO	
		Acc(%)	F1(%)	Acc(%)	F1(%)
ATAE-LSTM		87.11	82.79	79.02	78.78
MemNet		88.57	85.33	77.88	76.77
IAN		88.77	85.97	79.40	78.91
Ram		85.65	82.66	77.69	76.81
Tnet		87.32	83.47	79.77	79.14
AOA		88.36	85.52	79.58	79.21
MGAN		85.45	82.65	79.96	79.38
BERT		87.94	85.57	83.74	83.22
ERNIE		93.14	91.45	90.17	89.84
ERNIE-SPC		92.52	90.65	90.36	90.07
ERNIE-based	QA-B	92.41	92.41	89.23	89.22
	NLI-B	91.48	91.48	88.94	88.94
LCEM		**94.39**	**93.13**	**91.12**	**90.79**

Table 4: Results of ablation experiment

Datasets	COAE2013		NLPCC2014		SemEval16_CAM		SemEval16_PHO	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
Baseline(ERNIE)	95.77	95.74	80.89	80.88	93.14	91.45	90.17	89.84
$+RP$	95.96	95.94	80.76	80.75	92.93	91.77	89.60	89.17
$+\omega RP$	95.96	95.93	80.92	80.91	92.93	91.32	90.74	90.42
$+LRS$	96.51	96.49	80.40	80.40	92.31	90.74	90.17	89.83
$+RP\&LRS$	95.96	95.93	80.96	80.95	93.35	**93.35**	90.55	90.27
$+\omega RP\&LRS$(LCEM)	**96.69**	**96.68**	**81.08**	**81.08**	**94.39**	93.13	**91.12**	**90.79**

$+LRS$ represents only the learnable residual structure added to ERNIE. The comparison results also show that $+LRS$ has a slight improvement, indicating that the structure of the pre-trained language model, especially the residual structure, has the advantages of efficiency, stability, and universality.

Accuracy and macro-F1 of $+RP\&LRS$ are better than $+RP$, $+\omega RP$, and $+LRS$ in both datasets. This suggests that scaling within the residual structure can effectively adjust the enhanced multi-head attention as a branch of residual connection. In addition, the output of the previous layer serves as the skip connection branch of the residual structure of the next layer, and residual scaling can adjust the input branch and the residual branch adaptively. At the same time, it shows that enhanced attention by relative location features can capture both content and distance information, and learn richer context representation under the role of location information.

The proposed model LCEM($+\omega RP\&LRS$) achieves the highest accuracy and F1 in both sentence-level datasets. In the two datasets of SemEval16, the F1 improved by 1.68% and 0.95%, respectively, compared with baseline model ERNIE, and achieved the highest accuracy in both datasets. Compared with $+RP\&LRS$, the accuracy is significantly improved, indicating that weighted relative position encoding can achieve more effective optimization. The learnable weights during network training also reduce the noise effects introduced by relative position encoding, better capture the balance within the network and maximizing the gain of residual scaling.

5.4 Case Study

For further analysis of the model, the LCEM and ERNIE models are analyzed in this paper, as shown in Table 5.

For the adversative conjunction ”但是”, it serves as a transitional element between two sentences or clauses. It indicates a contrast or contradiction between the information presented before and after it. In the given context, the emotional tone of the sentence preceding the transition is predominantly negative. However, the emotional tone of the sentence following the use of ”但是” changes from negative to

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

Table 5: Case studies of LCEM and ERNIE models

Type of conjunction	Conjunction	Example	Model	Label
转折	但是	拿到的时候还觉得像盗版，但确实是正版的，很完整，非常不错	ERNIE	0
			LCEM	1
递进	而且	是真正的职场小说，感觉更像《圈子圈套》，而且厚厚的一大本，很值。	ERNIE	0
			LCEM	1

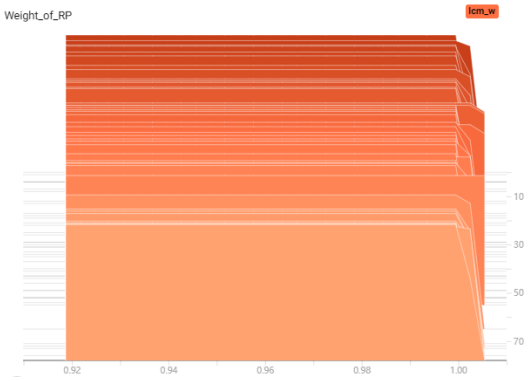


Figure 4: The parameter ω of RP over time.

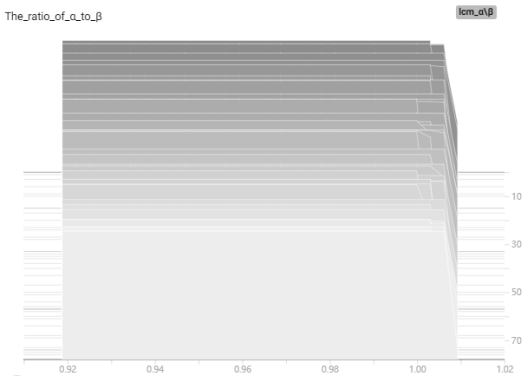


Figure 5: The ratio of α to β over time.

positive. Therefore, the emotional label of the first sentence in Table 5 is 1, signifying a shift from negative to positive emotion.

On the other hand, the coordinating conjunction ”而且” is used to connect two sentences or clauses to express a progression or addition of information. While the emotional information in the sentence before the conjunction may not be overtly expressed, it is more fully conveyed in the sentence that follows the use of ”而且.” Consequently, the emotional label of the second sentence in Table 5 is 1, indicating the enhanced expression of emotional content instead of label 0. When compared to ERNIE, LCEM, which incorporates conjunctive information, provides more accurate predictions of emotional labels.

6 Learnable Parameters Analysis

Figure 4 and Figure 5 show the changes of relative position parameter ω and α to β ratio over time. The X-axis represents the range of parameter values, while the Y-axis on the right represents the number of training steps. Each slice in the figure is a single histogram, representing the distribution of parameters in a training step. The number of training steps is gradually increased from back to front.

According to Figure 4, the learnable parameter ω of the relative position feature RP is more evenly distributed in $[0.919, 0.999]$, indicating that the relative position feature occupies a vital proportion of attention. Moreover, combined with the ablation experiment results in Section 5.4, relative location feature enhanced attention can capture both content and distance information and learn a richer context representation under the effect of location information.

Figure 5 shows that the ratio of α to β is evenly distributed in $[0.919, 1.01]$. In most cases, the proportion of input branches is smaller than that of residual branches. In each Transformer encoder, the proportion of representations from the previous layer is smaller than that of expressions enhanced by the relative position of the conjunctions. It demonstrates the significance of the semantic representation obtained through enhanced attention in the network. Moreover, input branches also play an important role in network. Through layer-by-layer propagation, the semantic representation acquired by each layer can be preserved in the lower layers and will participate in the attention mechanism to further extract abstract semantics. The learnable parameters greatly help the information transfer and optimization of network structure.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

7 Conclusion

In this paper, we introduce LCEM, a model that incorporates semantic information using relative position features of conjunctions, and guides the Chinese sentiment analysis task through adaptive residual structure.Specifically, weighted relative position features reduce the introduced noise and improve the learning ability of location-related syntactic features, which can better guide the self-attention mechanism and help the model focus on the critical sentences for semantic representation. At the same time, we propose a novel learnable residual structure based on pre-trained language models that can effectively handle the interaction between residual and input branches in an adaptive manner. Experimental results show that our method is effective in Chinese sentiment analysis, where relative position and adaptive residual structure complement each other. The relative position information helps the model to focus on crucial information for sentiment analysis, while the residual structure in each layer balances the learned knowledge within the network structure.

Acknowledgements

This research is supported by the key research and development and promotion project of Henan Provincial Department of Science and Technology in 2023: Research on Automatic Question Answering System in Science and Technology Management Based on Knowledge Graph (232102211041) and the National Social Science Foundation project: Research on Knowledge Base and Application of Modern Chinese Function Word Usage for Natural Language Processing (14BYY096). The author would like to thank the anonymous reviewers for their valuable comments and suggestions on the improvement of this paper.

References

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Yan Cheng, Z Ye, M Wang, Q Zhang, and G Zhang. 2019. Chinese text sentiment orientation analysis based on convolution neural network and hierarchical attention network. *Journal of Chinese Information Processing*, 33(01):133–142.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3433–3442.

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

Y He, S Zhao, and L He. 2018. Micro-text emotional tendentious classification based on combination of emotion knowledge and machine-learning algorithm. *J. Intell*, 37(5):189–194.

Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In Robert Thomson, Christopher Dancy, Ayaz Hyder, and Halil Bisgin, editors, *Social, Cultural, and Behavioral Modeling*, pages 197–206, Cham. Springer International Publishing.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Zhang Kunli, Zan Hongying, Chai Yumei, Han Yingjie, and Zhao Dan. Construction and application of the chinese function word usage knowledge base.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia, July. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6714–6721.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online, July. Association for Computational Linguistics.

Z Li, L Chen, and S Zhang. 2021. Chinese text sentiment analysis based on elmo and bi-san. *Application Research of Computers*, 38(8):2301–2307.

YJ Liu, SG Ju, SM Wu, and C Su. 2015. Classification of chinese texts sentiment based on semantic and conjunction. *Journal of Sichuan University Natural Science Edition*, 52:57–62.

Fenglin Liu, Meng Gao, Yuanxin Liu, and Kai Lei. 2019a. Self-adaptive scaling approach for learnable residual structure. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 862–870.

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

Fenglin Liu, Xuancheng Ren, Zhiyuan Zhang, Xu Sun, and Yuexian Zou. 2021. Rethinking skip connection with layer normalization in transformers and resnets. *arXiv preprint arXiv:2105.07205*.

Qun Liu. 2016. Study of conjunctive scope and its "special category" in modern chinese. *Journal of Jiangsu Normal University: Philosophy Social Science Edtion*, 42(85-90).

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. AAAI Press.

Navonil Majumder, Rishabh Bhardwaj, Soujanya Poria, Alexander Gelbukh, and Amir Hussain. 2022. Improving aspect-level sentiment analysis with aspectnbsp;extraction. *Neural Comput. Appl.*, 34(11):8333–8343, jun.

Xue Mengge, Bowen Yu, Tingwen Liu, Yue Zhang, Erli Meng, and Bin Wang. 2020. Porous lattice transformer encoder for Chinese NER. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3831–3841, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Haiyun Peng, Yukun Ma, Yang Li, and Erik Cambria. 2018. Learning multi-grained aspect target sequence for chinese sentiment analysis. *Knowledge-Based Systems*, 148:167–176.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana, June. Association for Computational Linguistics.

FU Yue SHI Wei. 2021. Microblog short text mining considering context:a method of sentiment analysis. *Computer Science*, 48(6A):158.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019a. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas, November. Association for Computational Linguistics.

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

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

R Xie and Y Li. 2020. Text sentiment classification model based on bert and dual channel attention. *Journal of Data Acquisition and Processing*, 35(4):642–652.

Jun Xu, Yu-Xin Ding, and Xiao-Long Wang. 2007. Sentiment classification for chinese news using machine learning methods. *Journal of Chinese Information Processing*, 21(6):95–100.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Jing Yang and Shiping Lin. 2011. Emotion analysis on text words and sentences based on svm. *Jisuanji Yingyong yu Ruanjian*, 28(9):225–228.

Hongying Zan, Kunli Zhang, Xuefeng Zhu, and Shiwen Yu. 2011. Research on the chinese function word usage knowledge base. *Int. J. Asian Lang. Process.*, 21(4):185–198.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

KL Zhang, HY Zan, YM Chai, Yingjie Han, and Dan Zhao. 2015. Survey of the chinese function word usage knowledge base. *Journal of Chinese Information Processing*, 29(3):1–8.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950.

Yan-Lan Zhu, Jin Min, Ya-qian Zhou, Xuan-jing Huang, and Li-De Wu. 2006. Semantic orientation computing based on hownet. *Journal of Chinese information processing*, 20(1):14–20.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 761-772, Harbin, China, August 3 - 5, 2023.

Improving Affective Event Classification with Multi-Perspective Knowledge Injection

Wenjia Yi, Yanyan Zhao*, Jianhua Yuan, Weixiang Zhao, Bing Qin
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{wjyi, yyzhao, jhyuan, wxzhao, qinb}@ir.hit.edu.cn

Abstract

In recent years, many researchers have recognized the importance of associating events with sentiments. Previous approaches focus on generalizing events and extracting sentimental information from a large-scale corpus. However, since context is absent and sentiment is often implicit in the event, these methods are limited in comprehending the semantics of the event and capturing effective sentimental clues. In this work, we propose a novel Multi-perspective Knowledge-injected Interaction Network (MKIN) to fully understand the event and accurately predict its sentiment by injecting multi-perspective knowledge. Specifically, we leverage contexts to provide sufficient semantic information and perform context modeling to capture the semantic relationships between events and contexts. Moreover, we also introduce human emotional feedback and sentiment-related concepts to provide explicit sentimental clues from the perspective of human emotional state and word meaning, filling the reasoning gap in the sentiment prediction process. Experimental results on the gold standard dataset show that our model achieves better performance over the baseline models.

1 Introduction

Affective Event Classification (AEC) aims at predicting the sentiment polarity of the given event. We consider events that have positive effects on people who experience them as positive events. For instance, typically positive events include *having a new crush*, *going to the bonfire*, *seeing a rainbow*. On the contrary, events that have negative effects on people who experience them are treated as negative events, such as *breaking a marriage*, *going to the funeral*, *hearing a loud noise*. Since events often trigger sentiments and sentiments are often implicit, recognizing affective events is of great values to various natural language processing applications, covering dialogue systems (Shi and Yu, 2018), question-answering systems (Oh et al., 2012), implicit sentiment analysis (Zhou et al., 2021) and opinion mining (Xu et al., 2022b).

The challenges of AEC lie in the limited context and the implicit sentiment of the event. To be specific, we often rely on the context to analyze sentiments, but there is no rich context to understand the event. Besides, traditional sentiment analysis methods rely on the occurrence of explicit sentiment words, but there are few explicit sentimental clues in the event. Many previous approaches have been devoted to cope with the challenges by extracting sentimental information from a large-scale corpus (Ding and Riloff, 2018; Saito et al., 2019; Zhuang et al., 2020). However, such attempts are not effective enough to understand the event and capture sentimental clues due to weak context modeling and insufficient sentimental information.

We believe that this task would benefit from multi-perspective knowledge injection. Specifically, context can provide additional semantic information to understand the event, and human emotional feedback as well as sentiment-related concepts can provide explicit sentimental information from two perspectives to fill the reasoning gap. Figure 1 shows an example of an affective event, demonstrating the significance

*Corresponding author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

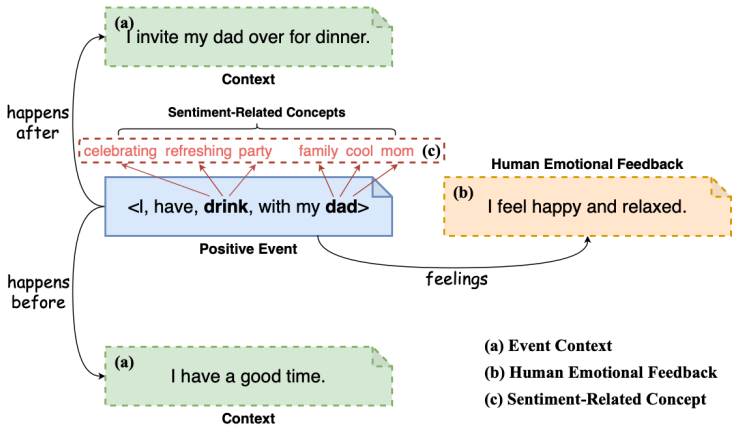


Figure 1: An example of an affective event for identifying the sentiment with the help of event contexts, human emotional feedback and sentiment-related concepts.

of contexts, human emotional feedback and sentiment-related concepts in understanding events and detecting implied sentiments. On the one hand, the given event ("⟨I, have, drink, with my dad⟩") could be thoroughly understood by supplying contexts ("I invite my dad over for dinner.", "I have a good time."). On the other hand, from people's positive emotional feedback towards the event ("happy", "relaxed"), we could know that the event would typically have a positive effect on them. Moreover, the meaning of "drink" and "dad" in the event are enriched by sentiment-related concepts ("celebrating", "party", "cool", etc.). Thus, the implicit sentiment of the event could be identified more easily via enhanced emotions and enriched concepts.

In this paper, to cope with the aforementioned challenges, we propose a novel Multi-perspective Knowledge-injected Interaction Network (MKIN) to thoroughly comprehend the event and precisely infer its sentiment. Specifically, we utilize a pre-trained generative commonsense reasoning model to create event contexts and human emotional feedback of the event. Meanwhile, a commonsense knowledge base and an emotion dictionary are adopted to retrieve sentiment-related concepts. Then, we devise a Multi-Source Text Encoding Module to encode these events and knowledge. To better integrate contextual information and sentimental clues, we construct a Semantic and Sentimental Fusion Module, which performs interaction as well as fusion of semantics and sentiments. Finally, we introduce a classifier to accurately classify affective events.

To evaluate the performance of MKIN, we conduct extensive experiments on the gold standard dataset for AEC. State-of-the-art performance is achieved by us compared with the baseline models.

The main contributions of our work are summarized as follows:

- For the first time, we propose to leverage commonsense knowledge to improve Affective Event Classification.
- We introduce a novel approach MKIN to perform context modeling and sentiment reasoning, which injects knowledge from multiple perspectives to meet the challenges of AEC.
- Extensive experimental results on the benchmark dataset demonstrate the superiority of MKIN. Our source code will be publicly available.

2 Related Work

Relevant work mainly includes two directions, one is affective event classification, and the other is incorporating external knowledge in sentiment analysis tasks.

2.1 Affective Event Classification

Prior work has focused on producing lexical resources of verbs or event phrases with corresponding sentiment polarity values. Goyal et al. (2010) created a new type of lexicon for narrative text comprehen-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

sion, consisting of patient polarity verbs that impart positive or negative states on their patients. Vu et al. (2014) created a manually-constructed dictionary of emotion-provoking events, then used seed expansion and clustering to automatically acquire and aggregate events from web data. Li et al. (2014) extracted major life events from Twitter by clustering tweets corresponding to speech act words, such as "congratulations" or "condolences". Ding and Riloff (2016) first defined stereotypical affective events as triples ⟨Agent, Verb, Object⟩ that are independent of context, and used a semi-supervised label propagation algorithm to discover affective events from Blogs.

More recently, many researches on affective event classification exert much effort to extract sentimental information from a large-scale corpus. Ding and Riloff (2018) expanded affective events as tuples ⟨Agent, Predicate, Theme, Prepositional Phrase⟩, and introduced a weakly supervised semantic consistency model for inducing a large collection of affective events from a personal story corpus. Saito et al. (2019) proposed to exploit discourse relations to propagate sentiment polarity from seed predicates. They extracted events that co-occur with seeds in a Japanese web corpus, and used discourse relations as constraints in the model learning process. Zhuang et al. (2020) first utilized the pre-trained model and presented a discourse-enhanced self-training method, which combines the classifier's predictions with information from local discourse contexts, and iteratively improves the classifier with unlabeled data. Another line of related work is Event-related Sentiment Analysis, which explicitly models events to improves sentiment analysis because events often trigger sentiments in sentences. Zhou et al. (2021) proposed a hierarchical tensor-based composition mechanism for event-centered text representation and develop a multi-task learning framework to improve sentiment analysis with event type classification.

However, existing methods only induce affective events based on semantic relations or discourse relations, or purely focus on sentimental information from local discourse contexts. Unlike the previous work, we consider multiple perspectives of knowledge, covering event contexts, human emotional feedback and sentiment-related concepts.

2.2 Sentiment Analysis with Knowledge

In recent years, there is a growing number of researches on incorporating external knowledge in various sentiment analysis tasks. Turcan et al. (2021) explored the use of commonsense knowledge via adapted knowledge models to understand implicitly expressed emotions and the reasons of those emotions for Emotion Cause Extraction. Sabour et al. (2021) leveraged commonsense knowledge to obtain more information about the user's situation and feelings to further enhance the empathy expression in the generated responses for Empathetic Response Generation. Zhao et al. (2022) utilized commonsense knowledge to provide causal clues to guide the process of causal utterance traceback for Emotion Recognition in Conversations. Peng et al. (2022) employed commonsense knowledge to obtain the psychological intention of the help-seeker to generate the supportive responses for Emotional Support Conversation. Xu et al. (2022a) used a knowledge graph to supplement a large amount of knowledge and common sense omitted in implicit emotional sentences for Implicit Sentiment Analysis.

There are also many studies on integrating external knowledge in other natural language processing tasks, but less studies on Affective Event Classification. To the best of our knowledge, this is the first attempt to introduce external knowledge into Affective Event Classification task.

3 Methodology

The problem of the AEC task could be formulated as follows. Given an event tuple ⟨ Agent: $agent = \{w_1, w_2, \cdots, w_{n_{agent}}\}$, Predicate: $pred = \{w_1, w_2, \cdots, w_{n_{pred}}\}$, Theme: $theme = \{w_1, w_2, \cdots, w_{n_{theme}}\}$, Prepositional Phrase: $prep = \{w_1, w_2, \cdots, w_{n_{prep}}\}$ ⟩ with the corresponding sentiment category, the goal of this task is to predict the sentiment distribution over three sentiment polarities.

The overall architecture of our proposed model MKIN is shown in Figure 2, which consists of four modules: Knowledge Acquisition Module, Multi-Source Text Encoding Module, Semantic and Sentimental Fusion Module, Sentiment Classification Module. Each one of the four modules will be elaborated in the rest of this section.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

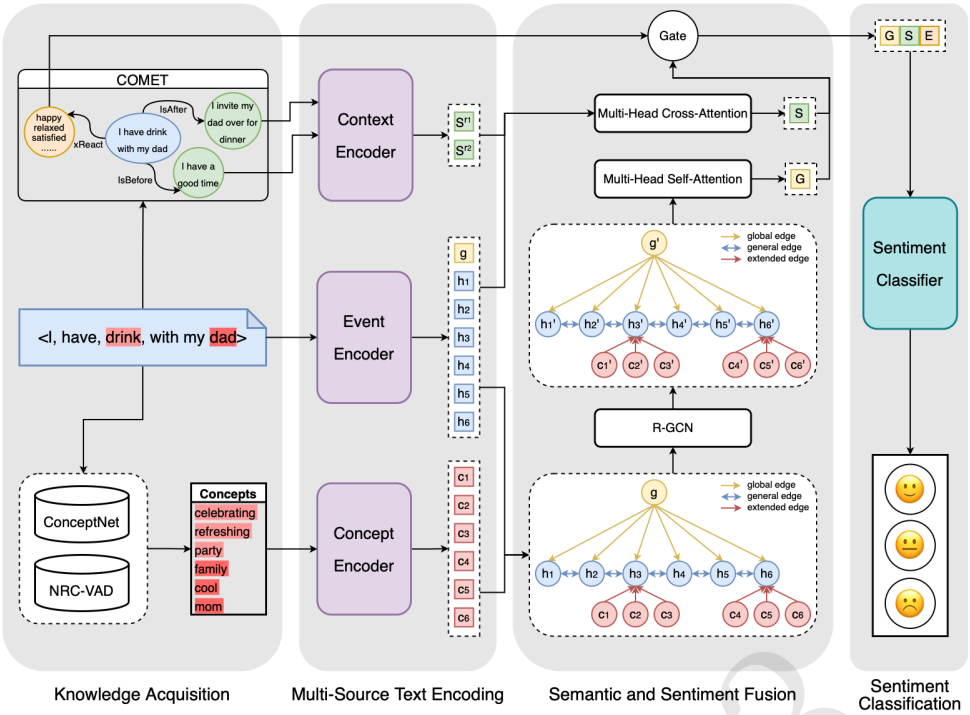


Figure 2: The overall architecture of our proposed model.

3.1 Knowledge Acquisition Module

Event Context Acquisition. Since retrieving context from corpus is expensive and noisy, we turn to the commonsense knowledge base to provide the context for the given event. In this work, we employ ATOMIC-2020 (Hwang et al., 2021) as our commonsense knowledge base, which is a commonsense knowledge graph of general-purpose everyday inferential knowledge covering social, physical, and event-centered aspects.

To be more specific, we explore two event-centered categories of commonsense knowledge from ATOMIC-2020, called *isAfter* and *isBefore*. These two relations provide reasoning about event scripts or sequences, respectively introducing events that can precede or follow an event. Therefore, we use *isAfter* and *isBefore* to introduce two events that happened before and happened after the given event, respectively. The two introduced events could form the context of a given event, and the three events can be treated as a narrative event chain. Then the given event could be fully understood via context awareness.

In order to acquire contexts for given events, we adopt a generative model COMET (Bosselut et al., 2019) which is a pre-trained GPT-2 model (Radford et al., 2018) finetuned on ATOMIC (Sap et al., 2019). More precisely, we use a BART-based (Lewis et al., 2020) variation of COMET, which is trained on ATOMIC-2020. This model can generate accurate and representative knowledge for new, unseen events. It is suitable and necessary for AEC task, because affective event has a broad scope and many events may not exist in the static ATOMIC-2020 dataset. An event is given to form the input format $(e, r, [GEN])$, where e is the sequence that comprise an event tuple. For instance, $\langle$I, have, drink, with my dad$\rangle$ is converted into the sequence "I have drink with my dad". And r is the relations we select, including *isAfter* and *isBefore*. Then we use COMET to generate five commonsense inferences for each relation r.

Human Emotional Feedback Acquisition. In this work, ATOMIC-2020 is also utilized to acquire human emotional feedback. We explore one type of social-interaction commonsense knowledge called *xReact*, which manifests the emotional states of the participants in a given event. The introduced emotion reactions could fill the reasoning gap between events and sentiments. We acquire human emotional

Dimensions	Values	Interpretations
Valence	[0,1]	Negative - Positive
Arousal	[0,1]	Calm - Excited
Dominance	[0,1]	Submissive - Dominant

Table 1: Interpretations of NRC_VAD dimensions

feedback in the same way that we acquire event contexts. As the commonsense inferences for *xReact* are usually emotion words (e.g., happy, sad, angry, etc.) rather than events or sentences, we simply adopt the hidden state representation from the last encoder layer of COMET as the human emotional feedback representation.

Sentiment-Related Concept Acquisition. Following (Li et al., 2022b), we use a commonsense knowledge base ConceptNet (Speer et al., 2017) combined with an emotion lexicon NRC_VAD (Mohammad, 2018) to obtain sentiment-related concepts.

ConceptNet is a large-scale multilingual semantic graph proposed to describe general human knowledge, allowing natural language applications to better understand the meanings behind the words. We introduce the tuple *(head concept, relation, tail concept, confidence score)* to represent the assertions in ConceptNet graph and their associated confidence scores, and denote the tuple as $\tau = (h, r, t, c)$. For instance, one such tuple from Conceptnet is *(birthday, RelatedTo, happy, 4.16)*. Let W be a collection of words in a given event tuple. For each non-stopword $h \in W$, we retrieve a set of tuples $T_i = \left\{\tau_i^j = \left(h_i, r_i^j, t_i^j, c_i^j\right)\right\}$ containing its immediate neighbors from ConceptNet, where i, j are indices of non-stopwords and the retrieved tuples.

To refine the retrieved concepts, we first remove tuples where concepts t_i^j are stopwords or not in our vocabulary. We further filter tuples where confidence scores c_i^j are smaller than 1 to reduce annotation noises. As many of the tuples are still useless for our AEC task, we select 10 relevant relations from 38 relations in ConceptNet as (Liao et al., 2022) did, they analyzed the effects of various relations on implicit sentiment analysis in detail. And then we remove the tuples where relations r_i^j belong to other relations.

To highlight sentimental information, we adopt NRC_VAD to measure sentimental intensity of the external concepts. NRC_VAD is a lexicon with valence, arousal, and dominance (VAD) scores. The interpretations of three dimensions are presented in Table 1. Such as the VAD score vector $[V_a, A_r, D_o]$ of word "happy" is $[1.000, 0.735, 0.772]$. Following (Zhong et al., 2019), sentimental intensity value of a concept x is computed as:

$$\eta(x) = min\text{-}max\left(\left\|V_a(x) - \frac{1}{2}, \frac{A_r(x)}{2}\right\|_2\right) \tag{1}$$

where $min\text{-}max()$ denotes min-max normalization, $\|.\|_k$ denotes L_k norm, $V_a(x)$ and $A_r(x)$ denote the valence and arousal scores in VAD vector of concept x, respectively. For concept x not in NRC_VAD, $\eta(x)$ will be set to 0. We rank the tuples according to the sentimental intensity values $\eta(t_i^j)$ of concepts t_i^j. Based on the order of tuples, we reserve at most three external concepts with adequate sentimental intensity values (i.e., $\eta(t_i^j) \geq 0.6$) for each word h.

3.2 Multi-Source Text Encoding Module

Multi-source text encoder considers text from three sources, including raw event, event context and external concepts. The event encoder, context encoder and concept encoder are the same encoder, which employ widely-used pre-trained model BERT (Devlin et al., 2019).

Firstly, the events are encoded. For each event $e = \{w_1, w_2, \cdots, w_n\}$, we concatenate two special tokens $[CLS]$ and $[SEP]$ to the beginning and end of the event. Then the sequence $\{[CLS], w_1, w_2, \cdots, w_n, [SEP]\}$ is fed to the encoder, leading to a series of hidden states:

$$h_i = \text{BERT}(\,[CLS], w_1, w_2, \cdots, w_n, [SEP]\,) \tag{2}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

where $h_i \in \mathbb{R}^{d_m}$ is the i-th token in the input sequence, d_m is the dimension of hidden states in BERT. And the vectorized representation of an event is H. It is worth noting that we specifically denote $[CLS]$ token as g.

Secondly, the contexts are encoded. For both *isAfter* and *isBefore* , we concatenate the five generated commonsense inferences to get a context sequence CS^r:

$$CS^r = cs_1^r \oplus cs_2^r \cdots \oplus cs_5^r \tag{3}$$

where $\oplus$ is the concatenation operation, r is the relation we select from ATOMIC-2020. Then we pass each context sequence in the same input format as $\{[CLS], CS_r, [SEP]\}$, to derive a series of hidden states from the last layer:

$$s_j^r = \text{BERT}(\ [CLS], CS_r, [SEP]\ ) \tag{4}$$

where $s_j^r \in \mathbb{R}^{d_m}$ is the j-th token in the input sequence. And the vectorized representation of a context sequence is S^r.

Thirdly, the concepts are encoded. For each concept x, we perform mean-pooling operation from the last hidden layer to obtain its representation $c \in \mathbb{R}^{d_m}$:

$$c = \text{Mean-pooling}(\ \text{BERT}(\ [CLS], x, [SEP]\ )\ ) \tag{5}$$

3.3 Semantic and Sentimental Fusion Module

As contexts and sentimental clues have been collected, Semantic and Sentimental Fusion Module is devised to perform interaction as well as fusion of contextual and sentimental information.

Semantic Interaction. In order to highlight the more important semantic features from the contexts, we utilize multi-head cross-attention mechanism (Vaswani et al., 2017) to achieve the interaction of contexts and the event. Then for each context sequence CS^r, a context-aware representation $S^{r\prime}$ is learned:

$$S^{r\prime} = \text{MH}(f(H), f(S^r), f(S^r)) \tag{6}$$

where f is a linear transformation, each vector is transformed to the dimension of d_h with f, and

$$\text{MH}(Q, K, V) = \text{Concat}(head_1, \cdots, head_h)W^O \tag{7}$$

$$head_i = \text{Att}(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

$$\text{Att}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{9}$$

where Q, K, and V are sets of queries, keys and values, respectively, the projections are parameter matrices $W^O \in \mathbb{R}^{md_v \times d_h}$, $W_i^Q \in \mathbb{R}^{d_h \times d_k}$, $W_i^K \in \mathbb{R}^{d_h \times d_k}$, $W_i^V \in \mathbb{R}^{d_h \times d_v}$, and $d_k = d_v = d_h/h$. The final context representation S is obtained by:

$$S = \bigoplus_{r \in \{isAfter, isBefore\}} \text{Max-pooling}(S^{r\prime}) \tag{10}$$

then S is transformed to the dimension of d_h with a linear projection.

Sentimental Interaction. We construct a graph network for modeling the event and relevant concepts. Specifically, each event token and concept are represented as vertices in the graph, including the $[CLS]$ token as a global vertex for aggregating information. Furthermore, three relation types of edges are applied to connect the vertices: (1) *global edge*, a directed edge which connects global node to each event node; (2) *general edge*, an undirected edge between two successive event nodes; (3) *extended edge*, a directed edge which connects a concept node to the corresponding event node.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ denotes our graph, where $\mathcal{V}$, $\mathcal{E}$, and $\mathcal{R}$ are sets of vertices, edges and relation types, respectively. We initialize each vertex with the corresponding encoded feature vector, and denote vertex features as $V = \{g, h_1, \cdots, h_n, c_1, \cdots, c_m\} = \{v_1, v_2, \cdots, v_N\}$.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

We feed the initial vertex features into a graph encoder to propagate semantic and sentimental information. Considering different relation types of edges, we adopt relational graph convolutional networks (Schlichtkrull et al., 2018) to update vertex representations. The convolutional computation for a vertex at the $(l+1)$-th layer which takes the representation $v_i^{(l)}$ at the l-th layer as input is defined as:

$$v_i^{(l+1)} = \text{ReLU}\left(\sum_{r\in\mathcal{R}}\sum_{v\in\mathbb{N}_i^r}\frac{1}{|\mathbb{N}_i^r|}W_r^{(l)}v_i^{(l)}\right) \tag{11}$$

where ReLU (Agarap, 2018) is an activation function, $\mathbb{N}_i^r$ is the set of neighbor vertices under relation type r, and $W_r^{(l)}$ are relation-specific learnable parameters at the l-th layer.

To selectively attend to the more important sentimental features within the enriched event representation, we pass the updated vertex features to a multi-head self-attention layer, then a sentiment-enhanced representation V' is learned:

$$V' = \text{MH}(V^{(L)}, V^{(L)}, V^{(L)}) \tag{12}$$

where $V^{(L)} = \left\{v_1^{(L)}, v_2^{(L)}, \cdots, v_N^{(L)}\right\} = \{g', h_1', \cdots, h_n', c_1', \cdots, c_m'\}$ are the outputs of the graph encoder. We take the global vertex feature as the final event representation G, and G is transformed to the dimension of d_h with a linear projection.

Feature Fusion. We first transform the human emotional feedback representation E to the dimension of d_h via a linear transformation. Inspired by (Liu et al., 2021), we fuse the three representations with a gated manner, including context representation S, event representation G, and human emotional feedback representation E. The gate is formulated as:

$$G_S = \text{ReLU}(\text{FC}([G, S, G-S, G\odot S])) \tag{13}$$

$$G_E = \text{ReLU}(\text{FC}([G, E, G-E, G\odot E])) \tag{14}$$

$$p = \text{Sigmoid}(\text{FC}[G_S, G_E]) \tag{15}$$

where FC is a fully-connected layer and $[.,.]$ means concatenation. Then the three features are fused as:

$$F = G + p\odot S + (1-p)\odot E \tag{16}$$

3.4 Sentiment Classification Module

Finally, taking the above fused representation as input, a sentiment classifier is applied to predict the sentiment of the event:

$$\hat{y} = \text{Softmax}(\text{MLP}(F)) \tag{17}$$

where MLP is a multi-layer perception.

Cross entropy loss is adopted to train the model, the loss function is defined as:

$$\mathcal{L} = -\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{C} y_i^j \cdot log(\hat{y_i}^j) \tag{18}$$

where T and C denote the number of training examples and the number of sentiment categories, respectively, and y_i^j represents the ground-truth label.

4 Experiments

In this section we present the dataset, evaluation metrics, baseline models, model variants, and other experimental settings.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

Category	Number
Negative Event	348
Neutral Event	717
Positive Event	435

Table 2: Dataset statistics

4.1 Dataset and Evaluation Metrics

We conduct experiments on the gold standard dataset for AEC. It is collected from Twitter Dataset with sentiment category labels annotated by (Zhuang et al., 2020), and the sentiment categories belong to negative, neutral and positive. Statistics of the dataset are shown in Table 2.

Following (Zhuang et al., 2020), we report the precision, recall and F1 score for each of the three categories, and weighted average results for each metric.

4.2 Baselines and Comparison Models

We compare our proposed model with the following method:

BERT-base/large (Devlin et al., 2019): BERT is a widely-used pre-trained language model with excellent performance in various natural language processing tasks. We adopt the base version and the large version of BERT as the basis for our classifier and perform fine-tuning during the training process.

RoBERTa-base/large (Liu et al., 2019): RoBERTa has the same model architecture as BERT but with a robustly optimized pre-training scheme allowing it to generalize better to downstream tasks. Similarly, we adopt the base version and the large version of RoBERTa for experiments.

DEST (Zhuang et al., 2020): DEST is a discourse-enhanced self-training model which is the state-of-the-art model for AEC. It introduces BERT-base model for classification and combines the classifier's predictions with information from local discourse contexts to iteratively assign high-quality labels to new training instances.

4.3 Implementation Details

Following (Zhuang et al., 2020), we performed 10-fold cross-validation over the dataset, where each of the 10 runs used 8 folds of the data for training, 1 fold of the data for validation and tuning, and 1 fold of the data for testing.

Base version of BERT is adopted as the encoder, and the dimension of hidden states d_m in the encoder is 768. For all representations in the rest of our model, the dimension d_h is set to 300. For the multi-head cross-attention layer and the multi-head self-attention layer, the number of attention head is 5 and 12, respectively. For sentiment classification, the dimensions of MLP are set to [300, 100, 3] and the dropout rate is set to 0.1. We train our model with AdamW optimizer in a learning rate of 1e-5 and a linear warmup rate of 0.1. And the batch size is set to 8. We implemented all models in PyTorch with a single Tesla V100 GPU. Reported results are medians over 5 times of 10-fold cross-validation with the same 5 distinct random seeds.

5 Results and Analysis

In this section we present model evaluation results, ablation study, and case study.

5.1 Overall Results

As depicted in Table 3, our proposed model achieves state-of-the-art results. Benefiting from the effective context modeling with event contexts and accurate sentiment reasoning with human emotional feedback and sentiment-related concepts, MKIN achieves the best results on each metrics and the highest F1 score in each category compared with the state-of-the-art model DEST and other baselines.

For the state-of-the-art model DEST, we reproduce the performance in the same setting as the original model. Although DEST utilizes a large number of coreferent sentiment expressions to provide explicit sentiment clues, it is unreliable because coreferent sentiment expressions are quite noisy due to imperfect

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

Model	NEG			NEU			POS			P	R	F1
	P	R	F1	P	R	F1	P	R	F1			
BERT-base (110M)	71.6	77.4	74.1	76.5	78.1	77.1	76.8	69.9	73	75.8	75.3	75.3
BERT-large (340M)	72.5	75.5	73.5	76.7	78.6	77.5	77.6	72	74.3	76.4	75.7	75.7
RoBERTa-base (125M)	73.4	74.9	73.6	78	79.6	78.7	78.3	74.5	76	77.3	76.9	76.8
RoBERTa-large (355M)	74.4	75.3	74.5	77.7	82.3	79.8	78.9	71.7	74.8	77.7	77.3	77.2
DEST (110M)	78.9	**77.6**	78	78	83.7	80.6	**80.2**	71.5	75	79.2	78.6	78.5
MKIN (ours) (110M)	**83.7**	75.9	**79.1**	**80.1**	**84.3**	**82**	80	**78.8**	**79.1**	**81.3**	**80.7**	**80.6**

Table 3: Performance of all models. The best results among all models are highlighted in **bold**.

Model	P	R	F1
MKIN	**81.3**	**80.7**	**80.6**
w/o Event Context	79.9	79.3	79.2
w/o Human Emotional Feedback	79.7	79	78.9
w/o Sentiment-Related Concept	79.4	78.8	78.8
w/o R-GCN	79.8	78.9	78.8
w/o Gate	79.7	79.1	79.1

Table 4: Results of ablation study on model components.

coreference and issues like sarcasm, which leads to low-quality pseudo labels, even if an additional event classifier is introduced. Instead of retrieving information from corpus, we turn to the commonsense knowledge base for context information and explicit sentiment clues. MKIN improves precision of negative events from 78.9 to 83.7 and improves recall of positive events from 71.5 to 78.8. The substantial gain demonstrates the effectiveness of injecting multi-perspective knowledge to improve affective event classification, and shows the strong ability of our Semantic and Sentimental Fusion Module in extracting important features for enriching the event representation.

For other baselines models, they are not comparable with our proposed model MKIN. It suggests that the event representations extracted by pre-trained language models are not sufficient for classification, and only slight improvements are gained when a larger model is adopted. Besides, two instructive conclusions can be derived. On the one hand, it is of great significance to perform context modeling and capture semantic relationships between events and contexts, which lead to the thorough understanding of the events. On the other hand, explicit sentiment clues provided by human emotional feedback and sentiment-related concepts can fill the reasoning gap between events and sentiments.

5.2 Ablation Study

To gain better insight into the performance of our proposed model MKIN, we conduct an ablation study to verify the contributions of its main components.

Results in Table 4 show that each component is beneficial to the final performance. First, when the Event Context component is removed, the semantic features of the context are not integrated in the final representation of the event. The performance of the model degrades to a certain extent, which proves that context modeling is crucial to AEC. Since there is very limited information in the event, the model needs additional semantic information from the context for better event representation learning. Second, when removing the Human Emotional Feedback component, human's feelings are not taken into account. The dropped results demonstrate that human emotional feedback are powerful sentimental signals. Third, when the Sentiment-Related Concept component is removed, external concepts are not introduced to expand the original word meaning. The performance of the model decreased even more, which suggests that sentiment-related concepts have a considerable impact on sentiment classification. Introducing external sentimental commonsense knowledge and enriching the meaning of words in events can help the model detect implicit sentiments. Besides, the use of R-GCN enables more accurate capturing of interactive information, while the gate can better fuse complementary information.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

Event & Label	Event Context	Human Emotional Feedback	Sentiment-Related Concept
⟨I, go, -, on date⟩ Positive	I meet a girl. I have a great time.	happy, excited, romantic	go → energy, travel, journey date → lover, engagement
⟨I, have been, -, at hospital⟩ Negative	I was in a car accident. I was released from hospital.	worried, sick, scared	hospital → death, disease, injury
⟨I, save, much money, -⟩ Positive	I work hard at my job. I buy a new car.	happy, satisfied, proud	save → rescue, protect money → rich, reward, earnings
⟨-, separate, child, from family⟩ Negative	Parents go to jail. Mother cries.	sad, unhappy, scared	separate → divorce, abduction child → cute, naughty, noisy family → fellowship, mother
⟨I, have, free weekend, -⟩ Positive	I work all week. I go to the beach to relax.	relaxed, happy, excited	free → fun, gift, independent
⟨I, hear, loud noise, -⟩ Negative	I am walking down the street. I call the police.	scared, alarmed, alert	loud → strong, nightclub, vulgar noise → explosion, bang, trouble

Table 5: Cases that our model makes the correct predictions.

5.3 Case Study

We provide several cases from the Twitter dataset to analyze the influence brought by event contexts, human emotional feedback and sentiment-related concepts. As illustrated in Table 5, the injected knowledge provide interpretable results for the prediction of our model. For events that do not contain sentiment words, such as those listed in Table 5, baseline models tend to classify them as neutral, whereas our model gives the correct predictions. From the cases, it can be observed that the three perspectives of knowledge injection play different roles in sentiment prediction. In most cases, intuitively, context information is of relatively little help in the reasoning process, because the model often does not get direct sentiment-related information from context. However, context information helps the model understand the event, which enriches the original event semantics. Moreover, compared with the other two kinds of information, human emotional feedback brings stronger sentimental signals. Especially when external concepts do not provide obvious sentimental clues, human emotional feedback plays a greater role. Finally, with the help of sentiment-related concepts, the model gains more profound insight into the meaning of words in the event. Since the sentiment of an event is often derived from its predicate and entities, important sentimental clues can be obtained from the extended concepts. Then the implied sentiment can be inferred more easily and more accurately.

6 Conclusion and Future Work

In this paper, we propose a novel Multi-perspective Knowledge-injected Interaction Network (MKIN) for affective event classification. MKIN models various aspects of information by considering contexts, human emotional feedback, and sentiment-related concepts, to fully comprehend the event as well as accurately predict its sentiment. To be more specific, in order to complement the semantic information of the event, we leverage context information and perform context modeling to capture the semantic association between the event and the context. To enhance the sentimental information of the event, we take advantage of human emotional feedback to provide sentimental clues from the perspective of people's emotional state. In addition, external sentiment-related concepts are introduced to enrich the word-level representations. Both emotional state information and concept information fill the reasoning gap between events and sentiments. Experiment results show that knowledge injection from all perspectives improve the model performance, and our model achieves 2.1% performance improvement over the state-of-the-art model on the gold standard dataset.

For future work, to apply the model in a variety of natural language processing applications, we would like to explore event-centered sentiment analysis. Affective event classification can be employed as an additional subtask to improve sentiment analysis.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key RD Program of China via grant 2021YFF0901602 and the National Natural Science Foundation of China (NSFC) via grant 62176078.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July. Association for Computational Linguistics.

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

Haibo Ding and Ellen Riloff. 2016. Acquiring knowledge of affective events from blogs using label propagation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Haibo Ding and Ellen Riloff. 2018. Weakly supervised induction of affective events by optimizing semantic consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86, Cambridge, MA, October. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from Twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1997–2007, Doha, Qatar, October. Association for Computational Linguistics.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022b. Knowledge bridging for empathetic dialogue generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10993–11001. AAAI Press.

Jian Liao, Min Wang, Xin Chen, Suge Wang, and Kai Zhang. 2022. Dynamic commonsense knowledge fused method for chinese implicit sentiment analysis. *Inf. Process. Manag.*, 59(3):102934.

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

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2021. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13406–13414. AAAI Press.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia, July. Association for Computational Linguistics.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang. 2012. Why question answering using sentiment analysis and word classes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 368–378, Jeju Island, Korea, July. Association for Computational Linguistics.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4324–4330. International Joint Conferences on Artificial Intelligence Organization, 7. Main Track.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. Cem: Commonsense-aware empathetic response generation. In *AAAI Conference on Artificial Intelligence*.

Jun Saito, Yugo Murawaki, and Sadao Kurohashi. 2019. Minimally supervised learning of affective events using discourse relations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5758–5765, Hong Kong, China, November. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia, July. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Elsbeth Turcan, Shuai Wang, Rishita Anubhai, Kasturi Bhattacharjee, Yaser Al-Onaizan, and Smaranda Muresan. 2021. Multi-task learning and adapted knowledge models for emotion-cause extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3975–3989, Online, August. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hoa Trong Vu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Acquiring a dictionary of emotion-provoking events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 128–132, Gothenburg, Sweden, April. Association for Computational Linguistics.

Minghao Xu, Daling Wang, Shi Feng, Zhenfei Yang, and Yifei Zhang. 2022a. KC-ISA: An implicit sentiment analysis model combining knowledge enhancement and context features. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6906–6915, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Ruoxi Xu, Hongyu Lin, Meng Liao, Xianpei Han, Jin Xu, Wei Tan, Yingfei Sun, and Le Sun. 2022b. ECO v1: Towards event-centric opinion mining. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2743–2753, Dublin, Ireland, May. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4524–4530. International Joint Conferences on Artificial Intelligence Organization, 7. Main Track.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China, November. Association for Computational Linguistics.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit sentiment analysis with event-centered text representation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2020. Affective event classification with discourse-enhanced self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5608–5617, Online, November. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 773-785, Harbin, China, August 3 - 5, 2023.

Enhancing Implicit Sentiment Learning via the Incorporation of Part-of-Speech for Aspect-based Sentiment Analysis

Junlang Wang, Xia Li*, Junyi He, Yongqiang Zheng and **Junteng Ma**
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China
{junlangwang,xiali}@gdufs.edu.cn

Abstract

Implicit sentiment modeling in aspect-based sentiment analysis is a challenging problem due to complex expressions and the lack of opinion words in sentences. Recent efforts focusing on implicit sentiment in ABSA mostly leverage the dependency between aspects and pretrain on extra annotated corpora. We argue that linguistic knowledge can be incorporated into the model to better learn implicit sentiment knowledge. In this paper, we propose a PLM-based, linguistically enhanced framework by incorporating Part-of-Speech (POS) for aspect-based sentiment analysis. Specifically, we design an input template for PLMs that focuses on both aspect-related contextualized features and POS-based linguistic features. By aligning with the representations of the tokens and their POS sequences, the introduced knowledge is expected to guide the model in learning implicit sentiment by capturing sentiment-related information. Moreover, we also design an aspect-specific self-supervised contrastive learning strategy to optimize aspect-based contextualized representation construction and assist PLMs in concentrating on target aspects. Experimental results on public benchmarks show that our model can achieve competitive and state-of-the-art performance without introducing extra annotated corpora.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) aims to identify the sentiment polarities towards specific aspects in sentences. For example, in the sentence "*The dessert is incredible but the service is terrible,*" the sentiment polarities towards the aspects "*dessert*" and "*service*" are *positive* and *negative* respectively.

Previous work on aspect-based sentiment analysis has focused on explicit sentiment expression for specific aspect terms. It means that the sentiment polarities towards the aspects can be explicitly revealed by opinion words. e.g., the sentence "*The dessert is incredible*" contains the opinion word "*incredible*" which carries the positive sentiment towards the corresponding aspect "*dessert*". Many studies have been proposed and achieved promising results towards this task, such as attention mechanism-based methods (Ma et al., 2017; Huang et al., 2018; Zhang et al., 2019a; Wu et al., 2022), graph neural network-based methods (Zhang et al., 2019b; Wang et al., 2020; Liang et al., 2022; Zheng et al., 2023), and pre-trained language model-based methods (Song et al., 2019; Phan and Ogunbona, 2020; Dai et al., 2021; Cao et al., 2022).

However, due to the diversity and flexibility of natural language, sentences containing implicit sentiment expressions are common in human speech. For implicit sentiment, we refer to the recognition of subjective textual units where no polarity markers, opinion words or obvious descriptions are present but people are still able to state whether the text portion under analysis expresses the sentiment (Russo et al., 2015). As shown in Table 1, the four sentences can clearly express the sentiment without any opinion words. Taking the second sentence as an example, no opinion words can be found to determine the sentiment polarities towards the aspects "*food*", but people can still recognize that its polarity is negative. Additionally, we find that some complex expressions, such as factual statements and rhetorical

*Corresponding author: xiali@gdufs.edu.cn

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Domain	Example	Polarity
Restaurant	(1) The **waiters** even forget their high-tipping regulars.	negative
	(2) They're a bit more expensive than typical, but then again, so is their **food**.	positive
Laptop	(3) My **voice recording** sounds like interplanetary transmissions in Star Wars.	negative
	(4) Can you buy any laptop that matches the **quality** of a MacBook?	positive

Table 1: Several examples of reviews with implicit sentiment expressions about laptops and restaurants where aspects are marked in bold. The "**Polarity**" column indicates the sentiment polarities of aspects.

techniques, are often used to express implicit sentiment, which always contains complex semantics. For example, sentence (1) and sentence (4) in Table 1 are factual statement and rhetorical question respectively. These complex expressions and the absence of opinion words make it more challenging to detect the implicit sentiment of sentences in the ABSA task.

Few previous studies have paid more attention to the implicit sentiment in ABSA. Among them, Yang et al. (Yang and Li, 2021) propose a local sentiment aggregation paradigm for learning the implicit sentiments in a local sentiment aggregation window. Li et al. (Li et al., 2021b) adopt supervised contrastive pre-training on large-scale sentiment annotated corpora to capture both implicit and explicit sentiment orientation towards aspects in reviews. Their results demonstrate promising performance. However, we argue that the complex implicit expressions can be handled with the help of linguistic knowledge. Motivated by the applications of Parts of Speech (POS) in ABSA (Phan and Ogunbona, 2020; Gong et al., 2020) and opinion mining (Dey and Haque, 2008), we suppose that POS-based linguistic knowledge has the potential to enhance implicit sentiment learning in ABSA. Intuitively, specific POS categories imply the orientation of sentiment polarity. As shown in Figure 1, although the sentence lacks opinion words, the verbs also carry rich sentiment information (Chesley et al., 2006; Nicholls and Song, 2009). The verb "*runs*" states the fact about "*virus scan*" without more related descriptions of this aspect. However, "*flickers*" shows the problem of the aspect "*display screen*". The polarities of the sentiments towards them should be neutral and negative, respectively. Such heuristics motivate us to incorporate POS-based linguistic knowledge into ABSA models for enhancing implicit sentiment prediction.

review:	My	computer	**runs**	a	*virus*	*scan*	but	the	*display*	*screen*	**flickers**
POS:	PRON	NOUN	**VERB**	DET	NOUN	NOUN	CONJ	DET	NOUN	NOUN	**VERB**

Figure 1: Review example with its corresponding POS sequence, marked with Universal POS tags (Petrov et al., 2012). The aspect terms and the verbs are marked in italics and bold.

Inspired by the exploitation of prompts (Li et al., 2021a; Ma et al., 2022) and linguistic knowledge in ABSA (Kiritchenko et al., 2014; Phan and Ogunbona, 2020), we propose a PLM-based, linguistically enhanced framework for aspect-based sentiment analysis that incorporates part-of-speech. We first design a template with POS sequences as PLMs' input. With the multi-head self-attention mechanism, PLMs based on the transformer architecture are able to pay attention to the POS tags and their context information (Vaswani et al., 2017), thereby acquiring potential sentiment knowledge from POS sequences. Considering that the POS sequences are essentially the ordered permutations of the POS tags corresponding to the input sequences and not independent natural language sentences, we leverage token-POS alignment to minimize the semantic impact of POS sequences. In addition, motivated by the applications of contrastive learning to optimize the sentence embeddings derived from BERT (Gao et al., 2021; Yan et al., 2021b; Jiang et al., 2022), an aspect-specific self-supervised contrastive learning strategy is proposed to enhance the construction of contextualized representations, which would focus on aspect-related words in context and the target aspects when handling reviews with multiple aspects. We carry out the experiments on the SemEval 2014 (Pontiki et al., 2014) and Twitter (Dong et al., 2014) benchmark datasets. The experimental results demonstrate the efficacy of our proposed framework.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

The main contributions of this work are as follows:

- We analyze the feasibility of incorporating Part-of-Speech to assist PLMs in modeling implicit sentiment and design an input template for PLMs to focus on both aspect-related contextualized features and POS-based linguistic features.
- We propose the token-POS alignment to reduce the influence of POS sequences on semantics. Additionally, the proposed aspect-specific self-supervised contrastive learning can optimize aspect-based contextualized representations construction and help PLMs concentrate on target aspects.
- Experimental results show the effectiveness of our method, which boosts PLMs to achieve competitive and state-of-the-art performance in ABSA with fewer additional parameters.

2 Related work

In this section, we will briefly review the studies on aspect-based sentiment analysis from three perspectives: methods based on attention mechanisms, graph neural networks (GNNs), and pre-trained language models (PLMs). Then we will introduce implicit sentiment study.

ABSA methods based on attention mechanism. The majority of early attention mechanism-based methods construct the relationship between context and aspects to tackle the ABSA task. Wang et al. (2016) and Ma et al. (2017) equip neural networks with attention mechanisms, promoting the model's ability to identify related information about aspects from input reviews. Li et al. (2018) propose a framework that combines contextual features with word representations. Except for concentrating on the relationship between context and aspects. Zhang et al. (2019b) exploit syntactic features from dependency and mark each word in reviews by proximity values.

ABSA methods based on GNNs. ABSA has demonstrated excellent performance in extracting syntactic features from graph structure since the development of graph neural networks. Sun et al. (2019) and Zhao et al. (2020) use the Graph Convolutional Network (GCN) with the dependency graph to model the dependencies of input sentences. Wang et al. (2020) leverage distances between words in the dependency tree and syntactic tags simultaneously to extract syntactic features by the Graph Attention Network (GAT). Xu et al. (2023) propose to divide sentences into structural scopes according to the results of constituency parsing, which improve the performance of GCN in ABSA.

ABSA methods based on PLMs. The emergence of pre-trained language models in recent years has given ABSA methods a new trend. On the one hand, in order to reduce the gap between pre-training and fine-tuning, numerous works propose sentiment-aware pre-training tasks (Yin et al., 2020; Ke et al., 2020; Fan et al., 2022) based on capturing sentiment semantics and incorporating external knowledge (Baccianella et al., 2010). On the other hand, recent efforts to help PLMs overcome the disadvantages of aspect-aware sentiment perception are flourishing. Cao et al. (2022) remove the sentiment bias of aspect terms and proposes a model trained with differential sentiment loss that is based on the model of Song et al. (2019). Ma et al. (2022) design three aspect-specific input transformations for BERT and RoBERTa that enable the enhancement of aspect-specific context modeling. Moreover, other PLM-based methods solve the ABSA task from the perspective of machine reading comprehension (Xu et al., 2019) and natural language generation (Yan et al., 2021a).

For handling implicit sentiment in ABSA, Li et al. (2021b) propose supervised contrastive pre-training that facilitates BERT in learning sentiment knowledge from large-scale sentiment-annotated corpora. The representation of implicit sentiment expressions is aligned with those of explicit sentiment expressions with the same sentiment polarities through supervised contrastive learning. Yang and Li (2021) build the local sentiment aggregation to model sentiment dependency, which promotes the model's ability to learn implicit sentiment by capturing sentiment information from adjacent aspects. A differentially weighted strategy is also proposed for controlling adjacent aspects that contribute different sentiment information. While these approaches improve the learning and modeling of implicit sentiment in ABSA, external large-scale annotated corpora for encoding adjacent aspects are required. In view of the limitations of these approaches, we propose to leverage POS-based linguistic knowledge to assist PLMs in learning and modeling implicit sentiment in ABSA.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

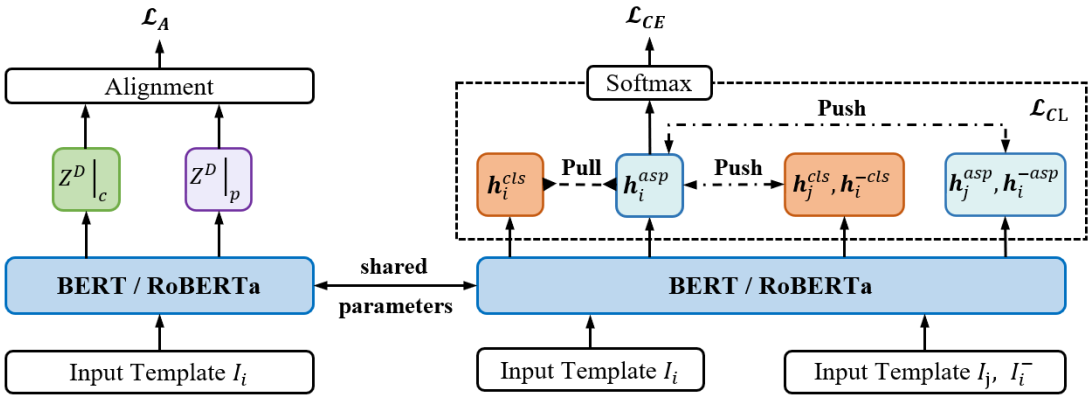


Figure 2: Overall architecture of our proposed framework. In a mini-batch, the input template I_i is derived from the i-th input sentence. I_j, I_i^- represent templates with the other sentence in the mini-batch and the disordered i-th input sentence. $Z^D|_c$ and $Z^D|p$ denote the representations of the input sentences subset and POS sequences subset. $\mathbf{h}_i^{cls}$, $\mathbf{h}_i^{asp}$, $\mathbf{h}_j^{cls}$, $\mathbf{h}_j^{asp}$, $\mathbf{h}_i^{-cls}$ and $\mathbf{h}_i^{-asp}$ are the representations from I_i, I_j and I_i^-, which are elaborated in Section 3.4.

3 Our method

3.1 Overall architecture

As mentioned above, in this paper, we propose a PLM-based linguistically enhanced framework for Aspect-based Sentiment Analysis. Our framework consists of three components: aspect-aware token-POS concatenation, token-POS alignment, and aspect-specific self-supervised contrastive learning. It is expected that POS-based linguistic knowledge will facilitate PLM's learning of implicit sentiment in ABSA. And self-supervised contrastive learning is applied to optimize the representation construction of the target aspect. Our method is shown in Figure 2.

Generally, an input sentence contains one or more aspect terms that correspond to multiple sentiments. In this paper, we focus on the sentiment analysis of a specific aspect. Given a sentence $\mathbf{x} = \{w_1, \ldots, w_t, a_1, \ldots, a_m, w_{t+1}, \ldots, w_n\}$ where w_i indicates the i^{th} word and $asp = \{a_1, \ldots, a_m\}$ denotes the target aspect in $\mathbf{x}$, an input template of our proposed framework I is composed of $\mathbf{x}$ and asp. We will elaborate on the detail of input template I in Section 3.2. The goal of ABSA is to predict the sentiment polarity towards asp according to the sentence $\mathbf{x}$.

3.2 Aspect-aware token-POS concatenation

Motivated by natural language prompt (Brown et al., 2020), we treat the POS sequence as a type of prompt with linguistic knowledge and change the input schema of the PLM. In a mini-batch, for each input sentence $\mathbf{x}_i$, we utilize spaCy[1] to perform part-of-speech tagging on it and combine POS tags into the POS sequence $pos_i = \{p_1, p_2, \ldots, p_n\}$ according to the order of $\mathbf{x}_i$. Instead of concatenating $\mathbf{x}_i$ and pos_i as the input template I_i directly, we additionally append the target aspect term asp_i to I_i following Song et al. (2019), which allows the PLM to capture dependencies between the context and the target aspect:

$$I_i = [CLS] + \mathbf{x}_i + [SEP] + pos_i + [SEP] + asp_i + [SEP] \quad (1)$$

The special tokens "[CLS]" and "[SEP]" of BERT should be "⟨s⟩" and "⟨/s⟩" in RoBERTa. After encoding I_i by BERT or RoBERTa, the pooled representation of asp_i is denoted as $\mathbf{h}_i^{asp} \in \mathbb{R}^{d \times l}$ $(l \geq m)$. Here d is the hidden size of the PLM and l is the length of the tokenized aspect by WordPiece (Wu et al., 2016) or Byte Pair Encoding (Sennrich et al., 2016).

[1] https://spacy.io/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786–800, Harbin, China, August 3 – 5, 2023.

3.3 Token-POS alignment

Unlike the discrete templates used in previous research, the POS sequence pos_i is not an independent natural language sentence but the ordered permutation of the POS tags corresponding to the given sentence $\mathbf{x}_i$. To reduce the effects of POS sequences on semantics and promote the interaction of POS sequences and input sentences, we design a token-POS alignment strategy referring to word patch alignment (Kim et al., 2021). As illustrated in Figure 2, in this method, the outputs of the PLM corresponding to the input sentences subset and POS sequences subset in each mini-batch are represented as $Z^D|_c$ and $Z^D|_p$ respectively. After the encoding, $Z^D|_c \in \mathbb{R}^{\mathcal{B}\times k\times d}$ and $Z^D|_p \in \mathbb{R}^{\mathcal{B}\times h\times d}$ can be treated as two different probability distributions, where $\mathcal{B}$ is the mini-batch size, h, k are the lengths of the tokenized input sentence and POS sequence, d is the hidden size. Thus, we convert the alignment into computing the statistical distance between $Z^D|_c$ and $Z^D|_p$, and the alignment score is optimized according to Optimal Transport theory (Peyré et al., 2019). Following such theory, we utilize Wasserstein distance (Vaserstein, 1969) to measure the statistical distance between $Z^D|_c$ and $Z^D|_p$:

$$W_p(Z^D|_c, Z^D|_p) := \mathbb{L}_{M^p}(Z^D|_c, Z^D|_p)^{\frac{1}{p}} \tag{2}$$

where p denotes the p-dimensional Wasserstein distance, $\mathbb{L}_{M^p}$ represents computing Wasserstein distance by Sinkhorn-Knopp algorithm (Knight, 2008) with the constraint of cost matrix $M \in \mathbb{R}^{d\times d}$, and the metric of Sinkhorn-Knopp algorithm is set to the cosine similarity considering the hidden size d of the PLM. Consequently, for $Z^D|_c$ and $Z^D|_p$ within a mini-batch B, the token-POS alignment loss can be defined as:

$$\mathcal{L}_A = \sum_{Z^D|_c, Z^D|_p \in B} W_p(Z^D|_c, Z^D|_p) \tag{3}$$

3.4 Aspect-specific self-supervised contrastive learning

Inspired by the applications of contrastive learning in ABSA (Li et al., 2021b; Liang et al., 2021), we propose to utilize self-supervised contrastive learning to enhance the representation construction of target aspects. According to the aim of contrastive learning (Hadsell et al., 2006), one of the keys is constructing the proper positive instances. Following the previous research in ABSA, both the embedding of the "[CLS]" token (Liang et al., 2021; Zhang et al., 2022) and the aspect features (Dai et al., 2021; Ma et al., 2022) can be used as the final representation for sentiment polarity classification. Hence, those two representations from the same instance can be treated as positives and others from different in-batch instances are taken as negatives. We denote $\mathbf{h}_i^{asp} = f_\theta^{asp}(\mathbf{x}_i)$ where $f_\theta(\cdot)$ represents the encoder. And the embedding of the "[CLS]" token from the same instance is represented as $\mathbf{h}_i^{cls} = f_\theta^{cls}(\mathbf{x}_i)$. Moreover, in order to further leverage the training data and improve the ability of the model to identify the aspect-related context, we construct hard negatives by disordering the input sentence as $\mathbf{x}_i^{dis} = \{w_{t+1}, \ldots, w_n, a_1, \ldots, a_m, w_1, \ldots, w_t\}$. Thus, the input template filled with the disordered input sentence I_i^- is defined as:

$$I_i^- = [CLS] + \mathbf{x}_i^{dis} + [SEP] + pos_i^{dis} + [SEP] + asp_i + [SEP] \tag{4}$$

where pos_i^{dis} is the POS sequence derived from $\mathbf{x}_i^{dis}$. The embedding of the "[CLS]" token and the pooled hidden vector of the aspect term from $\mathbf{x}_i^{dis}$ can be denoted as $\mathbf{h}_i^{-cls} = f_\theta^{cls}(\mathbf{x}_i^{dis})$ and $\mathbf{h}_i^{-asp} = f_\theta^{asp}(\mathbf{x}_i^{dis})$ respectively. Therefore, the aspect-specific self-supervised contrastive loss is defined as ($\mathcal{B}$ is the mini-batch size):

$$\mathcal{L}_{CL} = -\log \frac{e^{sim(\mathbf{h}_i^{asp}, \mathbf{h}_i^{cls})/\tau}}{\sum_{j=1}^{\mathcal{B}} (e^{sim(\mathbf{h}_i^{asp}, \mathbf{h}_j^{cls})/\tau} + e^{sim(\mathbf{h}_i^{asp}, \mathbf{h}_j^{-cls})/\tau} + e^{sim(\mathbf{h}_i^{asp}, \mathbf{h}_j^{-asp})/\tau})} \tag{5}$$

where τ is a temperature hyperparameter and $sim(\mathbf{h}_1, \mathbf{h}_2)$ is the function that computes the cosine similarity between $\mathbf{h}_1$ and $\mathbf{h}_2$.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

3.5 Joint training

Except for applying the two losses mentioned above to optimize the training of our proposed framework, we also use the cross-entropy loss $\mathcal{L}_{CE}$ as the fine-tuning object of the PLM for sentiment polarity prediction:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{\mathcal{B}}\sum_{j=1}^{N} y_i^j \log \hat{y}_i^j + \lambda||\theta||^2 \tag{6}$$

where N is the number of labels, $\mathcal{B}$ is the mini-batch size, λ and θ represent the L_2 regularization and the parameter of the model. As shown in previous studies (Ma et al., 2016), Dropout (Srivastava et al., 2014) may induce inconsistency between the training and inference stages of the model. We argue that such inconsistency will be severe when introducing POS sequences into the input sentences. In order to regularize Dropout, we use the bidirectional Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}$ based on R-Drop (Wu et al., 2021) in our models. The overall loss function $\mathcal{L}$ for joint training is:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1\mathcal{L}_A + \lambda_2\mathcal{L}_{CL} + \alpha\mathcal{L}_{KL} \tag{7}$$

where λ_1 and λ_2 are trainable parameters as the weights of token-POS alignment loss and aspect-specific self-supervised contrastive loss. The coefficient α is a hyperparameter.

4 Experiments

4.1 Datasets

We conduct the experiments using three publicly available benchmark datasets. They are Restaurant and Laptop from SemEval 2014 Task 4 (Pontiki et al., 2014) and Twitter (Dong et al., 2014). The statistics of the three datasets are shown in Table 2. Due to the lack of development sets, 10% of the items from the training sets are randomly selected and treated as development sets. Following previous research, we remove examples with conflicting sentiment polarities.

Dataset	Positive		Neutral		Negative		Total	
	Train	Test	Train	Test	Train	Test	Train	Test
Restaurant	2164	728	637	196	807	196	3608	1120
Laptop	994	341	464	169	870	128	2328	638
Twitter	1561	173	3127	346	1560	173	6248	692

Table 2: Statistics on three benchmark datasets of ABSA.

4.2 Implement details

We fine-tune the BERT-base-uncase (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019) models pre-trained by HuggingFace Transformers (Wolf et al., 2020) and implemented by PyTorch (Paszke et al., 2019). The learning rate is set as 2×10^{-5} and the batch size is 32. We adopt Dropout strategy and the drop probability is adjusted as 0.1. The model is trained with AdamW (Loshchilov and Hutter, 2017) optimizer and the L_2 regularization parameter λ is 10^{-5}. The temperature hyperparameter τ of aspect-specific self-supervised contrastive learning is 0.1. The coefficient α is set as 0.3. Following the work of Chen et al. (2020), we utilize the 2-dimensional Wasserstein distance for token-POS alignment. Since not all of the Universal POS tags exist in the vocabularies of BERT and RoBERTa, we map the tags to their complete names before encoding them to overcome the problem of out-of-vocabulary. We perform our proposed models three runs with different seeds and report their average performance.

4.3 Compared models

In order to demonstrate the effectiveness of our proposed method which can benefit various PLMs in ABSA, we compare the proposed models with several state-of-the-art baselines and models focusing on implicit sentiment in ABSA from the perspectives of BERT-based models and RoBERTa-based models:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

Category	Model	Laptop		Restaurant		Twitter	
		Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
BERT	BERT (Devlin et al., 2019)*	77.90	73.37	84.20	76.76	73.70	70.86
	BERT-SPC (Song et al., 2019)	78.99	75.03	84.46	76.98	74.13	72.73
	LCF-BERT (Zeng et al., 2019)*	80.09	76.42	85.65	78.68	74.32	73.32
	BERTAsp (Li et al., 2021b)	78.53	74.07	85.80	78.95	-	-
	BERT+AM (Ma et al., 2022)	76.33	71.93	84.71	78.07	-	-
	IPOS-BERT (Ours)	**80.56**	**76.99**	**85.83**	**79.41**	**76.11**	**74.52**
RoBERTa	RoBERTa (Liu et al., 2019)*	81.97	78.38	87.23	81.00	75.43	74.47
	ASGCN-RoBERTa (Dai et al., 2021)	83.33	80.32	86.87	80.59	76.10	75.07
	RGAT-RoBERTa (Dai et al., 2021)	83.33	79.95	87.52	81.29	75.81	74.91
	LSA_P-RoBERTa (Yang and Li, 2021)	83.39	80.47	88.04	82.96	-	-
	RoBERTa+AM (Ma et al., 2022)	82.07	78.50	86.41	79.58	-	-
	IPOS-RoBERTa (Ours)	**83.54**	**80.91**	**88.93**	**83.30**	**77.46**	**76.63**
SCAPT	BERTAsp+SCAPT (Li et al., 2021b)†	82.76	79.15	<u>89.11</u>	<u>83.79</u>	-	-

Table 3: Overall results (%) in three benchmark datasets where the "IPOS-BERT" and "IPOS-RoBERTa" are the proposed models that indicate combining BERT and RoBERTa with our method. The experimental results of the models we reproduced are marked by "∗". For a fair comparison, we mark BERTAsp+SCAPT by "†" and additionally list it in the category "SCAPT" because of its in-domain pre-training and underline its state-of-the-art performance. The best results within other models are highlighted in bold according to different categories.

- **BERT**, **RoBERTa** denote the vanilla BERT and RoBERTa proposed by Devlin et al. (2019) and Liu et al. (2019) respectively. We fine-tune them by ABSA datasets and keep their default settings.
- **BERT-SPC** (Song et al., 2019) transforms the input reviews into sentence-aspect pairs and takes the "[CLS]" token for sentiment polarity classification.
- **LCF-BERT** (Zeng et al., 2019) utilizes the local context focus mechanism to model the relation between global context and local context.
- **BERTAsp** and **BERTAsp+SCAPT** (Li et al., 2021b) are fine-tuned BERT for ABSA. The latter is pre-trained on large-scale annotated corpora by supervised contrastive learning before fine-tuning.
- **ASGCN-RoBERTa**, **RGAT-RoBERTa** are implemented by Dai et al. (2021). They are based on ASGCN (Zhang et al., 2019a) and RGAT (Wang et al., 2020) respectively and RoBERTa is applied including its induced tree and embeddings.
- **LSA_P-RoBERTa** (Yang and Li, 2021) aggregates local sentiments by BERT-SPC (Song et al., 2019) and models implicit sentiment by exploiting adjacent aspects' sentiment information.
- **BERT+AM** and **RoBERTa+AM** (Ma et al., 2022) uses the tokens "⟨asp⟩" and "⟨/asp⟩" to mark boundaries of aspects, which promotes PLMs to construct aspect-specific contextualized features.

4.4 Overall results and analysis

The experimental results of the aforementioned compared models and ours are shown in Table 3. Specifically, the accuracy and Macro-F1 score are utilized to evaluate the performance of models. According to the results, we have the following observations:

1) Incorporating linguistic knowledge improves the performance of ABSA models. Compared to the vanilla BERT and RoBERTa, on the one hand, incorporating syntactic knowledge by graph neural networks such as GCN and GAT promotes PLMs to capture the related information about the aspects, which is directly represented as the improvement of ASGCN-RoBERTa and RGAT-RoBERTa. On the other hand, leveraging Part-of-Speech to assist PLMs in modeling implicit sentiment benefits the ABSA task. By incorporating Part-of-Speech and aspect-specific self-supervised contrastive learning, both BERT and RoBERTa improve significantly on three ABSA benchmarks, achieving approximate 2.7%/1.6%/2.4% and 1.6%/1.7%/2.0% performance gains in accuracy as well as 3.6%/2.6%/3.7%

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

Models	Laptop-test		Restaurant-test	
	Accuracy	Accuracy-ISE	Accuracy	Accuracy-ISE
BERT-SPC (Song et al., 2019)	78.99	69.54	84.46	65.54
IPOS-BERT (Ours)	**80.56**	**76.00**	**85.83**	**66.66**
RoBERTa (Liu et al., 2019)	81.97	78.86	87.23	68.54
IPOS-RoBERTa (Ours)	**83.54**	**84.57**	**88.93**	**71.16**

Table 4: Model performance (%) on the Laptop and Restaurant benchmarks and their Implicit Sentiment Expression slices (ISE). The "Accuracy-ISE" column denotes the performance of models on ISE, which is measured by accuracy.

and 2.5%/2.3%/2.2% in Macro-F1 score on Laptop/Restaurant/Twitter benchmarks respectively.

2) Without introducing numerous additional parameters and extra corpora, our model can perform similarly to state-of-the-art models and even outperform them. The proposed model IPOS-RoBERTa has a similar number of parameters (125.2M) to the vanilla RoBERTa-base model (125M). The difference between them lies in the layer for sentiment polarity classification. However, IPOS-RoBERTa can achieve state-of-the-art performance on Laptop and Twitter benchmarks, demonstrating the effectiveness of our method. Unlike $\mathrm{LSA_P}$-RoBERTa and BERTAsp+SCAPT, our method optimizes the fine-tuning of RoBERT to learn implicit sentiment rather than introducing additional parameters for encoding adjacent aspects and extra corpora for pre-training. Specifically, the parameters of the compared models mentioned above are 138.2M and 133.3M respectively[2], indicating millions of parameters are added compared to our proposed model. However, on the test set of the Laptop benchmark, the Macro-F1 score of **IPOS-RoBERTa** is 80.91%, which is 1.76% higher than **BERTAsp+SCAPT** (Macro-F1=79.15%) and 0.44% higher than $\mathbf{LSA_P}$**-RoBERTa** (Macro-F1=80.47%). Though the results of RoBERT-based models on Twitter are not shown in (Yang and Li, 2021), the accuracy and Macro-F1 score of **IPOS-RoBERTa** are 0.55% and 0.73% higher than $\mathbf{LSA_P}$**-DeBERTa** (Accuracy=76.91%, Macro-F1=75.90%), which is based on a progressive PLM called DeBERTa (He et al., 2021). Though the difficulty of improving RoBERTa-based models in ABSA is indicated by Dai et al. (2021), these results prove that POS-based linguistic knowledge and aspect-specific self-supervised contrastive learning are actually beneficial for enhancing the performance of fine-tuned RoBERTa in this task.

4.5 Effectiveness on implicit sentiment learning

Besides conducting extensive experiments on three benchmark datasets mentioned above, we also report the results of the experiment on Implicit Sentiment Expression (ISE) slices of Laptop and Restaurant that are derived from the work of Li et al. (2021b). As shown in Table 4, on both two ISE slices, our proposed models IPOS-BERT and IPOS-RoBERTa outperform compared models based on the same PLMs with them. Though predicting sentiment polarities conveyed by implicit sentiment expressions is challenging, IPOS-RoBERTa's accuracy on ISE slices is higher than that of vanilla RoBERTa by large margins, which indicates the obvious improvement of 5.71% and 2.62%. And the other improvement (6.46% and 1.12%) of Accuracy-ISE can be observed by the comparison of IPOS-BERT and BERT-SPC on the ISE slices of Laptop and Restaurant respectively. Such progresses demonstrates the effectiveness of incorporating POS-based linguistic knowledge for learning implicit sentiment in ABSA.

5 Discussion

5.1 Ablation study

Considering that each component of the proposed framework plays a different role as well as the temperatures contribute variously, extensive ablation experiments are conducted on the Laptop benchmark and results are shown in Table 5. We find that removing the token-POS alignment degrades the proposed

[2]The statistics of parameters are derived from open-source repositories released by Yang and Li (2021) and Li et al. (2021b).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

Model Variant	Laptop	
	Accuracy	Macro-F1
IPOS-RoBERTa	83.54	80.91
w/o $\mathcal{L}_A$	81.97	78.80
w/o $\mathcal{L}_{CL}$	82.29	78.85
w/o $\mathcal{L}_{KL}$	82.60	79.25

(a) The ablation study of different components

Model Variant	Laptop	
	Accuracy	Macro-F1
IPOS-RoBERTa	83.54	80.91
$\tau = 0.01$	81.03	77.77
$\tau = 0.05$	83.23	80.65
$\tau = 0.5$	83.07	79.37

(b) The ablation study of different temperatures

Table 5: Ablation studies of different components and temperatures on the Laptop benchmark (%). "w/o $\mathcal{L}_A$, $\mathcal{L}_{CL}$, $\mathcal{L}_{KL}$" indicates the models without token-POS alignment, aspect-specific self-supervised contrastive learning and R-Drop respectively. In the ablation study of temperatures (τ), we compare the original setting ($\tau = 0.1$) with three variants.

Example and POS Sequence	RoBERTa	BERTAsp*	Ours
However, I can refute that OSX is FAST. ADV PRON AUX VERB SCONJ PROPN AUX ADJ	Pos (×)	Pos (×)	Neg (✓)
Fan only comes on when you are playing a game. NOUN ADV VERB ADP SCONJ PRON AUX VERB DET NOUN	Neg, Neu (×), (✓)	Neu, Neu (✓), (✓)	Neu, Neu (✓), (✓)
It has so much more speed and the screen is very sharp. PRON VERB ADV ADV ADJ NOUN CCONJ DET NOUN AUX ADV ADJ	Pos, Pos (✓), (✓)	Pos, Neg (✓), (×)	Pos, Pos (✓), (✓)
I did swap out the hard drive for a Samsung 830 SSD which I highly recommend. PRON AUX VERB ADP DET NOUN ADP DET PROPN PRON PRON ADV VERB	Neu, Neu (✓), (×)	Neu, Neu (✓), (×)	Neu, Pos (✓), (✓)

Table 6: A case study in the domain of laptops. For each case example, the original review and its POS sequence are shown. The model marked by "∗" denotes BERTAsp+SCAPT proposed by Li et al. (2021b) and the aspect terms are underlined. We use "Pos, Neu, Neg" to indicate three sentiment polarities ("Positive, Neutral, Negative"). The correct predictions are associated with the symbol "✓" and the wrong predictions are marked with "×".

model drastically and even leads to the suboptimal performance of the proposed model, which is similar to that of the vanilla RoBERTa. We suppose that the POS sequences imported from the external parser affect contextual semantics without the token-POS alignment (Similar visual examples are shown in the rows of "RoBERTa (with POS)" in Figure 3). Thus, though keeping the aspect-specific self-supervised contrastive learning and R-Drop, their effects are obscure while importing POS sequences directly. Such degradation indicates the importance of incorporating Part-of-Speech knowledge properly. Another noticeable performance degradation is caused by the absence of aspect-specific self-supervised contrastive learning since it promotes the model to concentrate on the target aspects. Similarly, our model benefits from R-Drop (Wu et al., 2021) due to the regularization of the predictions.

Moreover, in order to investigate the influence of different temperatures, we set the temperature $\tau \in \{0.01, 0.05, 0.1, 0.5\}$ and keep other settings of our model. Compared to the carefully tuned temperature ($\tau = 0.1$), the other lead to different degrees of impact. It is worth noting that an extremely small temperature ($\tau = 0.01$) causes an obvious drop in the performance, which makes the model focus much on hard negatives (Wang and Liu, 2021). However, a high temperature is also inappropriate. Specifically, both the accuracy and the Marco F1 score of the proposed model trained with a high temperature ($\tau = 0.5$) are lower than those of the model with a carefully tuned temperature by large margins.

5.2 Case study

To verify the effectiveness of our method, we select several cases in the laptop domain that contain implicit sentiment expressions, as shown in Table 6. According to these cases, the capabilities of modeling implicit sentiment and capturing syntactic features are demanded. Hence, BERTAsp+SCAPT (Li et al.,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

Model	Case Visualization											Asp
IPoS-RoBERTa	<s>	However	I	can	refute	that	OSX	is	FAST	</s>	ADV	OSX
	PRON	AUX	VERB	SCONJ	PROPN	AUX	ADJ	</s>	OSX	</s>		
RoBERTa (with PoS)	<s>	However	I	can	refute	that	OSX	is	FAST	</s>	ADV	
	PRON	AUX	VERB	SCONJ	PROPN	AUX	ADJ	</s>	OSX	</s>		
RoBERTa	<s>	However	I	can	refute	that	OSX	is	FAST	</s>	OSX	</s>
IPoS-RoBERTa	<s>	I	will	not	be	using	that	slot	again	</s>	PRON	AUX
	PART	AUX	VERB	DET	NOUN	ADV	</s>	slot	</s>			slot
RoBERTa (with PoS)	<s>	I	will	not	be	using	that	slot	again	</s>	PRON	AUX
	PART	AUX	VERB	DET	NOUN	ADV	</s>	slot	</s>			
RoBERTa	<s>	I	will	not	be	using	that	slot	again	</s>	slot	</s>

Figure 3: Visualization of two selected cases. Both two target aspects are expressed by implicit sentiment. The gradient saliency maps (Simonyan et al., 2014) for the embedding of input tokens are shown, including the words and corresponding POS tags. For each token, the darker color denotes the higher gradient saliency score. The "**Asp**" column indicates the aspect terms.

2021b) and RoBERTa (Liu et al., 2019) are chosen as strong compared models for the case study. Following the comparison results, both BERTAsp+SCAPT and RoBERTa fail to correctly predict all the case examples. For example, RoBERTa wrongly comprehends the semantics of the second review and predicts the sentiment polarity towards "**fan**" as negative, which is represented by implicit sentiment expression. Additionally, for the aspect "**screen**" in the third case, BERTAsp+SCAPT mistakes the opinion "sharp" and incorrectly infers the corresponding polarity as negative. However, when given some complicated cases carrying multiple aspects and intricate implicit sentiment, both of them improperly capture the aspect-related contextualized features such as the aspects "**OSX**" and "**Samsung 830 SSD**" in the first and the last cases.

Owing to the POS-based linguistic knowledge, the proposed IPOS-RoBERTa model can precisely predict all aforementioned cases. We suppose that POS sequences encourage the model to learn implicit sentiment and distinguish sentiment expressions about different aspects, as suggested by the good performance of our proposed model. For the first case, the adjective "FAST" is related to the aspect "**OSX**" from the view of syntax but it implies the contrary sentiment polarity due to the verb "refute", which helps to perceive the implicit sentiment. Moreover, when inferring multiple aspects "**hard drive**" and "**Samsung 830 SSD**" in the same sentence, IPOS-RoBERTa can distinguish the related information about them and predict the correct sentiment polarity towards "**Samsung 830 SSD**".

5.3 Visualization

Since it seems that the effect of appending POS tags to the input tokens is intricate, we visualize the gradient saliency scores of the embedding of input templates for two selected cases, which can be employed for model interpretation (Li et al., 2016). As shown in Figure 3, we compare our model with two backbones and keep the setting that appends the aspect terms to the input sequences for all of them. However, "RoBERTa (with POS)" denotes only employing aspect-aware token-POS concatenation to RoBERTa but ignoring the token-POS alignment and "RoBERTa" indicates the vanilla RoBERTa model.

In the first case, the words "refute" and "FAST" are assigned different saliency scores among the three models, signifying these words differently contributing to the predictions. Compared to another two models, we suppose that our model pays more attention to such important words in comprehending the semantics. Furthermore, due to the token-POS alignment, our model can distinguish the importance of

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786–800, Harbin, China, August 3 – 5, 2023.

different POS tags instead of treating them equally. Similarly, though the three selected models focus on the word "not", the neglect of the verb "using" leads to incorrect predictions of sentiment polarity towards the aspect "slot". In contrast, our model can precisely capture essential words and their POS for prediction, demonstrating the effect of aspect-aware token-POS concatenation and token-POS alignment.

6 Conclusion

In this paper, we propose a PLM-based linguistically enhanced framework for aspect-based sentiment analysis based on the analysis of the feasibility of incorporating Part-of-Speech into the ABSA task. Using POS-based linguistic knowledge, our method optimizes the PLMs' fine-tuning for implicit sentiment capturing. Aspect-specific self-supervised contrastive learning allows the model to concentrate on target aspects when handling sentences containing multiple aspect terms. Extensive experiments show that our proposed model can achieve competitive and state-of-the-art performance relative to baseline models without introducing extra corpora. Although the introduction of POS as linguistic knowledge can effectively improve the enhancement of implicit sentiment detection in ABSA, there are still limitations. If there are difficulties in deriving precise POS sequences in low-resource settings, the POS-based solution might not provide sufficient information. Further research can investigate approaches for integrating various linguistic knowledge into models for learning implicit sentiment without external sources.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61976062).

References

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

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

Jiahao Cao, Rui Liu, Huailiang Peng, Lei Jiang, and Xu Bai. 2022. Aspect is not you need: No-aspect differential sentiment framework for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1599–1609, Seattle, United States, July. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? a strong baseline for aspect-based sentiment analysis with RoBERTa. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1816–1829, Online, June. Association for Computational Linguistics.

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

Lipika Dey and S K Mirajul Haque. 2008. Opinion mining from noisy text data. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, AND '08, page 83–90, New York, NY, USA. Association for Computing Machinery.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland, June. Association for Computational Linguistics.

Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, JIan Guo, and Nan Duan. 2022. Sentiment-aware word and sentence level pre-training for sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4994, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7035–7045.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings 11*, pages 197–206. Springer.

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

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988.

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

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 437–442.

Philip A Knight. 2008. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California, June. Association for Computational Linguistics.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, et al. 2021a. Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis. *arXiv preprint arXiv:2109.08306*.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021b. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Bin Liang, Wangda Luo, Xiang Li, Lin Gui, Min Yang, Xiaoqi Yu, and Ruifeng Xu. 2021. Enhancing aspect-based sentiment analysis with supervised contrastive learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3242–3247.

Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

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

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

Xuezhe Ma, Yingkai Gao, Zhiting Hu, Yaoliang Yu, Yuntian Deng, and Eduard Hovy. 2016. Dropout with expectation-linear regularization. *arXiv preprint arXiv:1609.08017*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Fang Ma, Chen Zhang, Bo Zhang, and Dawei Song. 2022. Aspect-specific context modeling for aspect-based sentiment analysis. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pages 513–526.

Chris Nicholls and Fei Song. 2009. Improving sentiment analysis with part-of-speech weighting. In *2009 International Conference on Machine Learning and Cybernetics*, volume 3, pages 1592–1597. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Minh Hieu Phan and Philip O. Ogunbona. 2020. Modelling context and syntactical features for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3220, Online, July. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 443–450.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5679–5688.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

Leonid Nisonovich Vaserstein. 1969. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

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

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Haiyan Wu, Zhiqiang Zhang, Shaoyun Shi, Qingfeng Wu, and Haiyu Song. 2022. Phrase dependency relational graph attention network for aspect-based sentiment analysis. *Knowledge-Based Systems*, 236:107736.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1.

Lvxiaowei Xu, Xiaoxuan Pang, Jianwang Wu, Ming Cai, and Jiawei Peng. 2023. Learn from structural scope: Improving aspect-level sentiment analysis with hybrid graph convolutional networks. *Neurocomputing*, 518:373–383.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021a. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429.

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

Heng Yang and Ke Li. 2021. Improving implicit sentiment learning via local sentiment aggregation. *arXiv e-prints*, pages arXiv–2110.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019a. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4568–4578.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019b. Syntax-aware aspect-level sentiment classification with proximity-weighted convolution network. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1145–1148.

Kai Zhang, Kun Zhang, Mengdi Zhang, Hongke Zhao, Qi Liu, Wei Wu, and Enhong Chen. 2022. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3599–3610, Dublin, Ireland, May. Association for Computational Linguistics.

Pinlong Zhao, Linlin Hou, and Ou Wu. 2020. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowledge-Based Systems*, 193:105443.

Yongqiang Zheng, Xia Li, and Jian-Yun Nie. 2023. Store, share and transfer: Learning and updating sentiment knowledge for aspect-based sentiment analysis. *Information Sciences*, 635:151–168.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 786-800, Harbin, China, August 3 - 5, 2023.

Case Retrieval for Legal Judgment Prediction in Legal Artificial Intelligence

Han Zhang
School of Information,
Renmin University of China.
zhanghanjl@ruc.edu.cn

Zhicheng Dou
Gaoling School of Artificial Intelligence,
Renmin University of China.
dou@ruc.edu.cn

Abstract

Legal judgment prediction (LJP) is a basic task in legal artificial intelligence. It consists of three subtasks, which are relevant law article prediction, charge prediction and term of penalty prediction, and gives the judgment results to assist the work of judges. In recent years, many deep learning methods have emerged to improve the performance of the legal judgment prediction task. The previous methods mainly improve the performance by integrating law articles and the fact description of a legal case. However, they rarely consider that the judges usually look up historical cases before making a judgment in the actual scenario. To simulate this scenario, we propose a historical case retrieval framework for the legal judgment prediction task. Specifically, we select some historical cases which include all categories from the training dataset. Then, we retrieve the most similar Top-k historical cases of the current legal case and use the vector representation of these Top-k historical cases to help predict the judgment results. On two real-world legal datasets, our model achieves better results than several state-of-the-art baseline models.

1 Introduction

With the rapid development of artificial intelligence, it has become a trend to use artificial intelligence to help judicial personnel. Legal judgment prediction (LJP) is such an artificial intelligence task in legal artificial intelligence. As shown in Table 1, given the fact description of a legal case, the legal judgment prediction task can provide the judgment result of the case. The predicted result consists of three parts: relevant law article, charge and term of penalty. Legal judgment prediction can not only give the judgment results efficiently for reference for judicial personnel but also provide legal suggestions for ordinary people when there is a legal dispute (Zhong et al., 2020; Wang et al., 2019; Zhong et al., 2018; Zhang et al., 2021) in daily life.

With the application of deep learning in legal artificial intelligence, various methods have been proposed to improve the performance of legal judgment prediction. Some methods (Zhong et al., 2018; Yang et al., 2019) consider using the order information among the three subtasks of legal judgment prediction in reality to improve the representation of fact description. Further, some methods (Yue et al., 2021; Ma et al., 2021; Feng et al., 2022) consider a fine-grained division of the fact description to improve the fact representation. Additionally, some methods (Luo et al., 2017; Hu et al., 2018; Wang et al., 2019; Xu et al., 2020) consider the important role of law articles in reality and introduce them to improve the performance. These efforts have effectively improved the performance of legal judgment prediction. However, the existing methods are affected by the fact that the law articles are too concise and still have limitations in modelling the judgment process.

On the one hand, the law articles are very concise and lack specific details and some law articles have similar provisions and so easy to be confused. As shown in Figure 1, *Article #114* and *Article #115* both stipulate the same charge *Crime of Arson* and the provisions in the two law articles are very short. In order to distinguish them, the judge usually re-finds and analyzes historical cases because the fact description information of historical cases usually contains more detailed information than the law articles.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Fact Description On XX, XXX, the procuratorate accused the defendant Yang XX of taking gasoline out of his motorcycle fuel tank after quarrelling with his girlfriend Tang XX, putting it into a beer bottle, and **pouring gasoline through the crack of the door** into room X of the rental room opposite the XX Internet cafe in the XX community where Tang XX is located, and **using a lighter to ignite the gasoline**. **The fire spread to the room along with the gasoline and was extinguished** by Tang XX and other people in the room. On the morning of that day, the public security police arrested the defendant Yang XX ...
Relevant Law Article *Article #114* **[Crime of Arson] Whoever commits arson**, breaches a dike, causes an explosion, spreads toxic, radioactive, infectious disease pathogens and other substances or endangers public security by other dangerous methods, but has not caused serious consequences, shall be sentenced to **fixed-term imprisonment of not less than three years but not more than ten years.**
Charge: **Crime of Arson**
Term of Penalty: **A fixed-term imprisonment of thirty-six months**

Table 1: An example of the legal judgment prediction task.

On the other hand, most of the previous methods predict the judgment results mainly based on the fact description of a single case, however, they overlook the practical scenario that judges usually look up typical historical cases for reference before making a judgment. As we all know, historical cases are very important for making a judgment, whether in the Case Law system or the Statutory Law system.[0] In the Case Law system, judges mainly refer to historical cases to make a judgment. In the Statutory Law system, before making a judgment the judges should not only look up the law articles but also look up typical historical cases. Obviously, historical cases are indispensable references for judges in their work.

To solve these challenges, we propose a framework for legal judgment prediction based on a historical case retrieval module to simulate the actual legal scenario of looking up historical cases before making a judgment.

First, we consider that the number of cases looked up by judges in actual work is usually limited and select a part of cases from the training dataset as historical cases.

Second, in order to avoid the impact of highly unbalanced class distribution of the dataset on the model performance (Hu et al., 2018; Zhang et al., 2021), we consider selecting the same number of historical cases for each category.

Third, we retrieve the most similar Top-k historical cases of the current legal case and concatenate the vector representation of these Top-k cases and the fact description of the current case to predict the judgment results. Finally, we train our model with a cross-entropy loss function. We call our model **CR4LJP**, which stands for **C**ase **R**etrieval framework for **L**egal **J**udgment **P**rediction.

Our contributions are three-fold:

(1) We take into account that judges usually look up historical cases before making a judgment after investigating the human justice system.

(2) We propose a case retrieval framework for the legal judgment prediction task to use historical cases to help predict the judgment results.

(3) Experiment results of our framework with different encoders on two real large-scale legal datasets are better than the state-of-the-art models and verify the effectiveness of our framework. This study shows that case retrieval is an effective way to improve the performance of the legal judgment prediction task.

2 Related Work

2.1 Legal Judgment Prediction

The earliest legal judgment prediction (LJP) methods (Kort, 1957; Ulmer, 1963; Nagel, 1963; Segal, 1984; Gardner, 1984) mainly use mathematical and statistical tools. These methods are based on artificial features or rules, so they are difficult to extend. In recent years, some researchers have proposed a lot of models (Zhong et al., 2020; Zhong et al., 2018; Yang et al., 2019; Dong and Niu, 2021) based on

[0]The details of the Case Law and Statutory Law system can be found in https://en.wikipedia.org/wiki/Case_law and https://en.wikipedia.org/wiki/Statutory_law.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

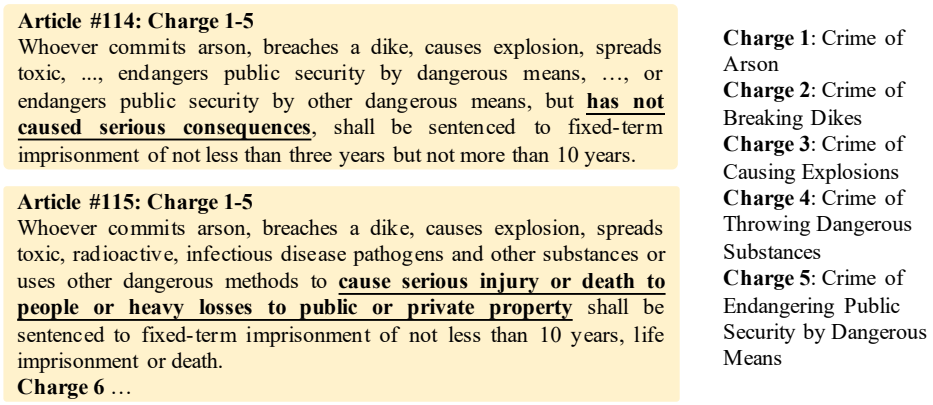


Figure 1: Article #114 and Article #115 both stipulate the same charges (Charge 1-5). There is little difference between the specific provisions of Article #114 and Article #115 on the Crime of Arson.

deep learning to predict judgment results. Specifically, some research works (Zhong et al., 2018; Yang et al., 2019) consider that the legal judgment prediction task is composed of three subtasks, and there are dependencies among them which are useful information. Some research works (Yue et al., 2021; Ma et al., 2021; Feng et al., 2022) consider that the fact description is usually long, and the fact description can be better represented by dividing or extracting the fine-grained information. Some research works (Luo et al., 2017; Hu et al., 2018; Wang et al., 2019; Xu et al., 2020) consider the important role of law articles in reality and then study how to make use of the information of law articles. These works improve the performance of LJP, but they fail to take into account that historical cases are also important information.

2.2 Retrieval Methods

For deep learning models, even the pre-trained models, such as Bert (Devlin et al., 2019), can not remember all samples. Therefore, it is worth considering using a retrieval model to obtain additional information. Generally, retrieval models can be divided into two types: sparse representation based on bag-of-word (BOW) (Chen et al., 2017) and dense vector representation based on neural networks (Karpukhin et al., 2020; Zhou et al., 2020). The retrieval models based on sparse representation have been applied in machine translation (Gu et al., 2018) and open domain question answering (Chen et al., 2017; Wang et al., 2018; Lin et al., 2018). The retrieval models based on dense vector representation (Karpukhin et al., 2020; Zhou et al., 2020) have received more attention in recent years. This method can achieve better recall performance than the sparse retrieval model on various Natural Language Processing (NLP) tasks, such as personalized search (Ma et al., 2020; Zhou et al., 2020) and domain question answering (Karpukhin et al., 2020; Guu et al., 2020; Yu et al., 2022). Considering that judges usually only need some typical cases and the good performance of dense vector representation, we use the dense vector representation retrieval method.

3 Problem Definition

Before introducing our model, we first introduce some concepts and definitions of legal judgment prediction.

A **legal case** in our paper consists of a fact description and three judgment results, which are made by human judges. The **fact description** is a text that describes the criminal facts of a suspect. As shown in Figure 2, our model uses f to represent it. The three **judgment results** are relevant law article, charge and term of penalty and we use y_1, y_2 and y_3 to represent them respectively. Then a legal case can be represented as:

$$\text{Case} = (f, y_1, y_2, y_3), \tag{1}$$

where f, y_1, y_2, y_3 are defined above.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

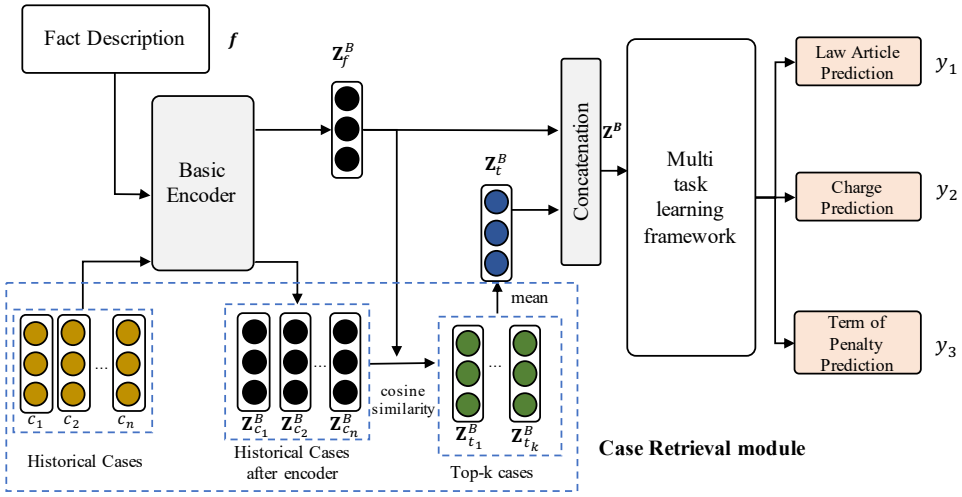


Figure 2: The framework of our model. The main module of our framework is the Basic Encoder and the Case Retrieval module.

Referring to previous studies (Zhong et al., 2018; Xu et al., 2020; Luo et al., 2017), we adopt a multi-task learning framework to solve the legal judgment prediction task. Our goal is to train a model $F(\cdot)$ which can be used to predict a case f_t in the test dataset with a given training dataset D, namely:

$$F(f_t) = (\hat{y}_1, \hat{y}_2, \hat{y}_3), \tag{2}$$

where $\hat{y}_1$, $\hat{y}_2$ and $\hat{y}_3$ are the predicted judgment results. Consistent with the existing works (Zhong et al., 2018; Xu et al., 2020), we only consider the legal cases with one relevant law article and one charge label.

4 Model Framework

In the actual judgment process, judges usually look up some typical historical cases for reference. To simulate this process, we propose a framework (CR4LJP) with a historical case retrieval module.

4.1 Overview

Our model framework is shown in Figure 2. In general, our model is a multi-task learning framework, which jointly solves three legal judgment prediction subtasks, with the case retrieval module we proposed. The main modules and training process of our model framework are as follows:

(1) The fact description f is converted into vector representation Z_f^B through the basic encoder.

(2) All selected historical cases are transformed into vector representations by the basic encoder. We select the Top-k cases which are most similar to the vector Z_f^B from these cases according to the cosine similarity. Then, we get the mean vector Z_t^B of these Top-k cases as auxiliary information.

(3) The representation vector Z_f^B and the mean vector Z_t^B of these Top-k cases are concatenated to solve the three legal judgment prediction subtasks.

(4) Our model is optimized by the losses of three subtasks. In the test phase, we also use the historical case vectors as auxiliary information to predict the judgment results.

4.2 Basic Encoder

As shown in Figure 2, our model framework uses the same encoder for the current case and historical cases. Considering the consistency with the previous models (Xu et al., 2020; Zhong et al., 2018; Yue et al., 2021) and the operation efficiency, we adopt the recurrent neural network (RNN) based encoder. Although we use RNN based encoder, our framework can flexibly select the neural network. Other neural networks, such as the current neural network (CNN) and pre-trained language models (PLMs), can also be used as encoders.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

Specifically, the fact description of a legal case with m words is represented as:

$$f = (w_1, \cdots, w_m), \tag{3}$$

where w_i is a word in the fact description. Then we convert it to a word embedding sequence $\mathbf{f}$ though looking up a pre-trained word embedding table $\mathbf{E}$:

$$\mathbf{f} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_m], \mathbf{e}_i \in \mathbf{E}, \tag{4}$$

where $\mathbf{f} \in \mathbf{R}^{m \times d_e}$, and $\mathbf{e}_i \in \mathbf{R}^{d_e}$ is the embedding vector of the i-th word w_i. Then we use Bi-GRU neural network to encode the fact description.

$$\mathbf{Z_f^B} = \text{Bi-GRU}(\mathbf{f}), \tag{5}$$

where $\mathbf{Z_f^B} = (h_1, \cdots, h_l) \in \mathbf{R}^{l \times d_h}$, d_h is the length of the hidden layer of Bi-GRU encoder.

After introducing RNN based encoder, we introduce an alternative neural network Bert (Devlin et al., 2019) as the encoder. First, the fact description $\mathbf{f}$ is set as the input of Bert after an embedding layer. After the multi-layer self-attention encoder, the output of "[CLS]" token of Bert is set as the vector representation of the fact description. It can also be represented as:

$$\mathbf{Z}_f^{\text{Bert}} = \text{BERT}(f)_{[\text{CLS}]}, \tag{6}$$

where "[CLS]" is one of the tokens ouput by the Bert model.

4.3 Case Retrieval Module

In the actual judgment process, judges usually look up some typical historical cases as references. So we design a case retrieval module to simulate the scenario. As the performance of the dense representation retrieval method is usually better, we choose the dense representation retrieval method for our retrieval module.

Case Selection. In reality, the number of historical cases is huge and judges usually only look up some cases as references. For the efficiency of the model, we consider only selecting part of the cases instead of all the cases in the training dataset as historical cases to be retrieved. It should be noted that some law articles stipulate the same charges as shown in Figure 1. And considering the unbalanced distribution of categories, we select the same number of cases for each charge under each law article.

Case Retrieval. In order to realize the historical case retrieval module, we first represent all historical cases $(c_1, c_2, \cdots, c_n)$ as word embedding sequences through Formula 4, and then represent them as n encoded vectors $(\mathbf{Z}_{c_1}^B, \cdots, \mathbf{Z}_{c_n}^B)$ through Formula 5. Then we calculate the similarity scores of these n historical cases and the fact vector representation $\mathbf{Z}_f^B$ of the current case according to cosine similarity. Finally, we select the **Top-k** most similar cases as the reference cases by ranking the similarity scores, and then we calculate the mean vector of these Top-k cases:

$$\mathbf{Z}_t^B = \text{Mean}(\mathbf{Z}_{t_1}^B, \cdots, \mathbf{Z}_{t_k}^B). \tag{7}$$

The final mean vector $\mathbf{Z}_t^B$ of these **k** historical cases is the output of the case retrieval module.

4.4 Prediction and Optimization

Before predicting the judgment results for calculating the losses of three legal judgment prediction subtasks, we concatenate the vector representation of the current case and the historical cases as follows:

$$\mathbf{Z}^B = [\mathbf{Z}_f^B; \mathbf{Z}_t^B], \tag{8}$$

and then we use a multi-layer perceptron layer to predict the results as follows:

$$y_i = \text{MLP}_i(\mathbf{Z}^B), \tag{9}$$

Table 2: The statistics of the CAIL dataset.

Dataset	CAIL-small	CAIL-big
# Training Set Cases	106,750	1,648,600
# Test Set Cases	25,652	200,449
# Law Articles	94	115
# Charges	109	129
# Term of Penalty	11	11

where i represent the i-th subtask of legal judgment prediction.

Total loss. The legal judgment prediction task includes three subtasks (relevant law article prediction, charge prediction and term of penalty prediction). We use the cross-entropy loss to calculate the loss of each subtask and train our model. The total loss is calculated as follows:

$$\mathcal{L}_{\mathrm{LJP}} = -\sum_{i=1}^{3} \alpha_i \sum_{j=1}^{|N_j|} y_{i,j} \log(\hat{y}_{\mathrm{i,j}}), \tag{10}$$

where $|N_{ij}|$ represent the number of labels of subtask i, and α_i is the weight of subtask i which is hyper parameter.

5 Experiments

5.1 Datasets and Preprocessing

Most of the state-of-the-art methods for legal judgment prediction are tested on the Chinese AI and Law challenge (CAIL2018) dataset (Xiao et al., 2018). The CAIL2018 dataset consists of a large of legal cases published by the Supreme People's Court of China and it has two sub-datasets, namely, CAIL-small and CAIL-big. Every case has a fact description and the judgment results given by human judges. The statistics of the dataset are shown in Table 2.

In addition, to be consistent with the baseline methods (Zhong et al., 2019; Xu et al., 2020; Yue et al., 2021), we first filter out the legal cases with multiple article/charge labels in the CAIL dataset, and then filter out the low-frequency law articles and charges which have less than 100 cases. Finally, we filter out the legal cases with missing or error labels (*e.g.* a small number of cases have no law article or charge labels, or the charge label is inconsistent with the law article label).

5.2 Baselines

In order to verify the effectiveness of our model, we select several representative legal judgment prediction models as the baselines.

(1) **FLA** (Luo et al., 2017) first considers the important role of law articles in the actual legal judgment process and uses the attention module to introduce the law article information.

(2) **Attribute-Att** (Hu et al., 2018) considers distinguishing the confusing charges is hard by introducing brief and concise law articles, and then designs ten common artificial attributes for charges.

(3) **TOPJUDGE** (Zhong et al., 2018) first takes into account the sequence dependency of the three subtasks of legal judgment prediction in the actual scenario. This model designs a topological multi-task learning framework to use the dependency information.

(4) **MPBFN-WCA** (Yang et al., 2019) takes into account that the judge needs to check again whether the relevant law articles, charges and term of penalty are suitable.

(5) **LADAN** (Xu et al., 2020) considers distinguishing the confusing law articles and design a graph distillation operator to learn the differences among law articles.

(6) **Neurjudge** (Yue et al., 2021) takes into account the circumstances in the actual scenario and use the intermediate results to separate the fact description vector representation. It is one of the state-of-the-art models.

(7) **CR4LJP** is our method.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

Method	Law Articles				Charges				Term of Penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
FLA	0.8853	0.8463	0.8067	0.8188	0.8732	0.8414	0.8134	0.8119	0.3566	0.3279	0.3176	0.3104
Attribute-Att	0.8910	0.8490	0.8357	0.8396	0.8896	0.8587	0.8343	0.8450	0.3686	0.3355	0.3288	0.3246
TOPJUDGE	0.8940	0.8578	0.8348	0.8430	0.8819	0.8513	0.8331	0.8379	0.3668	0.3296	0.3494	0.3275
MPBFN-WCA	0.8944	0.8600	0.8434	0.8478	0.8820	0.8537	0.8393	0.8425	0.3677	0.3417	0.3346	0.3357
LADAN	0.9016	0.8711	0.8556	0.8604	0.8871	0.8588	0.8451	0.8464	0.3718	0.3496	0.3488	0.3383
Neurjudge	0.9112	0.8853	0.8661	0.8720	0.8913	0.8663	0.8486	0.8512	**0.4064**	**0.3780**	**0.3641**	**0.3656**
CR4LJP	**0.9137**	**0.8868**	**0.8785**	**0.8791**	**0.8932**	**0.8675**	**0.8570**	**0.8596**	0.3802	0.3714	0.3398	0.3431

Table 3: Results with GRU-based encoder on CAIL-small dataset. The best results are in bold.

Method	Law Articles				Charges				Term of Penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
FLA	0.9436	0.8471	0.7870	0.8091	0.9383	0.8390	0.7765	0.7993	0.5338	0.4223	0.4033	0.4097
Attribute-Att	0.9512	0.8787	0.7849	0.8137	0.9469	0.8759	0.7821	0.8148	0.5503	0.4552	0.3941	0.4126
TOPJUDGE	0.9502	0.8648	0.8021	0.8246	0.9461	0.8643	0.7943	0.8201	0.5574	0.4583	0.4040	0.4206
MPBFN-WCA	0.9507	0.8733	0.8054	0.8291	0.9457	0.8656	0.7957	0.8189	0.5583	0.4429	0.4110	0.4221
LADAN	0.9530	0.8719	0.8141	0.8345	0.9427	0.8607	0.8070	0.8263	0.5799	0.4833	0.4334	0.4413
Neurjudge	0.9568	0.8841	0.8307	0.8497	0.9505	0.8707	0.8197	0.8356	**0.5805**	0.4851	**0.4611**	**0.4638**
CR4LJP	**0.9594**	**0.8850**	**0.8449**	**0.8576**	**0.9524**	**0.8800**	**0.8235**	**0.8436**	0.5801	**0.4864**	0.4537	0.4560

Table 4: Results with GRU-based encoder on CAIL-big dataset. The best results are in bold.

5.3 Experiment Setting

For the GRU-based encoder, we first use the tool THULAC (Sun et al., 2016) to do word segmentation for the fact description and pre-train the word embedding with the dimension of 200 using word2vec (Mikolov et al., 2013). The maximum text length of the fact description is set to 400 for all the models. The hidden size is set to 150 for all the models. The learning rate is set to 1e-3. For the Bert-based encoder, the learning rate is set to 1e-5. Our model is trained on one V100 GPU (2 V100 GPUs for Bert-based encoder) for 20 epochs and the batch size is 128. We set the hyperparameter α_i to 1 for three subtasks and we use the AdamW optimizer to train our model. For the case retrieval module, we select the same number of 20 cases for each charge under each law article and set the k of Top-k as 5. We use Accuracy (Acc.), Macro Precision (MP), Macro Recall (MR), and Macro F1 (F1) to measure all models following the previous works.

5.4 Overall Results

The experimental results of all the models on three legal judgment prediction subtasks are shown in Table 3 and Table 4. Compared with the best baseline model Neurjudge, our model CR4LJP increases F1 scores of the law article and charge prediction subtasks by 0.81% and 0.98% respectively on the CAIL-small dataset and increases F1 scores of these two subtasks by 0.93% and 0.96% respectively on CAIL-big dataset. The experimental results prove the effectiveness of the model. It should be noted that the results of the term of penalty prediction task of our model on the CAIL-small and CAIL-big datasets are still worse than those of Neurjudge. Neurjudge simulates the actual judicial process and makes fine-grained division of the case description which is based on human knowledge and proved very effective for the term of penalty prediction.

Compared with the results of other baseline models, we can draw the following conclusions:

(1) TOPJUDGE and MPBFN-WCA both make use of the dependencies among the three subtasks to improve the fact representation of a single case. The better results of our model show that retrieving

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

Method	Law Articles				Charges				Term of Penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
Bert	0.9238	0.8987	0.8822	0.8859	0.9139	0.8897	0.8759	0.8792	0.4083	0.3837	0.3486	0.3425
Bert-Crime	0.9235	0.8948	0.8875	0.8872	0.9145	0.8898	0.8844	0.8838	0.4100	0.4013	0.3409	0.3441
Neurjudge+Bert	0.9314	0.9112	0.9041	0.9064	0.9230	0.9065	0.8994	0.9010	**0.4126**	**0.3977**	**0.3594**	**0.3670**
CR4LJP+Bert	**0.9343**	**0.9140**	**0.9043**	**0.9070**	**0.9245**	**0.9067**	**0.9007**	**0.9022**	0.4072	0.3857	0.3447	0.3501

Table 5: Results with Bert-based encoder on CAIL-small dataset.

historical cases help the model get a better representation of the fact to improve the performance of judgment results.

(2) The performance of FLA is worse than those of other neural network models because it directly introduces the Top-k law articles, but the law articles are short and confusing, which may bring some noise. To solve the confusing law article problem, Attribute-Att designs ten artificial attributes and LADAN designs a graph distillation operator to improve the representation of introduced law articles. The better results of our model CR4LJP show that historical case information is more helpful to improve the performance of the legal judgment prediction task than law article information.

(3) For the results of all models on the relevant law article and charge prediction subtasks, the F1 scores on the CAIL-big dataset are worse than those on the CAIL-small dataset, and the accuracy is the opposite. The main reason is that the categories of law articles or charges in the CAIL-big dataset are more unbalanced than those in the CAIL-small dataset.

5.5 Results with Bert Based Encoder

The pre-trained language models achieve the best results on many NLP tasks, such as Bert(Devlin et al., 2019). These models can be used as encoders in our framework, which usually leads to better results. In order to show the flexibility of our model, we compare our model based on Bert encoder with other methods.

• **Bert** uses the fact description as the input and the output of "[CLS]" token as the representation of the fact description. We fine-tune it on the CAIL dataset for the legal judgment prediction subtasks.

• **Bert-Crime** (Zhong et al., 2019) pre-trains Bert on a larger legal dataset. The process of fine-tuning is the same as **Bert**.

• **Neurjudge+Bert** is the Bert-based Neurjudge model, which replaces the GRU encoder with Bert.

• **CR4LJP+Bert** is our Bert-based model.

Due to the limitation of computing resources and the huge amount of parameters of Bert, we only conduct experiments on the CAIL-small dataset. Specifically, on the CAIL-small dataset, the time required for one epoch of training CR4LJP+Bert is about 36 times that of GRU based model (41400s vs 1140s). The experimental results are shown in Table 5. CR4LJP+Bert achieves better results than the GRU-based encoder. This shows the flexibility and effectiveness of our framework.

Compared with other baseline Bert models, the following observations can be observed:

(1) Our model (CR4LJP+Bert) is better than Bert and Bert-Crime indicating that additional case information can improve the performance of the legal judgment prediction task.

(2) Our model is superior to Neurjudge+Bert, which proves once again the effectiveness of our case retrieval framework.

5.6 Ablation Study

In order to verify the effectiveness of the case retrieval module for three subtasks, we perform an ablation study. Specifically, we remove the Top-k historical cases' mean vector from each subtask to study the impact of the case retrieval module. The corresponding models are expressed as **w/o article**, **w/o charge**, and **w/o term**.

The ablation study results of CAIL-small dataset are shown in Figure 3. We can see:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

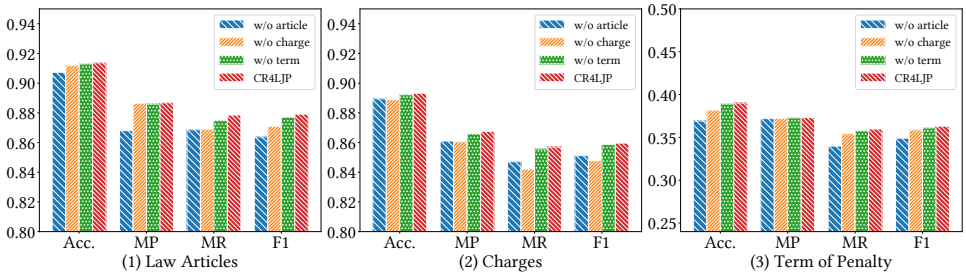


Figure 3: Ablation study on CAIL-small dataset. We remove the Top-k historical cases' mean vector from each subtask to study the impact of the case retrieval module.

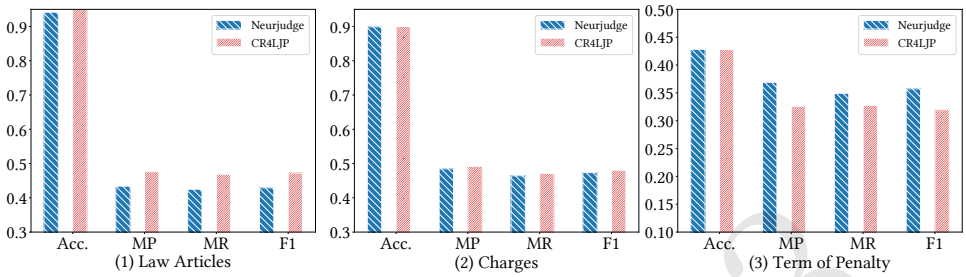


Figure 4: Case study on confusing charges.

(1) Removing the case retrieval module will degrade the performance of the three subtasks. This shows that the case retrieval module is effective.

(2) Removing the case retrieval module has the least impact on the term of penalty prediction subtask, which is in line with our expectations. In reality, the term of penalty prediction needs to be discussed and determined in more detail.

(3) In general, removing the case retrieval module from the law article prediction subtask has the most impact on the legal judgment prediction task. The underlying reason is that this subtask provides the basis for the other two tasks in reality, so it plays the most important role in the legal judgment prediction task.

5.7 Confusing Case study

As shown in Figure 1, Article #114 and Article #115 stipulate some similar charges. It is difficult to distinguish these cases with similar charge labels. In order to intuitively show the effect of models in distinguishing easily confusing cases, we select the cases related to Article #114 and Article #115 in the CAIL-small test set as a tiny dataset and test the baseline model Neurjudge and our model CR4LJP on the dataset.

From the experimental results in Figure 4, it can be seen that our model CR4LJP has better performance on the law article and charge prediction subtasks than Neurjudge, which shows that the case retrieval framework we proposed can effectively improve the ability to distinguish confusing cases.

6 Conclusion

In this paper, we first consider that judges usually look up some typical historical cases before making a judgment. We design a historical case retrieval model framework to simulate this scenario. For the current case, we retrieve the Top-k similar historical cases and get vector representation of these cases using the basic encoder, then we concatenate the mean vector of them to the fact description vector to predict the judgment results. Experimental results show that our method is effective.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

Acknowledgements

We thank all the reviewers for their insightful comments. Zhicheng Dou is the corresponding author. This work was supported by National Key R&D Program of China No. 2022ZD0120103, National Natural Science Foundation of China No. 62272467, Beijing Outstanding Young Scientist Program No. BJJWZYJH012019100020098, Public Computing Cloud, Renmin University of China, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. The work was partially done at Beijing Key Laboratory of Big Data Management and Analysis Methods, and Key Laboratory of Data Engineering and Knowledge Engineering, MOE.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

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

Qian Dong and Shuzi Niu. 2021. Legal judgment prediction via relational learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 983–992. ACM.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland, May. Association for Computational Linguistics.

Anne von der Lieth Gardner. 1984. *An artificial intelligence approach to legal reasoning*. Ph.D. thesis, Stanford University.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2018. Search engine guided neural machine translation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5133–5140. AAAI Press.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 487–498. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1736–1745. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2727–2736. Association for Computational Linguistics.

Zhengyi Ma, Zhicheng Dou, Guanyue Bian, and Ji-Rong Wen. 2020. PSTIE: time information enhanced personalized search. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1075–1084. ACM.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation learning in the real court setting. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 993–1002. ACM.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Tex. L. Rev.*, 42:1006.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. Thulac: An efficient lexical analyzer for chinese.

S Sidney Ulmer. 1963. Quantitative analysis of judicial processes: Some practical and theoretical applications. *Law and Contemporary Problems*, 28(1):164–184.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced ranker-reader for open-domain question answering. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5981–5988. AAAI Press.

Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019. Hierarchical matching network for crime classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 325–334. ACM.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3086–3095. Association for Computational Linguistics.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4085–4091. ijcai.org.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4961–4974. Association for Computational Linguistics.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 973–982. ACM.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

Han Zhang, Zhicheng Dou, Yutao Zhu, and Jirong Wen. 2021. Few-shot charge prediction with multi-grained features and mutual information. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Chinese Computational Linguistics - 20th China National Conference, CCL 2021, Hohhot, China, August 13-15, 2021, Proceedings*, volume 12869 of *Lecture Notes in Computer Science*, pages 387–403. Springer.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3540–3549. Association for Computational Linguistics.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Open chinese language pre-trained model zoo. Technical report.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5218–5230. Association for Computational Linguistics.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2020. Encoding history with context-aware representation learning for personalized search. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1111–1120. ACM.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 801-812, Harbin, China, August 3 - 5, 2023.

SentBench: Comprehensive Evaluation of Self-Supervised Sentence Representation with Benchmark Construction

Xiaoming Liu[1,3], **Hongyu Lin**[1]*, **Xianpei Han**[1,2]*, **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
{xiaoming2021,hongyu,xianpei,sunle}@iscas.ac.cn

Abstract

Self-supervised learning has been widely used to learn effective sentence representations. Previous evaluation of sentence representations mainly focuses on the limited combination of tasks and paradigms while failing to evaluate their effectiveness in a wider range of application scenarios. Such divergences prevent us from understanding the limitations of current sentence representations, as well as the connections between learning approaches and downstream applications. In this paper, we propose SentBench, a new comprehensive benchmark to evaluate sentence representations. SentBench covers 12 kinds of tasks and evaluates sentence representations with three types of different downstream application paradigms. Based on SentBench, we re-evaluate several frequently used self-supervised sentence representation learning approaches. Experiments show that SentBench can effectively evaluate sentence representations from multiple perspectives, and the performance on SentBench leads to some novel findings which enlighten future researches.

1 Introduction

Self-supervised representation learning is considered an important reason for breakthroughs in NLP (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019). And learning effective sentence representations has long been a fundamental challenge. (Kiros et al., 2015; Conneau et al., 2017; Cer et al., 2018). In recent years, various self-supervised sentence representation learning approaches leverage different self-constrained signals, e.g., sentence pairs in the same narratives (Devlin et al., 2019), sentence order (Lan et al., 2019), or sentence permutation (Lewis et al., 2020), to learn representations by training models to distinguish positive instances from negatives.

Even though current self-supervised sentence representation approaches have reached significant progress on some datasets like Semantic Textual Similarity (STS) (Ho and Nvasconcelos, 2020; Gao et al., 2021), benchmarks for evaluation lag far behind the development of methods (Wang et al., 2022). Currently, sentence representations are evaluated in limited tasks and specific paradigms. For example, the most commonly used SentEval benchmark (Conneau and Kiela, 2018) mainly focuses on single sentence classification and semantic similarity tasks. Unfortunately, prior literature shows that performance on STS cannot reflect the effectiveness of sentence representations on a wider range of tasks (Reimers et al., 2016; Zhelezniak et al., 2019; Wang et al., 2022). And available evaluation toolkits assess the same downstream task with a singular paradigm, limiting our perception of methods in different application scenarios. Moreover, current self-supervised sentence representation learning approaches are coupled with multiple factors, including diverse contrastive signals, training losses, and model architectures. Consequently, evaluating whether, where, and how a learning method will benefit the downstream tasks is difficult.

In this paper, we propose SentBench, a new benchmark to comprehensively evaluate sentence representations with various downstream tasks and evaluation paradigms. As shown in Figure 1, SentBench

*Corresponding authors.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

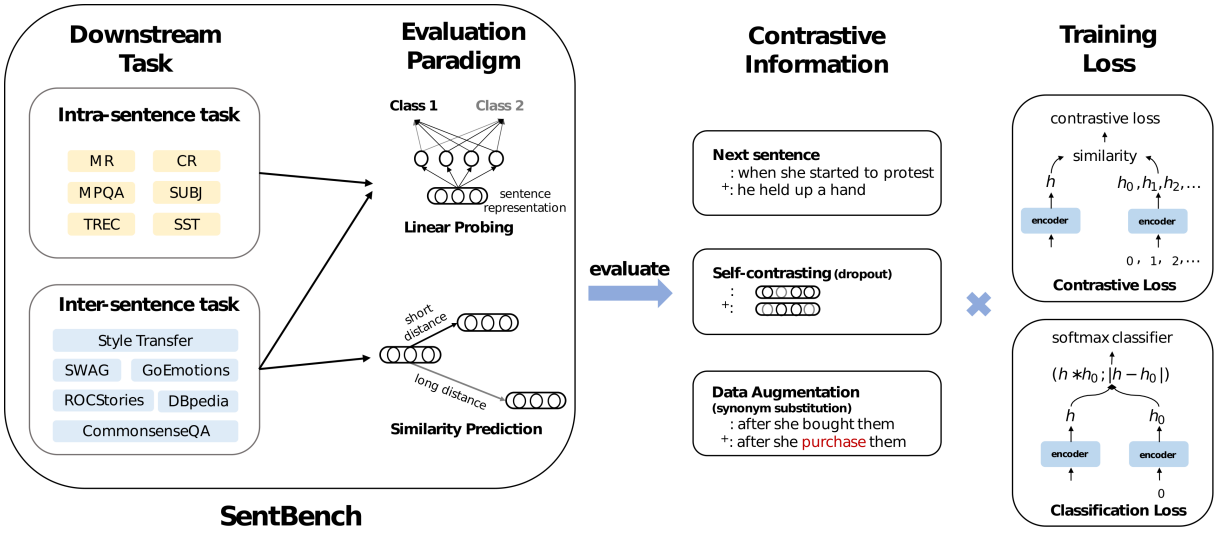


Figure 1: The framework of the paper (SentBench and decoupling analysis scheme).

contains 12 kinds of NLP tasks, including sentiment classification, question answering, story cloze, etc., and three evaluation paradigms, including single sentence classification, sentence pair classification and sentence pair contrasting (Zhu et al., 2018). The classification paradigm trains a simple additional classifier to assess information within representations for single sentence tasks or identify the connection between two candidate representations for pair-wise tasks. Besides, contrasting paradigm is similar to common retrieval or ranking scenario. Finally, SentBench constructs 18 datasets, which cover diverse tasks and common applications of sentence representations.

Based on SentBench, we re-evaluate several widely used self-supervised sentence representation learning approaches. We decouple previous approaches from two perspectives to identify critical factors: contrasting knowledge applied to construct positive instances and training losses used to optimize models. Specifically, we concentrate on three contrasting knowledge, including next sentence prediction (Devlin et al., 2019), self-contrasting (Yan et al., 2021; Gao et al., 2021) and data augmentation (Zhang et al., 2015; Feng et al., 2021), as well as two widespread training losses, including contrastive loss and classification loss. By thoroughly comparing different approaches on SentBench, we find that the advantages of the state-of-the-art methods can not be exhibited consistently to a broader range of downstream tasks and evaluation paradigms. Furthermore, the applied training loss leads to more significant impacts than contrasting knowledge. These findings shed some light on future research on sentence representation learning.

2 Benchmark Construction

2.1 Tasks

SentBench covers 12 downstream tasks for evaluating sentence representations, divided into single sentence and sentence pair tasks. In the following, we will briefly describe tasks in SentBench.

Single sentence tasks aim to classify sentence representations into corresponding categories. Because the previous SentEval[0] benchmark has covered extensive single sentence classification tasks, SentBench inherits all of them, including sentiment analysis (MR, SST) (Pang and Lee, 2005; Socher et al., 2013), Opinion Polarity (MPQA, SUBJ) (Wiebe et al., 2005; Pang and Lee, 2004), Question type (TREC) (Voorhees and Tice, 2000), product reviews (CR) (Hu and Liu, 2004).

Sentence pair tasks aim to identify sentence pairs with specific connections. We investigate six tasks covering various fields of downstream applications of NLP (Table 1):

[0] https://github.com/facebookresearch/SentEval

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Dataset	Classification			Contrasting
	Train size	Valid Size	Test Size	
SWAG	56,131	18,711	18,711	20,006
DBpedia	89,965	27,988	27,989	69,971
GoEmotions	54,535	18,178	18,179	4,590
ROCStories	2,513	-	629	1,571
StyleTransfer	24,986	8,328	8,330	2,500
CommonsenseQA	13,154	4,384	4,386	1,221

Table 1: The statistics of sentence pair tasks.

- **DBpedia** (Zhang et al., 2015), which identifies whether a pair of sentences come from the same category;
- **Style Transfer** (ST) (Jhamtani et al., 2017), which distinguishes whether modern English and Shakespearean English expresses same content;
- **GoEmotions** (GoEmo) (Demszky et al., 2020), which recognizes whether a sentence pair expresses similar fine-grained emotion;
- **ROCStories** (ROC) (Mostafazadeh et al., 2016), which predicts whether a given sentence is the proper ending to a four-sentence story;
- **CommonsenseQA** (CQA) (Talmor et al., 2019), which determines if candidate answers match a commonsense question;
- **SWAG** (Zellers et al., 2018), which predicts correct answer for a question about grounded situations.

2.2 Evaluation paradigm

We design three evaluation paradigms in SentBench:

- **single sentence classification** directly leverage sentence representations as features with a simple classifier to assess how much desirable information is contained in representations;
- **sentence pair classification** trains a simple classifier that determines whether there is a specific connection between candidate sentences, that is mapping a pair of sentence representation $(\mathbf{x_1}, \mathbf{x_2})$ into corresponding label;
- **sentence pair contrasting** distinguishes a sentence from candidates that are more likely to share a specific relationship with the given sentence, i.e., given a target sentence x and two candidates (x^+, x^-), sentence pair contrasting selects more suitable candidate based on the similarity between x, x^+, and x^-.

Note that the classification paradigm requires data to train additional classifier parameters, while sentence pair contrasting depends on the similarity between sentence pairs by directly calculating certain distance metrics (e.g., cosine similarity) without additional training instances. Therefore, we provide training and development sets for classification tasks.

3 Experiment Setup

Based on SentBench, we re-evaluate several most frequently used self-supervised sentence representation methods. Since contrasting knowledge and training losses are usually coupled, it is challenging to directly identify critical factors for successful sentence representations from previous works. To this end, this paper explores different combinations of contrasting knowledge and training losses to investigate the effects of distinct factors.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Contrasting Knowledge. We exploit three popular contrasting knowledge sources:

- **narrative contrasting**, which predicts whether a hypothesis sentence belongs to the same narrative with a premise, is also known as next sentence prediction (NSP);
- **self-contrasting**, which disturbs sentence representations at feature-level, tries to distinguish representations stemming from the same instance. SimCSE (Gao et al., 2021) is one of the most popular methods, which creates contrasting pairs via random dropout from neural networks;
- **data augmentation**, which modifies the original instances via some rule-based modification, and tries to distinguish original instances from others.

In this paper, we apply NSP (Devlin et al., 2019), two-times Dropout (Dropout) (Gao et al., 2021), and synonym substitution (DA) (Wu et al., 2020) as each knowledge sources, respectively.

Training Loss. Contrastive loss and classification loss are the most popular loss functions in self-supervised sentence representation learning. Given an instance $\mathbf{x}$, **contrastive loss** (CTR) (Van den Oord et al., 2018) aims to distinguish positive instance representation $\mathbf{x}^+$ from a batch of negatives:

$$\mathcal{L}_{CTR}(\theta) = -\log \frac{e^{sim(\mathbf{x},\mathbf{x}^+)/\tau}}{\sum_{\mathbf{x}_i \in batch} e^{sim(\mathbf{x},\mathbf{x}_i)/\tau}}$$

where τ is a temperature hyperparameter and sim is a similarity function (e.g., cosine similarity).

classification loss (CLS) classifies sentence pairs representation into corresponding semantic labels:

$$\mathcal{L}_{CLS}(\theta) = -\log P(y=1|\mathbf{x} * \mathbf{x}^+) - \sum_{\mathbf{x}^- \in batch} \log P(y=0|\mathbf{x} * \mathbf{x}^-)$$

where $*$ is the concatenation of representations.

Implementation Details. We implement the above-mentioned approaches based on $\text{BERT}_{\texttt{base}}$ (uncased) (Lan et al., 2019) and $\text{RoBERTa}_{\texttt{base}}$ (Liu et al., 2019). To compare the benefit of different approaches, we also implement two token-aggregation approaches without further learning as baselines, which regard average representations of all tokens or the [CLS][1] representation of the last layer of models as sentence representation.

In this paper, we use BookCorpus (Zhu et al., 2015) to construct the next sentence samples. Devlin et al. (2019) concatenate two sentences with [SEP] and feed the [CLS] representation into the classifier. A slight difference from the above approach is that we first obtain the [CLS] representations of two sentences separately and then concatenate them to learn the next sentence prediction. For self-supervised sentence representation learning with different combinations of loss functions and contrasting knowledge, we train models for one epoch on 10^6 sentences from BookCorpus and set batch size to 64. The temperature τ of contrastive loss is set to 0.05, and max sequence length is set to 32. Cosine similarity is the default distance metric and similarity function. All experiments are run in NVIDIA TITAN RTX GPUs. Following Gao et al. (2021) and Wu et al. (2020), the best checkpoint on the development set of STS is saved for evaluation. We use NLPAUG[2] for synonym substitution and take other sentences in the same mini-batch as negatives.

4 Empirical Findings

Table 2, 3 and 4 show the experiment results on three evaluation paradigms in SentBench, respectively. From these empirical results, we obtain the following findings.

[1]We discard the MLP layer over [CLS] for evaluation.
[2]https://github.com/makcedward/nlpaug

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Model	MR	CR	MPQA	SUBJ	SST	TREC	AVG
BERT-AVG	**82.24**^{1}	87.39^{1}	**88.71**^{2}	95.45^{3}	84.62^{4}	**91.80**^{1}	88.37^{2}
BERT-[CLS]	81.83^{2}	**87.39**^{1}	88.21^{6}	**95.48**^{2}	**86.91**^{1}	91.33^{2}	**88.53**^{1}
Dropout (CTR)	80.43^{4}	85.09^{5}	88.43^{4}	94.64^{6}	84.66^{3}	**90.67**^{3}	87.32^{4}
Dropout (CLS)	67.73^{8}	70.09^{8}	85.50^{7}	87.93^{8}	75.36^{8}	79.33^{8}	77.66^{8}
NSP (CTR)	**81.13**^{3}	**87.18**^{3}	88.34^{5}	**95.53**^{1}	**85.05**^{2}	89.67^{5}	**87.82**^{3}
NSP (CLS)	78.92^{6}	85.59^{4}	88.54^{3}	95.10^{4}	83.42^{6}	89.87^{4}	86.91^{6}
DA (CTR)	80.16^{5}	84.64^{6}	**89.33**^{1}	94.72^{5}	83.98^{5}	89.67^{5}	87.08^{5}
DA (CLS)	73.89^{7}	77.25^{7}	80.10^{8}	90.74^{7}	77.46^{7}	84.73^{7}	80.70^{7}
RoBERTa-AVG	**83.43**^{3}	**88.58**^{2}	**86.75**^{5}	**95.22**^{2}	**87.26**^{3}	**91.93**^{1}	**88.80**^{2}
RoBERTa-[CLS]	81.27^{4}	86.01^{5}	84.18^{6}	94.15^{4}	86.66^{4}	83.00^{6}	85.88^{6}
Dropout (CTR)	80.18^{5}	85.43^{6}	87.55^{2}	93.22^{6}	85.35^{5}	87.80^{5}	86.59^{5}
Dropout (CLS)	60.58^{7}	63.84^{8}	77.82^{7}	81.10^{7}	70.45^{7}	66.60^{7}	70.07^{7}
NSP (CTR)	**85.90**^{1}	**90.60**^{1}	**88.96**^{1}	**95.39**^{1}	**91.12**^{1}	**91.33**^{2}	**90.55**^{1}
NSP (CLS)	83.62^{2}	88.51^{3}	87.51^{3}	94.72^{3}	87.75^{2}	89.67^{3}	88.63^{3}
DA (CTR)	80.03^{6}	86.78^{4}	87.12^{4}	93.23^{5}	84.47^{6}	89.13^{4}	86.79^{4}
DA (CLS)	56.02^{8}	63.97^{7}	74.10^{8}	77.59^{8}	61.25^{8}	65.60^{8}	66.42^{8}

Table 2: Accuracies on single sentence classification tasks and corner markers represent the performance rank. CTR: contrastive loss; CLS: classification loss.

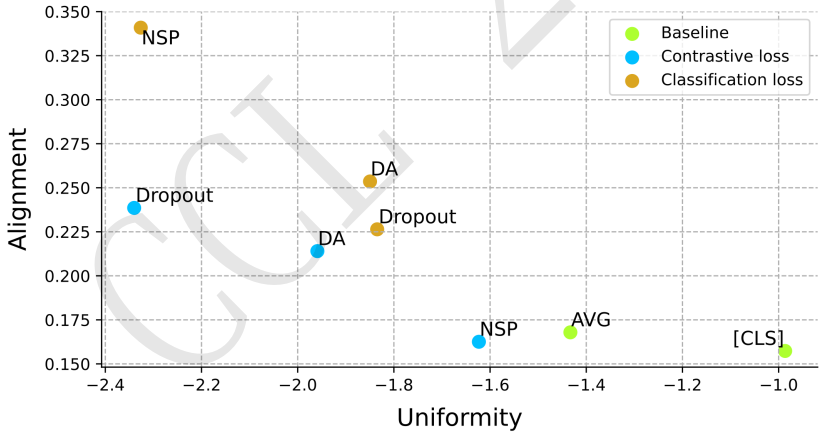


Figure 2: Alignment and uniformity plot of models based on BERT. For both alignment and uniformity, lower numbers are better.

Finding 1. Training loss is a more critical factor than contrasting knowledge. We find that the selection of training loss has more significant impacts than the selection of contrasting knowledge, and contrastive loss significantly outperforms classification loss across all contrasting knowledge, models, and evaluation paradigms. Note that previously NSP is commonly coupled with classification loss and therefore achieves little performance superiority (Liu et al., 2019). However, from our experiments, NSP trained with contrastive loss can bring significant performance improvements. To further investigate how contrasting knowledge and training loss influence sentence representations, we calculate the alignment and uniformity, two quantified quality evaluation metrics for sentence representations

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

(Wang and Isola, 2020). As shown in Figure 2, we can see that different contrasting information is essentially a trade-off between alignment and uniformity. And contrastive loss outperforms classification loss with better alignment and uniformity, which reveals the underlying reason for the superior performances.

Model	ST	DBpedia	GoEmo	ROC	CQA	SWAG	AVG
BERT-AVG	**86.03**^{1}	91.35^{6}	**56.64**^{5}	**63.12**^{2}	**58.38**^{3}	**65.81**^{2}	**70.22**^{2}
BERT-[CLS]	85.76^{3}	**91.57**^{5}	56.51^{6}	60.15^{4}	54.30^{6}	64.19^{3}	68.75^{6}
Dropout (CTR)	84.19^{6}	92.29^{4}	57.33^{3}	56.60^{6}	59.69^{2}	62.52^{5}	68.77^{5}
Dropout (CLS)	79.19^{8}	79.83^{8}	52.18^{8}	53.58^{8}	50.97^{7}	52.94^{8}	61.45^{8}
NSP (CTR)	**85.93**^{2}	**96.07**^{1}	**59.06**^{1}	**64.07**^{1}	**60.11**^{1}	**66.05**^{1}	**71.88**^{1}
NSP (CLS)	84.40^{5}	95.67^{2}	57.18^{4}	59.41^{5}	55.88^{5}	63.41^{4}	69.33^{4}
DA (CTR)	84.92^{4}	93.34^{3}	57.78^{2}	61.05^{3}	57.83^{4}	61.08^{6}	69.33^{3}
DA (CLS)	80.42^{7}	83.60^{7}	53.06^{7}	54.00^{7}	50.82^{8}	54.83^{7}	62.79^{7}
RoBERTa-AVG	**83.41**^{3}	89.17^{6}	**54.90**^{5}	**59.46**^{4}	**54.43**^{5}	**65.91**^{1}	**67.88**^{4}
RoBERTa-[CLS]	81.60^{5}	**89.78**^{5}	53.76^{6}	55.12^{6}	50.47^{7}	64.22^{2}	65.83^{6}
Dropoput (CTR)	82.09^{4}	92.74^{4}	55.38^{4}	55.75^{5}	56.72^{2}	60.46^{5}	67.19^{5}
Dropout (CLS)	75.16^{7}	69.62^{7}	50.72^{7}	53.15^{8}	49.93^{8}	51.76^{7}	58.39^{7}
NSP (CTR)	**84.83**^{1}	**96.49**^{1}	**58.95**^{1}	**66.93**^{1}	**60.41**^{1}	**63.85**^{3}	**71.91**^{1}
NSP (CLS)	83.46^{2}	95.74^{2}	56.70^{3}	63.01^{2}	55.82^{4}	61.54^{4}	69.38^{2}
DA (CTR)	81.47^{6}	94.69^{3}	57.88^{2}	59.62^{3}	55.83^{3}	59.15^{6}	68.11^{3}
DA (CLS)	74.12^{8}	66.97^{8}	50.16^{8}	53.21^{7}	50.52^{6}	51.30^{8}	57.71^{8}

Table 3: Accuracies on sentence pair classification tasks and corner markers represent the performance rank. CTR: contrastive loss; CLS: classification loss.

Finding 2. Narrative contrasting provides more useful information for a wide range of single sentence and sentence pair tasks. Experiments show that the NSP with contrastive loss achieves satisfactory performance in almost all settings. Besides, we can see that performance improvement on RoBERTa is more significant than that of BERT. This may be because the [CLS] representation of BERT has been pretrained with NSP signals and therefore already contain such kind of knowledge. Furthermore, we find that self-contrasting strategies, which are reported to achieve superior performance on STS benchmarks (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015; Agirre et al., 2016), do not perform well in SentBench. We believe that this is because, as previous findings have shown (Wang et al., 2022), STS tasks have a weak correlation with downstream tasks. Therefore, evaluations on STS benchmarks are not universal, revealing the necessity of building SentBench.

Finding 3. Self-supervised contrastive sentence representation learning leads to more significant improvements on sentence pair contrasting tasks. We can see that for BERT-AVG and RoBERTa-AVG, there are 6.2% and 12% of average performance improvements of all methods with contrastive loss, which is significantly higher than that on the other two tasks. We speculate that contrastive loss is more appropriate for similarity-based evaluation, which substantially improves the consistency between sentence representation distribution and downstream applications. Furthermore, single sentence and sentence pair classification tasks introduce an additional trainable classifier, which may weaken the effectiveness of self-supervised pretraining. Consequently, self-supervised contrastive sentence representation is more suitable for similarity-based scenarios without additional supervised signals, which is also consistent with recent advances of these methods on previous STS benchmarks (Gao et al., 2021).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Model	ST	DBpedia	GoEmo	ROC	CQA	SWAG	AVG
BERT-AVG	63.88[8]	**85.89**[5]	**57.02**[4]	58.75[4]	**52.99**[5]	**56.50**[5]	**62.50**[5]
BERT-[CLS]	**65.52**[6]	74.72[6]	53.09[6]	**59.90**[3]	52.09[6]	54.19[6]	59.92[6]
Dropout (CTR)	73.16[2]	91.43[4]	57.56[2]	60.53[2]	**67.49**[1]	62.01[2]	68.70[2]
Dropout (CLS)	73.16[2]	66.53[7]	52.96[7]	52.45[8]	51.68[7]	51.30[7]	58.01[7]
NSP (CTR)	71.84[4]	**94.68**[1]	**57.82**[1]	**62.70**[1]	65.85[3]	**63.24**[1]	**69.35**[1]
NSP (CLS)	64.72[7]	94.62[2]	56.27[5]	56.02[6]	61.51[4]	57.48[4]	65.10[4]
DA (CTR)	**75.52**[1]	91.76[3]	57.47[3]	57.73[5]	66.83[2]	59.64[3]	68.16[3]
DA (CLS)	71.48[5]	64.02[8]	52.14[8]	52.51[7]	49.80[8]	50.86[8]	56.80[8]
RoBERTa-AVG	61.20[8]	67.91[6]	50.11[8]	52.13[8]	55.61[6]	51.32[6]	56.38[7]
RoBERTa-[CLS]	**73.96**[3]	**86.20**[5]	**51.90**[5]	**58.82**[5]	**56.35**[5]	**60.32**[3]	**64.59**[5]
Dropout (CTR)	**75.68**[1]	90.76[4]	55.88[3]	60.09[3]	64.86[2]	61.98[2]	68.21[2]
Dropout (CLS)	70.60[6]	63.38[7]	51.35[6]	56.72[6]	52.25[7]	49.94[7]	57.37[6]
NSP (CTR)	69.64[7]	**96.78**[1]	**58.26**[1]	**64.74**[1]	**65.44**[1]	**62.96**[1]	**69.64**[1]
NSP (CLS)	70.80[4]	95.17[2]	55.53[4]	63.91[2]	61.43[4]	59.77[4]	67.77[3]
DA (CTR)	74.76[2]	94.17[3]	57.71[2]	59.01[4]	61.92[3]	57.00[5]	67.43[4]
DA (CLS)	70.72[5]	59.48[8]	50.13[7]	52.32[7]	46.85[8]	49.75[8]	54.87[8]

Table 4: Accuracies on sentence pair contrasting tasks and corner markers represent the performance rank.

5 Related Works

SentEval vs SentBench SentEval and SentBench are both benchmarks that evaluate the quality of sentence representations in natural language processing tasks. SentEval consists of a set of 17 downstream tasks and 10 probe tasks, including sentiment analysis, natural language inference, paraphrase detection, and text similarity. However, the tasks and methods in SentEval have fallen behind in recent years due to the rapid development of models and methods.

SentBench builds on SentEval, expanding the sentence-pair tasks to include six new datasets such as commonsense QA, story generation, and fine-grained sentiment analysis. Previous studies have shown that the performance of text semantic similarity tasks cannot reflect the effectiveness of sentence representations in more downstream tasks (Reimers et al., 2016; Zhelezniak et al., 2019; Wang et al., 2022). Unlike SentEval, SentBench replaces text similarity tasks with contrasting tasks, which can more objectively reflect the actual application performance of sentence representations. Additionally, SentBench adds different evaluation paradigms to enrich the evaluation forms of the data, which can provide different understanding perspectives for the same downstream task.

GLUE vs SentBench The General Language Understanding Evaluation (GLUE) benchmark is a collection of nine natural language processing tasks designed to assess the performance of language models in various natural language understanding tasks, including sentiment analysis, question answering, and natural language inference. Unlike SentBench, which aims to evaluate sentence representation models and methods, GLUE is designed to evaluate and analyze natural language understanding systems. Although both benchmarks contain sentence representation-related applications, the differences in their design goals result in differences in datasets and usage methods. While SentBench focuses on the generalization and universality of sentence representations, GLUE tests the overall ability of the model. Additionally, the datasets used in GLUE and SentBench are complementary, as SentBench does not currently collect data relevant to natural language inference tasks. Thus, SentBench could look to GLUE's

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

relevant content for future expansion.

Probing Researchers have not only focused on building more efficient evaluation benchmarks but also used various probing tasks to uncover the underlying principles of sentence representation, such as identifying syntactic and semantic information, as well as subtle perturbations. These evaluation tasks offer insights into which factors are challenging for sentence representation and which can better distinguish different models, driving the development of sentence representation. In their attempt to analyze sentence representation, Adi et al. (2016) designed three evaluation tasks that focused only on surface information, such as sentence length, sentence content, and word order, and experimented with popular methods. However, these evaluation tasks failed to reflect the syntactic, semantic, and other knowledge of sentence representation. To address this limitation, Conneau et al. (2018) designed and collected 10 probing tasks that were divided into categories of surface, syntactic, and semantic information, revealing differences and connections between different methods. Furthermore, Zhu et al. (2018) proposed a triplet evaluation framework that generated triplet sentences to explore how syntactic structure or semantic changes in a given sentence affected inter-sentence similarity. This approach not only evaluated the performance of different sentence representation methods in capturing different semantic attributes but also avoided bias from human annotation data, providing a better understanding of these methods. Our work is similar to the previously mentioned research in that we aim to investigate the underlying mechanisms of sentence representation learning through thorough more comprehensive evaluation and decoupling analysis.

6 Conclusion

In this paper, we propose a new universal sentence evaluation benchmark SentBench, which introduces more downstream tasks and evaluation paradigms. Furthermore, we decouple and analyze the effects of contrasting knowledge and training losses on sentence representations. Empirical findings show that training losses play a more critical role in self-supervised sentence representation learning and help us better understand and design sentence representation learning algorithms.

7 Limitations

Currently, SentBench mainly covers English datasets, and therefore can not evaluate whether self-supervised representation learning methods have some language-specific properties. Besides, due to the limitation of time, we mainly experiment with BERT and RoBERTa without evaluating more self-supervised sentence representations methods, such as Sentence-T5 (Ni et al., 2022). Finally, we mainly focus on the performance of models on SentBench without discussing more details of the training process, which is also an important aspect of self-supervised sentence representations.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. U1936207, 62122077 and 62106251.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM),*

Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, August. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July. Association for Computational Linguistics.

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

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, August. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Chih-Hui Ho and Nuno Nvasconcelos. 2020. Contrastive learning with adversarial examples. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17081–17093. Curran Associates, Inc.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

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

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland, May. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain, July.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Bin Wang, C.-C. Jay Kuo, and Haizhou Li. 2022. Just rank: Rethinking evaluation with word and sentence similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077, Dublin, Ireland, May. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

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

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, October-November. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation coefficients and semantic textual similarity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 951–962, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December.

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia, July. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Adversarial Network with External Knowledge for Zero-Shot Stance Detection

Chunling Wang[1], Yijia Zhang[1*], Xingyu Yu[1], Guantong Liu[2], Fei Chen[1], Hongfei Lin[3]
[1]College of Information Science and Technology, Dalian Maritime University / Dalian, China
[2]College of Artificial Intelligence, Dalian Maritime University / Dalian, China
[3]College of Computer Science and Technology, Dalian University of Technology / Dalian, China
{wangchunling, zhangyijia, 1120211509, lgt, chenf}@dlmu.edu.cn
hflin@dlut.edu.cn

Abstract

Zero-shot stance detection intends to detect previously unseen targets' stances in the testing phase. However, achieving this goal can be difficult, as it requires minimizing the domain transfer between different targets, and improving the model's inference and generalization abilities. To address this challenge, we propose an adversarial network with external knowledge (ANEK) model. Specifically, we adopt adversarial learning based on pre-trained models to learn transferable knowledge from the source targets, thereby enabling the model to generalize well to unseen targets. Additionally, we incorporate sentiment information and common sense knowledge into the contextual representation to further enhance the model's understanding. Experimental results on several datasets reveal that our method achieves excellent performance, demonstrating its validity and feasibility.

Keywords: Zero-shot stance detection , Adversarial learning , External knowledge , Contrastive learning

1 Introduction

Stance detection (Küçük and Can, 2020; Mohammad et al., 2016; Augenstein et al., 2016) is a significant task in NLP, focusing on identifying the stance (e.g., against, favor, or neutral) conveyed in the text towards a given target. It can be efficiently applied to social opinion analysis (Lai et al., 2020), rumor detection (Kumar and Carley, 2019), and other research fields by mining text opinions.

Traditional intra-target stance detection (Mohammad et al., 2016) has limited applications since it requires training and testing under the same target and depends heavily on labeled data to achieve excellent performance. With the frequent and vast updates of topics on social platforms, manually labeling new targets becomes expensive and time-consuming, making it impractical to create a labeled dataset with all potential targets (Wang et al., 2020). Therefore, the study of zero-shot stance detection (Allaway and Mckeown, 2020) for unseen targets is essential and promising.

To tackle the zero-shot stance detection task, existing works generally incorporate external knowledge (Liu et al., 2021) as support for inference or introduce attention mechanisms (Allaway and Mckeown, 2020) to capture the relationships between targets, which do not explicitly model of the transferable knowledge between source and destination targets. Some methods solely focus on employing adversarial training (Allaway et al., 2021; Xie et al., 2022) to learn a target-invariant representation of the text content, disregarding the possibility that the model may encounter challenges in correctly predicting sentences that contain implicit viewpoints or require more profound understanding.

For example 1 in Table 1, the document does not explicitly mention the target "Donald Trump." If the model is unaware that Donald Trump is affiliated with the Republican Party, it is easy to misclassify the stance as neutral. Therefore, by incorporating common sense knowledge into adversarial networks and supplementing the target-related concept representations in the knowledge base, we can help the model more efficiently understand the text content, thus improving its generalization. In addition, we find a

©*Corresponding Author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

Text	Target	Gold Label
I do not understand why the **Republicans** don't dismiss him.	Donald Trump	Against
@HillaryClinton **bad** wife, **bad** role model for women, **bad** lawyer, **bad** First Lady, **bad** Senator, **horrible** Secretary of State.	Hillary Clinton	Against

Table 1. Examples of zero-shot stance detection.

certain correlation between sentiment information and stance detection (Li and Caragea, 2019). For example 2 in Table 1, when a document contains some negative words, it generally implies an Against stance. Stance detection will perform better if some sentiment knowledge can be acquired concurrently.

Motivated, on the one hand, based on the knowledge transfer ability of pre-trained models, we jointly embed the text and target into BERT and sentiment-aware BERT (noted as SentiBERT), and employ a cross-attention module to integrate the sentiment information extracted by SentiBERT with the contextual representations, resulting in semantic feature representations of the text. Meanwhile, we impose supervised contrastive learning (Liang et al., 2022) to make the model learn to distinguish stance category features in the potential distribution space. We separate the target-specific and target-invariant representations using a feature separator, then feed the target-invariant representation into the target discriminator for adversarial training, which enables the model to learn robust and transferable representations that can generalize well across different targets. On the other hand, we extract document-specific subgraphs from ConceptNet, and obtain concept representations of the common sense graph by using a graph autoencoder trained on the ConceptNet subgraph, which is fused into the text representation to enhance the model's performance. Our contributions are as follows:

(1) Our proposed ANEK model utilizes semantic information, sentiment information and common sense knowledge for zero-shot stance detection, especially adding sentiment information to assist stance detection and implicit background knowledge to enhance the model's comprehension.

(2) We employ adversarial training to learn target-invariant information to transfer knowledge effectively. Stance contrastive learning is used to enhance the inference of the model.

(3) We experimentally demonstrate that ANEK obtains competitive results on three datasets, and the extension to target stance detection is also effective.

2 Related Work

2.1 Stance Detection

Stance detection is the study of determining a text's viewpoint on a prescriptive target. (Küçük and Can, 2020). Previous studies have primarily focused on scenarios where the training and testing sets share the same target, known as intra-target stance detection (Augenstein et al., 2016; Mohammad et al., 2016). However, when new topics emerge, there is insufficient labeled data. Some studies explore cross-target stance detection (Liang et al., 2021; Wei and Mao, 2019; Xu et al., 2018), which trains a model on one target and tests it on another related target. Xu et al. (2018) presented a self-attentive model to extract shared features between targets. Wei et al. (2019) further exploited the hidden topics between targets as transferred knowledge. In contrast, zero-shot stance detection does not rely on any assumption of target correlation and is a more general study that can handle irregular target emergence.

Allaway et al. (2020) developed a dataset containing multiple targets and presented a topic-grouping attention model to capture implicit relationships between them. Liu et al. (2021) utilized the structural and semantic information of the common sense knowledge graph to enhance the model's inference. Allaway et al. (2021) regarded each target as a domain and modeled the task as a domain adaptation problem, which successfully learnd the target-invariant representation. Liang et al. (2022) designed an agent task that distinguished stance expression categories and implemented hierarchical contrastive learning. These works are considered incomplete as they overlook the impact of external knowledge containing sentiment information on the model. Whereas, we not only learn transferable target-invariant knowledge, but also take into account the introduction of multiple knowledge to enhance semantic infor-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

mation, further improving the model's predictive ability. To the best of our knowledge, we are the first to systematically introduce external knowledge into adversarial networks and achieve good results.

2.2 Adversarial Domain Adaptation

Domain adaptation mainly aims to minimize domain differences, ensure available knowledge transfer, and increase the model's generalization ability. Adversarial loss methods, inspired by the generative adversarial network (GAN) (Goodfellow et al., 2014), have been commonly applied to domain adaptation. Ganin et al. (2016) proposed a domain adversarial neural network (DANN), which utilized a gradient reversal layer to obfuscate the domain discriminator and enable the feature extractor to capture domain-invariant knowledge. Tzeng et al. (2017) presented an adversarial discriminative domain adaptation (ADDA) model, which involved a discriminative method, GAN loss, and unshared weights to decrease the domain disparity. Therefore domain adaptation is an effective solution for the zero-shot stance detection task.

2.3 External Knowledge

Neural networks enhanced with external knowledge have been used for various NLP tasks, like dialogue generation, sentiment classification, and stance detection. Ghosal et al. (2020) employed a domain adversary framework to handle cross-domain sentiment analysis and further improved the performance by injecting common sense knowledge using ConceptNet. Zhu et al. (2022) incorporated target background knowledge from Wikipedia into the stance detection model. In addition, sentiment information is useful external knowledge for stance detection tasks. Li et al. (2019) designed a sentiment classification task as an auxiliary task and built sentiment and stance vocabularies to guide attention mechanisms. Hardalov et al. (2022) adopted a pre-trained sentiment model to generate sentiment annotations for text, which improved cross-lingual stance detection performance. Based on the above work, we simultaneously consider introducing common sense and sentiment knowledge to aid stance detection.

3 Method

The structure of our ANEK model is displayed in Figure 1, which mainly contains two parts. (1) Knowledge graph training: we train a graph autoencoder using ConceptNet relation subgraphs. (2) Stance detection: we obtain context and sentiment information with pre-trained models, use contrastive learning to improve representation quality, separate features and perform adversarial learning, and finally incorporate the extracted common sense knowledge graph features to implement stance detection.

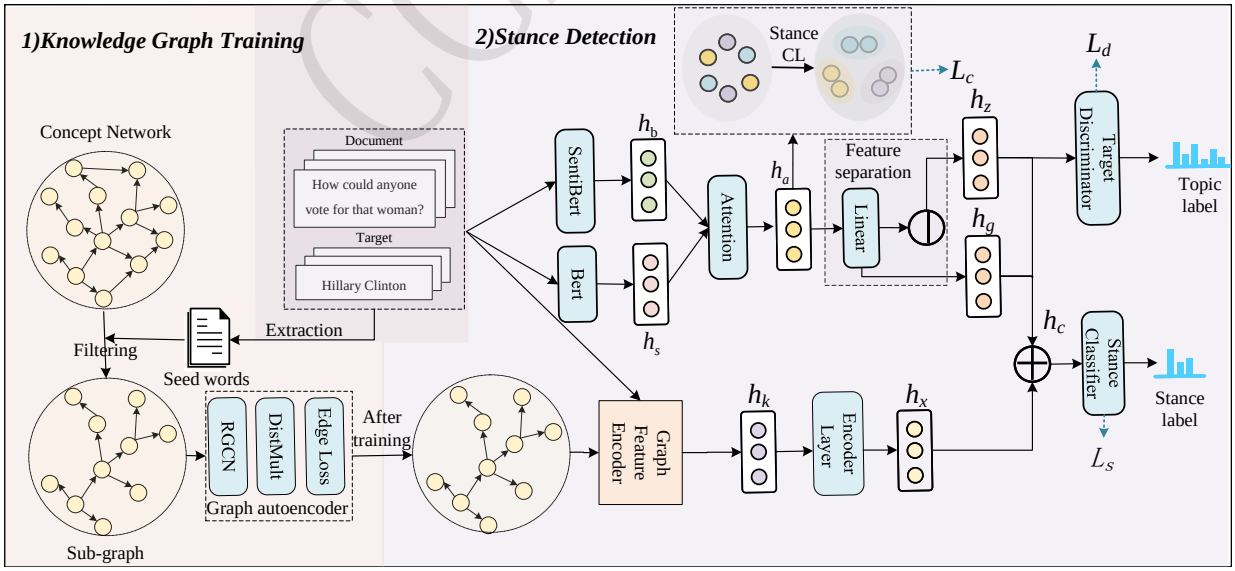


Figure 1. Overview of the ANEK model

3.1 Task Description

Suppose we are given an annotated dataset $D_s = \{x_s^i, t_s^i, y_s^i\}_{i=1}^{N_s}$ from source targets and an unlabeled dataset $D_d = \{x_d^i, t_d^i\}_{i=1}^{N_d}$ from a destination target (unknown target), where x is a document, t and y are its corresponding target and stance label, respectively, and N is the number of examples. The purpose of zero-shot stance detection is to train the model using labeled data from multiple source targets to predict the stance labels of the unknown target examples.

3.2 Knowledge Graph Training

3.2.1 Common Sense Subgraph Generation

ConceptNet is a common sense knowledge base denoted as a directed graph $G = (V, E, R)$, where concepts $v_p \in V$, edges $(v_p, r, v_q) \in E$, and $r \in R$ is the relation type of the edge between v_p and v_q. Given that ConceptNet contains tens of millions of triplet relations like (cake, IsA, dessert), we use it to construct our knowledge subgraph. To be specific, we extract unique nouns, adverbs, and adjectives from the datasets of all targets as seed words. We then extract all triples that are one edge distance away from these seed concepts to obtain a subgraph $G^{'} = \left(V^{'}, E^{'}, R^{'}\right)$.

3.2.2 Graph Autoencoder Pre-training

To integrate common sense knowledge into our model, we obtain the concept representations in the subgraph $G^{'}$ by training a graph autoencoder composed of a RGCN encoder and a DistMult decoder (Schlichtkrull et al., 2018). We feed the incomplete set of edges $\hat{E^{'}}$ from $E^{'}$ into the autoencoder. We then assigns scores to the potential edges (v_p, r, v_q) to ascertain the possibility of these edges being in $E^{'}$.

Encoder Module. To obtain enriched feature representations of the target-related concepts, we utilize two stacked RGCN encoders to compose our encoder module. RGCN can create a rich stance aggregated representation for each concept by combining related concepts in the process of neighborhood-based convolutional feature transformation. Specifically, we randomly initialize the hidden vector g_p of concept v_p and then transform it into the stance aggregated hidden vector h_p by a two-step graph convolution.

$$f(x_p, l) = \sigma(\sum_{r \in R} \sum_{q \in N_p^r} \frac{1}{a_{p,r}} W_r^{(l)} x_q + W_0^{(l)} x_p) \tag{1}$$

$$h_p = h_p^{(2)} = f(h_p^{(1)}, 2); h_p^{(1)} = f(g_p, 1) \tag{2}$$

where f denotes the encoder function with vector x_p and layer l as inputs, σ is the activation function, N_p^r indicates the neighbouring concepts of concept v_p with relation r, $a_{p,r}$ is a normalization constant, $W_r^{(l)}$, $W_0^{(l)}$ are trainable parameters.

Decoder Module. To reconstruct the edges of the graph to recover the triples' missing information, we utilize the DistMult factorization as a scoring function to calculate the score of a given triple (v_p, r, v_q).

$$s(v_p, r, v_q) = \sigma(h_p^T, R_r, h_q) \tag{3}$$

where σ is the logistic function, h_p^T is the transpose vector of concept v_p encoded by RGCN.

Training. We use negative sampling to train our graph autoencoder model (Ghosal et al., 2020). Specifically, for the triples in $\hat{E^{'}}$ (i.e., positive samples), we generate the same amount of negative examples by destroying the concepts or relation of links at random, resulting in the complete sample set Z. Our training goal is to perform binary classification between positive/negative triples with optimization using a cross-entropy loss function.

$$L_{G'} = -\frac{1}{2|\hat{E^{'}}|} \sum_{(v_p, r, v_q, y) \in Z} (y \log s(v_p, r, v_q) + (1 - y) \log(1 - s(v_p, r, v_q))) \tag{4}$$

where y is an indication that is set to 0 for negative triples and 1 for positive triples.

3.3 Stance Detection Training

3.3.1 Commonsense Feature Encoding

After training the graph autoencoder, we utilize it to generate common sense graph features for a specific target t and document x. Specifically, we extract all seed words in the document and denote them as the set K. Then the subgraph $G^{'}_{K}$ is extracted from $G^{'}$, where triples consist of concepts in K or around radius 1 of any concept in K. Next, we feed $G^{'}_{K}$ to the pre-trained RGCN encoder module and make a forward pass to get the feature representations. We calculate the average of the representations h_p for all concepts p of document x as its common sense graph features h_k. Finally, we input h_k to an encoder layer to obtain its hidden representation h_x.

$$h_x = W_x h_k + b_x \tag{5}$$

where W_x and b_x are trainable parameters.

3.3.2 Encoding with Sentiment Information

Considering that the stance of a text is influenced by sentiment information, we learn the sentiment knowledge of the text to increase prediction accuracy. Following Zhou et al. (2020), we exploit a perceptual sentiment language model (SentiBERT) to extract sentiment knowledge. We input the given document x and target t into the pretrained SentiBERT model in the form of "$[CLS]x[SEP]t[SEP]$" to obtain a hidden vector h_s with sentiment information.

$$h_s = SentiBERT([CLS]x[SEP]t[SEP]) \tag{6}$$

Moreover, to take advantage of the contextual information, we also adopt a pretrained BERT [11] model to jointly embed document x and target t to obtain a hidden vector h_b of each example.

$$h_b = BERT([CLS]x[SEP]t[SEP]) \tag{7}$$

Then h_b and h_s are concatenated, and the information of both is fused by the cross-attention module. Cross-attention can effectively capture the interdependencies between text and sentiment, facilitating the integration of knowledge and resulting in the generation of more accurate and meaningful features. The final output h_a is the hidden state of the $[CLS]$ token.

$$h_a = CrossAttention([h_b, h_s])[CLS] \tag{8}$$

3.3.3 Stance Contrastive Learning

Supervised contrastive learning can bring examples of identical categories closer together and push examples of distinct categories apart, thus learning a superior semantic representation space. To improve the generalization of the stance representation, based on the stance label information of the examples, we perform contrastive learning on their hidden vectors h_a (Liang et al., 2022). Specifically, given the hidden vectors $H = \{h_m\}_{m=1}^{N_b}$ of a batch of examples, for a specific anchor $h_m \in H$, if $h_n \in H$ and h_m have the same stance label, i.e., $y_n = y_m$, then h_n is considered to be a positive example of h_m, while other examples $h_o \in H$ are considered to be negative examples. The final contrastive loss is calculated over all positive pairs, including (h_m, h_n) and (h_n, h_m) in a batch:

$$L_c = \frac{1}{N_B} \sum_{h_m \in H} l(h_m) \tag{9}$$

$$l(h_m) = -\log \frac{\sum_{n=1}^{N_b} \mathbf{1}_{[n \neq m]} \mathbf{1}_{[y_m = y_n]} \exp(sim(\mathbf{h}_m, \mathbf{h}_n)/\tau)}{\sum_{o=1}^{N_b} \mathbf{1}_{o \neq m} \exp(sim(\mathbf{h}_m, \mathbf{h}_o)/\tau)} \tag{10}$$

$$sim(\mathbf{s}, \mathbf{t}) = \frac{\mathbf{s}^T \mathbf{t}}{||\mathbf{s}||||\mathbf{t}||} \tag{11}$$

where $1_{[m=n]} \in (0, 1)$ is an indicator function that evaluates to 1 iff $m = n$. $sim\,(\mathbf{s}, \mathbf{t})$ represents the cosine similarity of vectors $\mathbf{s}$ and $\mathbf{t}$. τ denotes a temperature parameter.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

3.3.4 Target Discriminator

The contextual representations generated by Bert and the fused sentiment information contain both target-specific and target-invariant information. Learning and exploiting transferable target knowledge is effective in enhancing the model's generalization to new targets. We separate and differentiate target-specific and target-invariant features by a simple linear transformation, which can decrease the transfer challenge with no removal of stance cues. We first extract target-specific features using a linear transformation layer (Xie et al., 2022):

$$h_g = W_g h_a + b_g \tag{12}$$

where W_g and b_g are trainable parameters. By subtracting target-specific features from h_a, the target-invariant features h_z can be obtained:

$$h_z = h_a - h_g \tag{13}$$

To further make the feature representation h_z target invariant and facilitate automatic adaptation of the model among different targets, we utilize a target discriminator to identify the target that the h_z comes from. If the discriminator cannot accurately predict the target label of h_z, we consider h_z has target-invariance. Our target discriminator is a linear network with softmax, which is trained with a cross-entropy loss function.

$$\hat{y_d} = Softmax(W_d h_z + b_d) \tag{14}$$

$$L_d = \sum_{x \in D_s} CrossEntropy(y_d, \hat{y_d}) \tag{15}$$

where W_d and b_d are the trainable parameters of the target discriminator, $\hat{y_d}$ and y_d are the predicted and true target labels. Specifically, h_z attempts to confound the target discriminator and increase the target classification loss L_d in order to learn the target-invariant features. Meanwhile, the discriminator itself struggles to decrease L_d. So we adopt the gradient reversal layer (GRL) technique, inspired by (Ganin et al., 2016), to achieve this adversarial effect by placing the GRL before the target discriminator. The essence of adversarial training is the minimum-maximum game:

$$\min_{\theta_Z} \max_{\theta_D} - \lambda \log f_D(h_z) \tag{16}$$

where θ_Z are the parameters of all network layers that generate h_z, including fine-tuned Bert, graph encoder, W_g and b_g, etc., θ_D is the discriminator parameters, and f_D is the discriminator function.

3.3.5 Stance Classifier

Since stances are essentially dependent on targets, target-specific information for each target is also indispensable. We concatenate the common sense knowledge graph features h_x, the target-invariant features h_z and the target-specific features h_g to obtain h_c, as the input for the stance classifier with softmax normalization. We minimize the stance classification loss using cross-entropy loss.

$$h_c = h_x \oplus h_z \oplus h_g \tag{17}$$

$$\hat{y} = Softmax(W_c h_c + b_c) \tag{18}$$

$$L_s = \sum_{x \in D_s} CrossEntropy(y, \hat{y}) \tag{19}$$

where W_c and b_c are the trainable parameters of the stance classifier, $\hat{y}$ and y are the predicted stance probability and ground-truth distribution.

The training goal of our proposed model is to minimize the overall loss, defined as follows:

$$L = L_s + \alpha L_c + \beta L_d \tag{20}$$

where α and β are hyperparameters.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

4 Experiments

4.1 Datasets

We conduct experiments on three publicly available datasets. 1)**SEM16** (Mohammad et al., 2016) is a Twitter dataset that contains six targets for stance detection, including the Legalization of Abortion (LA), Feminist Movement (FM), Hillary Clinton (HC), Donald Trump (DT), Atheism (A), and Climate Change is a Real Concern (CC). 2)**WT-WT** (Conforti et al., 2020) is a stance detection dataset in the financial domain. The dataset contains four targets, including ANTM_CI (AC), AET_HUM (AH), CVS_AET(CA), and CI_ESRX (CE). 3)**COVID-19** (Glandt et al., 2021) is a dataset related to COVID-19 health tasks, which includes four targets: Anthony S. Fauci, M.D. (AF), Wearing a Face Mask (WA), Keeping Schools Closed (SC), and Stay at Home (SH). Each text in the three datasets contains a stance (favor, against, neutral) for a specific target.

Following (Liang et al., 2022), we utilize the data from one target as the test set and the remaining targets as the training set. Moreover, we report the F1_avg (the Macro-averaged F1 of against and favor) as evaluation metrics.

Table 2 represents the statistics for the three datasets, listing all targets under each dataset and the number of samples labeled "favor, against, neutral, unlabeled" (where WT-WT and COVID-19 have no unlabeled samples) for each target.

Dataset	Target	Favor	Against	Neutral	Unlabeled
SEM16	DT	148	299	260	2,194
	HC	163	565	256	1,898
	FM	268	511	170	1,951
	LA	167	544	222	1,899
	A	124	464	145	1,900
	CC	135	26	203	1,900
WT-WT	CA	2,469	518	5,520	-
	CE	773	253	947	-
	AC	970	1,969	3,098	-
	AH	1,038	1,106	2,804	-
COVID-19	WA	515	220	172	-
	SC	430	102	85	-
	AF	384	266	307	-
	SH	151	201	396	-

Table 2. Statistics of the SEM16, WT-WT and COVID-19 datasets.

4.2 Experimental Implementation

We employ the pretrained SentiBERT and BERT models as the encoder, whose maximum sequence length is 85. Adam (Kingma and Ba, 2014) is used to optimize the model. In the graph autoencoder training stage, the graph batch size is 10000, the learning rate is 0.01, the dropout rate is 0.25, and we apply gradient clipping to 1.0. In the stance detection training stage, the batch size is 8, the learning rate is 1.5e-5, the dropout rate is 0.1, we train up to 50 epochs, the patience is 5, the temperature parameter for contrastive loss is 0.07. We use different seeds to train our model and record the best results.

4.3 Baselines

We compare the ANEK with several strong baselines, including **BiCond** (Augenstein et al., 2016) bidirectional conditional encoding model, **CrossNet** (Xu et al., 2018): BiCond with topic-specific attention, **TOAD** (Allaway et al., 2021): BiCond with adversarial learning, **BERT** (Kenton and Toutanova, 2019): pretrained language model, **BERT-GCN** (Liu et al., 2021): BERT with GCN for node information aggregation, **TGA Net** (Allaway and Mckeown, 2020): Bert with topic-group attention, **TPDG** (Liang

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

Model	SEM16(%)						WT-WT(%)				COVID-19(%)			
	DT	HC	FM	LA	A	CC	CA	CE	AC	AH	WA	SC	AF	SH
BiCond	30.5	32.7	40.6	34.4	31.0	15.0	56.5	52.5	64.9	63.0	30.1	33.9	26.7	19.3
CrossNet	35.6	38.3	41.7	38.5	39.7	22.8	59.1	54.5	65.1	62.3	38.2	40.0	41.3	40.4
TOAD	49.5	51.2	54.1	46.2	46.1	30.9	55.3	57.7	58.6	61.7	37.9	47.3	40.1	42.0
BERT	40.1	49.6	41.9	44.8	55.2	37.3	56.0	60.5	67.1	67.3	44.3	45.1	47.5	39.7
BERT-GCN	42.3	50.0	44.3	44.2	53.6	35.5	67.8	64.1	70.7	69.2	-	-	-	-
TPDG	47.3	50.9	53.6	46.5	48.7	32.3	66.8	65.6	74.2	73.1	48.4	**51.6**	46.0	37.3
TGA Net	40.7	49.3	46.6	45.2	52.7	36.6	65.7	63.5	69.9	68.7	-	-	-	-
PT-HCL	50.1	54.5	54.6	**50.9**	**56.5**	38.9	**73.1**	69.2	**76.7**	76.3	**58.8**	44.7	41.7	**53.3**
ANEK	**50.3**	**54.7**	**55.0**	49.0	54.1	**39.2**	71.4	**69.8**	74.8	**76.3**	52.9	49.8	**48.6**	50.3

Table 3. Experimental results on three datasets. Bold indicates the best score for each test target.

et al., 2021): GCN-based model for designing target-adaptive pragmatic dependency graphs, **PT-HCL** (Liang et al., 2022): hierarchical contrastive learning model.

4.4 Main Results

We implemented comparison experiments on three datasets and show the F1_avg results (Percentage System) in Table 3. Our proposed ANEK model presents superior performance compared to the baseline models on most target datasets. Specifically, BiCond and CrossNet perform the worst overall, as they do not consider the target invisibility to learn transferable information. Although TOAD also adopts an adversarial strategy to learn target-invariant information, its use of BiLSTM encoding is prone to poor performance in case of an unbalanced target distribution. It can be observed that it performs even less efficiently than Bert on multiple targets. As a strong baseline in NLP, BERT has good generalization because it learns rich semantic information in a large corpus, despite ignoring transferable information between targets. However, when it is applied to target transfer, it causes performance degradation due to its tendency to fit the source data. Our model explores adversarial learning based on pre-trained models, which can learn enhanced target-invariant features and improve the model's transferability.

Table 3 shows that relying solely on the introduction of common sense knowledge to help the model understand is not enough for Bert-GCN, and our model also accounts for learning sentiment information to enhance the discriminative capability of the model. We can find that ANEK slightly outperforms the PT-HCL method with hierarchical contrastive learning. Although PT-HCL obtains excellent generalization by identifying the invariant stance expressions from specific syntactic levels, it requires preprocessing the data to generate pseudo-labels, which increases the complexity of the model. Moreover, the noise brought by pseudo-labels may affect the prediction results. In contrast, our model has stronger generality and interpretability.

4.5 Ablation Study

We further designed several variants of ANEK for ablation experiments to analyze the effects of different components on the model, where "w/o CL", "w/o SK", "w/o CK", "w/o TD" denote the removal of contrastive learning, sentiment information, common sense knowledge and adversarial learning, respectively.

We report the F1_avg scores (Percentage System) of the ablation study in Table 4. The experimental results indicate that removing stance contrastive learning ("w/o CL") significantly decreases the model's performance, which suggests that we perform stance contrastive learning on the text representation assists the encoder in learning better category representations from samples, leading to better generalization. The removal of sentiment information ("w/o SK") reduces model performance, implying that the model may learn the potential relationship between stance and sentiment and make judgments with the help of sentiment knowledge. Removing common sense knowledge ("w/o CK") leads to poor performance in stance detection, indicating that introducing common sense knowledge can indeed help the model

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

understand text information and improve its reasoning ability. "w/o TD" indicates that the removal of the target discriminator becomes less effective on multiple targets, demonstrating the success of adversarial learning applied to zero-shot scenarios, generalizing to unseen targets by encouraging the encoder to generate target-invariant representations.

Model	SEM16(%)						WT-WT(%)				COVID-19(%)			
	DT	HC	FM	LA	A	CC	CA	CE	AC	AH	WA	SC	AF	SH
ANEK	**50.3**	**54.7**	**55.0**	**49.0**	**54.1**	**39.2**	**71.4**	**69.8**	**74.8**	**76.3**	**52.9**	**49.8**	**48.6**	**50.3**
w/o LC	49.2	52.8	52.9	47.8	53.2	38.0	69.2	66.5	73.2	75.2	51.3	48.2	48.1	49.2
w/o SK	48.7	51.8	53.4	47.2	52.0	37.8	68.1	67.5	71.3	74.0	51.0	49.3	47.2	48.0
w/o CK	48.0	52.4	53.0	46.8	51.1	36.5	67.6	66.8	72.0	73.8	49.7	48.7	46.5	47.9
w/o TD	47.8	51.2	52.3	46.5	52.9	37.8	69.0	68.8	72.6	73.3	50.4	47.9	47.8	47.2

Table 4. Experimental results of the ablation study.

4.6 Generalizability Analysis

We further performed experiments on the SEM16 dataset for cross-target stance detection and report the F1_avg results (Percentage System) in Table 5. The cross-target stance detection task is treated as a particular zero-shot setting, as we need to train using data from a source target related to the test target. Table 5 illustrates that our ANEK model achieves better performance. We can also find that the cross-target setting outperforms the zero-shot setting, which indicates that knowing the relationship between targets in advance can learn more reliable target-invariant representations to generalize to unseen targets, illustrating the challenges of zero-shot stance detection. Additionally, enhancing the understanding and generalization of the model by introducing external knowledge is also effective.

Model	SEM16(%)			
	FM→LA	LA→FM	HC→DT	DT→HC
BiCond	45.0	41.6	29.7	35.8
CrossNet	45.4	43.3	43.1	36.2
BERT	47.9	33.9	43.6	36.5
TPDG	58.3	54.1	50.4	52.9
PT-HCL	**59.3**	54.6	53.7	55.3
ANEK	58.5	**54.8**	**54.3**	**56.4**

Table 5. Experimental results of cross-target stance detection. "FM→LA" indicates training on FM, testing on LA, etc.

Text	Target	Gold Label	BERT	TOAD	ANEK
Your have to wonder if Hillary will attempt to replace #ObamaCare with #HillaryCare.	Donald Trump	Against	Neutral	Against	Against
Donald trump is way better than ANY candidate out there. Because he's real, not a lobbyist backed puppet.	Donald Trump	Against	Favor	Favor	Against
I do not understand why the Republicans don't dismiss him.	Donald Trump	Against	Neutral	Neutral	Against
.....and some, I assume, are good people.	Donald Trump	Against	Favor	Favor	Favor

Table 6. Four cases of the predictions by BERT, TOAD and ANEK.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

4.7 Case Study

To qualitatively analyze our model, we conduct a case study and error analysis. We select four cases from the test data of SEM16 and compare our results to the predictions of BERT and TOAD. Table 6 reports these results. In the first case, our model and TOAD with adversarial learning output the correct labels, while the output of BERT is wrong. We believe that because the training data contains the target "Hillary Clinton," the model learns the election relationship between the two targets and transfers the knowledge, and semantically focuses more on the stance-related words rather than the target words, with a robust target generalization. In the second case, only our method makes the correct prediction, demonstrating that depending only on contextual information is insufficient. Adding sentiment information strengthens the model's comprehension of texts with a sarcastic sentiment. In the third case, our method still correctly predicts the outcome. Although no words about Trump appear in the text, we speculate that the model learns the hidden connection between "Republican" and "Donald Trump" and understands the implied meaning of the text, further confirming the validity of common sense knowledge.

In the fourth case, all models output incorrect results. We suspect that this is because the text is too brief, resulting in less valid information being learned, and the background knowledge is too complex, which reveals that we can explore data augmentation methods in the future to improve the performance of zero-shot stance detection by expanding the data.

5 Conclusion

This paper proposes an adversarial network with external knowledge (ANEK) to handle the zero-shot stance detection task. The model applies adversarial learning based on pre-trained models to ensure knowledge transferability, and introduces common sense knowledge and sentiment information to enhance the model's deep understanding and assist stance detection. In addition, stance contrastive learning is used to improve the model's generalization. The experimental results on three benchmark datasets indicate that our method performs competitively on some unseen targets. In future work, we will design a data enhancement method to alleviate the data scarcity problem in zero-shot settings and improve performance.

Acknowledgements

This work is supported by a grant from the Social and Science Foundation of Liaoning Province (No. L20BTQ008)

References

Emily Allaway and Kathleen Mckeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Emily Allaway, Malavika Srikanth, and Kathleen Mckeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. Kingdom: Knowledge-guided domain adaptation for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks (2014). *arXiv preprint arXiv:1406.2661*, 1406.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-shot cross-lingual stance detection with sentiment-based pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10729–10737.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Sumeet Kumar and Kathleen M Carley. 2019. Tree lstms with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5047–5058.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6299–6305.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the Web Conference 2021*, pages 3453–3464.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2738–2747.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176.

Zhen Wang, Qiansheng Wang, Chengguo Lv, Xue Cao, and Guohong Fu. 2020. Unseen target stance detection with adversarial domain generalization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176.

Feng Xie, Zhong Zhang, Xuechen Zhao, Jiaying Zou, Bin Zhou, and Yusong Tan. 2022. Adversarial learning-based stance classifier for covid-19-related health policies. *arXiv e-prints*, pages arXiv–2209.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics*, pages 568–579.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2070–2075.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 824-835, Harbin, China, August 3 - 5, 2023.

The Contextualized Representation of Collocation

Daohuan Liu
Huazhong University of
Science and Technology
liudh@hust.edu.cn

Xuri Tang
Huazhong University of
Science and Technology
xrtang@hust.edu.cn

Abstract

Collocate list and collocation network are two widely used representation methods of collocations, but they have significant weaknesses in representing contextual information. To solve this problem, we propose a new representation method, namely the contextualized representation of collocate (CRC), which highlights the importance of the position of the collocates and pins a collocate as the interaction of two dimensions: association strength and co-occurrence position. With a full image of all the collocates surrounding the node word, CRC carries the contextual information and makes the representation more informative and intuitive. Through three case studies, i.e., synonym distinction, image analysis, and efficiency in lexical use, we demonstrate the advantages of CRC in practical applications. CRC is also a new quantitative tool to measure lexical usage pattern similarities for corpus-based research. It can provide a new representation framework for language researchers and learners.

1 Introduction

Collocation is an important concept in the fields of linguistics and computational linguistics (Firth, 1957; Halliday and Hasan, 1976; Sinclair, 1991), which can be widely used in language teaching, discourse analysis and other fields. Currently, there are two widely used representation methods of collocation, namely collocate list and collocation network. However, they are both flawed.

Collocate list takes the list of collocate words as the main form and generally provides the correlation strength, co-occurrence frequency, etc. between the node word and the collocate word[1]. Sometimes a collocate list may also include information such as the total frequency of collocates, the frequency of appearing on the left and right sides, etc. Table 1 shows the collocation list of the node word *importance* in a small news corpus.

Collocate	PMI	Co-occur Frequency
attach	11.216	29
underscore	9.223	2
emphasize	8.821	4
stress	8.811	19
aware	8.528	3
awareness	8.386	2
great	7.555	28

Table 1: Sample Collocate List of *importance* as a node. Pointwise Mutual Information is adopted to measure the association strength between two words (`measure=PMI`); Only collocates with 2 or more co-occurrences are considered (`min_freq=2`); Only collocates with an association strength greater than 7.5 are displayed (`thresh=7.5`).

[1]We follow the names by Sinclair (1991), and call the focal word in the collocation a "node word" (Node), and call the word appearing in the other position in the collocation a "collocate word" (Collocate).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

The expression capability is very limited through collocate lists, as they could neither present the interaction between collocates nor be visually friendly to readers. However, connectivity is an important feature of collocation knowledge (Phillips, 1985). In order to improve these weaknesses, Brezina et al. (2015) implemented the representation method of collocation graph and network[2] (see Figure 1). In a collocation graph, the collocates are scattered around and connected to the central word (node). The closer a collocate is linked to the node, the stronger it is associated with it. Compared to collocate list, collocation graph improves the visualization and enables the interaction of multiple collocations through node connection and graph extension. Brezina (2018) also demonstrates the possible applications of collocation networks with cases including discourse analysis, language learning, and conceptual metaphor research.

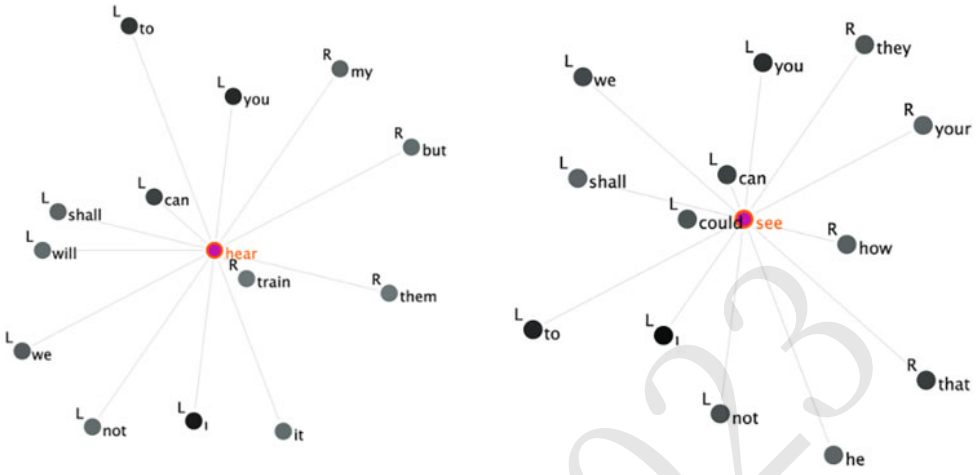


Figure 1: Sample Collocation Graphs of *hear* and *see* for image comparison, based on the corpora of World War I poems (Taner and Hakan, 2021).

However, these two traditional representation methods both have many critical flaws. The fundamental problem is that they neglect the natural language as a kind of sequence data. Collocate list regards collocation as a simple juxtaposition of tokens, and collocation graph regards each word as a free discrete data point in the space. Nevertheless, the context information is not only related to the semantics of the collocates surrounding a node but also related to the order and the position of the words.

First of all, they only tell the semantic relations but ignore the syntactic relations between nodes and collocates. The semantic association between nods and collocates is direct and clear. Firth's (1957) defines collocation as a container for semantic associations between the two words; and the meaning of a word comes not only from itself, but also from other words that co-occur with it. The association scores shown in both collocate list and collocation graph are an evaluation of co-occurrence, in other sense, a reflection of the semantic relationship. Nevertheless, Tang (2018) addresses that the syntactic association between the node and the collocate is also an essential part, acknowledging collocation as a complex of syntactic and semantic knowledge. He also sorts out the key role of syntactic relations in semantic theories and their applications through related studies (Katz and Fodor, 1963; Petruck, 1996). The syntactic nature of collocations is mainly reflected in the fact that collocations have direction and span. For instance, the two semantically related words *student* and *diligent* generally do not appear as "student diligent" in actual language use, which is syntactically incorrect in most cases; while "diligent student" or "student is diligent" is much more common and intuitively correct application. This shows that the collocation knowledge is actually an overall model that is restricted both by semantic relations and grammatical relations.

Secondly, they fail to reflect the relative position (relative distance) between nodes and collocates. Qu (2008) pointed out that the two components in a collocation tend to have fixed positions, i.e., one word

[2]A collocation network is a connected network of multiple collocation graphs. The two terms "collocation network" and "collocation graph" are used interchangeably in this paper, referring to the same representation method.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

always appears on the left or right side of the other word. For example, among the collocates in Table 1, *attach* mostly appears on the left side of the node *importance*. In the case that the positions of these two words are reversed, the syntactic relationship between the two words should also change. According to the data samples in the corpus, *importance* mostly acts as the object in *attach-importance* collocations, while *importance* often serves as the subject in *importance-attach* collocations.

Finally, these two representations cannot reflect the freedom of choice of collocates, which would restrict the usage of collocation in practice. This is not conducive to group collocates and find patterns with respect to semantic or syntactic relations. For example, if we want to find other predicates to substitute *attach* for the node *importance*, it is hard to tell from a raw collocate list or collocation graph. In order to satisfy this requirement, an extra screening operation such as Part-of-Speech (POS) tagging or syntactic analysis is required.

To overcome the above-mentioned shortcomings, we propose a new representation and visualization method to describe collocation, called Contextual Representation of Collocations (CRC), which makes improvements in syntactic representation and visualization abilities. Key features of CRC include:

- Applying conflated linear representation with respect to the nature of language;
- Foregrounding positional information of the collocates;
- Using spatial and visual symbols to indicate strength.

We will include three case studies of CRC respectively applied to synonym distinction, image analysis, and language teaching. These practical applications should reveal the advantages of this approach. In terms of knowledge representation, it can present more detailed grammatical information; in terms of knowledge application, it can achieve an accurate comparison of collocation distributions. CRC can provide a new representation framework for language researchers and language learners, as well as facilitate language research and teaching.

2 Contextualized Representation of Collocation (CRC)

While retaining the dimension of association strength, CRC promotes the relative position of collocates to another major feature dimension, so that each collocate can present those two important attributes at the same time. Therefore, the essence of the CRC is a two-dimensional scatter plot.

Figure 2 is an instance of CRC, which is based on the same data source as Table 1. The key features of this visualization will be explained in detail.

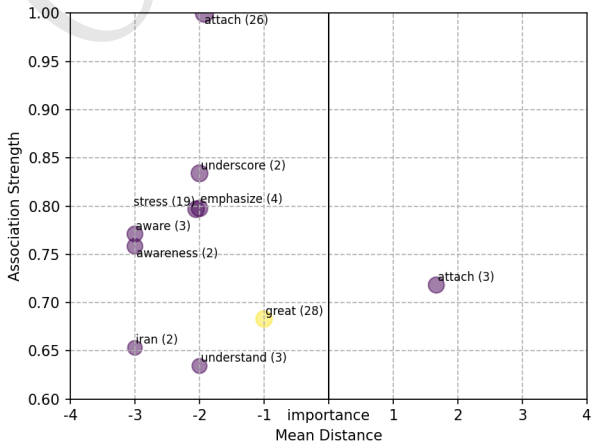


Figure 2: CRC of node *importance* (`measure=PMI`, `min_freq=2`, `thresh=0.6`).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

2.1 Conflated Linear Representation

Compared with the network structure of collocation graph, CRC follows the linear characteristics of natural language and uses conflated linear representation in expressing the relations of all the collocate-node pairs. In Figure 2, collocation relations are described as parallel horizontal lines, and compressed in a certain space range. This parallel and linear presentation helps to visually compare the commonalities and differences of different collocations, which is the basis of all the other advantages of CRC.

In our CRC implementation, we arrange the longitudinal spatial distribution of all collocate-node pairs according to their association score[3]. For the convenience of drawing and reading, the scores are normalized to [0,1]. The closer to 1, the higher the collocation strength.

2.2 Positional Information

In Figure 2, the horizontal dimension represents grammatical relations in terms of direction and distance. In this instance, the distance of each collocate is the average of the distances of all the collocate-node pair occurrences. `Distance=0` represents the position of the node word (*importance*).

Foregrounding positional information is the core contribution of CRC, as well as the key feature that distinguishes CRC from the other two representation methods. It is easy to understand that if the positional information is removed from Figure 2, all the collocate points will appear on the same vertical line, hence, it will degenerate into a simple visualization of the collocate list. Instead, if the Cartesian coordinate system is transformed into a polar coordinate system, it then becomes a collocation graph with fixed node positions.

Positional Information reflects the order of words, and the order of words further reflects the syntactic relationship. This can benefit CRC users with plenty of straightforward linguistic knowledge. Taking Figure 2 as an example, we could observe at least the following facts:

- The word *attach* tends to appear on the left of *importance* (`frequency=26`, `strength=1.00`) rather than on the right (`frequency=3`, `strength=0.72`), which might imply the *attach-importance* collocation is more likely to be used in active voice instead of passive voice.
- The word *importance* has a strong right-leaning tendency (Wang et al., 2007), which means it expects to be modified by a modifier prior to it.
- Collocates like *attach*, *underscore*, *emphasize*, *stress*, and *understand* might all play similar grammatical roles in the relationship with the node importance, because they all appear in the -2 position.
- People tend not to say "attach importance" but to use "attach great importance", which can be reckoned from their positions (*attach*=-2, *great*=-1, *importance*=0). This shows that CRC could also be used to recognize continuous word clusters and phrase patterns, making CRC more prospective in the application of analyzing and teaching. And it is capturing common contexts that the node is often used in. And this is also the reason why this representation of collocation is termed "contextualized".

2.3 Visualization Strategy

In addition to its advantage in context modelling of the node, as a representation method, CRC could be easily visualized with many visualization strategies. It can combine many spatial methods and visual symbols to expand the expression and presentation of collocation knowledge. As mentioned before, in addition to implementing the CRC in the plane Cartesian coordinate system, it can also be realized in the polar coordinate system; the size, color, and grayscale (transparency) of the data points in the figure can all be useful tools to group or describe collocates.

In general, compared with existing collocation representation methods, i.e., collocate list and collocate network, CRC can intuitively present richer context information and provide more convenience for researchers who use collocation analysis. In the following sections, we will apply CRC in three case studies of recent years to demonstrate the superiority of the new representation method.

[3]We use various measuring algorithms to calculate association scores and pick the intuitively best one from all the results. The adopted measuring method for each case is described in the captions of the figures.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

3 Case Study 1: CRC in Synonym Distinction

Many researchers utilize collocation analysis to distinguish synonyms. Liu has studied the usage differences of many synonym sets in English with the COCA corpus, such as *Actually*, *Genuinely*, *Really*, *Truly* (2012) and *Chief*, *Main*, *Major*, *Primary*, *Principal* (2010) using a behavioral profile approach. He also analyzed the learners' misuse of three synonym groups of *circumstance*, *demand*, and *significant* by comparing the use of these words in a second-language learner corpus (2018). Xiong and Liu (2022) compared the usage patterns and semantic differences of the two synonyms, *absolutely* and *utterly*, with the help of collocation lists and the Key-Word-In-Context (KWIC) function provided by the COCA corpus. Their conclusions are largely based on random sampling and qualitative analysis. Obviously, the above research methods take a large manual workload in observation and statistics. The use of CRC can not only carry out descriptive conclusions easily and clearly but also provide quantitative measures to differentiate synonyms.

In this case study, we use CRC to restudy the differences among the synonyms *Actually*, *Really*, *Truly*, and *Genuinely*, and try to verify some findings reported by Liu and Espino (2012). We choose a smaller corpus, BROWN, instead of COCA as the data source of this case study for its availability. The frequencies of each adverb in the BROWN corpus are shown in Table 2. It can be found that these four adverbs have a similar frequency proportion although the total data amount of BROWN is much smaller than COCA. Since *Genuinely* is not found in BROWN, we only study the other three adverbs.

Corpus	#actually	#really	#truly	#genuinely
COCA	105,039	263,087	20,504	3,065
BROWN	166	275	57	0

Table 2: Frequencies of *Actually*, *Really*, *Truly*, and *Genuinely* respectively in COCA and BROWN.

We retrieve the free collocates of *actually*, *really*, and *truly* from the corpus and generate three CRC instances (Figure 3a, Figure 3b and Figure 3c). Based on the frequency ratio of these three words, we set the thresholds of association strength to 0.3, 0.5 and 0.1 respectively for better visualization effects.

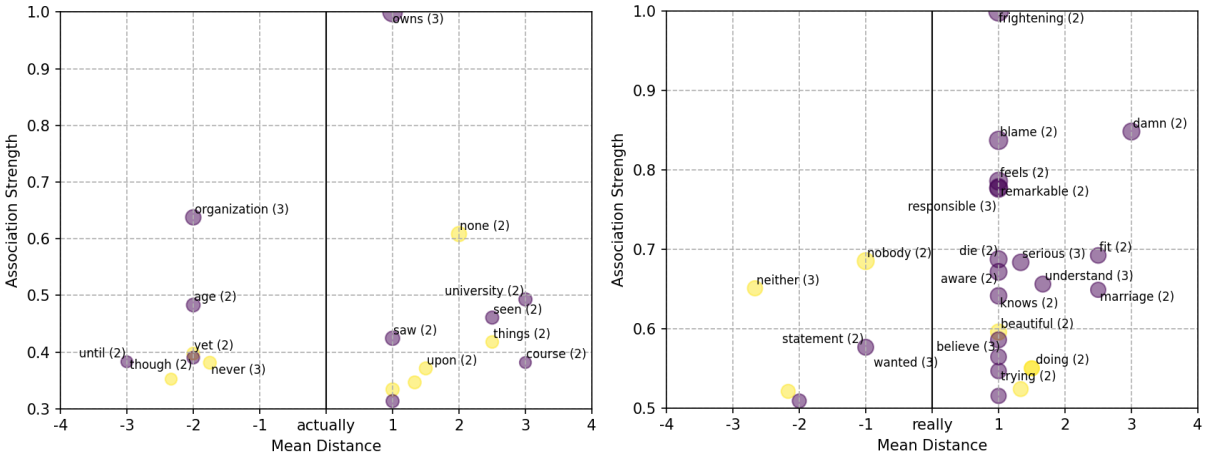
The usage pattern of *really* (Figure 3b) is significantly different from the other two words. There are much more highly associated collocates to the right of *really* (especially at `position=1`) than *actually* and *truly*, suggesting that *really* is more often used as verb and adjective modifiers. From Figure 3a, it is hard to find clear usage patterns of *actually* from the distribution of collocates on both sides. However, when compared to *truly* (Figure 3c), it could be observed that *actually* is surrounded by more content words and has many adversative words on the left side (e.g. *never*, *yet*, *though*, *until*), indicating that it may be used more as a disjunct. Figure 3c presents few meaningful collocates, but we can also reckon that *truly* is prone to occur as an adjective modifier from the limited samples. Moreover, from its potential context (*will/be* + *truly* + *fine/great*, etc.) we can infer that it is prone to be used for attitude emphasis and enhancement.

The above inferences are entirely based on CRC figures and are basically consistent with the main findings of Liu and Espino (2012) (see Figure 4) who adopted rigorous and systematic statistical methods (Hierarchical configural frequency analysis, HCFA). This demonstrates that CRC is a fast and handy tool in certain lexical studies. Of course, it would also be encouraged that researchers use other corpus approaches such as concordance and HCFA as auxiliary verification methods.

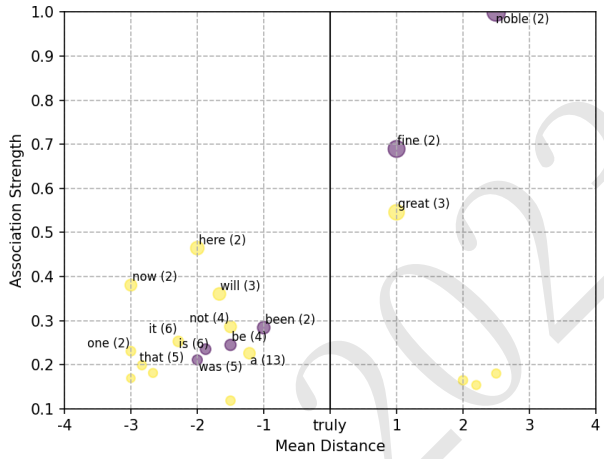
The network structure in Figure 4 is artificially constructed through subjective analysis. Yet CRC can make this relationship network more objective and accurate by quantifying the differences between the usage patterns of these words.

CRC preserves precise location and strength information of the surrounding words, thereby it is a collocate distribution of the node. This allows us to compute the degree of difference between distributions, namely the distance between CRCs. The smaller the distance is, the more similar the usage patterns of the two words. To compute the distance between two distributions, we are free to apply any suitable distance algorithm. Here we use a simple processing flow:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.



(a) *actually*. (b) *really*.



(c) *truly*.

Figure 3: CRCs of node *actually*, *really* and *truly* (`measure=PMI`, `min_freq=2`, `thresh` differs to accommodate to the number of data points for a better visualization).

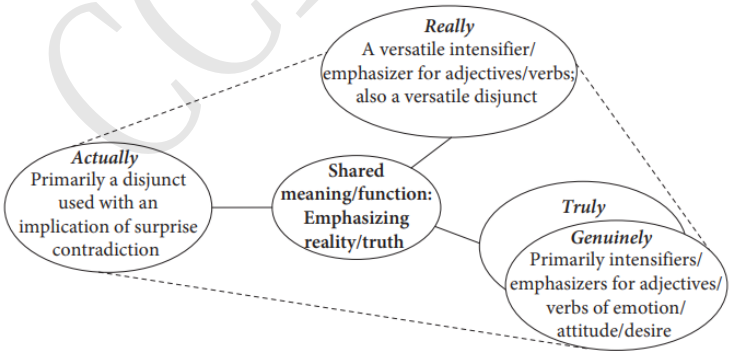


Figure 4: The internal semantic structure of the four synonymous adverbs by Liu and Espino (2012).

(1) For a common collocation word, calculate the Euclidean distance between the coordinates of the word on the two graphs.

(2) For unique collocations, they are not included in the distance calculation.

(3) Finally, average the Euclidean distance between all points to obtain the comprehensive distance.

Using the above algorithm, we obtained three distances (Figure 5). Our results show slight divergence

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

from Figure 4, as Liu and Espino considers *really* more similar to *truly* but CRC distance indicates that *really* is closer to *actually* (D(*really*, *actually*)=0.1641) rather than *truly* (D(*really*, *truly*)=0.1853). This difference might be due to BROWN's insufficient amount of data. It is also interesting to see CRC applied to COCA and verify if the semantic structure in Figure 4 is accurate.

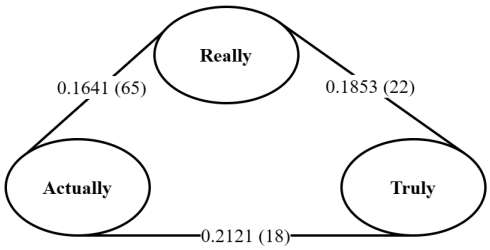


Figure 5: The collocation distribution distance of *actually*, *really*, and *truly* in the BROWN corpus. The number of valid collocation words actually involved in the calculation is indicated in brackets (`measure=PMI`, `min_freq=2`).

4 Case Study 2: CRC in Image Analysis

As an analytical tool, collocation is also widely used in other fields besides linguistic research. In the area of journalism and information communication, collocations are also used in assistance to discourse analysis, image analysis, and sentiment analysis (Koteyko et al., 2013). For example, Pan and Hei (2017) inspected the verb collocates on the right side of *we* in the interpreting corpus of press conferences, so as to analyze the image construction strategy of the government. In the field of digital humanities, collocation could also facilitate the style analysis of writing and author (Vickers, 2012; Wijitsopon, 2013; Taner and Hakan, 2021), as well as the image analysis of characters in the literary works. We select an image analysis task in a literary work and use CRC as an analyzing tool for research and discussion.

We pick the classic fiction "Lord of the Flies" (Golding, 1954) as our research subject, because the characters in this novel have distinctive characteristics and the language is simple and straightforward. Its characters and images have been heatedly discussed through book reviews and literary interpretations (Oldsey and Weintraub, 1963; Spitz, 1970), yet CRC may reveal the character-building methods from a corpus perspective.

This fiction narrates the story of a group of teenagers surviving on a desert island. The cooperation and confrontation among those children are interpreted as the epitome of human antagonism and political game. We select three main characters: Ralph, Jack and Simon as research objectives. Their right-side verb collocates are retrieved and described through CRC (Figure 6a, Figure 6b and Figure 6c). Before extracting collocations, the text is lower-cased and POS tagged. Different thresholds of association strength are applied in order to show a similar amount of collocates.

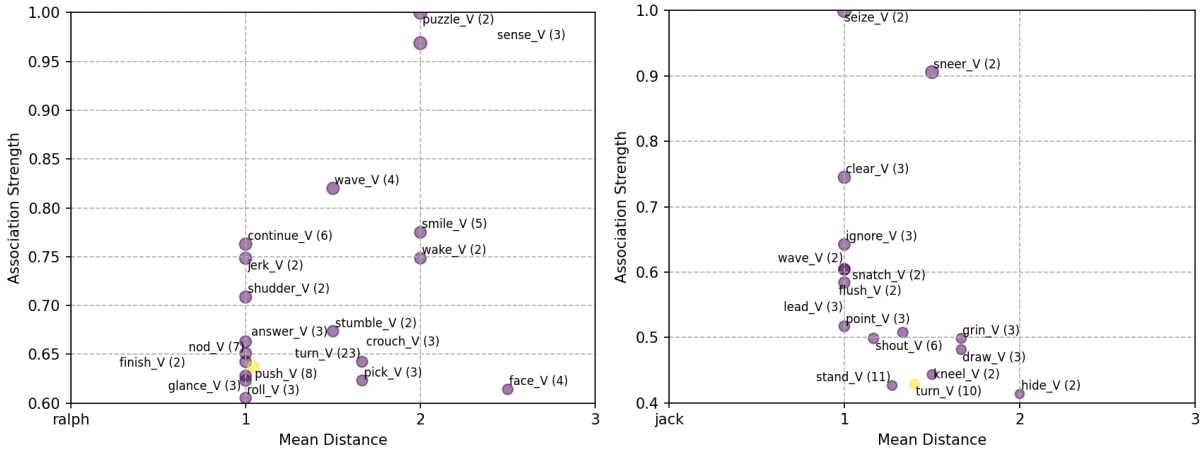
By comparing the three CRC figures, it can be found that when describing different characters, different verbs are used to shape the image of the characters. From the sentiment of verbs used to describe the character, we can infer the author's tendency in the image-building of each character.

Ralph is the main positive leader in the fiction, representing civilization and democracy before and after World War II. Figure 6a shows many clues in favor of that description. Verbs such as *puzzle*, *sense*, and *shudder* place Ralph on the opposite side of chaos and violence; others like *answer* and *nod* depict Ralph actively affirming and responding to others' opinions. These behaviors together shape Ralph as a "democratic man, the symbol of consent" (Spitz, 1970).

Jack is the representative of brutality and power. His unique behaviors include *seize*, *snatch*, *clear*, and *ignore* (Figure 6b), which show Jack's tendency to command and enforce, in consistent with the evaluation by Spitz (1970): "Jack then, is authoritarian man ... like Hitler and Mussolini".

Simon is regarded as "the Christ-figure, the voice of revelation" (Spitz, 1970). From his unique behaviors *lower*, *walk*, *speak*, *feel*, etc. (Figure 6c), readers envisage a sanctified image with calm,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.



(a) *ralph*. (b) *jack*.

(c) *simon*.

Figure 6: CRCs of node *ralph*, *jack* and *simon* (`measure=PMI`, `min_freq=2`, `thresh` differs to accommodate to the number of data points for a better visualization).

humility, detachment, and transcendence.

Apart from unique behaviors, similar behaviors are also described with verbs with different semantic polarities. For example, the author uses *smile* for Ralph but *sneer* and *grin* for Jack to express laugh; this further consolidates the contrasting images of the two characters.

5 Case Study 3: CRC and Efficiency in Lexical Use

Collocations could also play a role in second language acquisition and language teaching. The mastery of collocation is considered to be the decisive factor for the naturalness of a language learner's expression (Oktavianti and Sarage, 2021). One of the cases of collocation network illustrated by Brezina (2018) is the analysis and evaluation of different-level second language learners' expression. The selected corpus is Trinity Lancaster Corpus (TLC) of spoken L2 English (Gablasova et al., 2017), a transcribed corpus of English interview responses. Brezina divided the corpus into three sub-groups according to the speakers' language proficiency levels: Pre-intermediate (B1), Intermediate (B2) and Advanced and Proficiency (C1/C2). The most common collocates of the three verbs *make*, *take* and *do* used by students at these levels are shown with collocation graphs. Figure 7 shows the situation of *make*, from which we can observe a rise in the richness of collocates as the proficiency level lifts.

However, Brezina pointed out that no such "clear relationship between increasing proficiency and a higher number of collocates" was found on *take* and *do* (p.73). This implies that the increase in

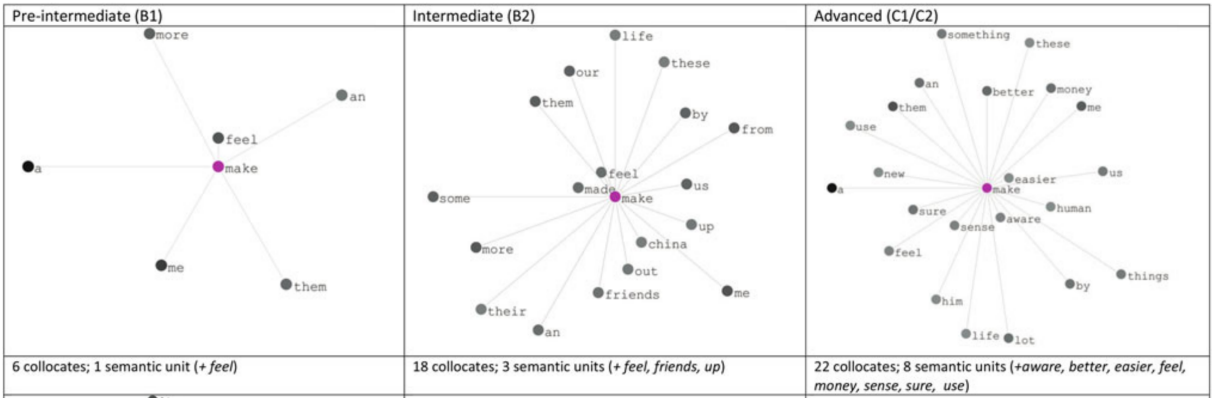


Figure 7: Collocation graphs of *make* for band B1, B2, and C1/C2 (Brezina, 2018).

collocate richness is not the decisive factor in measuring one's language proficiency. When talking about the language learner's communicative language competencies, *Common European framework of reference for languages: Learning, teaching, assessment* (CEFR) points out that a competitive language learner should not only "has a good command of a very broad lexical repertoire" (vocabulary range) but also masters "idiomatic expressions and colloquialisms" and use words "correct and appropriate" (vocabulary control) (2001)(pp.112,114). In other words, the collocation network alone cannot reveal the relationship between the group's collocation performance and their language competency, because collocation network cannot tell the above aspects, i.e., the naturalness and accuracy of the collocations.

While CRC with its quantitative ability (as used in Chapter 3) is a solution to the above problem. To examine whether the speakers' language competency truly matches their labeled level, we select the native speaker sub-group (NS) as a reference corpus, and respectively compute the CRC distances between NS with B1, B2, and C1/C2, so as to evaluate the usage patterns between different levels. Intuitively, we may expect the gap becomes smaller from B1 to C1/C2, because native speakers usually produce the most natural expressions.

The CRC distances of B1-NS, B2-NS and C1/C2-NS are shown in Table 3, including those of *take* and *do*. It can be seen that the speakers' usage pattern of *make* is approaching that of native speakers from B1 to C1/C2. However, statistics on the other two verbs do not present a similar trend; *do* even displays a totally opposite attitude, showing an increasing discrepancy from native speakers as "language level" rises. A possible explanation might be the insufficient data samples. For instance, only 30 common collocates are used to calculate the CRC distance between C1/C2 and NS, most of which are trivial words with low association strength such as *not*, *what*, and *want*.

Distance	B1-NS	B2-NS	C1/C2-NS
make	0.19 (48)	0.1453 (101)	0.1223 (93)
take	0.2349 (32)	0.1743 (49)	0.2077 (45)
do	0.0915 (116)	0.1371 (66)	0.1436 (30)

Table 3: The collocation distribution distance of *make*, *take* and *do* in the three bands and NS corpus. The number of valid collocation words actually involved in the calculation is indicated in brackets (`measure=Log-Likelihood`, `min_freq=2`).

Nevertheless, the findings are basically in line with that of Brezina (2018) but in a more comprehensive and more precise manner. Besides, the distances on word pairs disclose the most misused collocates of the node, which might be helpful in language evaluation and grammar correction. To be specific, CRC could tell which collocates are most distantly distributed in the usage pattern of language learners and of native speakers, so as to improve the learners' worst-acquired collocation knowledge.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

6 Conclusions and Future Work

This paper re-examines two widely used representation methods of collocation, i.e., collocate list and collocation network. In view of their weakness in expressing contextual information, we propose a new representation method, namely the contextualized representation of collocation (CRC). CRC adopts conflated linear representation and highlights the importance of the position of the collocates. It pins a collocate as the interaction of two dimensions, i.e., association strength and co-occurrence position. With a full image of all the collocates surrounding the node word, CRC carries the contextual information and makes the representation much more informative and intuitive. We did three case studies to demonstrate the advantages of CRC in practical applications, covering synonym distinction, image analysis, and efficiency in lexical use. Besides, CRC provides a new quantitative tool to measure lexical usage pattern similarities for corpus-based research.

We believe that the potential power of CRC is far beyond the cases we have discussed. As an auxiliary corpus tool, it may also be used directly in teaching activities. The importance of corpus tools in language teaching is investigated by Boldarine and Rosa (2018), who used some of the searching functions provided by COCA, mainly the collocation tool, to improve students' phrasal integrity. Their survey shows that most students are happy to use corpus tools to test their language intuition, though the aids are less effective for students with poor performance. CRC can be used as a good visualized presentation tool for phrase, collocation and idiom studying, and should intuitively be more friendly to "weaker students" because it is much more easy-reading and more informative compared with collocate list and collocation network.

In addition to involving CRC in teaching, we can also extend or adapt its visualization to fit more needs and scenarios. For example, CRC is also suitable for visualizing constructions (Fillmore, 1988) and collostructions (Stefanowitsch and Rosa, 2003) because it follows the sequential nature of the language. For instance, a CRC-like visualization of the construction "It be ADJ that" could set the slot *ADJ* to position 0, and respectively fix *it*, *be*, *that* at the positions -2, -1, and 1. This requires the support of construction searching algorithms.

In summary, we hope that CRC can provide a new representation framework for language researchers and learners, and will lead them to address the importance of contextual information in research and learning. More applications of CRC in teaching and research are worthy of further empirical study in the future.

References

Brezina, V., McEnery, T., and Wattam, S 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2):139–173.

Brezina, V. 2018. Collocation graphs and networks: Selected applications. In *Lexical collocation analysis* (pp. 59-83). Springer, Cham.

Boldarine, A. C. and Rosa, R. G. 2018. Prepping a prep course: a corpus linguistics approach. *BELT-Brazilian English Language Teaching Journal*, 9(2), 379-394.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*, Cambridge University Press.

Fillmore, C. J. 1988. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 14, pp. 35-55).

Firth, John R. 1957. *Modes of meaning. In: Papers in Linguisticss, 1934-1951*. Oxford: Oxford University Press.

Gablasova, D., Brezina, V., Mcenery, T. and Boyd, E. 2017. Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, 38(5), 613–637.

Golding, W. 1954. *Lord of the flies*. Faber & Faber.

Halliday, M. A. K. and Hasan. R 1976. *Cohesion in English*. London: Longman Group Ltd.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

Katz, J. J. and Fodor, J. A. 1963. The Structure of a Semantic Theory. *Language*, 39(2), 170–210. https://doi.org/10.2307/411200.

Koteyko, N., Jaspal, R. and Nerlich, B. 2013. Climate change and 'climategate' in online reader comments: A mixed methods study. *The geographical journal*, 179(1), 74-86.

Liu, D. 2010. Is it a chief, main, major, primary, or principal concern?: A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics*, 15(1), 56-87.

Liu, D. and Espino, M. 2012. Actually, Genuinely, Really, and Truly: A corpus-based Behavioral Profile study of near-synonymous adverbs. *International Journal of Corpus Linguistics*, 17(2), 198-228.

Liu, D. 2018. A corpus study of Chinese EFL learners' use of circumstance, demand, and significant: An in-depth analysis of L2 vocabulary use and its implications. *Journal of Second Language Studies*, 1(2), 309-332.

Oktavianti, I. N. and Sarage, J. 2021. Collocates of 'great' and 'good' in the Corpus of Contemporary American English and Indonesian EFL textbooks. *Studies in English Language and Education*, 8(2), 457-478.

Oldsey, B. and Weintraub, S. 1963. Lord of the Flies: Beezlebub Revisited. *College English*, 25(2), 90–99. https://doi.org/10.2307/373397.

Pan, F. and Hei, Y. 2017. Government Image-building in Chinese-English Press Interpretation: A Case Study of the collocation of Personal Pronoun "we". *Foreign Languages and Their Teaching* (5), 8.

Petruck, M. R. 1996. Frame semantics. *Handbook of pragmatics*, 2.

Phillips, M. 1985. *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam, Netherlands: North-Holland.

Qu, W. 2008. *Research on Automatic Word Disambiguation of Modern Chinese*. Beijing: Science Press.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Spitz, D. 1970. Power and Authority: An Interpretation of Golding's "Lord of the Flies." *The Antioch Review*, 30(1), 21–33. https://doi.org/10.2307/4637248.

Stefanowitsch, A. and Gries, S. T. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2), 209-243.

Tang, X. 2018. *Collocation and Predicate Semantics Computation*. Wuhan University Press.

Taner C. and Hakan C. 2021. A warring style: A corpus stylistic analysis of the First World War poetry. *Digital Scholarship in the Humanities*, fqab047, https://doi.org/10.1093/llc/fqab047.

Vickers, B. 2012. Identifying Shakespeare's additions to The Spanish Tragedy (1602): A new (er) approach. *Shakespeare*, 8(1), 13-43.

Wang, D., Zhang, D., Tu, X., Zheng, X. and Tong, Z. 2007. Collocation Extraction Based on Relative Conditional Entropy. *Journal of Beijing University of Posts and Telecommunications*, 30(6), 40.

Wijitsopon, R. 2013. A corpus-based study of the style in Jane Austen's novels. *Manusya: Journal of Humanities*, 16(1), 41-64.

Xiong, Y. H. and Liu, D. F. 2022. A Corpus-based Analysis of English Near-synonymous Adverbs: Absolutely, Utterly. *Journal of Literature and Art Studies*, 12(4), 359-365.

Training NLI Models Through Universal Adversarial Attack

Jieyu Lin, Wei Liu, Jiajie Zou, Nai Ding *
Key Laboratory for Biomedical Engineering of Ministry of Education,
College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, China
{jieyu_lin,liuweizju,jiajiezou,ding_nai}@zju.edu.cn

Abstract

Pre-trained language models are sensitive to adversarial attacks, and recent works have demonstrated universal adversarial attacks that can apply input-agnostic perturbations to mislead models. Here, we demonstrate that universal adversarial attacks can also be used to harden NLP models. Based on NLI task, we propose a simple universal adversarial attack that can mislead models to produce the same output for all premises by replacing the original hypothesis with an irrelevant string of words. To defend against this attack, we propose Training with UNiversal Adversarial Samples (TUNAS), which iteratively generates universal adversarial samples and utilizes them for fine-tuning. The method is tested on two datasets, i.e., MNLI and SNLI. It is demonstrated that, TUNAS can reduce the mean success rate of the universal adversarial attack from above 79% to below 5%, while maintaining similar performance on the original datasets. Furthermore, TUNAS models are also more robust to the attack targeting at individual samples: When search for hypotheses that are best entailed by a premise, the hypotheses found by TUNAS models are more compatible with the premise than those found by baseline models. In sum, we use universal adversarial attack to yield more robust models.

1 Introduction

Pre-trained models have achieved impressive performance among natural language processing (NLP) tasks, including natural language inference (NLI) and machine reading comprehension (MRC) (Liu et al., 2019; He et al., 2020). Nevertheless, these models are vulnerable under adversarial attacks (Behjati et al., 2019). For most adversarial attack methods, the adversarial samples are input-specific, i.e., the adversarial perturbation is targeted at a specific input. More recently, however, studies have also shown the existence of universal adversarial attacks, which are input-agnostic (Wallace et al., 2019; Behjati et al., 2019). Multiple methods have been proposed to find universal adversarial samples. One method is to append an input-agnostic string of words to any input to convert the input into an adversarial sample. For example, Wallace et al. (2019) use gradient-based search to find strings that, when concatenated to any input, could result in specific model output. For instance, for models trained on SNLI, prepending "nobody" to the hypothesis could cause >99% of the samples to be judged as being contradictory to the premise, even when all the tested hypotheses are in fact entailed by the premises. Another method is to randomly sample a large number of sentences and screen for universal adversarial samples. For example, Lin et al. (2021) use such a method to find sentences that a model always judges as the correct answer to multiple-choice MRC questions.

The mainstream method to increase the robustness of models against adversarial attacks is adversarial training (Madry et al., 2018; Goodfellow et al., 2015; Zhang et al., 2019). In this process, adversarial samples are generated and injected into the training batch. Adversarial training generally focuses on input-specific attacks, which involve small perturbations and targeting at individual samples. Therefore, models fine-tuned with these methods still fail in universal adversarial attacks (Shafahi et al., 2020). Besides, unlike input-specific attacks, universal attacks use single perturbation to cause the model fail in lots of samples, making it more effective to generate adversarial samples. Recently, in the domain

*Corresponding author: Nai Ding

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

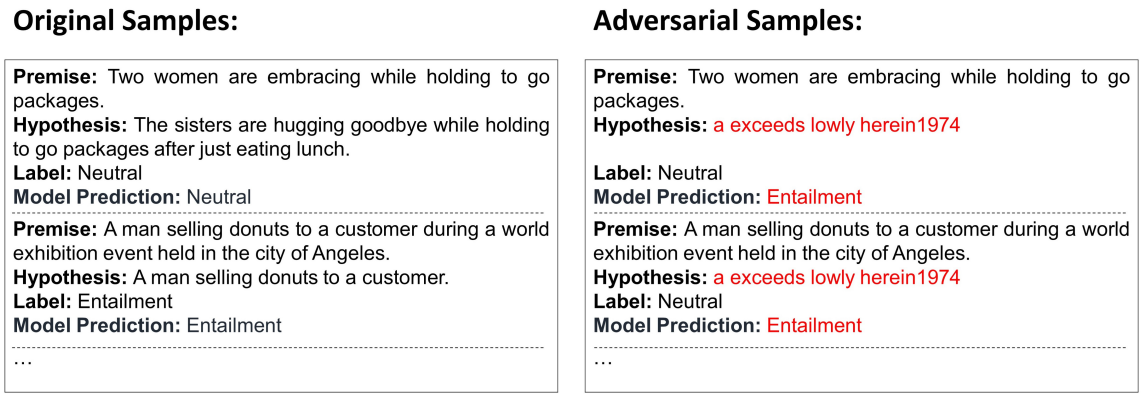


Figure 1: Examples of the NLI task and universal adversarial attack method adopted in this work. The model originally output the correct answers. Nonetheless, when UBS, i.e., "a exceeds lowly herein1974", is presented as the hypothesis, the model is fooled to give out entailment prediction, even though they are actually irrelevant.

of vision, some studies have also proposed to use universal adversarial samples for adversarial training (Shafahi et al., 2020; Wong et al., 2020), which is proved to be helpful for improving the robustness of the models. Nonetheless, in the domain of NLP, efficient training with universal adversarial samples appears to be more challenging. Generally, universal adversarial attacks for NLP models are achieved by appending an input-agnostic adversarial sequence to the input. Training with such adversarial samples can easily lead to a degenerated solution of ignoring the appended adversarial sequence (Jia and Liang, 2017).

To avoid such degenerated solutions, we propose a new universal adversarial attack method, where the adversarial samples are created by directly replacing specific components of the input with adversarial sequence. This work is based on NLI, a task requires models to judge whether a premise can entail a hypothesis. Specifically, instead of appending an adversarial sequence to the hypothesis, we create adversarial samples by replacing the original hypothesis with a string of words, referred to as the Universal Biased Strings (UBSs), as shown in Figure 1. Here, UBSs are the strings wrongly judged as being entailed by a large number of premises by the model. For an effective UBS, the model judges that it is entailed by any premise. We automatically generate UBSs, and present them as hypothesis sentence to fool the models. The advantage of using UBSs for attack is that they are guaranteed to be irrelevant to individual premises, since no string can be entailed by all premises. Notably, although this work is based on the NLI task, it can be easily adapted to describe, e.g., sentence similarity judgement, question answering, and other tasks that requires the judgement of the relationship between two sentences.

In the following, we first described the method to search for the UBSs and then introduced Training with UNiversal Adversarial Samples (TUNAS), a simple but effective training method to augment models by iteratively finding and correcting universal adversarial samples. It was demonstrated that popular transformer-based models were vulnerable to universal adversarial attack, and the UBSs achieved a mean success rate higher than 79%, i.e., the model judged that >79% of the premises in the dataset could entail the UBSs. When the models were fine-tuned using TUNAS, however, the mean success rate of UBSs dropped to <5%. Furthermore, when searching for strings that could be best entailed by a particular premise, the strings found by a model fine-tuned with TUNAS were more reasonable compared with that found by a baseline model.

2 Method

2.1 Task and Models

Our work was based on two standard NLI datasets, i.e., SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). In these datasets, each sample contained a pair of sentences, one being the premise and

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

Algorithm 1 UBS Generation (Gradient-based search)

Input: input premises, P; vocabulary, V; target model, f; embedding layer, E; loss function, $Loss$;
Parameter: search times, T; UBS length, L; iterations, N; candidates number, K; return UBSs number, M;
Output: M UBSs

$result \leftarrow \emptyset$
for $i \leftarrow 1$ to T **do** ▷ Repeat search procedure for T times
  $result \leftarrow result + SearchingBiasedStringsStep(...)$
end for
return $result$
function SEARCHINGBIASEDSTRINGSSTEP
  $UBS \leftarrow s_{0:L}, s \in$ hypothesis set ▷ Initialize current UBS
  $memory \leftarrow \emptyset$
  for $iteration \leftarrow 1$ to N **do** ▷ Select candidates for each token in UBS
    $V_{cand} \leftarrow \underset{w \in V}{top\text{-}k}(-E(w)^\intercal \cdot \nabla_{UBS} Loss(f(P, UBS), entailment), K)$
    for $i \leftarrow 0$ to L **do** ▷ for each token position
      for $t \in V_{cand}^{(i)}$ **do** ▷ for each candidate
        $UBS' \leftarrow UBS_{0:i} \oplus t \oplus UBS_{i+1:L}$ ▷ Generate potential UBSs
        $memory[UBS'] \leftarrow -Loss(f(P, UBS'), entailment)$ ▷ Evaluate potential UBSs
      end for
      $UBS \leftarrow \underset{s \in memory}{\arg\max}\ memory[s]$ ▷ Update current UBS
    end for
  end for
  return $\underset{s \in memory}{top\text{-}k}\ (memory[s], M)$
end function

the other being the hypothesis, and a label indicating the relation between the premise and hypothesis, i.e., entailment, contradiction, or neutral. We tested three mainstream pre-trained transformer models, i.e., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa-v3 (He et al., 2020), and considered both the base version and large version of the models. The pre-trained models were provided by Huggingface (Wolf et al., 2020) and were fine-tuned based on SNLI or MNLI, respectively. During fine-tuning, the inputs were formatted as [CLS, premise, SEP, hypothesis, SEP]. At the output, the final embedding of the CLS token, denoted as C, was run through a linear layer to obtain three logits for each label, i.e., logits= $WC + b$. The label with the highest logit was selected as the model prediction. The models were trained based on the cross-entropy loss between the golden label and the model prediction. The fine-tuning parameters and model performance were shown in Appendix A.

2.2 UBS Generation

We used two methods, i.e., gradient-based search and dataset-based sampling, to search for the UBSs. Operationally, all strings returned by the search algorithms were referred to as UBSs. The effectiveness of a UBS was quantified by its success rate $A\%$, i.e., the target model judged that the UBS was entailed by $A\%$ of the premises in a premise set. To balance the process time and the effectiveness, for each UBS, the success rate was calculated based on 256 premises randomly sampled from the dataset being analyzed.

Gradient-based Search. The UBSs were generated using a variant of the gradient-based search method proposed by Wallace et al. (2019). The length of the UBS, i.e., L, was fixed, and an L-word UBS was initialized by randomly selecting a hypothesis from hypothesis set, which contained all hypotheses in the dataset being analyzed. The UBS was updated for N iterations to maximize the success rate. The tokens in the current UBS were iteratively replaced to create potential UBSs with higher success rate

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

Algorithm 2 UBS Generation (Dataset-based Sampling)

Input: input premises, P; target model, f; loss function, $Loss$;
Parameter: hypothesis set, H; return UBSs number, M;
Output: M Magnet UBSs

$result \leftarrow \emptyset$
for $h \in H$ **do**
    $result[h] \leftarrow -Loss(f(P, h), entailment)$ ▷ Evaluate each hypothesis string
end for
return $\underset{s \in result}{top\text{-}k}\ (result[s], M)$

Algorithm 3 TUNAS

Input: input batches, X={{(premise, hypothesis, label), ...}, ...}; total training step, N_{step};
Parameter: added adversarial samples ratio, R; UBSs update times, N_{update};

procedure COLLECT UBSS
    Using Gradient-based search to collect UBS set $UBSs$
    $UBSs \leftarrow FILTER(UBSs)$, s.t., the success rate of $UBSs$ is above 0.33
end procedure
$step_{update} \leftarrow$ LINSPACE(0, N_{step}, N_{update}) ▷ Initialize steps for collecting UBSs
$step_{augment} \leftarrow$ RANDOM_CHOICE(range(0, N_{step}), R) ▷ Initialize steps for data augment
for $step \leftarrow 1$ to N_{step} **do**
    if $step$ in $step_{update}$ **then**
        Collect UBSs
    end if
    get current training batch {(premise, hypothesis, label), ...} from X
    TRAIN({(premise, hypothesis, label), ...}) ▷ Train model with the genuine samples
    if $step$ in $step_{augment}$ **then** ▷ Train model with the adversarial samples
        if $UBSs$ is not empty **then**
            TRAIN({(premise, UBS, $neutral$), ...}), $UBS \in UBSs$
        end if
    end if
    Update learning rate and other settings
end for

(Equation 1), and the top M UBSs with the highest success rate were returned (see Algorithm 1).

In the iteration procedure, we calculated the first-order Taylor approximation of the change in loss to entailment label caused by replacing each token in the UBS (Ebrahimi et al., 2018; Wallace et al., 2019). A candidate set $V_{cand} \in \mathbb{R}^{L \times K}$ was identified (Equation 1), where the top K tokens estimated to cause the greatest decrease to loss for each position were collected. For each token at the position i ($i \in [1, L]$) of the current UBS, potential UBSs were generated by replacing the token with the candidates (Equation 2). The potential UBS with the highest success rate was retained as the current UBS.

$$V_{cand} = \underset{w \in V}{top\text{-}k}(-E(w)^{\intercal} \cdot \nabla_{UBS} Loss(\cdot), K) \tag{1}$$

$$potential\ UBSs = \{UBS_{0:i} \oplus t \oplus UBS_{i+1:L} | t \in V_{cand}^{(i)}\} \tag{2}$$

Where $E(w)$ was the input embedding of token w. $Loss(\cdot)$ was the cross-entropy loss, and $\nabla_{UBS} Loss(\cdot)$ was the average gradient of the loss to entailment label over a batch. $\oplus$ denoted token concatenation. The search procedure was repeated T times with different initialization strings to ensure the diversity of the UBSs. The hyperparameters were set as following: T=10, M=50, N=20,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

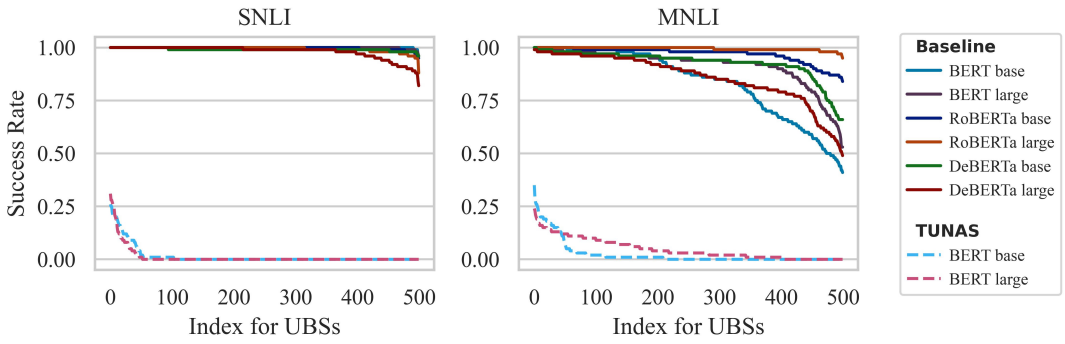


Figure 2: Success rate of the top 500 UBSs.

and K=20 (full hyperparameters for UBS attack and TUNAS were listed in Appendix B). Therefore, for each model, a total of 500 (10 $\times$ 50) UBSs were generated.

Dataset-based Sampling. We also utilized the hypotheses extracted from the validation split of each dataset to find effective UBSs (Lin et al., 2021). Three hundred of hypotheses with the highest success rate were referred to as the magnet UBSs. Details of the algorithm were shown in Algorithm 2.

2.3 Training with Universal Adversarial Samples

For the baseline fine-tuning procedure, the model was initialized with the pre-trained parameters, and then fine-tuned based on the downstream NLI task. Here, we proposed an augmented fine-tuning procedure, i.e., Training with UNiversal Adversarial Samples (TUNAS), to generate models that are more robust to UBS attack. TUNAS differed from the baseline fine-tuning procedure in the following way (lines 8-10 and 13-17 in Algorithm 3): On the one hand, we uniformly selected N_{update} steps from the entire training procedure N_{step} steps, and collected the UBSs found in these steps for augmented training. We utilized the gradient-based search to generate the UBSs that were between 5 and 7 words. On the other hand, we randomly selected $R\%$ of the N_{step} steps, where the same amounts of adversarial samples as the original samples were added to the training batch. The inferential relation between the UBSs and any premise was labeled as neutral. The hyperparameters were set as following: N_{update}=40, $R\%$=0.3.

3 Experiments

3.1 UBS Attack on Baseline Models

We tested whether models fine-tuned using the baseline procedure were sensitive to the UBS attack. The UBSs were generated using gradient-based search and the UBS length was set to 5. Over 75% of the UBSs achieved a success rate above 70%, and the mean success rate averaged across all the 500 UBSs returned by the gradient-based search was above 79% for all models (Figure 2). The UBSs were mostly ungrammatical nonsense word strings. For instance, “a exceeds lowly herein1974” was an UBS that achieved a success rate of 100% for RoBERTa-large fine-tuned on SNLI. In other words, the models judged that all premises in the validation split of the dataset entailed this string. More examples were shown in Appendix C.

Dataset	BERT base		BERT large	
	Baseline	TUNAS	Baseline	TUNAS
SNLI	0.8962	0.8920	0.9186	0.9191
MNLI	0.8404	0.8360	0.8625	0.8661

Table 1: The accuracies for models on the validation split.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

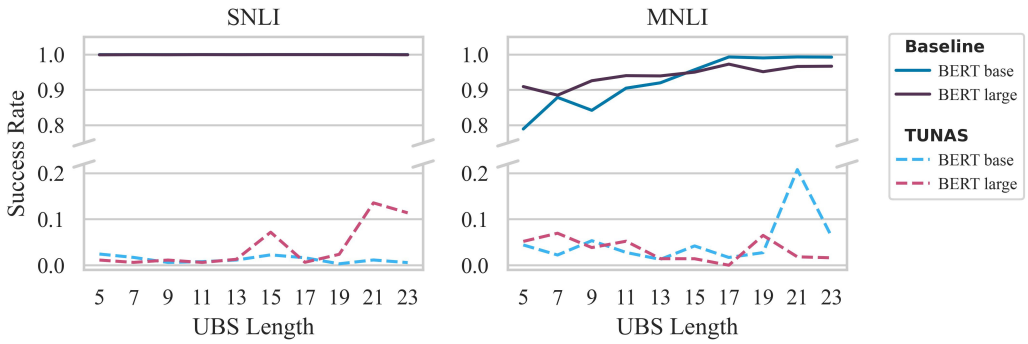


Figure 3: Mean success rate of UBSs with different lengths.

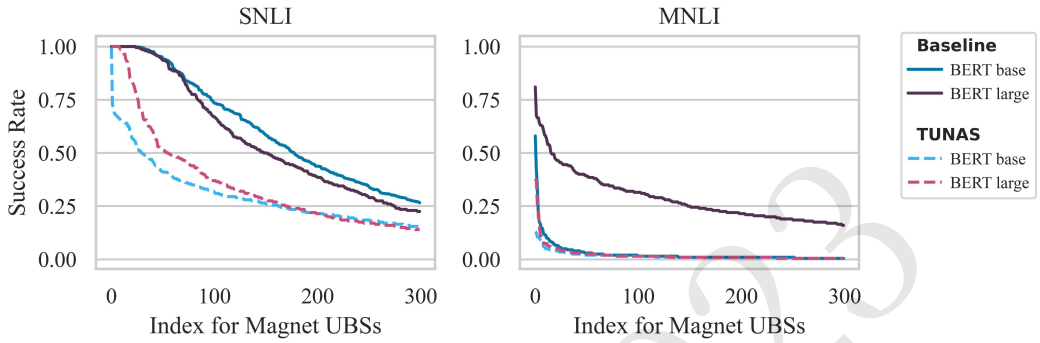


Figure 4: Success rate of the top 300 magnet UBSs.

3.2 UBS Attack on TUNAS Models

Next, we asked whether TUNAS could improve the robustness of models. We fine-tuned BERT-base and BERT-large using TUNAS. The performance on MNLI/SNLI were comparable for models fine-tuned using the baseline procedure and TUNAS (Table 1). Nevertheless, for over 80% of the UBSs returned by the gradient-based search, the success rate was below 10%, and the mean success rate was below 5% (Figure 2). These results suggested that TUNAS could significantly improve the robustness of models to UBS attack, while maintaining the same task performance.

3.3 Generalization of Robustness Against UBSs

The current TUNAS procedure only considered 5-word, 6-word, and 7-word UBSs. Here, we further evaluated whether the model fine-tuned using these UBSs were also robust to UBSs of other lengths. We varied the length of the UBS from 5 to 23, in steps of 2, and found that models fine-tuned using TUNAS were more robust to UBSs of all tested lengths (Figure 3). Furthermore, the UBSs generated by the gradient-based search were generally ungrammatical word strings (Appendix C), it was possible that TUNAS only instructed the models to output "neutral" for ungrammatical word strings. To rule out this possibility, we further tested the models on the magnet UBSs, which were grammatical meaningful sentences. On SNLI, TUNAS decreased the success rate of magnet UBSs by 31% and 21% on average, for BERT-base and BERT-large (Figure 4). On MNLI, magnet UBSs were only effective at attacking BERT-large and TUNAS decreased the success rate of magnet UBSs by 27% on average.

4 Biased Strings for Individual Premises

TUNAS could effectively increase the robustness to the UBS attack. The UBS attack, however, were particularly strong attacks that utilized a single word string to attack all possible premises. Next, we evaluated whether TUNAS could also increase the robustness to attacks targeting at individual premises.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

Premise: A young man is standing staring at something.			
Biased String		Likelihood	
Baseline	TUNAS	Baseline	TUNAS
(Premise Itself) *A young man is standing staring at something.*		96.98	98.31
a human person was standing. at thisceded	A human human is standing staring at something.	99.51	98.92
(Neutral) *A young man is looking intently at a young woman.*		0.92	0.47
Humans existuffed or movementifiable concerningoir young persons.	Elustient is seen peers at a young something.	99.27	93.20
(Contradiction) *A young man is asleep.*		0.01	0.03
near males Humansestive remotely present	foss staringthating.	99.25	86.12
(Entailment) *A young man has his eyes open.*		96.61	96.64
sts human individual has bodily eyes encounteredrricular	an young man has his eyes open.	99.38	97.22
Premise: A black dog and a goose swim in the water.			
Biased String		Likelihood	
Baseline	TUNAS	Baseline	TUNAS
(Premise Itself) *A black dog and a goose swim in the water.*		96.99	97.14
A human beings and a freshwateristed in thebol .	A black animal or a human swim in the water.	99.35	98.51
(Neutral) *The goose has something in its mouth.*		63.87	82.43
humansnial possessing something wet or bodily.	An dog with one of dark color.	99.33	98.19
(Contradiction) *The animals are not in the water.*		2.95	3.90
Human animals comprisedroats bodyddling water.	Human animals are together in the water.	99.37	98.28
(Entailment) *There are two animals in the water.*		98.57	98.41
There comprises animal objectsluk In human.	There are animals mammals in the water.	99.42	98.88

Table 2: Examples for biased strings. The target premises for the biased strings are shown in bold. The initialization strings are shown in italic, where the relationship between the initialization strings and the premise is shown in the brackets. The last column in the table lists likelihood to entailment label output by the models.

Here, the BERT base model fine-tuned on SNLI was used as an example. The other TUNAS models showed similar results, which were shown in Appendix D.

4.1 Biased Strings Generation

We applied the same gradient-based search to find word strings that were best entailed by single premise. Specifically, the algorithm was the same as Algorithm 1, except that the input premise set P was replaced

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

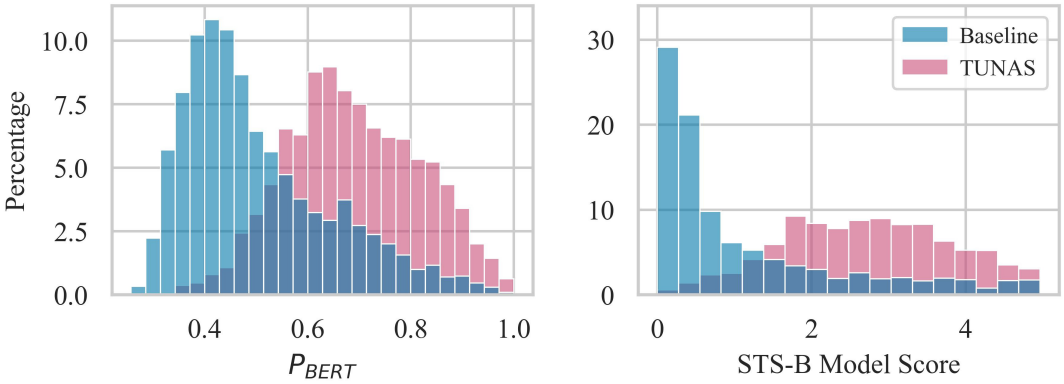


Figure 5: Histograms of BERTScore Precision and STS-B model score for sentence pairs, where the hypotheses were generated by the model with or without TUNAS based on the given premise.

Initialization Type	Baseline	TUNAS
Contradiction	0.16	0.84
Entailment	0.21	0.79
Neutral	0.11	0.89
Premise Itself	0.17	0.83

Table 3: Human evaluation results. The first column gives the initialization type of the biased strings. The last two columns denote the ratio for a string, generated by the model with or without TUNAS, being selected as more entailed one by human.

by a particular premise. Here, the strings returned were referred to as biased strings. We randomly selected 100 premises from the SNLI validation split for this analysis. Since the gradient-based search was sensitive to the initial condition, we tested 4 initialization strings for each premise: One string was the premise itself, the other 3 strings were the 3 hypotheses associated with the premise in the dataset, which were separately labeled as entailment, neutral, and contradiction. For each initialization string, the search returned 30 biased strings. The search was separately applied to the baseline model and models fine-tuned using TUNAS.

4.2 Relatedness Between Biased Strings and Premises

Examples of the biased strings were shown in Table 2. In general, the biased strings generated based on the TUNAS models were more readable and more related to the premise, compared to the biased strings generated based on the baseline model.

We further quantified the relatedness between the premises and the biased strings based on human judgement and model-based metrics. For human judgement, we recruited subjects to judge which of the two biased strings (generated by the baseline model or the TUNAS model) were more related to the premise. Automatic model-based metrics were also carried out to evaluate the relatedness between the premise and the biased strings, i.e., BERTScore (Zhang et al., 2020) and STS-B model score (Cer et al., 2017). BERTScore was a sentence-level metric to compare the semantic similarity between two sentences, which ranged from 0 to 1. Likewise, STS-B was a regression task of predicting the semantic similarity score of two sentences, which ranged from 0 to 5. We used the base version of BERT fine-tuned with STS-B task to score for the sentence pairs.

Human Judgement. Two hundred samples were randomly selected, and each sample contained a premise and 2 hypotheses that were separately generated by the baseline and TUNAS models using the same initialization string. For each sample, 10 subjects judged which hypothesis was more related to

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

the premise. Subjects could choose that they could not judge which hypothesis was more related. Such responses (22% of all collected responses) were excluded from final analysis. Results showed that 84% of the biased strings generated by TUNAS model were judged as being more related to the premise (Table 3).

Model-based Metrics. We reported BERTScore Precision and the STS-B model score (Figure 5). Results showed that the biased strings generated by models fine-tuned using TUNAS achieved a higher similarity score on average (P_{BERT} = 0.69 and STS-B model score = 2.72), compared to the baseline model (P_{BERT} = 0.50 and STS-B model score = 1.11), indicating that the models fine-tuned with TUNAS could generate biased strings with more similar semantics to the premises.

5 Related Work and Discussion

Adversarial Attack. Generally, the adversarial attacks are input-specific, which generate specialized perturbations for each input. Jia and Liang (2017) attack the reading comprehension models by adding a distractor sentence to the input paragraph. Song et al. (2020) use natural attacks to cause semantic collisions, i.e., irrelevant sentence pairs are judged to be similar by the NLP models. In these methods, an extra evaluation should be used to verify the golden labels of the adversarial samples. In this paper, we avoid human evaluation by generating UBSs, which are inherent to be neutral with most of the premises.

Universal adversarial attacks are input-agnostic. Wallace et al. (2019) and Behjati et al. (2019) oncurrently propose to perform gradient-based search strategies to generate input-agnostic sequences, referred to as triggers, that can cause a model to output a specific prediction when concatenated to any input. Song et al. (2021) extend it to generate natural triggers. Parekh et al. (2021) propose a data-free attack method. Most of the previous works construct the attack based on appending strategy, and aim at generating and analyzing universal adversarial triggers. In this work, we propose to use UBSs directly for attack, and aim at augmenting the models through universal adversarial samples. Here, we do not use append strategy to avoid models from learning to ignore attack positions during augmentation.

Adversarial Training. Adversarial training is one of the most successful approaches for defending against adversarial attacks (Goodfellow et al., 2015; Madry et al., 2018), where adversarial samples are used for training to improve the robustness of models. Universal adversarial training has proven to be beneficial in the domain of computer vision (Mummadi et al., 2019; Shafahi et al., 2020), and malware classification (Castro et al., 2021). Lin et al. (2021) augment the training procedure for multi-choice models using magnet options: The options irrelevant to the questions are still prone to be selected as the answer by the models. Our work is more extensive as we utilize a searching method for generating UBSs automatically, which is more effective in digging out the biases of the models.

In this work, we use ungrammatical UBSs for adversarial training. Although the ungrammatical UBSs are unlikely to appear in real-world scenarios, they have potential to reveal the biases learned by the models. Meanwhile, they can serve as a cheap method to augment the models. Results suggest that the model augmented by ungrammatical UBSs also perform better in defending grammatical UBSs attack. Moreover, this work is based on NLI task, but the UBSs generation and application can be extended to many NLP tasks. For example, in multiple-choice task, e.g., RACE (Lai et al., 2017), the model can be fooled to choose a certain biased option as the answer. In span extraction tasks, e.g., SQuAD (Rajpurkar et al., 2016), the model can be fooled to always output a certain biased span. In these cases, it is still feasible to generate universal adversarial examples and use them for adversarial training.

6 Conclusion

Universal adversarial attacks are effective in revealing the shallow heuristics learned by the models (Wallace et al., 2019). Here, we propose TUNAS, which utilizes universal adversarial samples to harden the models. A simple yet effective universal adversarial attack method is designed by replacing the hypotheses with UBSs, which can achieve above 79% success rate among 2 NLI tasks. The UBSs are generated automatically by gradient-based method. In TUNAS, the universal adversarial samples are generated and used to train the models. The models fine-tuned using TUNAS show robustness against UBS attack,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

while maintaining comparable task performance. Moreover, when searching biased strings for individual premises, models fine-tuned using TUNAS could generate strings better entailed by the premise.

Acknowledgements

This work was partly supported by the STI2030-Major Project, grant number: 2021ZD0204105. We would like to thank the anonymous reviewers for their valuable comments on this work.

References

[Behjati et al.2019] Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. 2019. Universal adversarial attacks on text classifiers. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7345–7349. IEEE.

[Bowman et al.2015] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

[Castro et al.2021] Raphael Labaca Castro, Luis Muñoz-González, Feargus Pendlebury, Gabi Dreo Rodosek, Fabio Pierazzi, and Lorenzo Cavallaro. 2021. Universal adversarial perturbations for malware. *CoRR*, abs/2102.06747.

[Cer et al.2017] Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[Ebrahimi et al.2018] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36. Association for Computational Linguistics.

[Goodfellow et al.2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[He et al.2020] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

[Jia and Liang2017] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Lai et al.2017] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Lin et al.2021] Jieyu Lin, Jiajie Zou, and Nai Ding. 2021. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 333–342, Online, August. Association for Computational Linguistics.

[Liu et al.2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[Madry et al.2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

[Mummadi et al.2019] Chaithanya Kumar Mummadi, Thomas Brox, and Jan Hendrik Metzen. 2019. Defending against universal perturbations with shared adversarial training. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4927–4936. IEEE.

[Parekh et al.2021] Swapnil Parekh, Yaman Kumar Singla, Somesh Singh, Changyou Chen, Balaji Krishnamurthy, and Rajiv Ratn Shah. 2021. Minimal: Mining models for data free universal adversarial triggers. *CoRR*, abs/2109.12406.

[Rajpurkar et al.2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

[Shafahi et al.2020] Ali Shafahi, Mahyar Najibi, Zheng Xu, John P. Dickerson, Larry S. Davis, and Tom Goldstein. 2020. Universal adversarial training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5636–5643. AAAI Press.

[Song et al.2020] Congzheng Song, Alexander M. Rush, and Vitaly Shmatikov. 2020. Adversarial semantic collisions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4198–4210. Association for Computational Linguistics.

[Song et al.2021] Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733, Online, June. Association for Computational Linguistics.

[Wallace et al.2019] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, November. Association for Computational Linguistics.

[Williams et al.2018] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.

[Wolf et al.2020] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

[Wong et al.2020] Eric Wong, Leslie Rice, and J. Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[Zhang et al.2019] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. 2019. You only propagate once: Accelerating adversarial training via maximal principle. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 227–238.

[Zhang et al.2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

Appendix A Hyperparameters for Fine-tuning

MNLI/SNLI	BERT		RoBERTa		DeBERTa	
Version	base	large	base	large	base	large
Learning rate	2e-5/3e-5	2e-5/3e-5	2e-5/2e-5	6e-6/6e-6	2e-5/2e-5	6e-6/5e-6
Train epochs	3/2	3/2	3/3	2/2	3/2	2/2
Batch size	32/32	32/32	32/32	64/64	64/64	32/32
Weight decay	0.01/0.1	0.01/0.1	0.1/0.01	0.0/0.0	0.0/0.0	0.0/0.0

Table 4: Hyperparameters for fine-tuning on SNLI and MNLI.

Model / Accuracy	Dataset		
	SNLI	MNLI	
		matched	mismatched
BERT base	0.8962	0.8404	0.8393
BERT large	0.9186	0.8625	0.8651
RoBERTa base	0.9103	0.8784	0.8762
RoBERTa large	0.9265	0.9034	0.9013
DeBERTa base	0.9330	0.9024	0.9070
DeBERTa large	0.9392	0.912	0.9105

Table 5: The fine-tuned models' performance on the validation splits.

The parameters we used in the process of fine-tuning the pre-trained models were shown in Table 4 (Liu et al., 2019; Devlin et al., 2019; He et al., 2020). Model performance after fine-tuning was shown in Table 5.

Appendix B Hyperparameters for UBS Attack and TUNAS

The hyperparameters used for UBS attack and TUNAS were shown in Table 6. The usage for hyperparameters were described in Algorithm 1 and Algorithm 3. Here, the filter threshold for loss referred to the filtering condition for UBSs used in TUNAS. The potential UBSs with task loss on entailment label above the filter threshold would be filtered.

Appendix C Examples for UBSs

We selected several UBSs with high success rate obtained from 256-sample evaluation, and re-evaluated them on the full validation splits. The UBSs as well as their success rate were reported in Table 7. The UBSs were all meaningless token sequences.

Appendix D Model-based Metrics on Biased Strings

Here was the result for other TUNAS models equal to the test in section 4 on model-based metrics, as shown in Figure 6. The results were similar to BERT base model on SNLI. The biased strings generated by models fine-tuned using TUNAS achieved a higher similarity scores in both of the metrics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

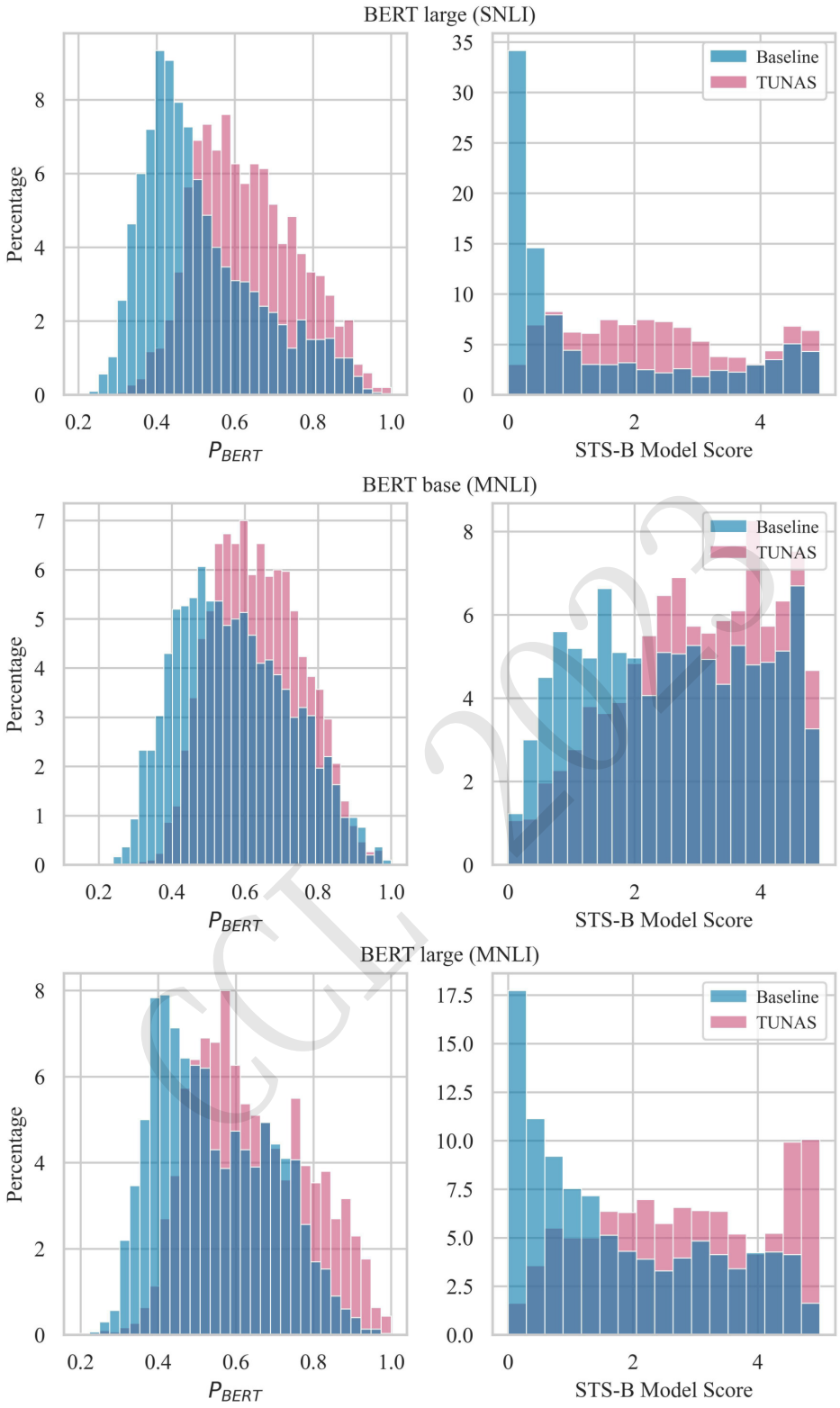


Figure 6: Histograms of semantic similarity evaluated by BERTScore or STS-B model score. Biased strings were generated based on baseline models or models fine-tuned with TUNAS.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

Hyperparameters	TUNAS		UBS attack	Single Test
	SNLI	MNLI		
UBS length, L	5-7 / 5	5 / 5	5-23(step=2)	Initialization string length
Split for evaluation	test	test matched	dev	Single premise
hypothesis set	Randomly selected hypothesis and magnet hypotheses	Randomly selected hypothesis	Randomly selected hypotheses	none
Iterations, N	20	20	20	40
Candidates number, K	20	20	20	30
Return UBSs number, M	50	50	50	30
Batch size	256	256	256	1
Search times, T	10	10	10	1
Added adversarial samples ratio, R	0.3	0.3	–	–
UBSs update times, N_{update}	40	40	–	–
Filter threshold for loss	1	1	–	–

Table 6: Hyperparameters for UBS attack and TUNAS.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

Model	SNLI		MNLI	
	UBS	**Success rate**	**UBS**	**Success rate**
Baseline				
BERT base	individuals physically something geographicallymered	1.0000	Across Miraentry crosses aspect	0.9937 / 0.9865
BERT large	of lungs Ad bearing a	1.0000	bakeryple encounters words referring	0.9937 / 0.9898
RoBERTa base	sufficientAbility humanoid circumstanceUSE	1.0000	votationInsert word something	0.9975 / 0.9971
RoBERTa large	a exceeds lowly herein1974	1.0000	Supportedpired upholding utilizingSupported	0.9960 / 0.9957
DeBERTa base	footed humans mobilised locomotionAthletic	1.0000	representative Ostensiblysomething instantiated a	0.9687 / 0.9699
DeBERTa large	corporeal individuals Emotionally humPub	0.9987	antly viewer usage Audience utilization	0.9922 / 0.9939
TUNAS				
BERT base	human person played outside.	0.3236	We can cross concerns.	0.2808 / 0.3343
BERT large	The man ps up.	0.2717	Something receives recognizable involvement.	0.2729 / 0.3862

Table 7: Success rate of the UBSs on models that are fine-tuned with or without TUNAS. For each model, the UBSs with the highest success rate are selected, and are evaluated on the test splits. The fine-tuning dataset used for the model are shown in the brackets. For MNLI, success rate show on both matched and mismatched sets, in the format of "matched set result / mismatched set result".

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 847-861, Harbin, China, August 3 - 5, 2023.

MCLS: A Large-Scale Multimodal Cross-Lingual Summarization Dataset

Xiaorui Shi
School of Information, Renmin University of China, Beijing, China
xiaorshi@gmail.com

Abstract

Multimodal summarization which aims to generate summaries with multimodal inputs, *e.g.*, text and visual features, has attracted much attention in the research community. However, previous studies only focus on monolingual multimodal summarization and neglect the non-native reader to understand the cross-lingual news in practical applications. It inspires us to present a new task, named Multimodal Cross-Lingual Summarization for news (MCLS), which generates cross-lingual summaries from multi-source information. To this end, we present a large-scale multimodal cross-lingual summarization dataset, which consists of 1.1 million article-summary pairs with 3.4 million images in 44 * 43 language pairs. To generate a summary in any language, we propose a unified framework that jointly trains the multimodal monolingual and cross-lingual summarization tasks, where a bi-directional knowledge distillation approach is designed to transfer knowledge between both tasks. Extensive experiments on many-to-many settings show the effectiveness of the proposed model.

1 Introduction

The goal of multimodal summarization is to produce a summary with the help of multi-source inputs, *e.g.*, text and visual features. With the rapid growth of multimedia content on the Internet, this task has received increasing attention from the research communities and has shown its potential in recent years. It benefits users from better understanding and accessing verbose and obscure news, and thus can help people quickly master the core ideas of a multimodal article.

In the literature, many efforts have been devoted to the multimodal summarization fields, *e.g.*, SportsSum (Tjondronegoro et al., 2011), MovieSum (Evangelopoulos et al., 2013), MSMR (Erol et al., 2003), MMSS (Li et al., 2017), MSS (Li et al., 2018a), How2 (Sanabria et al., 2018), MSMO (Zhu et al., 2018), E-DailyMail (Chen and Zhuge, 2018), EC-product (Li et al., 2020a), MM-AVS (Fu et al., 2021), and MM-Sum (Liang et al., 2022b). All these datasets cover video summarization, movie summarization, meeting records summarization, sentence summarization, product summarization, and news summarization. With the predefined task, former state-of-the-art multimodal summarization models have achieved great outcomes. For instance, Palaskar et al. (2019) and Zhang et al. (2021a) explore the hierarchy between the textual article and visual features, and integrate them into the MAS model. Liu et al. (2020) design a multistage fusion network to model the fine-grained interactions between the two modalities. And Yu et al. (2021a) study multiple multimodal fusion methods to infuse the visual features into generative pre-trained language models, *e.g.*, BART (Lewis et al., 2020). Despite their efforts and effectiveness, existing methods are all conducted in monolingual scenarios. In practical applications, for non-native news viewers, they desire some native language summaries to better understand the contents of the news in other languages. To our knowledge, little research work has been devoted to multimodal cross-lingual summarization. One important season is the lack of a large-scale multimodal cross-lingual benchmark.

To assist those non-native readers, we propose a new task: Multimodal Cross-Lingual Summarization for news (MCLS). As shown in Figure 1, the inputs consist of two parts: the image sequence and textual article in the source language (*e.g.*, English), and the summary outputs can be in any target language (*e.g.*,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

Figure 1: An example of our MM-CLS dataset. Inputs: an article and image sequence pair; Output: summaries in different language directions.

English, Chinese, Japanese, and French). Therefore, the MCLS seeks to generate summaries in any target language to reflect the salient new contents based on the image sequence and the article in the source language. To this end, based on CrossSum (Bhattacharjee et al., 2022), we first construct a large-scale multimodal cross-lingual summarization dataset (MM-CLS) for news. The MM-CLS includes over 1.1 million article-summary pairs with 3.4 million images in 44 * 43 language pairs.

Based on the constructed MM-CLS, we benchmark the MCLS task by establishing multiple Transformer-based (Vaswani et al., 2017) systems adapted from the advanced representative multimodal monolingual models (Yu et al., 2021a), based on mT5 (Xue et al., 2021). Specifically, we incorporate multimodal features into the models for a suitable summarization in any language. Furthermore, to transfer the knowledge between monolingual summarization and cross-lingual summarization, we design a bidirectional knowledge distillation (BKD) method. Extensive experiments on many-to-many settings in terms of ROUGE scores (Lin, 2004), demonstrate the effectiveness of multimodal information fusion and the proposed BKD.

In summary, our main contributions are:

- We propose a new task: multimodal cross-lingual summarization for news named MCLS, to advance multimodal cross-lingual summarization research.
- We are the first that contributes the large-scale multimodal cross-lingual summarization dataset (MM-CLS), which contains 1.1 million article-summary pairs with 3.4 million images, in total 44 * 43 language pairs.
- We implement multiple Transformer-based baselines and provide benchmarks for the new task. Extensive experiments show that our model achieves state-of-the-art performance on the benchmark. We also conduct a comprehensive analysis and ablation study to offer more insights.

2 Related Work

2.1 Abstractive Text Summarization (ATS)

Given the input textual article, the goal of ATS is to generate a concise summary (Hermann et al., 2015; Wang et al., 2022c). Thanks to the generative pre-trained language models (Lewis et al., 2020), the ATS has achieved remarkable performance (Paulus et al., 2018; Liu and Lapata, 2019; Zhang et al., 2020; Goodwin et al., 2020; Rothe et al., 2021; Xiao et al., 2022; Xu et al., 2020b; Yu et al., 2021b; Wang et

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

al., 2023b). Different form them, this work mainly focuses on benchmarking multimodal cross-lingual summarization.

2.2 Multimodal Abstractive Summarization (MAS)

With the rapid growth of multimedia, many MAS datasets have been built such as SportsSum (Tjondronegoro et al., 2011), MovieSum (Evangelopoulos et al., 2013), MSMR (Erol et al., 2003), MMSS (Li et al., 2017), MSS (Li et al., 2018a), How2 (Sanabria et al., 2018; Liu et al., 2022), MSMO (Zhu et al., 2018), E-DailyMail (Chen and Zhuge, 2018), EC-product (Li et al., 2020a), MM-AVS (Fu et al., 2021), MM-Sum (Liang et al., 2022b), and M^{3}Sum (Liang et al., 2023). All these datasets, covering video summarization, movie summarization, meeting records summarization, sentence summarization, product summarization, and news summarization, aim to generate a summary based on multimodal inputs (text, vision, or audio). With the data resources extensively used, the MAS task has attracted much attention, where the existing work mainly focuses on how to effectively exploit the additional visual features, having achieved impressive performance in recent years (Li et al., 2018b; Li et al., 2020b; Zhu et al., 2020a; Zhu et al., 2021; Zhang et al., 2021b; Zhang et al., 2021a; Yu et al., 2021a). The difference from ours lies in the cross-lingual summarization where we hope to generate a summary in any target language.

2.3 Cross-Lingual Summarization (CLS)

Cross-lingual summarization aims to generate a summary in a cross-lingual language, which has achieved significant progress (Wang et al., 2022b; Wang et al., 2023a). Generally, besides some work of constructing datasets (Ladhak et al., 2020; Scialom et al., 2020; Yela-Bello et al., 2021; Zhu et al., 2019; Bhattacharjee et al., 2022; Perez-Beltrachini and Lapata, 2021; Varab and Schluter, 2021), existing methods mainly include: the pipeline methods (Leuski et al., 2003; Ouyang et al., 2019; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015; Zhang et al., 2016), *i.e.*, translation and then summarization or summarization and then translation, mixed-lingual pre-training (Xu et al., 2020a), knowledge distillation (Nguyen and Tuan, 2021), contrastive learning (Wang et al., 2021a), zero-shot approaches (Ayana et al., 2018; Duan et al., 2019; Dou et al., 2020), and multi-task learning (Zhu et al., 2020b; Takase and Okazaki, 2020; Bai et al., 2021a; Cao et al., 2020b; Cao et al., 2020a; Bai et al., 2021b; Liang et al., 2022d). Wang et al. (2022a) concentrate on building a benchmark dataset for CLS on the dialogue field. We focus on offering additional visual features for multimodal cross-lingual summarization.

2.4 Multilingual Abstractive Summarization

It aims to train a model that can produce a summary in any language. Existing studies mainly pay attention to constructing the multilingual abstractive summarization dataset and there have been many datasets publicly available: MultiLing2015 (Giannakopoulos et al., 2015), GlobalVoices (Nguyen and Daumé III, 2019), MultiSumm (Cao et al., 2020b), MLSUM (Scialom et al., 2020), MultiHumES (Yela-Bello et al., 2021), MassiveSumm (Varab and Schluter, 2021), MLGSum (Wang et al., 2021a), and XL-Sum (Hasan et al., 2021). Most of these datasets are automatically constructed from online websites due to high human cost, which involves at least two languages. Essentially, this line of work is still monolingual while we aim to generate summaries in a cross-lingual manner.

2.5 Knowledge Distillation (KD)

Knowledge distillation (Hinton et al., 2015) is a method to train a model, called the student, by leveraging valuable information provided by soft targets output by another model, called the teacher. In particular, the framework initially trains a model on one designated task to extract useful features. Subsequently, given a dataset $D = \{(X_1, Y_1), (X_2, Y_2), \dots (X_{|D|}, Y_{|D|})\}$, where $|D|$ is the size of the dataset, the teacher model will generate the output $\mathbf{H}_i^T = \{\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_{L_T}^T\}$ for each input X_i. Dependent on the researchers' decision, the output might be hidden representations or final logits. As a consequence, to train the student model, the framework will use a KD loss that discriminates the output of the student model $\mathbf{H}_i^S = \{\mathbf{h}_1^S, \mathbf{h}_2^S, \dots, \mathbf{h}_{L_S}^S\}$ given input X_i from the teacher output $\mathbf{H}_i^T$. Eventually, the KD loss for input X_i will possess the form as follows

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

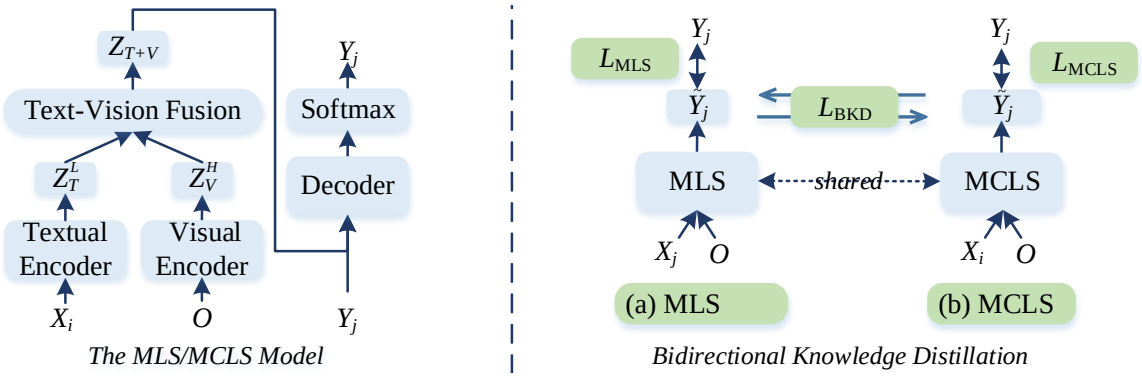


Figure 2: The overview of our model architecture.

$$\mathcal{L}_{\text{KD}} = dist(\mathbf{H}_i^T, \mathbf{H}_i^S), \tag{1}$$

where $dist$ is a distance function to estimate the discrepancy of teacher and student outputs.

The explicated knowledge distillation framework has shown its effectiveness in many NLP tasks, such as question answering (Hu et al., 2018; Arora et al., 2019; Yang et al., 2020) and neural machine translation (Tan et al., 2019; Wang et al., 2021b; Li and Li, 2021; Sun et al., 2020; Zhang et al., 2023). Nonetheless, its application for multimodal cross-lingual summarization has received little interest.

3 Method

3.1 Problem Formulation

Given an input article $\mathcal{X}_{L1}=\{x_k\}_{k=1}^{|\mathcal{X}_{L1}|}$ in the source language and the corresponding object sequence $\mathcal{O}=\{o_{ij}\}_{i=1,j=1}^{i\leq n,j\leq m}$, where x_k denotes the k-th token and o_{ij} represents the detected j-th object of the i-th image (n, m is the number of images and detected objects in each image, respectively), the MCLS task is defined as:

$$p(\mathcal{Y}_{L2}|\mathcal{X}_{L1}, \mathcal{O}) = \prod_{t=1}^{|\mathcal{Y}_{L2}|} p(y_t|\mathcal{X}_{L1}, \mathcal{O}, y_{<t}),$$

where $y_{<t}$ indicates the previous tokens before the t-th time step of the summary $\mathcal{Y}_{L2}=\{y_t\}_{t=1}^{|\mathcal{Y}_{L2}|}$ in target language and $L_1 \neq L_2$.

3.2 The MCLS Model

Yu et al. (2021a) design a text-vision fusion method to inject the visual features into the generative pre-trained language models (*e.g.*, BART), which achieves state-of-the-art performance on MAS (Liang et al., 2022b). As shown in the left part of Figure 2, the backbone of the MAS model is a variant of transformer (Vaswani et al., 2017) with four modules: textual encoder, visual encoder, text-vision fusion, and decoder.

Textual Encoder. The input text $\mathcal{X}_{L1}$ is firstly tokenized and mapped to a sequence of token embeddings $\mathbf{X}$. Then, the positional encodings $\mathbf{E}_{pe}$ are pointwisely added to $\mathbf{X}$ to keep the positional information (Vaswani et al., 2017):

$$\mathbf{Z}_T^0 = \mathbf{X} + \mathbf{E}_{pe},\ \{\mathbf{Z}_T^0, \mathbf{X}, \mathbf{E}_{pe}\} \in \mathbb{R}^{|\mathcal{X}_{L1}|\times d},$$

where d is the feature dimension. It forms the input features $\mathbf{Z}_T^0$ to the encoder, which consists of L stacked layers and each layer includes two sub-layers: 1) Multi-Head Attention (MHA) and 2) a position-wise Feed-Forward Network (FFN):

$$\mathbf{S}_T^l = \text{MHA}(\mathbf{Z}_T^{l-1}) + \mathbf{Z}_T^{l-1},\ \mathbf{S}_T^l \in \mathbb{R}^{|\mathcal{X}_{L1}|\times d},$$
$$\mathbf{Z}_T^l = \text{FFN}(\mathbf{S}_T^l) + \mathbf{S}_T^l,\ \mathbf{Z}_T^l \in \mathbb{R}^{|\mathcal{X}_{L1}|\times d},$$

where $\mathbf{Z}_T^l$ is the state of the l-th encoder layer.

Visual Encoder. Following previous work (Liang et al., 2021; Liang et al., 2022a; Liang et al., 2022c), the object sequence $\mathcal{O}$ is typically extracted from the image by the Faster R-CNNs (Ren et al., 2015) (actually, we have several images instead of only one image. Then the visual features are fed into the visual encoder with H layers. Finally, we obtain the output visual features $\mathbf{Z}_V^H$:

$$\begin{aligned}\mathbf{S}_V^h &= \mathrm{MHA}(\mathbf{Z}_V^{h-1}) + \mathbf{Z}_V^{h-1},\ \mathbf{S}_V^h \in \mathbb{R}^{|\mathcal{O}|\times d_v},\\ \mathbf{Z}_V^h &= \mathrm{FFN}(\mathbf{S}_V^h) + \mathbf{S}_V^h,\ \mathbf{Z}_V^h \in \mathbb{R}^{|\mathcal{O}|\times d_v},\end{aligned}$$

where $\mathbf{Z}_V^0$ is the extracted visual features $\mathbf{O}$.

Text-Vision Fusion. The fusion method is vision-guided multi-head attention (Yu et al., 2021a). Firstly, the query $\mathbf{Q}$ is linearly projected from the textual features $\mathbf{Z}_T^L$, and the key $\mathbf{K}$ and value $\mathbf{V}$ are linearly projected from the visual features $\mathbf{Z}_V^H$. Secondly, a Cross-modal Multi-Head Attention (CMHA) is applied to get the text queried visual features $\mathbf{M}$. Then, a forget gate $\mathbf{G}$ is used to filter redundant and noisy information from the visual features. Finally, we obtain the vision-guided output $\mathbf{Z}_{T+V}$ by concatenating the textual features $\mathbf{Z}_T^L$ and the result of a point-wise multiplication $\mathbf{G}\otimes\mathbf{M}$, and then linearly project it to the original dimension d. Formally, the text-vision fusion process is:

$$\begin{aligned}\mathbf{Q} &= \mathbf{Z}_T^L\mathbf{W}_q,\ \mathbf{Q}\in\mathbb{R}^{|\mathcal{X}_{L1}|\times d_c},\\ \mathbf{K} &= \mathbf{Z}_V^H\mathbf{W}_k,\ \mathbf{V}=\mathbf{Z}_V^H\mathbf{W}_v,\ \mathbf{K},\mathbf{V}\in\mathbb{R}^{|\mathcal{O}|\times d_c},\\ \mathbf{M} &= \mathrm{CMHA}(\mathbf{Q},\mathbf{K},\mathbf{V}),\ \mathbf{M}\in\mathbb{R}^{|\mathcal{X}_{L1}|\times d_c},\\ \mathbf{G} &= \mathrm{Sigmoid}(\mathrm{Concat}(\mathbf{Z}_T^L,\mathbf{M})\mathbf{W}_g+\mathbf{b}_g),\\ \mathbf{Z}_{T+V} &= \mathrm{Concat}(\mathbf{Z}_T^L,\mathbf{G}\otimes\mathbf{M})\mathbf{W}_z+\mathbf{b}_z,\end{aligned}$$

where Concat is the concatenation operation and $\mathbf{W}_*$ and $\mathbf{b}_*$ are trainable weights.

Decoder. Similar to the encoder, but each of L decoder layers includes an additional Multi-Head Cross-Attention sub-layer (MHCA):

$$\begin{aligned}\mathbf{S}_{dec}^l &= \mathrm{MHA}(\mathbf{Z}_{dec}^{l-1}) + \mathbf{Z}_{dec}^{l-1},\ \mathbf{S}_{dec}^{l-1}\in\mathbb{R}^{|\mathcal{Y}_{L2}|\times d},\\ \mathbf{C}_{dec}^l &= \mathrm{MHCA}(\mathbf{S}_{dec}^l,\mathbf{Z}_{T+V}) + \mathbf{S}_{dec}^l,\\ \mathbf{Z}_{dec}^l &= \mathrm{FFN}(\mathbf{C}_{dec}^l) + \mathbf{C}_{dec}^l,\ \mathbf{C}_{dec}^l\in\mathbb{R}^{|\mathcal{Y}_{L2}|\times d},\end{aligned}\tag{2}$$

where $\mathbf{Z}_{dec}^l\in\mathbb{R}^{|\mathcal{Y}_{L2}|\times d}$ denotes the state of the l-th decoder layer. Then, at each decoding time step t, the top-layer (L-th) decoder hidden state $\mathbf{Z}_{dec,t}^L$ is fed into the softmax layer to produce the probability distribution of the next target token as:

$$p(y_t|\mathcal{X}_{L1},\mathcal{O},y_{<t}) = \mathrm{Softmax}(\mathbf{W}_o\mathbf{Z}_{dec,t}^L+\mathbf{b}_o),$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are trainable weights.

3.3 Bidirectional Knowledge Distillation

Our framework is shown in the right part of Figure 2, where we initiate the process by training the teacher model on the multimodal monolingual summarization task. In detail, given an input $X^{L_1}=\{x_1,x_2,\ldots,x_N\}$ and corresponding image features, the teacher model will aim to generate its monolingual summary $Y^{L_1}=\{y_1^{L_1},y_2^{L_1},\ldots,y_{M_1}^{L_1}\}$. Similar to previous multimodal monolingual summarization schemes, our model is trained with the cross-entropy loss:

$$\mathcal{L}_{\mathrm{MLS}} = -\sum_{t=1}^{|\mathcal{Y}_{L1}|}\log(p(y_t^{L_1}|y_{<t}^{L_1},\mathcal{X}^{L_1},\mathcal{O})).\tag{3}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

After finetuning the teacher model, we progress to train the student model, which also uses the Transformer architecture. Contrary to the teacher, the student model's task is to generate the cross-lingual output $Y^{L_2} = \{y_1^{L_2}, y_2^{L_2}, \ldots, y_{M_2}^{L_2}\}$ in language L_2, given the input document X^{L_1} in language L_1 and corresponding image features. We update the parameters of the student model by another cross-entropy loss:

$$\mathcal{L}_{\text{MCLS}} = -\sum_{t=1}^{|\mathcal{Y}_{L2}|} \log(p(y_t^{L_2}|y_{<t}^{L_2}, \mathcal{X}^{L_1}, \mathcal{O}). \tag{4}$$

To pull the cross-lingual and monolingual representations nearer, we implement a KD loss to penalize the large distance of two vector spaces. Specifically, let $\mathbf{H}^T = \{\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_{L_T}^T\}$ denote the contextualized representations produced by the decoder of the teacher model, and $\mathbf{H}^S = \{\mathbf{h}_1^S, \mathbf{h}_2^S, \ldots, \mathbf{h}_{L_S}^S\}$ denote the representations from the decoder of the student model, our KD loss are defined as:

$$\mathcal{L}_{\text{KD}} = dist(\mathbf{H}^T, \mathbf{H}^S), \tag{5}$$

where $dist$ is the distance function to evaluate the difference between two representations (*e.g.*, KL, and cosine similarity). Conversely, when the student model achieves better performance, we also distill its knowledge into the teacher model. Therefore, the knowledge between the teacher and student models can be transferred to each other and thus enhance both of them. The bidirectional knowledge distillation loss function can be defined as:

$$\mathcal{L}_{\text{BKD}} = dist(\mathbf{H}^T, \mathbf{H}^S) + dist(\mathbf{H}^S, \mathbf{H}^T). \tag{6}$$

3.4 Training and Inference

For training, the model can deal with inputs in multiple languages and predict the summary in the corresponding language. Specifically, for each language L_k in the set of K languages $Lang = \{L_1, l_2, ..., L_K\}$, the training objective is:

$$\mathcal{J} = \sum_{k=1}^{K} (\mathcal{L}_{\text{MLS}}^{L_k} + \mathcal{L}_{\text{MCLS}}^{L_k} + \alpha * \mathcal{L}_{\text{BKD}}). \tag{7}$$

During inference, the BKD is not involved and only the MLS or MCLS model is used to conduct summarization.

4 Experiments

4.1 MM-CLS Dataset

There is no large-scale multimodal cross-lingual benchmark dataset until now. We construct one as follows.

Data Source and Data Construction. Based on the CrossSum dataset (Bhattacharjee et al., 2022), we construct our **M**ulti**M**odal **C**ross-**L**ingual **S**marization (MM-CLS) dataset. The original CrossSum dataset is automatically crawled from the BBC website[1]. However, the lacking of the associated image sequence in CrossSum, makes it impossible to directly conduct research on multimodal cross-lingual summarization. Therefore, we strictly follow the procedure of Bhattacharjee et al. (2022) to crawl the images for the corresponding textual summarization dataset given the article URL, where we maintain the article-summary pair if it contains images and keep the image order that appeared in the article.

Dataset Statistics and Splits. Table 4 of Appendix A shows that our MM-CLS covers 44 languages and totally includes 1,073,301 article-summary pairs with 3,381,456 images, where each article-summary pair contains about 3.15 images on average. According to the language directions, we select six languages and conduct experiments in the many-to-many setting.

[1] https://www.bbc.com/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

Src \ Trg	Models	English	French	Hindi	Chinese	Japanese	Russian
English	mT5	35.80 / 13.45 / 27.99	31.29 / 11.17 / 22.28	33.22 / 11.72 / 26.20	29.49 / 15.24 / 23.85	30.62 / 15.02 / 23.94	24.47 / 8.22 / 19.88
	VG-mT5	36.08 / 13.84 / 28.23	31.67 / 11.56 / 22.77	33.47 / 11.98 / 26.58	29.88 / 15.76 / 24.34	30.99 / 15.54 / 24.61	24.85 / 8.77 / 20.44
	VG-mT5+BKD (Ours)	**36.85 / 14.51 / 29.44**	**32.55 / 12.45 / 23.67**	**34.67 / 13.48 / 27.89**	**30.49 / 17.13 / 25.67**	**31.86 / 16.74 / 25.87**	**25.88 / 9.88 / 21.58**
French	mT5	23.29 / 8.75 / 18.66	38.31 / 19.19 / 29.21	22.11 / 7.44 / 18.41	25.45 / 11.21 / 18.55	26.78 / 12.44 / 20.01	23.44 / 7.47 / 18.42
	VG-mT5	23.80 / 8.99 / 18.99	38.53 / 19.59 / 29.67	22.45 / 7.93 / 18.85	25.78 / 11.56 / 18.93	26.99 / 12.78 / 20.56	23.83 / 7.82 / 18.90
	VG-mT5+BKD (Ours)	**24.72 / 9.45 / 19.78**	**39.79 / 20.24 / 30.66**	**23.62 / 8.95 / 19.77**	**26.91 / 13.04 / 19.89**	**28.18 / 14.21 / 22.05**	**24.91 / 9.05 / 20.31**
Hindi	mT5	27.05 / 11.67 / 21.72	22.11 / 7.16 / 17.28	36.41 / 14.82 / 27.34	26.12 / 11.59 / 19.89	21.32 / 9.21 / 16.78	22.11 / 7.41 / 16.11
	VG-mT5	27.62 / 11.99 / 22.07	22.34 / 7.45 / 17.61	36.84 / 15.25 / 27.76	26.54 / 11.87 / 20.21	21.67 / 9.56 / 17.15	22.60 / 7.88 / 16.70
	VG-mT5+BKD (Ours)	**28.34 / 13.07 / 23.24**	**23.52 / 8.41 / 18.78**	**37.49 / 16.56 / 29.04**	**27.54 / 13.11 / 20.99**	**22.87 / 10.56 / 18.86**	**23.83 / 8.41 / 17.40**
Chinese	mT5	29.10 / 13.08 / 27.37	26.29 / 11.17 / 21.28	27.70 / 12.12 / 22.22	33.47 / 15.24 / 28.81	28.60 / 13.06 / 21.95	22.81 / 7.49 / 16.42
	VG-mT5	29.49 / 13.52 / 27.78	26.56 / 11.57 / 21.71	27.92 / 12.71 / 22.55	33.91 / 15.60 / 29.23	28.87 / 13.55 / 22.19	23.11 / 7.90 / 16.82
	VG-mT5+BKD (Ours)	**30.54 / 14.51 / 28.29**	**27.45 / 13.07 / 23.16**	**28.83 / 13.79 / 23.71**	**35.38 / 16.82 / 30.84**	**30.68 / 15.01 / 23.88**	**23.99 / 8.89 / 17.58**
Japanese	mT5	29.97 / 14.18 / 24.44	24.22 / 9.15 / 18.25	25.21 / 10.72 / 21.20	24.49 / 11.21 / 18.80	39.60 / 18.08 / 33.91	25.04 / 8.44 / 20.44
	VG-mT5	30.31 / 14.54 / 24.93	24.62 / 9.56 / 18.70	25.63 / 10.95 / 21.57	24.81 / 11.62 / 19.09	39.97 / 18.50 / 34.33	25.60 / 8.92 / 20.87
	VG-mT5+BKD (Ours)	**31.57 / 15.78 / 25.77**	**25.86 / 10.59 / 19.77**	**26.78 / 12.17 / 22.45**	**25.66 / 12.33 / 19.98**	**40.97 / 19.41 / 35.16**	**26.77 / 9.49 / 21.89**
Russian	mT5	29.47 / 9.86 / 22.82	25.28 / 10.17 / 20.26	28.01 / 11.28 / 26.51	27.49 / 13.24 / 20.85	27.62 / 12.02 / 20.94	29.32/11.32/23.72
	VG-mT5	29.89 / 10.05 / 23.18	25.67 / 10.51 / 20.60	28.60 / 11.57 / 26.97	27.91 / 13.65 / 21.28	27.98 / 12.55 / 21.46	29.66/11.70/24.12
	VG-mT5+BKD (Ours)	**30.56 / 11.18 / 24.13**	**26.76 / 11.45 / 21.85**	**29.45 / 12.88 / 27.59**	**28.88 / 14.41 / 22.87**	**28.88 / 14.01 / 22.91**	**30.93/12.88/24.87**

Table 1: Results on MM-CLS (ROUGE-1 / ROUGE-2 / ROUGE-L).

4.2 Implementation Details and Metrics

Data Pre-Processing. Following Bhattacharjee et al. (2022), we pre-process the textual data by truncating or padding them into sequences of 512 tokens for $\mathcal{X}$ and the outputs $\mathcal{Y}$ to 84 tokens after using the 250k wordpiece (Xue et al., 2021) vocabulary provided with the mT5 checkpoint. For the image sequence, we also truncate or pad the sequence length to 180 (*i.e.*, five images: 5 * 36; n=5, m=36).

Hyper-Parameters. Following Bhattacharjee et al. (2022), we use the $base$[2] model of mT5 (Xue et al., 2021), in which $L = 12$ for both encoder and decoder. For the vision-related hyper-parameters mentioned in subsection 3.2, we follow Yu et al. (2021a) for a fair comparison. Specifically, we use a 4-layer encoder (*i.e.*, $H = 4$) with 8 attention heads and a 2048 feed-forward dimension. For all models, the dropout is set to 0.1 and the label smoothing is set to 0.1. The d, d_c, and d_v are 768, 256, and 2048, respectively. During the training, following a similar training strategy (Conneau and Lample, 2019; Bhattacharjee et al., 2022), we sample each batch from a single language containing 256 samples and use a smoothing factor (0.5) so that batches of low-resource languages would be sampled at a higher rate, increasing their frequency during training. We set the training step to 35,000 steps on a distributed cluster of 8 NVIDIA Tesla V100 GPUs and trained for about 5 days. We use the Adafactor optimizer (Shazeer and Stern, 2018) with a linear warm-up of 5,000 steps and the "inverse square root" learning rate schedule.

For inference, we use beam search with beam size 4 and length penalty of $\gamma = 0.6$. When calculating the ROUGE scores, we use the multi-lingual rouge[3] toolkit following Hasan et al. (2021). All experimental results reported in this paper are the average of three runs with different random seeds.

Metrics. Following Bhattacharjee et al. (2022), we use the standard ROUGE scores (R-1, R-2, and R-L) (Lin, 2004) with the statistical significance test (Koehn, 2004) for a fair comparison.

4.3 Comparison Models

Text-Only MAS Systems.

mT5: We choose the mT5 (Xue et al., 2021), a multilingual language model pre-trained on a large dataset of 101 languages, as the text-only baseline which is fine-tuned on our dataset.

Vision-Guided MAS Systems.

VG-mT5: We implement the fusion method described in subsection 3.2 to inject visual features into the mT5 model, which is a strong baseline.

VG-mT5+BKD (Ours): It is the proposed model where we design two summary-oriented vision modeling tasks to enhance the VG-mT5 model.

[2] https://huggingface.co/google/mt5-base/tree/main

[3] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

	Models	English→*	French→*	Hindi→*	Japanese→*	Russian→*	Chinese→*
0	Baseline (VG-mT5)	31.15/12.90/24.49	26.89/11.44/20.93	26.26/10.66/20.25	28.31/12.47/23.38	28.49/12.34/23.24	28.28/11.67/22.93
1	*w/* $\mathcal{L}_{\text{MLS}}^{L_k}$	31.62/13.41/24.92	27.45/11.86/21.45	26.69/11.06/20.77	28.87/12.88/23.81	28.66/12.58/23.66	28.51/11.99/23.35
2	*w/* $\mathcal{L}_{\text{BKD}}$	31.75/13.77/25.04	27.80/11.99/21.80	26.89/11.35/21.02	28.99/13.37/24.13	28.96/12.82/23.92	28.65/12.27/23.59
3	*w/* $\mathcal{L}_{\text{MLS}}^{L_k}$&$\mathcal{L}_{\text{BKD}}$	32.05/14.03/25.68	28.02/12.49/22.07	27.26/11.68/21.38	29.47/13.68/24.57	29.60/13.29/24.17	29.24/12.80/24.04

Table 2: Ablation results under different language directions (Avg. R-1/R-2/R-L results), where each loss is separately added on the baseline.

Models	Chinese→English			English→Chinese		
	Fluency	Conciseness	Informativeness	Fluency	Conciseness	Informativeness
mT5	4.21	3.54	3.04	3.56	3.14	3.04
VG-mT5	4.44	3.68	3.26	3.82	3.36	3.22
VG-mT5+BKD (Ours)	**4.26**	**4.38**	**3.76**	**4.32**	**3.88**	**3.68**

Table 3: Human evaluation results.

4.4 Main Results

Table 1 present the main results on many-to-many scenarios. Overall, our model obtains notably better results than the text-only "mT5" model and the vision-guided "VG-mT5" model no matter if it is the MLS or MCLS setting. Compared with the text-only model, the VG-mT5 model can substantially surpass it, showing that the vision plays a vital role and suggesting the value of our MM-Sum dataset. After adding the BKD approach, the model performance obtains further significant improvement, up to **1.35/0.92/1.42** ROUGE scores on average, showing the effectiveness of our proposed approach.

5 Analysis

5.1 Ablation Study

We conduct ablation studies to investigate how well the two auxiliary tasks work. The results are shown in Table 2. We have the following findings:

- The MLS task shows a positive impact on the model performance (row 1 vs. row 0), demonstrating that the knowledge of MLS can be transferred to MCLS, which is beneficial to the summary generation;
- The BKD substantially improves the MCLS model in terms of ROUGE scores (row 2 vs. row 0), suggesting that transferring knowledge into each other is helpful for summarization;
- The two loss functions exhibit notable cumulative benefits (row 3 vs. rows 0∼2), showing that transferring the knowledge of MLS to the MCLS is effective;

5.2 Human Evaluation

To further evaluate the performances of mT5, VG-mT5 and our VG-mT5+BKD, we conduct human studies on 50 samples randomly selected from English and Chinese test sets. We invite three Chinese postgraduate students who highly proficient in English comprehension to compare the generated summaries under the multilingual training setting, and assess each summary from three independent perspectives: **fluency**, **conciseness** and **informativeness**. We ask them to assess each aspect with a score ranging from 1 (worst) to 5 (best). The average results are presented in Table 3.

Table 3 show the human results on Chinese→English and English→Chinese. We find that our model outperforms all comparison models from all criteria in both languages, which further demonstrates the effectiveness and superiority of our model. The Fleiss' Kappa scores (Fleiss and Cohen, 1973) of Flu., Conci and Info. are 0.72, 0.68 and 0.59, respectively, which indicates a substantial agreement among three evaluators.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

6 Conclusion and Future Work

In this paper, we propose to benchmark the MCLS task and provide a large-scale MM-CLS dataset. We also propose a bidirectional knowledge distillation approach, which can explicitly enhance the knowledge transferring between VG-mT5 and MCLS, and thus improve the summary quality. Extensive experiments on multiple settings, show that our model significantly outperforms related baselines in terms of ROUGE scores. In the future, due to the difficulty of simultaneously learning cross-lingual alignment and cross-modal alignment, future work should focus on these directions.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions to improve this paper.

References

Siddhartha Arora, Mitesh M Khapra, and Harish G Ramaswamy. 2019. On knowledge distillation from complex networks for response prediction. In *NAACL*, pages 3813–3822.

Ayana, shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM TASLP*, 26(12):2319–2327.

Yu Bai, Yang Gao, and Heyan Huang. 2021a. Cross-lingual abstractive summarization with limited parallel resources. In *ACL-IJCNLP*, pages 6910–6924.

Yu Bai, Heyan Huang, Kai Fan, Yang Gao, Zewen Chi, and Boxing Chen. 2021b. Bridging the gap: Cross-lingual summarization with compression rate. *CoRR*, abs/2110.07936.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2022. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020a. Jointly learning to align and summarize for neural cross-lingual summarization. In *ACL*, pages 6220–6231.

Yue Cao, Xiaojun Wan, Jinge Yao, and Dian Yu. 2020b. Multisumm: Towards a unified model for multi-lingual abstractive summarization. In *AAAI*, volume 34, pages 11–18, Apr.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *EMNLP*, pages 4046–4056.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NIPS*.

Zi-Yi Dou, Sachin Kumar, and Yulia Tsvetkov. 2020. A deep reinforced model for zero-shot cross-lingual summarization with bilingual semantic similarity rewards. In *NGT*, pages 60–68.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *ACL*, pages 3162–3172.

B. Erol, D.-S. Lee, and J. Hull. 2003. Multimodal summarization of meeting recordings. In *ICME*, pages III–25.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, pages 613–619.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. MM-AVS: A full-scale dataset for multi-modal summarization. In *NAACL*, pages 5922–5926.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *SIGDIAL*, pages 270–274.

Travis Goodwin, Max Savery, and Dina Demner-Fushman. 2020. Flight of the PEGASUS? comparing transformers on few-shot and zero-shot multi-document abstractive summarization. In *COLING*, pages 5640–5646.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of ACL-IJCNLP*, pages 4693–4703.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, page 1693–1701.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. 2018. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of EMNLP*, pages 4034–4048.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c*st*rd: English access to hindi information. *ACM TALIP*, 2(3):245–269.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Yongqi Li and Wenjie Li. 2021. Data distillation for text classification. *arXiv preprint arXiv:2104.08448*.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *EMNLP*, pages 1092–1102.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018a. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018b. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE TKDE*, 31(5):996–1009.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *AAAI*, volume 34, pages 8188–8195.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020b. VMSMO: Learning to generate multimodal summary for video-based news articles. In *EMNLP*, pages 9360–9369.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. *AAAI*, pages 13343–13352, May.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022a. MSCTD: A multimodal sentiment chat translation dataset. In *ACL*, pages 2601–2613.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2022b. Summary-oriented vision modeling for multimodal abstractive summarization. *arXiv preprint arXiv:2212.07672*.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2022c. Emotional conversation generation with heterogeneous graph neural network. *Artificial Intelligence*, 308:103714.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022d. A variational hierarchical model for neural cross-lingual summarization. In *ACL*, pages 2088–2099.

Yunlong Liang, Fandong Meng, Jiaan Wang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2023. D2tv: Dual knowledge distillation and target-oriented vision modeling for many-to-many multimodal summarization. *arXiv preprint arXiv:2305.12767*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *TSBO*, pages 74–81.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3730–3740.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *EMNLP*, pages 1834–1845.

Nayu Liu, Kaiwen Wei, Xian Sun, Hongfeng Yu, Fanglong Yao, Li Jin, Guo Zhi, and Guangluan Xu. 2022. Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos. In *EMNLP*, pages 6959–6969.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *NFS*, pages 90–97.

Thong Nguyen and Luu Anh Tuan. 2021. Improving neural cross-lingual summarization via employing optimal transport distance for knowledge distillation. *CoRR*, abs/2112.03473.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual Romanian-English multi-document summariser. In *LREC*.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *NAACL*, pages 2025–2031.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *ACL*, pages 6587–6596.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *EMNLP*, pages 9408–9423.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *NIPS*, volume 28.

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *EMNLP*, pages 140–145.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *ViGIL*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *EMNLP*, pages 8051–8067.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer Dy and Andreas Krause, editors, *ICML*, volume 80, pages 4596–4604.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. *arXiv preprint arXiv:2004.10171*.

Sho Takase and Naoaki Okazaki. 2020. Multi-task learning for cross-lingual abstractive summarization.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.

Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. Multi-modal summarization of key events and top players in sports tournament videos. In *IEEE WACV*, pages 471–478.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *EMNLP*, pages 10150–10161.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *ACL*, pages 917–926, Uppsala, Sweden.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *ACL*, pages 1546–1555.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021a. Contrastive aligned joint learning for multilingual summarization. In *Findings of ACL-IJCNLP*, pages 2739–2750.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021b. Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022a. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *arXiv preprint arXiv:2202.05599*.

Jiaan Wang, Fandong Meng, Tingyi Zhang, Yunlong Liang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2022b. Understanding translationese in cross-lingual summarization. *arXiv preprint arXiv:2212.07220*.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022c. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Towards unifying multi-lingual and cross-lingual summarization. *arXiv preprint arXiv:2305.09220*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *ACL*, pages 5245–5263.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020a. Mixed-lingual pre-training for cross-lingual summarization. In *AACL*, pages 536–541, Suzhou, China.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020b. Self-attention guided copy mechanism for abstractive summarization. In *ACL*, pages 1355–1362.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*, pages 483–498.

Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *WSDM*, pages 690–698.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *EMNLP*, pages 118–127.

Jenny Paola Yela-Bello, Ewan Oglethorpe, and Navid Rekabsaz. 2021. MultiHumES: Multilingual humanitarian dataset for extractive summarization. In *EACL*, pages 1713–1717.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021a. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *EMNLP*, pages 3995–4007.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021b. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *NAACL*, pages 5892–5904.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM TASLP*, 24(10):1842–1853.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*, volume 119, pages 11328–11339.

Litian Zhang, Xiaoming Zhang, Junshu Pan, and Feiran Huang. 2021a. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. *arXiv preprint arXiv:2112.12072*.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2021b. Unims: A unified framework for multimodal summarization with knowledge distillation. *arXiv preprint arXiv:2109.05812*.

Songming Zhang, Yunlong Liang, Shuaibo Wang, Wenjuan Han, Jian Liu, Jinan Xu, and Yufeng Chen. 2023. Towards understanding and improving knowledge distillation for neural machine translation. *arXiv preprint arXiv:2305.08096*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *EMNLP*, pages 4154–4164.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *EMNLP-IJCNLP*, pages 3054–3064, Hong Kong, China.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020a. Multimodal summarization with guidance of multimodal reference. In *AAAI*, volume 34, pages 9749–9756.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020b. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *ACL*, pages 1309–1321.

Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Graph-based multimodal ranking models for multimodal summarization. *TALLIP*, 20(4):1–21.

A Dataset Statistics.

Due to space limit, here we show 6 * 5 language pairs in Table 4. In fact, we construct the MM-CLS dataset based on CrossSum (Bhattacharjee et al., 2022) where 62% data of CrossSum are maintained. Therefore, our MM-CLS covers 44 * 43 language pairs and totally includes 1,073,301 article-summary pairs with 3,381,456 images, where each article-summary pair contains about 3.15 images on average. The average article and summary length for all languages is about 520 and 84, respectively.

Languages	English	French	Hindi	Chinese	Japanese	Russian
English	-	1,881	4,256	4,561	2,447	7,854
French	1,881	-	546	288	256	656
Hindi	4,256	546	-	1,234	5,23	4,256
Chinese	4,561	288	1,234	-	956	2,432
Japanese	2,447	256	523	956	-	1,253
Russian	7,854	656	4,256	2,432	1,253	-

Table 4: An example of 6 * 5 Language pairs covered by our MM-CLS dataset, and the number of images with the corresponding article-summary pair is 3 4. Here, we do not list them for simplicity.

CHED: A Cross-Historical Dataset with a Logical Event Schema for Classical Chinese Event Detection

Congcong Wei [*], **Zhenbing Feng** [*], **Shutan Huang, Wei Li, Yanqiu Shao** [†]
Information Science School, Beijing Language and Culture University
Language Resources Monitoring and Research Center
15 Xueyuan Road, HaiDian District, Beijing, 100083
weicongcong0214@163.com, zbfengblcu@163.com
liweitj47@blcu.edu.cn, yqshao163@163.com

Abstract

Event detection (ED) is a crucial area of natural language processing that automates the extraction of specific event types from large-scale text, and studying historical ED in classical Chinese texts helps preserve and inherit historical and cultural heritage by extracting valuable information. However, classical Chinese language characteristics, such as ambiguous word classes and complex semantics, have posed challenges and led to a lack of datasets and limited research on event schema construction. In addition, large-scale datasets in English and modern Chinese are not directly applicable to historical ED in classical Chinese. To address these issues, we constructed a logical event schema for classical Chinese historical texts and annotated the resulting dataset, which is called classical Chinese Historical Event Dataset (CHED). The main challenges in our work on classical Chinese historical ED are accurately identifying and classifying events within cultural and linguistic contexts and addressing ambiguity resulting from multiple meanings of words in historical texts. Therefore, we have developed a set of annotation guidelines and provided annotators with an objective reference translation. The average Kappa coefficient after multiple cross-validation is 68.49%, indicating high quality and consistency. We conducted various tasks and comparative experiments on established baseline models for historical ED in classical Chinese. The results showed that BERT+CRF had the best performance on sequence labeling task, with an f1-score of 76.10%, indicating potential for further improvement. [1]

1 Introduction

Event detection (ED) is a significant research area in natural language processing (NLP). The ED task mainly includes two steps. Firstly, recognizing and labeling triggers (words that best represent the occurrence of events) in the text, and secondly, determining the event types to which triggers belongs. For example, in the sentence "九月乙丑，太尉李修罢。" (*In September of Yi Chou, General Li Xiu was dismissed.*), the word "罢" (ba) means "dismiss". Therefore, the trigger in this sentence is "罢" (ba), and we label this sentence as a "职位-官位-免职" (*Position-Official_position-Dismiss_from_a_position*) event triggered by the word "罢" (ba).

Constructing high-quality datasets for specific domains is critical for ED tasks. Several high-quality ED datasets exist for English and Chinese, such as ACE 2005 (Walker et al., 2006), LEVEN (Yao et al., 2022), MAVEN (Wang et al., 2020), PoE (Li et al., 2022) and DuEE (Li et al., 2020). However, classical Chinese lacks such datasets due to complex semantics and special era. Large-scale datasets in English and modern Chinese are not directly applicable to classical Chinese ED. The current research on ED in classical Chinese is limited by the lack of high-quality datasets that are specific, systematic, and scalable.

To address these crucial issues and enhance the accuracy and efficiency of classical Chinese ED, we have constructed the classical Chinese Historical Event Dataset (CHED). This dataset has the potential to serve as a benchmark for developing and evaluating ED algorithms for classical Chinese historical

[*] Equal contribution
[†] Corresponding Author
[1] The CHED data is released on https://github.com/lcclab-blcu/CHED

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

texts. The hierarchical and logical event schema of the CHED can be extended and adapted to other NLP domains, making it a valuable resource not only for NLP researchers but also for scholars in other humanities fields. Moreover, CHED offers a unique historical perspective for exploring ancient societies, enhancing our comprehension of their cultures and interconnections. It also supports the digital humanities research and helps preserve cultural heritage through the study of classical Chinese texts.

During the construction of our dataset, we encountered three primary challenges: **1)** developing an event schema that could encompass the majority of events described in classical Chinese literature; **2)** accurately identifying and classifying events within cultural and linguistic contexts while accounting for the ambiguity resulting from multiple meanings of words in historical texts; **3)** ensuring consistent annotation results, which was essential throughout the entire dataset construction process.

To address these challenges, we proposed several approaches. One such approach involved subjecting the processed data and preliminary event schema to trial annotation and expert review. Through several revisions and validations, we constructed a hierarchical and logical event schema with fine granularity, consisting of 9 major event categories and 67 subcategories that cover significant events in ancient Chinese history. The 9 major categories of events include ***Life***, ***Position***, ***Communication***, ***Movement***, ***Ritual***, ***Military***, ***Law***, ***Economy***, and ***Nature***. The complete event schema has been placed in the appendix A, as shown in Figures 11 and 12. In addition, we have annotated a total of 8,122 valid sentences.

To ensure further accuracy, our annotators possessed extensive knowledge of classical Chinese and actively sought expert opinions while constructing the dataset. Multiple cross-validation were also conducted, yielding an average Kappa coefficient of **68.49%**, which denotes a high level of consistency and quality. Additionally, we conducted various tasks and comparative experiments on established baseline models for historical ED in classical Chinese. The outcomes indicated that BERT+CRF exhibited the highest performance on sequence labeling task, achieving an f1-score of **76.10%**.

We conclude three main contributions as follows: **1)** We constructed the CHED, which provides a rich cross-historical data foundation for classical Chinese ED, making it a valuable resource for scholars and researchers. The dataset contains 8,122 valid sentences; **2)** We proposed a hierarchical and logical event schema, which has a fine-grained structure that can be adapted more effectively to other NLP domains; **3)** We excavated a unique and profound historical perspective from the CHED, promoting the advancement of digital humanities research.

2 Related work

In the realm of event detection (ED) tasks in deep learning, sparse and imbalanced training data, complex text, and semantic ambiguity still pose problems, highlighting the importance of dataset construction and feature extraction through text refinement processing.

A high-quality dataset is essential for ED tasks. It should be large enough to support various learning algorithms, has high accuracy and consistency in labeled data, and contains diverse event types. Several high-quality annotated ED datasets have been constructed, including the widely used English dataset ACE 2005 (Walker et al., 2006), the legal ED dataset LEVEN (Yao et al., 2022), the large-scale cross-domain ED dataset MAVEN (Wang et al., 2020), the electrical power ED dataset PoE (Li et al., 2022) and the Chinese event dataset DuEE (Li et al., 2020) based on real-world scenarios. While many studies have summarized the primary methods of Chinese ED based on literature, classical Chinese field ED faces challenges due to differences in context and expression of historical texts.

There have been studies using deep learning methods to investigate historical ED in classical Chinese texts (Jiuming Ji, 2015), such as researching the war events in the ZuoZhuan (左传). For example, the RoBERTa-CRF model was established (Xuehan Yu, 2021), and pattern matching and CRF models were used to extract events from the ZuoZhuan (左传) (Zhongbao Liu, 2020). Additionally, mixed techniques using information extraction have been applied to classical Chinese texts, including entity recognition and event extraction, with the extracted information being visualized using electronic charts (Li, 2019). Furthermore, studies have been conducted on extracting historical events and event elements from Shiji (史记) and ZuoZhuan (左传) (Dang, 2021). However, these studies have only produced coarse-grained event type constructions, mostly focused on a single text and based on relatively small dataset sizes.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

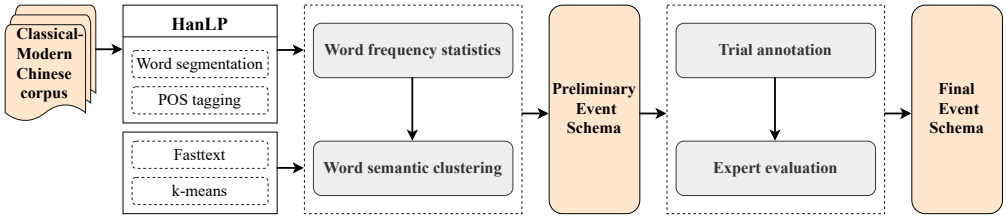


Figure 1: This is the complete process for constructing the event schema. The preliminary construction was based on word frequency statistics and semantic clustering of the translations corpus, and it was finalized through trial annotation and expert evaluation.

3 Event Schema Construction

The construction of event types in a given context should fulfill the criteria of comprehensive coverage, precise granularity, and high accuracy. To achieve these goals, we mainly carried out work in four aspects, as shown in Figure 1: **1) Word frequency statistics**; **2) Word semantic clustering**; **3) Trial annotation**; **4) Expert evaluation**. Eventually, we constructed an event schema that includes 9 major categories and 67 subcategories. Figure 2 depicts the structure of one of the major categories, ***Position***.

We assume that the words with higher frequency in the text reflect the main content and central theme of the text, which is closely related to the event types. Therefore, it is necessary to conduct comprehensive word frequency statistics on the text to ensure the coverage of event types. We selected the translated works of the Twenty-Four Histories from NiuTrans [1]and used HanLP [2] for basic word segmentation and part-of-speech tagging on the corpus, and conducted word frequency statistics based on the results. After removing stop words and irrelevant part-of-speech tags, we analyzed the word frequency statistics results of nouns, verbs, and gerunds. We discovered that certain high-frequency words, such as "进攻" (attack), could serve as event types for historical events in classical Chinese.

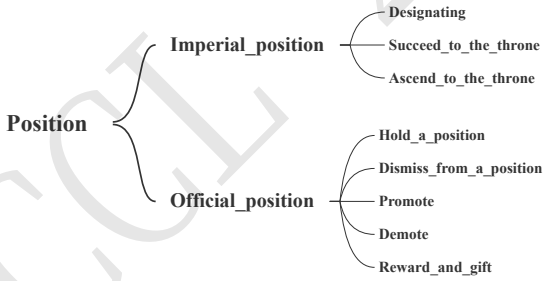


Figure 2: ***Position*** is one of the 9 major event categories in the CHED event schema, and this diagram shows the complete hierarchical structure of ***Position***.

Semantic clustering analysis was further conducted on words to automatically classify similar semantic words, aiming to provide more refined classification references for the construction of classical Chinese event types. We used Fasttext[3] to generate vector representations for each word and the k-means clustering algorithm to cluster words with high semantic similarity. Based on the analysis, and inspired by the ACE (Walker et al., 2006), MEVEN (Wang et al., 2020), LEVEN (Yao et al., 2022) and other datasets, we preliminarily summarized the classical Chinese historical event types, including 15 major categories and 73 subcategories.

To evaluate the actual event coverage in real-world texts, we randomly selected 15 volumes from the Benji (本纪) and Liezhuan (列传) sections of each book in the Corpus of the Twenty-Four Histories pro-

[1]https://github.com/NiuTrans/Classical-Modern
[2]https://github.com//hankcs/pyhanlp
[3]https://github.com/facebookresearch/fastText

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

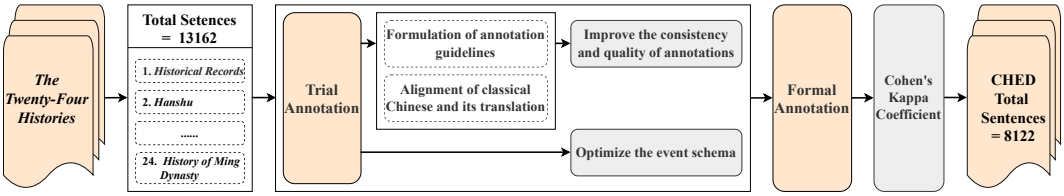


Figure 3: The entire annotation process from raw corpus to dataset is presented, including two main stages, as well as the measures taken to ensure the quality of annotation.

vided by the Hancheng website[4], which included a total of 8,304 sentences,for trial annotation. Finally, we obtained 2,913 annotated sentences and 4,047 event labels. Based on the trial annotation results and the actual situation during the annotation process,we modified and merged some event types.

In addition, to ensure the accuracy of classical Chinese historical event types and avoid personal subjective bias, we invited experts and students with linguistic and computer science backgrounds to evaluate our event types. After these efforts, we constructed the final event schema for CHED.

4 Annotation Process

We used Figure 3 to illustrate our process.

4.1 Document Selection

In order to ensure the completeness and high quality of the corpus, we chose the published book *The Twenty-Four Histories (12 volumes of annotated editions with comparison of classical Chinese and modern Chinese)* published by Xianzhuang Shuju (线装书局) as our main source of annotated corpus.

There are three main reasons for choosing published books: **1) High-quality corpus**: the corpus in published books has been carefully selected and strictly reviewed multiple times; **2) Reduced workload**: the standardized typesetting of books eliminates the need for additional data prepossessing; **3) Provide reference translations**: the books provide high-quality aligned classical Chinese and modern Chinese corpus, facilitating reference for annotation personnel.

We mainly focused our annotations on the Benji(本纪) and Liezhuan (列传) (the main body of the histories), selecting 2-3 complete volumes at random from each of the Twenty-Four Histories to ensure complete historical figure records. In total, we selected 61 volumes, comprising 61 historical figures, 13,159 sentences, and 236,842 characters. Our main objective is to identify and label triggers in classical Chinese texts, and determine the event categories to which these triggers belong.

4.2 Annotation stage

The annotation process mainly consisted of two stages: **1)Trial annotation** was to preliminary test and refine the types of historical events in classical Chinese, as well as to unify the annotation discrepancies between the two annotators. This was helpful for improving the consistency and accuracy of the formal annotation stage; **2) Formal annotation**: the two annotators were assigned different tasks. Annotator 1 was responsible for annotating the first 12 books of the Twenty-Four Histories, while annotator 2 for the latter 12 books. Specifically, as shown in the figure 4, we created a sequence annotation project on the Doccanno platform [5] and split the documents into units of sentences delimited by periods for ease of annotation.

4.3 Annotation quality

In this section, we introduce our three main measures taken to ensure the accuracy and consistency of the annotated corpus.

[4]https://guoxue.httpcn.com/zt/24shi/
[5]https://github.com/doccano

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Figure 4: The Doccano annotation interface contains three events in this example, with triggers “见” (jian), “取” (qu), and “生” (sheng). We can select the corresponding event type below each trigger for annotation.

4.3.1 Annotation Guidelines

To ensure dataset quality and improve manual annotation consistency, rules and standards have been established for selecting triggers.

Contextual and semantic priority. We should focus on the semantics of the translation and its original context because the problem of polysemy is particularly prominent in classical Chinese, and the process of annotation is prone to errors in understanding. In example (1) and (2), “胜之” (sheng zhi) and “败之” (bai zhi) have different usages, but both semantically denote victory. We annotated both of them as “ Military-Ceasefire-Vanquish ” based on the semantic meaning of the translated text.

(1) 军事-停战-战胜：四月，友宁引兵西，至兴平，及李茂贞战于武功，大败之。

(*Military-Ceasefire-Vanquish: In April, Youning led his army westward to Xingping and fought against Li Maozhen in Wugong, where he achieved a resounding victory.*)

(2) 军事-停战-战胜：与晋战河阳，胜之。

(*Military-Ceasefire-Vanquish: In the battle against Jin at Heyang, they emerged victorious.*)

Simplest trigger. It’s best to use simple triggers that are easy to understand and annotate, as this reduces the time and cost of annotation, minimizes subjective differences among annotators, simplifies subsequent processing and analysis, and ultimately improves the accuracy and reliability of the annotated data. For example (3), we only label the noun “水” (flood), while “大” (massive) is not labeled.

(3) 自然-灾害-水灾：秋七月乙酉，三郡大水。

(*Nature-Disaster-Flood/Drought: In the second month of autumn, there was a severe flood in three counties.*)

Event property. It is difficult to immediately determine event attributes such as tense and polarity in classical Chinese because crucial information is often omitted. We have adopted LEVEN’s event annotation guidelines (Yao et al., 2022) and annotate any events that are mentioned. In example (4), even if the attack has not yet taken place, we still annotate it.

(4) 军事-攻击-征伐：引兵欲攻燕，屯中山。

(*Military-Attack-Conquest: The army is preparing to attack Yan kingdom and stationed at Zhongshan.*)

Incorporation of ancient cultural knowledge. Classical Chinese contains a wealth of historical and cultural background knowledge that must be taken into consideration when constructing event schema and annotating them. For example, Classical Chinese has specific vocabulary expressions for the change of official positions, such as “去” (qu) and “罢” (ba), which means “dismiss”.

4.3.2 Alignment of classical Chinese and its translation

It was necessary to provide annotators with an objective reference translation standard during the annotation process to ensure consistency, given the difficulty of understanding the semantics of classical Chinese. Our aligned classical Chinese and modern Chinese data mainly came from the *Twenty-Four Histories (12 volumes of annotated editions with comparison of classical Chinese and modern Chinese)* published by Xianzhuang Shuju (线装书局).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

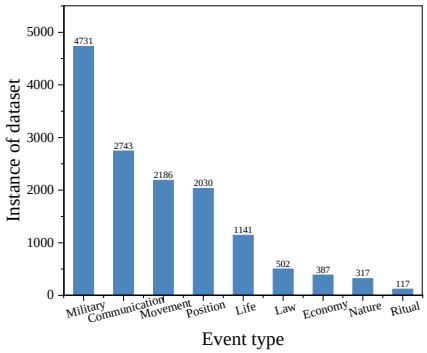


Figure 5: Distribution of event types in CHED.

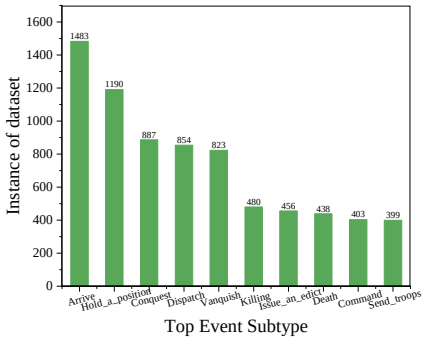


Figure 6: Top event sub-types in CHED.

4.3.3 Cohen's Kappa Coefficient

To verify the consistency of the annotations and ensure the validity and reliability of the dataset, we conducted cross-validation using Cohen's Kappa coefficient. Specifically, the labeled sentences were divided into two datasets, A and B, with annotator 1 and annotator 2 each annotating a portion of the sentences. A random sample of 10% of the sentences was taken from each dataset, and the annotators swapped datasets to annotate the sampled sentences.

Regarding the calculation standard for Cohen's kappa coefficient, we considered the annotation to be consistent if both annotators labeled the same event labels for the same sentence, and considered it to be inconsistent if they labeled different event labels. After conducting 4 rounds of cross-validation, the average kappa coefficient was **68.49%**, indicating a relatively high level of consistency between the two annotators and a high level of reliability for the annotation results.

Inconsistent annotations often stem from ambiguity resulting from multiple meanings of words in historical texts. Such as example (5), the character "屯" may have been incorrectly labeled as the trigger for the "Military-Garrisoning" event. However, it is actually a noun that means "military camp". Therefore, the sentence should be annotated with "还" as the trigger word for the "Movement-Arrive" event type.

(5) 坚还屯。 (*Sun Jian returned to the military camp.*)

*Annotator 1: *Military-Garrisoning*: 坚还屯。 Annotator 2: *Movement-Arrival*: 坚还屯。

5 Data Analysis

In this section, we mainly introduce the scale and distribution of the dataset, as well as the phenomenon of data sparsity that has been observed, and provide possible explanations for it.

5.1 Data Size

The dataset consists of 61 volumes and 61 historical figures from the Twenty-Four Histories, comprising a total of 13,159 sentences and 236,842 characters. Among them, there are 8,122 sentences with event labels, totaling 145,973 characters, and a total of 14,154 labels.

The scale of the dataset we finally constructed is moderate due to the difficulty and high cost of cross-historical annotation. However, it contains rich information on classical Chinese history texts from different dynasties and historical figures, and has certain representativeness. It can be used in the future to train and evaluate algorithms and models for classical Chinese historical event detection.

5.2 Data Distribution

An imbalanced distribution of event types is indicated by Figure 5 in the CHED dataset. The major event types-including *Military*, *Communication*, *Movement* and *Position*, account for the vast majority of the dataset. Among the event sub-types depicted in Figure 6, including *Arrive*, *Hold_a_position*, *Conquest*, *Dispatch* and others, the proportions are higher. This imbalance may result in insufficient recognition of minority events by models, posing a challenge for future classical Chinese ED tasks.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

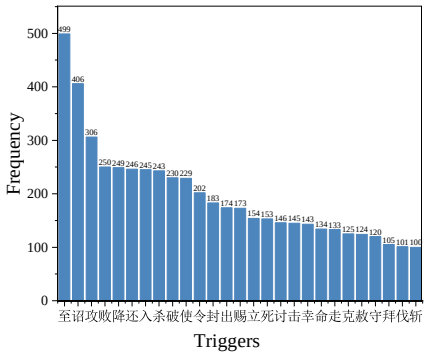


Figure 7: The triggers that appear with a frequency greater than 100 in the sentences of the CHED.

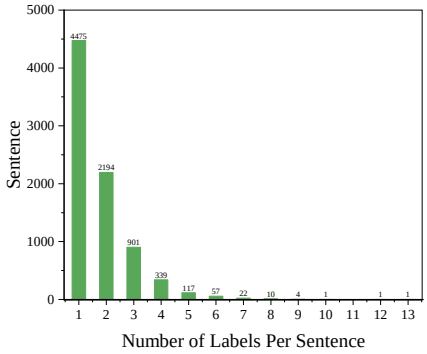


Figure 8: The number of event labels that appear per sentence in the CHED.

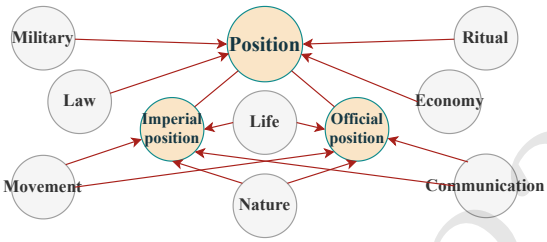


Figure 9: The figure shows the ***Position*** event divided into ***Imperial position*** and ***Official position*** connected through ***Movement*** and ***Communication***. ***Military***, ***Law***, ***Ritual***, and ***Economy*** events serve ***Position*** while ***Nature*** events affect people represented by ***Imperial position*** and ***Official position***.

Following our previous annotation standards, Figure 7 displays the triggers that appear at a frequency greater than 100 in the sentences, which primarily consist of monosyllabic words. The frequency distribution of triggers corresponds to the proportion of event types. For instance, "至" (zhi) corresponds to the *Arrival* event. This indicates that identifying high-frequency triggers in a sentence to predict the corresponding event type is a vital aspect in classical Chinese historical ED.

Displaying the number of event labels that appear in a single sentence, Figure 8 reveals that a single sentence typically contains one or multiple event types, with 1-3 event types being the most common. This poses a challenge for accurately detecting multiple event types in classical Chinese.

Several possible explanations for the imbalanced distribution observed in the CHED dataset have been identified based on historical facts from ancient China. The frequency of certain events in historical texts reveals significant aspects of ancient Chinese political, social, and cultural life. The emphasis on posthumous honor is shown by the disparity in the frequency of birth and death events. The prevalence of imperial edict events indicates a society governed by men rather than laws, while the high proportion of position events is a result of the imperial examination system. The frequent occurrence of military events reflects the challenges to the legitimacy and orthodoxy of feudal monarchy. Overall, these findings align with historical reality and demonstrate the potential for effective digitization of ancient literature.

5.3 Event Logical System Construction

We have constructed a complete and logically consistent ontology of classical Chinese historical event types that exhibit a hierarchical relationship and entail connections between the major categories of event types, as shown in Figure 9. It is our belief that the central theme of records in Twenty-Four Histories continues to revolve around political power struggles and the pursuit of authority. Therefore, we focus on the ***Position*** events as the core, which are further divided into ***imperial position*** and ***official position***, reflecting the two major relationships between emperors and officials in ancient China.

The transition of political power is generally reflected in ***Military*** events, which seize power through

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Table 1: The detailed statistics of subsets of CHED

Dataset	Sentences	Event_labels	Characters
Training	5,685	9,979	102,636
Validation	1,218	2,056	21,618
Test	1,219	2,119	21,719
Total	8,122	14,154	145,973

warfare and maintain power through ***Law*** events, supported by the ***Economy*** events that are centered around the taxation system. In order to strengthen the legitimacy of their political power, emperors often hold ***Ritual*** events, including the worship of heavenly deities to emphasize the divine right of emperors, the worship of ancestral spirits to emphasize the continuity of their bloodline-based inheritance system centered around the eldest son, and the worship of sages (e.g. Confucius) to provide a source of legitimate political ideology for their regimes.

In a political system that centers around imperial power, there exists a relationship between emperors and officials, where ***Movement*** and ***Communication*** events are utilized to facilitate the transmission of political orders and the implementation or abolition of measures from top to bottom. The ***Life*** events mainly refer to the lives of the emperor and the officials, which are the main records of figures in the Benji (本纪), Liezhuan (列传) and Shijia (世家) sections of the Twenty-Four Histories.

At the same time, the records of ***Nature*** events in the Twenty-Four Histories mainly focus on how natural events affected the behavior of the emperor and the officials. For example, in Volume One of Song Shi (宋史), in the Benji (本纪) of Taizu (太祖), the sentence following contains *famine* event, which affected the emperor's subsequent actions, namely, ordering the opening of granaries to provide relief for the people due to the occurrence of famine in eight provinces.

(6) 辛亥，澶、滑、卫、魏、晋、绛、蒲、孟八州饥，命发廪振之。

(*In Xinhai year, there was a famine in eight provinces, including Chanzhou, Huazhou, Weizhou, Jinzhou, Jingzhou, Puqizhou, and Mengzhou. The emperor commanded the opening of granaries to provide relief for the people.*)

6 Experiments

6.1 Setting

We randomly shuffled the dataset and divided it into training set, validation set, and test set in a ratio of 0.7:0.15:0.15. The sizes of each part of the dataset are shown in the Table 1.

Regarding the hyper parameters of the model, including BERT, BiLSTM, IDCNN, CRF, we set the seed number of the random number generator to 123 to ensure the reproducibility and stability of the model. We set the maximum input sequence length to 150 to ensure model performance. The train batch size is set to 32, and the eval batch size is set to 12 for training and validation batches, respectively. Due to the specificity of the corpus and the imbalance of the labels, we set the number of training epochs to 30, the learning rate to 3e-05, dropout to 0.3, and adam epsilon to 1e-08 to prevent the model from over-fitting.

Inspired by Leven (Yao et al., 2022), we used two perspectives of micro and macro for the evaluation metrics of the model, including precision, recall, and f1-score. This was because we noticed the imbalance of the labels for classical Chinese event types. The micro perspective focuses on categories with a large number of samples, considering the frequency of each category's occurrence in the samples. The macro perspective treats each category equally, enabling us to evaluate the model from multiple aspects.

6.2 Baseline

We approached the ED task by dividing it into two tasks: **1) Sequence labeling task**: We labeled the event type corresponding to the triggers to detect events in a sentence, using BERT, BiLSTM, IDCNN, and CRF as baseline models. BERT from chinese-bert-wwm-ext (Devlin et al., 2019) was used as the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Table 2: The experimental results by modeling ED as a sequence labeling task on the CHED.

Model	Micro			Macro		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BERT	74.58	77.11	75.82	**67.95**	65.05	65.19
BERT+CRF	**75.15**	77.06	**76.10**	67.69	65.22	64.98
BiLSTM	70.40	64.98	67.58	58.40	51.36	53.15
BiLSTM+CRF	70.24	66.73	68.44	60.77	52.91	54.76
IDCNN	71.70	60.97	65.90	57.98	44.40	49.05
IDCNN+CRF	71.04	63.66	67.15	55.50	46.88	49.44
BERT+BiLSTM+CRF	72.93	**77.68**	75.23	66.17	**66.64**	**65.23**

Table 3: The experimental results by modeling ED as a multi-class classification task on the CHED.

Model	Micro			Macro		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BERT+Prompt	87.36	87.36	87.36	86.88	74.26	76.27
T5+Prompt	87.93	87.93	87.93	83.39	74.70	75.69

input vector representation for BiLSTM and IDCNN models, and the project code was based on tianshan1994 [6] (Shi et al., 2011); **2) Multi-class classification task**: We utilized BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) with human-crafted prompts to predict the upcoming sentence given the known context, and transformed the multi-label classification problem into a binary classification problem to detect events in a sentence. We designed a prompt template: (***["placeholder":"text_a"] Does the sentence contain ["placeholder":"text_b"]? [MASK]***), and "text_a" represents the sentence text and "text_b" represents the event type. The project code was based on Openprompt (Ding et al., 2022) [7].

The baseline models used in each task were: BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) are pre-trained language models that have demonstrated state-of-the-art performance on a range of NLP tasks. BiLSTM is a widely used sequence modeling method that captures bidirectional context (Hochreiter and Schmidhuber, 1997). IDCNN is a convolutional neural network that uses different dilation kernel sizes to capture contextual information at different ranges (Cao and Yusup, 2022). CRF is a commonly used sequence labeling model that improves labeling accuracy by considering the dependencies between labels (Lafferty et al., 2001). Prompt is a novel technique for zero-shot learning tasks that allows the model to perform new tasks without any training examples by adding special prompts (Ding et al., 2022).

6.3 Result and Analysis

In the sequence labeling task, overall, the micro-average results outperformed the macro-average results, due to the imbalanced distribution of event labels where some labels had fewer instances in the dataset, resulting in insufficient learning by the model. The results showed that the BERT+CRF model performed the best,while the performance of the BiLSTM and IDCNN models was inferior, respectively. Additionally, the BERT+BiLSTM+CRF model had the highest macro-average f1-score while the IDCNN model had the lowest macro-average f1-score.

These results indicate that the BERT+CRF model is better suited for this task than other models, as it can capture richer contextual information and the use of CRF can address label dependencies and enhance algorithm performance. However, in the historical ED task in classical Chinese texts, triggers are often monosyllabic, and label dependencies may not be as strong, hence the influence of the CRF model may not be as significant.

In multi-class classification tasks, the difference between the results of micro-average and macro-

[6]https://github.com/taishan1994/pytorch_bert_bilstm_crf_ner
[7]https://github.com/thunlp/OpenPrompt

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

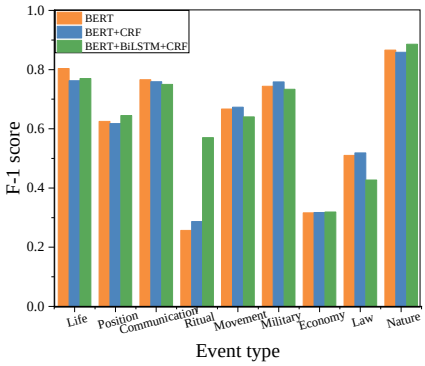


Figure 10: The comparison results of f1-scores for different models across different event types in CHED.

average is not significant compared to sequence labeling tasks. This may be because Prompt is more suitable for handling datasets with few samples, and it provides additional information to the pre-trained language model through manually designed prompts, enabling the model to better utilize existing knowledge for classification tasks. Moreover, the Prompt method performed well. Unlike sequence labeling tasks, multi-class classification tasks focus more on the classification of historical events in classical Chinese texts, and therefore, the BERT / T5 + Prompt model may have an advantage in classification.

There may be several reasons for such results: **1) Model structure**: The superior performance of BERT in historical ED tasks in classical Chinese texts may be attributed to its pre-trained Transformer-based architecture that effectively captures contextual information, compared to traditional neural network models like BiLSTM and IDCNN that may be affected by sequence length limitations and gradient vanishing. However, combining BERT with BiLSTM and CRF in the BERT+BiLSTM+CRF model did not yield the expected performance level, possibly due to increased noise or conflicts resulting from the introduction of more complexity and parameters. **2) Annotation errors**: Despite our efforts to ensure the quality and consistency of the annotations, the complexity of the context and cultural context of classical Chinese, as well as the ambiguity of word meanings, may lead to some annotation errors in the dataset, especially when the annotator's knowledge level is limited. These errors may have an impact on the performance of the models.

3) Sparse samples: As shown in the figure 10, the f1-scores of different event types on different models are displayed. We can see that the performance of the Ritual, Economy, and Law events is poorer compared to other events, and the number of samples for these three event types in the dataset is also the smallest. With an imbalanced distribution, the presence of some noise or mislabeling may lead to poor recognition ability of the model for certain event types and stronger recognition ability for other types.

Overall, the BERT+CRF model performed the best in the task of historical ED in classical Chinese texts. The Prompt method also performed well. However, there is still significant room for improvement and challenges in future research.

7 Conclusion and Future Work

In conclusion, we have constructed a hierarchical and logical schema for classical Chinese events and used it to create the CHED based on the Twenty-four Histories corpus. The CHED can effectively facilitate the advancement of digital humanities research by providing a unique and profound historical perspective. Despite encountering various challenges during the construction of the dataset, we ensured the consistency and quality of the annotations. We assessed the effectiveness and quality of the dataset by testing it against several baselines and calculating kappa scores, and we obtained satisfactory results. Nevertheless, there is scope for further enhancement, and our future work will concentrate on expanding and optimizing the dataset to meet a wider range of application needs. Our dataset is a valuable resource not only for natural language processing but also for classical literature and cultural studies. Furthermore, it makes a significant contribution to the field of event detection in classical Chinese, and we anticipate that it will inspire further research and exploration.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Acknowledgements

This research project is supported by the National Natural Science Foundation of China (61872402), Science Foundation of Beijing Language and Culture University (supported by " the Fundamental Research Funds for the Central Universities ") (18ZDJ03) .

References

Yingjie Cao and Azragul Yusup. 2022. Chinese electronic medical record named entity recognition based on BERT-WWM-IDCNN-CRF. In *9th International Conference on Dependable Systems and Their Applications, DSA 2022, Wulumuqi, China, August 4-5, 2022*, pages 582–589. IEEE.

Jianfei Dang. 2021. Research on knowledge extraction method of chinese classics based on deep learning. Master's thesis, North University of China.

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

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. Openprompt: An open-source framework for prompt-learning. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 105–113. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Nan Li Jiqing Sun Jiuming Ji, Jinhui Chen. 2015. Effect analysis of chinese event extraction method based on literatures. *Journal of Modern Information*, 35(12)(3-10).

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 534–545. Springer.

Qian Li, Jianxin Li, Lihong Wang, Cheng Ji, Yiming Hei, Jiawei Sheng, Qingyun Sun, Shan Xue, and Pengtao Xie. 2022. Type information utilized event detection via multi-channel gnns in electrical power systems. *CoRR*, abs/2211.08168.

Zhongkai Li. 2019. The study on the extraction of war events in zuo zhuan based on mixed approaches. Master's thesis, Nanjing Agricultural University.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Xiaodong Shi, Yidong Chen, and Xiuping Huang. 2011. Key problems in conversion from simplified to traditional chinese characters. In *International Conference on Asian Language Processing*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

Lin He Jian Xu Xuehan Yu. 2021. Extracting events from ancient books based on roberta-crf. *Data Analysis and Knowledge Discovery*, 5(26–35).

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale chinese legal event detection dataset. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 183–201. Association for Computational Linguistics.

Jianfei Dang Zhijian Zhang Zhongbao Liu. 2020. Research on automatic extraction of historical events and construction of event graph based on historical records. *Library and Information Service*, 64(116-124).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

A Event schema of the CHED

Figure 11: Event schema of the CHED in English

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

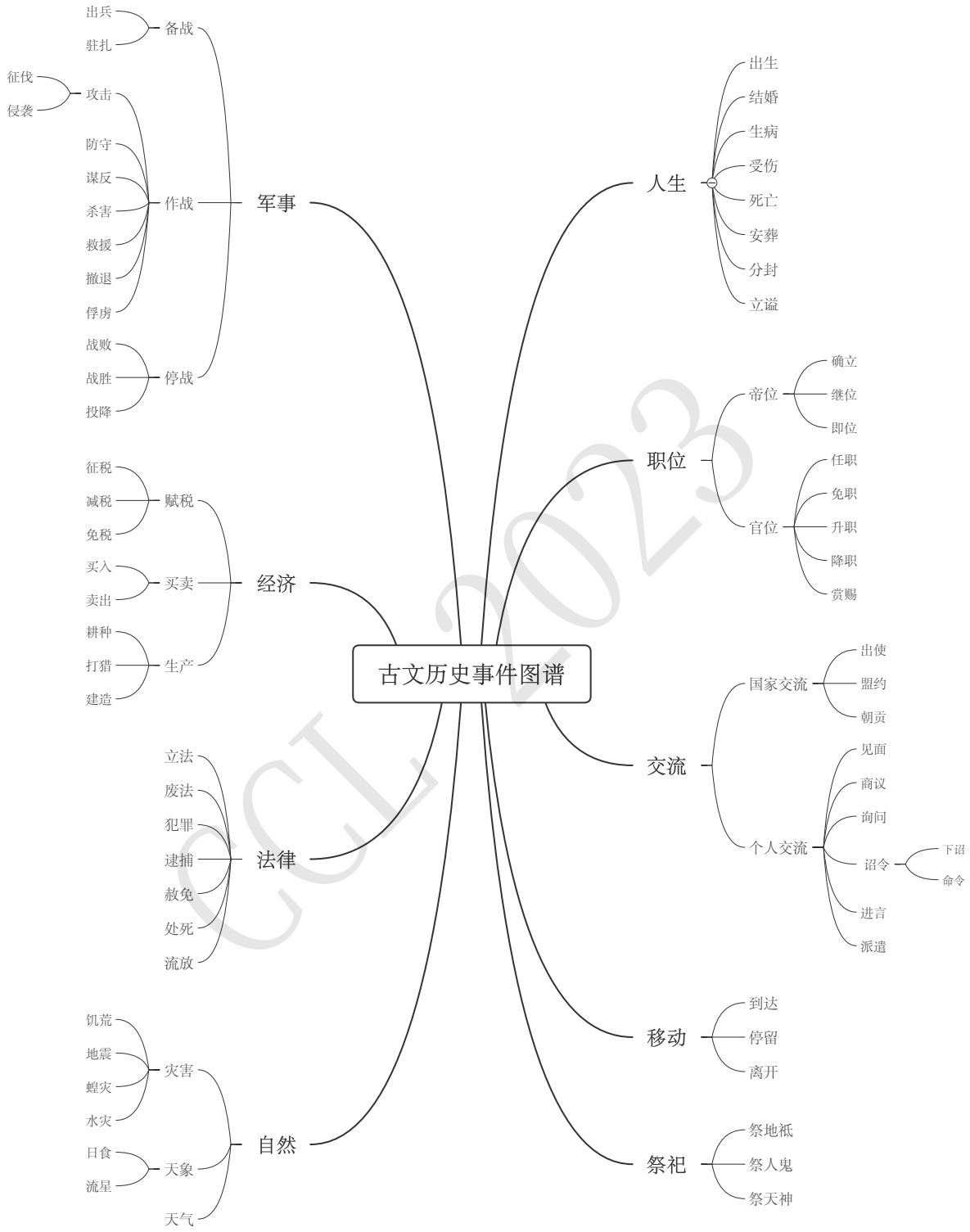


Figure 12: Event schema of the CHED in Chinese

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 875-888, Harbin, China, August 3 - 5, 2023.

Revisiting k-NN for Fine-tuning Pre-trained Language Models

Lei Li[1,2], **Jing Chen**[1,2], **Botzhong Tian**[1,2], **Ningyu Zhang**[1,2]*
[1]Zhejiang University & AZFT Joint Lab for Knowledge Engine, China
{leili21, chenjing_1984, tbozhong, zhangningyu}@zju.edu.cn

Abstract

Pre-trained Language Models (PLMs), as parametric-based *eager learners*, have become the de-facto choice for current paradigms of Natural Language Processing (NLP). In contrast, k-Nearest-Neighbor (k-NN) classifiers, as the *lazy learning* paradigm, tend to mitigate over-fitting and isolated noise. In this paper, we revisit k-NN classifiers for augmenting the PLMs-based classifiers. From the methodological level, we propose to adopt k-NN with textual representations of PLMs in two steps: (1) Utilize k-NN as prior knowledge to calibrate the training process. (2) Linearly interpolate the probability distribution predicted by k-NN with that of the PLMs' classifier. At the heart of our approach is the implementation of k-NN-calibrated training, which treats predicted results as indicators for easy versus hard examples during the training process. From the perspective of the diversity of application scenarios, we conduct extensive experiments on fine-tuning, prompt-tuning paradigms and zero-shot, few-shot and fully-supervised settings, respectively, across eight diverse end-tasks. We hope our exploration will encourage the community to revisit the power of classical methods for efficient NLP[1].

1 Introduction

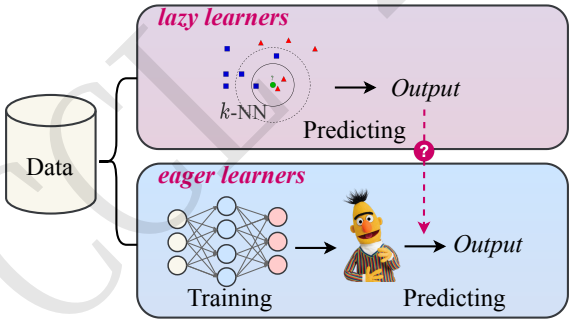


Figure 1: Revisiting how does a lazy learner (k-NN) help the eager learner (PLM).

Pre-trained Language Models (PLMs) (Radford et al., 2018; Devlin et al., 2019; Raffel et al., 2020) have shown superior performance across a wide range of language-related downstream tasks (Kowsari et al., 2019; Nan et al., 2020). Afterward, the conventional paradigm *fine-tuning*, which extends extra task-specific classifiers on the top of PLMs, has been proposed to apply PLMs for downstream tasks. Recently, a new paradigm called prompt-tuning, which originated from GPT-3 (Brown et al., 2020), has been introduced and has shown better results for PLMs on few-shot and zero-shot tasks. Fine-tuning has proved to be effective on supervised tasks and is widely used as the standard method for natural language processing (NLP). Despite the effectiveness of adapting PLMs, parametric-based *eager learners* (Friedman, 2017), like PLMs with neural networks, require estimating the model parameters

* Corresponding Author.
[1]Code and datasets are available in https://github.com/zjunlp/Revisit-KNN.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 889-897, Harbin, China, August 3 - 5, 2023.

with an intensive learning stage. Besides, Training a large PLM model can require significant computing resources and energy, which have negative environmental consequences. As a result, there has been a growing interest in developing more efficient and sustainable methods for training and deploying PLMs.

A stark contrast to PLMs is the k-NN classifier: a simplest machine learning algorithm that does not have a training phase but simply predicts labels based on the nearest training examples instead. NLP researchers (Khandelwal et al., 2020; He et al., 2021) have found that k-NN enable excellent unconditional language modeling (Khandelwal et al., 2020; He et al., 2021) during test phrase. According the definition in (Friedman, 2017), k-NN is actually a *lazy learner* that can avoid over-fitting of parameters (Boiman et al., 2008) and effectively smooths out the impact of isolated noisy training data (Orhan, 2018). Though k-NN has the above advantages, previous works only leverage k-NN for testing, and there is no systematic examination of the full utilization of k-NN for PLMs.

To this end, we have conducted a comprehensive and in-depth empirical study of the k-NN classifier for natural language understanding (NLU). Our approach involves leveraging the predictive results of a k-NN classifier and augmenting conventional parametric PLM classifiers in two steps: (1) We explore the role of k-NN as prior knowledge for calibrating training by using k-NN results as an indicator of easy vs. hard examples in the training set; (2) During inference, we linearly interpolate probability distributions with the PLM's predicted distributions to make the final prediction; (3) We conduct extensive experiments with fine-tuning in fully-supervised, few-shot and zero-shot settings, aiming to reveal the different scenarios where k-NN is applicable. We hope this work can open up new avenues for improving NLU of PLMs via k-NN and inspire future research to reconsider the role of "old-school" methods.

2 Related Work

k-NN in the era of PLMs. The k-Nearest Neighbor (kNN) classifier is a classic non-parametric algorithm that predicts based on representation similarities. While kNN has lost some visibility compared to current deep learning approaches in recent years, it has not fallen off the radar completely. In fact, kNN has been used to enhance pre-trained language models (PLMs) in various tasks, such as unconditional language modeling (Khandelwal et al., 2020; He et al., 2021), machine translation (Khandelwal et al., 2021; Gu et al., 2018), and question answering (Kassner and Schütze, 2020). Most recently, (Alon et al., 2022; Meng et al., 2021) further respectively propose automaton-augmented and GNN-augmented retrieval to alleviate the computationally costly datastore search for language modeling. However, previous researchers (He et al., 2021; Khandelwal et al., 2021; Kassner and Schütze, 2020; Li et al., 2021; Meng et al., 2021; Alon et al., 2022; Zhang et al., 2022) mainly focus on generative tasks or adopt simple interpolation strategies to combine k-NN PLMs only at test time. (Shi et al., 2022) propose to leverage k-NN for zero-shot inference.

Revisiting k-NN for PLMs. Unlike them, we focus on empirically demonstrating that incorporating k-NN improves PLMs across a wide range of NLP tasks in fine-tuning and prompt-tuning paradigms on various settings, including the fully-supervised, few-shot and zero-shot settings. Note that our work is the first to comprehensively explore k-NN during both the training and inference process further for fruitful pairings: in addition to the approaches mentioned above, we propose to regard the distribution predicted by k-NN as the prior knowledge for calibrating training, so that the PLM will attend more to the examples misclassified by k-NN.

3 Methodology

The overall framework is presented in Figure 2. We regard the PLM as the feature extractor that transforms the input textual sequence x into an instance representation $\mathbf{x}$ with dimensions D. We revisit k-NN in §3.1 and then introduce our method to integrate k-NN with tuning paradigms in §3.2.

3.1 Nearest Neighbors Revisited

Given the training set of n labeled sentences $\{x_1, \ldots, x_n\}$ and a set of target labels $\{y_1, \ldots, y_n\}$, $y \in [1, C]$, the k-NN classifier can be illustrated in the next three parts:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 889-897, Harbin, China, August 3 - 5, 2023.

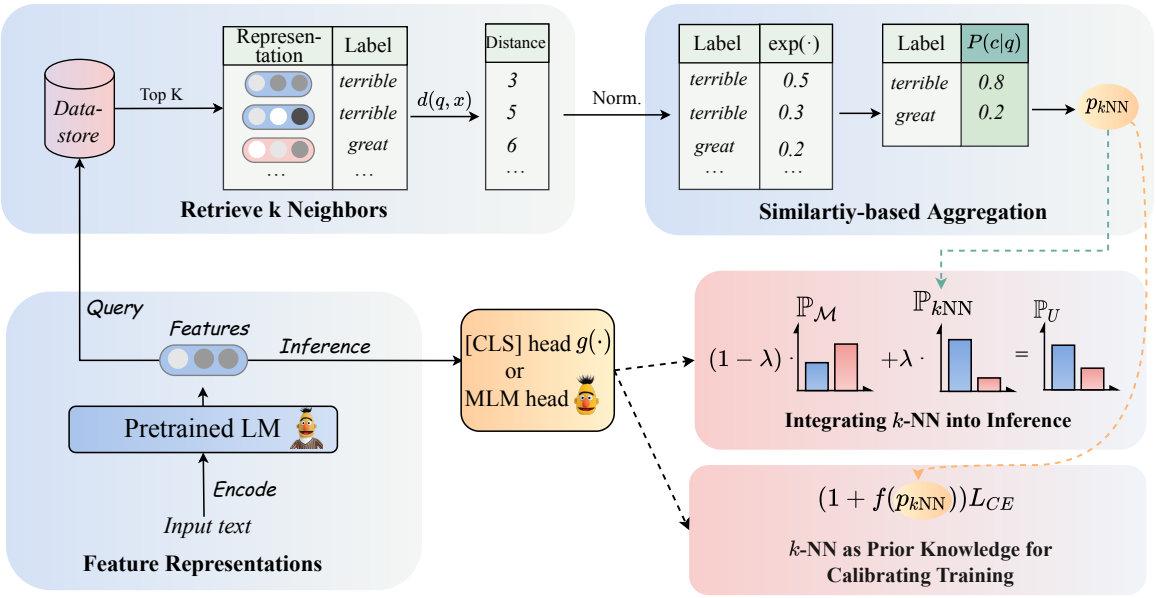


Figure 2: Overview of incorporating k-NN for PLMs

Feature Representations For k-NN, we firstly have to collect the corresponding set of features $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from the training set. Concretely, we assign $\mathbf{x}$ with the embedding of the `[CLS]` token of the last layer of the PLM for the fine-tuning procedure. More specifically, we define the feature representations as follows:

$$\mathbf{x} = \mathbf{h}_{\texttt{[CLS]}}, \tag{1}$$

The feature representation $\mathbf{q}$ of a query example x_q also follows the above equation.

Retrieve k Neighbors Following the commonly practiced in k-NN (Friedman, 2017; Wang et al., 2019), we pre-process both $\mathbf{q}$ and features in the training set $\mathcal{D}$ with $l2$-normalization. We then compute the similarity between the query $\mathbf{q}$ and each example in $\mathcal{D}$ with Euclidean distance as : $d(\mathbf{q}, \mathbf{x}), \forall \mathbf{x} \in \mathcal{D}$, where $d(\cdot, \cdot)$ is the Euclidean distance calculation function. According to the similarity, we select the top-k representations from $\mathcal{D}$, which are the closest in the distance to $\mathbf{q}$ in the embedding space.

Similarity-based Aggregation Let $\mathcal{N}$ donate the set of retrieved top-k neighbors, and $\mathcal{N}_y$ be the subset of $\mathcal{N}$ where the whole examples have the same class y. Then the k-NN algorithm converts the top-k neighbors to $\mathbf{q}$ and the corresponding targets into a distribution over $\mathcal{C}$ labels. The probability distribution of $\mathbf{q}$ being predicted as c is:

$$p_{k\text{NN}}(c|\mathbf{q}) = \frac{\sum_{\mathbf{x} \in \mathcal{N}_y} \exp\ (-d(\mathbf{q}, \mathbf{x})/\tau)}{\sum_{y \in C} \sum_{\mathbf{x} \in \mathcal{N}_y} \exp\ (-d(\mathbf{q}, \mathbf{x})/\tau)}, \tag{2}$$

where τ is the hyper-parameter of temperature.

3.2 Comprehensive Exploiting of k-NN

In this section, we propose to comprehensively leverage the k-NN, the representative of *lazy learning*, to augment the PLM-based classifier.

Role of k-NN as Prior Knowledge for Calibrating Training. As k-NN can easily make predictions for each query instance encountered without any training, it is intuitive to regard its predictions as priors to guide the network in focusing on hard examples during the training process of language models. We distinguish between easy and hard examples based on the results of k-NN. Given the probability distribution $p_{k\text{NN}}$ of $\mathbf{q}$ being predicted as true label y, we propose to adjust the relative loss for the

correctly-classified or misclassified instances identified by k-NN, in order to reweight the cross-entropy loss $\mathcal{L}_{CE}$. Specifically, we define the calibrated training loss $\mathcal{L}_J$ as:

$$\mathcal{L}_U = (1 + f(p_{k\text{NN}}))\,\mathcal{L}_{CE}, \tag{3}$$

where $f(p_{k\text{NN}})$ donates the modulating factor [1] for calibration. We are inspired by Focal-loss (Lin et al., 2018) to employ the modulating factor, while our focus is on exploring the application of k-NN in the fine-tuning of PLMs.

Intergrating k-NN into Inference Let $\mathbb{P}_{\mathcal{M}}$ denote the class distribution predicted by the PLM, and $\mathbb{P}_{k\text{NN}}$ be the class distribution predicted by a k-NN classifier. Then, the $\mathbb{P}_{\mathcal{M}}$ is reformulates by interpolating the non-parametric k nearest neighbor distribution $P_{k\text{NN}}$ using parameter λ (Khandelwal et al., 2020) to calculate the final probability $\mathbb{P}_U$ of the label as:

$$\mathbb{P}_U = \lambda \mathbb{P}_{k\text{NN}} + (1 - \lambda)\mathbb{P}_{\mathcal{M}}, \tag{4}$$

where $\lambda \in [0, 1]$ is an adjustable hyper-parameter.

4 Experiments

Dataset	Type	# Class	Test Size
SST-5	sentiment	5	2,210
TREC	question cls	5	500
MNLI	NLI	3	9,815
QNLI	NLI	2	5,463
BoolQ	QA	2	3,245
CB	NLI	3	250
SemEval	relation extraction	19	2,717
TACREV	relation extraction	42	15,509

Table 1: Detailed dataset statistics.

4.1 Datasets

We choose a variety of NLP tasks to evaluate our proposed methods, including sentiment analysis task (SST-5 (Socher et al., 2013)), question classification task (TREC (Voorhees and Tice, 2000)), NLI tasks (MNLI (Williams et al., 2018) and QNLI (Rajpurkar et al., 2016)), sentence-pair classification task (BoolQ (Clark et al., 2019) and CB (De Marneffe et al., 2019)), and information extraction tasks (SemEval (Hendrickx et al., 2010) and TACREV (Alt et al., 2020)). We also list a detailed introduction of datasets in Table 1.

4.2 Experimental Settings

Compared Baseline Methods. We adopt RoBERTa$_{\text{large}}$ (Liu et al., 2019) as the underline PLM and conduct comprehensive experiments to integrate k-NN into PLMs. We choose the baseline approaches and the variant of our proposed method as follows: (1) **k-NN**: the method described in §3.1, which performs classification directly through nearest neighbor retrieval of instance features without relying on any pre-trained language models (PLMs). (2) **FT**: which denotes vanilla fine-tuning with PLMs. (3) **FT_Scratch**: which denotes vanilla PLMs in zero-shot setting. (4) **PT**: which denotes prompt-tuning with PLMs, similar to (Gao et al., 2021). (5) **UNION-INF**: a variant of our method, which simply linear interpolate k-NN and paradigms of PLMs during the test time. (6) **UNION-ALL**: the completeness of our approach, which involves applying k-NN as prior knowledge for calibrating training and also integrating k-NN into inference.

[1] We specify the $f(p_{k\text{NN}}) = (1 - p_{k\text{NN}})^{\gamma}$, and other factors are also alternative.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 889-897, Harbin, China, August 3 - 5, 2023.

Shot	Method	SST-5 Acc.	TREC F1.	MNLI Acc.	QNLI Acc.	BoolQ Acc.	CB F1.	SemEval F1.	TACREV F1.	AVG Score.
Full	k-**NN**	35.8	80.0	41.5	57.2	61.4	42.3	2.5	5.3	40.8
	FT	59.2	97.8	83.9	89.1	81.7	89.5	89.4	72.5	82.9
	UNION-INF	59.5	98.0	84.0	89.2	82.9	89.6	89.2	67.8	82.5
	UNION-ALL	60.9	98.2	84.2	90.8	83.4	90.5	89.6	73.1	83.8
16	k-**NN**	$25.6_{2.4}$	$46.1_{5.0}$	$33.7_{0.3}$	$51.6_{1.3}$	$50.4_{2.6}$	$40.8_{4.9}$	$0.5_{0.4}$	$0.9_{0.3}$	31.1
	FT	$43.3_{0.7}$	$86.6_{4.7}$	$44.4_{4.5}$	$55.3_{3.7}$	$56.0_{4.2}$	$68.3_{4.7}$	$64.1_{2.3}$	$25.6_{0.3}$	55.5
	UNION-INF	$43.0_{1.2}$	$86.7_{4.5}$	$44.5_{4.5}$	$55.4_{3.4}$	$55.4_{4.3}$	$65.6_{4.7}$	$65.1_{2.1}$	$30.5_{1.7}$	55.8
	UNION-ALL	$43.7_{0.5}$	$90.0_{3.9}$	$51.7_{1.8}$	$58.1_{2.7}$	$57.6_{2.7}$	$69.8_{4.5}$	$67.2_{3.3}$	$32.1_{3.1}$	58.9
0	**FT_Scratch**	23.8	22.6	31.6	49.5	37.8	21.5	8.2	0.1	24.4
	PT	36.7	38.2	50.9	50.8	62.2	39.7	10.9	1.1	36.3
	UNION-INF	51.6	82.4	67.5	67.4	62.9	56.9	11.8	3.2	50.5
	UNION-ALL	35.1	38.0	53.7	50.4	62.4	50.3	11.3	1.4	37.8

Table 2: Results on eight NLP tasks across the fully-supervised, few-shot (16-shot) and zero-shot settings. For the 16-shot setting, we provide the mean and standard deviation across three different random seeds. Scores that are marked with an underline signify the best results among all methods.

Settings. We test the above methods in full-supervised, few-shot and zero-shot experiments, we assign different settings, respectively: (1) **Full-supervised setting:** We use full trainsets to train the PLMs and as neighbors to retrieve. (2) **Few-shot setting:** We follow LM-BFF (Gao et al., 2021) to conduct 16-shot experiment and test the average performance with a fixed set of seeds $\mathcal{S}_{\text{seed}}$, across three different sampled $\mathcal{D}_{\text{train}}$ for each task. In this setting, we use the few-shot training set as k-NN neighbors to retrieve. (3) **Zero-shot setting:** We directly evaluate the vanilla FT and UNION-INF on the test set **without training**. As for UNION-ALL, we take the prompt tuning (Gao et al., 2021) to tag the pseudo labels on **unlabeled** trainsets and apply untrained k-NN in the training and inference.

4.3 Hyper-parameter Settings

We report the hyper-parameters in Table 3. For the GLUE and SuperGLUE datasets, we follow LM-BFF[2] to construct templates and verbalizer for prompt-tuning. While for RE datastes, we follow KnowPrompt (Chen et al., 2021) to construct templates and verbalizer. We utilize Pytorch to conduct experiments with 1 Nvidia 3090 GPUs. We used the AdamW optimizer for all optimizations, with a linear warmup of the learning ratefollowed by a linear decay over the remainder of the training. The hyper-parameter settings used in our experiments are listed below.

Hyper-parameter	Value
maximum sequence length	{128, 256}
max training step	1000
evaluation step	100
learning rate	{1e-5, 2e-5, 5e-5}
batch size	8
gradient accumulation step	{2, 4, 8}
adam epsilon	1e-8
k	{16, 32, 128}
λ	{0.1 : .1 : 0.9}
τ	{0.01, 0.1, 1, 10}

Table 3: Hyper-parameter settings.

[2] https://github.com/princeton-nlp/LM-BFF

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 889-897, Harbin, China, August 3 - 5, 2023.

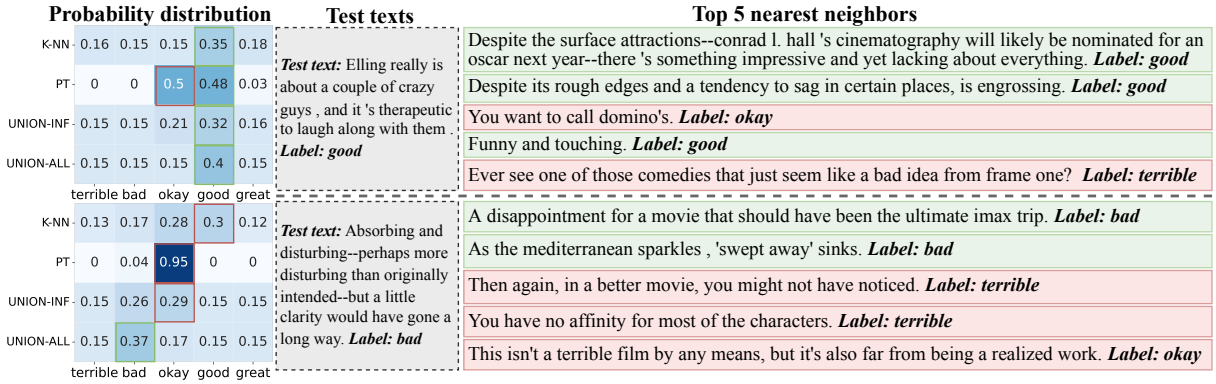


Figure 3: Case analysis to show how k-NN benefits the prediction of PLMs. We illustrate the test texts, the predicted probability distribution, and the top-5 nearest neighbors from the 16-shot training set of the SST-5 dataset.

4.4 Main Results

k-NN features result in performance gains. We compare the specific results with baseline models and provide comprehensive insights of k-NN on different paradigms and different settings. The results as shown in Table 1. Leverage k-NN features results in performance gains in both few-shot and fully-supervised settings. In the zero-shot setting, PT-based methods outperform FT-based and k-NN features further enhance the performance of PT-based methods, which demonstrates that it is flexible and general to integrate k-NN for PLMs.

Calibrating training vs. Incorporating into inference. It is necessary to study the different application scenarios of incorporating k-NN during the training and testing phases. From Table 2, we observe the following: (1) Leveraging k-NN during the test phrase is especially helpful for the zero-shot setting. While **UNION-ALL** performs worse due to the noise brought from the pseudo-labels on unsupervised data. (2) **UNION-INF** is not doing as well in the fully-supervised and few-shot setting. In contrast, **UNION-ALL** outperforms **UNION-INF** in these settings, especially in the few-shot setting. These findings reveal to us the applicable scenarios of incorporating k-NN and inspire further studies to utilize k-NN classifier more practically for efficient NLP.

4.5 Analysis

Q1: How does the lazy learner benefit eager learner? To further understand how does the *lazy learner* (k-NN) benefit the *eager learner* (PLM), we manually check cases in which k-NN, PT, UNION-INF and UNION-ALL produce different results. As shown in the example of the upper row of Figure 3, k-NN and UNION-ALL predict correctly when PT fails. This result is because UNION-ALL produces a more confident probability for the correct class via calibrating the attention on the easy vs. hard examples identified by the k-NN classifier. Note that the bottom row shows that UNION-ALL predicts correctly even when k-NN predicts wrongly, possibly due to the robustness of k-NN calibration.

Q2: Does the similarity metric matter? In the above experiments, we mainly utilize negative $L2$ distance to measure the similarity between the query $\mathbf{q}$ and the instance representation of the data store. It is intuitive to estimate the impact of different similarity metrics, such as cosine similarity. Thus, we present the performance of UNION-ALL using both metrics with the same hyperparameters as below.

Similarity Metric	$L2$	cos
16-shot SST-5 (%)	**43.7**	42.8
16-shot TREC (%)	**90.0**	89.4
16-shot QNLI (%)	**58.1**	57.2

We can find that UNION-ALL with cosine distance achieves nearly the same performance as those trained with $L2$, revealing that our UNION-ALL is robust to the similarity metric.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 889-897, Harbin, China, August 3 - 5, 2023.

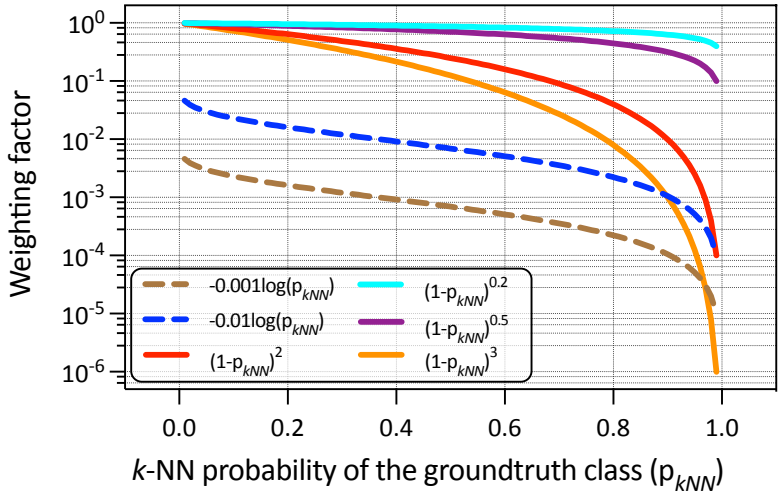


Figure 4: Comparison between the modulating factors NLL and Focal.

Q3: How dose the modulating factor $f(p_{k\mathbf{NN}})$ works? Since we adopt focal loss (Focal) as the modulating factor for main experiments, we further explore other functions as modulating factors, such as negative log-likelihood (NLL). As shown in Figure 4, we visualize two modulating factors with different settings of α and γ, where α donates a scalar that represent the proportion of the term of NLL, and γ is the exponential coefficient for Focal. We can find that NLL and Focal produce large weights for the misclassified examples, demonstrating the diversity of modulating factor selection.

5 Limitations

We only explore leveraging the training data for k-NN search, while various external domain data are also suitable for k-nearest neighbor retrieval. Moreover, incorporating k-NN also faces the following limitations: (1) the requirement of a large memory for retrieval; (2) hyper-parameters (such as λ and α) used for retrieval have an impact on the performance of model training; (3) if the number of nearest neighbors k is too large, it will also affect the efficiency.

6 Conclusion and Future Work

In this paper, we propose a novel method to enhance PLM-based classifiers using k-NN. Specifically, we introduce a calibration process and linear interpolation of inference phrases to effectively integrate k-NN into the training pipeline. To evaluate the effectiveness of our approach, we conduct a comprehensive and in-depth analysis of the role of k-NN in various NLU tasks and tuning paradigms. Our results demonstrate that the integration of k-NN is flexible and can significantly enhance the performance of large models. Future work should explore the combination of k-NN and LLMs such as (1) Inject external knowledge into the LLMs with k-NN. Specifically, k-NN can be used to retrieve relevant knowledge from an external database during the reasoning process, which can help correct errors and reduce the prevalence of gibberish output and factual errors that are common in LLMs. (2) Retrieve contextual information to enhance LLMs. k-NN algorithms can automatically retrieve relevant information based on the input sentence, such as instructions or other relevant context. (3) Augment the training data for LLMs. k-NN is a powerful tool for identifying similar instances in a large dataset, which can help overcome the limitations of data scarcity and improve the performance LLMs.

References

Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of ACL 2020*.

Oren Boiman, Eli Shechtman, and Michal Irani. 2008. In defense of nearest-neighbor based image classification. pages 1–8. IEEE.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS 2020*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *CoRR*, abs/2104.07650.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*.

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

Jerome H Friedman. 2017. *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of ACL*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2018. Search engine guided neural machine translation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5133–5140. AAAI Press.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proc. of EMNLP*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of SemEval*, pages 33–38.

Nora Kassner and Hinrich Schütze. 2020. Bert-knn: Adding a knn search component to pretrained language models for better QA. In *Findings of EMNLP*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 889-897, Harbin, China, August 3 - 5, 2023.

Linyang Li, Demin Song, Ruotian Ma, Xipeng Qiu, and Xuanjing Huang. 2021. KNN-BERT: fine-tuning pre-trained models with KNN classifier. *CoRR*, abs/2110.02523.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

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

Yuxian Meng, Shi Zong, Xiaoya Li, Xiaofei Sun, Tianwei Zhang, Fei Wu, and Jiwei Li. 2021. GNN-LM: language modeling based on global contexts via GNN. *CoRR*, abs/2110.08743.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of ACL*.

Emin Orhan. 2018. A simple cache model for image recognition. 31:10107–10116.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. *CoRR*, abs/2205.13792.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*.

Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. 2019. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Ningyu Zhang, Xin Xie, Xiang Chen, Shumin Deng, Chuanqi Tan, Fei Huang, Xu Cheng, and Huajun Chen. 2022. Reasoning through memorization: Nearest neighbor knowledge graph embeddings. *CoRR*, abs/2201.05575.

Adder Encoder for Pre-trained Language Model

Jianbang Ding[1]*, **Suiyun Zhang**[1], **Linlin Li**[2]
[1]Huawei Technologies Co., Ltd
[2]Huawei Noah's Ark Lab
{dingjianbang1, zhangsuiyun, lynn.lilinlin}@huawei.com

Abstract

BERT, a pre-trained language model entirely based on attention, has proven to be highly performant for many natural language understanding tasks. However, pre-trained language models (PLMs) are often computationally expensive and can hardly be implemented with limited resources. To reduce energy burden, we introduce adder operations into the Transformer encoder and propose a novel AdderBERT with powerful representation capability. Moreover, we adopt mapping-based distillation to further improve its energy efficiency with an assured performance. Empirical results demonstrate that AddderBERT$_6$ achieves highly competitive performance against that of its teacher BERT$_{\mathrm{BASE}}$ on the GLUE benchmark while obtaining a 4.9x reduction in energy consumption.

1 Introduction

The last five years have seen great success achieved by large-scale pre-trained language models, such as BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), and GPT3 (Brown et al., 2020). By modeling long-distance dependencies based on self-attention, they can learn powerful language representations from the unlabeled corpus.

While these models lead to significant improvement on many downstream tasks (eg., the GLUE benchmark (Wang et al., 2019)), the growing computation costs have impaired their deployment, especially on limited-resource devices such as mobile phones, AR glasses, and smartwatch. Since attending to all tokens yields a complexity of $O(n^2)$ with respect to sequence length, prior works aim to investigate efficient Transformers with lower complexity. Kitaev et al. (2020) replaces dot-product attention with one using locality-sensitive hashing. Wang et al. (2020) decomposes the original attention into multiple smaller ones by linear projections. However, they can only solve the problem partway, for the consumption except self-attention has not been changed in the encoder.

Various attempts also focus on model compression techniques, including quantization (Gong et al., 2014), weights pruning (Han et al., 2015), and knowledge distillation (KD) (Romero et al., 2015). As one of the most popular methods, KD aims to transfer knowledge from a large teacher network to a small student network, employed by DistilBERT (Sanh et al., 2019), BERT-PKD (Sun et al., 2019), TinyBERT (Jiao et al., 2020), and FastBERT (Liu et al., 2020). Beyond these methods, Chen et al. (2020) proposed Adder Neural Network (AdderNet), which replaced massive multiplications with cheaper additions to reduce computation costs, and achieved better performance on the ImageNet dataset compared to CNNs. Then researchers attempt to build efficient deep-learning models based on AdderNet for computer vision tasks (Xu et al., 2020; Shu et al., 2021). Inspired by this, it is an interesting idea to investigate the feasibility of replacing multiplications with additions in pre-trained language models like BERT.

In this paper, we first present AdderBERT, a pre-trained model consisting of several adder encoders, in which key modules including multi-head attention and feed-forward network are implemented with cheaper adder operations. As shown in Figure 1, it also has a unique mapping-based distillation that could make it to be more energy-efficient with an assured performance. Finally, we conduct full experi-

*Corresponding author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 898-905, Harbin, China, August 3 - 5, 2023.

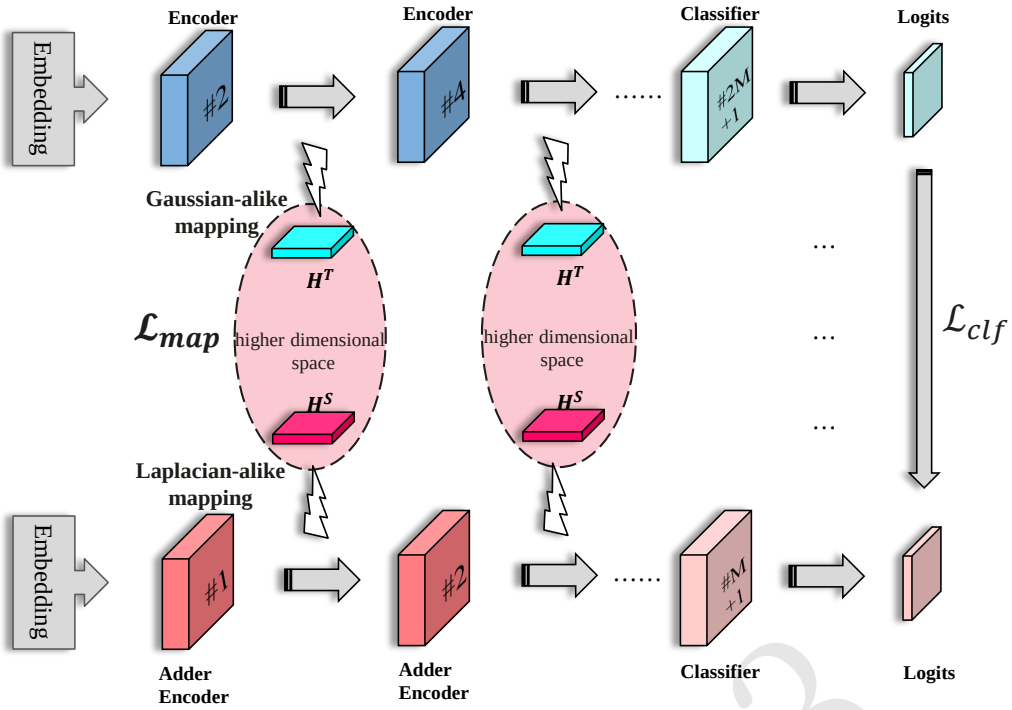


Figure 1: Depiction of AdderBERT learning. AdderBERT implements the encoder block with cheaper adder operations, and it has a unique mapping-based distillation. $\mathbf{H}^S$ and $\mathbf{H}^T$ are the hidden states of the student and teacher networks, respectively. M denotes the number of adder encoders.

ments on the GLUE benchmark. Empirical results demonstrate that our method can achieve comparable performance with the baselines in much lower energy consumption.

The contributions are summarized as follows:

- We propose AdderBERT, which introduces adder operations into the mechanism of self-attention and feed-forward network.
- We adopt a novel mapping-based distillation to encourage that linguistic knowledge can be adequately transferred from the teacher network to AdderBERT.
- Experimental results show that AddderBERT$_6$ can achieve highly competitive performance against that of its teacher BERT$_{\text{BASE}}$ on the GLUE benchmark while obtaining a 4.9x reduction in energy consumption.

2 Preliminary

In this section, we revisit the related works including AdderNet and knowledge distillation. We are motivated by them to design AdderBERT.

2.1 Adder Neural Networks (AdderNet)

Denote the input feature as $\mathbf{X} \in \mathbb{R}^{h \times w \times c_{in}}$, in which h and w are the height and width of the feature map, respectively. Consider a filter $\mathbf{W} \in \mathbb{R}^{d \times d \times c_{in} \times c_{out}}$ in an arbitrary layer of AdderNet, where d is the kernel size, c_{in} and c_{out} are the number of input channels and output channels, respectively. The original adder operation is defined as:

$$\mathbf{Y}(m,n,v) = -\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{u=1}^{c_{in}} \left|\mathbf{X}(m+i,n+j,u) - \mathbf{W}(i,j,u,v)\right|, \quad (1)$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 898-905, Harbin, China, August 3 - 5, 2023.

where $|\cdot|$ is the absolute value function. m and n are the spatial locations of features. v denotes the index of output channels. Given that Equation 1 has been proven to be used to replace the traditional convolution operation, it is an interesting idea to transport this success of CNNs to PTMs.

2.2 Knowledge Distillation (KD)

As one of the most popular compression techniques, KD was used to help a small student network S mimic the behavior of a large teacher network T for better performance. Given f^T and f^S represent the mapping functions of teacher and student networks, respectively. The student network can be optimized with the following objective function:

$$\mathcal{L}_{\text{KD}} = \sum_{x \in \Omega} \mathcal{L}(f^S(x), f^T(x)), \tag{2}$$

where $L(\cdot)$ is an arbitrary loss function, x is the input sequence and Ω denotes the training dataset, $f^S(x)$ and $f^T(x)$ are the outputs of student network and teacher network, respectively. Based on Equation 2, we adopt a unique kernel-based distillation to encourage that linguistic knowledge can be adequately transferred from the teacher network to the student AdderBERT.

3 Method

This section describes our proposed AdderBERT as well as its training method. Concisely, AdderBERT implements the encoder block with cheaper adder operations, and it takes advantage of mapping-based distillation to be better in performance and efficient in energy.

3.1 Adder Encoder

Given that linear transformation is equivalent to 1×1 convolution with fixed input size in mathematical, in this paper, the adder operation can be redefined as:

$$\mathbf{Y}(l, v) = -\sum_{u=1}^{d_{embd}} |\mathbf{X}(l, u) - \mathbf{W}(u, v)| = \mathbf{X} \oplus \mathbf{W}, \tag{3}$$

where l is the sequence length, d_{embd} is the dimension of embedding, and $\oplus$ denotes the adder operation between matrices.

Following the original Transformer (Vaswani et al., 2017), We first consider creating output queries $\mathbf{Q} \in \mathbb{R}^{l\times d_q}$, keys $\mathbf{K} \in \mathbb{R}^{l\times d_k}$, and values $\mathbf{V} \in \mathbb{R}^{l\times d_v}$ by weight matrices $\mathbf{W}_Q \in \mathbb{R}^{d_{embd}\times d_q}$, $\mathbf{W}_K \in \mathbb{R}^{d_{embd}\times d_k}$, $\mathbf{W}_V \in \mathbb{R}^{d_{embd}\times d_v}$ in the projection layer of a single-head self-attention. d_q, d_k, d_v is the dimension of queries, keys, and values, respectively. We employ Equation 3 to measure the ℓ_1-distance between embedding and the weight matrices as:

$$\mathbf{Q} = \mathcal{LN}(\mathbf{X} \oplus \mathbf{W}_Q), \quad \mathbf{K} = \mathcal{LN}(\mathbf{X} \oplus \mathbf{W}_K), \quad \mathbf{V} = \mathcal{LN}(\mathbf{X} \oplus \mathbf{W}_V), \tag{4}$$

where $\mathbf{X} \in \mathbb{R}^{l\times d_{embd}}$ is the input embedding and $\mathcal{LN}(\cdot)$ denotes layer normalization (Ba et al., 2016). Chen et al. (2020) first indicated that the output values of the adder operation should be followed by batch normalization. We also apply layer normalization to stabilize the hidden state dynamics for better learning. Similarly, Equation 3 can be easily modified for a batch matrix-matrix product to realize self-attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q} \oplus \mathbf{K}^T}{\sqrt{d_k}}) \oplus \mathbf{V}, \tag{5}$$

where $\text{softmax}(\cdot)$ is the normalized exponential function and d_k is used for scaling. The attention matrix is calculated from the similarity of $\mathbf{Q}$ and $\mathbf{K}$ by adder operation and acts as the weighted sum factor to $\mathbf{V}$ to get the final output. Multi-head self-attention concatenated different heads from different representing subspaces as follows:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 898–905, Harbin, China, August 3 – 5, 2023.

$$\text{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathcal{LN}(\text{Concat}(\text{head}_1,...,\text{head}_n) \oplus \mathbf{W}_O), \tag{6}$$

where head_i is the i-th attention head obtained by Equation 5 and n is the number of heads. Then we use $\mathbf{W}_O$ to realize an adder projection for dimensional transformation followed by layer normalization. Finally, the feed-forward network can also be reformulated as:

$$\text{FFN}(X) = \mathcal{LN}(\text{ReLU}(\mathbf{X} \oplus \mathbf{W}_1) \oplus \mathbf{W}_2). \tag{7}$$

The new FFN consists of two adder linear transformations, one ReLU activation, and one layer normalization.

3.2 Mapping-based Distillation

Since the basic calculation paradigm of AdderBERT is completely different from that of the original BERT, we adopt a novel mapping-based distillation to adequately transport linguistic knowledge from Teacher BERT to student AdderBERT.

Specifically, we distill the output of each encoder block, and the objective function is as follows:

$$\mathcal{L}_{map} = \text{MSE}(\mathbf{H}^S, \mathbf{H}^T), \tag{8}$$

where $\mathbf{H}^S$ and $\mathbf{H}^T$ are the hidden states of the student and teacher networks, respectively. $\text{MSE}(\cdot)$ denotes the mean square error loss function. As discussed in AdderNet (Chen et al., 2020), the weight distribution in a well-trained ANN is Laplacian distribution rather than Gaussian distribution. Thus we attempt to map the inputs and weights to a higher dimensional space to minimize the distribution gap between $\mathbf{H}^S$ and $\mathbf{H}^T$.

Given $\{\mathbf{X}^S, \mathbf{W}_1^S, \mathbf{W}_2^S\}$, $\{\mathbf{X}^T, \mathbf{W}_1^T, \mathbf{W}_2^T\}$ are the inputs and weights of the FFN of the student and teacher network, respectively. During the distillation process, we transform the hidden states by feature mapping as follows:

$$\begin{aligned} \mathbf{H}^S &= k_1 < \mathbf{X}^S, \mathbf{W}_1^S, \mathbf{W}_2^S >= e^{-\frac{\mathbf{x}^S \oplus \mathbf{w}_1^S \oplus \mathbf{w}_2^S}{\sigma_s}}, \\ \mathbf{H}^T &= k_2 < \mathbf{X}^T, \mathbf{W}_1^T, \mathbf{W}_2^T >= e^{-\frac{\mathbf{x}^T \mathbf{w}_1^T \mathbf{w}_2^T}{2\sigma_t^2}}, \end{aligned} \tag{9}$$

where σ_s and σ_t are two learnable smoothing factors. $k_1 < \cdot >$ is a designed Laplacian-alike kernel that takes the adder operation of two matrices, while $k_2 < \cdot >$ is a Gaussian-alike kernel. After applying Equation 9, the inputs and weights are mapped into a higher dimensional space, thus we can calculate the hidden states by the new smoothing representation.

We also use the cross-entropy loss for classifier distillation $\mathcal{L}_{clf}$ as in previous work (Hinton et al., 2015). Then the final loss function is defined as:

$$\mathcal{L}_{model} = \alpha\mathcal{L}_{map} + \mathcal{L}_{clf} = \sum_{x\in\Omega}(\sum_{m=1}^{M} \alpha\mathcal{L}_{map}^m(x) + \mathcal{L}_{clf}(x)), \tag{10}$$

where M is the number of encoder blocks of AdderBERT, and m denotes the m-th block. α is the hyper-parameter for seeking the balance between $\mathcal{L}_{map}$ and $\mathcal{L}_{clf}$.

4 Experiment

In this section, we verify the effectiveness of AdderBERT on three tasks with different model settings.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 898–905, Harbin, China, August 3 – 5, 2023.

4.1 Datasets

We evaluate our method on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019). For Sentiment classification, we test on CoLA (Warstadt et al., 2019), SST-2 (Socher et al., 2013). For similarity matching, we conduct on QQP [1], MRPC (Dolan and Brockett, 2005), and STS-B (Cer et al., 2017). For language inference, we use MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), WNLI (Levesque et al., 2012) and RTE (Bentivogli et al., 2009).

4.2 AdderBERT Settings

For a fair comparison, We build AdderBERT$_{12}$ with the same configuration as the original BERT$_{\mathrm{BASE}}$ (the number of layers is 12, the hidden size is 768, the feed-forward size is 3072, the number of heads is 12). BERT$_{\mathrm{BASE}}$ uses pre-trained parameters released by Google, and we train AdderBERT$_{12}$ from scratch with the same pre-training settings. We fine-tune them both using the AdaMod (Ding et al., 2019) optimizer for better performance, the learning rate is set to 2e-5, and the batch size is set to 32. We then select the model with the best accuracy in 3 epochs.

We also use the fine-tuned BERT$_{\mathrm{BASE}}$ as the teacher model and use 6 and 3 layers of AdderBERT as the student models (i.e. AdderBERT$_6$ and AdderBERT$_3$). The student models learn from every 2 and 4 layers of the teacher model, respectively. We increase the learning rate to 5e-5 and distill them for 5 epochs. All the experiments are conducted on NVIDIA Tesla-V100 GPUs.

Given that hyperparameters can exert a great impact on the ultimate result, we report the full details about them. We follow the grid search method until the best-performing parameters are at one of the middle points in the grid. For fine-tuning, we tune over hyperparameters to work well across all tasks about batch size: {8, 16, 32, 64}, the initial learning rate of AdaMod:{2e-5, 3e-5, 5e-5, e-4}, and the number of epochs:{2, 3, 4, 5}.

4.3 Baselines

We compare AdderBERT against two strong baselines as follows:

- **BERT** The 12-layer BERT$_{\mathrm{BASE}}$ model, which was pre-trained on Wiki corpus and released by Google (Devlin et al., 2019).
- **DistilBERT** The most famous distillation version of BERT with 6 layers, which was released by Huggingface (Sanh et al., 2019). In addition, we use the same method to distill the DistilBERT with 3 layers.

4.4 Experiments on GLUE

We submitted our model predictions to the official GLUE evaluation server to get results on the test data, as reported in Table 1. Note that values in both models are 32-bit floating numbers and the energy consumptions for a 32-bit multiplication and addition are 3.7pJ and 0.9pJ, respectively (Dally, 2015). The original BERT$_{\mathrm{BASE}}$ achieves a 79.7 score on average with 11.27B multiplications and 11.27B additions, and AdderBERT$_{12}$ achieves a 79.0 score with 0.31B multiplications and 22.23B additions. By replacing massive multiplications with additions, our proposed model obtains about a 2.5x reduction in energy consumption from 51.8BpJ to 21.1BpJ at the cost of a little performance loss (0.7 drops relative to BERT$_{\mathrm{BASE}}$ on the average score). This demonstrates that AdderBERT performs a powerful representation capacity like BERT even with few multiplications.

We then evaluate the distillation versions of our model against the strong KD baselines, respectively. For 6 layers, DistilBERT$_6$ achieves a 75.8 score on average with 5.64B multiplications and 5.64B additions while AdderBERT-6 achieves a higher 79.6 score with 0.16B multiplications and 11.12B additions. With mapping-based distillation, our proposed model AdderBERT-6 significantly outperforms the baseline DistilBERT$_6$ by a margin of 3.8 on average and obtains a 2.5x reduction in energy consumption as well. Compared to BERT$_{\mathrm{BASE}}$, AdderBERT$_6$ is in much lower energy consumption (4.9x reduction) while maintaining competitive performance (79.7 vs 79.6). This indicates that our proposed KD method

[1] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 898-905, Harbin, China, August 3 - 5, 2023.

Model	#Mul.	#Add.	Energy (pJ)	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg
$\text{BERT}_{\text{BASE}}$ (T)	11.27B	11.27B	51.8B	83.8/83.1	71.0	90.7	93.9	52.5	85.5	89.1	68.1	**79.7**
AdderBERT_{12}	0.31B	22.23B	**21.1B**	84.3/83.4	70.4	90.9	92.5	50.7	83.7	89.0	65.9	79.0
DistilBERT_6	5.64B	5.64B	25.9B	82.1/81.3	69.9	88.9	92.4	47.4	76.2	88.1	56.3	75.8
AdderBERT_6	0.16B	11.12B	**10.6B**	84.9/83.1	71.3	91.2	93.8	51.3	83.2	87.9	69.5	**79.6**
DistilBERT_3	2.82B	2.82B	13.0B	73.4/72.9	66.0	81.3	85.6	27.5	75.1	80.2	61.0	69.2
AdderBERT_3	0.08B	5.56B	**5.3B**	80.7/80.9	68.6	88.0	90.7	45.9	79.3	87.2	65.2	**76.3**

Table 1: Results from the GLUE test server. The best results for each group are in-bold. All models are learned in a single-task manner. The energy consumption is calculated from the number of multiplications and additions, respectively. Nothing that T denotes the teacher model, and all the 3-layers and 6-layer models are distilled from it while AdderBERT_{12} is undistilled.

can adequately transport linguistic knowledge from the teacher model to the student model. For 3 layers, AdderBERT_3 is consistently better than DistilBERT_3 (a large improvement of 8.1 on average), especially on the challenging CoLA dataset, and it only consumes less than one-tenth of the energy of the teacher model. In conclusion, empirical results validate our motivation that AdderBERT combines the advantage of both BERT and AdderNet, that is, it could obtain comparable results with the teacher model but substantially reduce the energy burden.

4.5 Ablation Study

We further investigate the effectiveness of different distillation objectives on AdderBERT learning. The baselines include without mapping-based distillation (w/o map) or classification distillation (w/o clf), respectively. The results are summarized in Table 2. We can find the performance without mapping-based distillation drops significantly from 76.9 to 71.8, which demonstrates that our proposed method plays the most important role of the two objectives. The reason for the significant drop lies in the distribution gap between H^S and H^T. Linguistic knowledge is hard to transport across completely different representations.

Model	MNLI-m	MNLI-mm	MRPC	CoLA	Avg
AdderBERT_6	84.3	83.4	89.0	50.7	76.9
w/o map	80.5	77.8	84.3	44.6	71.8
w/o clf	82.0	79.3	88.6	46.9	74.2

Table 2: Ablation studies of different distillation objectives in the AdderBERT learning. The results are validated on the dev set.

5 Conclusion

In this paper, we propose an energy-efficient version of BERT, called AdderBERT. Specifically, AdderBERT consists of several adder encoders implemented by cheap addition operations but has a powerful representation capacity. It adopts a unique mapping-based distillation method to narrow the gap in feature distribution between the teacher and student model. Empirical results on the GLUE benchmark demonstrate that our method can achieve highly competitive performance to the teacher $\text{BERT}_{\text{BASE}}$ while reducing energy consumption significantly.

Acknowledgements

We are grateful to Yunhe Wang, and Yixing Xu for their helpful discussions. This work is supported by Huawei Technologies Co., Ltd.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC*. NIST.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval@ACL*, pages 1–14. Association for Computational Linguistics.

Hanting Chen, Yunhe Wang, Chunjing Xu, Boxin Shi, Chao Xu, Qi Tian, and Chang Xu. 2020. Addernet: Do we really need multiplications in deep learning? In *CVPR*, pages 1465–1474. Computer Vision Foundation / IEEE.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*. OpenReview.net.

B. Dally. 2015. High-performance hardware for machine learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Jianbang Ding, Xuancheng Ren, Ruixuan Luo, and Xu Sun. 2019. An adaptive and momental bound method for stochastic learning. *CoRR*, abs/1910.12249.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *IWP@IJCNLP*. Asian Federation of Natural Language Processing.

Yunchao Gong, Liu Liu, Ming Yang, and Lubomir D. Bourdev. 2014. Compressing deep convolutional networks using vector quantization. *CoRR*, abs/1412.6115.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural network. In *NIPS*, pages 1135–1143.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174. Association for Computational Linguistics.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *ICLR*. OpenReview.net.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *KR*. AAAI Press.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. 2020. Fastbert: a self-distilling BERT with adaptive inference time. In *ACL*, pages 6035–6044. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392. The Association for Computational Linguistics.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *ICLR (Poster)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Han Shu, Jiahao Wang, Hanting Chen, Lin Li, Yujiu Yang, and Yunhe Wang. 2021. Adder attention for vision transformer. In *NeurIPS*, pages 19899–19909.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642. ACL.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 898-905, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *EMNLP/IJCNLP (1)*, pages 4322–4331. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR (Poster)*. OpenReview.net.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Trans. Assoc. Comput. Linguistics*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pages 1112–1122. Association for Computational Linguistics.

Yixing Xu, Chang Xu, Xinghao Chen, Wei Zhang, Chunjing Xu, and Yunhe Wang. 2020. Kernel based progressive distillation for adder neural networks. In *NeurIPS*.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 898-905, Harbin, China, August 3 - 5, 2023.

FinBART: A Pre-trained Seq2seq Language Model for Chinese Financial Tasks

Hongyuan Dong[†], **Wanxiang Che**[*,†], **Xiaoyu He**[‡], **Guidong Zheng**[‡], **Junjie Wen**[‡]
[†]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
[‡]China Merchants Bank AILAB
{hydong, car}@ir.hit.edu.cn
{hexiaoyu42, zhengguidong, wenjunjieee}@cmbchina.com

Abstract

Pretrained language models are making a more profound impact on our lives than ever before. They exhibit promising performance on a variety of general domain Natural Language Processing (NLP) tasks. However, few work focuses on Chinese financial NLP tasks, which comprise a significant portion of social communication. To this end, we propose FinBART, a pretrained seq2seq language model for Chinese financial communication tasks. Experiments show that FinBART outperforms baseline models on a series of downstream tasks including text classification, sequence labeling and text generation. We further pretrain the model on customer service corpora, and results show that our model outperforms baseline models and achieves promising performance on various real world customer service text mining tasks.

1 Introduction

From making investment decisions to managing personal finances, financial text data is crucial in helping individuals and organizations stay up-to-date with the latest financial trends. With the development of information technology, the volume of available financial information looms so large that the use of machine learning methods to facilitate financial text processing becomes increasingly important. However, financial text data differs from general domain data as it contains a large number of technical terms and financial concepts, which require abundant expert knowledge to analyze and annotate. As a result, the application of powerful deep learning technology in financial text processing tasks is limited.

In this work, we apply the fast-developing pretraining techniques to Chinese financial domain. Although a number of works attempt to adapt transformer-based [1] pretrained language models to financial text [2, 3, 4], few studies focus on the application on Chinese financial text data. What's worse, these works adopt transformer encoder architecture, leading to limited ability in modeling the dependency between future tokens and poor performance in text generation tasks. To this end, we propose FinBART, which is an encoder-decoder transformer language model pretrained on Chinese financial corpora. FinBART extracts bi-directional semantic information with its encoder, and generates text with the decoder part. In this way, FinBART is capable of tackling both understanding and generation tasks in Chinese financial domain.

In real world scenarios of the financial industry, customer service is mainly conducted through text interaction. Therefore, there is a great demand for text analysis tools. Customer service corpora differ from financial ones for their sample text is highly colloquial. Although FinBART learns professional knowledge from the financial field, it cannot handle customer service tasks as well as financial ones without adaptation. To this end, we further pretrain FinBART model with customer service corpora, which are comprised of customers' queries about financial services. Further pretraining on customer service corpora adapts FinBART to text interaction scenarios, enabling FinBART to capture accurate semantic information of customer queries.

In summary, the main contributions of this work are listed as follows:

- We pretrain a transformer encoder-decoder language model named FinBART on Chinese financial corpora to infuse professional knowledge into the model. Experiments show that FinBART

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

outperforms baseline models on a variety of Chinese financial tasks, validating the necessity of domain-specific adaption.

- We futher pretrain FinBART on customer service corpora with a newly proposed pretraining objective. Experiments show that further pretrained FinBART-CS achieves higher performance on a series of customer service text mining tasks, facilitating text information filtering and processing in the customer service field.

2 Related Works

Unsupervised pretraining technique is regarded as one of the most significant breakthroughs in recent Natural Language Processing (NLP) research. Taking advantage of unsupervised pretraining methods, Pretrained Language Models (PLMs) achieve remarkable performance on a variety of natural language understanding and generation tasks. Decoder-only language models, represented by GPT [5, 6], possess a remarkable capability for text generation, but lack language understanding ability because they model semantic information in a causual way. Encoder-only language models like BERT [7], ELECTRA [8] and RoBERTa [9] are proposed to introduce bidirenctional semantic information to obtain more representative word embeddings for downstream tasks. However, Encoder-only models cannot model the dependency between `[MASK]` tokens and therefore suffer from limited generation capability. Encoder-decoder transformer models are more versatile because they extract bidirenctional semantic information with the encoder part and generate text smoothly with their decoders. Representative encoder-decoder models are BART [10] and T5 [11], which adopt different pretraining objectives to gain general language modeling ability.

Although it is not far from trivial to adapt language models pretrained on general corpus to domain-specific downstream tasks, their performance may be limited because of the lack of professional knowledge. This problem looms larger for domains involving large amounts of technical terms and concepts like finance. To this end, researchers propose to pretrain language models on domain-specific corpus to endow models with professional skills for downstream tasks. BioBERT [12] collects large biomedical corpora and pretrains a transformer encoder model to produce contextualized word representation for biomedical text. SciBERT [13] trains a BERT model for scientific NLP tasks with scientific publication corpora. In financial industry, researchers also seek to train domain-specifc language models to facilitate financial text meterials processing. FinBERT [2] designs a series of pretraining objectives for English financial text to train the model more effectively. Mengzi-BERT-fin [14] further trains Chinese BERT model to adapt to Chinese financial text. However, these works are mainly designed for Engligsh natural lnaguage understanding tasks and lack the ability to tackle Chinese NLP tasks involving generation. Our work fills this gap by pretraining a seq2seq transformer-based language model on Chinese financial corpora.

3 Methods

In this section, we introduce the pretraining procudure of the proposed FinBART and FinBART-CS. We choose transformer encoder-decoder architecture as the backbone, reconciling language understanding and generation ability. We use financial corpora and customer service corpora for model pretraining, and then finetune FinBART and FinBART-CS in their domain-specific downstream tasks, respectively. The overall data collecting, model pretraining and downstream task adapting pipeline is shown in Figure 1.

3.1 Copora

3.1.1 Chinese Financial Copora

We purchase a number of Chinese financial corpus data from Datayes, which is a financial technology data provider company. We also crawl a large amount of financial news from news portals and financial websites such as EastMoney[1], Ji Wei Net[2], and CNStocks Net[3]. In total, we collect an amount of 20

[1]https://www.eastmoney.com/
[2]https://m.laoyaoba.com/
[3]https://www.cnstock.com/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

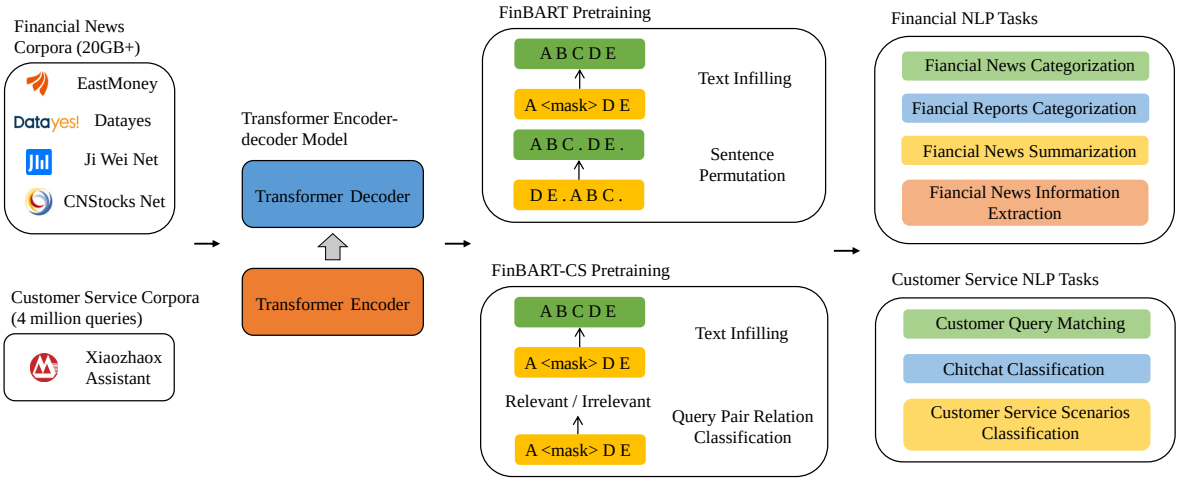


Figure 1: The overall pipeline of data collecting, FinBART and FinBART-CS model pretraining and downstream tasks adaptation.

GB Chinese financail news data, including approximately 12 million pieces of news and 1 billion tokens. These news are published between October 2020 to April 2022.

We further filter the financial new corpus with CCNet [15] toolkit, which extracts high quality datasets from web crawl data. To be specific, we use the LM filtering tool provided in CCNet to compute the perplexity of each financial news sample using a Chinese language model. News samples which score over 12,000 perplexity are filtered out. We also discard news samples which consist of more than 52% digits and puctuations, which are highly probable to be structured data and not expected to appear in the pretraining data.

3.1.2 Customer Service Query Copora

In customer service scenarios, the text content is highly colloquial and the language style and form are distinct from written financial news. To tackle real world customer service text mining tasks, we collect a large amount of customer service query data to further pretrain the model. We collect and mannually compile customer queries, and annotate each query with a primary question. Each query is semantic equivalent to its primary question. In this way, the customer service queries can be clustered into several semantic equivalent question groups, and each cluster is represented with a single primary question. We obtain approximately 4,038,304 customer service queries in total.

3.2 FinBART pretraining

We adopt BART [10] architecture as the model's backbone. BART is an encoder-decoder transformer model. It captures bidirenctional semantic information via self-attention mechanism with its encoder. Global context information is injected to the contextualized word embeddings produced by the encoder, which are then referred to with cross-attention during decoding.

As shown in Figure 2, we choose the combination of text infilling and sentence permutation as the pretraining objectives of our model, which is proven to be the most effective pretraining strategy for BART model [10]. For text infilling objective, we conduct whole word masking on Chinese financial documents to raise the difficulty, preventing the model from degrading to predict the masked word with only its neighbour words. We use Jieba, an open-source Chinese word segmentation toolkit, to segment financial documents into whole Chinese words. We mask a total of thirty percent of tokens and replace the contiguous whole word spans as single `[MASK]` tokens. For sentence permutation objective, we split financial documents into sentences, and shuffle the order of the sentences randomly. The model is trained to recover the sentence order on the decoder side, and is therefore driven to model the semantic information of the whole document.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

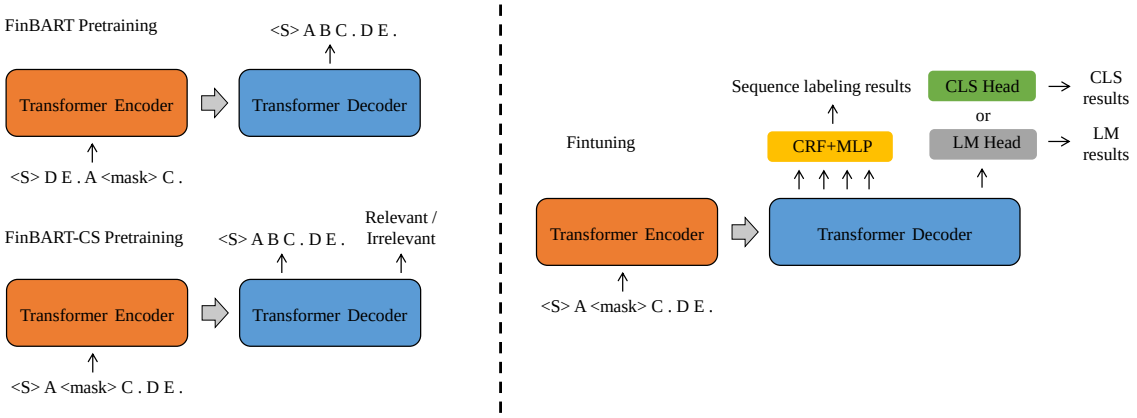


Figure 2: A detailed illustration of the pretraining (left) and finetuning procedure (right) of FinBART and FinBART-CS. "CLS" stands for classfication, and "LM" stands for language modeling. "CRF+MLP" is the abbreviation of conditional random field and multi-layer perceptron.

3.3 FinBART-CS pretraining

Customer service requires a large amount of text interaction, leading to an urgent need of automatic text mining tools. To this end, we futher pretrain FinBART with customer service corpora to facilitate domain-specific text analysis.

Pretraining data construction. We reorganize the customer service corpus into query pairs. To be specifc, we denote a query sample as $d = (\mathbf{x}_d, d^p)$, where $\mathbf{x}_d$ stands for the text content of the query and d^p is its primary question. Each query is of the same meaning as its primary question, and only one query is marked as primary question in its cluster. We denote a cluster with d^p as its primary question as follows:

$$\mathcal{C}_{d^p} = \{d_i | d_i^p = d^p\}. \tag{1}$$

For each query in the corpus $\mathcal{D}$, we sample five other queries pointing to the same primary question randomly from $\mathcal{C}_{d^p}$. These samples are paired with d to form five query pairs of equivalent semantic meaning. We denote the set of paired semantic equivalent query pairs as $\mathcal{D}_d^{eq}$. Similarly, we sample five queries from other clusters to form query pairs of two unrelated queries. We represent the unrelated query pairs as $\mathcal{D}_d^{dif}$. When pairing queries, we set the order of the two queries randomly. In this way, we obtain ten query pairs for each single query in the corpus. We aggeragate each query sample d's corresponding $\mathcal{D}_d^{eq}$ and $\mathcal{D}_d^{dif}$ to form the final pretraining dataset:

$$\mathcal{D}_{paired} = \bigcup_{d \in \mathcal{D}} \mathcal{D}_d^{eq} \cup \bigcup_{d \in \mathcal{D}} \mathcal{D}_d^{dif}. \tag{2}$$

Pretraining objectives. To utilize the knowledge in the customer service corpus effectively, we add a supervised training objective, which is query pair relation classification, in addition to BART's pretraining objective. The pretraining objectives of FinBART-CS is illustrated in Figure 2 (left). We extract the semantic representation of the given query pairs with the FinBART-CS model. Each query pair is represented as the hidden states of the last token extracted from the final decoder layer. An classification head, which is contructed with 2-layer neural network, is then used to map the representation to a binomial distribution indicating whether the pair of queries are of the same meaning. Model parameters are optimized with cross entropy loss between the predicted binomial distribution and ground truth label. We also conduct whole word masking for the text infilling objective, but discard sentence permutation objective because customer service queries are relatively short and not as coherent as written language. As a result, introducing sentence permutation objective cannot not promote the model to capture overall

semantic information, but instead confuses the model and leads to worse quality of semantic representations produced by the encoder model.

4 Experimetns

In this section, we introduce the experimental settings and resources used in the pre-training and downstream task adaptation stages of FinBART and FinBART-CS.

4.1 FinBART

4.1.1 Pretraining Settings

The proposed FinBART is a 12-layer transformer encoder-decoder model with 768-dimensional inner representation. The encoder and decoder of FinBART both consist of 6-layer transformer blocks while conducting different attention mechanism. FinBART adopts WordPiece [16] tokenizer to segment Chinese text.

We train two versions of FinBART, which are training from scratch and continual training. For training from scratch, we initialize model parameters randomly and optimize the parameters with pretraining loss computed on Chinese financial corpora. We set learning rate as 7e-4 with a batch size of 2048 and weight decay of 0.01. We use Adam optmizer for model pretraining with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and warm up the pretraining procedure for the first 10,000 steps. After warming up, the learning rate decays linearly for 60,000 steps to a minimum learning rate of 1e-6. For continual training, we start from bart-base-chinese [17] checkpoint, and futher train the model for 800,000 steps. The learining rate is set as 2e-5, which warms up for the first 10,000 steps and decays for 500,000 steps liniearly.

We implement FinBART pretraining pipeline with Megatron-LM [18] framework. We use data parallel strategy on 8 NVIDIA V100 GPUs to accelerate FinBART pretraining. Mixed precision training is also adopted in the pretraining procedure, where model parameters and garadients are stored in FP16 precision, while the accumulated gradients and parameter optimization are computed in FP32 precision. A gradient upscaling operation is also adopted to alleviate the parameter underflow problem.

4.1.2 Downstream Tasks

We collect four downstream tasks to evaluate FinBART's capabilities in Chinese financial NLP tasks. For each task, we finetune the model on the training set with a learning rate of 1e-5 and a batch size of 8, and evaluate the model's performance on validation set with a fixed interval. When the validation performance does not increase for more than 20 times of valiadation, we stop finetuning and test the model's final performance on the test set.

Financial news categorization. We leave out a small proportion of the Chinese financial news pretraining corpus to evaluate FinBART's natural language understanding ability. These financial news are collected from news portals and financial websites from October 2020 to April 2022. Financial news categorization dataset is a 29-class text classification task, which contains 4,292 training samples, 537 validation samples and 537 testing samples belonging to different industires. The average length of sample texts is 430 words. During finetuning, we extract sentence embedding as the hidden states of the last token, and implement a 2-layer classfication head to transform the representation to classification results. We use categorization accuracy as the evaluation metric.

Financial reports categorization. We collect a financial reports categorization dataset to evaluate model's performance on more professional financial documents. Financial reports categorization dataset is a 41-class text classification task, with a training set of 3,330 samples, a validation set of 416 sampels and a test set of 417 samples. The average length of sample texts is 430 words. We process financial reports categorization in the same way as financial news categorization task and use categorization accuracy as the evaluation metric.

Financial news summarization. We also leave out a small proportion of financial news corpus to test model's text generation ability. For each financial news document, we use the article body as input text, and select its abstraction as the target summarization. Financial news summarization dataset contains

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Model	News Cat Accuracy	Reports Cat Accuracy	News Sum BLEU	DUEE-fin Slot F1
MENGZI-BERT-BASE-FIN [14]	79.14	76.74	–	**81.01**
BART-BASE-CHINESE [17]	78.21	76.98	0.6141	76.27
FINBART (SCRATCH)	**81.19**	79.62	0.6087	76.08
FINBART (CONTINUAL)	80.07	**80.10**	**0.6206**	76.44

Table 1: Raw experiment results of FinBART on four Chinese financial downstream tasks. Bold numbers indicate the best result on the task.

Model	News Cat	Reports Cat	News Sum	DUEE-fin	Average
MENGZI-BERT-BASE-FIN [14]	+1.19	−0.31	–	**+6.21**	–
BART-BASE-CHINESE [17]	0.00	0.00	0.00	0.00	0.00
FINBART (SCRATCH)	**+3.81**	+3.43	−0.88	−0.25	+1.53
FINBART (CONTINUAL)	+2.38	**+4.05**	**+1.06**	+0.22	**+1.93**

Table 2: Normalized relative performance improvement of FinBART over BART-base-chinese baseline on four Chinese financial downstream tasks. The results are given in percentage. Bold numbers indicate the best result on the task.

39,694 trainin samples, 4,962 validation sampels and 4,962 test samples. The average length of the article body is 367 words, and reference summarizations' average length is 82 words. We adopt greedy decoding strategy to generate summarizations, and restrict the vocab to tokens appeared in the original news text. The results are evaluated with BLEU [19] score between the generated text and the reference.

Duee-fin. Duee-fin is a publicly available financial event extraction dataset [20], where each sample contains a Chinese financial news, corresponding event type and argument roles. In total, there are 13 event types and 61 slot labels appearing in Duee-fin dataset. Duee-fin contains 9,425 training samples, 1,518 validation samples and 273,876 test samples. Because we cannot get access to labels of the test set, we sample training, validation and test set randomly from the union of original training and validation sets according to a ratio of 8:1:1. We use the official demo pipeline to preprocess the event, extracting the trigger sentence and its corresponding sequence labels in BIO format. We use the sequence labels to evaluate FinBART's capability in token-level financial NLP tasks. As shown in Figure 2 (right), we implement a linear transformation to map each token's last layer representation to a distribution over sequence labels. A Conditional Random Field (CRF) is then used to model the dependency between sequence labels. We use macro F1 score over all slot types as the evaluation metric.

4.1.3 Main Results

We report the raw experimental results for each task in Table 1. Since each task adopts different metrics, the overall improvement of the proposed models is hard to be quantified. To address this issue, we follow UL2 [21] to use the *normalized relative gain with respect to baselines* as the overall metric. We use BART-base-chinese as the baseline and compute the relative performance improvement of other models. The normalized results are listed in Table 2.

As listed in Table 1, compared with BART-base-chinese [17] trained on generic Chinese corpora and Mengzi-bert-base-fin [14] trained on financial corpora, FinBART not only tackles both financial language understanding and generation tasks, but also achieves better performance on these datasets. For News categorization dataset, the best model FinBART trained from scratch gains 81.19 classification accuracy, which is 3.81% higher than BART-base-chinese baseline and 2.59% higher than Mengzi-bert-base-fin baseline. For Reports categorization dataset, continually trained FinBART achieves the best 80.10 classfication accuracy, outperforming BART-base-chinese with 3.12 accuracy score (4.05%↑) and leading Mengzi-bert-base-fin with 4.38% higher accuracy. For financial news summarization dataset,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Training strategy	Bsz	Lr	# Steps	News Cat	Reports Cat	News Sum	DUEE-fin
–	–	–	–	78.21	76.98	0.6141	76.27
SCRATCH	256	1e-4	1e6	**81.75**	77.70	0.5729	**78.51**
	2048	7e-4	2e5	81.19	79.62	0.6087	76.08
CONTINUAL	256	2e-5	8e5	80.07	80.10	**0.6206**	76.44
	2048	1e-4	8e4	78.21	**80.34**	0.6024	77.94

Table 3: The influence of hyper parameters to FinBART's performance on downstream Chinese financial NLP tasks. The first line shows the performance of BART-base-chinese baseline. Bold number is the best result on the task.

Training strategy	Bsz	Lr	# Steps	News Cat	Reports Cat	News Sum	DUEE-fin	Average
–	–	–	–	78.21	76.98	0.6141	76.27	0.00
SCRATCH	256	1e-4	1e6	**+4.53%**	+0.94%	−6.71%	**+2.94%**	+0.42%
	2048	7e-4	2e5	+3.81%	+3.43%	−0.88%	−0.25%	+1.53%
CONTINUAL	256	2e-5	8e5	+2.38%	+4.05%	**+1.06%**	+0.22%	**+1.93%**
	2048	1e-4	8e4	0.00%	**+4.36%**	−1.91%	+2.19%	1.16%

Table 4: Relative performance gain of FinBART with different hyper parameter settings over BART-base-chinese baseline. The first line shows the performance of BART-base-chinese baseline. Bold number is the best result on the task.

continually trained FinBART achieves the best performance with 0.6206 BLEU score, surpassing BART-base-chinese with a 1.06% margin. Mengzi-bert-base-fin cannot tackle text generation tasks because of its transformer encoder architecture, so its performance on news summarization task is not listed in the table. For Duee-fin task, Mengzi-bert-base-fin achieves the highest 81.01 slot F1 score. We attribute its advantage to its bi-directional language modeling mechanism. Among language models with encoder-decoder architecture, continually trained FinBART performs the best with 76.44 slot F1 score, which is 0.22% higher than BART-base-chinese baseline.

To facilitate comparison, we reorganize the experiment results into relative performance improvement over BART-base-chinese baseline. The reorganized results are shown in Table 2. We show the average relative performance gain in the last column to compare the overall performance of each model. Generally speaking, FinBART models trained on Chinese financial corpora achieve better results than BART-base-chinese baseline, indicating the necessity of adaptation training on domain-specific corpora. FinBART continually trained from BART-base-chinese checkpoint gains the most relative performance improvement, leading BART-base-chinese baseline with a 1.93% higher relative performance averagely on the four Chinese financial NLP tasks. FinBART trained from scratch on Chinese financial corpus outperforms BART-base-chinese baseline with 1.53% higher overall performance, but fails to win BART-base-chinese on 2 out of 4 tasks. We ascribe its unsatisfactory performance to the small size of the pretraining corpus, which results into limited general language modeling ability. Continually trained FinBART contains both general knowledge and financial domain-specific knowledge, and therefore outperforms BART-base-chinese baseline with a larger margin on all four downstream tasks.

4.1.4 The Impact of Hyper Parameters

In Section 4.1.1, we introduce the different sets of hyper parameters when training FinBART from scratch and continually training from BART-base-chinese checkpoint. To understand how hyper parameters influence the model performance, we train FinBART under different hyperparameter settings. We set varying pretraining batch size, learning rate and the number of pretraining steps for pretraining. For FinBART trained from scratch, we train the model with 1e-4 and 7e-4 learning rate respectively, and set batch size and optimization steps accordingly for fair comparison. For FinBART continually trained from BART-base-chinese checkpoint, we use smaller learning rates as suggested for BERT domain adaptation

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

training code. The results are shown in Table 3 and Table 4.

Overall, FinBART models continually trained from BART-base-chinese checkpoint achieve better results than those trained from scratch. We attribute the leading performance of continually trained FinBART models to the large corpora used in the pretraing process. The two stage pretraining (general domain pretraining & financial domain continual pretraining) procudure injects both general knowledge and financial domain-specific knowledge into the model. Therefore, continually trained FinBART models show impressive ability in Chinese financial NLP tasks.

For FinBART models trained from scratch, we find that the combination of large batch size and large learning rate obtains better model performance, which is consistent with the conclusion of RoBERTa [9]. For continually trained FinBART, we set smaller learning rates for domain adaptation training. In this circumstance, the model trained with small batch size and more training steps achieves the best performance.

4.2 FinBART-CS

4.2.1 Pretraining Settings

FinBART-CS shares the same architecture with FinBART, which is a 12-layer encoder-decoder transformer model with 768-dimensional inner representation. We train FinBART-CS with additional customer service corpora from FinBART checkpoint. We set the learning rate as 2e-5 and train the model for 1,000,000 steps with a 10,000 steps warming up and 500,000 steps linear learning rate decay. The minimum learning rate is 1e-6. We use 256 batch size for FinBART-CS pretraining, and optimizes model parameters with Adam optimizer. FinBART-CS's optimization involves two pretraining objectives, which are text infilling and query pair relation classification. We give equal weights to the two objective losses, and optimize FinBART-CS parameters and classification head parameters jointly.

4.2.2 Downstream Tasks

We collect four downstream tasks to evaluate FinBART-CS's capabilities in customer service NLP tasks. For each task, we finetune the model on the training set with a learning rate of 1e-5, and monitor model performance with the validation set to avoid overfitting. We set the early stopping threshold as 20.

BQ Corpus. BQ Corpus is a publicly available customer service query matching dataset [22]. Each sample consists of a pair of customer service queries and a label indicating whether the two queries have the same semantic meaning. We concatenate the query pair and feed it to the model, converting it to a binary text classification task. BQ Corpus dataset contains 88,000 training sampels, 11,000 validation samples and 11,000 testing sampels. The average length of the query pairs is 34. We set the finetuning batch size as 8 and use classification accuracy as the evaluation metric.

Query match. Query match dataset's formulation is akin to that of BQ Corpus. Each sample consists of a pair of customer queries and a binary label indicating whether the query pair shares the same semantic meaning. Query match dataset consists of a 2,967,316-sample training set, a 370,914-sample validation set and a 370,915-sample test set. The average sample text length is 36. We set the finetuning batch size as 32 and use classification accuracy as the evaluation metric.

Chitchat classification. Chitchat classification dataset is a binary text classfication task. Sample text is obtained from real world Xiaozhaox smart assistant customer service scenario. Each sample contains a piece of customer query text and a label indicating whether the query is asking for services or just chitchatting. Chitchat classification dataset contains 102,459 samples, and we split it into training, validation and testing parts randomly according to a ratio of 8:1:1. The average length of the sample text is 14. We set the finetuning batch size as 8 and use classification accuracy as the evaluation metric.

Xiaozhaox categorization. Xiaozhaox categorization is a 1,046-class text categorization task. Sample text is obtained from real world Xiaozhaox smart assistant customer service scenario. Note that although a small proportion of samples of Xiaozhaox categorization dataset are contained in FinBART-CS pretraining corpora, categorization label information is not introduced during pretraining. Each sample belongs to a certain class of 1,046 catrogories. Xiaozhaox categorization dataset contains 1,837,722

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Model	BQ Copus Accuracy	Query Match Accuracy	Chitchat Cls Accuracy	Xiaozhaox Cls Accuracy
MENGZI-BERT-BASE-FIN [14]	92.17	88.38	99.29	65.40
BART-BASE-CHINESE [17]	88.82	87.58	99.07	63.38
FINBART (SCRATCH)	90.75	87.15	99.08	62.08
FINBART (CONTINUAL)	91.65	87.43	99.07	63.63
FINBART-CS (CONTINUAL)	**93.32**	**90.97**	**99.36**	**67.94**

Table 5: Raw experiment results of FinBART-CS on four Chinese financial customer service downstream tasks. Bold numbers indicate the best result on the task.

Model	BQ Copus	Query Match	Chitchat Cls	Xiaozhaox Cls	Average
MENGZI-BERT-BASE-FIN [14]	+3.77%	+0.91%	+0.22%	+3.19%	+2.02%
BART-BASE-CHINESE [17]	0.00	0.00	0.00	0.00	0.00
FINBART (SCRATCH)	+2.17%	−0.49%	+0.01%	−2.05%	−0.09%
FINBART (CONTINUAL)	+3.19%	−0.17%	0.00%	+0.39%	+0.85%
FINBART-CS (CONTINUAL)	**+5.07%**	**+3.87%**	**+0.29%**	**+7.19%**	**+4.11%**

Table 6: Normalized relative performance improvement of FinBART-CS over BART-base-chinese baseline on four Chinese financial downstream tasks. The results are given in percentage. Bold numbers indicate the best result on the task.

training samples, 30,000 valiadation samples and 6,029 testing samples. The average length of the text samples is 13. We set the finetuning batch size as 32 and use classification accuracy as the evaluation metric.

4.2.3 Results

We report the raw experimental results for each task in Table 5. We also reorganize the experiment results into *normalized relative gain with respect to baselines* to compare the model's overall performance. The normalized model performance scores are listed in Tabel 6.

As shown in Table 5 and Table 6, FinBART-CS continually trained on customer service corpus achieves the highest accuracy across all four Chinese financial customer service NLP tasks, and outperforms BART-base-chinese baseline with 5.07%, 3.87%, 0.29% and 7.19% accuracies, respectively. Averagely, continually pretraining on customer service corpus leads to a 4.11% overall performance improvement. Meanwhile, we have following observations based on Table 5 and Table 6:

- FinBART-CS is the only model outperforming Mengzi-bert-base-fin. Mengzi-bert-base-fin adopts transformer encoder-only architecture, which is regarded more suitable for natural language understanding tasks. None of the listed encoder-decoder model gains higher scores on the four datasets except FinBART-CS, indicating the superiority of FinBART-CS in customer service NLP tasks. Compared to Mengzi-bert-base-fin, FinBART-CS learns to understand customer service text data via continual pretraining, making up the lack of bi-directional semantic information.

- Models trained with financial corpora achieve better performance on financial customer service tasks. Continual pertraining on financial corpora injects financial domain-specifc knowledge into the model, facilliating the understanding and processing of financial customer service text. That being said, customer service text data is highly colloquial, and its distribution is different from that of written financial news data. As a result, FinBART trained on both general corpora and financial corpora outperforms FinBART trained from scratch with only financial corpora, which may overfit to the distribution of financial document data.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Model	BQ Copus	Query Match	Chitchat Cls	Xiaozhaox Cls
BART-BASE-CHINESE	88.82	87.58	99.07	63.38
FINBART-CS	**93.32**	**90.97**	**99.36**	**67.94**
FINBART-CS (ABLATION)	91.65	87.43	99.07	63.63

Table 7: FinBART-CS and classfication loss ablated FinBART-CS's performance on four financial customer service NLP tasks. Bold numbers indicate the best result on the task.

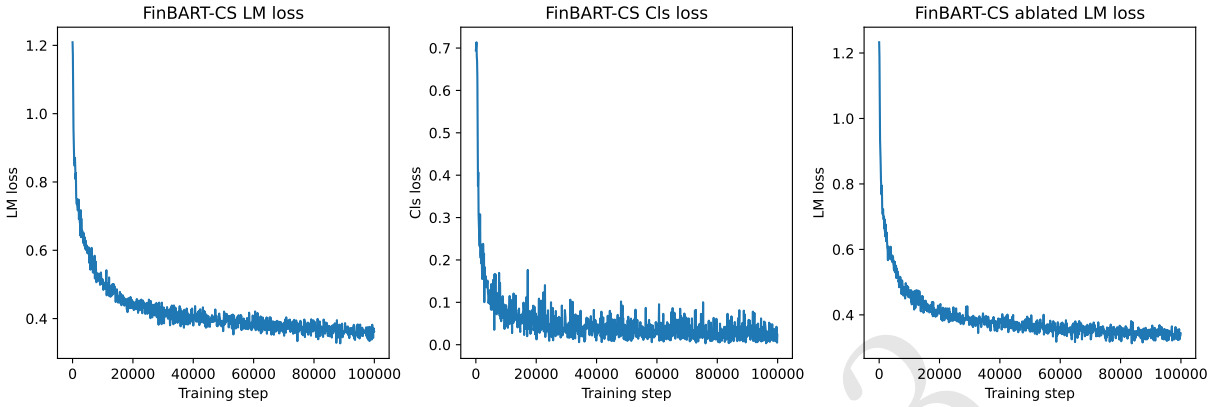


Figure 3: The loss curves of FinBART-CS's LM loss (left), FinBART-CS's Cls loss (mid) and FinBART-CS's ablated LM loss (right). "LM" stands for language modeling, and "Cls" stands for classfication. We curate the first 1e5 steps of the pretraining procedure for clarity.

4.2.4 Ablation Study

To validate the effectiveness of the proposed pretraining objective for FinBART-CS, we conduct abaltion study for the query pair relation classification objective. We train the ablated FinBART-CS with the same set of parameters as FinBART-CS with only query pair relation classification loss removed from the pretraining procedure. Table 7 shows the performance of FinBART and FinBART-CS on four Chinese financial customer service NLP tasks. When query pair relation classification objective is removed, the ablated FinBART-CS suffers from a large performance drop. FinBART-CS outperforms the ablated model on all four downstream tasks, validating the effectiveness of query pair relation classification pretraining objective.

We also investigate how the newly introduced classification loss influences FinBART-CS's pretraining procedure. As shown in Figure 3, the LM (Language Modeling) loss of FinBART-CS converges smoothly regardless of whether the classification loss is introduced. The query pair relation classification loss also converges fast and smoothly during the pretraining process. Supervised signals are injected to the model without hindering the model from learning general language knowledge. The loss curves show the compatibility of LM loss and classification loss, validating the effectiveness of our proposed query pair relation classification pretraining objective.

5 Conclusions

In this work, we propose FinBART and FinBART-CS, which are Chinese encoder-decoder language models pretrained on Chinese financial corpora and customer service corpora, respectively. The proposed models fill the blank of Chinese seq2seq language model for financial and customer service NLP tasks. Experiments on a series of Chinese financial NLP tasks and customer service NLP tasks indicate the promising performance of the proposed models. We also conduct analysis experiments to show the rationality of pretraining hyper parameter selection and the effectiveness of the proposed pretraining objective for FinBART-CS.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

References

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

[2] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519, 2021.

[3] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

[4] Vinicio DeSola, Kevin Hanna, and Pri Nonis. Finbert: pre-trained model on sec filings for financial natural language tasks. *University of California*, 2019.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

[8] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

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

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[13] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, 2019.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

[14] Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*, 2021.

[15] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

[16] Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144, 2016.

[17] Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*, 2021.

[18] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[20] Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. Duee-fin: A large-scale dataset for document-level event extraction. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pages 172–183. Springer, 2022.

[21] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022.

[22] Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 906-917, Harbin, China, August 3 - 5, 2023.

Exploring Accurate and Generic Simile Knowledge from Pre-trained Language Models

Shuhan Zhou[1] **Longxuan Ma**[2] **Yanqiu Shao**[1*]
[1] School of Information Science, Beijing Language and Culture University
[2] Research Center for Social Computing and Information Retrieval,
Faculty of computing, Harbin Institute of Technology
blcu_zsh@163.com lxma@ir.hit.edu.cn yqshao163@163.com

Abstract

A simile is an important linguistic phenomenon in daily communication and an important task in natural language processing (NLP). In recent years, pre-trained language models (PLMs) have achieved great success in NLP since they learn generic knowledge from a large corpus. However, PLMs still have hallucination problems that they could generate unrealistic or context-unrelated information. In this paper, we aim to explore more accurate simile knowledge from PLMs. To this end, we first fine-tune a single model to perform three main simile tasks (recognition, interpretation, and generation). In this way, the model gains a better understanding of the simile knowledge. However, this understanding may be limited by the distribution of the training data. To explore more generic simile knowledge from PLMs, we further add semantic dependency features in three tasks. The semantic dependency feature serves as a global signal and helps the model learn simile knowledge that can be applied to unseen domains. We test with seen and unseen domains after training. Automatic evaluations demonstrate that our method helps the PLMs to explore more accurate and generic simile knowledge for downstream tasks. Our method of exploring more accurate knowledge is not only useful for simile study but also useful for other NLP tasks leveraging knowledge from PLMs. Our code and data will be released on GitHub.

1 Introduction

A simile is a figure of speech that compares two things from different categories (called the tenor and the vehicle) via shared properties (Paul, 1970). A tenor and a vehicle are usually connected with comparator words such as "like" or "as". For example, the sentence "The girl is as pretty as an angel." is a simile where the tenor is "The girl", the vehicle is "an angel", the comparator is "as ... as" and the shared property is "pretty". Simile plays an important role in human language to make utterances more vivid, interesting, and graspable (Zhang et al., 2021), comprehending similes is essential to appreciate the inner connection between different concepts and is useful for other natural language processing (NLP) tasks (Song et al., 2021; He et al., 2022).

In recent years, pre-trained language models (PLMs) have achieved great success in NLP since they learn generic knowledge from a large corpus and could serve as a knowledge base (Devlin et al., 2019; Radford et al., 2019). Considerable attention has been paid to exploring simile knowledge from PLMs to solve downstream simile tasks, such as recognition, interpretation, and generation (Chen et al., 2022; He et al., 2022). However, PLMs are known to suffer from hallucination problems (Shuster et al., 2021; Dziri et al., 2022; Liu et al., 2022), they could generate unrealistic or unfaithful information about the provided source content, which will impact their performance on downstream tasks. For example, when completing the blank in a simile sentence "Are you feeling ill ? You are as __ as a ghost .", a PLM may generate "creepy" instead of the expected shared property "pale".

In this paper, we study how to explore more accurate and generic simile knowledge from PLMs. Specifically, we first train PLMs with three main simile tasks (recognition, interpretation, and generation). In this way, the PLMs can learn the shared semantic feature among different tasks and gain a better understanding of the simile knowledge. However, this understanding may be limited by the distribution

*Corresponding author

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

Metaphor Category	Example	Is a simile?
Noun phrase	The judge is like *an angel*.	Yes
Adjective	The boy has a warm heart.	No
Verbal	He kills the seeds of peace.	No
Adverb-Verb	The child speaks France fluidly.	No
Verbal phrase	Raising little cats is like *taking care of children*.	Yes
Sentence	The man walks into the crowd like *a fish swims into the ocean*.	Yes

Table 1: Different metaphor categories. For similes, we use underline font to show **tenors** and use italic font to show *vehicles*.

of the training data. The performance of the model will drop when applied to unseen domains. To explore more generic simile knowledge, we further add semantic dependency features in the fine-tuning process. The semantic dependency feature serves as a global signal, helps the model learn simile knowledge shared among similar syntax structures, and enhances the model's performance on unseen domains. During tests, we conduct experiments on both seen and unseen test sets to verify the effectiveness of our method. To sum up, our contributions are:

- We propose a novel method to explore more accurate and generic simile knowledge from PLMs.
- We test our model with both seen and unseen test sets. Experimental results demonstrate the effectiveness of our method and we give a detailed analysis of the results.
- Our code and data (including a new manually annotated simile data set) will be released on GitHub[1].

2 Related Work

In this section, we will introduce previous work related to this paper.

2.1 Simile and Metaphor

Metaphor is often used in human language to make speech more vivid and easy to understand (Niculae and Danescu-Niculescu-Mizil, 2014). Bizzoni and Lappin (2018) categorized metaphor into Noun phrases, Adjectives, Verbs, and Multi-word. Li et al. (2022) defined metaphor as Nominal, Verbal (Subject-Verb-Object), Adjective-Noun, and Adverb-Verb. Table 1 shows examples of these categories. The Noun phrase metaphor is usually defined as a simile (Li et al., 2022; He et al., 2022; Chen et al., 2022). In this paper, we not only study the Noun phrase metaphor. Meanwhile, to test whether the trained model performs well on unseen domains, we construct a new test set. In this new test set, the tenor and vehicle can be verbal phrases/sentences that perform a similar role to Noun phrases. The examples of verbal phrases and sentences as simile components are shown in Table 1.

2.2 Tasks in Simile

The current simile study usually focus on recognition (Birke and Sarkar, 2006; Liu et al., 2018), interpretation (Su et al., 2016), and generation (Li et al., 2022). The recognition task (Tsvetkov et al., 2014; Mohler et al., 2016; Steen, 2010; Li et al., 2022) is judging whether a triplet or a sentence contains a simile. The interpretation (Liu et al., 2018) assigns an appropriate interpretation to a simile expression (Bizzoni and Lappin, 2018) or infers the shared properties of the tenor and the vehicle (Song et al., 2021; He et al., 2022; Chen et al., 2022). The generation task generates a simile sentence (Li et al., 2022; Chakrabarty et al., 2020; Stowe et al., 2021; Zhang et al., 2021) or the vehicle (Song et

[1]https://github.com/realZsh/simile-tasks

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

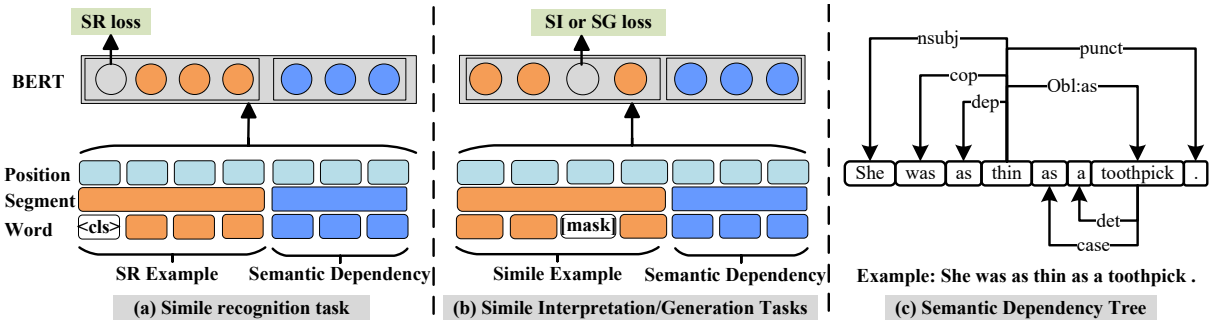


Figure 1: Demonstration of the training method and semantic dependency.

al., 2021; Chen et al., 2022). In this paper, we follow previous work and study the simile recognition/interpretation/generation (SR/SI/SG) tasks. Since there are not enough simile data that can be used for all three simile tasks. We construct the data we need based on existing SI data.

2.3 Exploring Simile Knowledge in PLMs

Previous simile work usually exploited the simile knowledge from PLMs for resolving downstream tasks. Song et al. (2021) fine-tune BERT (Devlin et al., 2019) for simile recognition and simile component (tenor, shared property, and vehicle) extraction. Chakrabarty et al. (2020) fine-tune BART (Lewis et al., 2020) on the literal-simile pairs to generate novel similes given a literal sentence. He et al. (2022) design a simile property probing task to let the PLMs infer the shared properties of similes for the interpretation task. Chen et al. (2022) propose an Adjective-Noun mask Training method to explore simile knowledge from BERT for simile interpretation and generation tasks. Li et al. (2022) fine-tune a GPT-2 (Radford et al., 2019) model for simile generation. In this paper, we also study how to explore simile knowledge from PLMs. However, different from previous work, we investigate how to leverage three simile tasks to explore more generic simile knowledge from PLMs.

3 Our Proposed Method

In this section, we formalize the simile recognition/simile interpretation/simile generation (SR/SI/SG) tasks and introduce our method in detail. For a fair comparison with previous work (He et al., 2022; Chen et al., 2022), we use BERT-base (Devlin et al., 2019) as the backbone of our model. Figure 1 shows the model structure of SR/SI/SG tasks.

3.1 Training of Simile Recognition (SR) Task

We follow previous work (Liu et al., 2018; Li et al., 2022) and define SR as a binary classification task. The SR model needs to distinguish whether an input sequence contains a simile. The input to the SR model is a sequence and the output is a binary label: True for simile and False for literal. The only common feature between simile data and literal data is that they both contains the comparator words (Liu et al., 2018). For example, the sentence "the boy runs like a deer." is a simile, but the sentence "the girl looks like her mother." is literal.

Following the original BERT paper, we use the first output position (a special token <cls>) to calculate the classification score, such as (a) part in Figure 1. We denote the corresponding output vector of <cls> as E_{cls}. Then the final score $\mathcal{S}$ of the input sequence is calculated as follows:

$$\mathcal{S} = \sigma(W_2 \cdot \mu(W_1 \cdot E_{cls} + b_1) + b_2), \tag{1}$$

where $W_{1,2}$ and $b_{1,2}$ are training parameters; σ/μ is the sigmoid/tanh function, respectively. The example with $\mathcal{S} \geq 0.5$ is classified as a simile, otherwise literal. The training loss is cross-entropy between predicted labels y_i and ground-truth label $\bar{y}_i$:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918–929, Harbin, China, August 3 – 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Task	Example	Candidates
SI	My client is as [MASK] as a newborn lamb.	**A**. innocent. **B**. delicious. **C**. legal. **D**. guilty.
SG	The participant swims like a [MASK].	**A**. dolphin. **B**. plait. **C**. depiction. **D**. pod.

Table 2: Examples for simile interpretation/generation tasks. We place the correct answer in the first position in these examples. In real data, the position of the correct answer is randomly placed. During training, the model learns to recover the [MASK] word. During the test, the model needs to select one answer from the 4 candidates.

$$\mathcal{L}_{SR} = -\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i log P(y_i)) \tag{2}$$

Where N is the number of training examples. After this fine-tuning, we can test the model on the SR test sets. We input an example and verify whether the SR model gives a correct classification for it.

3.2 Training of Simile Interpretation (SI) and Simile Generation (SG) Tasks

Following the previous simile interpretation (SI) and simile generation (SG) work (Song et al., 2021; He et al., 2022), we define the training of SI and SG as a masked language model task where the BERT learns to recover the masked words, such as (b) part in Figure 1. Two examples are shown in Table 2. In SI, the masked word is the shared property. In SG, the masked word is the vehicle.

During the test, we also follow the previous work (Song et al., 2021; He et al., 2022) and define SI/SG as a multi-choice task which chooses an answer from 4 candidates. Given an input simile sentence or dialogue with a masked shared property/vehicle, the SI/SG model needs to select the correct property/vehicle from the candidates, respectively. We use the masked-word-prediction heads of BERT to compute the probability for each candidate. The candidate with the highest probability will be chosen as the final choice.

3.3 Training with Semantic Dependency Features

Through the training process with SR/SI/SG, the PLM learns to use simile knowledge for three different simile tasks. However, the distribution of the training data may restrict the model's performance when applied to unseen domains. To this end, we enhance the PLM with global semantic dependency information, which can help the model learn simile knowledge across different syntax structures. This more generic simile knowledge can help the model's performance on unseen domains.

We adopt the semantic dependency tool[1] to get the semantic dependency tree of each input sequence. One example is shown in (c) part of Figure 1. The dependency tree for "She was as thin as a toothpick ." is a list of tuples: "[('ROOT', '.', 'thin'), ('nsubj', 'thin', 'She'), ('cop', 'thin', 'was'), ('dep', 'thin', 'as'), ('case', 'toothpick', 'as'), ('det', 'toothpick', 'a'), ('obl', 'thin', 'toothpick'), ('punct', 'thin', '.')]". The word "thin" is the root of this tree and please refer to Manning et al. (2014) for the definition of each semantic dependency relation.

For the SR task, we can directly use the semantic dependency results. However, in SI or SG task, key simile component such as the vehicle "toothpick" of the above example is masked. We change the example to "She was as thin as a UNK .", where UNK represents the [MASK] vehicle. Then the output semantic dependency tree changes to "[('ROOT', '.', 'thin'), ('nsubj', 'thin', 'She'), ('cop', 'thin', 'was'), ('dep', 'thin', 'as'), ('case', 'UNK', 'as'), ('det', 'UNK', 'a'), ('obl', 'thin', 'UNK'), ('punct', 'thin', '.')]". In this way, the model is aware of the semantic dependency tree of the input sentence but does not see the masked word.

The final input to BERT is the concatenation of the semantic dependency tree and the original sentence. We use different segment embedding to distinguish the data example and its semantic dependency information, such as the (a)/(b) part of Figure 1.

[1]https://stanfordnlp.github.io/CoreNLP

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

Dataset	Train / Dev / Test	Words/Example	Data Format
MSP-original (for SI)	4,510 / - / 1,633	12.2	sentence
MSP-modified for SG	4,510 / - / 1,633	12.3	sentence
MSP-modified for SR	7,216 / 902 / 902	12.3	sentence
New test set	- / - / 957	30.6	three-turn dialogue

Table 3: Statistics of datasets.

Relation:	*Definition*
RelatedTo:	*The most general relation. There is some positive relationship between A and B, but ConceptNet can't determine what that relationship is based on the data. Symmetric. exercise <-> fit*
IsA:	*A is a subtype or a specific instance of B; every A is a B. This can include specific instances; the distinction between subtypes and instances is often blurry in language. This is the hyponym relation in WordNet. car -> vehicle; Mexico -> Country*
Causes:	*A and B are events, and it is typical for A to cause B. run -> tired*
Desires:	*A is a conscious entity that typically wants B. Many assertions of this type use the appropriate language's word for "person" as A. person -> respect*
DistinctFrom:	*A and B are distinct member of a set; something that is A is not B. Symmetric. red <-> blue; June <-> May*
SymbolOf:	*A symbolically represents B. blue -> cold*
MannerOf:	*A is a specific way to do B. Similar to "IsA", but for verbs. auction -> sale*
LocatedNear:	*A and B are typically found near each other. Symmetric. computer <-> table*
CausesDesire:	*A makes someone want B. hungry -> eat food*
MadeOf:	*A is made of B. porcelain -> ceramic*

Table 4: Relations in ConceptNet we used to find distractors. "<->" means Symmetric relation for A and B. "->" means Asymmetric relation that A entails B.

After training, we test with two different settings, one is the MSP test set, and the other is an unseen test set that is newly constructed by us. Next, we will introduce the data sets.

4 Experimental Setup

4.1 Datasets

We use simile data sets with "as ... as" comparator since the shared property naturally exists in the comparator, which is suitable for our experiments since we want conduct all SR/SI/SG tasks with this data. This kind of simile data can be used for all three simile tasks. The data statistics are shown in Table 3 and we introduce the data details next.

4.1.1 MSP dataset (for SI task)

Since we could not find enough data for all three simile tasks, we construct the required data based on a recently released simile benchmark. The multi-choice simile probe (MSP) data (He et al., 2022) is originally proposed for SI task. It has a total of 5,410 training examples and 1,633 test examples. All examples in MSP are simile sentences with comparator "as ... as". Each example in the MSP test set has three distractors for the shared property. During training, the model learns to recover the masked property in MSP training data. During the test, the model needs to choose the correct answer from 4 candidates in the MSP test set.

4.1.2 MSP-modified data (for SG task)

To perform the SG task, we introduce a modified version of MSP. During training, we mask the vehicle and train the model to recover it. During the test, we provide 4 vehicle candidates for the multi-choice task. Besides the real vehicle, the other 3 distractors are constructed with ConceptNet (Speer et al.,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

2017). The ConceptNet is a knowledge graph that connects words and phrases of natural language with labeled relations (Speer et al., 2017). We show 10 relations of ConceptNet in Table 4. They are used to find the related concepts to the vehicle as the distractors. For the example "She was as thin as a toothpick .", the vehicle is the word "toothpick". We find that "toothpick" is usually located near to (LocatedNear) "food" and can be made of (MadeOf) "plastic" or "wooden". So the three distractors can be "food, plastic, wooden". When we find more than three distractors with the relations in Table 4, we randomly choose 3 of them as the final distractors. Notice that there are a few cases we could not find enough distractors, we manually construct distractors for these cases.

4.1.3 MSP-modified data (for SR task)

Similarly to the SG task, we introduce another modified version of MSP for the SR task. Since the SR task needs both simile examples and literal examples (Liu et al., 2018; Li et al., 2022), we use certain relations in ConceptNet to obtain the literal data we need. For example, we replace the tenor "his muscle" in the simile example "his muscle is as hard as a rock" with the phrase "a stone", the Synonym concept of "a rock", then we get a literal sentence "a stone is as hard as a rock". This is different from replacing "his muscle" with a random word such as "air". Because the sentence "air is as hard as a rock" does not have a practical meaning. If we use "air is as hard as a rock" as a literal sample to train an SR model. The model may classify this sample as literal by identifying that it is against common sense. Instead, when we use the literal sentence "a stone is as hard as a rock", the SR model needs to use simile knowledge to judge whether this example is a simile. The knowledge is that simile only exists when comparing things from different categories. "stone" and "rock" are in the same category so this sentence is literal. Besides the Synonym relation, we can also use other relations of the vehicle including DistinctFrom/IsA/RelatedTo/SimilarTo in ConceptNet to find a concept to replace the tenor. When we find more than one distractor, we randomly choose one of them as the literal sentence. By this method, we not only obtain the required training literal data but also has more difficult literal data. Because the syntax structure of the literal data is the same as the original simile example but the semantic information is different. These literal examples will help the model to learn more accurate simile knowledge. Finally, we obtain 9020 examples. We randomly split this data into train/dev/test (8:1:1) to train our model. During training, the model learn to give a higher/lower score for the simile/literal data. During the test, the model assigns a score for the input. In both training and testing, an example with a score ≥ 0.5 will be set as simile, <0.5 will be set as literal.

4.1.4 A new test data (for SR/SI/SG task)

After the above data set construction, we now have the training/testing MSP sets for SR/SI/SG tasks. We denote the MSP test sets as a seen set because the training and testing data are in a similar domain and similar range of length. To test whether our method can help to explore more generic simile knowledge, we provide unseen test sets for SR/SI/SG tasks.

The new test data is collected from Reddit-dialogue corpus (Dziri et al., 2018) which has $\sim$15 million English dialogues. The dialogues are comments from the Reddit forum and each dialogue has three turns. We extract 1,000 dialogue examples from the Reddit dataset with three rules. First, the dialogue length is around 30 tokens so it is informative and not too long. Second, the last turn must contain a comparator "as ... as" with an adjective word in the comparator. Third, we use the semantic dependency tool to ensure that the tenor and vehicle are in the response. Then we manually annotate whether they are similes or literal. For the simile sentences, we further check whether the tenor and vehicle labeled by the semantic dependency tool are correct. Notice that we do not make any change to the data. Therefore, for dialogue examples that tenor or vehicle is missing, we withdraw this example even it contains a simile. We make sure that all simile components are in the example so that we can use it for all simile tasks. We finally have 486 simile examples and 471 literal examples, total 957 examples. When testing on SI/SG, we construct the distractors using the same method as we construct MSP-modified data. For the examples in this new test set that we could not find enough vehicle distractors, we randomly choose the vehicles from other dialogues as the distractors.

The new test set is different from the training data (MSP) in the following respects: 1) the data format

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

is dialogue and the length is much longer than data in MSP; 2) the tenor and vehicle in dialogue can be verbal phrase or sentence, which is different from the noun phrase in MSP. We use the new test set to verify whether our method can perform well on a different simile distribution compared to MSP.

4.2 Baselines

We introduce the baselines we used in this section.

4.2.1 Baselines for SR

BERT-base is fine-tuned on the MSP modified SR training set. The checkpoint for test is selected based on the performance on the corresponding dev set.

4.2.2 Baselines for SI/SG

The first baseline is a BERT-base model without fine-tuning with the data sets in this paper. It takes the input with key simile component masked and predicts the masked words. The second baseline is BERT-ANT (Chen et al., 2022) which is trained with masked word prediction with a number of metaphor data. It is based on a BERT-large-uncased model and can solve the SI and SG tasks in a unified framework of simile triple completion. For example, when giving tenor=fireman and vehicle=bull, BERT-ANT can generate a list of words including the shared property like ”strong” or ”brave”. When performing our SI/SG tasks, we match the candidates of each example with the output list of BERT-ANT. An example is counted correct if the ground truth answer is listed before the other three distractors. The BERT-Probe baseline is from (He et al., 2022) that fine-tuned BERT with MSP-original data for simile interpretation task. To compare both SI and SG tasks with this baseline, we further fine-tuned the BERT-Probe model with MSP-modified SG training data and report its results on the MSP-modified SG test data.

4.2.3 Our models

Besides the fully fine-tuned model, we also provide several settings for our model. (- SR training) means we remove the simile recognition data in the unified training process. Similarly, (- SI training) and (- SG training) means we remove the SI and SG data in training, respectively. (- Semantic Dependency) means we do not use syntax features. These settings can reflect the contribution of the removing part.

4.3 Evaluation Metrics

Following previous work (Liu et al., 2018), we use macro Precision/Recall/F1 and Accuracy to measure the simile recognition results. Following previous work on simile interpretation and generation (Chen et al., 2022), we use Hit@1 to measure the multi-choice accuracy.

4.4 Implementation Details

Our model is implemented by PyTorch (Paszke et al., 2019). The implementations of the pre-trained models in this paper are all based on the public Pytorch implementation[2]. During the training, the maximum input length is set to 512. We use a single Tesla v100s GPU with 32gb memory for experiments. The batch size is all set to 24. The model is optimized using the Adam optimizer with a learning rate of 5e-6. The learning rate is scheduled by a warm-up and linear decay. A dropout rate of 0.1 is applied for all linear transformation layers. The gradient clipping threshold is set as 10.0. Early stopping on the corresponding validation data is adopted as a regularization strategy. The training epochs are ∼3. For SI/SG testing on the new unseen set, if the masked position is a single word, we select the answer with the highest probability of the masked position; if there are multiple masked words, we encode the predicted words and the candidates into dense vectors with a sentence-transformer (https://www.huggingface.co/sentence_transformers/all-MiniLM-L6-v2). Then we compute the cosine similarity between the predicted words and each of the candidates. The candidate with the highest similarity is chosen as the answer. We use Hit@1 to measure the accuracy.

[2]https://github.com/huggingface/transformers

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

Model	Precision	Recall	F1	Accuracy
MSP-modified SR Test set				
BERT-base	0.7127	0.6981	0.6939	0.6996
Ours	**0.7904***	**0.7905***	**0.7905***	**0.7905***
(- SR training)	0.5000*	0.5000*	0.3768*	0.5000*
(- SI training)	0.7712*	0.7725*	0.7718*	0.7717*
(- SG training)	0.7774*	0.7801*	0.7781*	0.7779*
(- Semantic Dependency)	0.7822*	0.7805*	0.7836*	0.7821*
Our Proposed Test set				
BERT-base	0.4949	0.4963	0.4559	0.4922
Ours	**0.5419***	**0.5393***	**0.5332***	**0.5413***
(- SR training)	0.4927	0.4968	0.4179	0.5026
(- SI training)	0.5030*	0.5020*	0.4532*	0.4974*
(- SG training)	0.5152*	0.5136*	0.4985*	0.5110*
(- Semantic Dependency)	0.5325*	0.5284*	0.5114*	0.5256*

Table 5: Simile recognition results. The BERT-base (fine-tuned with MSP-modified SR train set) is the base model to do the significant test for our models (* means statistically significant with $p<0.01$).

5 Results and Analysis

In this section, we introduce the experimental results and provide our analysis of the results.

5.1 Simile Recognition

Table 5 shows the simile recognition results.

5.1.1 Comparing with Baseline

The BERT-base model is fine-tuned with the MSP-modified SR train set and is tested with two test sets. One is the MSP-modified SR test set and the other is our new test set. We can see that on both test sets, our model performs better than the baselines. On the MSP-modified SR test set, our model surpasses BERT-base by around 7.8% on accuracy. On our proposed test set, our model outperforms BERT-base by around 4.9% on accuracy. On Macro Precision/Recall/F1, our model also outperforms the BERT-base model. The results show that our method not only can help PLM to use a more accurate simile knowledge but also perform better on a more difficult unseen test set. The results on the new test set are much lower than the MSP-modified SR test set, which indicates the new test set is much harder. Although our method helps the PLM to obtain a better performance on this new test set, there is still a lot of room to improve.

5.1.2 Ablation Study on SR

We also report the ablation study in Table 5. We can see that on both the MSP test set and the new test set, removing the key component of our model will cause declines. On the MSP test set, (- SR training) is exactly 50% because the model does not understand the SR task without the SR training. On the new test set, similar results are observed. The results are also around 50% and are not statistically significant.

On both test sets, (- SI training) performs worse than (- SG training). The results indicate that the SI fine-tuning task (recovering the masked property) is more useful than the SG fine-tuning task (recovering the masked vehicle) for the model to learn SR knowledge. It is because the shared property usually serves as the root of the semantic dependency tree. As shown in the (c) part of Figure 1, the shared property connects most words in a simile sentence and the vehicle only connects a few words. When training with SI, the model learns more semantic relations between words than training with SG, so that the model can better leverage this semantic dependency knowledge for the SR task.

(- Semantic Dependency) causes more declines on the new test set (from 0.9 $\sim$ 2.2% on all metrics) than on the MSP test set (from 0.7 $\sim$ 1.0% on all metrics). It means the semantic dependency information

Model	Interpretation	Generation
MSP-original SI Test set and MSP-modified SG Test set		
BERT-base (without fine-tuning)	0.7436	0.8155
BERT-Probe (He et al., 2022)	0.8015	0.8667
BERT-ANT (Chen et al., 2022)	0.8020	0.8675
Ours	**0.8101***	**0.8986***
(- SR training)	0.8006*	0.8819*
(- SI training)	0.7273*	0.8608*
(- SG training)	0.7832*	0.8113*
(- Semantic Dependency)	0.8089*	0.8799*
Our proposed Test set (the simile data)		
BERT-base (without fine-tuning)	0.5905	0.4510
BERT-Probe (He et al., 2022)	0.6454	0.5031
BERT-ANT (Chen et al., 2022)	0.6521	0.5094
Ours	**0.6642***	**0.5232***
(- SR training)	0.6584*	0.5189*
(- SI training)	0.6401*	0.4976*
(- SG training)	0.6525*	0.4888*
(- Semantic Dependency)	0.6531*	0.5022*

Table 6: Simile interpretation and generation results (Hit@1) on MSD-En. The BERT-Probe is the base model to do the significant test for our models (* means statistically significant with $p<0.01$).

helps the PLM to learn a more generic simile knowledge. This generic simile knowledge brings more gains in an unseen domain.

To sum up, experimental results on SR verify that 1) our method can explore more accurate and generic simile knowledge; 2) each fine-tuning task and the semantic dependency signal contributes to the performance.

5.2 Simile Interpretation and Generation

Table 6 shows the simile interpretation and simile generation results. The SI task uses the MSP-original SI test set and our new test set. The SG task uses the MSP-modified SG test set and our new test set.

5.2.1 Comparing with Baselines

The first baseline is the BERT-base model without any fine-tuning. We can see that BERT-Probe performs better than BERT-base on both SI/SG tasks. The results are reasonable since BERT-Probe benefits from the fine-tuning of MSP-original/MSP-modified data on SI/SG tasks, respectively.

Different from the above two baselines, BERT-ANT is based on BERT-large and trained with a large corpus through Adjective-Noun mask Training. Benefiting from both a larger parameter size and the training process, BERT-ANT outperforms the BERT-Probe on both SI/SG tasks.

On the other hand, our model surpasses the strong BERT-ANT on both SI/SG even though our model uses BERT-base as the backbone. The results again verify that our method can enhance PLM with more accurate and generic simile knowledge.

The results on the new test set are still lower than the MSP test sets. One notable result is that the gap between results on the SG task is much larger than the gap on the SI task. The results show that the MSP-modified SG test set is easier than the MSP-original SI test set. The Hit@1 results are 89.86% and 81.01%, respectively. This may also be one of the reasons why SI training contributes more than SG training in Table 5. We can try constructing more difficult SG training data to improve the learning efficiency of our model.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

5.2.2 Ablation Study on SI/SG

We also report the ablation study in Table 6. We can see that on both MSP test sets and the new test set, removing the training component of our model will cause declines.

On the MSP-original SI test set, (- SI training) causes ∼8.3% declines. On the new test set, (- SI training) only has ∼2.4% declines. The results are reasonable since the unseen test set is not as sensitive to the training data as the seen test set. A similar trend can be observed with the SG task. On the MSP-modified SG test set, (-SG training) causes ∼8.7% declines. On the new test set, (- SG training) only entails ∼3.4% declines.

On all test sets, (- SR training) only causes a little decline, which indicates that the SR fine-tuning contributes little to SI/SG tasks. This is different from the experimental results in Table 5, where SI/SG training contribute more to the SR task. How to leverage SR training to improve the SI/SG tasks requires further study.

Similar to the SR experiments, (- Semantic Dependency) causes more declines on the new test set (∼1.1% on SI and ∼2.1% on SG) than on MSP test sets (∼0.1% on SI and ∼1.9% on SG). The results mean the semantic dependency information helps more on an unseen set than the seen set, which is consistent with the results of the SR task.

To sum up, experimental results on SI/SG again verify that 1) our method can explore more accurate and generic simile knowledge; 2) each fine-tuning task and the semantic dependency signal have positive effects on the performance.

6 Conclusion

We propose a novel method to explore more accurate and generic simile knowledge from PLMs. We fine-tune PLM with three simile tasks (recognition, interpretation, and generation) to explore local simile knowledge between key simile components (tenor, shared property, vehicle). Then we use the semantic dependency feature for global simile knowledge among different examples. This global simile knowledge can help our model perform well across domains. Experiments with seen and unseen test sets verify the effectiveness of our method. Our exploring method may be useful for other NLP tasks that leverage knowledge from PLMs. Since our method does not need an expensive pre-training process, it may also be useful for leveraging more large-scaled PLMs. Future works include but are not limited to **1)** testing our method on other knowledge-intensive tasks; **2)** verifying whether our method can be transferred to auto-regressive-based PLMs.

Acknowledgements

This research project is supported by the National Natural Science Foundation of China (61872402), Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities") (18ZDJ03)

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In Diana McCarthy and Shuly Wintner, editors, *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee Leong, editors, *Proceedings of the Workshop on Figurative Language Processing, FigLang@NAACL-HLT 2018, New Orleans, Louisiana, 6 June 2018*, pages 45–55. Association for Computational Linguistics.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6455–6469. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5875–5887. Association for Computational Linguistics.

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

Nouha Dziri, Ehsan Kamalloo, Kory W. Mathewson, and Osmar R. Zaïane. 2018. Augmenting neural response generation with context-aware topical attention. *CoRR*, abs/1811.01063.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R. Zaïane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5271–5285. Association for Computational Linguistics.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7875–7887. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6468–6479. International Committee on Computational Linguistics.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1543–1553. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6723–6737. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc T. Tomlinson. 2016. Introducing the LCC metaphor datasets. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 2008–2018. ACL.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Anthony M Paul. 1970. Figurative language. In *Philosophy & Rhetoric*, page 225–248.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. A knowledge graph embedding approach for metaphor processing. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:406–420.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Gerard Steen. 2010. A method for linguistic metaphor identification: From mip to mipvu. volume 14. John Benjamins Publishing.

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021. Exploring metaphoric paraphrase generation. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 323–336. Association for Computational Linguistics.

Chang Su, Jia Tian, and Yijiang Chen. 2016. Latent semantic similarity based interpretation of chinese metaphors. *Eng. Appl. Artif. Intell.*, 48:188–203.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 248–258. The Association for Computer Linguistics.

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and A neural approach. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14383–14392. AAAI Press.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 918-929, Harbin, China, August 3 - 5, 2023.