CCL 2022

The 21st Chinese National Conference on Computational Linguistic

Proceedings of the 21st Chinese National Conference on Computational Linguistics

October 14 - October 16, 2022

Nanchang, China

©The 21st Chinese National Conference on Computational Linguistic

Order copies of this and other CCL proceedings from:

Chinese National Conference on Computational Linguistic (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 twentieth China National Conference on Computational Linguistics (21st CCL). The conference and symposium were hosted and co-organized by Inner Mongolia University, 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 forum for disseminating new scholarly and technological work in computational linguistics, with a major emphasis on computer processing of the languages in China such as Mandarin, Tibetan, Mongolian, and Uyghur.

The Program Committee selected 86 papers (64 Chinese papers and 22 English papers) out of 293 submissions for publication. The acceptance rate is 29.35%. The 86 papers cover the following topics:

- Linguistics and Cognitive Science (10)
- Fundamental Theory and Methods of Computational Linguistics (6)
- Information Retrieval, Dialogue and Question Answering (6)
- Text Generation and Summarization (4)
- Knowledge Graph and Information Extraction (11)
- Machine Translation and Multilingual Information Processing (6)
- Minority Language Information Processing (6)
- Language Resource and Evaluation (10)
- Social Computing and Sentiment Analysis (8)
- NLP Applications (19)

The final program for the 21st 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 ACL Anthology 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 in beautiful Nanchang.

August 2022

Maosong Sun, Yang Liu, Wanxiang Che
Yang Feng, Xipeng Qiu, Gaoqi Rao, Yubo Chen

Organizers

Conference Chairs

Maosong Sun	Tsinghua University, China
Yang Liu	Tsinghua University, China

Program Chairs

Wanxiang Che	Harbin Institute of Technology, China
Yang Feng	Institute of Computing Technology, CAS, China
Xipeng Qiu	Fudan University, China

Area Co-Chairs

Weidong Zhan	Peking University, China
Jingxia Lin	Nanyang Technological University, Singapore
Kewei Tu	ShanghaiTech University, China
Tao Yu	The University of Hong Kong, China
Shaochun Ren	Shandong University, China
Jie Yang	Technische Universiteit Delft, Netherlands
Xiaocheng Feng	Harbin Institute of Technology, China
Lifu Huang	Virginia Polytechnic Institute and State University, America
Shizhu He	Institute of Automation, CAS, China
Yangqiu Song	The Hong Kong University of Science and Technology, China
Jingsong Su	Xiamen University, China
Jiatao Gu	Meta, USA
Aishan Wumaier	Xinjiang University, China
Quecairang Hua	Qinghai Normal University, China
Muyun Yang	Harbin Institute of Technology, China
Yunfei Long	University of Essex, Britain
Tieyun Qian	Wuhan University, China
Lidong Bing	Alibaba DAMO Academy, China
Richong Zhang	Beihang University, China
Meng Jiang	University of Notre Dame, America
Liang Pang	Institute of Computing Technology, CAS, China

Local Arrangement Chairs

Mingwen Wang	Jiangxi Normal University, China
Jiali Zuo	Jiangxi Normal University, China

Evaluation Chairs

Hongfei Lin	Dalian University of Technology, China
Zhenghua Li	Soochow University, China

Publications Chairs

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

CCL Chair of Frontier Forum

Zhiyuan Liu	Tsinghua University, China

CCL Workshop Chairs

Jiajun Zhang	Institute of Automation, CAS, China
Rui Yan	Peking University, China

CCL Sponsorship Chairs

Qi Zhang	Fudan University, China
Tong Xiao	Northeastern University, China

CCL Publicity Chair

Ruifeng Xu	Harbin Institute of Technology, China

CCL Website Chair

Shujian Huang	Nanjing University, China

CCL System Demonstration Chairs

Min Peng	Wuhan University, China
Weinan Zhang	Harbin Institute of Technology, China

CCL Student Seminar Chairs

Xianpei Han	Institute of Software, CAS, China
Zhuosheng Zhang	Shanghai Jiao Tong University, China

CCL Finance Chair

Yuxing Wang	Tsinghua University, China

Table of Content

中国语言学研究 70 年：核心期刊的词汇增长

王珊
澳门大学人文学院中国语言文学系
珠海澳大科技研究院
shanwang@um.edu.mo

詹润哲
澳门大学科技学院计算机和信息科学系

姚双云
华中师范大学语言与语言教育研究中心

摘要

建国以来我国语言学经过 70 年的发展取得了瞩目的成就，已有研究主要以回顾主要历史事件的方式介绍这一进程，但尚缺少使用量化手段分析其历时发展的研究。本文以词汇增长为切入点探究这一主题，首次创建大规模语言学中文核心期刊摘要的历时语料库，并使用三大词汇增长模型预测语料库中词汇的变化。本文选择拟合效果最好的 Heaps 模型分阶段深入分析语言学词汇的变化，显示出国家政策的指导作用和特定时代的语言生活特征。此外，与时序无关的验证程序支撑了本文研究方法的有效性。

关键词：中国语言学；词汇增长；核心期刊；摘要；语料库；历时发展

70 Years of Linguistics Research in China: Vocabulary Growth of Core Journals

Shan Wang
Department of Chinese Language and Literature,
Faculty of Arts and Humanities,
University of Macau, Macau, SAR, China

Zhuhai UM Science & Technology Research Institute,
Zhuhai, China
shanwang@um.edu.mo

Runzhe Zhan
Department of Computer and Information Science,
Faculty of Science and Technology,
University of Macau, Macau SAR, China

Shuangyun Yao
Research Center for Language and Language Education,
Central China Normal University,
Wuhan, China

Abstract

Since the founding of P.R. China, linguistics in China has made remarkable achievements after 70 years of development. The existing studies have mainly introduced the development of linguistics by reviewing historical events, but no research has used quantitative means to analyze its longitudinal development. This article has explored this topic from the perspective of vocabulary growth. For the first time a large-scale diachronic corpus of abstracts from Chinese core linguistic journals is created. Subsequently, the analysis is conducted on this corpus with the help of three vocabulary growth models. Then the Heaps model with the best fitting effect is selected to further analyze the changes of linguistic vocabulary in different times, showing the guiding role of national policies and the characteristics of language life in specific era. Furthermore, a time independent validation procedure is performed, which supports the effectiveness of the proposed methodology of this study.

Keywords: Linguistics in China; Vocabulary Growth; Core Journals; Abstract; Corpus; Diachronic Development

1. 引言

建国以来，我国语言学研究经过 70 年的发展，从筚路蓝缕到开拓创新，取得了瞩目成

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第1页-第13页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

就。已有的研究主要以回顾主要历史事件的方式介绍这段历史，对我们了解前辈学者和当代时贤对语言学的贡献起着重要作用。但是，尚无研究对这段时期语言学词汇的演变进行历时考察。汉语语言学核心期刊能够代表在不同时期中国语言学的最新发展动向和关注热点，而论文摘要是整篇文章中最具代表性的部分，具有牵引文章脉络的作用，展现了极高的信息密度，能够在一定程度上直观地反映语言学的发展。

词汇增长（Vocabulary Growth）模型体现了文本中独特性用词的比例。对于以时间序列组织的文本，该指标能够反映文本在一定时间范围内是否有新词加入、新词增加的比例或速度等特征。词种（types）与词例（tokens）数量的比值 TTR（type-token-ratio）刻画了一定长度的文本内非重复用词的比例，常被用作建模词汇增长问题的指标之一。在历时语料中，以时间序递进式扩大计算长度窗口，统计不同采样点的 TTR 能够反映出新词的增长率变化情况；词汇增长模型能够反映 TTR 的变化趋势，并预测在语料数量继续扩大情况下新词数量将以何种趋势增长。词汇增长模型的有效性已在不同类型的文本上得以证实（Savoy, 2015; 王珊、王会珍, 2019）。对于学术领域的文本，新词的出现或增长情况能够直观地反映出学科的发展，故选取词汇增长模型对学术文本的 TTR 变化特征进行建模拟合，能够体现某个学科的历时发展与演化特征。目前尚无采用词汇增长模型的方法探究学科领域演化的研究。本文选取建国以来的语言学中文核心期刊摘要，构建大规模历时语料库，进行词汇增长建模对约 70 年间语言学领域的词汇演化进行定量与定性分析，探究词汇的变化趋势与学科发展、社会因素之间的关系，从而进一步折射七十年来我国语言生活的变化情况。

2. 相关研究

2.1 词汇增长研究

词汇增长已用于判定作者身份、评定语言能力、分析施政风格等诸多研究中。Hoover（2003）对十二位作者的作品词种数量进行统计，发现词汇增长可以用于判断作者的身份。Yu（2010）指出词汇增长与写作和口语的质量在统计学上具有显著的正相关性，证明对于语言能力较高的人，其词汇增长指标也会相对较高，是考察学习者对一门语言掌握情况的指标。Mellor（2010）发现词汇增长可以用于衡量说话者与写作者的语言程度，语言程度越高，则使用的低频词较多。X. Wang（2014）分析了英语二语学习者的电子邮件中的词汇增长与写作熟练度之间的关系。Savoy（2015）分析了 1790 年到 2014 年历任美国总统发表的 225 篇演讲中的词汇增长情况，对比了 Heaps、Hubert-Labbe 两个词汇增长模型的适用性并对不同的词汇增长的变化作出了分析，作者将整个时间段分为低于模型预期值与高于模型预期值的时间段，联系历任总统的施政风格与对应时代的政治经济背景进行定性分析。王珊、陈钊、张昊迪（2021）利用词汇增长模型刻画十余年间澳门新闻报刊内容的词汇历时演变，结果表明词汇增长的倾向性与施政时期方针、人民生活的变化有极大关联性。

2.2 学术汉语的词汇研究

Swales（1985）提出了专门用途英语（ESP, English for specific purpose）与学术用途英语（EAP, English for academic purpose）的概念。学术词汇是学术语言的重要组成部分，是除了核心词汇外使用频率较高的词汇种类（Paquot, 2010）。学术用途英语研究取得了丰硕的成果，但目前对学术汉语词汇的研究仍旧十分稀少。涉及学术汉语的研究可分为三大类：一是文本特征研究，例如吴格奇、潘春雷（2010）参考 Hyland（2005）提出的立场分析框架，分析立场标记词汇以探究学术写作者的语用策略与身份建构；张赪、李加、申盛夏（2020）基于自建的多学科语料库，分析了不同学科的论文写作者对实词与虚词的使用特征；朱宇、胡晓丹（2021）基于自建的人文社科语料库，利用多维度分析法针对连词在论文中的语言功能进行了考察。二是英汉对比研究，其研究多以中文为母语者为考察对象，包括对立场信息一致性的对比（赵永青等, 2019）、身份指称的差异对比（李志君, 2014）、衔接用词的差异对比（胡芳、陈彧, 2005）等。三是学术词表的研制，刘锐、王珊（2017）用小规模知网学术论文语料构建学术词表，王笑然、王佶旻（2022）则通过自建经贸类的学术语料库，研制经贸类学术汉语词表。从现有研究来看，目前公开的大型综合性汉语语料库（如 BCC、CCL、Chinese

(c) 2022 中国中文信息学会计算语言学专业委员会

Gigaword 语料库等）中缺少学术领域的语料，进行有关研究仍需自建共时或历时语料库，所需时间代价与人力成本较高，影响了学术汉语研究的开展。

2.3 建国后的中国语言学发展

建国以来，我国的语言文字工作取得了瞩目的成就。建国初期国家就对文字改革提出了三大方向：简化汉字、推广普通话与制定推行汉语拼音方案，间接推动了语言学多个领域的发展，例如推广普通话所需要的调研任务为汉语方言学的研究打下了扎实的材料基础，在制定汉语拼音方案的过程中对普通话音位系统的描写奠定了汉语音系学研究的基础。尽管中间历经了文化大革命的动荡使语言学研究工作有所停滞，但改革开放后，在国家各行业逐步与国际环境接轨的大环境下，外来理论和方法的引入为汉语语言学研究带来更多新的视角；在吸收外来理论的基础上，现在我国汉语语言学研究工作正逐步迈入自主创新的阶段。

对于建国以来的语言学工作的开展，陆俭明（1999）从学科建设和学术发展的角度对 21 世纪之前的语言学工作与研究进行了梳理，刘丹青（2019）对我国语言学研究各个分支领域近 70 年内代表性的理论与应用进行了介绍，国家语言文字工作委员会（2019）则整理了纪年史料，侧重对党和国家在语言文字工作上的重要事件进行了汇编总结。这些研究回顾了语言学主要的历史事件，但尚未有借助量化统计方法对建国以来的语言学发展的分析。

综上所述，在大规模语料库基础上，对学术汉语中语言学的词汇增长进行考察，能够进一步完善现有的研究，丰富语言学发展的研究成果。语料库中的词汇增长情况能够使用数学模型进行展示，故对以时间关系组织的文本序列进行增长情况建模，能够得到目标文本历时的 TTR 变化信息。籍由此，我们能够进一步分析相应时间段之内和不同时间段之间的文本特征及其变化情况。本文选取自建国以来语言学中文核心期刊现存所有电子化收录的语料来创建语料库，主要采用语料库驱动、定量与定性相结合的研究方法，旨在解决以下问题：⑴如何构建具有代表性的语言学中文期刊语料库？⑵如何对语料库进行词汇增长模型建模并分析其反映的语言学发展特点？⑶如何验证词汇增长模型对语言学领域的词汇分析的准确性？

3. 创建语言学核心期刊语料库

现时对中文核心期刊的认定有认可度较高的索引，例如中文社会科学引文索引（Chinese Social Sciences Citation Index，CSSCI）、北京大学中文核心期刊要目总览与中国科学引文数据库（Chinese Science Citation Database，CSCD）等。其中 CSSCI 索引作为国家教育部的重点课题攻关项目，采取定量与定性评价相结合的方法筛选出人文社科领域的标志性期刊[1]，具有较大的影响力与公信力（马费成, 2000; 苏新宁, 2012; 邹志仁, 2000），被知网、万方等电子化文献数据库收录。CSSCI 数据库来源期刊的遴选工作遵循以下原则：公开、公平、公正；总量控制，动态调整；定量（文献计量指标）评价与定性（学科专家）评价相结合；质量优先，兼顾地区与学科平衡。所有入选期刊必须具备以下基本条件：刊载人文社会科学原创学术论文和学术评论等一次文献为主的中文学术期刊；中国大陆出版的期刊应具有 CN 号；按既定出版周期准时出版，符合期刊编辑出版规范，文献信息著录完整、规范。此外，CSSCI 索引不但针对人文社科研制，还对人文社科下属若干个子学科具有较完备的细化分类。

本文依据《CSSCI 来源期刊（2019-2020）目录》中“语言学”子类所收录的 24 个期刊，收集已电子化的论文元信息和摘要作为语料的来源。其中，论文的元信息包括期刊名、年份、期（卷）、作者、页码等信息。论文摘要高度凝练了学术论文的背景、动机、观点与结论，是整篇文章中最具代表性的部分（S. Wang, Liu, & Zhou, 2022）；就阅读过程而言，摘要具有牵引文章脉络的作用，展现了极高的信息密度；摘要中的词汇、句法均带有密集且丰富的特定领域话语特征，故本研究以论文摘要作为研究对象。

本研究构建的语言学中文核心期刊语料库收录了自 1957 年 6 月至 2020 年 8 月共计 71988 篇论文的信息。本文对所有能收集到的语料进行了预处理：第一，期刊发布的信息有一些不属于学术论文的内容，例如征稿通知、编委会信息、会议通知等与语言学的词汇增长情况并

[1] CSSCI 来源期刊遴选标准: https://cssrac.nju.edu.cn/gywm/lxbz/20200102/i64328.html

(c) 2022 中国中文信息学会计算语言学专业委员会

无关联，故本研究将其视其为噪声信息并采取关键词过滤的方式进行排除。第二，个别期刊的早期文献以繁体中文书写，例如 1957 年《外语教学与研究》中大部分论文以繁体中文收录，而在计算中，简繁词语字形不一致将被视作两个单位进行统计，本文使用 OpenCC 工具[2]将所有繁体字形式统一转换为简体中文，例如“英語詞彙学”转换为“英语词汇学”。第三，标点不属于词汇，因此统计时根据 Python 中的 zhon[3]与 string[4]库分别去除了中英文标点符号。

为了进行词汇信息的提取，预处理后的语料由 pkuseg 多领域中文分词工具[5]进行词语切分。该工具由北京大学语言计算与机器学习组研制（Luo, Xu, Zhang, Ren, & Sun, 2019），与 jieba、THULAC 等分词工具包相比，在细领域分词的 F-Score 与跨领域测试的平均分上均占据优势。默认分词模型在混合领域上训练，并支持在细领域上对预训练模型进行调优以取得更高的准确率。由于学术期刊涉及议题所含的领域较为复杂，语言学不仅仅涉及本体理论研究，更与各学科与社会现象紧密相关，故本研究采取该工具的默认分词模型进行词语切分工作[6]。

表 1 列出了“语言学中文核心期刊摘要语料库”（1957-2020）的信息，包括期刊名称、所搜集到的期刊的时间跨度、记录总条数、有摘要的文献数量，以及词例数、词种数、字数等统计信息。其中时间跨度、记录总条数、摘要数量是在经上文所述预处理方法清洗后的文本上得出的，词例数、词种数、字数的统计方法为：基于所获得的摘要，统计含中外文的词例、词种、字数，例如，分词后的文本“HSK 四 级”，计为 3 个词例、3 个词种、5 个字。该语料库共涵盖 24 个语言学核心期刊的 65791 个摘要，语料规模高达 5949428 词例，198241 词种，10813606 字。

表 1 语言学中文核心期刊摘要语料库概况

期刊	时间跨度	记录总条数	摘要数量	词例数	词种数	字数
外语教学与研究	1957-2020	3783	3443	308826	23640	573421
当代语言学	1962-2020	2310	1969	196766	18505	385596
现代外语	1978-2020	2648	2575	245678	19943	462929
外国语	1978-2020	3567	3355	316703	28601	648638
方言	1979-2020	1824	1702	126130	18454	224487
上海翻译	1979-2020	2976	2795	221340	20206	413212
中国翻译	1979-2020	5509	5171	468122	39394	878834
外语教学	1979-2020	4294	4117	394671	25443	715272
民族语文	1979-2020	3181	2783	179859	19534	324393
外语电化教学	1979-2020	3784	3614	321875	19196	588461
语言教学与研究	1979-2020	2852	2677	207521	15503	359776
外语教学理论与实践	1979-2020	2174	2151	217189	15612	395126
外语界	1980-2020	2786	2533	226961	13476	416330
汉语学习	1980-2020	3980	3488	351913	26654	597033
语文研究	1980-2020	2076	1973	171294	19724	290294
当代修辞学	1982-2020	5578	5034	532995	43538	907909

[2] https://github.com/BYVoid/OpenCC，本文选取“t2s.json Traditional Chinese to Simplified Chinese 繁体到简体”

[3] https://pypi.org/project/zhon/

[4] https://docs.python.org/3/library/string.html

[5] https://github.com/lancopku/pkuseg-python

[6] pkuseg 性能评估：https://github.com/lancopku/pkuseg-python/blob/master/readme/comparison.md

(c) 2022 中国中文信息学会计算语言学专业委员会

外语与外语教学	1984-2020	4664	4298	377824	24639	698950
世界汉语教学	1987-2020	2059	1592	158581	14522	311380
古汉语研究	1988-2020	2617	2433	175187	25101	293994
语言文字应用	1992-2020	2733	2315	187747	13808	333344
中国语文	1994-2020	2692	2176	213085	24228	367221
语言科学	2002-2020	1493	1243	125248	12876	219472
中国外语	2004-2020	1615	1588	156500	11401	292117
汉语学报	2004-2020	793	766	67413	8214	115417
总计	/	71988	65791	5949428	198241[7]	10813606

4. 以词汇增长建模语料库中的词汇历时变化

4.1 三种词汇增长模型

词汇增长模型假设词例与词种之间存在着某种函数关系，该关系能够预测在不同词例下的词种的数量。通过分别分析词种的预测值与观测值之间的差值与 TTR，能够反映不同时间段下词汇丰富度的变化情况。词汇增长模型主要有以下三种：

Guiraud（1954）提出对语料库的词汇增长进行建模，即转换为词种对词例数量的比值之间的关系。词种数量V与词例数量N的平方根的比值为常数c。若将该法则转换为预测模型，预测词种数量V'与当前词例数量n之间的关系可被表示为：

$$V'(n) = c \cdot \sqrt{n}$$

Heaps（1978）则提出，在双对数空间内，预测词种数量V'与当前词例数量n之间存在线性关系，并可通过如下关系式表达：

$$V'（n） = k \cdot \ n^b$$

其中，对一般的英文语料库来说，典型的k, b值处于以下区间内（Manning, Schütze, & Raghavan, 2008）：$30 \leq k \leq 100, b \approx 0.5$。例如，在 Reuters-RCV1 数据集（含有约10^5个词例）下，拟合得到的参数值为$k = 44, b = 0.49$。

Hubert and Labbe（1988）等人将词频因素作为建模的考量，提出了另一种基于词频高低的增长模型（Hubert & Labbe, 1988; Labbé, Labbé, & Hubert, 2004）。给定当前语料占总语料的比例u，则预测词种数量V'与参数u之间存在如下关系：

$$V'(u) = p \cdot u \cdot V + （1 - p）[V - \sum_{i=1}^{i=freq} V_i \, (1-u)^i]$$

其中，V_i代表词汇出现的频次为i的词汇的数量， p为预估参数，代表出现频次较低的词所占的比例。p的值受到文本的词汇多样性影响较大，能够体现不同文本的风格，Savoy（2015）应用该模型分析了不同时期美国总统演讲的语料。

上述三种模型均用于本文的实验以进行词汇增长拟合效果的对比。

4.2 词汇增长模型设置

本节对上面提到的三种模型使用非线性最小二乘法（Non-linear Least Squares）进行拟合。给定一组词汇增长采样点$\{(x_1, y_1), (x_2, y_2), \ldots, （x_n, y_n）\}$，用于拟合增长的模型$y = f(x; \boldsymbol{\theta})$,其中$\boldsymbol{\theta}$为若干模型参数集合$\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$，拟合过程中采用的残差为$r_i$，迭代过程中根据残差平方和进行拟合，形式化如下：

[7] 因不同期刊之间的词种有重复，因而此处不是每个期刊的词种数量相加。其他表格相同。

(c) 2022 中国中文信息学会计算语言学专业委员会

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{r_i^2}{2\sigma_i^2}$$

$$r_i = y_i - f\ (x_i; \boldsymbol{\theta})$$

实验通过拟合得到Heaps模型中参数 $k = 24.04, b = 0.58$，Guiraud模型中参数$c = 81.26$, Hubert-Labbe 模型中参数 $p = 0.00$（拟合参数触碰至搜索下界）。在下文中所得出的数据均基于这些参数。

在文本处理中，我们依时间对语料库中的文本每隔 1000 个点进行采样。此外，为了分析特定时间段内的词汇增长，我们在特定时间段内进行了新词检测实验。本文的新词是指在下一个时间段未出现在之前所有时间段的词，例如“新冠”一词首次在 2020 年的摘要中出现，但在 2020 年之前的摘要中没有这个词。新词检测由当前时间段的累积词汇集合减去之前所有的时间段的累积词汇集合得出，其中检测出的新词按照词汇出现频次进行排序。由于各区间段结果数量分布不一致（新词数量如表 4 所示），分析时统一筛选了排行前 50 的词语。

4.3 语言学核心期刊摘要语料库的模型拟合

不同词汇增长模型对语料实际增长情况的预测模式不一，但均无法处理由于现实语言生活因素带来的增长突变。在本例中，Heaps 模型对前期增长拟合较好，但在后期存在高估的趋势，Guirauds 模型则反之；Hubert-Labbe 模型与前两者相比整体趋势为低估增长，原因在于拟合得出的超参数p触碰到了搜索下界 0 值，使原本模型中低频词的部分失效，从而造成了较差的拟合表现。图 1 显示了不同模型对实际观测值（采样点）的拟合情况，曲线代表词种与词例数量之间的增长关系，残差r_i总和显示 Heaps 模型对现实增长拟合最好。在增长起始与中间阶段（20 万词例与 30 万词例间），各个模型的拟合性均出现了较大的偏差。图 2 则将图 1 中 Heaps 模型拟合得到的增长曲线与关键的年份时间点结合进行可视化，关键年份点的划分方法与依据将在本文第 5 节进行详细阐述。

图 1 不同模型拟合与现实词汇增长的曲线

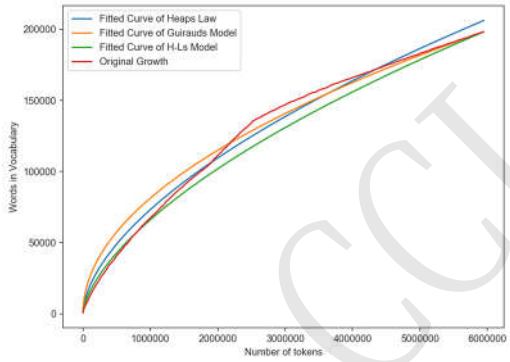


图 2 现实词汇增长情况在年份区间上的映射

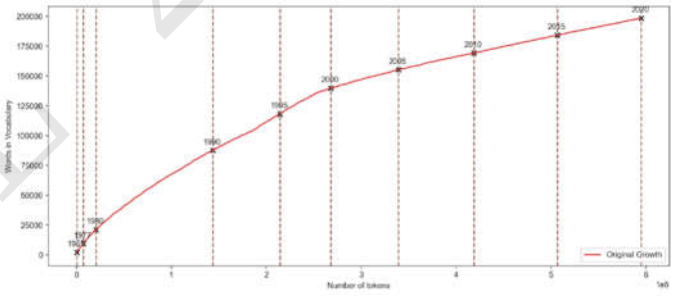
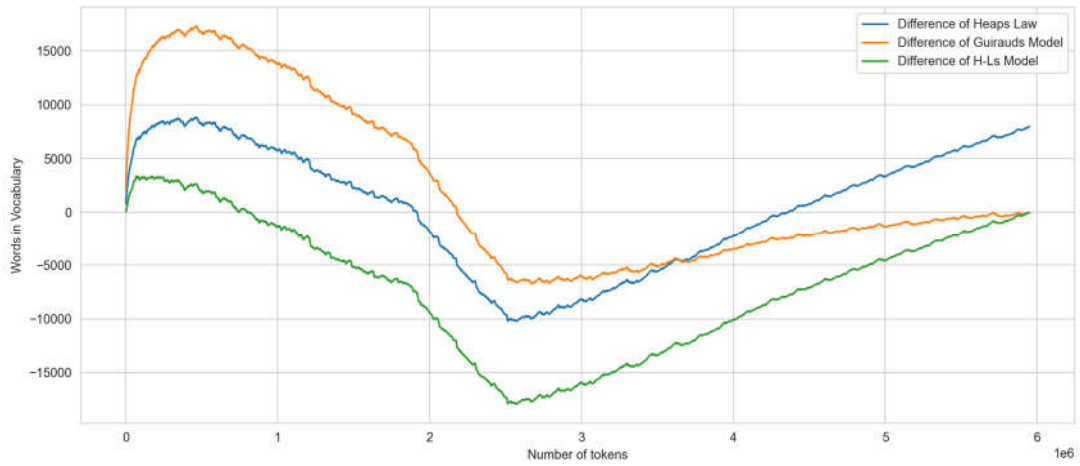


图 3 显示了模型在词汇增长预测中与实际观测值的偏差情况。从总体上看，增长起始阶段大部分模型预测值高于现实文本的词汇增长，现实增长的词种无法达到模型预期，这可能是由于学科发展缓慢造成的。而在中期预测拟合值则低于实际观测值，显示该阶段现实语言生活的中的重大事件可能较多，学科的快速发展引入的新词数量远高于模型预测。后期二者之间的偏差趋于稳定。可能为学科发展稳定亦或是特殊事件导致的语言生活各方面发展稳定。总体上看，利用三大模型的实验结果显示，Heaps 模型的拟合效果最好。下文将结合该模型的结果，对语言学发展的各个时期进行分析。

图 3 不同词汇模型的预测误差曲线



5. 语言学核心期刊摘要的词汇增长的历时分析

学科的发展具有一定的阶段性，若使用每一年作为分析的周期，容易受该年份随机性的影响，结果常具有偶然性。现有的研究也都采用分期的方式，例如刘丹青（2019）基于理论与应用研究发展历程，将中国语言学工作的发展总结为三段，即“封闭自足—对外开放—自主创新”；郭熙（2019）在考察政策、文献、年份代表性词语、语言产品的基础上对 70 年来中国的语言生活总结为两个阶段：前三十年（1949～1978 年）和后四十年（1979～2018 年）。本文先利用词汇增长模型对大规模语料得出词汇历时演化的基本趋势（第 4.3 节），再针对宏观趋势结合时代关键节点分析影响演化趋势的主要原因。

本研究采用的分期情况如下：（1）将 1957~1980 年视作特殊的学科发展起始阶段。一方面，建国后各项科学文化工作百废待兴，加之 1966~1976 年间文化大革命的爆发令学科发展停滞，大部分期刊停刊或未保有出版记录，可供观测分析的样本较少，1978 年前的期刊情况如表 2 所示。另一方面，改革开放初期大量刊物创刊并开始出版，1978-1980 年的期刊情况如表 3 所示。（2）1980 年后期刊多样且数据较为平衡，该段以十年为周期进行分段，分为 1980 年代（即 20 世纪 80 年代）、1990 年代、2000 年代、2010 年代，具有整体性与稳定性。语言学是一门与社会关联性紧密的学科，这样分期便于指称不同时代的特点。每阶段的观测值与 Heaps 模型的预测值数据如表 4 所示，预测值与观测值的差值变化及在年份上的映射情况则如图 4 所示（与 Hubert-Labbe 模型相对比），整体趋势同 4.3 中所述一致。

罗常培（1989）指出，“（88 页）语言的内容足以反映出某一时代社会生活的各方面的侧影响。社会的现象，由经济生活到全部社会意识，都沉淀在语言里面”。故透过计量得到对预测影响较大的词汇，从语言与社会的关系出发，进行语言生活中的重要事实的分析，是探究词汇增长模型预测差异的恰当的切入点，而学术文献也可折射语言学这一学科发展的不同阶段。

表 2 1957-1977 年考察期刊概况

期刊	时间跨度	文献数量	词例数	词种数	字数
外语教学与研究	1957-1977	568	45204	6390	78861
当代语言学	1962-1977	268	26328	4779	49651
总计	/	836	71532	9057	128512

表 3 1978-1980 年考察期刊概况

期刊	时间跨度	文献数量	词例数	词种数	字数
语文研究	1980-1980	15	1787	781	3073
方言	1979-1980	54	8104	2529	14440
外语教学理论与实践	1979-1980	90	9633	2423	17244
外语界	1980-1980	24	1952	921	4052

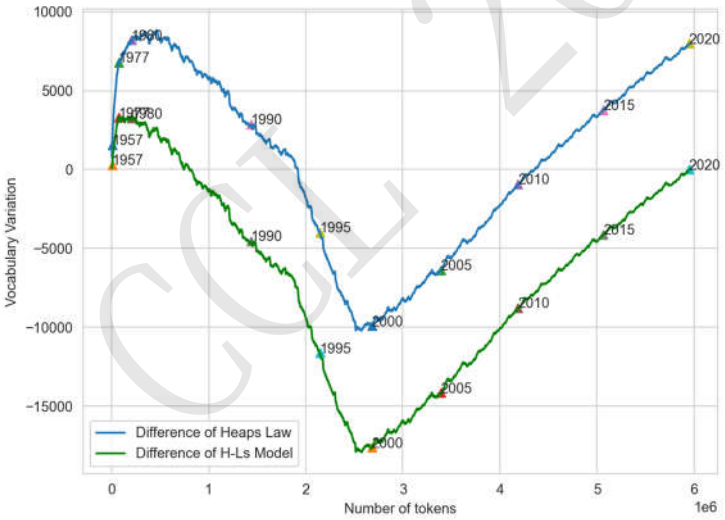
(c) 2022 中国中文信息学会计算语言学专业委员会

外语电化教学	1979-1980	90	8953	2454	16072
外国语	1978-1980	206	19206	4465	36232
汉语学习	1980-1980	87	13589	3099	22111
外语教学与研究	1978-1980	164	12200	3187	22662
中国翻译	1979-1980	76	6660	2451	13371
现代外语	1978-1980	149	9872	3202	21157
语言教学与研究	1979-1980	110	6158	1845	10469
当代语言学	1978-1980	279	24638	5065	48818
民族语文	1979-1980	99	6901	2153	12423
外语教学	1979-1980	53	4270	1644	8154
总计	/	1496	133923	16894	250278

表 4　每阶段现实观测值与 Heaps 预测值

时间段	累积词例	累积词种	Heaps 预测值	Heaps 预测值-观测值/累积词种	观测值 TTR	预测值 TTR	新词语数量（在之前所有阶段未出现的词）	累积文献数量
1957-1977	71532	9057	15833	6776	0.12661	0.22134	9057	836
1957-1980	205455	21001	29215	8214	0.10222	0.14220	11944	2332
1957-1990	1436173	87482	90352	2870	0.06091	0.06291	66481	15279
1957-2000	2679564	139687	129779	-9908	0.05213	0.04843	52205	30211
1957-2010	4187052	169098	168171	-927	0.04039	0.04016	29411	48353
1957-2020	5949428	198241	206222	7981	0.03332	0.03466	29143	65791

图 4 词汇增长模型预测误差曲线在年份区间上的映射



在初创阶段（1949~1980 年），区间终点采样点的模型预测值高于实际观测值 8214 个词种。由于建国初期百业待兴、经济基础薄弱，全民文化教育水准也普遍较低，文盲率高达 80%，该阶段我国较为重要的语言生活事件集中在国家所开展的文字改革，因此 50 年代我国语言学科发展较为缓慢。由于普通话、汉语拼音方案的推广，该阶段的学科研究热点关注汉语语音研究，例如 1958 年全国人大批准颁布《汉语拼音方案》（中华人民共和国第一届全国人民代表大会, 1958）前，语言学界已于 1957 年发表《北京语音音位简述》（徐世荣, 1957）对普通话音位系统给出了理论描述，为普通话音位系统深化研究奠定了基础。从计量结果来看，“变调”作为词频最高的新词出现，其他出现次数较高的词语还包括“声”、“韵”、“调”、“音标”、

“上声”、“阳平”、“阴平”等。但随之而来的十年文化大革命扰乱了语言生活的正常秩序，更令刚起步的学科受到重创，该区间段的高频新词中出现了“四人帮”，间接表明文化大革命以及社会因素对该阶段学科发展的影响之大。另一方面，在语言生活的国际交流层次与前苏联关系密切，使得语言学研究领域分类格局等受到苏联影响较大，早期文献（B·A·阿尔乔莫夫、郑昌荣、周毅, 1958; 林学洪, 1958; 刘世沐, 1958; 王鸿斐, 1958） 中出现的如“资产阶级”、“社会主义”、“苏联”、“斗争”等，印证了该时期语言生活的意识形态与语言学发展受到苏联等社会因素的制约。

1981~1990 年间，区间终点模型预测值高于观测值 2870 个词种，且从趋势上看模型的预测正偏差开始急剧地向负偏差增长，表明该阶段学科的发展开始恢复，并具有增长的趋势。相较于之前 30 年间政治运动因素是影响社会语言生活的重要因素（郭熙, 1999），改革开放带来的思想开放与新概念、新事物的引入使得语言生活中的政治成分下降，所涉及的主题逐渐丰富。首先，语言生活学术环境极大改善，文革后期刊的复刊（如《中国语文》）与雨后春笋般出现的新刊物（如《方言》等，见表 3）为语言学工作者构造了良好的空间，“辞格”、“喻体”、“义项”、“借代”等新增词占据高频榜单，代表着语言生活中对学术研究探求的正常化。而该阶段新增高频词中“测试”、“考生”、“试卷”、“托福”等说明了语言教育生活的正常化及进阶化，尤其是 1986 年《中华人民共和国义务教育法》的通过使得越来越多儿童能够接受正规的语言教育，且恢复高考后高等教育体系的不断完善使得中国语言学的硕博士点不断增加，进而为语言学科发展输送了大量人才。与此同时，学界对外关闭的大门得以敞开，不少论文介绍西方语言学理论，使语言学的理论基础与研究范式得以提升，丰富了汉语研究方法，研究的视角不再只聚焦于印欧语系的语言特点，例如该阶段徐烈炯对生成语法的介绍（徐烈炯, 1988）。此外，新的研究方法与视角也不断涌现，例如 80 年代初在《中国语文》期刊上对析句方法的讨论（华萍, 1981; 陆俭明, 1981）。 “国家教委”是该阶段后五年出现频率最高的词，体现了核心期刊选题方向依然以国家重要机构的指导政策作为参考，但同时语言学的发展同时也受改革开放的时代因素影响。一方面，人们语言生活的接触面增加，例如“录像机”、“录像带”等词在后五年文献语料中的关注度极高；另一方面，语言生活的国际接触面变广，催生了语言生活中新的教育需求，“HSK”、“二语”等作为新增高频词出现在后五年的文献中，这与汉语国际教育的兴起有直接联系。众多大学开设汉语国际教育学院，并逐渐形成为一个独立的学科，有《世界汉语教学》、《语言教学与研究》等期刊作为高质量学术交流平台，涌现出大量的基础性工作，奠基了该领域理论研究的基础。

1991~2000 年间，模型预测呈现负偏差，即预测值相对于语言学领域期刊的实际词汇增长情况低 9908 个词种。该段时间的语言生活可视作上一阶段的领域持续深化，主要为学术生活发展与教育需求强化。该时间段汉语语言学理论研究与国际语言学前沿接轨，部分基础理论的介绍已跟进国外较新的研究成果。例如上一时间段已有对生成语法的介绍，“Chomsky”在该年份间就出现在新词检测的结果中，进一步的数据调阅则显示已有基于 Chomsky 提出的最简程序（Chomsky, 1995）解释汉语中“得”字句的工作（杨寿勋, 1998）。同时改革开放深化的影响持续体现在语言生活中，“市场经济”、“韩国”、“流行语”位列前五年词频前三的新词，体现语言生活中的新现象与大环境的改变；后五年中，“因特网”、“互联网”等极富时代特点的词出现，表现出新的沟通方式扩展了语言生活的媒介。同时 1993 年《中国教育改革和发展纲要》、《关于重点建设一批高等学校和重点学科点的若干意见》等教育领域的重大文件相继发布，随即这一时代的新词中包括“CET-4”、“EFL”等与我国高等教育密切的词，高校培养新时代高等人才的迫切需求推动了在大环境下对英语二语教育相关的研究，表明语言生活中国民高等教育的进一步发展。

2001~2010 年，模型预测值比观测值低 927 个词种，预测误差开始出现由负转正的趋势，现实观测值增长速率有所放缓，但模型预测值依旧偏低。主要原因是语言生活与时俱进使语言生活的内容形式不断丰富，学界发展仍有诸多新话题可供探究，但经过前阶段的介绍和引入，整体学科体系与基础理论趋于完善。随着加入世贸组织带来的更多机遇与挑战，语言生活中的诸多社会现象受到语言学界关注，如“WTO”、“农民工”分别作为前、后五年词频最高的新词出现，从“女性主义”、“以人为本”等词也同样可见一斑。而随着语言生活的沟通媒介不断丰富，“手机”、“博客”、“聊天室”等关键词的出现表明已有一部分工作将新时代的语言

(c) 2022 中国中文信息学会计算语言学专业委员会

模式作为研究对象。另一方面，就语言生活的研究方法而言，现代化分析手段被引入，如“SPSS”、“体验式”、 “语料库”与“PPT”等。与语言教学特别是二语教学相关的研究也进一步发展，尤其是标准语言测试的主题增加，英语等级考试（CET）相较于前十年词频上升，并得到持续关注，“TEM-4”作为新的专业英语等级考试也开始被取材为研究对象；二语领域新增词汇术语也频繁出现在该时代的文献中，如 “二语词”、“ 二语朗读”等。

2011~2020 年，该阶段的发展相较于前两个十年的增长速率放缓，模型的预测高于观测值 7981 个词种，可能是由于前二十年较多概念与观点的介绍，使我国语言学研究打下了理论基础。从这个时期出现的新词来看，此时语言生活相伴发展的时代特征较为突出，透过新词频率排行可发现，该阶段新媒体媒介的使用对语言生活的影响较大，例如“翻转”、“MOOC”、“微课”等教学工具的引入体现了新时代对语言教学方法的思考、而“大数据”、“微博”等词则体现了对于大数据时代对语言生活新环境下的网络用语关注及研究方法上对语料库范式的利用。而随着语言生活领域的扩展，政治、医疗等领域开始重视语言问题及探究语言生活与该领域的关联与影响，例如“习近平”、“十九大”、“领导力”等词关注了国家政治人物和政治用语，新型冠状病毒即使作为 2020 年的医疗新事件，在该时间段的文献中“新冠”也作为新词出现（12 次），在本语料库中，该词最早出现于《世界汉语教学》（田列朋, 2020）。

综上所述，模型能够从理论上预测一定时间段内语料库的词汇增长趋势，但现实中词汇如何增长会受到受到现实因素影响而呈现不同的趋势，语言学的词汇变化也是受语言生活与国家语言规划政策的影响而变化的通过预测偏差分析能够体现这这一进程，便于我们找出整体词汇增长趋势的关键节点，如图 5 所示。一方面，因为国家的语言政策直接影响着语言生活与语言的时代特征，对语言文字的研究活动也属于语言生活的范畴（李宇明, 1997），故语言学研究会受到国家政策的影响；另一方面，这种影响并非是单向的，在语言学的研究中，对语言生活的研究亦会影响国家的语言政策。“就语言生活为语言生活而研究语言和语言生活”的学术主张（李宇明, 2016）促成了中国语言生活派的形成，语言生活派的研究成果与关注点也为和谐语言生活的建立与语言学研究带来了不可忽视的影响。李宇明（2010a）对外语规划和公民外语素养的提升提出了思考和建议，在对应年份的文献高频词中便有二语教育的内容，正如前文中 2001～2010 阶段中新词内容所体现的那样，2013 年教育部亦启动《大学英语教学指南》的研制工作；李宇明（2010b）指出需要关注城市化进程的语言问题，并从语言规划层面着重考虑，同样是 2006～2010 年份区间段的研究热点。此外，《中国语言生活状况报告》（又称“中国语言生活绿皮书”）的逐年发布也能够为国家政策提供参考（李宇明, 2007），为学界、社会与国家语言规划之间建立起了桥梁。由此可见，词汇增长模型的误差分析对语言学发展的刻画不但能够大致勾勒所分析的区间段的热点，也能够体现出国家语言规划等对语言学研究的指引。

图 5 词汇增长模型预测误差分析框架图

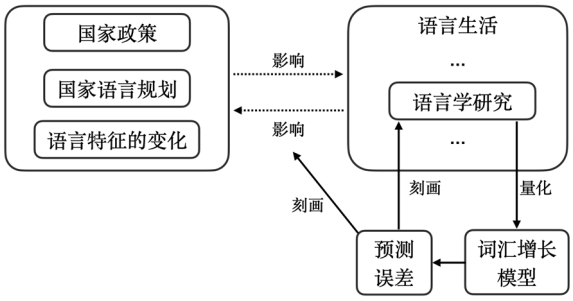
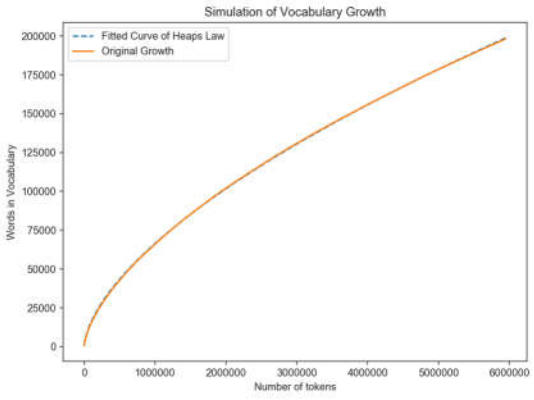


图 6 随机文本增长与预测曲线



6. 词汇历时变化的建模有效性验证

上一节在 Heaps 模型预测结果的基础上，从定量结合定性的角度分析了词汇实际增长情况，阐释了模型拟合的差异受社会发展、语言生活变化等因素的影响。上述分析基于模型的

理论预测值与现实观测值的误差进行分析，因此合理的模型理论预测值是保障研究结论正确的一大重要前提。本节将从定量角度进一步验证所使用模型的理论正确性，参考 Savoy（2015），本文采用随机文本的方法进行验证：即将文本以词语为单位随机打乱，使其失去原有的词汇分布特征和时间信息，从而得到一个新的学术文本语料库，该语料库既不具备时间序列上的增长关系，也不具备词语之间语义分布的关联，能够在消除外部影响因素的情况下，仅从语料库计量的角度验证模型是否能够成功拟合重新组织后的文本。对乱序后的文本，我们采用与 4.2 节一致的设定，每隔 1000 个词例进行采样，对采样点集合采用 Heaps 模型进行拟合，得到的参数为$k = 13.64, b = 0.61$，拟合曲线如图 6 所示，显示出 Heaps 模型在随机文本上较高的拟合精度。为了进一步探究该拟合结果，采用 Z-Score 指标进行拟合差异的评估，公式如下:

$$Diff（n） = V'(n) - V(n)$$

$$Z_{score} = \frac{Diff(n) - \mu_{Diff}}{\sigma_{Diff}}$$

图 7 和图 8 分别显示了随机文本与现实文本下不同采样点的 Z-Score 分布情况，两个文本的标准评分均处于（-2，2），即数据均在可接受的范围$[-3\sigma, 3\sigma]$内，符合正态分布，证明拟合模型所带来的差异是可以接受的。此外，随机文本中的预测误差显示出一定的随机性，而现实文本的预测误差 Z-Score 分布与上述年份的分析变化保持一致。就总体而言，现实文本差异的变化范围较随机文本大，而这种差异可能是由于现实因素所造成的，而非拟合模型的失误。

图 7　随机文本的预测误差 Z-Score 分布

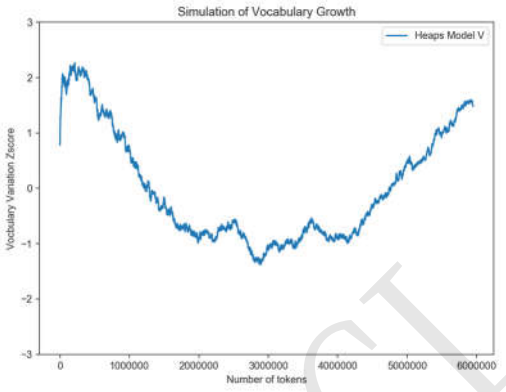
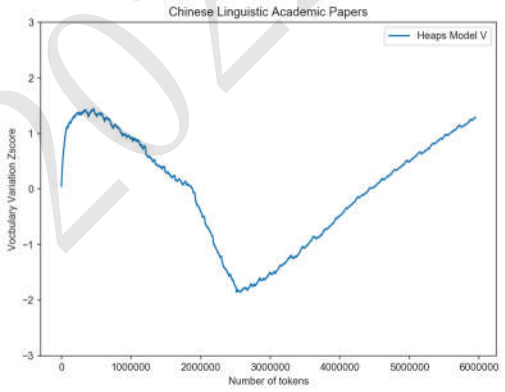


图 8　随机文本的预测误差 Z-Score 分布



通过在随机文本下的拟合实验，一方面证明前面实验中选择的模型能够较好地对语言学领域文本的词汇增长进行拟合，同时拟合的鲁棒性受到重大政治事件、经济变化、热点社会问题等现实因素的影响较小。另一方面，在实际拟合过程中由外部因素与学科发展所造成的预测误差是不可避免的，而无论定性还是定量角度均支撑了这些误差存在的合理性。

7. 结论

建国以来我国语言学历经 70 年的发展取得了瞩目成就，已有研究对语言学发展史上的重要事件进行了回顾和介绍，但尚未有使用量化手段对其历时演变的分析。为探究建国以来中国语言学的发展，首先，本文搜集 CSSCI（2019-2020）收录的中文核心期刊于 1957~2020 年间的摘要，构建了 5949428 词例、198241 词种、10813606 字的大规模语料库。第二，在该大型语料库基础上，本文采用三种词汇增长模型作为建模手段进行拟合，其中 Guiraud 模型（Guiraud, 1954）利用语料库中词种与词例的平方比值关系进行建模，Heaps 模型（Heaps, 1978）则基于双对数空间下二者的线性关系，Hubert-Labbe 模型（Hubert & Labbe, 1988; Labbé *et al.*, 2004）在 Heaps 模型（Heaps, 1978）基础上融入了对词频高低的考量。第三，本文选取拟合误差较小的 Heaps 模型对拟合结果以每 10 年为周期，借助量化统计结果与不同阶段的新词进行分析，研究结果体现出语言学发展与社会发展的关联性。根据模型的预测差值的变化趋势不同，可以看到我国语言学的发展经历了三阶段：起步阶段封闭停滞、改革开放后蓬勃发展、近十年逐步迈入自主创新阶段。这一结果符合实际发展趋势，从定性角度支撑了

本文所采取的研究方法的可行性。第四，本文打乱语料库的词序，使之失去时序信息，在随机文本上进行的验证程序证实了 Heaps 模型在探究中文词汇演化这一研究主题上的有效性与合理性。在未来研究中，我们将开展以下工作：1）根据本文收集的历时语料探索能够对未来学术文本增长趋势以及潜在话题进行预测的模型；2）根据现有研究框架进一步对语言学之外的人文社科其他领域以及科学领域核心期刊的词汇增长进行历时考察；3）利用已建立的学术汉语语料库，将提取语言学摘要的学术词表、核心词表等，为学术写作的研究和教学提供资源；4）构建英语学术论文摘要语料库，开展汉英学术论文摘要写作的对比研究，为一语和二语学术写作提供语料库驱动的教学建议。

致谢

本研究受国家语委（项目号：YB135-159）和澳门大学（项目号：MYRG2019-00013-FAH）科研项目资助。

参考文献

B·A·阿尔乔莫夫, 郑昌荣, 周毅. 1958. 苏联外语教学心理学四十年(1917—1957). 《外语教学与研究》(02), 129-140.

郭熙. 1999. 《中国社会语言学》，南京: 南京大学出版社.

郭熙. 2019. 七十年来的中国语言生活. 《语言战略研究》(04),14-26.

国家语言文字工作委员会. 2019. 《新中国语言文字事业发展 70 年纪事》，北京：语文出版社.

胡芳, 陈彧. 2005. 英汉学术期刊论文摘要衔接手段的对比研究. 《武汉科技大学学报(社会科学版)》(03), 83-86.

华萍. 1981. 评“暂拟汉语教学语法系统”. 《中国语文》(02).

李志君. 2014. 汉英学术语篇中作者身份指称语使用的调查与分析. 《外国语言文学》31(02), 81-89.

林学洪. 1958. 苏联的翻译事业. 《外语教学与研究》(04), 434-439.

刘丹青. 2019. 《新中国语言文字研究 70 年》，北京：中国社会科学出版社.

刘世沐. 1958. 必须坚决地清除语言学研究中的资产阶级立场观点和方法——评李赋宁同志“英国民族标准语的形成与发展”一文. 《外语教学与研究》(03), 261-265.

陆俭明. 1981. 分析方法刍议——评句子成分分析法. 《中国语文》(03).

陆俭明. 1999. 新中国语言学 50 年. 《当代语言学》(04), 3-5.

李宇明. 1997. 语言保护刍议. 《双语双方言》（五)，香港: 汉学出版社.

李宇明. 2007. 关于《中国语言生活绿皮书》，《语言文字应用》(1), 12-19

李宇明. 2010a.《关注中国城市化进程中的语言问题》.《中国语言生活状况报告 2009》，北京：商务印书馆。

李宇明. 2010b. 中国外语规划的若干思考. 《外国语（上海外国语大学学报)》(1).

李宇明. 2016. 语言生活与语言生活研究. 《语言战略研究》（3), 15-23.

刘锐, 王珊. 2017. 《两岸中文学术常用词对比研究》.《第三届国际汉语教学研讨会论文集》, 香港教育大学，香港.

罗常培. 1989. 《语言与文化》，语文出版社.

马费成. 2000. CSSCI 与社会科学评价. 《南京大学学报(哲学.人文科学.社会科学版)》(04), 155-160.

苏新宁. 2012. 中文社会科学引文索引(CSSCI)的设计与应用价值.《中国图书馆学报》38(05), 95-102.

田列朋. 2020. “战疫语言服务团”用语言利器助力抗疫. 《世界汉语教学》 34(02), 231.

王鸿斐. 1958. 科学研究一定要政治挂帅. 《外语教学与研究》(04), 422-423.

王珊, 陈钊, 张昊迪. 2021. 《近十年来澳门的词汇增长》. 《第 20 届中国计算语言学大会论文集》, 内蒙古大学，内蒙古.

(c) 2022 中国中文信息学会计算语言学专业委员会

王珊, 王会珍. 2021. 中文词汇增长研究. 《中文信息学报》, 35(1), 17-24.
王笑然, 王佶旻. 2022. 经贸类本科专业学术汉语词表研究. 《语言教学与研究》(04), 9-19.
吴格奇, 潘春雷. 2010. 汉语学术论文中作者立场标记语研究. 《语言教学与研究》(03), 91-96.
徐烈炯. 1988. 《生成語法理論》, 上海外語教育出版社.
徐世荣. 1957. 北京语音音位简述. 《语文学习》(08), 22-24.
杨寿勋. 1998. “得”的生成语法研究. 《现代外语》(01), 52+51+53-73.
张赫, 李加, 申盛夏. 2020. 学术汉语的词汇使用特征研究. 《语言教学与研究》(06), 19-27.
赵永青, 薛舒云, 邓耀臣, 徐建伟, 丁科家.（2019). 同一期刊论文英汉摘要作者立场信息一致性研究. 《解放军外国语学院学报》 42(03), 73-81+160.
中华人民共和国第一届全国人民代表大会. 1958. 《中华人民共和国第一届全国人民代表大会第五次会议关于汉语拼音方案的决议》. http://www.moe.gov.cn/ewebeditor/uploadfile/2015/03/02/20150302165814246.pdf
朱宇, 胡晓丹. 2021. 汉语连词在不同学术语域的聚合：多维度定量分析. 《语言教学与研究》(02), 57-69.
邹志仁. 2000. 中文社会科学引文索引(CSSCI)之研制、意义与功能. 《南京大学学报(哲学.人文科学.社会科学版)》(04), 145-154.
Chomsky, N. 1995. *The minimalist program*. Cambridge: The MIT Press.
Guiraud, P. 1954. *Les caractères statistiques du vocabulaire: essai de méthodologie*: Presses universitaires de France.
Heaps, H. S. 1978. *Information retrieval, computational and theoretical aspects*: Academic Press.
Hoover, D. L. 2003. Another perspective on vocabulary richness. *Computers and the Humanities, 37*(2), 151-178.
Hubert, P., & Labbe, D. 1988. A model of vocabulary partition. *Literary and Linguistic Computing, 3*(4), 223-225.
Hyland, K. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse studies, 7*(2), 173-192.
Labbé, C., Labbé, D., & Hubert, P. 2004. Automatic segmentation of texts and corpora. *Journal of Quantitative Linguistics, 11*(3), 193-213.
Luo, R., Xu, J., Zhang, Y., Ren, X., & Sun, X. 2019. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. *CoRR, abs/1906.11455*.
Manning, C. D., Schütze, H., & Raghavan, P. 2008. *Introduction to information retrieval*: Cambridge university press.
Mellor, A. 2010. *Automatic essay scoring for low level learners of English as a second language.*（Ph.D.). Swansea University（United Kingdom), Ann Arbor.
Paquot, M. 2010. *Academic vocabulary in learner writing: From extraction to analysis*: Bloomsbury Publishing.
Savoy, J. 2015. Vocabulary growth study: an example with the State of the Union addresses. *Journal of Quantitative Linguistics, 22*(4), 289-310.
Swales, J. 1985. *Episodes in ESP: A source and reference book on the development of English for science and technology*（Vol. 1): Pergamon.
Wang, S., Liu, X., & Zhou, J. 2022. Readability is decreasing in language and linguistics. *Scientometrics*. doi:https://doi.org/10.1007/s11192-022-04427-1
Wang, X. 2014. The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL, 9*.
Yu, G. 2010. Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics, 31*(2), 236-259.

(c) 2022 中国中文信息学会计算语言学专业委员会

一个适合汉语的带有范畴转换的组合范畴语法

王庆江，陈淑娴
华北水利水电大学 信息工程学院，河南 郑州 450046
wangqingjiang@ncwu.edu.cn, 1152825510@qq.com

摘要

为使汉语句子中词或短语的范畴对应其句法功能，在组合范畴语法中添加范畴转换。把词类和短语结构的范畴分别按出现率和是否由结合规则得到分为典型和非典型，建立短语结构中词类和短语结构的范畴转换规则。实词或短语结构通过范畴转换与虚词搭配，让虚词的句法功能趋于明确。树库显示，35%的短语结构形成需要范畴转换，使用范畴转换的短语直接成分中99.67%是实词或短语结构，范畴转换使组合范畴语法适合缺乏屈折的汉语。

关键词： 组合范畴语法 ； 范畴转换 ； 范畴类型透明性 ； 树库

A Chinese-Suitable Combinatory Categorial Grammar with Categorial Conversions

Qing-jiang Wang, Shu-xian Chen
School of Information Engineering
North China University of Water Resource and Electric Power
Zhengzhou 450046, China
wangqingjiang@ncwu.edu.cn, 1152825510@qq.com

Abstract

To make categories of words or phrases in Chinese sentences correspond with their syntactic functions, categorial conversions are added into Combinatory Categorial Grammar. Categories of parts of speech and phrasal structures are divided into the classical and the non-classical respectively by occurrence rate and whether obtained via combinatory rules, category conversion rules are established for parts of speech and phrasal structures in phrasal structures. Notional words or phrasal structures collocate with functional words by categorial conversions, making syntactic functions of functional words tend to definite. Treebank shows, 35% of phrasal structure formations require categorial conversions, and 99.67% of phrasal immediate components using categorial conversions are notional words or phrasal structures, and categorial conversions make CCG adapt to inflectional-absent Chinese.

Keywords: Combinatory Categorial Grammar , categorial conversion , categorial type transparency , treebank

1 引言

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：河南省重点研发与推广专项（212102210495）

在深度学习推动自然语言处理全面发展的时代，语言语法的研究仍然重要(冯志伟, 2021)。组合范畴语法[†] (Steedman, 2019)（Combinatory Categorial Grammar，缩写CCG）是范畴语法的一种扩展，用类型提升和函数组合解释宾语提取、状中、中补等短语结构，用斜线类型将句法类型结合的精确控制由规则一侧转到词汇一侧，使规则跨语言通用，且仍具有句法类型结合伴随语义组合的范畴语法亮点 (Jayant and Mitchell, 2014)，这些使CCG具有重要的计算语言学价值 (陈鹏, 2016; 满海霞, 2022)。

CCG给词指派句法类型（即范畴）及其关联的语义解释，例如 (Steedman, 2019)给sees指派及物动词范畴和解释其语义的λ-项，sees := (S\NP$_{3s}$)/NP: $\lambda x\lambda y$.sees' x y，这里句法类型(S\NP$_{3s}$)/NP指右结合NP得句法类型S\NP$_{3s}$，后者左结合NP$_{3s}$得句法类型S，语义式$\lambda x\lambda y$.sees' x y在先后应用于右侧句法类型NP相伴的语义式x'、左侧句法类型NP$_{3s}$相伴的语义式y'后得到谓词-论元结构sees' x' y'。一个句法类型对应一种句法功能，而词类是词按句法功能分的类，故可按词类考虑词的句法类型。但是汉语里"词有定类"和"类有定职"两难 (胡明扬, 1995)，词类问题复杂，厘清词类问题后才能基于词类给词指派范畴。

另一方面，范畴语法的规则通常指两个范畴结合为一个范畴的规则，但其实还可以是一个范畴转为另一个范畴的规则 (Carpenter, 1991)，前者是跨语言通用的范畴结合规则，后者是因语言而异的范畴转换规则。若结合衍生的范畴不对应短语结构充当的句法成分，就需要转换范畴实现对应。汉语里短语结构缺乏屈折，存在大量的这种不对应。

关于词类及其句法功能，语言学界经过漫长争论，逐渐倾向词有定类而类无定职，并可按出现率把词类的句法功能分为典型和非典型，而短语结构的范畴也可按是否由结合规则得到分为典型和非典型，这样就可以对词类和短语结构统一讨论范畴从典型到某个非典型的转换，建立一个带有范畴转换的组合范畴语法（CCG with Category Conversions, 缩写CCC-C^2）：词典里力争词与范畴一一对应，在句子结构里可以通过范畴转换给同一词类或短语结构指派不同范畴。CCC-C^{2}中的范畴转换是句法层面的，与文献 (Carpenter, 1991)"词:=范畴"闭包中范畴转换的不同在于，后者是词法层面的。要使CCG-C^{2}适合汉语的各种句法特征，只能通过建立树库，才能逐渐明确词类和特殊词的典型范畴，形成词类和短语结构的范畴转换规则体系。

本文的创新性工作有两个方面：（1）在句法层面统一考虑汉语词类和短语结构的范畴转换，即词类和短语结构都可以按所在短语结构的需要由典型范畴改用非典型范畴。（2）通过构建树库，形成适合汉语的范畴转换规则体系。

本文第2节给出组合范畴语法的基本定义；第3节阐述范畴转换规则的汉语言学依据；第4节按范畴结合规则建立句法成分到范畴的一一映射，使实词类或短语结构有典型范畴，由虚词短语的典型范畴得到虚词的典型范畴，并论述一些特殊词的范畴设立依据；第5节举例说明短语结构中的各种范畴转换规则；第6节由树库统计评价这个带有范畴转换的CCG；最后总结工作，指出下一步的研究方向。

2 组合范畴语法的基本定义

英语的基本范畴一般有n、np、pp和s，分别对应名词、名词短语、介宾短语和句子，若考虑数、格、屈折等特征，可有更多基本范畴。汉语里数、格、屈折不明显，本着能归于其他基本范畴就不单设的想法，基本范畴可只有np和s。范畴（Category）用巴科斯范式（BNF）定义如下，其中斜线（Slash）类型'.'、'$\diamond$'、'$\times$'、'$\star$'实现词对规则选择的控制 (Steedman, 2019)，含有斜线的范畴是衍生范畴。若斜线左侧或右侧是衍生范畴，要用圆括号括起来，以保持二分性，例如(s\.np)/.np。衍生范畴可看作函数，斜线左侧为函数结果，右侧为函数参数，方向反映范畴序列中参数位于函数的哪一侧。

SyntaxType ::= np | s | SyntaxType Slash SyntaxType

Slash ::= /. | \. | /$_\diamond$ | \$_\diamond$ | /$_\times$ | \$_\times$ | /$_\star$ | \$_\star$

范畴语法的规则有关联的语义运算，且有类型透明性 (Steedman, 2019)，即句法类型决定语义类型。高阶组合规则（>B^n、<B^n、>B$^n_\times$、<B$^n_\times$）可以无限枚举，但目前也只发现二阶组合的作用，如前向二阶组合规则（>B^2）允许副词或能愿动词与双宾语动词的结合。提升规则

[†]早期对范畴语法（CG）的扩展，即对相邻范畴的包裹（Wrap）、组合、提升、替换等函数运算，本质上都是结合的 (Steedman, 2011)，都源于Moses Schönfinkel的结合子（Combinator），故Combinatory Categorial Grammar应该是这一类范畴语法扩展的总称或者是基于各种结合子的范畴语法，并译为"结合范畴语法"。Mark Steedman的CCG主要是新增了结合子Z（Haskell Brooks Curry称之为B）对应的函数组合，这也许是国内将CCG译为"组合范畴语法"的原因。

(c) 2022 中国中文信息学会计算语言学专业委员会

（<T、>T）总是和组合规则连用。用下列规则形成汉语各种短语结构，‘⇒’左边匹配短语内部结构，右边得到短语整体功能。在范畴关联的语义部分，a是常量，z、w是变量，f、g是函数，函数也可表示为λ-抽象。

$$\mathrm{X}/_{\star}\mathrm{Y}{:}f \quad \mathrm{Y{:}a} \Rightarrow \mathrm{X}{:}f\ \mathrm{a} \qquad (>)$$
$$\mathrm{Y{:}a} \quad \mathrm{X}\backslash_{\star}\mathrm{Y}{:}f \Rightarrow \mathrm{X}{:}f\ \mathrm{a} \qquad (<)$$
$$\mathrm{X}/_{\diamond}\mathrm{Y}{:}f \quad \mathrm{Y}/_{\diamond}\mathrm{Z}{:}g \Rightarrow \mathrm{X}/_{\diamond}\mathrm{Z}{:}\lambda z.\, f\,(g\,z) \qquad (>\mathrm{B})$$
$$\mathrm{Y}\backslash_{\diamond}\mathrm{Z}{:}g \quad \mathrm{X}\backslash_{\diamond}\mathrm{Y}{:}f \Rightarrow \mathrm{X}\backslash_{\diamond}\mathrm{Z}{:}\lambda z.\, f\,(g\,z) \qquad (<\mathrm{B})$$
$$\mathrm{X}/_{\diamond}\mathrm{Y}{:}f \quad (\mathrm{Y}/_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z{:g}} \Rightarrow (\mathrm{X}/_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z}{:}\lambda z\lambda w.\, f\,((g\,z)\,w) \qquad (>\mathrm{B}^2)$$
$$(\mathrm{Y}\backslash_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z{:g}} \quad \mathrm{X}\backslash_{\diamond}\mathrm{Y{:f}} \Rightarrow (\mathrm{X}\backslash_{\diamond}\mathrm{W})/_{\diamond}\mathrm{Z}{:}\lambda z\lambda w.\, f\,((g\,z)\,w) \qquad (<\mathrm{B}^2)$$
$$\mathrm{X}/_{\times}\mathrm{Y}{:}f \quad \mathrm{Y}\backslash_{\times}\mathrm{Z}{:}g \Rightarrow \mathrm{X}\backslash_{\times}\mathrm{Z}{:}\lambda z.f\,(g\,z) \qquad (>\mathrm{B}_{\times})$$
$$\mathrm{Y}/_{\times}\mathrm{Z}{:}g \quad \mathrm{X}\backslash_{\times}\mathrm{Y}{:}f \Rightarrow \mathrm{X}/_{\times}\mathrm{Z}{:}\lambda z.f\,(g\,z) \qquad (<\mathrm{B}_{\times})$$
$$\mathrm{X{:a}} \Rightarrow \mathrm{T}/_{i}(\mathrm{T}\backslash_{i}\mathrm{X}){:}\lambda f.f\,\mathrm{a} \qquad (>\mathrm{T})$$
$$\mathrm{X{:a}} \Rightarrow \mathrm{T}\backslash_{i}(\mathrm{T}/_{i}\mathrm{X}){:}\lambda f.f\,\mathrm{a} \qquad (<\mathrm{T})$$

不同语言的语法区别仅在词法层面，按词法形成的词汇通过“结合”规则[‡]映射到语言的句子 (Steedman, 2019)，即确定句子里每个词的范畴，按规则一步步结合相邻范畴，就可以得到句子结构。由“结合”规则得到的范畴再与相邻范畴结合，有是否引入范畴转换的两种做法 (王庆江, 张琳, 2020)。词库里可记录每个词的典型和非典型范畴，但词选择哪个非典型范畴即范畴转换，只发生于范畴转换规则表示的上下文，这避免了范畴转换对语法表达力的过度增强。范畴转换其实是把句子的词范畴歧义转化为句法歧义，是词或短语句法功能转换的客观反映，仍属于词法层面。

3 范畴转换规则的汉语言学依据

能不能绕过词类，直接考虑词或短语的范畴转换，本质上是语法中能否去掉词类这一概念的问题，汉语言学里该问题的争论止于上世纪五十年代，之后就是争论如何划分词类 (吕叔湘, 1954)。划分词类的目的是为了进行句法分析，划分词类的标准必须考虑句法分析的需要 (胡明扬, 1995)。上世纪五十年代前，词类划分和给词定类都采用意义标准；五十至八十年代，词类划分渐趋句法标准，而给词定类仍坚持意义标准。九十年代及以后，给词定类趋于句法标准，即只根据词参与构建短语和充当句子成分的能力 (沈家煊, 2009)。

给词定类必须先于句法分析，否则词类对句法分析起不到任何作用 (胡明扬, 1995)。然而，“离句无品”，词性通过词所在的例句体现，先确定词在句中充当的句子成分再确定词性的反向逻辑长期存在。《现代汉语词典》第五版才对词做词类标注 (徐枢, 谭景春, 2006)，可见语言学界对词语是否有固有词性是多么纠结。“词有定类”指每个单词力求归于一类，“类有定职”指词类与句子成分一一对应。汉语里，“类有定职”则“词无定类”，词类失去存在意义，故只能是“词有定类”，让词类与句子成分的关系错综复杂 (朱德熙, 1985)。

“词有定类”与词兼类、词同形可能是相容的。兼类词在作不同词类时，词的意义虽然相关但已经不同，而同形词的意义更是毫无联系。词类有意义基础 (张斌, 2005)，“词有定类”可能蕴涵词在一种意义下只属于一个词类，这与范畴转换发生在词类内部是一致的，即发生范畴转换时词的意义并未改变。若该假设成立，兼类词或同形词在每个义项下的词类就是唯一的。

与印欧语词类分立不同，汉语词类句法功能可以重叠甚至包含。在汉语实词类包含模式 (沈家煊, 2009)中，凡动词皆名词，即动词有名词的所有功能，“这本书的出版”中“出版”是动词也是名词，从动词角度讲符合简约原则，即不增加不必要的步骤和名目，从名词角度讲符合中心扩展，中心扩展指以一个成分为中心加以扩展，扩展后结构的语法性质跟中心成分的语法性质一致。

基于词类充当不同句子成分的出现率，可以把词类功能限制为充当出现率较高的句子成分 (胡明扬, 1995)，也可以把出现率最高的句子成分作为词类的典型功能，把其它出现率的句子成分作为词类的非典型功能 (沈家煊, 1997; 沈家煊, 1999)。如果按词有定类、类对应典型功能建立“词:=范畴”词典，则词在短语结构里需要使用非典型功能时，就需要转用非典型功能对应的范畴。

汉语缺乏系统的形态标记，不仅导致词类多功能，也造成语法的词组本位特征 (张伯江, 2011)。汉语句子的构造原则跟词组的构造原则基本一致，可以把各类词组作为抽象的句法格式

[‡]这里指实现范畴结合的所有规则，而非特指Schönfinkel结合子对应的那些规则。

(c) 2022 中国中文信息学会计算语言学专业委员会

来描写它们的内部结构以及每一类词组作为一个整体在更大的词组里的分布状况，把各类词组的结构和功能描写清楚了，句子的结构也就描写清楚了 (朱德熙, 1985)。词组本位思想表现为结构包孕，即短语结构的基本类型虽然很有限，但每一种结构都可以包孕与它自身同类型或不同类型的结构 (朱德熙, 1982)，这种结构包孕也被称做结构套叠 (陆俭明, 1990)。结构套叠对应到CCG里，就是短语结构的范畴由其直接成分的范畴结合而来，这样得来的范畴如果不是短语结构作为整体在更大短语里应该采用的范畴，就转换范畴然后再范畴结合，依此递归下去。

4 词类和特殊词的典型范畴

首先为句法成分指派范畴，使基本句法结构能按范畴结合规则形成，然后由句法成分的范畴确定词类的典型范畴。给定句子范畴 s 和名词短语范畴 np，由后向应用（<）、前向应用（>）、前向组合（>B）、后向组合（<B）可得谓语范畴s\.np、述语范畴(s\.np)/.np、定语范畴np/.np、状语范畴$(s\backslash.np)/_{\diamond}(s\backslash.np)$和$(np/.np)/_{\star}(np/.np)$、补语范畴$np\backslash_{\star}np$、$(s\backslash.np)\backslash_{\times}(s\backslash.np)$和$(np/.np)\backslash_{\star}(np/.np)$，使主谓（Subject-Predicate，SP）、定中（Attribute-Headword，AHn）、状中（aDverbial-Headword，DHv或DHa）、中补（Headword-Complement，HnC、HvC或HaC）、述宾（Verb-Object，VO）等结构的范畴是其成分的范畴按范畴结合规则得到的结果，这里中心成分H的语法性质可以是名词（n）、动词（v）或形容词（a）。

实词类按短语结构需要选择范畴，例如主语位置上的形容词选用范畴np，定语位置上的名词选用范畴np/.np。出现率最高的范畴是词类的典型范畴，如名词、形容词的典型范畴分别是np和np/.np。与一级词类相比，二级词类的典型范畴更明显，如谓语动词、单宾语动词、双宾语动词的典型范畴分别是s\.np、(s\.np)/.np和((s\.np)/.np)/.np。用范畴中斜线类型精确反映词类的句法功能，如数词、量词的典型范畴分别是$np/_{\star}np$和$(np/_{\star}np)\backslash_{\star}(np/_{\star}np)$，因类型'$_{\star}$'只匹配前向应用（>）、后向应用（<）中的线性类型，故数词可单独做定语，也可与量词结合后再做定语，而量词除与数词结合没有其它句法功能。

虚词不单独充当句法成分，虚词、虚词附着的实词（短语）、附着形成的虚词短语三者之间存在范畴结合规则的约束。实词类有典型和非典型范畴，不妨让虚词附着的实词（短语）使用虚词典型搭配词类的范畴，使虚词的范畴按虚词短语的范畴确定。虚词数量少，句法功能差异大，需要对每个虚词专门考虑其句法范畴。例如'的'、'地'、'得'字短语的典型功能分别是做定、状、补语，让'的'附着的实词（短语）无论是否名词性都使用范畴np，'地'和'得'附着的实词（短语）无论是否形容词性都使用范畴np/.np，'的'、'地'、'得'的典型范畴就可分别明确为$(np/_{\star}np)\backslash_{\star}np$、$((s\backslash.np)/_{\diamond}(s\backslash.np))\backslash_{\star}(np/.np)$和$((s\backslash.np)/_{\times}(s\backslash.np))/_{\star}(np/.np)$。

个别词有特殊的句法功能。助词'所'接述语形成'所'字短语，表述转指称，如"所说"指说的话，'所'的范畴可令为$np/_{\star}((s\backslash.np)/.np)$。介词'把'接宾语，再结合述语，形成谓语，如"我把馍吃了"中的"把馍吃"，故'把'的范畴可令为$((s\backslash.np)/_{\star}((s\backslash.np)/.np))/_{\star}np$，其用法如图1，其中"Desig"表示词按词典"词:=范畴"取得指派的范畴，'X'匹配任意范畴，'α'匹配任意λ-项，语义式中的函数都写为λ-抽象。范畴结合规则具有范畴类型透明性，即结合前后都有范畴类型决定语义类型。按λ-应用的左结合优先，图1句子的语义式也可写为"了' (((把' 馍) '吃') 我')"。

```
--Desig ---------------------------Desig ----------Desig ----------Desig ----------Desig
np      ((s\.np)/⋆((s\.np)/.np))/⋆np    np              (s\.np)/.np     X\.X
:我,    :λxλfλy.把' x f y               :馍,            :λxλy.吃' x y   :λα.了' α
        ------------------------------------------->
        (s\.np)/⋆((s\.np)/.np)
        :λfλy.把' 馍' f y
        ------------------------------------------------------------->
        s\.np
        :λy.把' 馍' 吃' y
-------------------------------------------------------------------------<
s
:把' 馍' 吃' 我'
-------------------------------------------------------------------------------<
s
:了' (把' 馍' 吃' 我')
```

图 1: 介词'把'的范畴及其用例

(c) 2022 中国中文信息学会计算语言学专业委员会

介词‘被’是被动标记，被动句可分为无施事的短被动句和有施事的长被动句 (姚从军, 俎孟晨, 2022)。在缺少出现率依据情况下，不妨设介词‘被’大多用在长被动句中，典型功能是接引宾语提取，形成句子谓语，如“饭被我吃了”中‘被’字短语是“被我吃”而非“被我”。‘被’与介宾结构中介词的功能不同，宾语提取的范畴为s/.np，故‘被’的范畴是(s\.np)/$_\star$(s/.np)，其用法如图2。若‘被’用在短被动句中，如“饭被吃了”，就让及物动词‘吃’转用其非典型范畴s/.np。

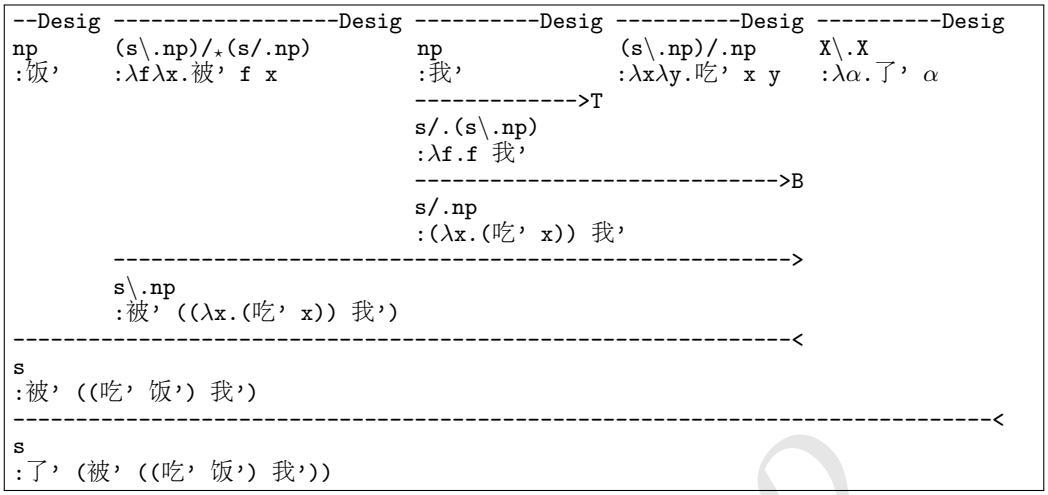


图 2: 介词‘被’的范畴及其用例

图2中，‘饭’的范畴与‘被’字短语的范畴结合，形成“饭被我吃”的范畴，而从伴随的语义结合看，语义项“吃’”是与语义项“饭’”结合，产生这一现象的根源是使用了组合规则。组合规则允许相邻的两个函数范畴先组合，预留的参数空位最终被相距较远的参数范畴填补 (满海霞, 2022)。这种参数占位可保持语义项二元结合的顺序不变，如在规则“>B”关联的语义层面，函数f和g组合为函数λx.f(g x)，没改变函数g先应用到参数x、函数f再应用到参数(g x)的顺序。

5 词类和短语结构的范畴转换规则

对于实词类或向心结构，转换标记“句法成分/词类”表示词类或以该词类为中心成分的短语结构使用句法成分对应的范畴。在构建树库过程中收集词类和短语结构的范畴转换规则，典型范畴为np的名词（短语）的范畴转换规则如表1，其中一些是时间、方位等二级名词特有的，示例下标是词或短语的范畴，‘⇒’下标是规则标记。动词、形容词、副词等实词类或以这些词类为中心的短语结构也有相应的范畴转换规则。

范畴转换规则	标记	适用短语示例	转换与结合连用（短语类型）
np np⇒np s\.np	P/n	今天$_{\text{np}}$星期一$_{\text{np}}$	$\Rightarrow_{\text{P/n}}$ np s\.np$\Rightarrow_{<}$ s (SP)
np np⇒(s\.np)/.np np	V/n	组织$_{\text{np}}$学生$_{\text{np}}$	$\Rightarrow_{\text{V/n}}$ (s\.np)/.np np$\Rightarrow_{>}$ s\.np (VO)
np np⇒np/.np np	A/n	门$_{\text{np}}$把手$_{\text{np}}$	$\Rightarrow_{\text{A/n}}$ np/.np np$\Rightarrow_{>}$ np (AH_n)
np np⇒np np\$_\star$.np	C_n/n	商城$_{\text{np}}$沃尔玛$_{\text{np}}$	$\Rightarrow_{\text{C}_\text{n}/\text{n}}$ np np\$_\star$np$\Rightarrow_{<}$ np (H_nC)
s\.np np⇒s\.np (s\.np)\$_\star$(s\.np)	C_v/n	走$_{\text{s\.np}}$一天$_{\text{np}}$	$\Rightarrow_{\text{C}_\text{v}/\text{n}}$s\.np (s\.np)\$_\star$(s\.np)$\Rightarrow_{<}$ s\.np (H_vC)
np s\.np⇒(s\.np)/$_\diamond$(s\.np) s\.np	D/n	寒假$_{\text{np}}$返校$_{\text{s\.np}}$	$\Rightarrow_{\text{D/n}}$(s\.np)/$_\diamond$(s\.np) s\.np$\Rightarrow_{>}$ s\.np (DH_v)

表 1: 名词（短语）的范畴转换规则

对于非向心结构，转换标记“句法成分/短语结构”表示短语结构使用句法成分对应的范畴，标记“词类/短语结构”表示短语结构使用词类的典型范畴，目前发现的非向心结构有主谓（s）、宾语提取（oe）和谓语提取（pe），宾语提取即“主述”短语，谓语提取即“主状”短语，如表2，其中删除线部分为短语结构的上下文，“U1P”指‘的’字短语，方位名词的典型范畴是np\$_\star$np，谓语提取的典型范畴是s/$_\diamond$(s\.np)。

(c) 2022 中国中文信息学会计算语言学专业委员会

范畴转换规则	标记	适用短语示例	转换与结合连用（短语类型）
s s\.np ⇒ np s\.np	S/s	桃花开$_{s}$是在春天$_{s\backslash.np}$	⇒$_{S/s}$ np s\.np ⇒$_{<}$ s (SP)
np s ⇒ np s\.np	P/s	解放军$_{np}$意志坚定$_{s}$	⇒$_{P/s}$ np s\.np ⇒$_{<}$ s (SP)
(s\.np)/.np s ⇒ (s\.np)/.np np	O/s	认为$_{(s\backslash.np)/.np}$你很努力$_{s}$	⇒$_{O/s}$ (s\.np)/.np np ⇒$_{>}$ s\.np (VO)
s np ⇒ np/.np np	A/s	他上班$_{s}$时间$_{np}$~~不长$_{s\backslash.np}$~~	⇒$_{A/s}$ np/.np np ⇒$_{>}$ np (AH_n)
s np\⋆np ⇒ np np\⋆np	H_n/s	课程安排$_{s}$上$_{np\backslash\star np}$	⇒$_{H_n/s}$ np np\⋆np ⇒$_{<}$ np (H_nC)
s (np/⋆np)\⋆np ⇒ np (np/⋆np)\⋆np	N/s	他上班$_{s}$的$_{(np/\star np)\backslash\star np}$	⇒$_{N/s}$ np (np/⋆np)\⋆np ⇒$_{<}$ np/⋆np (U1P)
(s\.np)/.np s/.np ⇒ (s\.np)/.np np	O/oe	让$_{(s\backslash.np)/.np}$你做$_{s/.np}$	⇒$_{O/oe}$ (s\.np)/.np np ⇒$_{>}$ s/.np (VO)
s/.np np\⋆np ⇒ np np\⋆np	H_n/oe	学生学习$_{s/.np}$上$_{np\backslash\star np}$	⇒$_{H_n/oe}$ np np\⋆np ⇒$_{<}$ np (H_nC)
s/.np (np/⋆np)\⋆np ⇒ np (np/⋆np)\⋆np	N/oe	他读$_{s/.np}$的$_{(np/\star np)\backslash\star np}$	⇒$_{N/oe}$ np (np/⋆np)\⋆np ⇒$_{<}$ np/⋆np (U1P)
s/$_{\diamond}$(s\.np) (np/⋆np)\⋆np ⇒ np (np/⋆np)\⋆np	N/pe	学生在班里$_{s/_{\diamond}(s/.np)}$ 的$_{(np/\star np)\backslash\star np}$~~表现~~	⇒$_{N/pe}$ np (np/⋆np)\⋆np ⇒$_{<}$ np/⋆np (U1P)

表 2: 非向心结构的范畴转换规则

虚词短语使用非典型句法功能，可以通过范畴转换，如‘的’字短语的典型范畴是np/⋆np，通过“S/a”可转用非典型范畴np，也可通过虚词转用非典型范畴得到虚词短语的非典型范畴，如助词‘得’典型功能是接形容词做述语补语，如“干得好”，非典型功能是接修饰形容词的副词（简称形容词副词）做形容词补语，如“灵得很”，存在范畴转换规则“U3d/u3”，“u3”指助词‘得’，“U3d”指右接形容词副词时‘得’的范畴((np/.np)\⋆(np/.np))/⋆((np/.np)/⋆(np/.np))。附带地，副词典型功能是修饰动词，“D_a/d”表示副词使用形容词副词的范畴，“U3d/u3”总是和副词的范畴转换规则“D_a/d”连用。

短语的两个直接成分同时使用范畴转换规则，这样的情况如表3。

范畴转换规则	标记	适用短语示例
s np/.np ⇒ np s\.np	S/s-P/a	学生住宿$_{s}$方便$_{np/.np}$
s (s\.np)/.np ⇒ np/.np np	A/s-H_n/v	社会主义现代化$_{s}$建设$_{(s\backslash.np)/.np}$
s np ⇒ np np\⋆np	H_n/s-C_n/n	人懒$_{s}$这个现象$_{np}$
np s\.np ⇒ (s\.np)/.np np	V/n-O/v	组织$_{np}$教学$_{s\backslash.np}$
np s ⇒ np/.np np	A/n-H_n/s	这$_{np}$成绩好$_{s}$~~是必然的~~
np np/.np ⇒ np/.np np	A/n-H_n/a	职务$_{np}$方便$_{np/.np}$
np (s\.np)/.np ⇒ np/.np np	A/n-H_n/v	思政$_{np}$教育$_{(s\backslash.np)/.np}$
np s\.np ⇒ np/.np np	D_a/n-A/v	这样$_{np}$好$_{(s\backslash.np)/.np}$~~的规定~~
s\.np np/.np ⇒ np s\.np	S/v-P/a	出门$_{s\backslash.np}$方便$_{np/.np}$
s\.np (s\.np)/.np ⇒ np/.np np	A/v-H_n/v	毕业$_{s\backslash.np}$设计$_{(s\backslash.np)/.np}$
(s\.np)/.np (s\.np)/$_{\diamond}$(s\.np) ⇒ np/.np np	A/v-H_n/d	~~少数有~~录取$_{(s\backslash.np)/.np}$可能$_{(s\backslash.np)/\diamond(s\backslash.np)}$~~的学生~~
(s\.np)/.np s\.np ⇒ s\.np (s\.np)\×(s\.np)	P/vt-C_v/v	分配$_{(s\backslash.np)/.np}$到基层工作$_{s\backslash.np}$
np/⋆np (s\.np)/.np ⇒ np s\.np	S/a-P/vt	留级的$_{np/\star np}$也算上$_{(s\backslash.np)/.np}$
((s\.np)\×(s\.np))/⋆(np/.np) (s\.np)/$_{\diamond}$(s\.np) ⇒ ((np/.np)\⋆(np/.np))/⋆((np/.np)/⋆(np/.np)) (np/.np)/⋆(np/.np)	U3d/u3-D_a/d	~~灵得~~$_{((s\backslash.np)\times(s\backslash.np))/\star(np/.np)}$ 很$_{(s\backslash.np)/\diamond(s\backslash.np)}$

表 3: 短语两个成分都使用范畴转换规则

(c) 2022 中国中文信息学会计算语言学专业委员会

范畴转换规则也具有范畴类型透明性，即范畴转换后范畴类型仍然决定语义类型，如图3。

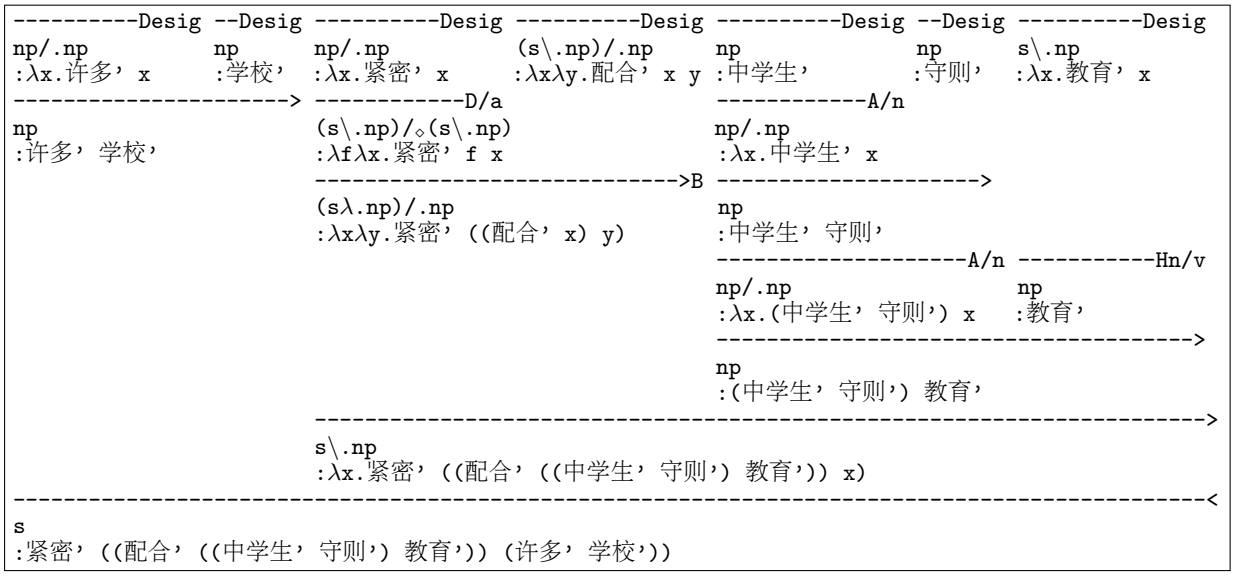


图 3: 范畴转换中的范畴类型透明性

6 语法树库的统计分析

句法分析时，若保留句法歧义，即每次传递时允许使用所有范畴转换，传递后不删除造成歧义的短语，短语集合规模大致会随传递次数指数增长，故句法分析中必须及时消解句法歧义。在构建树库过程中已记录遇到的每一个句法歧义片段，将来可通过挖掘消歧模式，并引入常识和世界知识，探索完全靠机器消歧获得人脑预期分析树的可能性。

句法歧义的主要成因是句法结构层次、句法结构关系的不同 (何洪峰, 2016)，映射到CCG分析里，就是一个范畴既可与左边的范畴结合，又可与右边的范畴结合，毗连的两个范畴有多条结合途径可用，结合途径指范畴结合经过的范畴转换与范畴结合规则。范畴转换使句法歧义严重，是汉语缺乏形态标记、句法成分之间大多可以套叠 (陆俭明, 1990)的直接后果。为限制和及时消解句法歧义，每次计算范畴结合传递时，由用户按需选择范畴转换，然后由机器完成范畴结合，再由用户消解句法歧义，使最后形成的传递闭包只含一棵分析树，这样的句法分析其实是一个计算结合-消歧传递闭包的过程。

由语法的二分性，分析树只有度为0的结点（即词）和度为2的结点（即短语）。把词看作跨度0的短语，使词和短语有统一的数据抽象。短语有外在的位置、句法类型和语义表示，也有内在的结合途径、语法关系，定义为三元组:

$$((Start, Span), (SyntaxType, Tag, Seman, PhraStru, Act), SecStart)$$

其中$Start$是短语在句中的起始位置，$Span$是短语跨度，即所含词数减一，$SyntaxType$是短语的句法类型，Tag是短语的句法类型结合途径的标记，词的Tag为“Desig”，取指定（Designate）之意；$Seman$是短语的语义式，由两个成分的语义式通过函数运算得到，词的语义用词加撇表示；$PhraStru$是短语的语法关系，可以是实词性语法单位间语法关系，如主谓（SP）、动宾（VO），也可以是与虚词有关的语法单位间关系，如宾语抽取（OE）、介宾（PO）、‘的’字结构（U1P），词的$PhraStru$为“DE”，取指定（Designated）之意。已参与形成短语时，Act为False，否则为True。$SecStart$是短语中第二组成成分在句中的起始位置。分析树存储为纯文本，从中可统计句子长度、短语个数、每个短语的外在和内部属性，并可生成树状结构图。

在平衡语料库（http://corpus.zhonghuayuwen.org/CnCindex.aspx）按“学生”检索词类标注语料，对前200句进行小句分割 (邢福义, 1995; 李艳翠, 2013)，共得到727个小句。对每个小句，将词类标注替换为范畴标注，然后进行人机交互的句法分析，得到华水树库1.0（https://github.com/wangqingjiang-ncwu/my-ccg/tree/master/doc）。

(c) 2022 中国中文信息学会计算语言学专业委员会

华水树库1.0共使用67个不同的范畴转换规则，占转换规则总数的91%，使用各规则的次数如图4，次数为1时不显示。动词（v）、名词（n）、形容词（a）、副词（d）使用范畴转换规则的次数比例分别是33%、32%、14%、7%，若算上连词（c）、量词（q）、非向心结构短语（s、oe和pe），以及短语两个成分都使用范畴转换，则使用次数比例合计99.67%，说明使用范畴转换的几乎都是实词或短语。平均每棵分析树使用范畴转换规则2.46个，形成7.05个短语，故35%的短语形成用到范畴转换。

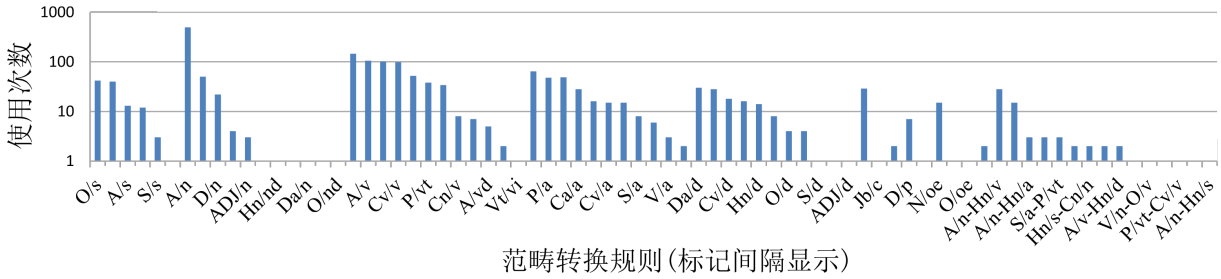
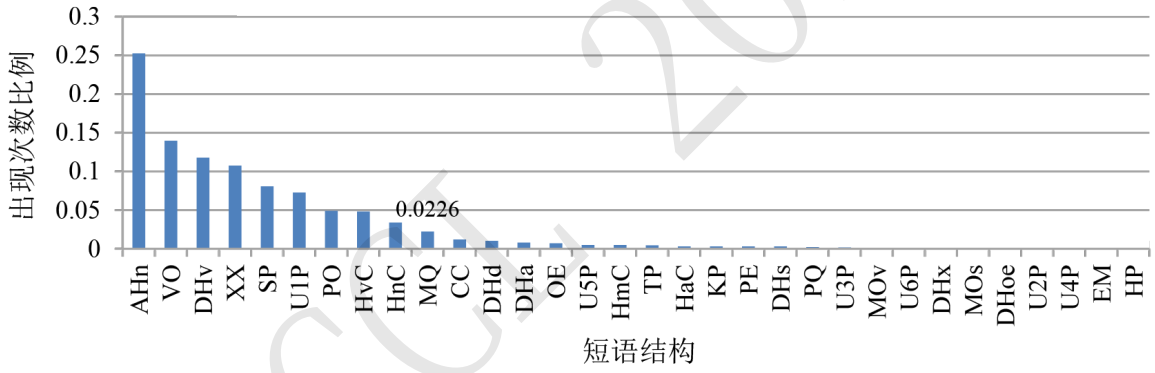


图 4: 华水树库1.0中范畴转换规则的使用次数

华水树库1.0中分析树共包含27种、5127个短语，覆盖全部短语结构类型的84%，短语类型出现次数比例如图5。因结合途径中范畴结合规则相同，连谓并入中补（HvC），兼语归入动宾（VO），复指并入中补（HnC），方位并入中补（HnC），能愿并入状中（DHv），这时定中、动宾、状中、并列、主谓，中补等基本结构可合称广义基本结构，其次数比例合计81%。'的'字短语（U1P）、介宾短语（PO）是广义基本结构外使用比例最高的两种。次数比例最高的前9种短语类型的次数比例总计90%，其它短语类型的次数比例均低于3%。



AH_n: 定中$_{名}$　CC: 并列小句　DH_a: 状中$_{形}$　DH_d: 状中$_{副}$　DH_{oe}: 状中$_{宾语提取}$　DH_s: 状中$_{句}$
DH_v: 状中$_{动}$　DH_x: 状中$_{趋向动词}$　EM: 语气短语　H_aC: 中$_{形}$补　H_mC: 中$_{数}$补　H_nC: 中$_{名}$补
HP: 前接短语，如“老三”、“第一”　H_vC: 中$_{动}$补　KP: 后接短语，如“工作者”、“中式”
MO_s: 移动宾语到主语前，即‘被’字短语　MO_v: 移动宾语到述语前，即‘把’字短语
MQ: 数量　OE: 宾语提取　PE: 谓语提取　PO: 介宾　PQ: 代量，如“这筐”
SP: 主谓短语　TP: 语调短语　U1P: ‘的’字短语　U2P: ‘地’字短语　U3P: ‘得’字短语
U4P: ‘着’、‘了’、‘过’字短语（已归入 H_vC）　U5P: 以‘等’、‘似的’结尾的比况短语
U6P: 所’字短语　VO: 动宾　XX: 并列短语

图 5: 华水树库1.0中短语类型的出现次数比例

7　结论

与印欧语相比，汉语缺乏屈折，使一种词类或一种短语结构能充当多种句法成分，对应到组合范畴语法中，就是有多个范畴，而这种多功能出现在一个句子里时，必须完成与印欧语屈折对应的范畴转换，范畴语法分析才能继续下去。为此，把词类的范畴按出现率分为典型和非典型，把短语结构作为整体充当句法成分时需要的范畴按是否由范畴结合规则得到分为典型和

(c) 2022 中国中文信息学会计算语言学专业委员会

非典型，把使用非典型看作源自典型的转换，可建立一套适应短语结构需要的范畴转换规则，而且这样的范畴转换具有范畴类型透明性。

在构建树库过程中，逐渐明确词类和特殊词的典型范畴，形成词类和短语结构的范畴转换规则体系。统计发现，短语直接成分使用非典型范畴的概率是35%，这是迄今已知的现代汉语句法语料中句法功能屈折的首次测量。使用范畴转换的短语直接成分中99.67%是实词或短语结构，说明虚词范畴力争明确，让虚词附着的实词或短语结构通过范畴转换与虚词搭配，这样的范畴转换规则体系是成功的。

范畴转换的上下文是短语结构，这种转换对组合范畴语法的生成能力带来了什么影响，需要做理论分析；分析树刻画的是小句结构，若句子不能正确分割为小句，则一些范畴的转换上下文就会不准确，需要对小句理论特别是小句识别进一步研究。在做好小句识别基础上还要分析更多的句子，纠正词类和特殊词的典型范畴，完善词类和短语结构的范畴转换规则，调整短语类型体系，逐渐让这三方面收敛，使这个带有范畴转换的CCG更简洁、稳定。另外，句子结构与句子语义式的关系，以及基于语义式的理解模型也是一个研究方向。

致谢

感谢匿名审稿人对论文的评审，评审意见在论文补充研究和进一步完善上发挥重要作用。

参考文献

Bob Carpenter. 1991. The generative power of Categorial Grammars and Head-Driven Phrase Structure Grammars with lexical rules. Computational Linguistics, 17(3):301-313.

陈鹏. 2016. 组合范畴语法(CCG)的计算语言学价值. 重庆理工大学学报(社会科学), 30(8):5-11.

冯志伟. 2021. 神经网络, 深度学习与自然语言处理. 上海师范大学学报(哲学社会科学版), (2):110-122.

何洪峰. 2016. 句法结构歧义成因的思考. 语言研究, 23(4):26-31.

胡明扬. 1995. 现代汉语词类问题考察. 中国语文, 1995(5):381-389.

Jayant K and M Mitchell T. 2014. Joint syntactic and semantic parsing with combinatory categorial grammar. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Stroudsburg, Pennsylvania, pages:1188-1198.

李艳翠, 冯文贺, 周国栋等. 2013. 基于逗号的汉语子句识别研究. 北京大学学报(自然科学版), 49(1):7-14.

陆俭明. 1990. 汉语句法成分特有的套叠现象. 中国语文, (2):81-90.

吕叔湘. 1954. 关于汉语词类的一些原则性问题. 中国语文, (9):6-14.

Mark Steedman, Jason Baldridge. 2011. Combinatory Categorial Grammar. Non-Transformational Syntax. Blackwell:181-224.

Mark Steedman. 2019. Combinatory Categorial Grammar. Current Approaches to Syntax – A Comparative Handbook. De Gruyter Mouton:389-420.

满海霞. 2022. 组合范畴语法: 通向人工智能语义理解的一种逻辑经验主义路径. 哲学动态, (1):119-125.

沈家煊. 1997. 形容词句法功能的标记模式. 中国语文, (4):242-250.

沈家煊. 1999. 不对称和标记论. 江西教育出版社, 南昌.

沈家煊. 2009. 我看汉语的词类. 语言科学, 8(1):1-12.

王庆江, 张琳. 2020. 支持中文句法结构套叠的组合范畴语法. 中文信息学报, 34(1):17+22.

邢福义. 1995. 小句中枢说. 中国语文, (6):420-428.

徐枢, 谭景春. 2006. 关于现代汉语词典（第5版）词类标注的说明. 中国语文, (1):74-86.

姚从军, 俎孟晨. 2022. 语言、逻辑与计算互动视角下汉语直接被动句的MMCCG处理. 湖南科技大学学报（社会科学版）, 25(1):42-50.

(c) 2022 中国中文信息学会计算语言学专业委员会

张斌. 2005. 现代汉语语法十讲. 复旦大学出版社, 上海.

张伯江. 2011. 关于现代汉语词典（第5版）词类标注的说明. 汉语学习, (2):3-12.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

朱德熙. 1985. 语法答问. 商务印书馆, 北京.

CCL 2022

双重否定结构自动识别研究

王昱
北京大学中文系/ 北京100871
wangyustu@pku.edu.cn

袁毓林
澳门大学人文学院中文系/ 澳门
北京大学中文系/中国语言学研究中心
/教育部计算语言学重点实验室/ 北京100871
yulinyuan@um.edu.mo/yuanyl@pku.edu.cn

摘要

双重否定结构是一种“通过两次否定表示肯定意义”的特殊结构，其存在会对自然语言处理中的语义判断与情感分类产生重要影响。本文以“¬ ¬ P==>P”为标准，对现代汉语中所有的“否定词+否定词”结构进行了遍历研究，将双重否定结构按照格式分为了3大类，25小类，常用双重否定结构或构式132个。结合动词的叙实性、否定焦点、语义否定与语用否定等相关理论，本文归纳了双重否定结构的三大成立条件，并据此设计实现了基于规则的双重否定结构自动识别程序。程序实验的精确率为98.85%，召回率为98.90%，F1值为98.85%。同时，程序还从96281句语料中获得了8640句精确率约为99%的含有双重否定结构的句子，为后续基于统计的深度学习模型提供了语料支持的可能。

关键词： 双重否定 ； 自动识别程序 ； 语义分析

The Research on Automatic Recognition of the Double Negation Structure

Wang Yu
Department of Chinese Language and Literature, Peking University / Beijing 100871
wangyustu@pku.edu.cn

Yuan Yulin
Department of Chinese Language and Literature, University of Macau / Macau
Department of Chinese Language and Literature, Center for Chinese Linguistics, Key Laboratory of Computational Linguistics, Peking University / Beijing 100871
yulinyuan@um.edu.mo/yuanyl@pku.edu.cn

Abstract

The double negation structure is a special structure of ”expressing positive meaning through two negations”, in which the two negations have an important impact on the semantic analysis and emotional classification in natural language processing. Taking ”¬ ¬ P==>P ” as the standard, this paper makes an ergodic study on the ”negation word + negation word” structures in modern Chinese; According to the formal features and semantic expressions, the double negation structure is divided into three categories, 25 sub-categories and 132 common structures or constructions. Then, based on the theories of factuality, negation focus, descriptive truth-functional negation and non-truth-functional negation, this paper investigates the double negation structure, summarizes the three conditions for the establishment of the double negation structure, and designs the program of automatically recognizing the double negation structure. The accuracy rate of the recognition of the double negation structure is 98.80%, the recall rate is 98.90%, and the F1 value is 98.85%. Meanwhile, the program also obtains 8640 sentences which 99% contains double negation structure from 96281 sentences, which provides corpus for the subsequent statistical based deep learning model.

Keywords: Double negation , Automatic recognition program , Semantic Analysis

第二十一届中国计算语言学大会论文集，第24页-第34页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

在否定用法中，有一种特殊的用法——双重否定。丁声树先生在《现代汉语语法讲话》中将其概括为：“一句话先后用两个否定词，如‘不能不去’，‘没有人不去’，‘非去不可’之类，都是双重否定的句法。双重否定意思上是肯定的，不过跟单纯肯定不全一样”。例如，“我不得不喜欢他”指“我得喜欢他”，“我不一定不同意这个观点”指“我可能同意这个观点”[0]。虽然句子使用的是否定形式，但其表达的却是肯定语义，且句子的语气还与单纯肯定有一定的区别（例如，“不得不喜欢”的语气比“喜欢”强，而“不一定不同意”的语气比“同意”弱，等等）。可见，双重否定结构的存在会对句子的语义真值与语气强度产生重要影响。

对于自然语言处理来说，双重否定是处理语料时必须考虑的内容。例如，下图展示了目前常用的自然语言处理应用对双重否定结构的处理情况。左图为微软小冰聊天机器人。如图所示，小冰无法有效识别以否定形式的“不是不”表示肯定的句子的语义与情感。即使是对于“我不是不喜欢你，而是超级喜欢你”这种明显表示积极情感的句子，小冰也仍然会将其误判为消极。右图是百度情感分析应用。我们的输入是“我不认为这件事情不值得称赞”，这句话应该是中性或者偏积极。但是由于无法识别双重否定结构，程序将其误判为了百分之百消极。可见，目前自然语言处理应用尚无法有效识别双重否定结构，且双重否定结构的存在会对程序的语义识别与情感分析产生重要影响。**综上所述，无论是从汉语本体研究的理论意义考虑，还是从自然语言处理的实用价值出发，我们都有必要对双重否定结构进行更全面更深入的研究与探索。**

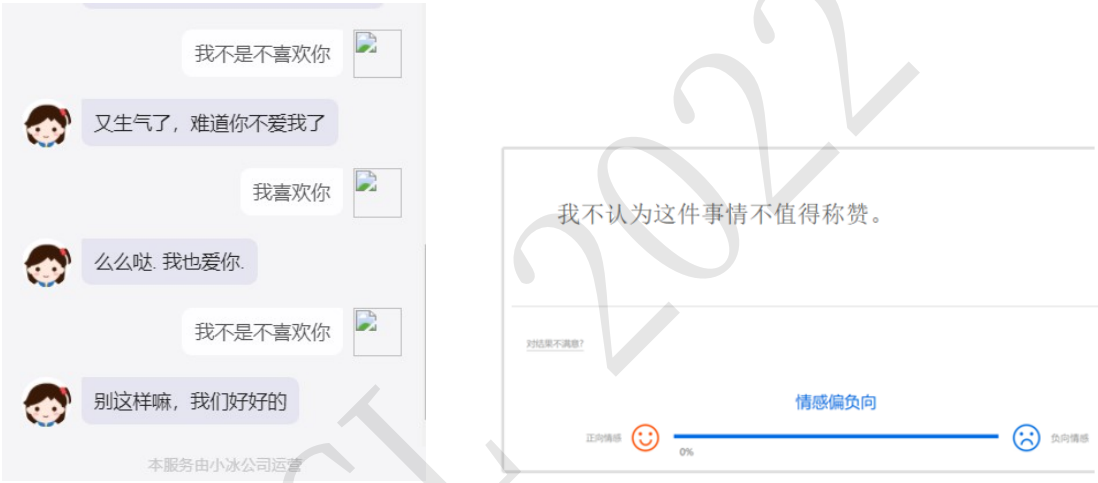


Figure 1: 双重否定结构处理情况示例图

1.1 文献综述

目前国内关于双重否定结构的理论研究主要集中在双重否定结构的定义、范围、格式和语义等领域。关于双重否定结构的定义与范围，学界存在不少的争论，学者们争议较多的问题有：含有否定词的反诘句是否为双重否定，紧缩复句是否为双重否定，双重否定是否包含复句，等等。关于双重否定结构的格式，已有研究都是从分类与举例的角度进行讨论，尚未有一篇论文从形式的角度对双重否定格式进行具体详细的遍历分析。目前对双重否定格式概括最为全面的是芜崧（1987）所划分的八大类型，25个格式，然而，仍有许多语料中出现的双重否定结构未被芜崧收入其中，如“不可能不”“不应该不”“不是...不...”，等等。关于双重否定结构的语义，叶文曦（2013）、方绪军（2017）、何爱晶（2019）等引入了Ladusaw（1997）的形式语义学，Horn（1985）的元语否定等理论，对一些典型的结构进行了分析，得出了具有解释

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：本课题得到国家科技创新2030“新一代人工智能”重大项目《以自然语言为核心的语义理解理论、模型与方法》（编号：2020AAA0106701）和国家社科基金重大项目《基于“互联网+”的国际汉语教学资源与智慧教育平台研究》（项目编号：18ZDA295）的资助，谨此致谢。

[0]丁声树等.现代汉语语法讲话[M].北京：商务印书馆，2004.200-202.

力的成果。然而，由于双重否定的范围、格式还未确定，目前学者只集中分析了几个典型的结构，覆盖面十分有限，缺乏系统性地总结与梳理。

目前关于双重否定结构的应用研究主要集中在情感分析领域，具体根据研究方法可以分为以下两种。

一、通过搜集典型的双重否定结构（如图2所示），构建双重否定词典，以服务相关的情感分析，如王勇等（2014），吴杰胜、陆奎（2019），等等。**这种方法的准确率很高，但是覆盖面不足。**

双重否定词典	绝非不、并非不、不是不、不能不、不会不、不可不、不要不、不得不、没有不、无不、不无

Figure 2: 王勇等（2014）双重否定词典

二、在否定词的基础上，通过统计修饰每一个情绪词的否定词个数来判断双重否定，并以系数的形式将双重否定的语气功能加入到情感分析的结果当中，如封洋（2016），等等。**这种方法的涵盖范围广，但是错误率很高，因为任何含有两次否定的结构都会被判断为表示肯定的双重否定结构。**

1.2 本文选题及目标

综上所述，目前学界对汉语双重否定结构的研究成果颇丰但仍然存在一些不足之处，例如，双重否定结构的格式与范围尚不完整；双重否定语料资源匮乏；系统化、全面化的双重否定结构自动识别尚未实现，等等。鉴于此，本研究将“双重否定结构”作为研究对象，试图通过遍历分析与语料考察相结合的方法，对双重否定结构进行以下探索：

1. 梳理双重否定结构格式，使其能够全面覆盖CCL语料库；
2. 总结双重否定结构成立条件，并据此提出相应的计算机识别策略；
3. 建立高F1值的双重否定结构自动识别程序；
4. 进一步验证语言学知识在双重否定结构自动识别过程中的贡献，通过程序测试上述成立条件在双重否定结构识别过程中的作用；
5. 搜集双重否定语料资源，为基于统计的双重否定识别深度学习模型提供支持。

2 双重否定结构的定义标准与考察范围

鉴于语义真值识别和情感极值判断是计算机对否定结构进行语义识别时所面临的主要问题，本文借鉴形式语义学，为双重否定拟定了一个工作定义：**只要两次否定与肯定在语义真值上相同，即“¬ ¬ P==>P”，即属于双重否定。**

目前我们的考察范围为所有“否定词（+...）+否定词”中双重否定表肯定的结构。暂不考虑下列情况：

1. 否定词为隐性否定词（即本身语义内含有否定意思的动词，如“讨厌”“拒绝”等。）
2. 否定类型为语用否定的结构（比如“我不是不喜欢你，而是恨你”中的“不是不喜欢”。）
3. “反问句+否定词”结构（比如“难道...不...”等，具体参见刘彬、袁毓林（2017）。）

我们结合吕叔湘（1956/1944）、王力（1985/1943）、朱德熙（1982）等前贤研究，梳理出了十三个否定词，前十个为否定副词，后三个为否定动词，具体如下：

“非、不、别、甭、未、莫、勿、没、没有、休、无、没、没有”

结合语料，我们对“否定词+否定词”组合中符合要求的双重否定结构的格式进行了遍历梳理，整理出了25种“否定词+否定词”可表肯定的结构，具体如表1所示（其中三角形表示该组合不出现/极少出现于实际语料中）。

下面，我们将对上述25种“否定词+否定词”格式展开具体分析，梳理每一种格式中双重否定结构的成立条件，并设计与之相应的计算机识别策略，在此基础上总结所有现代汉语中常用的双重否定结构。

(c) 2022 中国中文信息学会计算语言学专业委员会

	不	没	没有	无	非	莫	别
不	不…不	不…没…	不…没有…	不无	▲	▲	▲
没	没不 没…不…	没…没…	没…没有…	▲	▲	▲	▲
没有	没有不 没…不…	没有…没	没有…没有…	▲	▲	▲	▲
无	无不 无…不…	▲	▲	▲	▲	▲	▲
非	非不 非…不…	非没 非…没…	非没有 非…没有…	非无	▲	非…莫…	▲
莫	莫不	▲	▲	▲	▲	▲	▲
别	别不	别没	▲	▲	▲	▲	▲

Table 1: 双重否定结构格式调查情况表

3 双重否定结构的成立条件与识别策略

通过初步考察，我们发现，双重否定结构“不V1...不V2”的成立条件最为复杂，需要同时满足以下三个条件：

①“不V1”与“不V2”构成述宾关系；

②“V1”动词有限制（为非叙实动词）；

③“不V1”的否定焦点在“不V2”上。

我们发现，除了“不是...不...”需要区别语义否定与语用否定外，其他双重否定结构的成立条件都已囊括在了上述三个条件中，只是部分细节存在差异。因此，我们先对“不V1...不V2”与“不是...不...”的成立条件与识别策略进行详细分析，再在此基础上，对其他双重否定结构进行讨论。

3.1 “不V1...不V2”双重否定结构的成立条件与识别策略

3.1.1 第一个条件：“不V1”与“不V2”构成述宾关系

“不V1...不V2”的结构类型有并列、主谓、紧缩、述宾等。在各类结构类型中，只有述宾结构的“不+V1+（...）+不+VP”存在表示双重否定的可能。具体讨论如下：

并列结构的“不V1...不V2”，指“不哭不闹”、“不高不低”这一类表达。袁毓林（1999）指出，并列结构“通常不能通过直接在这种谓词性并列结构的前面加上‘不、没有’等否定词来构成否定式,而是要在这种并列结构的各个直接成分之前分别加上‘不、没有’等否定词。”因此，“不V1...不V2”只是“V1V2”并列结构的单重否定结构，不属于双重否定结构。如“不哭不闹”不等于“哭闹”。

主谓结构的“不V1...不V2”语料数量很少，指“不隐藏不代表泄露”、“不买票不是我的决定”这一类表达。在该类结构中，“不V1”只是一个命题陈述，是交流中的旧信息。“不V1”的“不”与“不V2”的“不”，并没有语义上的关联，并不构成“双重”否定的结构。“不V1...不V2”只是“不V1V2”结构的否定，不属于双重否定结构。如“不隐藏不代表泄露”不等于“隐藏代表泄露”，“不买票不是我的决定”不等于“买票是我的决定”。

紧缩结构的“不V1...不V2”指“不给钱不办事”、“不买票就不让进”这一类表达。紧缩结构虽然在语义上有条件性，但在句法上仍是并列关系，前后并不构成从属结构。关于紧缩条件类的结构是否为双重否定未有定论。本文主要从形式语义学的角度对其进行讨论。

以“不给钱不办事”为例。“给钱办事”语义为“如果给钱，那么办事”。P命题可以分解为q1“给钱”，q2“办事”，逻辑式为蕴含式q1→q2,它的等值式为¬ q2→¬ q1。而“不给钱不办事”语义为“如果不给钱，那么不办事”，逻辑式应为蕴含式¬ q1→¬ q2。从下列真值表本文可以看出，q1→q2与¬ q1→¬ q2的语义真值不一致，不符合“¬ ¬ P==>P”的标准，因此从形式语

(c) 2022 中国中文信息学会计算语言学专业委员会

义学来看，紧缩语义结构不是双重否定结构。

q1	q2	¬q1	¬q2	q1→q2	¬q2→¬q1	¬q1→¬q2
T	T	F	F	**T**	T	**T**
T	F	F	T	**F**	F	**T**
F	T	T	F	**T**	T	**F**
F	F	T	T	**T**	T	**T**

Table 2: 紧缩结构语义真值表

最后述宾结构的“不V1...不V2”指“不觉得不好”、“不知道你不来”等V1为述语，“不V2”为宾语的结构。该类结构中的“不...不...”可以理解为肯定。例如，“不认为他明天不会来”可以理解为“认为他明天会来”,“不觉得这件事不妥”可以理解为“觉得这件事妥”，等等。由此，我们可以得出**“不V1...不V2”构成双重否定结构的第一个条件：“不V1”与“不V2”构成述宾关系。**

3.1.2 第二个条件：V1为非叙实动词

述宾结构的“不V1...不V2”中只有一部分成员，其“不V1”对“不V2”有管辖作用，属于双重否定结构，其他成员仍只表示单纯的否定。试看下例：

1a.我不知道他不来　　1b.*我知道他来
2a.我不幻想他不来　　2b.*我幻想他来
3a.我不认为他不来　　3b.我认为他来

通过例句，可以发现，当V1为“认为”时，“不V1...不V2”可以理解为“V1...V2”，而当V1为“知道”、“幻想”时却不能。同样是动词，“知道”、“幻想”、“认为”却存在着区别。本文认为，“不V1”对“不VP”是否有管辖作用与V1的语义有关，具体来说与V1的叙实性有关。

李新良（2015）将叙实性定义为“叙实性是动词的一种语义功能，即动词预设其宾语小句真值的能力。具体来说，肯定式和否定式都预设其宾语小句为真的动词是叙实动词.....肯定式和否定式都不预设其宾语小句为真，也不预设其宾语小句为假的动词是非叙实动词.....肯定式和否定式都预设其宾语小句为假的动词是反叙实动词”。对于叙实动词和反叙实动词来说，由于其预设固定，无论主句有无否定，宾语小句的真值都不变，因此主句无法影响宾语小句的真值，不构成“¬ ¬ P==>P”。而对于非叙实动词（如：认为）来说，由于非叙实动词对宾语小句并没有预设，在述宾结构中，主句中的V1可以对宾语的真值造成影响，存在“¬ ¬ P==>P”的可能。因此，我们可以得出**“不V1...不V2”构成双重否定结构的第二个条件：V1为非叙实动词。**

3.1.3 第三个条件：“不V1”的否定焦点包含V2

除上述两个条件外，结构中否定焦点的情况也会对“不V1...不V2”是否为双重否定造成影响。试看下例：

4a.我不认为他不来。　　4b.我认为他来。
5a.我不认为他故意不来。　　5b. *我认为他故意来。
6a.我不相信他不喜欢我。　　6b.我相信他喜欢我。
7a.我不相信他不喜欢我到了看见我就恶心的地步。
7b.*我相信他喜欢我到了看见我就恶心的地步。

在例句中，4a、6a可以转换为4b、6b，而5a、7a却不能转换为5b、7b。本文认为这主要与否定的焦点有关。袁毓林(2000)指出“有的成分表达的是句子的预设意义，属于旧信息，事实上它们的意义在否定的情况下仍然得以保持;有的成分表达的是句子的焦点意义，属于新信息，它们是真正被否定的。”当“不V1”的否定焦点不落在“不V2”上时，“不V1”对“V2”不

(c) 2022 中国中文信息学会计算语言学专业委员会

造成否定，不能构成“¬ ¬ P”结构，因此无法满足“¬ ¬ P==P”的条件，不属于双重否定。如“我不认为他故意不来”中的“不V1”否定的是“故意”而不是“不来”，其中“不来”是预设成分，属于旧信息。无论是“认为他故意不来”还是“不认为他故意不来”，语义都是“他不来”。“V1认为”的否定无法影响到“不来”的真值，无法构成“否定+否定”的语义结构。因此，我们可以得出**“不V1...不V2”构成双重否定结构的第三个条件：“不V1”的否定焦点落在“不V2”上。**

综上所述，“不V1...不V2”需要同时满足①“不V1”与“不V2”构成述宾关系、②V1为非叙实动词、③“不V1”的否定焦点落在“不VP”，三大成立条件，才可构成双重否定结构，表示肯定语义。

3.2 “不是...不V2”双重否定结构的成立条件与识别策略

如章节开头所述，“不是...不V2”若要表示双重否定，除需满足上述“不V1...不V2”的成立条件外，还需保证“不是”为描述性真值否定（descriptive truth-functional negation，又称语义否定），而不是元语否定（non-truth-functional negation，又称语用否定）。“所谓元语否定，就是用元语言对对象语言所描述的非真值语义的否定，...是一种非真值意义否定；与之相对应的是真值否定，否定的是句子的真值条件（truth condition）”[1] 这种否定常常是引述性否定，是对之前对话中已出现的内容的否定。例句如下：

8a. “可现在杀他不容易啊。”有人说。不是不容易，是根本不可能。
8b. * “可现在杀他不容易啊。”有人说。是容易，是根本不可能。

例中否定形式的8a“不是不容易，是根本不可能”不能理解为相应的肯定形式8b“是容易，是根本不可能”。这是因为例8中的“不是”是语用否定，并不对语义真值产生影响，不构成“否定+否定”的语义结构，无法通过两次否定表示肯定。因此，为了识别“不是...不V2”中的双重否定结构，首先需要区分“不是”是语义否定还是语用否定。为此，我们对大量语料进行了考察。我们发现，“不是”的否定情况具体有以下三种:

1. 当“不是”的上文中没有出现“不是”否定的观点时，“不是”只能是语义否定，而不能是语用否定。例如:

9a. 甲：你明天来不来参加生日聚会?
9b. 乙：不是我不乐意，而是我这边实在没时间。
语义否定✓ → 我是乐意，（但）是我这边实在没时间。
语用否定× → *“不乐意”这个表述不恰当，是我这边实在没时间。

2. 当“不是”的上文中出现了“不是”否定的观点，且“不是”所在句的后文与被否定观点的真值一致，则“不是”只能是语用否定，而不能是语义否定。例如:

10a. 甲：你不乐意明天来参加生日聚会?
10b. 乙：我不是不乐意，而是超级反感。
语义否定× → *我是乐意，是超级反感。
语用否定✓ → “不乐意”这个表述不恰当，（应该）是超级反感。

3. 当“不是”的上文中出现了“不是”否定的观点，且“不是”所在句的后文与被否定观点的真值不一致，则“不是”既可看作语义否定，又可看作语用否定。对于这种情况，从计算机处理的角度出发，我们可以统一将其处理为“不是”表示语义否定。

11a. 甲：你不乐意明天来参加生日聚会?
11b. 乙：不是我不乐意，而是我这边实在没时间。
语义否定✓ → 我是乐意，（但）是我这边实在没时间。
语用否定✓ → “不乐意”这个表述不对，（应该）是我这边实在没时间。

综上所述，只有当“不是”的上文中出现了“不是”否定的观点，且“不是”所在句的后文与被否定观点的真值一致时，“不是”才会是语用否定，而其余情况，皆可被计算机视为语义否定。

根据上述条件，我们可以通过计算情感真值的方法，来约束“不是”为语义否定。在文本里，绝大多数表示语用否定的“不是”都只出现在“不+是+不+x，（而）是+y”的结构中。因此对于“不是”是语义否定还是语用否定，我们可以拟定识别策略如下：

[1]何爱晶.反叙的非真值义否定和真值义肯定[J].外语研究,2019,36(04):第25页.

(c) 2022 中国中文信息学会计算语言学专业委员会

提取“不+是+不+x”中的x,并匹配y与“不+x”的情感真值。若“不+是+不+x，（而）是+y”中“y”的情感真值与“不+x”不一致，则句中的“不是不”属于双重否定结构；若一致，则不属于双重否定结构。

3.3 双重否定结构的成立条件

以上即为“不V1...不V2”与“不是...不...”双重否定成立条件的基本情况。本文所有双重否定结构的成立条件皆在上述成立条件的基础上进行一定的调整。我们按照“组合类”、“构式类”、“粘合类”三个大类类别[2]，对第二节中所提出的25小类双重否定结构的成立条件进行了详细梳理，在此过程中，本文还结合成立条件，从25小类中梳理出了双重否定常用结构或构式132个。与芜崧（1987）的结果相比，我们的分类更系统，类型更丰富，覆盖更全面，所涉及的常用双重否定结构格式约为芜崧（1987）的5倍。具体结果如图3所示。

大类	双重否定结构格式小类	双重否定结构常用格式	具体限制情况	语料具体示例
组合类	1. 不+助动词+不+VP	不得不+VP 不能不+VP 不可不+VP …共计 23 个	助动词范围：能、能够、可能、会、可以、应该、应、应当、一定、要、得、愿意、愿、肯、可、想、要、敢、该、当、准、许、容	不得不四处筹钱 不得不惴惴地沉默不语 不能不从我国的国情出发 不可不仔细用功
	2. 不+是（……）+不+VP	不+是（……）+不+VP 共计 1 个	1. “不是”与“不”必须是述宾关系，“不是”后的“不VP”为否定焦点； 2.如果出现在“不+是+x，（而）是+y”结构中，若y的情感真值与“不x”的情感真值不一致，则“不是不”为双重否定结构，否则，不为双重否定结构。	不是不读书 不是不重要 不是不愿意丰富自己的生活 不是审计师不小心
	3. 不+V1（……）+不+VP	不认为（……）+不+VP 不觉得（……）+不+VP 不希望（……）+不+VP …共计 11 个	1. 不+V1+不+VP 为述宾结构 2. V1 属于非叙实动词，且 V1 后的“不VP”为否定焦点。V1 具体为：认为、说、感到、觉得、允许、同意、相信、愿意、希望、考虑、打算。	不觉得不尊重 不认为不适当 不觉得不应该捡 不认为自己的点子不好 不希望有个不愉快的开始
	4. 不+助动词+V1（……）+不+VP	不会感到（……）不+VP 不会觉得（……）不+VP 不应该说（……）不+VP …共计 15 个	1. 助动词范围 2. 不+V1+不+VP 为述宾结构 3. V1 属于非叙实动词，且 V1 后的“不VP”为否定焦点。	不能说不好 不能说不可以 不会觉得不愉快不
	5. 不+助动词+没/没有+NP/VP	不可没/没有+NP/VP 不能没/没有+NP/VP 不可能没/没有+NP/VP …共计 23 个	同 1	不得没规矩 不可没有梦想 不能没带耳机和钱
	6. 不+是（……）+没/没有+NP/VP	不+是（……）+没+NP/VP 不+是（……）+没有+NP/VP 共计 2 个	同 2	不是他没有这个能耐 不是男人没有责任 不是人民还没有战胜这些灾害
	7. 不+V1（……）+没/没有+ NP/VP	不认为（……）+没/没有+NP/VP 不觉得（……）+没/没有+NP/VP 不希望（……）+没/没有+NP/VP …共计 11 个	同 3	不认为院长没有注意到手术中发生的过失 我不认为自己就没有生存力 我不相信没有主持公道的地方 你不办，我还不相信没有人办了。
	8. 不+助动词+V1（……）+没/没有+NP/VP	不能认为…+没/没有+NP/VP 不会觉得…+没/没有+NP/VP 不能说…+没/没有+NP/VP …共计 15 个	同 4	不会觉得这世界没有鱼的容身之地 不能认为本案已经没有价值。 不能说他没有责任。
	9. 没/没有+NP/VP+不+VP/AP	没有人不+VP/AP 没有什么不+VP/AP 没有一个不+VP/AP 共计 3 个	/	没人不喜欢 没人不辛苦 没有人不会走 没有一个学校不采用应试教育

大类	双重否定结构格式小类	双重否定结构常用格式	具体限制情况	语料具体示例
组合类	10. 没/没有+NP+没/没有+VP	没/没有+NP+没/没有+VP 共计 4 个	/	没有人没完成任务 没有白羊座没有见过他 没有一个人没有写过好作品
	11. 无+NP+不+VP/AP	无一不、无人不、无一个不 共计 3 个	/	无人不开宫 无人不受奖 无一人不会唱
	12. 非+…+没/没有/不+VP/AP/NP	并非/绝非+…+没有/没/不+VP/AP/NP 共计 6 个	/	并非他没有尽职 并非对生活没有要求 并非分部不精当
构式类	13. 不…不行/不好	不…不行、不…不好 共计 2 个	“不…不行/不好”为构式短语，在一个分句中出现。	不降价不行、不孝顺不行、不卖不行、不借不好
	14. 非…不可/不行	非…不可、非…不行 共计 2 个	“非…不可”为构式短语，在一个分句中出现。	非推新不可、非彻底改革不可、非着喜马不行、非阿永不行
	15. 非…莫属	非…莫属，共计 1 个	“非…莫属”为构式短语，在一个分句中出现。	非马云莫属、非他莫属、非邓丽君莫属、非机器人莫属
粘合类	16. 没不…	没不… 共计 1 个	“没”、“不”皆为独立的否定词	没不严肃、没不成的、没不让你追求
	17. 没有不…	没有不… 共计 1 个	“没有”、“不”皆为独立的否定词	没有不对、没有不同意、没有不失败的
	18. 不无…	不无… 共计 1 个	“不”、“无”皆为独立的否定词	不无道理、不无裨益、不无参考意义、不无关系
	19. 无不…	无不… 共计 1 个	“无”、“不”皆为独立的否定词	无不佩服、无不为之动容、无不能写、无不悲伤欲绝
	20. 非不…	非不… 共计 1 个	“非”为“并非”、“绝非”等否定词，“不”为独立的否定词	并非不懂政治、并非不知道、并非不合理、绝非不让人接近
	21. 非没…	非没… 共计 1 个	“非”为“并非”、“绝非”等否定词，“没”为独立的否定词	并非没意见、并非没听到、并非没感觉、并非没钱
	22. 非没有…	非没有… 共计 1 个	“非”为“并非”、“绝非”等否定词，“没有”为独立的否定词	并非没有依据、并非没有想法、绝非没有道理、绝非没有可能
	23. 非无…	非无… 共计 1 个	“非”为“并非”、“绝非”等否定词，“无”为独立的否定词	并非无次序、并非无计划、绝非无好处、绝非无意义
	24. 别不…	别不… 共计 1 个	“别”、“不”皆为独立的否定词	别不信、别不吭声、别不满意、别不玩
	25. 别没…	别没… 共计 1 个	“别”、“没”皆为独立的否定词	别没良心、别没钱
合计	25 小类	132 个常用结构/构式		

Figure 3: 双重否定结构格式与成立条件示例图

3.4 双重否定结构的成立条件

针对上述成立条件，我们设计了相应的计算机识别策略，具体整理如下：

成立条件	识别策略
① 否定词 1 所在结构与否定词 2 所在结构构成述宾关系	依存句法分析，检测句法结构
② 否定词 1 所在结构对动词有限制	根据情况建立词表（如，助动词词表）+字符串匹配
③ 否定词 1 的否定焦点在否定词 2 所在结构上	依存句法分析，检测句法结构+ 字符串匹配，排除否定焦点在“否定词 2 结构”修饰语上的情况
④ “不是”为语义否定	构建情感词典，计算情感真值+字符串匹配

Table 3: 双重否定结构成立条件与识别策略对应表

[2]组合类：否定词与否定词之间不相连，扩充了其他成分；构式类：特定的否定词与否定词之间组成构式，结构复杂，形式固定；粘合类：否定词与否定词之间没有其他成分，二者紧连；

(c) 2022 中国中文信息学会计算语言学专业委员会

4 双重否定自动识别程序的建立

4.1 词库的建立

为了使计算机能够识别助动词、非叙实动词、情感真值，本文对助动词、非叙实动词与情感词进行了梳理，在常用的基础词表中补充了助动词词表、非叙实动词词表与情感词表。助动词方面，本文以郑贵友（1989）整理的助动词范围为基本，结合鲁晓琨（2004）等前人的研究以及现代汉语的使用情况，选取了23个助动词，构成常用助动词词表。具体如下：

能、能够、可能、会、可以、应该、应、应当、一定、要、得、愿意、愿、肯、可、想、要、敢、该、当、准、许、容

非叙实动词方面，结合袁毓林（2014）、李新良（2015）等人对非叙实动词的研究，本文认为非叙实动词多为心理动词。因此本文对心理动词进行了考察。若一个心理动词的宾语的真值无法确定，则该心理动词为非叙实动词。以此为标准，本文对心理动词进行了考察，筛选出了11个非叙实动词[3]。整理如下：

认为、说、感到、觉得、允许、同意、相信、愿意、希望、考虑、打算

情感词方面，我们结合知网hownet情感词典、台湾大学NTUSD简体中文情感词典与清华大学李军中文褒贬义词典，设计了情感词表，共收纳正面词语10323个，负面词语9411个。

根据上述双重否定结构的识别策略，我们设计编写了双重否定结构自动识别程序。通过该程序，对含有两个否定词的语料文件，进行自动识别实验，输出其中存在的双重否定句以及相应的双重否定结构。程序输出结果示例如下（在程序中，我们还做了双重否定结构的语气识别，由于篇幅原因，本文暂不对其进行讨论。）：

Figure 4: 程序输出结果示例图

5 双重否定自动识别实验

5.1 实验语料来源

我们从CCL语料库中，按照各类结构的情况，进行了同等分布提取（即根据各结构在CCL语料库中的语料数量比例进行提取），收集了100000条初始语料。然后，通过程序，对上述语料进行筛选，排除了所有未被成功分句，长度超过150字的句子，最终获得测试语料96281句。

5.2 初始实验

我们使用双重否定结构自动识别程序对96281句语料进行识别，获得了8640句计算机认为含有双重否定结构的句子。由于人力有限，我们无法对8640个句子都进行人工检校。因此，为了计算精确率，我们从程序识别出的句子中随机抽取了1000句进行检测，经过人工检校，以上1000个句子中判断正确的句子为992句，由此我们可以计算，程序识别的精确率约为99.20%；在CCL“否定词+否定词“语料中，双重否定句的比例约为8.9%。

为了进一步验证，我们从96281句语料中随机抽取了1000语料进行检测。通过人工校验，上述1000句测试语料中存在92句含有双重否定结构的句子。然后，我们将上述1000句测试语料输入到双重否定自动识别程序中，程序从中识别出了90句含有双重否定结构的句子。根据计

[3]非叙实动词的界限并不是完全清晰的。非叙实动词与叙实动词、反叙实动词之间还存在一定的渗透性。由于其情况较为复杂，且对本文研究的影响较小，故暂不讨论。

(c) 2022 中国中文信息学会计算语言学专业委员会

算，程序的召回率约为97.83%。根据F1值公式，该程序的F1值约为98.51%。CCL“否定词+否定词”语料中双重否定句的比例仍在9%左右。

5.3 正式实验

初始实验表明，上述96281句语料中双重否定结构与非双重否定结构的比例差距很大（约为9：91），因此为了更好地计算程序的精确率与召回率，我们人工构建了2000句双重否定语料，其中1000句为含有双重否定结构的正例，1000句为不含双重否定结构的负例。我们将语料输入双重否定结构程序中，具体结果如下表：

	1000 正例	1000 负例	总精确率	总召回率	F1 值
	识别出的双重否定句数				
本文程序	989	12	98.80%	98.90%	98.85%
王勇等（2014）	777	10	98.72%	77.70%	86.96%
封洋（2016）	1000	1000	50.00%	100%	66.67%

Table 4: 实验结果数据表

本文程序识别的精确率约为98.80%，召回率约为98.90%，F1值约为98.85%，实验结果较王勇等（2014）、封洋（2016）的结果有较明显提升。需要说明的是，本文的实验为封闭测试，检验数据的方式是抽样，且文章对于双重否定结构的判断均来自于作者本人，因此结果会存在一定的偏差。后续我们还会投入更多的时间与人力，来获取更为准确的数据。

5.3.1 实验结果分析

无论是精确率还是召回率，实验的准确率都是在百分之九十多，未达到百分之百。通过分析，我们发现程序识别与召回错误的主要原因与句子的分词与句法分析错误有关。由于分词与句法分析等基础自然语言处理工具的问题，程序对一些句子的句法判断错误，导致一些原本应被判为并列关系、因果关系的成分，被误判为述宾关系，从而使整个双重否定结构的判断错误。示例如下：

34743:...不大紧；有的急于求成把将来要办的事情，拿到今天来办，由于条件不允许迟迟开展不了。

（程序识别结果：双重否定结构为“不+非叙实动词+...不”）
（实际情况：“条件不允许”与“迟迟开展不来”是因果并列关系）

6893:...要的网络公司均未能达到阿尔诺的预期。为此，去年6月底，阿尔诺不得不刹车。他说，他要考虑”战略调整”。

（程序分词：阿尔诺不得不/刹车　　实际分词：阿尔诺/不得不/刹车）

当我们输入人工修改后的分词与句法分析结果后，程序的错误得到纠正，精确率与召回率皆可达到100%。

6 双重否定结构成立条件的测试实验

为了测试上述三个条件在双重否定结构识别过程中的作用，进一步验证语言学知识在双重否定结构自动识别过程中的贡献，我们将人工构建的2000句双重否定语料作为输入，测试在取消某一条件后，双重否定结构识别程序的识别情况与召回情况。具体结果如下表：

可以看到，“构成述宾结构”与“动词为非叙实动词”对整个双重否定结构的识别造成较大影响。尤其在召回率方面，相较应我们提供的标准正确数据，没有“构成述宾结构”条件约束的程序较原始结果将额外召回约37倍的错误结构（460：12），而没有“动词为非叙实动词”条件约束的程序也将额外召回1.4倍的错误结构（29：12）。

相较来说，“否定焦点”条件会对双重否定结构识别的影响最为轻微。我们认为，这是因为人们在实际语言交流中很少会使用非常复杂的句子（例如：“我不相信他不喜欢我到了看见我就恶心的地步”）。当我们扩大检测数据，用程序对96281句原始语料进行测试时，缺少否定焦点

(c) 2022 中国中文信息学会计算语言学专业委员会

	1000 正例	1000 负例
	识别出的双重否定句数	
原始数据	989	12
取消“述宾结构”条件限制	1078	460
取消“非叙实动词”条件限制	1031	29
取消“否定角度”条件限制	989	13

Table 5: 双重否定结构成立条件测试实验结果

的程序将会比标准程序额外召回140句双重否定结构，进一步说明否定焦点会对双重否定结构识别程序造成影响，只是由于语料中复杂的句子很少，因此影响较轻微。这种情况也符合我们日常表达的经济性原则。

7 结语

本文以“¬ ¬ P==>P”为标准，借助计算机程序与CCL语料库，对现代汉语中所有的“否定词+否定词”结构进行了遍历研究，实现了以下目标：

1. 将双重否定结构按照格式分为了3大类，25小类，常用双重否定结构或构式132个，进一步地揭露了双重否定结构的全貌；
2. 总结得出了双重否定结构成立的三大条件；并对其进行了实验测试分析，进一步验证了语言学知识在双重否定结构中的作用；
3. 补充了助动词表、非叙实动词表、情感词表等基础词表，编写实现了双重否定结构自动识别程序。程序的识别的精确率约为98.80%，召回率约为98.90%，F1值约为98.85%；
4. 获得了8640句精确率约为99%的标明双重否定结构的句子，为后续建立双重否定语料库提供了支持。具体文件烦请参见脚注链接[4]。

本文还有许多不足未尽之处，例如：本文的识别程序是在规则的基础上建立的，而人为编写的规则未免有不尽之处；本文对双重否定结构的判断皆基于作者个人的语感，未免有疏漏之处；对于一些已经固化的双重否定结构，如“非...不...”、“无非”、“莫非”等，本文的处理还较为粗糙，有待进一步分析与探索，等等。

未来我们拟在本文的基础上，继续展开与深入，具体计划有：

1. 建立双重否定语料库，在程序的辅助下构建数万句级的双重否定语料库，工程量应能控制在数十小时；
2. 探究双重否定结构自动识别的深度学习模型。
3. 对双重否定结构的语用方面进行进一步研究与探索。
4. 对双重否定结构中的构式，例如“非...不可”、“非...莫属”等等，进行进一步的研究与探索。

参考文献

Horn,L. 1989. *A Nature History of Negation.* University of Chicago Press, Chicago, US, 311-312.

Ladusaw, W. A. 1997. *Negation and polarity items. In S. Lappin (ed.),The Handbook of Contemporary Semantic Theory.* Blackwell Publishing Ltd, Oxford, UK, 321-341.

丁声树等. 2004. 现代汉语语法讲话. 商务印书馆, 北京, 200-202.

方绪军. 2017. *“不是不X”、“不是没(有)X”*和 *“没(有)不X”*. 语言科学, 16(05):511-521.

封洋. 2016. 中文微博情绪分析. 上海交通大学, DOI:10.27307/d.cnki.gsjtu.2016.000526.

符达维. 1986. 对双重否定的几点探讨. 福建论坛*(*文史哲版*)*,(6):78-81.

[4]https://download.csdn.net/download/qq_43342081/86399912

(c) 2022 中国中文信息学会计算语言学专业委员会

何爱晶. 2019. 反叙的非真值义否定和真值义肯定. 外语研究, 36(04):24-29.

郎桂青. 1989. 双重否定句表示肯定的条件. 语文研究,（1）:28.

李新良, 王明华. 2015. 汉语动词的叙实性研究的应用前景. 对外汉语研究, (02):120-129.

鲁晓琨. 2004. 现代汉语基本助动词语义研究. 中国社会科学出版社, 北京.

吕叔湘. 1956. 中国文法要略. 商务印书馆, 北京.

王力. 1943. 中国现代语法. 商务印书馆, 北京.

王勇, 吕学强, 姬连春. 2014. 基于极性词典的中文微博客情感分类. 计算机应用与软件,(01):40-43+132.

芜崧. 1987. 双重否定句的种类与功能. 荆州师专学报（哲社版）,(3)：52-57.

叶文曦. 2013. 否定和双重否定的多维度研究. 语言学研究,(2):20-31.

袁毓林. 1999. 并列结构的否定表达. 语言文字应用,(03):42-46.

袁毓林. 2000. 论否定句的焦点、预设和辖域歧义. 中国语文,(02):99-108+189.

袁毓林. 2014. 隐性否定动词的叙实性和极项允准功能. 语言科学,13(6):575-586.

郑贵友. 1989. 汉语“助动词”的研究刍议. 汉语学习,(06):23-27.

CCL 2022

单项形容词定语分布考察及“的”字隐现研究*

宋锐[1], 王治敏[2]
[1] 沈阳师范大学文学院，沈阳 110034
[2] 北京语言大学汉语国际教育研究院，北京 100083
songrui1990@126.com

摘要

本文以2019-2021年《人民日报》文章中单项形容词定语77845个词例为研究对象，从实用性的角度考察了粘合式与组合式定语词例的分布特征、音节组配模式及“的”字的隐现倾向性。通过研究我们发现，粘合式定语的词例数量明显少于组合式定语词例数量，但使用频数却高出组合式定语的4-5倍。两种定语结构中，形容词和名词重复使用的比例很高，但其共现组合的比例偏少，同时，真实文本中“的”字的隐现具有“两极分化”的特征，绝大部分词例在使用过程中带“的”或不带“的”都具有很强的倾向性，“的”字出现具有区分词义和突显信息的作用，“的”字隐藏能促使语义更加凝练，进一步固化句式结构，使得某些句式形成了特指或隐喻的表达方式。本文为形容词定语结构的词汇语义研究提供依据和参考。

关键词： 形容词定语；分布特征；“的”字隐现；使用倾向

Study on Distribution of Single Item Adjective Attributives and Appearance and Disappearance of “de”

Rui Song[1], Zhimin Wang[2]
[1]School of Literature Shenyang Normal University，Shenyang 110034
[2]Research Institute of International Chinese Language Education,
Beijing Language and Culture University, Beijing 100083
songrui1990@126.com

Abstract

This paper studies 77,845 single-item adjective attributives from articles in "People's Daily" from 2019 to 2021. Distribution characteristics, syllable combination patterns of adhesive attributives and combined attributives, and appearance and disappearance tendency of the word "de" are studied from practical perspective. We find the number of adhesive attributives is significantly less than that of combined attributives, but the frequency of use is 4-5 times higher. In the two attributive structures, adjectives and nouns have a high proportion of repeated usage, but the proportion of their co-occurrence is relatively small. Besides, there is a strong tendency to use "de" or not in the process of usage. "De" has the function of distinguishing meaning and highlighting information. Disappearance of "de" can make the meaning more concise and further solidify the sentence structure, forming a specific or metaphorical expression. This paper provides a basis and reference for the lexical semantic research of adjective attributive structure.

* 基金项目：国家社科基金重大项目（18ZDA295）；国家语委科研项目（ZDI135-139）；中央高校基本科研业务费（19PT03）
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

Keywords: Adjective attributives , Distribution characteristics , Appearance and disappearance of "de" , Usage tendency

1 引言

单项定语是定语成分仅为一个修饰语的偏正结构，朱德熙(1982)把体词性偏正结构分为粘合式和组合式两大类，本文所研究的单项定语结构仅为形容词作定语成分的情况，即粘合式定语结构指形容词直接（不带“的”字）作定语的形式，组合式定语结构指定语成分为形容词（带“的”字）的偏正结构。二者形式相近，但表达意义却有不同的情况，例如：

（1）他发现自己脖子后面长了一块硬的骨头。

（2）西吉全县将上下总动员、发起总攻势，坚决啃下硬骨头，打赢脱贫攻坚战。

例（1）（2）中的“硬的骨头”和“硬骨头”分别为组合式和粘合式单项定语结构，虽然仅一个“的”字之差，但二者所表达的意义却明显不同。又如：

（3）炎炎夏日中，劳动者们仍然保持乐观的心态，坚守在一线。

（4）残疾人运动员的乐观心态让人难忘。

例（3）（4）中的“乐观的心态”和“乐观心态”，虽然也有“的”字的差别，但表达的意义基本相同。

那么，单项定语结构中，“的”字的隐现对词语意义表达有什么制约和影响？“的”字的隐现与哪些因素和条件相关？在大规模语料中，粘合式与组合式单项定语结构的分布情况如何，有多少词例是能够互相替换使用的，又有哪些词例分别倾向于出现在粘合式或组合式定语结构中，以上问题是本文主要分析和讨论的内容。

前人学者们有很多关于定语类型、定语语序及特点的研究(朱德熙, 1956; 刘月华, 1984; 袁毓林, 1999; 傅远碧, 2001; 崔应贤, 2002; 李先银, 2016)，也有关于定语结构中“的”字的作用和隐现的研究(吕叔湘, 1966; 锐, 2000; 蕾, 2004; 郑远汉, 2004; 王光全、柳英绿, 2006; 王远杰, 2006; 徐阳春, 2011; 雷友芳, 2012; 裴泓镔, 2020)，前人的研究成果主要侧重于理论层面和定语结构本身，缺乏语言现象的佐证和量化的数据分析，因此，本文结合大规模新闻文本语料，从实用性的角度，针对单项形容词定语结构[1]的词例现象，探究其分布特征、组合规律和使用倾向性，为汉语词汇研究和教学提供参考借鉴。

2 粘合式与组合式单项定语的分布特点考察

本文以2019-2021年的《人民日报》文章内容作为研究语料，共计一亿两千万余字。基于Python语言环境自编程序，利用哈工大语言技术服务平台（LTP4.0）进行分词和词性标注，提取出所有粘合式与组合式定语结构的词例，并进行人工校对和筛选处理，为了确保提取词例数据的真实性、准确性和客观性，我们对词例的筛选主要以《现代汉语语法信息词典》[2]作为词性的参照和判定标准。

在处理过程中，我们发现了一些不符合本文研究范围的词例，例如：

（5）李克强说，近期我国局地极端强降雨，造成重大人员伤亡和财产损失，令人痛心。

（6）新冠肺炎疫情是第二次世界大战结束以来最严重的全球公共卫生突发事件。

例（5）中的“重大人员伤亡”和例（6）中的“严重的全球公共卫生突发事件”，中心语成分为名词词组或NP结构，由于本文所限定的定语结构中心语均为光杆名词的情况，因此将此类词例去除。

另外，由于《现代汉语语法信息词典》的年代比较久远，我们在统计和筛选语料的过程中，发现有一些形容词和名词没有收录进来，如形容词“精准”“惊艳”“高清”“严苛”“酷炫”等，名词“变局”“抓手”“经济体”“志愿者”“互联网”等，经过与《现代汉语词典》（第七版）和《形容词分类词典》的对照，共增加未收录的形容词87个，未收录的名词120个，将其所组成的定语结构划归到本文的研究范围中，例如：

粘合式“an”结构：大变局、重要抓手、高清屏幕、高端智库……等等。

组合式“a的n”结构：凝练的词语、逼仄的走廊、清丽的歌声、舒缓的旋律……等等。

[1]本文的研究对象界定为：定语成分仅为形容词，且中心语成分仅为光杆名词的单项定语结构（限于篇幅，中心语为名词短语或NP结构的情况有待后续研究）。

[2]《现代汉语语法信息词典》是1998年由清华大学出版社出版适用于计算机对自然语言进行分类和处理的词典。

(c) 2022 中国中文信息学会计算语言学专业委员会

依据上述筛选原则和判定标准，经过3-4轮的人工判断和筛选后，我们共得到粘合式单项形容词定语结构词例34805例，组合式单项形容词定语结构词例43040例，词例数量和使用频数的整体分布情况如图1所示。

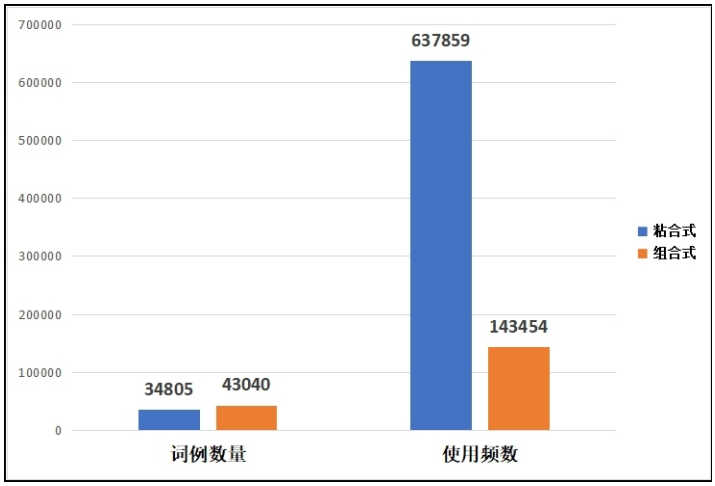


图 1: 粘合式与组合式词例数据对比

我们发现，粘合式“an”结构的词例数量占组合式“a的n”结构词例数量的4/5左右，但粘合式“an”结构的词例使用频数却是组合式“a的n”结构词例使用频数的4-5倍之多，差距非常明显，为什么会出现这样的现象呢?

我们认为，粘合式定语结构的形式简洁、语义凝练，定语成分和中心语成分的连接紧密度更高，对能够进入定语结构内的形容词和名词的搭配关系要求更高，词例所承受的内部限制程度更大，但由于语言的经济性原则，人们更倾向于使用简短的话语和文字进行交流和表达，所以粘合式词例的使用频率远远高于组合式定语结构，如“新时代”“高质量”等词例的反复使用频率非常高。相比而言，组合式定语结构由于加入了“的”字，弱化了定语成分和中心语之间的连接强度，使得组合式定语结构的形式相对松散，各成分之间即形容词和名词的语义搭配关系比较宽泛，因此组合式定语结构的词例数量较多，词例的使用范围较广，但由于组合式定语结构内部各项成分之间的组配模式比较复杂，所以组合式词例的整体使用数量远远低于粘合式定语结构的使用数量。

具体来看，粘合式结构与组合式结构的高频词例（排序前10位）及出现概率[3]如表1所示。

粘合式词例	出现概率	组合式词例	出现概率
新时代	0.338724388	新的历史	0.027323528
高质量	0.253503757	新的时代	0.008003083
重要讲话	0.115049136	新的征程	0.00752778
高级工程师	0.083614876	重要的作用	0.006898324
高水平	0.082946882	重要的意义	0.005485259
贫困人口	0.07471257	新的起点	0.004920033
新征程	0.074558417	美丽的世界	0.003918042
特别行政区	0.073903269	美丽的社会主义	0.003905196
贫困地区	0.071462522	坚实的基础	0.003673967
重要作用	0.059310168	新的活力	0.003661121

表 1: 粘合式与组合式单项定语排序前10位的词例

可见，从实用性的角度来看，粘合式与组合式各项序位的高频词例及出现概率各不相同，如排序第一位的分别为“新时代”和“新的历史”。又如，粘合式词例中排序仅第

[3]本文词例出现概率的计算方式为：个别词例的使用频数与所有词例频数之和的比值。

(c) 2022 中国中文信息学会计算语言学专业委员会

十位的“重要作用”在组合式词例“重要的作用”排序上升至第四位，二者的出现概率分别为“0.059310168”和“0.006898324”，这说明，虽然仅一个“的”字之差，但在实际语料中，“重要（的）作用”更倾向于粘合式的用法，即两种定语结构中，词例的使用倾向性不同。

从定语结构内部各项成分的使用情况来看，粘合式定语结构中，定语成分形容词共有1351例，使用频数为637859次，组合式定语结构中，定语成分形容词共有1821例，使用频数为143454次。在不考虑名词的情况下，粘合式与组合式两种定语结构中的形容词（交集）有1231例，占粘合式形容词数量的91.12%，占组合式形容词数量的67.60%，如图2所示。

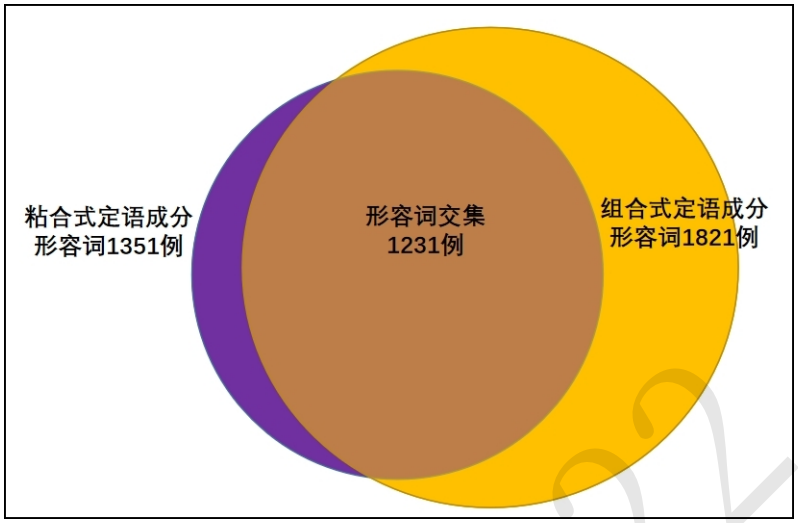


图 2: 粘合式与组合式定语成分形容词的分布

形容词交集的词例比例很大，说明两种结构中，大部分的形容词都能够通用，例如：

粘合式：新时代、新征程、新技术、新能源、新模式、新篇章……（交集为“新”）

组合式：新的历史、新的时代、新的起点、新的贡献、新的机遇、新的形势……

此处的交集为形容词相同，名词相同或不同均有的情况，如“新时代”和“新的时代”这样形容词和名词都相同，只是“的”字隐现的情况将在下文详述。

可以看出，粘合式定语结构中，绝大部分形容词（90%以上）都能进入组合式定语结构中，而组合式结构中的形容词仅有六成左右能进入到粘合式结构中，说明粘合式定语结构对形容词的限制程度更高。形容词交集中的高频词例（排序前10位）及出现概率如表2所示。

排序	定语成分形容词（交集）词例	出现在粘合式结构的概率	出现在组合式结构的概率
1	新	0.11706448	0.015223092
2	重要	0.075625262	0.005568831
3	高	0.0421662	0.002561073
4	大	0.035619528	0.008026233
5	重大	0.036665203	0.00037373
6	贫困	0.026895751	0.00074618
7	好	0.021760805	0.005104228
8	高级	0.020077741	0.000017919
9	安全	0.016716732	0.002338371
10	基本	0.017412996	0.000430045

表 2: 粘合式与组合式定语成分形容词（交集）排序前10位的词例

从出现概率的比较来看，交集的高频形容词（前10位）出现在粘合式结构中的概率更大。

中心语成分名词的分布情况如下：粘合式结构中名词的词例数量有6809例，使用频数为637859次；组合式结构中名词的词例数量为7334例，使用频数为143454次。在不考虑形容词的情况下，能共同进入到两种定语结构中的名词（交集）有4648例，占粘合式名词数量的68.26%，占组合式名词数量的63.38%，如图3所示。

(c) 2022 中国中文信息学会计算语言学专业委员会

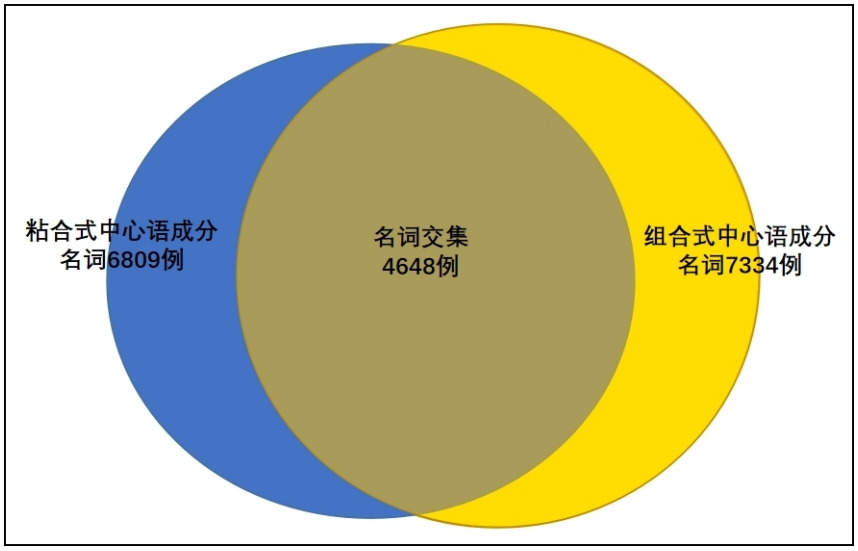


图 3: 粘合式与组合式中心语成分名词的分布

相比而言，名词交集的比例比较均衡，即两种定语结构都有65%-70%的名词能通用，例如：

粘合式：新时代、伟大时代、旧时代、美好时代、落后时代……（交集为“时代”）

组合式：新的时代、鲜明的时代、辉煌的时代、有趣的时代……

此处的交集为名词相同，形容词相同或不同均有的情况，如“新时代”和“新的时代”这样形容词和名词都相同，只是“的”字隐现的情况将在下文详述。

名词交集中的高频词例（排序前10位）及出现概率如表3所示。

排序	定语成分名词（交集）词例	出现在粘合式结构的概率	出现在组合式结构的概率
1	时代	0.034819592	0.001524357
2	质量	0.025508343	0.00012287
3	问题	0.017556344	0.001923685
4	贡献	0.011503712	0.000861371
5	水平	0.011946557	0.00011903
6	作用	0.010469556	0.001159586
7	讲话	0.011478114	0.000008959
8	地区	0.009455877	0.000766658
9	技术	0.008617545	0.001016238
10	社会	0.00719174	0.00173042

表 3: 粘合式与组合式中心语成分名词（交集）排序前10位的词例

通过对比研究发现，两种定语结构中，形容词和名词都具有位置的倾向性，即使是在能通用的交集中，词例的出现概率也具有明显地高低之分，也就是说，具体词例的实际运用过程中，出现在粘合式还是组合式结构并不是随意的搭配和组合，而是有一定程度的区分性和概率倾向。

3 单项定语中“的”字的隐现类型

上文考察了粘合式与组合式定语词例的分布特征和使用倾向特征，本节将对单项定语中“的”字的隐现情况进行统计和分析。通过梳理语料我们发现，两种定语结构中“的”字的“隐藏”与“出现”可以分为以下四种类型。

①单项定语中“的”字必须隐藏。例如：

（7）扎西是西藏昌都市第一小学的高级教师，主教科目是藏语文。

（8）发展是硬道理，是党执政兴国的第一要务，也是解决一切问题的关键。

以上两例为粘合式“an”定语结构，例（7）中的“高级教师”是指取得高级职称资格的教师，“高级教师”作为一个语义整体，此处“的”字必须隐藏，一般不能用“高级的教师”。例（8）中的“硬道理”源于邓小平在1992年南方谈话时提出的经济思想“发展才是硬道理”，属于政治话

(c) 2022 中国中文信息学会计算语言学专业委员会

语，并且经过时间的累积和沉淀，“硬道理”的语义已经具有凝固性，此处“的”字必须隐藏，不能说成“硬的道理”。目前汉语语法学界对形容词的分类问题还存在着分歧，有些学者将“高级”“低级”此类只能作定语不能作谓语的形容词区分为非谓形容词或区别词(吕叔湘、饶长溶, 1981; 朱德熙, 1982)等，有些学者则按传统的形容词大类进行分类，本文根据分词程序的标注为准，以实用性原则的角度，暂不对形容词内部进行细致的区分。

②单项定语中“的”字必须出现。例如：

（9）老村长说：“过去，我们从来没想过会喝到这么干净的水。”

（10）一位感染新冠肺炎的产妇顺利产下一名男婴，在场的每一个人都流下激动的泪水。

以上两例为组合式“a的n”定语结构，例（9）中“干净的水”，“干净”作为“水”的修饰成分，具有突出强调的作用，意思是现在喝的水比之前的水干净很多，此处“的”字必须出现，一般不能说成“喝到这么干净水”，会造成句意不通顺。例（10）中“激动的泪水”，形容词“激动”表明了强烈的内心情感活动，具有很强的修饰描写性质，此处的“的”字必须出现，一般不能替换为“激动泪水”。

③“的”字可隐藏也可出现，但是表达意义明显不同。例如：

（11）作为新加坡乃至东盟最大的银行，星展银行将继续推动中新金融合作……。

（12）校园市场……，对全国性大银行来说，这块业务很难在战略层面引起重视。

例（11）中“大的银行”是指某一区域中个体形态或占地面积最大的银行，比如“这条街上最大的银行是工商银行”，而例（12）中“大银行”的意思是指银行的性质或业务经营范围是全国性的，而不是地方性或区域性的，虽然“的”字“可隐可现”，但二者的意义差别明显。一些单音节形容词短语的个别用例如“最大”“最硬”等，常与副词连用，其中副词起到补齐音步的作用，本文按照实用性原则将其纳入研究范围。

④单项定语中“的”字可隐藏也可出现，并且表达的意义相近。例如：（13）董事长薛蕾说：“现在产能远远满足不了订单的需求，新的厂房正在加紧建设。”

（14）在厂区一侧，两栋新厂房正在抓紧建设，预计在今年5月底建成。

例（13）（14）中，“新厂房”与“新的厂房”，虽然定语结构不同，但均指新修建或新投入使用的厂房，结构中“的”字“可隐可现”，并且意义差别不大。

依据现有的语料规模，我们分别统计出“的”字必须隐藏（只隐不现）、“的”字必须出现（只现不隐）和“的”字“可隐可现”三种类型的词例数量分布，如图4所示。

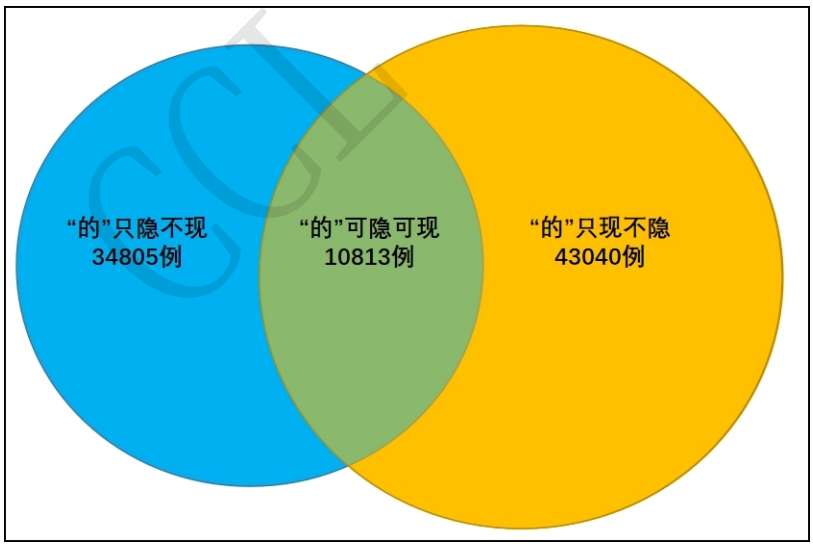


图 4: “的”字隐现词例的分布

从整体分布来看，“的”字“只现不隐”的词例数量最多，“的”字“可隐可现”的词例数量最少，分别占“只隐不现”词例数量的31.07%，占“只现不隐”词例数量的25.12%。

我们发现，真实文本中“的”字的隐藏和出现具有“两极分化”的特征，即大部分词例在使用过程中究竟带“的”还是不带“的”都具有很强的倾向性，如“新的时代、新时代”“硬的骨头、硬骨头”“乐观的心态、乐观心态”“新的厂房→新厂房”这样“的”字“可隐可现”词例数量的比例仅

(c) 2022 中国中文信息学会计算语言学专业委员会

占1/4-1/3之间。

4 音节组配及“的”字的隐现倾向

经统计，两种定语结构中，定语成分形容词和中心语成分名词的音节组配情况如下：

粘合式“an”结构，定语成分形容词共有1351例，其中单音节形容词115例，双音节形容词1236例；中心语成分名词6809例，单音节名词154例，双音节名词5408例，多音节名词[4]1247例。组合式“a的n”结构，共有形容词821例，其中单音节形容词136例，双音节形容词1685例；共有名词7334例，单音节名词183例，双音节名词5602例，多音节名词1549例。

两种定语结构具有共性特点，即均为双音节形容词和双音节名词的数量和比例占优，为了考察各项成分的音节组配模式，我们将“小楼”“小城”“新车”此类词例按音节标记为“单形+单名”，将“重要讲话”“优秀传统”此类词例按音节标记为“双形+双名”，以此类推，具体的音节组配分布及比例如表4所示。

粘合式“an”结构音节组配			
组配模式	词例数量	所占比例（100%）	词例示例
单形+ 单名	229	0.66%	新车
双形+ 单名	101	0.29%	平凡人
单形+ 双名	7160	20.57%	新时代
双形+ 双名	24660	70.85%	重要讲话
单形+ 多名	707	2.03%	大博物馆
双形+ 多名	1948	5.60%	高级工程师
组合式“a的n”结构音节组配			
组配模式	词例数量	所占比例（100%）	词例示例
单形+ 单名	279	0.65%	高的山
双形+ 单名	790	1.84%	可爱的人
单形+ 双名	6330	14.71%	新的历史
双形+ 双名	31310	72.75%	重要的意义
单形+ 多名	1044	2.43%	新的里程碑
双形+ 多名	3287	7.64%	正确的价值观

表 4: 组合式定语结构的音节组配模式

通过对比得知，两种定语结构中，“双形+双名”的词例数量和比例均为最高，我们认为，这与语料的体裁和表达风格相关，《人民日报》作为国家重点新闻网站，内容包括要闻、评论、观察、国际、经济、政治、文化、社会、广告等多个方面，语言内容具有书面、客观、简洁、工整的风格特点，因此双音节词例的数量比例非常高，双音节形容词与双音节名词的组配模式最为普遍。

在“双形+双名”的组配模式中，“的”字“只隐不现”的词例有16889例，占30.18%，高频词例如“高级教师”“高级技师”“先进个人”等；“的”字“只现不隐”的词例有23539例，占42.06%，高频词例如“年幼的孩子”“激动的泪水”“温暖的双手”等；“的”字“可隐可现”的词例仅有7771例，占13.89%，高频词例如“重要（的）讲话”“美丽（的）世界”“热烈（的）掌声”等。

值得注意的是，当中心语成分为单音节名词时，粘合式与组合式结构的音节组配模式有很大差别。我们发现，由单音节名词作中心语所构成的粘合式“an”结构中，“单形+单名”的组配模式占优，并且此类结构“的”字倾向于“只隐不现”的用法，如图5所示。

[4]本文多音节名词指三音节和三音节以上的名词，如“工程师”“行政区”“共产党员”等。

(c) 2022 中国中文信息学会计算语言学专业委员会

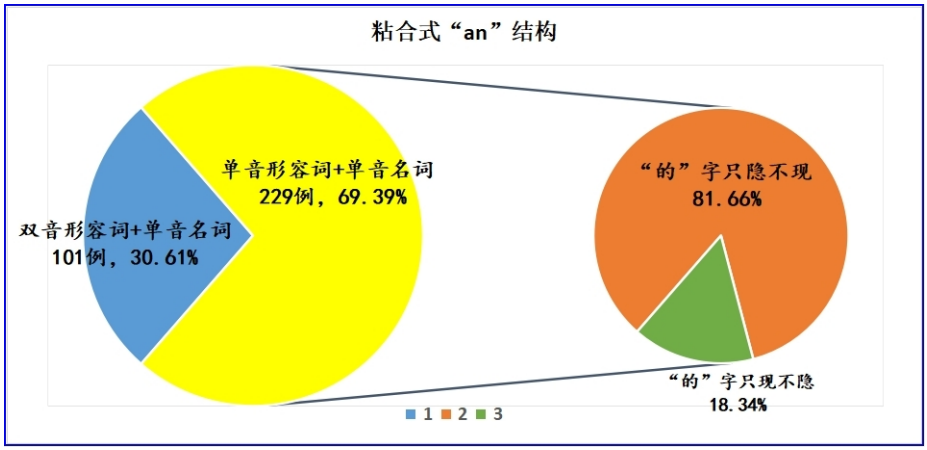


图 5: 单音名词作中心语的粘合式结构音节组配

由图5可见，“单形+单名”的词例比例已经接近70%，并且其中有80%以上都倾向于“的”字“只隐不现”的用法，具体的高频词例（前50例）如下：

大省、新路、大病、小城、小楼、大县、小病、小店、大球、新车、大类、大市、新家、好茶、大山、好菜、好酒、好药、新歌、红线、老城、红墙、长假、老屋、小岛、小县、好店、假酒、老店、小省、好官、小猫、好车、活鱼、小虾、好货、好课、大国、大幕、古井、古村、小树、好字、大沟、小坑、好马、好诗、好米、好兵、好剧

使用实例如下：

（15）中国是农业大国，四川是农业大省。

（16）希望通过与中国的合作，为贫困地区农民增收寻找一条新路。

例（15）（16）中的“大省”“新路”均为“单形+单名”的组配模式，一般情况下，“的”字倾向于“只隐不现”的用法，当然，在一些特殊的条件和语境下，也可以用作“大的省”或“新的路”，但此类词例的种类非常稀少，用法的数量极为有限。

相比而言，由单音节名词作中心语所构成的组合式“a的n”结构中，“双形+单名”的组配模式占优，并且此类结构“的”字倾向于“只现不隐”的用法，如图6所示。

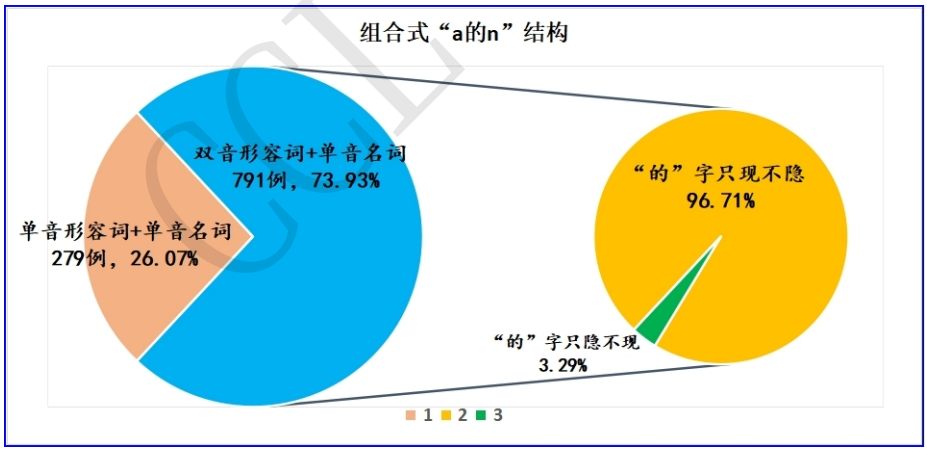


图 6: 单音名词作中心语的组合式结构音节组配

图6中，“双形+单名”的词例比例达到73.93%以上，并且其中绝大部分96.71%都倾向于“的”字“只现不隐”的用法，具体的高频词例（前50位）如下：

可爱的人、温暖的家、有用的人、容易的事、幸福的事、平凡的人、贫困的县、开心的事伟大的党、平凡的事、干净的水、高尚的人、重要的事、幸福的家、善良的心、高兴的事完整的家、快乐的事、抠门的人、优秀的人、简单的事、勇敢的人、粗糙的手、温暖的光洁净的水、轻松的事、光明的路、正确的事、可敬的人、幸福的人、温馨的家、麻烦的事平凡的路、伟大的人、正常的人、清澈的水、健康的人、谦逊的人、健全的人、赤诚的心

(c) 2022 中国中文信息学会计算语言学专业委员会

年轻的脸、年轻的心、勤奋的人、甘甜的水、美丽的花、幸福的梦、炽热的心、勤劳的人
善良的人、有趣的事

使用实例如下：

（17）一线的医务工作者最辛苦，……，是新时代最可爱的人。

（18）她看着可怜的侄子侄女，……，用柔弱肩膀给他们一个温暖的家。

例（17）（18）中“可爱的人”“温暖的家”，均为“双形+单名”组配模式，一般情况下，“的”字倾向于“只现不隐”的用法，“的”字的“出现”起到了突显信息和加深语义情感程度的作用，此处一般不能说成“*可爱人”或“*温暖家”。

5 “的”字的隐现规律分析

上文指出，两种单项定语结构具有“两极分化”的使用特征，即“的”字“可隐可现”的词例数量和比例非常少，那么“的”字的隐现究竟与那些因素相关，在“可隐可现”的词例中，是否具有倾向性？通过对语料数据的梳理和语言现象的观察，我们总结出以下四个特征。

①“的”字的出现具有区分词义的作用。即隐去“的”字后，形容词与名词共同组建成为一个新的词汇或词组，并且具有新的意义。例如：

纯净的水→ 纯净水

优秀的学生→ 优秀学生

高跟的鞋→ 高跟鞋

小的程序→ 小程序

此类词例中“的”字“可隐可现”，但倾向于“隐”的用法。因为隐去“的”字后所形成的“an”结构形式更加简洁，词语的含义范围缩小，如“纯净水”“高跟鞋”等的语义已经凝固，人们倾向于使用“形式简洁、意义明确”的词语来表达和交流，根据词例和数据的具体使用表现来看，“隐”的用例占据绝大部分比例。

②在特定的语境和条件下，“的”字的出现能够起到信息突显和情感强调的作用。例如：

激动的泪水→ *激动泪水

干净的水→ *干净水

温暖的双手→ *温暖双手

优异的中医→ *优异中医

此类词例中“的”字必须出现（或倾向于“现”的用法），比如“激动的泪水”一般不说成“*激动泪水”，因为“激动”一词具有强烈的描写性，表示说话者的内心情绪非常高涨，此时需要“的”字的加入来突显形容词的描写性，即“怎样的泪水？”是“激动的泪水”。前人学者们也对此进行过细致的研究，定语中的“的”字受信息突显原则的影响，“的”字倾向于出现在说话人着意表达的定语性成分之后(谢成名, 2008; 徐阳春, 2011)等。

同样，“温暖的双手”、“优异的中医”中的“的”字都能够起到突显和强调的作用，如果隐去“的”字，容易造成词例结构不完整或语义残缺的现象。

③“的”字隐藏能够促使语义凝练，进一步固化句式结构。当“的”字隐藏时，形容词与名词所形成的某些词例具有了隐喻的意义，实际意义与隐喻意义均可以使用，但倾向于隐喻意义的用法。例如：

实际意义隐喻意义

硬的骨头→ 硬骨头

牛的脾气→ 牛脾气

花的架子→ 花架子

落水的狗→ 落水狗

此类词例中“的”字可隐可现，但其隐现所表达的词义范围有所不同，“的”字出现时表达的是实际意义，词义范围较窄，而“的”字隐去后表达的是隐喻意义，词义范围会得到延伸，相比而言，隐喻意义的用例更多。我们认为，隐喻是人们对已有经验所产生的新的理解和认知，将“的”字隐去后，赋予了词汇或词组新的、更深层的含义，同时，通过隐喻的刻画后，把复杂的词义用更加形象生动地、直观地词例表现出来，让人们在使用过程中更容易理解和便于记忆，因此这类“结构简单、易于理解、含义深远”的词例的使用频率更高。

④“的”字的隐现所指称的对象不同。部分具有区别性意义的形容词后，隐去“的”字所形成的词例具有特指的功能或指称某一种特定的词例类型。例如：

(c) 2022 中国中文信息学会计算语言学专业委员会

小的战士→ 小战士
老的同志→ 老同志
大的货车→ 大货车
大的银行→ 大银行

此类词例中“的”字可隐可现，但其隐现所表达的指称功能不同，比如“小的战士”可以指称年龄小或个子小的战士，而“小战士”则具有特指的功能，经常用于“这位小战士”、“那个小战士”等话语中。又如，“大的货车”意义宽泛，可以指体积大、运载量大的货车，隐去“的”字后的“大货车”意义变窄，专指大型货车这种系列类型的车。从词例的出现概率来看，倾向于“隐”的用法，即特指功能表述的词例数量占优。

综上所述，我们总结了单项定语结构中“的”字隐现的四种规律和特征，但是在区分词例使用情况或“的”字隐现情况是否成立时，一方面可以参照大规模语料库中的使用实例，另一方面可以采用内省的方法，正如任何事物都不是完全绝对的，每个人的语感和判定尺度都不同，每个词语的使用条件和语境都不同，不能说某些词语就绝对不能如何去使用，因此本文探讨定语结构中“的”字的隐现情况，以及各项定语成分的组配规律时，给出的是基于语料使用上的概率倾向。

6 结语

本文以2019-2021年《人民日报》的文章内容作为研究语料，对粘合式与组合式单项定语结构词例的分布情况、音节组配模式和“的”字隐现情况进行了考察和研究。本文通过对粘合式与组合式单项定语词例分布的对比，发现粘合式“an”结构的词例数量明显低于组合式“a的n”结构的词例数量，但粘合式结构词例的使用频数却远远高于组合式结构词例的使用频数，并从受限程度和使用倾向性上给予分析和解释。本文从实用性的角度，对两种结构的定语成分形容词和中心语成分名词的分布情况、共有词例及其出现概率进行了统计和分析，发现形容词和名词重复使用的比例很高，但其共现组合的比例偏少，即使是在能通用的交集中，词例的出现概率也具有明显地高低之分，粘合式或组合式结构的使用并不是随意的搭配，而是有一定程度的区分性和概率倾向。同时，本文还从音节组配的角度对两种定语结构词例进行了分类考察，指出音节组配最常见的、使用最普遍的为“双形+双名”模式，并针对中心语成分为单音节名词时，两种定语词例的不同组配特征和“的”字隐现倾向性进行了总结和分析。另外，本文将单项定语“的”字的隐现情况分为“只隐不现”“只现不隐”和“可隐可现”三种类型，我们发现，真实文本中“的”字的隐现具有“两极分化”的特征，绝大部分词例在使用过程中带“的”或不带“的”都具有很强的倾向性，“的”字“可隐可现”词例比例仅占1/3左右，而且“的”字出现具有区分词义和凸显信息的作用，“的”字隐藏能促使语义更加凝练，进一步固化句式结构，使得某些句式形成了特指或隐喻的表达方式。本文为形容词定语结构的词汇语义研究提供依据和参考，未来的研究计划中，我们将考虑扩展语料规模，收纳其他体裁风格的语料并进行更加深入地对比分析。

参考文献

傅远碧. 2001. 定语的类型. 绵阳师范高等专科学校学报，第*04*期.

刘月华. 1984. 定语的分类和多项定语的顺序. 合肥：安徽教育出版社.

吕叔湘. 1966. 单音形容词用法研究. 中国语文，第*02*期.

吕叔湘、饶长溶. 1981. 试论非谓形容词. 中国语文，第*02*期.

崔应贤. 2002. 现代汉语定语的语序认知研究. 北京：中国社会科学出版社.

徐阳春. 2011. 版块、凸显与“的”字的隐现. 语言教学与研究，第*06*期.

朱德熙. 1956. 现代汉语形容词研究. 语言研究，第*01*期.

朱德熙. 1982. 语法讲义. 北京：商务印书馆.

李先银. 2016. 定名组合的指称功能与汉语多项定语的顺序. 语言与翻译，第*01*期.

王光全、柳英绿. 2006. 定中结构“的”字的隐现规律. 吉林大学社会科学学报，第*02*期.

(c) 2022 中国中文信息学会计算语言学专业委员会

王远杰. 2006. 定语标记“的”的隐现研究. 首都师范大学博士学位论文.

张 蕾. 2004. 定名结构中“的”字隐现规律探析. 湖北大学学报（哲学社会科学版）第*04*期.

袁毓林. 1999. 定语顺序的认知解释及其理论蕴涵. 中国社会科学，第*02*期.

裴泓镔. 2020. 汉语定中结构“的”字隐现问题研究综述. 汉字文化，第*13*期.

谢成名. 2008. 多项定语定中结构中“的”字隐现规律考察. 北京语言大学硕士学位论文.

郑远汉. 2004. 定语后面“的”字的用和不用问题. 修辞学习，第*01*期.

郭 锐. 2000. 表述功能的转化和“的”字的作用.

雷友芳. 2012. 多项定语与“的”字隐现的定量研究. 北京大学硕士学位论文.

CCL 2022

基于GPT-2和互信息的语言单位信息量对韵律特征的影响

郝韵[1]　解焱陆[1]　林炳怀[2]　张劲松[1]
[1]北京语言大学信息科学学院，北京100083
[2]腾讯科技，北京100083
haoyun7725@163.com　xieyanlu@blcu.edu.cn　binghuailin@tencent.com
jinsong.zhang@blcu.edu.cn

摘要

基于信息论的言语产出研究发现携带信息量越大的语言单位，其语音信号越容易被强化。目前的相关研究主要通过自信息的方式衡量语言单位信息量，但该方法难以对长距离的上下文语境进行建模。本研究引入基于预训练语言模型GPT-2和文本-拼音互信息的语言单位信息量衡量方式，考察汉语的单词、韵母和声调信息量对语音产出的韵律特征的影响。研究结果显示汉语中单词和韵母信息量更大时，其韵律特征倾向于被增强，证明了我们提出的方法是有效的。其中信息量效应在音长特征上相比音高和音强特征更显著。

关键词： GPT-2；信息量；韵律；音长；互信息

Prosodic Effects of Speech Unit's Information Based on GPT-2 and Mutual Information

Yun HAO[1]　**Yanlu XIE**[1]　**Binghuai LIN**[2]　**Jinsong ZHANG**[1]
[1]School of Information Science, Beijing Language and Culture University, Beijing, 100083, China
[2]Smart Platform Product Department, Tencent Technology Co., Ltd, Beijing, 100083, China
haoyun7725@163.com　xieyanlu@blcu.edu.cn　binghuailin@tencent.com
jinsong.zhang@blcu.edu.cn

Abstract

Research has shown that linguistic units carrying more information tend to be realized with enhanced speech signals. Most previous studies measure the information that a linguistic unit carries with its surprisal. However, such measurement lacks the ability to model long-distance contextual effects. The current study proposes novel measures of linguistic unit' s information by incorporating the GPT-2 pre-trained language model and mutual information (MI) between text and its phonemic transcription. We examine the prosodic effects of word surprisal and MI-based information of final and tones in Mandarin Chinese. Results show that more information of both words and finals enhance prosodic prominence, proving the validity of our proposed measurements. Besides, the effects of information are more notable on duration feature compared with pitch and intensity feature.

第二十一届中国计算语言学大会论文集，第46页-第55页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

Keywords: GPT-2 , information , prosody , duration , mutual information

1 引言

在语言交流过程中，各语言单位（如词、词素、音素等）所携带的信息量会影响我们感知与产出的难易程度。早在1929年，Zipf (1929)就发现音素的频率与其语音复杂度之间存在着反比关系：音素的频率越高，其语音复杂度就越低。后来的心理语言学与实验语音学领域的研究都提供了类似的证据：Howes and Solomon (1951) 发现辨认单词所需要的视觉呈现时间与根据语料库计算的词频有关：词频越高，被试辨认出所呈现的单词的时间越短。Lieberman (1963)通过填空任务的正确率衡量了英语句子中词的可预测性，并通过语音产出实验发现可预测性更强的词在时长上更短、音高和音强更弱。

真正意义上把语言和信息理论相结合的研究始于Shannon (1949)的信息论。信息论将语言交流视为信息传输的系统：在语言信息传输过程中，说话人将想要传递的信息编码成语音信号，而听者基于噪声下的语音信号对信息进行解码。当信息率为均匀分布且接近信道容量时，可以达到信息传输的最小冗余(Genzel and Charniak, 2002)。在信息论的基础上，Jaeger and Levy (2006)提出了语言传递的均匀信息密度假设(Uniform Information Density Hypothesis)，认为说话人致力于维持一段话语中每单位的Shannon自信息服从均匀分布。Aylett and Turk (2004)也提出了类似思想的平稳信号冗余度假设(Smooth Signal Redundancy Hypothesis)，不同之处在于他们的理论更加强调在语音产出中韵律突显特征对信息量的调节作用：当文本的信息量局部过小或过大时，通过音长、音高或音强的方式弱化或强化语音信号以实现整体上平稳的信息量分布。

韵律突显(prosodic prominence)指一段话语中的某个语音单位在声学或感知上相对突出的特性(Terken and Hermes, 2000; Aylett and Turk, 2004)。韵律突显的声学特性一般体现在时长、音高、音强或其他频谱特征上(Terken and Hermes, 2000)。韵律突显的主要功能便是在话语中突出更加重要、信息量更大的语言单位(Calhoun, 2007)。随着大规模语音语料库的出现，许多研究开始定量探究基于统计的语言单位信息量与韵律特征的关系。Jurafsky et al. (2001) 基于Switchboard 语料库发现英语中单词的频率及Bigram 概率对元音时长有显著的影响。Van Son et al. (2004) 基于荷兰语、芬兰语和俄语语料库发现根据单词中音素的条件概率计算的音素信息量越大，该音素的时长、音强及频谱特征就越容易被加强。

对信息量的韵律效应的研究早期主要集中在印欧语系语言，近年来逐渐拓展到了其他语言如Kaqchikel Mayan语(Tang and Bennett, 2018),日语(Shaw and Kawahara, 2019; Hashimoto, 2021)及包含各语系语言的跨语言比较(Pimentel et al., 2021)。对于汉语中语言单位的信息量，早在20世纪60 年代就有中国科学院声学研究所对汉语的单词出现频率、声韵母及声调的出现频率、声韵母结合概率等进行了统计分析(张家騄, 2010)。关于汉语中信息量与韵律的关系，周韧 (2007)主张句法组合中信息量大的成分将得到重音，而信息量小的成分得不到重音，但并未进行定量的统计分析。Tang and Shaw (2021)基于语料库和Bigram 语言模型发现汉语中词的信息量对时长、音高和音强均有显著的影响，但他们仅探究了词层级的信息量效应，对更细粒度的语言层级(如音素、声母或韵母等)的信息量没有涉及。

传统方法多使用N-gram概率的方法衡量单词或音素等语言单元的信息量，但此类方法无法对长距离的上下文语义关系进行建模(Daland and Zuraw, 2018)。为了解决此问题，本研究提出

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

两种改进的语言单位信息量的计算方法：一种是引入预训练语言模型计算字词的信息量，另一种是引入文本- 拼音互信息的方法计算音位层级的信息量。相比传统的N-gram 语言模型，基于大数据训练的预训练语言模型具有更强的泛化能力，且模型结构中的深度注意力机制可以学习到长距离的上下文语义依赖关系。文本- 拼音互信息方法建立在文本-音位-文本传输模型(Zhang et al., 2008; Zhang et al., 2010)的基础上。在音位的功能负载研究中，该方法可以量化特定音位对辨别语义的贡献程度(Zhang et al., 2010; Wu et al., 2014; Chen et al., 2016; Zhang et al., 2021)。该方法的另一个优势是可以在同样标准下量化不同语音范畴的信息量，包括声韵母、声调、韵律边界等(Wu et al., 2014; Chen et al., 2016)。因此，我们提出使用特定音位信息丢失时文本- 拼音互信息的损失来量化韵母与声调的信息量。基于语音语料库的实验结果证明了汉语的单词和韵母信息量更大时，其韵律特征倾向于被增强，且本研究引入的方法对韵律参数有更好的回归效果。

2 相关工作

2.1 基于自信息的语言单位信息量

目前的信息量的语音产出效应相关研究主要采用基于自信息的方式衡量语言单位的信息量，如式(1)所示。

$$SI_{unit_i} = -\log_2 P(unit_i|context) \tag{1}$$

其中$context$表示该语言单位出现的环境条件。单词自信息的研究中一般计算单词给定前n个词条件下的自信息(Jurafsky et al., 2001; Bell et al., 2009; Tang and Shaw, 2021)。音素自信息的研究中，有些考虑单词中某音素在给定所有前接音素的条件概率(Van Son et al., 2004; Priva, 2015)；也有研究仅考虑给定前一个音素条件下的音素概率(Malisz et al., 2018; Shaw and Kawahara, 2019)。以上方法虽然可以反映单词或音素在给定局部环境条件下的可预测性，但无法对话题、新旧信息等更长距离的上下文依赖关系进行有效的建模(Daland and Zuraw, 2018)。

2.2 文本-拼音互信息理论

文本-拼音互信息理论在Zhang et al. (2008)中首次被提出，以用于为汉语语音识别设计音素集。后来该理论被应用于音系学相关研究，用来计算音位的功能负载(Zhang et al., 2010; Wu et al., 2014; Zhang et al., 2021)。基于该理论的功能负载衡量了音位在受到上下文语境影响的条件下对语言信息传递的重要程度，对我们将要研究的语音单位信息量具有重要启示意义。计算互信息的文本-音位-文本传输模型如图(1)所示。

其中W表示原始文本，即说话人想要传达的信息；F表示W对应的拼音形式；$\hat{W}$表示对F解码得到的所有文本的集合。如果信息编码与解码过程是无损的，应该满足$W = \hat{W}$。然而语言传输过程中可能由于噪声、同义词等因素而产生信息损失。W编码为F需要依赖该语言的音素词典Φ，而F 解码为W需要依赖词典Φ和语言模型LM。W与其拼音形式F之间的互信息定义为式(2)：

$$MI(W;F) = H(W) - H(W|F) \tag{2}$$

互信息量化了一个随机变量在已知另一个随机变量的情况下减少的不确定性。$MI(W;F)$表示根据音素序列F解码出原始文本W的可能性。文本-拼音互信息越大，说明

(c) 2022 中国中文信息学会计算语言学专业委员会

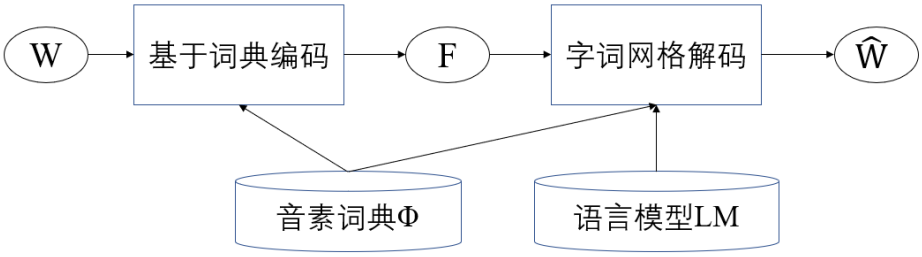


Figure 1: 文本-音位-文本传输模型

越容易从拼音还原出正确的文本内容。根据Shannon-McMillan-Breiman 定理，如果(W, F) 同时是平稳的(stationary)并且是各态经历的(ergodic)，公式(2)可以推导为(3)：

$$MI(W;F) = \lim_{n\to\infty} -\frac{1}{n}\log\sum_{i=1}^{m} P(W_i') \tag{3}$$

其中W_i'表示所有拼音形式为F的文本，即$\hat{W}$中的元素。我们可以根据语言模型计算得到$P(W_i')$。式(3)表明发音为F的文本概率和越大，文本-拼音互信息越小。

3 本文方法

3.1 预训练语言模型GPT-2

随着深度学习及预训练语言模型的兴起，语言研究者们开始关注其对人类语言处理表现上的预测能力。GPT-2是由OpenAI提出的第二代基于Transformer结构的大规模预训练语言模型(Radford et al., 2019)。其模型结构为Transformer的解码器部分，训练目标是对于一段给定文本预测下一个单词的概率分布。GPT-2在生成类似人类创作的文本任务上表现突出，并且在多项预测人类语言处理任务(如眼动数据)上表现优于其他预训练语言模型。例如，Wilcox et al. (2020) 比较了N-gram, LSTM, RNNGs 和GPT-2 模型在预测句子加工时长及眼动表现上的效果，发现GPT-2模型的表现最佳。Hao et al. (2020) 也发现GPT-2 模型在预测阅读句子的眼动数据上表现优于XLM、Transformer-XL等其他预训练语言模型。基于以上背景，本研究尝试将GPT-2预训练语言模型应用于估计汉语单词及声韵母、声调层级信息量，进而探究信息量对语音产出中韵律特征的影响。作为自回归语言模型，GPT-2可以基于给定的上文输入预测单词出现的概率。我们将基于式(4) 计算句子中第t 个单词w_t的自信息，并基于式(5)计算长度为N的句子信息量以用于后续计算文本-拼音的互信息。

$$SI(w_t) = -\log_2 P(w_t|w_1, ..., w_{t-1}) \tag{4}$$

$$SI(s) = -\sum_{t=1}^{N} \log_2 P(w_t|w_1, ..., w_{t-1}) \tag{5}$$

3.2 基于文本-拼音互信息的信息量

我们基于文本-拼音互信息理论提出一种计算语音单位信息量的方法。该方法基于这样的假

(c) 2022 中国中文信息学会计算语言学专业委员会

设：某个语音单位对信息传递的贡献度可以被假设为当该语音单位在传输过程中丢失时（即听话人没有听到该声音），文本-拼音互信息的减少程度。即某语音单位基于互信息的信息量定义为公式(6):

$$MIloss(p) = \frac{MI(W;F) - MI(W;F')}{MI(W;F)} \tag{6}$$

其中p可以表示任何语音单位，包括声韵母、声调等。F表示文本W的规范发音，而F'表示p丢失时的发音。式(6)量化了p丢失的情况下文本-拼音互信息减少的程度，互信息损失越大，说明该语音丢失造成的混淆程度越大，即该语音越重要。图(2)展示了同样的文本内容“你好”在三种不同的语音编码情况下的信息传递过程，其中F表示拼音形式的发音，括号中为声调。当发音为规范发音“ni(3) hao(3)”时可能解码得到包括“你好”、“拟郝”...等的文本序列集合。而当“你”的韵母或声调信息丢失时，解码文本集合$\hat{W}$扩大，可能得到其他发音的文本如“女好”、“哪好”或“尼好”、“逆好”等。在韵母或声调的发音丢失的情况下，由于文本数量增加，文本- 拼音互信息减少。如果增加的文本概率小，那么互信息减少的程度小，说明该语音单位的信息贡献较小；反之如果增加的文本概率大，那么该语音的信息贡献大。

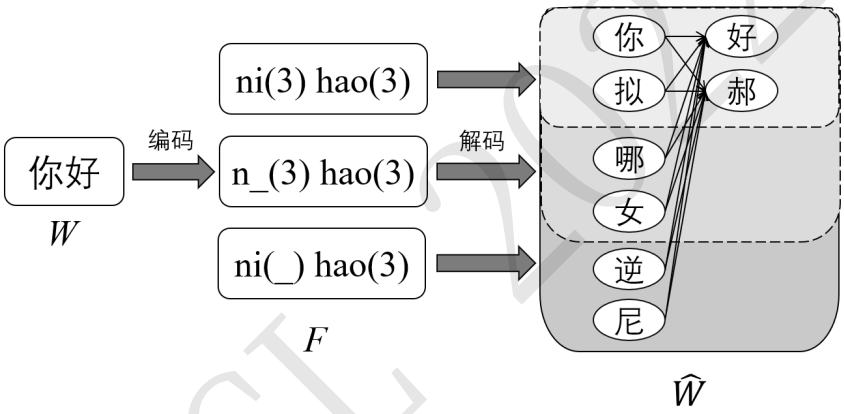


Figure 2: 示例“你好”在拼音信息无损/损失韵母/损失声调情况下的编码与解码过程

计算基于互信息的语音单位信息量的具体过程如下：首先，一个句子被转录成拼音，并进一步转录成音位序列。接下来基于音素词典，原始音位序列被解码成所有可能的文本序列。对于句子中的每一个音节，得到该音节中某个语音单位p信息丢失后的音位序列对应的所有文本序列。最后通过语言模型获得所有文本序列的概率，并根据式(6)计算丢失p后的互信息损失。

4 实验设置

4.1 语音语料

本实验的语音语料来自北京语言大学汉语中介语语料库(曹文and 张劲松, 2009)中的母语部分。语料文本包含选自对外汉语教材的301句话，发音人包含12个中国母语者（6个男性，6 个女性）。我们的实验选取了其中的189个字数为8个字以内的句子。语料库中包含对音节及声韵母边界的标注。我们使用Praat软件在边界标注的基础上提取了每个音节韵母段的时长、音高最大值和最小值、音强最大值。去除未能成功提取音高的音节之后，本实验的最终数据为12459个音

(c) 2022 中国中文信息学会计算语言学专业委员会

节。在后续分析中，我们对所有语音数据进行了发音人归一化处理，并将时长和音高数据取对数以使它们更接近正态分布。

4.2 语言模型

本研究使用的中文预训练GPT-2模型为Du (2019)训练并发布在Github和Huggingface上的GPT2-chinese-cluecorpussmall模型，模型结构包含12 个层，12个注意力头和768个隐藏层节点。模型的训练数据为CLUECorpusSmall(Xu et al., 2020)，包含新闻语料、社区互动语料、维基百科语料和评论数据语料四个部分，总数据量超过14G、50亿字。我们同时训练了先行研究中常用的Bigram语言模型以便与GPT-2模型的结果进行对比：使用KenLM工具包(Heafield, 2011)进行训练，使用modified Kneser-Ney方法(Heafield et al., 2013)进行平滑处理；训练语料为CLUECorpusSmall中的评论数据部分，包含2.3G左右文本、约10亿字。

4.3 统计方法

本研究采用线性混合模型的方法，基于R统计软件的lmerTest库对相关变量进行回归分析。我们对每个因变量基于Bigram 和GPT-2 语言模型的信息量分别进行了线性混合模型的回归。每个回归模型中包含了除信息量之外其他可能影响韵律特征的控制变量，并包含了不同发音人的随机效应。模型的因变量、控制变量和信息量变量如下所示。

- **因变量** 韵母段时长，音高最大值，音高范围，音强最大值。
- **控制变量** 当前音节声母/韵母/声调，前接/后接音节声调，后接音节声母，单词内前/后音节个数，标点符号分割的短句内前/后音节个数，句子内前/后音节个数，句子语速（音节/每秒），由THULAC工具包(孙茂松et al., 2016)得到的词性标注。
- **信息量变量** 单词自信息(Bigram/GPT-2)，韵母合并后的互信息损失(Bigram/GPT-2)，声调合并后的互信息损失(Bigram/GPT-2)。对互信息损失变量取对数，并对所有信息量变量都进行了归一化处理。

在回归过程中，首先对每个因变量建立只对控制变量进行回归的基线模型；再通过向后剔除方法去掉不显著的控制变量；最后在筛选控制变量后的基线模型上分别加入基于Bigram模型的信息量和基于GPT-2 模型的信息量，即对每个因变量最终得到基线、Bigram和GPT-2 三个回归模型。对各模型进行方差膨胀系数检验发现VIF<2.5，说明各模型均不存在多重共线性。

5 实验结果

5.1 信息量对韵律特征的影响

由线性混合模型统计得到各信息量衡量方法分别对4种韵律特征的固定效应及显著性如表5.1所示。其中固定效应β值表示信息量对韵律特征影响的大小和方向，p 值表示影响的显著性，加粗表示$p < 0.05$，即固定效应显著。

对韵母词长的统计结果显示，两种单词自信息量均有正向的显著影响（Bigram：$\beta = 0.037$，$p = 0.01$，GPT-2：$\beta = 0.038$，$p = 0.04$），即单词信息量越大，韵母时长越长。这与前人对其他语言中信息量的语音效应的结论相符，也与Tang and Shaw (2021)对汉语的研究结论相一致。我们提出的基于Bigram 和GPT-2 的韵母互信息损失对时长都有有正向的显著影响（Bigram:

(c) 2022 中国中文信息学会计算语言学专业委员会

信息量	语言模型	音长		音高最大值		音高范围		音强最大值	
		β值	p值	β值	p值	β值	p值	β值	p值
单词自信息	Bigram	0.027	**0.01**	0.012	0.18	0.022	**0.04**	0.003	0.81
	GPT-2	0.038	**0.04**	0.019	0.1	0.033	0.07	0.038	0.82
韵母互信息损失	Bigram	0.062	**<0.001**	-0.008	0.55	0.027	0.06	0.014	0.31
	GPT-2	0.05	**<0.001**	0.011	0.23	0.009	0.44	0.024	0.05
声调互信息损失	Bigram	0.012	0.33	-0.037	**<0.01**	-0.019	0.27	-0.003	0.87
	GPT-2	-0.013	0.14	-0.034	**<0.001**	-0.044	**<0.001**	-0.018	0.25

注：显著性水平$p < 0.001$ ***，$p < 0.01$ **，$p < 0.05$ *，$p < 0.10$.

Table 1: 信息量变量对韵律特征的固定效应及显著性

$\beta = 0.062$，$p < .001$，GPT-2: $\beta = 0.050$，$p < 0.001$)；但声调互信息损失对时长影响不显著(Bigram: $\beta = 0.012$，$p = 0.33$，GPT2: $\beta = -0.013$，$p = 0.14$)。

对于音高特征，信息量效应则多数不显著。基于Bigram和GPT-2的声调互信息损失对音高最大值有显著的负向影响（Bigram: $\beta = -0.037$，$p < 0.01$，GPT-2: $\beta = -0.034$，$p < 0.001$)，这与我们的预期相反。基于Bigram的单词自信息对音高范围由有显著的正向影响（$\beta = 0.022$，$p = 0.04$)，基于GPT-2的单词自信息对音高范围也有接近显著的正向影响（$\beta = 0.033$，$p = 0.07$)，该结果的趋势与前人的发现Tang and Shaw (2021)一致。

对于音强最大值，结果显示只有基于GPT-2 的韵母互信息损失有接近显著的正向影响($\beta = 0.024$，$p = 0.05$)。以上实验结果表明，在三种韵律特征中时长最容易受到信息量效应的影响，而音高和音强较少受到信息量效应的影响。单词自信息和韵母互信息损失对韵律特征的影响都是正向的，这与前人的研究结果及我们的预期相同。声调互信息损失对韵律参数的影响多数不显著，只对音高呈现负向的效应，与我们的预测不符。前人研究很少涉及超音段单位的信息量，我们提出的声调互信息损失是对超音段单位信息量与语音产出之间关系的初步探索，还需要未来进一步探究和讨论。

5.2 对数似然值比较

为了比较加入基于Bigram和GPT-2语言模型的各信息量变量对韵律特征回归的贡献程度，我们引入了对数似然值的变化($\Delta logLikelihood$)，表示加入某变量后与基线模型相比对数似然值的提升（5.2)。对数似然值越大，说明模型对韵律特征的拟合效果越好。正的$\Delta logLikelihood$表明该信息量变量对韵律特征的回归结果有提升。表中加粗显示了效果更优的语言模型。

表(5.2)的结果显示多数信息量变量都可以提升对韵律参数的拟合效果，尤其是对音长参数的拟合效果。其中，基于GPT-2的单词自信息对所有韵律参数的拟合都有帮助，且全部优于基于Bigram 的单词自信息，这说明我们提出的基于预训练语言模型的单词自信息是有效的。基于文本-拼音互信息的韵母和声调互信息损失也有部分可以显著提升模型的拟合效果，说明了我们提出的基于互信息的信息量的有效性。在基于文本- 拼音互信息的韵母和声调信息量中，可以看到韵母互信息损失对韵律参数的贡献较大，优于声调互信息的损失；其中基于Bigram的韵母互信息损失对所有韵律参数的拟合都有帮助；基于GPT-2 的韵母互信息损失在解释音强最大值

(c) 2022 中国中文信息学会计算语言学专业委员会

信息量	语言模型	音长	音高最大值	音高范围	音强最大值
单词自信息	Bigram	29.49	-3.18	0.67	-6.42
	GPT-2	**47.64**	**0.25**	**7.73**	**4.76**
韵母互信息损失	Bigram	**115.5**	**66.7**	**74.17**	70.13
	GPT-2	41.77	-2.99	-3.06	**74.62**
声调互信息损失	Bigram	**37.57**	**6.74**	-1.31	-1.51
	GPT-2	16.44	1.7	**3.14**	**2.67**

Table 2: 各信息量贡献的对数似然值$\Delta logLikelihood$

时优于Bigram 模型，在解释其他韵律参数时效果弱于Bigram 模型。声调互信息损失在音长参数上有较明显的效果，且基于Bigram模型的信息量优于GPT-2，对其余韵律特征的回归贡献则较小。

6 总结与讨论

基于信息传递效率的语言研究认为语音的韵律突显与语言文本的语境信息量呈正相关(Lieberman, 1963; Aylett and Turk, 2004)，且该现象的语音实现存在跨语言的差异(Malisz et al., 2018)。本研究通过基于语料库的实验探究了汉语中语言单位（单词、韵母和声调）的信息量对韵律声学特征（音长、音高和音强）的影响。为了更好地对语境信息进行建模，我们提出了基于预训练语言模型GPT-2 和文本- 拼音互信息的语言单位信息量的衡量方式。在我们提出的两种方法中，基于GPT-2估计的单词信息量相比Bigram模型对韵律参数的拟合有明显提升，这说明了相对于传统方法，预训练语言模型得到的单词信息量可以更好地解释人类语言产出的有关现象；基于文本-拼音互信息的韵母信息量对韵律特征尤其是时长也有显著的正向影响，说明了我们提出方法的有效性。我们还考虑了汉语声调信息量对韵律特征的影响，但其效应普遍不显著，在部分特征上结果与预期不符。对超音段层级信息量的语音效应相关定量研究目前较少，只有一些关于重音可预测性与相关声学参数的讨论(Athanasopoulou et al., 2017)。本研究对于声调信息量的研究是对该方向的初步探索，目前的结果受限于语音语料及语言模型性能等因素的影响，还需未来进一步的探索与考察。

我们的研究结果支持了平稳信号冗余度假设(Aylett and Turk, 2004)，即韵律突显在语音信号中起到了平滑语言的信息量的作用，并且发现时长是汉语中主要体现信息量效应的韵律特征。先行研究对基于N-gram统计的信息量与言语产出/感知的关系研究已有足够相关实验证据及理论，但对深度学习模型训练得到的单词表示与人类语言能力之间关系的探究还在起步中。我们的实验中，基于GPT-2 模型计算的信息量与基于Bigram模型的信息量对韵律特征的效应在显著性上得到了相似的结论，但对各因变量的对数似然值贡献上存在一定差异。在未来的研究中，我们将尝试进一步探索汉语中的声调范畴与语音信息量在韵律表现上的交互作用，并通过引入其他预训练语言模型和进行模型微调等改进信息量的计算方式，继续探索语言单位信息量与人类语言产出/理解的关系。

致谢

本工作得到中央高校基本科研业务专项资金（20YJ040002）、北京语言大学梧桐创新平

(c) 2022 中国中文信息学会计算语言学专业委员会

台(19PT04)、以及语言资源高精尖中心项目“面向智能语音教学的汉语中介语语音多模态语料库研究”(KYR17005)的资助，张劲松是本文的通讯作者。

参考文献

曹文and 张劲松. 2009. 面向计算机辅助正音的汉语中介语语音语料库的创制与标注. 语言文字应用, (4):10.

孙茂松, 陈新雄, 张开旭, 郭志芃, and 刘知远. 2016. Thulac：一个高效的中文词法分析工具包. https://github.com/thunlp/THULAC-Python.

周韧. 2007. 信息量原则与汉语句法组合的韵律模式. 中国语文, (3):15.

张家騄. 2010. 汉语人机语音通信基础. 上海科学技术出版社, 上海.

Angeliki Athanasopoulou, Irene Vogel, and Hossep Dolatian. 2017. Acoustic properties of canonical and non-canonical stress in french, turkish, armenian and brazilian portuguese. In *INTERSPEECH*, pages 1398–1402.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1):31–56.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.

Sasha Calhoun. 2007. *Information structure and the prosodic structure of English: A probabilistic relationship*. Ph.D. thesis, University of Edinburgh.

Yue Chen, Yanlu Xie, Bin Wu, and Jinsong Zhang. 2016. A study on functional load of chinese prosodic boundaries under reduction of syllable information. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

Robert Daland and Kie Zuraw. 2018. Loci and locality of informational effects on phonetic implementation. *Linguistics Vanguard*, 4(s2).

Zeyao Du. 2019. Gpt2-chinese: Tools for training gpt2 model in chinese language. https://github.com/Morizeyao/GPT2-Chinese.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954*.

Daiki Hashimoto. 2021. Probabilistic reduction and mental accumulation in japanese: Frequency, contextual predictability, and average predictability. *Journal of Phonetics*, 87:101061.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Davis H Howes and Richard L Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of experimental psychology*, 41(6):401.

T Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

(c) 2022 中国中文信息学会计算语言学专业委员会

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, 45:229–254.

Philip Lieberman. 1963. Some effects of semantic and grammatical context on the production and perception of speech. *Language and speech*, 6(3):172–187.

Zofia Malisz, Erika Brandt, Bernd Möbius, Yoon Mi Oh, and Bistra Andreeva. 2018. Dimensions of segmental variability: Interaction of prosody and surprisal in six languages. *Frontiers in Communication*, 3:25.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. *arXiv preprint arXiv:2109.15000*.

Uriel Cohen Priva. 2015. Informativity affects consonant duration and deletion rates. *Laboratory phonology*, 6(2):243–278.

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.

Claude E Shannon. 1949. Communication theory of secrecy systems. *The Bell system technical journal*, 28(4):656–715.

Jason A Shaw and Shigeto Kawahara. 2019. Effects of surprisal and entropy on vowel duration in japanese. *Language and speech*, 62(1):80–114.

Kevin Tang and Ryan Bennett. 2018. Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan). *The Journal of the Acoustical Society of America*, 144(2):997–1017.

Kevin Tang and Jason A Shaw. 2021. Prosody leaks into the memories of words. *Cognition*, 210:104601.

Jacques Terken and Dik Hermes. 2000. The perception of prosodic prominence. In *Prosody: Theory and experiment*, pages 89–127. Springer.

Rob Van Son, Olga Bolotova, Louis CW Pols, and Mietta Lennes. 2004. Frequency effects on vowel reduction in three typologically different languages (dutch, finish, russian). In *Interspeech*. Citeseer.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Bin Wu, Jinsong Zhang, and Yanlu Xie. 2014. A clustering analysis of chinese consonants based on functional load. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–4. IEEE.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *ArXiv*, abs/2003.01355.

Jinsong Zhang, Xinhui Hu, and Satoshi Nakamura. 2008. Using mutual information criterion to design an efficient phoneme set for chinese speech recognition. *IEICE TRANSACTIONS on Information and Systems*, 91(3):508–513.

Jinsong Zhang, Wei Li, Yuxia Hou, Wen Cao, and Ziyu Xiong. 2010. A study on functional loads of phonetic contrasts under context based on mutual information of chinese text and phonemes. In *2010 7th International Symposium on Chinese Spoken Language Processing*, pages 194–198. IEEE.

Yuqing Zhang, Zhu Li, Bin Wu, Yanlu Xie, Binghuai Lin, and Jinsong Zhang. 2021. Relationships between perceptual distinctiveness, articulatory complexity and functional load in speech communication. *Proc. Interspeech 2021*, pages 1733–1737.

George Kingsley Zipf. 1929. Relative frequency as a determinant of phonetic change. *Harvard studies in classical philology*, 40:1–95.

(c) 2022 中国中文信息学会计算语言学专业委员会

人文社科学术论文语言变异的多维度分析*

袁亮杰[1]，王治敏[1]，朱宇[2†]
[1]北京语言大学汉语国际教育研究院，北京100083
[2]厦门大学国际中文教育学院，福建厦门361102
{yljlarry,wangzm000}@qq.com, zhuyu@xmu.edu.cn

摘要

通过自建人文和社科领域中文学术期刊论文语料库（逾920万字），运用多维度分析法对111项语言特征的频次数据进行因子分析和维度识别，发现人文社科领域学术论文具有7个维度的语言特征共现模式：描述性vs.阐释性、概念判断vs.行为再现、铺陈与发展、已然性表述、计数与测量、模糊性表达、顺序与连接。进而，对语料在上述各维度的量化表现施以统计检验和聚类分析，发现学术语体在人文与社科两大领域的语言变异显著体现于除“计数与测量”、“顺序与连接”以外的其他5个维度；人文领域和社科领域内部学科的语言变异，各在6个维度上存在显著差异。本研究为学术汉语写作、汉语语体语法等提供一定启示。

关键词： 语言变异；多维度分析；学术汉语

A multi-dimensional analysis of register variations in Chinese academic papers of Humanities and Social Sciences

Yuan Liangjie[1], Wang Zhimin[1], Zhu Yu[2]
[1]Research Institute of International Chinese Language Education,
Beijing Language and Culture University, Beijing 100083
[2]Chinese International Education College,
Xiamen University, Xiamen, Fujian 361102
{yljlarry,wangzm000}@qq.com, zhuyu@xmu.edu.cn

Abstract

Through self-built corpus of Chinese academic journal papers in the field of humanities and social sciences(more than 9.2 million words), using Multi-Dimensional Analysis to conduct factor analysis and dimension identification on the frequency data of 111 language features, it is found that Chinese academic papers in the field of humanities and social sciences have 7 co-occurrence patterns: descriptive vs. interpretative, conceptual judgment vs. behavior reappearance, spreading and development, existentiality statement, counting and measuring, expression of hedging, sequence and connection. Furthermore, one-way MANOVA and cluster analysis reveals variations of academic Chinese journals between humanities and social sciences exist in 5 dimensions, also reveals variations exist within 6 dimensions among the disciplines of humanities and social science. This research provides more objective and pertinent suggestions for academic Chinese writing, Stylistic-Resister Grammar, etc.

Keywords: register variation , Multi-Dimensional Analysis , academic Chinese

* 基金项目：国家社科基金重大项目（18ZDA295）；国家社科基金后期资助项目（21FYYB010）；教育部人文社科研究规划基金项目（21YJA740058）；陕西省教育厅人文社科专项项目（20JK0086）；陕西理工大学校级科研项目（SLGKY2024）

† 通讯作者 corresponding author

(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

学术期刊论文是学术成果的重要载体，其语体功能是准确记述自然、社会及人类思维现象，严密论证其内在规律等(王德春, 2004)，与文学、新闻等相区分。语体动因的功能类型塑造了语法(方梅, 2007)，而恰切的语体观察，是说明语法规律的最佳途径(张伯江, 2012)。语域是与特定使用情境相联系的一种语言变体(Biber and Conrad, 2019)。在学术语体内部，不同学科领域的语言使用情境不同，从而形成各具差异的学科语域。考察不同学科语域的语言特征及其变异情况，能够概括汉语学术语体的语法规律。

随着多维度分析法（Multi-Dimensional Analysis）的日臻完善，近年来学者不断借鉴并跟随其首倡者Douglas Biber(1985; 1986; 1988; 1995; 2006; 2014)的研究路径，借助大规模真实语料库和因子分析的统计方法，得到语言特征在语料中的若干聚合模式（即维度），从而以量化视角探索跨语体/语域的语言变异情况(Biber and Kurjian, 2006; Biber et al., 2016; Biber and Egbert, 2016; Biber and Conrad, 2019)。雷秀云和杨惠中(2001)以及武姜生(2001)是国内较早系统介绍多维度分析法的文献，并籍此引领中国学者对学术英语和其他书面英语的语言变异情况进行描写(武姜生, 2004; 潘, 2012; 胡春雨, 谭金琳, 2020)。

在现代汉语语言变异的多维度分析方面，Zhang(2012)对汉语书面语的语言变异情况进行考察；朱晓楠(2014)考察16个汉语口笔语语域并提取出现代汉语语言变异的5个维度；刘艳春(2019)识别出汉语语体的7个变异维度；范楚琳和刘颖(2020)采用随机森林和k-means聚类算法定量比较了鲁迅三种文体的语言特征；朱宇和胡晓丹(2021)定量考察141项连词的共现模式，创新汉语虚词的量化研究路径。这些文献将多维度分析法运用于汉语研究和自然语言处理领域，是对汉语语体和语言变异进行实证考察的有益尝试，拓展了汉语语言变异研究的视野。但总体来看，国内学界罕见运用多维度分析法对学术汉语进行中观和微观视角下的专门描写与研究。因此，本文基于大规模中文学术期刊语料，尝试分领域、分学科考察学术汉语书面语的语言变异情况。研究问题如下：

（1）在人文领域和社科领域之间，中文学术期刊论文的语言变异有哪些异同?

（2）在人文领域内部和社科领域内部，中文学术期刊论文的语言变异各有哪些异同?

2 研究设计

2.1 语料库建设

为体现样本的典型性和代表性，本文选取CSSCI各学科排名前两位的期刊，建立中文学术期刊语料库（人文社科子库）。其中，人文领域选取中国文学、历史学和哲学，社科领域选取经济学、社会学和法学，共六门学科。具体抽样方法为：确定CSSCI上述六门学科排名前两位的12种期刊，记录其在相同时段内发表的全部学术论文（不含简讯、书讯、征稿启事等），从各学科随机抽取25%作为研究语料，共计970篇，约920万字，见表1。

2.2 语言特征体系建立

本文结合现代汉语特点，参考语体研究成果，建立现代汉语语言特征体系，共21类111项（详见附录A），涉及现代汉语词汇、语法、韵律等多个层面。该特征体系内的语言特征并非独立，而是存在交叉重叠、容纳包含关系。例如名词中既有按使用频次的分类，又有按抽象或具体特征的分类，还有按常用领域的分类等。旨在从多元视角，多方位审视学术汉语语言变异情况。

2.3 语料清洗和处理

为确保数据准确性，删除各篇学术论文的作者简介、基金项目、图表、附录、参考文献、英文摘要、英文关键词等内容。使用NLPIR汉语分词系统[0]对语料进行分词和词性标注。使用AntConc软件和自编程序Text Analysis对每篇文本中语言特征的频数进行检索与统计，并做归一化处理，得到每项语言特征在每篇文本内每千字的出现频率，作为语言特征频率数据。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

[0]https://github.com/NLPIR-team/NLPIR

领域	学科	选取期刊	篇数	字数
人文	中国文学	《文学评论》	104	1 327 487
		《文学遗产》	68	811 081
	历史学	《历史研究》	57	924 922
		《近代史研究》	41	721 352
	哲学	《哲学研究》	160	507 775
		《哲学动态》	130	508 306
社科	经济学	《经济研究》	109	684 100
		《世界经济》	68	755 321
	社会学	《社会学研究》	47	713 853
		《中国人口科学》	53	440 443
	法学	《中国法学》	72	922 832
		《法学研究》	61	874 337
总计			970	9 200 000

表 1: 语料库构成及规模

2.4 因子分析与维度建立

使用SPSS 22.0软件对语言特征频率数据进行因子分析，建立语言特征维度。KMO取样适切性量数0.884，Bartlett球形检验显著性$p<0.001$（Approx.$\chi^2 = 57813.725$，df=4656），表明数据适合进行因子分析。运用未加权的最小二乘法提取因子，并以Promax斜交方式进行因子旋转。经过对比研判和综合考量，参考碎石图曲线，选取前7个因子建立维度（总方差的累积解释为42.913%[1]），排除载荷绝对值小于0.3的特征项，共得到60项语言特征及其载荷值。

3 学科语域文本的维度分析

通过因子分析，本文成功识别出人文社科中文期刊论文的7个维度语言特征集及其载荷值。按照Biber(1988),载荷值越高说明其对该维度的贡献越大，相关性越强。语言特征根据其载荷值的正负分为正向特征和负向特征，代表其在同一维度中的两个方向。每个维度既可包含方向相同的一组特征集（均为正向或负向特征），也可包含方向不同的两组特征集（既有正向也有负向特征），且该两组特征集在某些特定语域中往往呈现互补分布。

进而，根据语言特征频率数据的标准分，按照各维度不同载荷值进行定量加权，得到每篇文本在7个维度上的维度分数。该分数是标准化之后的值，不受文本篇幅等影响，故能够在同一基础上进行量化比较。以下参照各维度语言特征及学科语域维度分均值，结合具体文本，分别对人文社科语域的7个维度进行分析和命名。

3.1 维度一：描述性vs.阐释性

维度一包含33项语言特征，见表2。其中，正向特征的数量和种类较多，如“嵌偶单音词”常见于状语，满足庄典语体需求(黄梅, 冯胜利, 2009)；“副词：中度常用”[2]“形容词：非常用”“副词：最常用”等以形容词和副词为主，多用于描写和修饰；“动词‘有’，表存现”和“存现动词”常用在描述话语；“特指疑问句”、“第一人称代词‘我’”指代确定的人或事物。以上载荷值较高的语言特征均具有一定的描写和叙述功能。

负向特征中，“词长”和“平均句长”是对文本词句的度量，体现词汇和句子的复杂度和丰富度。“名动词”“名词：最常用”“抽象名词”和“集体名词”等承载较高信息密度。“形式动词”在书面

[1]关于本文中该数值的大小是否合理的问题，笔者曾向多维度分析法的首倡者、美国北亚利桑那大学英语系Dr. Douglas Biber教授发邮件请教，得到肯定答复。在此，特向Dr. Douglas Biber教授致以诚挚感谢！

[2]根据国家语委《现代汉语语料库词语分词类频率表》，本文将名词、动词、形容词、副词这四类词语，按照出现频率排序，各分为三个水平：最常用（前30%）、中度常用（31%-65%）、非常用（后35%）。该12项语言特征仅从频率角度划分，与其他语言特征存在交叉重叠。

(c) 2022 中国中文信息学会计算语言学专业委员会

维度	语言特征	载荷值	语言特征	载荷值
维度一	嵌偶单音词	0.974	存现动词	0.326
	语气词	0.855	假设复句	0.324
	其他第一人称代词	0.726	动词：非常用	0.323
	动词“有”，表存现	0.697	名词：非常用	0.320
	目的复句	0.694	特指疑问句	0.320
	副词：中度常用	0.671	第一人称代词“我”	0.312
	指示代词	0.587	状态词	0.307
	形容词：非常用	0.544	集体名词	-0.313
	动词：中度常用	0.476	抽象名词	-0.318
	副词：非常用	0.457	形式动词	-0.372
	副词：最常用	0.445	名词：最常用	-0.392
	交流动词	0.421	助词“的”	-0.457
	并列复句	0.418	平均句长	-0.573
	形容词：中度常用	0.407	合偶双音词	-0.586
	时间副词	0.399	名动词	-0.659
	增强语	0.341	词长	-0.930
	介词短语	0.328		

表 2: 维度一的语言特征及其载荷值

语中“经常用于庄重、典雅的文体”(刁晏斌, 2004)，“合偶双音词”则是体现书面正式语体功能的语法结构手段(王永娜, 2015)。这些特征体现出较强的信息聚集，通常用于阐说和解释。以下通过具体文本分析：

语料（1）节选自历史学期刊论文，其中用不同符号标注出维度一的部分正向特征（形容词、副词、指示代词、**介词**）。该语料包含许多形容词和副词，描述性强；亦含表示对象、范围、方式的介词，以保证描述表达的精确和严密(袁晖, 李熙宗, 2005)。

（1）这样惨切与残败之余的社会生活感受，我们读到《聊斋》故事，即知其**所**描写的环境气氛绝不是偶然的胡思乱想。……具见其**以**己**作为**作品中人物之作，即**以**自己真切深挚的情感意绪**以为**书中人物的描写。即**在**人物创作与描绘**中**，**把**自己的思想感情化了进去。

语料（2）节选自经济学期刊论文，标注出部分负向特征（名动词、名词：最常用、**合偶双音词**）。该语料共两个句子，平均句长71，平均词长1.95，实词占57.53%，用以解释“企业生产效率分化”的作用。可见社科论文对概念和术语的运用较多，解析力求准确、严谨，具有较高信息负荷，体现出较强的阐释功能。

（2）企业之间生产效率的分化，进而激发了产业层面的重组和跨企业的资源再配置过程：一方面，生产率较低的企业逐渐**收缩**，并最终退出市场；另一方面，新企业大规模**进入**，生产率较高的企业也开始扩张。同时，通过**促进**跨企业资源配置效率的**改**善，银行部门的市场化也**成为**中国工业部门全要素生产率增长的重要源泉。

对六门学科和两大领域的文本维度分取均值，绘制出图1，可知：人文领域，三门学科在维度一的均值为正，而社科领域三门学科均值为负，呈现出明显分野。根据4.1节事后检验结果，人文与社科两大领域及各领域内学科之间在维度一均具有显著性差异，这说明学术期刊论文在维度一“描述性vs.阐释性”存在普遍的语言变异情况，也是其最基本的功能。

3.2 维度二：概念判断vs.行为再现

共有15项语言特征聚合成维度二，见表3，其中3项负向特征均属动词词类。“副动词”和“形式动词”用以增强句中谓语的陈述性和宾语的指称性，构成以合偶词为基础的“双必合双”结构(冯胜利等, 2008; 王永娜, 2016)，如“**设法**完成”“**加以**警告”等，呈现“对行为的再现或重构”功能，一般用于说明类和实证类学术论文。

(c) 2022 中国中文信息学会计算语言学专业委员会

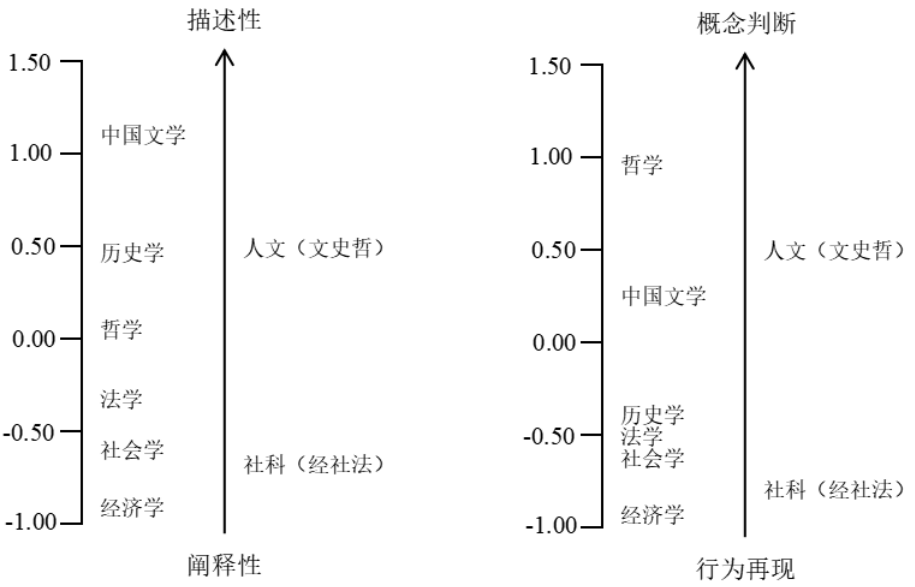


图 1: 维度分均值在维度一和维度二的分布

维度	语言特征	载荷值	语言特征	载荷值
维度二	第三人称代词	0.852	心理名词	0.359
	动词“是”作主要动词	0.744	推测性动词	0.339
	介词“把”	0.744	趋向动词	0.319
	代词：它	0.691	插入语	0.315
	助词“的”	0.570	形式动词	-0.309
	数量词	0.473	副动词	-0.310
	大学学科专业分类词	0.392	动词：非常用	-0.368
	比况助词	0.387		

表 3: 维度二的语言特征及其载荷值

维度二的正向特征“第三人称代词”“动词‘是’作主要动词”“代词‘它’”等，在书面语中用于对客观的人、事物或现象进行概念判断或特点概括，是一种“下定义”式的表达，以突出其固有的、区别于他者的特点。维度二的其他正向特征，如“数量词”、“大学学科专业分类词”、“比况助词”等起到限定、指代、说明等作用，让一些抽象概念被解析得更加明确。

语料（3）节选自哲学期刊论文，标注出维度二的部分正向特征（代词：它、**动词“是”作主要动词**、介词“把”、助词“的”）。语料（4）节选自社会学期刊论文，标注出维度二的负向特征（动词：非常用、**副动词**、形式动词）。

（3）第二种理论主张可称为有意识的倾向主义。它坚持认为，一个命题态度**是**由某些对意识显现状态倾向所构成的。第三种可称为自然种类主义。它坚持认为，一种命题态度**是**无论什么自然种类发挥的所有作用，既**是**行为的又**是**现象的，并且我们把这种作用相关于这种态度。

（4）在使用年龄偏好指数进行**相应**计算时有一个前提假设条件。同时，单岁组计算时分母**相应**增多，也稀释了流动人口集中带来的影响。对此，本文提出对联合国综合指数进行修正，以单岁组代替5岁组为单位进行分析计算。

从图1的维度分均值分布可知，中国文学和哲学在维度二均值为正，尤以哲学表现出较强的“概念判断”特征；而历史学和社科领域的三门学科为负值，表现出与之对照的“行为再现”特征。从某种意义上来说，历史学也是基于史料的论证和探索，包含对先前研究行为的再现，故一些历史学论文在该维度上呈现出“行为再现”的功能。这体现了历史学在表达“行为再现”功能时区别于其他人文学科的独特性。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.3 维度三：铺陈与发展

维度三的语言特征包括四类复句和两种副词，均为正向，且绝对值均大于0.4，见表4。在汉语中，复句各分句之间关系复杂，但表述层次清楚，语意周密(吴礼权, 2012)。该维度的复句中，三种属于偏正复句，一种（顺承复句）属于联合复句，使成分之间联系紧密，铺陈接续。

维度	语言特征	载荷值	语言特征	载荷值
维度三	因果复句	0.771	时间副词	0.498
	条件复句	0.715	副词：表态度	0.474
	顺承复句	0.626	假设复句	0.432

表 4: 维度三的语言特征及其载荷值

哲学期刊论文在该维度的均值最高（图2），节选语料（5）标注出两种复句（因果复句、**条件复句**），体现出较强的逻辑关系。

（5）在基督教中，**无论**是否亲历基督复活之事，一个人**只要**相信此事，他**就**在精神上成为了圣保罗的当代人，……之所以是复活而不是回忆，因为历史虽可以回想，却无法在苏格拉底意义上被回忆。……**只有**开天辟地的造物主**才**能在无宾语状态下仅凭自身而存在。

基于以上特征共同表达的功能，将维度三命名为“铺陈与发展”。这是一种对客观事实的描述和看法，强调时间或逻辑上的紧密联系，包含一定程度的主观性与非理性因素，具有人为参与感。

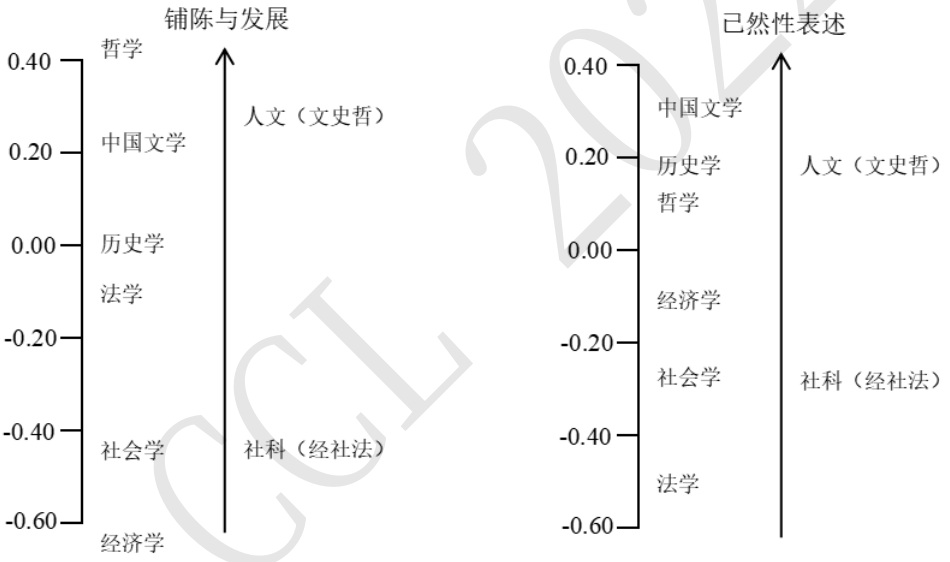


图 2: 维度分均值在维度三和维度四的分布

3.4 维度四：已然性表述

维度四由“过去式动词”和“助词‘了’”组成，均具有极高载荷值，见表5。二者关联性较强，都是对过去发生的动作或行为的表述。

维度	语言特征	载荷值	语言特征	载荷值
维度四	过去式动词	1.030	助词“了”	0.988

表 5: 维度四的语言特征及其载荷值

以汉语为代表的孤立语主要通过虚词的语法手段来表达时间范畴。本文中，“过去式动词”的检索方式为“表示过去的副词+动词”，较多分布在陈述性的话语中。助词“了”是汉语的三

(c) 2022 中国中文信息学会计算语言学专业委员会

个动态助词之一，除了表示过去的意义，还可表达完成的意义。由此可见，维度四的两项语言特征表达相似的时和体意义，具有高度相关的功能。

语料（6）节选自中国文学论文，标注了维度4的语言特征（**助词"了"**、过去式动词），更加清晰地看出作者对已然性的表述。文中通过对莫言人生经历的回顾，表现作者对莫言及其作品研究的历程，为文章增添说服力。

（6）理论上讲，作家莫言早已成了都市人。……近年来，有诸多研究者已发现莫言与沈从文经历的相近，……这一写作模式深刻影响了几代中国作家的回乡写作。……事实上，批评家们已然注意到，以《红高粱》为标志，莫言已经寻找到他的"精神家园"……她嫁了个哑巴丈夫，生了个哑巴孩子。

由图2可知，人文领域的三门学科在该维度均为正值，而社科领域的三门学科均为负值，呈现出与维度一类似的学科分野。事后检验结果表明，该维度中人文领域的三门学科之间均不具有显著性差异，说明"已然性表述"亦是人文领域的常见表达功能。

3.5 维度五：计数与测量

维度五共4项特征，均为数词和量词，见表6。在汉语中，二者通常组合成数量短语作句法成分，因而其出现频率和分布具有正相关。量词具有一定的感情色彩义，对于相同的对象，使用不同量词会产生不同的感情色彩。由于科学语体在词语运用上排斥描绘形象色彩和感情色彩等(袁晖, 李熙宗, 2005)，因而法学期刊论文中数词和量词出现频率较低，见图3。同理，哲学概念往往由于其抽象性而无对应的量词，故哲学语域表现出与法学类似的情形。

维度	语言特征	载荷值	语言特征	载荷值
维度五	动量词	-0.315	数词	-0.822
	时量词	-0.678	量词	-0.867

表 6: 维度五的语言特征及其载荷值

语料（7）和（8）分别来自经济学和历史学，标注了维度五的全部四项语言特征（量词、数词、**时量词**、动量词），从中可见对"计数与测量"表达的严密性和客观性。

（7）利用中国85个大中城市的环境污染和房价数据，……这是首次运用发展中国家的数据检验城市化进程中……近10年来，中国经历了快速的房价上涨，……这4个城市的房价分别上涨了2.18倍、1.89倍、2.16倍和1.81倍。

（8）每在黄道上移动15度时，即定为一节气的日期。……以致有可能出现一月之中有三气，偶而也会出现一年之内有两个无中气之月的现象。……清初以来三百五十年间，平均每半个世纪才发生一次的闰八月，……将是社会天文学史研究的另一重要内容。

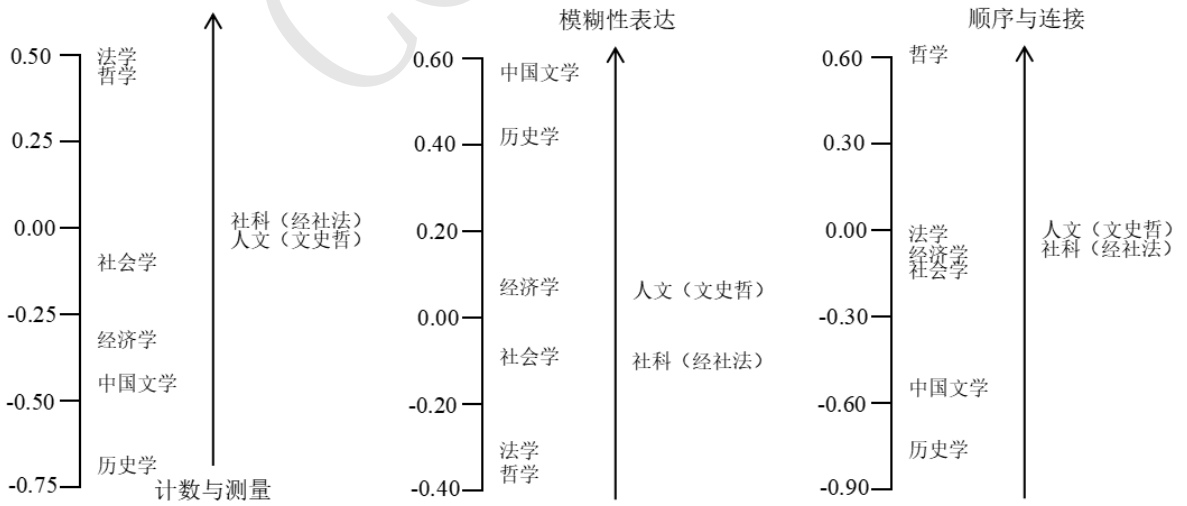


图 3: 维度分均值在维度五、维度六和维度七的分布

(c) 2022 中国中文信息学会计算语言学专业委员会

3.6 维度六：模糊性表达

维度六包含4项语言特征，见表7。其中“模糊限制语”和“副词：表可能”载荷值均大于0.5，二者分别是对模糊性和可能性的表达，且存在一定重叠，在使用中具有较高关联性和替代性。“转折复句”通过关联词语表达转折关系范畴，在语义特征方面具有相对性和补充性，体现信息的翻转和婉转(赵岩, 2021)，反映交互双方的信息不确定或不对称。“插入语”属于独立语范畴，用于补足句义，使句子表意更加严密，如“据说”“在我看来”等。

维度	语言特征	载荷值	语言特征	载荷值
维度六	模糊限制语	0.868	转折复句	0.491
	副词：表可能	0.643	插入语	0.350

表 7: 维度六的语言特征及其载荷值

以上特征共同实现对事物发展变化的模糊性推断和判断，表示不肯定或不确切的意味。由图3，中国文学和历史学在该维度体现明显。语料（9）节选自历史学期刊论文，标注维度六的全部特征（模糊限制语、**副词：表可能**、转折复句、插入语），表达因年代久远而对历史事件的推测或假设。

（9）这当然只是胡适和陈垣个人之间的争论，其中**或许**有胡适和陈垣之间的瑜亮心结，所以话题变得**比较**敏感。但是，在现代中国的佛教道教研究史中……不过，事实上彼此也还是有些在意。4月10日，陈垣写了一封**似乎**是高挂“免战牌”的信，……换句话说，即佛教传到中国，……是否还**可能**经由南海到达交、广地区这一途径?

3.7 维度七：顺序与连接

维度七由载荷值均较高的“顺序词”和“句内并列连词”组成，见表8。在汉语中，二者都可归为连词范畴，都是虚词。顺序词体现先后的逻辑，并列连词是对具有并联关系成分间的连接。该两项特征均属于明晰化手段，能让文本的层次、语义关系变得明白确切(袁晖, 李熙宗, 2005)，增强表达的精确性。

维度	语言特征	载荷值	语言特征	载荷值
维度七	顺序词	1.097	句内并列连词	0.897

表 8: 维度七的语言特征及其载荷值

由图3，哲学论文在该维度的表现尤为突出。语料（10）节选自哲学期刊论文，标注维度七的语言特征（顺序词、**句内并列连词**）。句中并列连词和顺序词对文本架构进行梳理和推动，使行文在叙述抽象概念时更有逻辑和层次。

(10)对于确定性的追问首先表现在对世界存在、变化**和**发展的始基、根据**和**解答上。……其次，经过实在**与**表象、本质**与**现象二元论的剥离，……最后，以同一性排斥差异性，**并**最终导致……人类认识的起始**和**真理标准**及**意义界限等……

4 学科语域的差异检验及聚类分析

4.1 基于方差分析的差异检验

通过维度分均值对比和语料分析，本文从7个维度对不同学科语域的差异进行分析。为了验证以上差异在哪些方面具有统计学意义的显著性，本文使用多变量方差分析（MANOVA）对各文本的7个维度分数据进行差异检验。

首先，使用基于马氏距离（Mahalanobis distances）的方法对多元异常值进行检测并剔除。结果显示，数据中有17项异常值，剔除后剩余953篇有效语料。同时，使用皮尔逊相关系数对多因变量的多重共线性进行检验[3]，发现各因变量间相关系数的最大值r=0.591＜0.90，即因变量间的多重共线性不强，满足进行后续统计分析所需条件。

[3]Tabachnick and Fidell(2013)指出，具有较强相关性的因变量之间存在多重共线性，其相关系数为r=0.90或更高。故使用皮尔逊相关系数对多因变量进行多重共线性检验，若r值小于0.90，则可认为因变量之间的多重共线性较弱，对结果的影响较小。

(c) 2022 中国中文信息学会计算语言学专业委员会

接着，以领域（人文、社科）、人文学科（文、史、哲）和社科学科（经、社、法）分别作为分组变量，对各组内的文本维度分进行三次协方差矩阵齐性假设检验，发现三个Box's M协方差齐性检验结果均显著。进而采用Pillai's trace校正统计量[4]分别再次检验，发现三次Pillai's trace检验的p值均接近0，说明三种分组情况下的文本维度分在本文建立的7个维度上具有显著性差异。

随后，为验证各单一维度下的学科语域差异是否显著，使用单变量方差分析，将单一维度下的误差方差进行分解。方差齐性检验表明，各单一维度下的方差同质性Levene's检验结果均显著。为此，使用修正方法Welch's ANOVA对三种分组方式下每个维度的文本维度分数据进行组间差异性检验，领域分组检验结果如表9所示，学科分组主体间效应检验结果见附录B。

来源	因变量	III类平方和	自由度	均方	F值	p值	η^2
领域	维度一分数	300.253	1	300.253	491.105	0.00	0.341
分组	维度二分数	365.710	1	365.710	656.067	0.00	0.408
	维度三分数	108.221	1	108.221	144.646	0.00	0.132
	维度四分数	55.468	1	55.468	58.004	0.00	0.057
	维度五分数	0.069	1	0.069	0.087	0.768	0.00
	维度六分数	3.808	1	3.808	4.538	0.033	0.005
	维度七分数	0.188	1	0.188	0.196	0.658	0.00

表 9: 人文社科领域7个维度文本维度分的主体间效应检验结果

结果显示，人文和社科领域之间的语域变异显著存在于5个维度（p值小于0.05），分别是：维度一"描述性vs.阐释性"、维度二"概念判断vs.行为再现"、维度三"铺陈与发展"、维度四"已然性表述"和维度六"模糊性表达"。同时，文、史、哲期刊论文或经、社、法期刊论文各有6个维度呈现显著性差异，说明这两大领域内部学科亦存在凸出的语言变异情况。

为了探究人文和社科两大领域内部的学科在7个维度上各自存在哪些差异，本文以上述结果为基础，在95%置信区间内采用Dunnett's T3检验法[5]进行事后检验，各学科的多重比较显著性结果可概括为表10，详细结果见附录C。

	中国文学	历史学		经济学	社会学
历史学	I,II,III	/	社会学	I,II,V	/
哲学	I,II,V,VI,VII	I,II,III,V,VI,VII	法学	I,II,III,IV,V,VI	I,III,IV,V

表 10: 事后检验呈现显著差异的维度简表

以上结果表明，人文领域内部学科在除维度四"已然性表述"外均具有显著差异，而社科领域内部学科在除维度七"顺序与连接"外均具有显著差异。其中，哲学与中国文学在5个维度存在显著差异，与历史学在6个维度存在显著差异，说明哲学语域跟同属人文领域的中国文学、历史学语域存在较大差异；法学与经济学在6个维度上存在显著差异，与社会学在4个维度上存在显著差异，说明法学语域与同属社科领域的经济学、社会学语域差异较大。

4.2 学科语域的聚类分析

为了进一步量化比较不同学科语域之间的差异，本文使用Minitab软件进行聚类分析。具体方法为：以构成7个维度的60项语言特征的均值为依据，通过计算值间距离，对六门学科语域进行聚类。相关参数设定为：联结法为最长距离法，距离度量为Euclidean平方，标准化变量，相似性水平为95%。聚类结果如图4所示。

[4]对于Pillai's trace校正统计量，Pillai(1955)指出该方法通常被认为是最有效且稳健的，是在不满足协方差矩阵同质性假设的情况下进行MANOVA的最佳选择。

[5]根据Sauder and DeMars(2019)，在不满足方差齐性的情况下，使用Dunnett's T3进行事后检验能够有效控制第一类错误，是一种比较保守的多重比较方法。

(c) 2022 中国中文信息学会计算语言学专业委员会

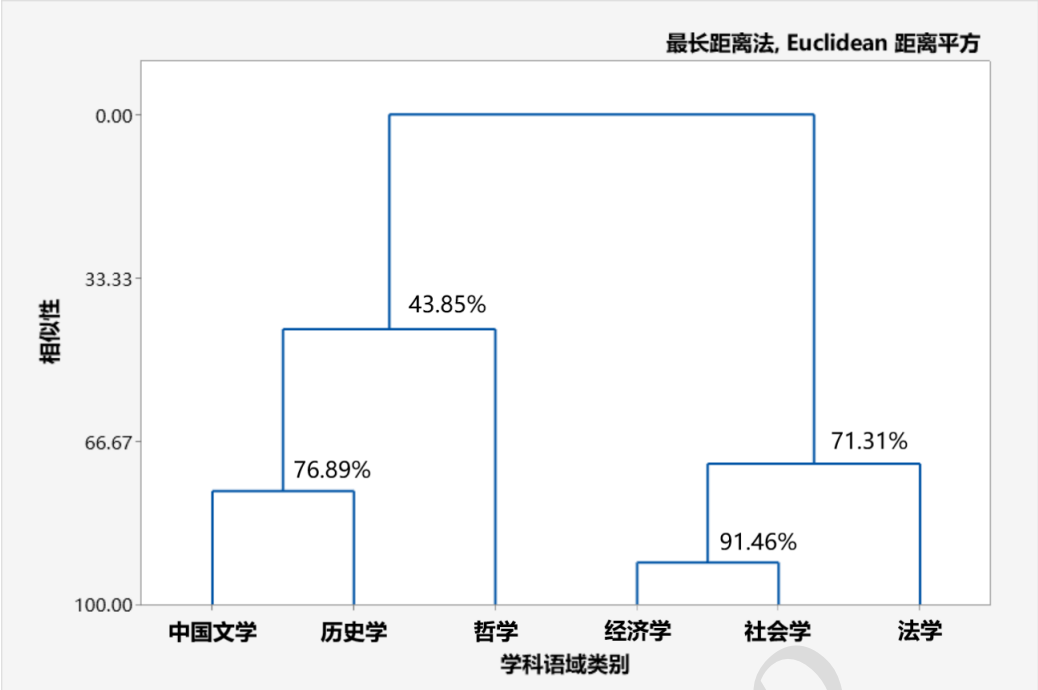


图 4: 学科语域的聚类

由图可知，六门学科的聚类结果分明，共有两个层次。首先，六门学科被分为两大聚类集合：第一大聚类集合包括中国文学、历史学、哲学，第二大聚类集合包括经济学、社会学、法学——这与本文选取的“人文”和“社科”两大领域下的学科分布完全一致。其次，在人文领域内部，中国文学与历史学的相似度为76.89%，哲学与其相似度较低，仅为43.85%；在社科领域内部，经济学与社会学的相似度为91.46%，法学与其相似度为71.31%。以上数据实现了对六门学科之间文本相似性的度量，相关结论与事后检验结果一致，从而量化反映出人文社科领域语域及其内部学科语域之间的差异大小。

5 结语及展望

语体语法是现代汉语领域新近形成的重要语法研究流派。该流派秉持“大语法观”，认为语体语法的特征体现于“音系、韵律、词法、句法、篇章、语义”各层面语言表达中(冯胜利, 施春宏, 2018)，凡语言系统中某种结构形式的变化和某项特征的有无或差异能够体现为语体功能的对立或改变，都可成为实现语体的表现手段和方式。

以上观点与本文所使用的语域多维度分析法的底层逻辑存在高度一致性。具体而言，本文发现：汉语学术语体内部存在显著的语言形式与功能差异，仅以其中的人文社科期刊语域而言，其语法范畴层面的语言特征在使用频数上具有明显差异，这些差异经过因子分析可概括为7个语言功能差异维度（即：描述性vs.阐释性、概念判断vs.行为再现、铺陈与发展、已然性表述、计数与测量、模糊性表达、顺序与连接）。并且，人文与社科语域在其中五个维度的表征上存在显著差异，人文或社科内部各自在其中六个维度的表征上存在显著差异。

这为语体语法理论的发展完善提供了具体的资料与例证。尤其是考虑到当前汉语语体语法的研究更多采用讨论某一语法特征在两种特定语体中“合法”或“不合法”的两极判断范式，而以本文所示的研究范式，则可有助于构建语法特征在不同的语体/语域中由“不合法”到“合法”的连续统，同时也可以看到多个语法特征使用频数的差异是如何实现不同的差异维度（语言功能），而若干语言功能的定量表达又是怎样构建了不同的语域/语体，这对于丰富和发展汉语语体语法理论有重要启示作用。

此外，本文的发现对人文社科领域的人才培养，特别是学术写作教学实践也具有指导作用。当前国内高校对学术写作的教学指导多停留在学位论文写作方面。诚然，学位论文固然重要，但其仅仅是学术写作诸多形式之一，况且当前学位论文写作的教学重点更多强调学位论文的程序性和规范性，而非学术汉语语言特征的使用。本文通过人文社科中文学术期刊语言差异

(c) 2022 中国中文信息学会计算语言学专业委员会

研究的发现，凸显了学术汉语语言表征的多样化功能及形式的丰富性，在一定程度上揭示出人文和社科领域六个具体学科的学术语言表征特点，对于学术写作教学有具体的参照借鉴意义。

后续工作可从以下方面继续深入：1）在学科语域选择方面，引入更多领域和学科进行对比，如人文社科与理工学科的对比，语言学、心理学等跨领域多栖学科的对比等；2）在语料库建设方面，除了继续扩充学术期刊、学位论文、学术教材等书面语语料外，亦可增添学术讲座、课堂授课、研讨会发言等学术口语语料，丰富学术汉语语料库的类型和规模，从而更加全面地考察学术汉语语域变异情况。3）另外，相关应用研究也可开展，如：基于学科语域变异规律的国内高校学术写作教学研究、汉语作为第二语言的学术语体教学研究、“中文+职业教育”分科教材编写研究等。

参考文献

Barbara G. Tabachnick, Linda S. Fidell. 2013. *Using multivariate statistics(6th edition)*. Boston: Pearson Education, Inc.

Derek C. Sauder, Christine E. DeMars. 2019. An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science*. 2(1):26-44.

Douglas Biber. 1985. Investigating macroscopic textual variation through multi-feature/multi-dimensional analysis. *Linguistics*. (2):337-360.

Douglas Biber. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*. (2):384-414.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Douglas Biber. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

Douglas Biber. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins Publishing Company.

Douglas Biber. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*. (1):7-34.

Douglas Biber, Bethany Gray, Shelley Staples. 2016. Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*. 37(5): 639-668.

Douglas Biber, Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *English Linguistics*. 44(2):95-137.

Douglas Biber, Jerry Kurjian. 2006. Towards a taxonomy of web registers and text types: A multidimensional analysis. *Language and Computers*. (1):109-131.

Douglas Biber, Susan Conrad. 2019. *Register, Genre and Style (2nd edition)*. Cambridge: Cambridge University Press.

K. C. S. Pillai 1955. Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics*. 26(1):117-121.

Zheng-sheng Zhang. 2012. A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory*. 8(1):209-240.

刁晏斌. 2004. 试论现代汉语形式动词的功能. 宁夏大学学报. (3):33-38.

范楚琳, 刘颖. 2010. 基于多维度分析法的鲁迅三种文体比较研究. 中文信息学报. 34(10):94-104.

方梅. 2007. 语体动因对句法的塑造. 修辞学习. (6):1-7.

冯胜利, 王洁, 黄梅. 2008. 汉语书面语体庄雅度的自动测量. 语言科学. (2):113-126.

冯胜利, 施春宏. 2018. 论语体语法的基本原理、单位层级和语体系统. 世界汉语教学. (3):302-325.

胡春雨, 谭金琳. 2020. 中美企业致股东信语域特征的多维分析. 外语与外语教学. (6):66-75.

(c) 2022 中国中文信息学会计算语言学专业委员会

黄梅, 冯胜利. 2009. 嵌偶单音词句法分布刍析——嵌偶单音词最常见于状语探因. 中国语文. (1):32-44.

雷秀云, 杨惠中. 2001. 基于语料库的研究方法及MD/MF模型与学术英语语体研究. 当代语言学. (3):143-158.

刘艳春. 2019. 汉语语体变异的多维度分析——基于17个语体72项语言特征的考察. 江汉学术. (3):100-110.

潘. 2012. 中国非英语专业本科生和研究生书面语体的多特征多维度调查. 外语教学与研究. (2):220-232.

王德春. 2004. 大学修辞学. 福州：福建人民出版社.

王永娜. 2015. 汉语合偶双音词. 北京：北京语言大学出版社.

王永娜. 2016. 汉语书面正式语体语法的泛时空化特征研究. 北京：中国社会科学出版社.

武姜生. 2001. 语域变异的多维向分析模式简介. 解放军外国语学院学报. (3):6-9.

武姜生. 2004. “学术交流e-mail”文体特征的多维度分析. 外语与外语教学. (2):53-57.

吴礼权. 2012. 现代汉语修辞学. 上海：复旦大学出版社.

袁晖, 李熙宗. 2005. 汉语语体概论. 北京：商务印书馆.

张伯江. 2012. 以语法解释为目的的语体研究. 当代修辞学. (6):13-22.

赵岩. 2021. 现代汉语转折关系范畴研究. 长春：吉林大学博士学位论文.

朱晓楠. 2014. 汉语普通话的多维语域变异研究. 杭州：浙江大学硕士学位论文.

朱宇, 胡晓丹. 2021. 汉语连词在不同学术语域的聚合：多维度定量分析. 语言教学与研究. (2):57-69.

附录A.现代汉语语言特征体系

名词及其特殊属类：1.名词：最常用；2.名词：中度常用；3.名词：非常用；4.抽象名词；5.立场名词；6.心理名词；7.大学学科专业分类词；8.指人名词；9.集体名词；10.具象名词；11.具象科技名词；12.度量衡名词

动词及其特殊属类：13.动词：最常用；14.动词：中度常用；15.动词：非常用；16.动作行为动词；17.使役动词；18.存现动词；19.心理动词；20.肯定动词；21.交流动词；22.推测性动词；23.可能性动词；24.趋向动词；25.副动词；26.形式动词

形容词及其特殊属类：27.形容词：最常用；28.形容词：中度常用；29.形容词：非常用；30.大小形容词；31.常用形容词：相关性；32.状态词；33.区别词

数词和量词：34.数词；35.数量词；36.量词；37.动量词；38.时量词

代词和代动词：39.第一人称代词“我”；40.第一人称代词“我们”；41.其他第一人称代词；42.第二人称代词；43.第三人称代词；44.代词：它；45.普通名词+们；46.指示代词；47.不定代词；48.句首代词；49.代动词

副词及其特殊属类：50.副词：最常用；51.副词：中度常用；52.副词：非常用；53.副词：表必然性；54.副词：表可能；55.副词：表态度

介词及其短语：56.句首介词；57.介词短语

助词：58.助词“的”；59.助词“地”；60.助词“得”；61.助词“等”/“等等”；62.比况助词

其他词汇属类：63.语气词；64.顺序词；65.句内并列连词；66.小品词；67.口语词；68.儿化词；69.嵌偶单音词；70.合偶双音词；71.古语词；72.增强语；73.模糊限制语

名词形式：74.名物化；75.名动词；76.名形词

状态形式：77.动词“是”作主要动词；78.动词“有”，表存现

时态和体标记：79.进行式动词；80.过去式动词；81.助词“着”；82.助词“了”；83.助词“过”

情态动词：84.必要性情态动词；85.情态动词：表未来

地点和时间状语：86.时间副词；87.副词：表处所；88.处所词；89.方位词

(c) 2022 中国中文信息学会计算语言学专业委员会

缩略形式：90.缩略语
否定形式：91.否定词
独立语：92.叹词；93.拟声词；94.插入语
疑问句：95.特指疑问句
被动形式：96.被动；97.无施事者被动句；98.介词“把”
复句：99.并列复句；100.顺承复句；101.解说复句；102.选择复句；103.递进复句；104.条件复句；105.假设复句；106.因果复句；107.目的复句；108.转折复句
词汇丰度：109.类符/形符比；110.词长；111.平均句长

附录B.学科分组主体间效应检验结果

文史哲三学科7个维度文本维度分的主体间效应检验结果

来源	因变量	III类平方和	自由度	均方	F值	p值	η^2
人文	维度一分数	122.943	2	61.471	87.686	0.00	0.245
分组	维度二分数	135.431	2	67.761	137.736	0.00	0.338
	维度三分数	15.489	2	7.744	8.979	0.00	0.032
	维度四分数	3.645	2	1.823	1.582	0.206	0.006
	维度五分数	138.147	2	69.074	96.284	0.00	0.263
	维度六分数	111.104	2	55.552	66.571	0.00	0.198
	维度七分数	211.572	2	105.786	119.934	0.00	0.308

经社法三学科7个维度文本维度分的主体间效应检验结果

来源	因变量	III类平方和	自由度	均方	F值	p值	η^2
社科	维度一分数	15.041	2	7.520	47.179	0.00	0.188
分组	维度二分数	11.418	2	5.709	19.728	0.00	0.008
	维度三分数	21.096	2	10.548	20.521	0.00	0.092
	维度四分数	11.395	2	5.698	8.516	0.00	0.040
	维度五分数	49.154	2	24.577	56.753	0.00	0.218
	维度六分数	11.403	2	5.701	10.314	0.00	0.048
	维度七分数	0.066	2	0.033	0.06	0.942	0.00

CCL 2022

附录C.人文社科领域内部学科在7个维度的事后多重比较显著性结果（p值）

维度一：描述性vs.阐释性					
	中国文学	历史学		经济学	社会学
历史学	0.000	/	社会学	0.000	/
哲学	0.000	0.000	法学	0.000	0.021
维度二：概念判断vs.行为再现					
	中国文学	历史学		经济学	社会学
历史学	0.000	/	社会学	0.008	/
哲学	0.000	0.000	法学	0.000	0.368
维度三：铺陈与发展					
	中国文学	历史学		经济学	社会学
历史学	0.037	/	社会学	0.074	/
哲学	0.106	0.000	法学	0.000	0.016
维度四：已然性表述					
	中国文学	历史学		经济学	社会学
历史学	0.521	/	社会学	0.543	/
哲学	0.238	1.000	法学	0.000	0.039
维度五：计数与测量					
	中国文学	历史学		经济学	社会学
历史学	0.214	/	社会学	0.009	/
哲学	0.000	0.000	法学	0.000	0.000
维度六：模糊性表达					
	中国文学	历史学		经济学	社会学
历史学	0.282	/	社会学	0.310	/
哲学	0.000	0.000	法学	0.000	0.063
维度七：顺序与连接					
	中国文学	历史学		经济学	社会学
历史学	0.251	/	社会学	0.999	/
哲学	0.000	0.000	法学	0.993	0.984

(c) 2022 中国中文信息学会计算语言学专业委员会

基于语料的“一+形容词+量词+名词”构式语义考察

吴宁
汉语国际教育研究院/ 北京语言大学
1261066160@qq.com

王治敏
汉语国际教育研究院/ 北京语言大学
wangzm000@qq.com

摘要

“数形量名”构式是我们日常语言交流中大量使用的结构。本文在北京语言大学BCC在线语料库5710条语料的基础上考察“一形量名”结构，寻求影响构式成立与否的的关键性因素。本文研究了语义限制下进入构式形容词的语义特点、“物理抽象度”对构式名词成分的限制以及量词在构式形成过程中的作用。研究表明，具备高拆分计量性等语义特征的形容词更易进入此构式，进入构式形容词中90%以上项目都可由单一变化物理量进行衡量，此部分形容词在同一意义层面上与构式内的量词互相和谐；“一形量名”构式对“物理抽象度（[+易量化、+低有机活性、+形状易概括]）”赋值低的名词包容性更高；此外，本文还发现集合量词的出现可降低整体构式的物理抽象度，从而增强“一形量名”构式成立可能性。

关键词： 构式语义 ； 语义限制 ； 物理抽象度 ； 语料分析

A Semantic Study of “One-Adjective-Quantifier-Noun” Based on Corpus

Ning Wu
Research Institute Of International Chinese Language Education / Beijing Language and Culture University
1261066160@qq.com

Zhimin Wang
wangzm000@qq.com

Abstract

The “Numeral-Adjective-Quantifier-Noun” construction is a type of structure that is widely used in our daily language communication. Based on the analysis and observation of 5710 corpus in BCC corpus, this paper studies the semantic characteristics of constructional adjectives under semantic restrictions, the restriction of physical abstraction on constructional noun components, and the role of quantifiers in the formation of constructions by means of word frequency statistics and Pearson correlation analysis, by which to find the key factors that affect the establishment of the “One-Adjective-Quantifier-Noun” construction. The results of statistical analysis show that adjectives with a high degree of separation and measurement as their main semantic feature and nouns form a high degree of semantic harmony with them, that is, nouns with low physical abstraction are more likely to enter this construction. In addition, the appearance of collective quantifiers can reduce the physical abstraction of the overall construction, thereby enhancing the possibility of the establishment of the “One-Adjective-Classifier-Noun” construction.

Keywords: Constructional semantics , Semantic restriction , Physical abstraction , Corpus analysis

第二十一届中国计算语言学大会论文集，第70页-第78页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

无论是口头陈述还是书面表达，“数量名”结构广泛存在于我们的语言生活之中。例如“一块肉”“一张纸”等，而在“数量名”结构的使用过程之中，出于不同目的，有一部分数量结构可被拆开以插入单音节形容词，构成“数词+形容词+量词+名词”的形式（以下简称为“数形量名”结构）。像“一小块肉”、“一大张纸”等都是典型的“数形量名”结构。参见以下例句，它们表现为在“数量名”结构的基础上加入单音节形容词。但是这种操作又不仅仅是添加了成分把句子拉长这么简单，形容词的加入让句子表达的含义产生微妙的变化。在例（1）a的基础上插入形容词“小”，给句子增添了一层“微不足道”的含义。例（2）a插入“大”则显示了说话者的“周到”、“全面”。在整体语义上，“数形量名”结构可以认作是进行限定之后的“数量名”结构，在原始义的基础上，新插入的部分给与意义的进一步限定或细节的加深。

（1）a. 我拇指没了一块肉。→ b. 我拇指没了一小块肉。

（2）a. 我每次都是在上面垫一张纸。→ b. 我每次都是在上面垫一大张纸。

形容词插入量名结构带来新的特点，而且通过简单的同类词替换，我们发现并非所有的量词、名词都可以进入此结构，进入此结构的形容词也受到特定的语义限制。见例（3）-（4），我们同样在包含“数量名”结构的句子中插入形容词，但得到的句子并不符合语言规范。我们常常说“一个人”，但是“一大个人”在正常语境下下很难出现，原本稳固的数量名结构“一个人”在插入形容词“大”后，数词、形容词、量词、名词各成分无法同时和谐，难以成立，“一大头驴”、“一大管枪”和“五十大个法朗”听起来也十分怪异。

（3）a. 一个人在另一个岗位上却可能光彩照人。

b. *一大个人在另一个岗位上却可能光彩照人。

（4）a. 他提出的代价是一头驴、一管枪和五十个法朗。

b. *他提出的代价是一大头驴、一大管枪和五十大个法朗。

1982年，朱德熙(1982)先生在其《语法讲义》中提出，少数个体量词可以受到前置单音节形容词修饰，但此类形容词仅有“大”、“小”、“长”、“方”等有限的几个，构成短语如“一大张纸”、“一长条肥皂”、“一小块冰”、“一方块冰”等。但朱先生并未对此具体计量和形成原因展开具体探讨。对此，陆俭明(1987)先生对数量词中间插入形容词的情况进行了具体考察，发现只有129个数量词构成的数量短语间可以插入形容词，而此类形容词的数量则仅有7个，但陆先生同样未对考察到的现象作深入解释。对其成因，刘殊墨(2018)在其研究中认为，数量名结构中是否能插入形容词，受到该结构修饰名词的制约和影响，取决于其中数词、形容词、量词、名词之间组合的可能性，她提出，名词物体凸显的空间维度尺寸变化以及此名词成分的语义特征同样影响着此构式的形成。在以上文章及许多未提及的研究中，陆俭明先生对此结构中的量词、形容词进行了数据统计，其他大部分文章或采用旧文献中的数据，或采用经验推测。再看“数形量名”构式源结构的“数量名”短语结构，在句法语义、类型学等方面受到了学者们的深入研究。在句法语义层面，贺颖(2019)考察了“数量名”短语的句法结构同逻辑语义联系的方式，探究影响汉语“数量名”语义特征的因素。杨烈祥(2019)则对数量短语的内部结构进行解构，探讨了其句法生成和语义解读等问题。围绕“数量名”结构的讨论和研究十分丰富，许多学者采用各种方法对其进行了深入的探讨，而同样在汉语环境中出现的“数形量名”构式则并未受到太多关注，其从词语搭配、语义限制方面还有待于进一步探究，还缺少在大规模语料直接的证据。因此，本文借助北京语言大学BCC语料库，在进行人工分词排除错误后，抽取了5710条相关有效语料，重新对“数形量名”构式短语进行分词类计量统计，在前人研究的基础上，根据全新的统计结果，对此构式的进入限制条件进行分析推断。

2 “数形量名”结构分布特点分析

在“数形量名”结构中，对不同词类所包含的具体词语进行分析，对解决本文论题有关键性作用。在选择语料方面，本文所使用的例句均来源于BCC语料库，涵盖文学、报刊、对话等大量语料。BCC汉语语料库总字数约150亿字，包括：报纸（20亿）、文学（30亿）、微博（30亿）、科技（30亿）、综合（10亿）和古汉语（20亿）等多领域语料，是可以全面反映当

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家社科基金重大项目（18ZDA295）；国家语委科研项目（ZDI135-139）；中央高校基本科研业务费（19YCX057）；北京语言大学汉语国际教育研究院研究生创新基金（21YJY007）

今社会语言生活的大规模语料库。本文将基于BCC语料库抽取语料进行研究，获取包含“数形量名”构式的有效真实语料，基于真实语料进行分类统计，运用数据统计分析软件SPSS进行推理验证，从而探索该构式形成的规律。本文采用“m a q n（0<m<10, m∈N）”（数形量名）作为检索式，即从“一”到“九”分别带入，构成“一||二（两）||三||...||九a q n”检索式在BCC语料库中进行检索。检索所得语料数量如下表 1。其中用“一a q n”（一形量名）检索式抽取到5956条语料，占比达71.79%，能基本反映“数形量名”结构的整体整体规律，故本文主要选取“一+形容词+量词+名词”（以下简称“一形量名”结构）所对应语料进行分析研究。

检索式	一aqn	（二+两）aqn	三aqn	四aqn	五aqn
六aqn	七aqn	八aqn	九aqn		
结果数量（条）	5956	532	510	174	118
27	27	21	6		
占比	71.79%	6.41%	6.15%	2.10%	1.42%
0.33%	0.33%	0.25%	0.07%		

表 1: 语料检索表

对基于“一形量名”抽取的5956条语料进行观察发现，并非所有项目符合研究要求，大规模语料库带来便利的同时也带来数据易出错等相关问题，如例（5）中的句子。（5）a中的“一顺位继承人”或是（5）b中的“一大盘冲高”，在拆分开来看似符合要求，但从句法层面上看，这些句子显然不属于“数形量名”结构，是机器错误分词的结果，应该予以排除。

（5）a. 所以吴叔才成为第一顺位继承人。

b. 周一大盘冲高回落。

针对这一情况，笔者对5956条原始语料进一步人工数据清洗，对不符合要求的项目进行一一排除，最终得到有效语料5710条。其数量远低于“一量名”结构和“一形名”结构，三种结构数量差异情况如图 1所示：

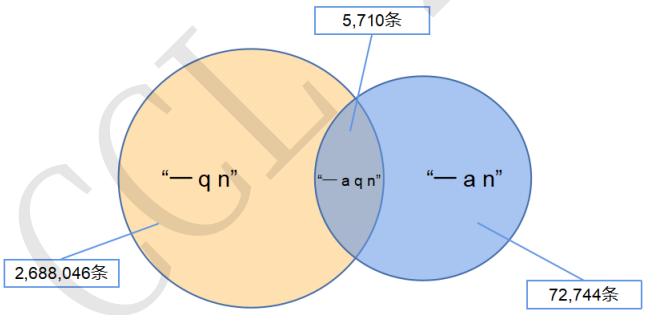


图 1: “一形量名”相关结构分布示意图

在观察阶段，不加分类直接对大量语料进行审视不仅耗时耗力，而且难以准确发现问题所在。本文以形容词、量词、名词为分类依据，对有效语料做进一步拆分归纳，并对相同的项目进行合并统计得形容词14个、量词90个、名词2047个。这里的14、90、2047能说明什么呢？为了更加直观得展示数据差异，我们利用BCC语料库分别对其包含的形容词、量词、名词进行了简单统计，结果显示：BCC语料库中共包含不同形容词3135个、量词685个、名词357766个。由此可以粗略得出上述三类词针对“一形量名”结构的进入率，分别为：形容词0.45%、量词13.14%、名词0.57% ，可以看到，进入本结构的量词占总量词比例相对较高，而形容词和名词占比非常低，换句话说，少数特定的形容词和名词在近六千条语料中频繁出现。看似“数形量名”结构对量词来得更加“包容”，那么是什么原因导致了这种结果呢，为什么偏偏是总量很小的量词更容易进入“一形量名”结构？而形容词、名词这两个开放词类却仅有不到1%的词可以进入到本结构当中。我们可以合理地认为，形容词、名词进入此结构受到一定

(c) 2022 中国中文信息学会计算语言学专业委员会

限制，而这个限制远远高于量词进入结构的限制。这种限制导致了大量形容词和名词被“一形量名”结构拒绝，因此我们难以在语料中找到相应的句子。简言之，进入“一形量名”结构的形容词、名词受限严重，量词同样受限但情况相对稍好，因此本文认为对语料中出现的词语进行分词类统计分析，研究高频词与低频词之间的联系与差异，有助于解释特定词语进入“数形量名”构式所遵循的语言规律。

3 形容词语义限制分析

“一形量名”结构中，并非所有的形容词都能进入此结构。其中特定词语的使用存在明显的倾向性，现存“一形量名”构式短语在用词上向某些词语倾斜。如“大”“小”等特定形容词，它们出现的十分频繁；而有些词，即使属于同一词类，也鲜少出现。例（6）a、（7）a、（8）a皆为BCC语料库中的句子，而（6）b、（7）b、（8）b则是我们进行单一同类词替换得到的句子。原本正确的句子在进行替换之后，皆变成了错误的句子。首先看例（6），两者的差异在于量词“束”前的形容词，（6）a中“一大束玫瑰”变成（6）b“一美束玫瑰”之后则难以在语义层面符合汉语的使用规范，成为错句。此类错误可以比较直观地进行考察，因为形容词“大”和“美”之间，无论是从语义看，或是观其修饰对象，都有显著的差异。参见苏新红(2013)的《现代汉语分类词典》，苏在性质与状态一节下将形容词分为形貌、直觉、性状、性质、才品、情状六类，“大”属于形貌类，“美”属于性状类，他们不属于同一分类。

（6）a. 同事在光棍节收到了一大束玫瑰花。

b. *同事在光棍节收到了一美束玫瑰花。

在例（6）中设置用单音节形容词“美”代替“大”可能在某些层面看有些牵强，因为我们可以单凭直觉就可以分辨出这两个形容词之间的区别，另外又可结合《现代汉语分类词典》得知“美”“大”两形容词分属于不同类别。但是在有些情况下，即使属于同一小类的形容词，它们之间也不能互换，像例（7）中的“厚”“长”都属于形貌类形容词，但彼此之间却不能随意互换。此时，“数形量名”结构的成立与否与形容词本身和量词的搭配有着重要联系。“一厚沓剪报”中“厚”的使用取决于其后量词“沓”的语义限制。参见“沓”字的概念：“沓”作为量词表示叠起来的纸张或其他薄的东西。“沓”在数量上有厚度的差别，故一般不能用“长”“短”修饰。

（7）a. 当时我们这些文字记者，带的是一支笔和一厚沓剪报。

b. *当时我们这些文字记者，带的是一支笔和一长沓剪报。

通过进一步考察，我们发现即使排除上述两个因素，即同时满足两个形容词不仅属于同一分类，而且两者指称的是同一类型的量，句子中的形容词也并非都可互换，如上文例（8）所示。形容词“大”是指面积、体积、容量、数量、强度、力量超过一般或超过所比较的对象，与“小”相对。“大”“小”这对形容词都是来修饰规模程度的词汇，常常搭配成对出现且存在对应关系，如“这件衣服大小怎么样？”“你想要大的还是小的？”但是像句（8）a中的“一小点利益”，看起来用“一大点”代替“一小点”，只是用描述同一类现象的相反对应词予以替换，从语法层面上看没有问题，但实际上句（8）b在语义层面上同样难以成立。

（8）a. 不要因某一小点利益限制了自己的自由。

b. *不要因某一大点利益限制了自己的自由。

而有时同样的“大”“小”两词又是可以互换的，像“一大块蛋糕、一小块蛋糕都是正确的说法，由此我们可以合理的认定此结构所暗含的某种语义限制，使得此结构中形容词的互换有一定的限制，并且有一部分形容词暂时或永远被排除在此结构之外。针对此现象，笔者对抽取的5710条语料进行分词统计，结果如图 2所示：

“一形量名”结构中形容词数量从高到底分别为“大、小、长、薄、厚、满、窄、短、全、扁、净、宽、实、瘦”，但窄、短、全、扁、净、宽、实、瘦的占比皆小于10%。观察数据可以发现，此结构中出现的高频形容词皆有某种共性：从词义角度出发，像“大、小、长”等词皆有较强的可计量性，在具体情形下都能进行二维定量描述，与不能出现在该结构中的“美、丑、对、错、好”等强描述判断性形容词存在显著差异。取决于形容词类本身所具备的特点，虽然“大小”类形容词和“美丑”类形容词所表达的含义都有一定相对的意味，但前者的计量性更强一些：我们可以设定某具体尺寸的电视机为中等，超过此尺寸的是“大”电视，小于设定尺寸的为“小”电视，但却很难用一个具体的度来判断一个人到底属于“美”还是“丑”之列。因此，本文认为后组所具备的拆分计量特征不如前组显著。参见陆俭明(2010)在借鉴语音和谐律之基础上提出的语义和谐律：句子要求其各组分的语义处于一种和谐的状态，这其中包含三种形式的和

(c) 2022 中国中文信息学会计算语言学专业委员会

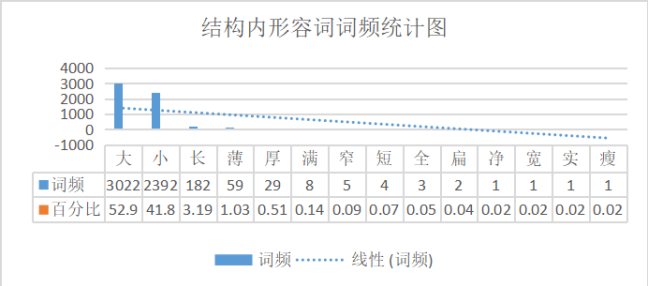


图 2: 形容词词频统计图

谐：整体构式义与其组成成分之间的语义、构式内部词语与词语之间的语义、构式内部所使用的词语和构式外部使用词语的语义。像“拔出来”“拔出去”等短语，由于构式内各词语之间的语义必须保持一种和谐的状态，故不允许像“拔进来”“拔进去”这种形式的出现。反观量词的定义：量词，顾名思义，即用来进行数量描述的词，是用来表示人、事物、动作之数量单位的词。在进入本结构的形容词中，除去仅占极小部分“瘦”等词，90%以上的词都可由单一变化的物理量进行测量或衡量，这部分形容词在同一意义层面上与构式内的量词互相和谐，而其他无法和量词的语义相和谐的形容词则不易或者是不能进入本构式。这也解释了为什么在14.4%常用量词可进入的情况下，在大量形容词中却仅有14个单音节形容词进入了“一形量名”构式。通过分析，我们发现，进入构式的形容词具有计量性特征，而比如前文提到的形容词“大”和“小”都具有较强的计量特性，但它们不能任意互换，其从根本上受到构式内其他成分的制约。再比如：“一小点利益”是由“一点利益”衍生出来的一种形式，这里的“小”不能换成“大”在一定程度上受其后量词“点”的语义制约。“点”的语义侧重于表示“少量”，例如液体的小滴“雨点儿”以及小的痕迹“斑点”。进入构式的形容词“小”与量词“点”互相和谐，并且进一步加深“少量”的概念，而“大”却不可以达到相应的效果。

4 名词语义限制分析

在大量名词之中，几乎所有的名词都能前加数量词构成数量短语，像“一张纸”“一头驴”“一个人”等等，但是并非所有此类结构都能在数词后添加形容词，对上述例子加以更改，分别在数词后添加形容词“大”，只能得到“一大张纸”一个完全合乎规范的短语，而对于另外两个例子，则无法找到可以前加得形容词。在第三节中，本文探讨了与量词紧密结合的形容词，其蕴含的语义特点直接影响形容词能否进入本结构，而与量词结合更加紧密的名词本身是否也决定了构式的成立与否呢？本文合并提取出进入“一形量名”构式的所有名词共2046个项目，下图 3为统计结果。

图 3: 名词词频

从短语层面看，5710条语料合并出3097项“一形量名”短语。对其进行词频统计、词义标记等进一步研究。从名词的语义可切分性，我们将这3097个项目分为A、B、C三类，如下表 ??所示。为更清晰的描写这些名词之间的关系，我们不妨按照其语义进行标记：本体是否为单个物体（记作[±singular]）、本体降格后是否还与本体共享同一性质（记作[±identity]）。

A类中的短语都具备可简单切分的性质。所谓可简单切分，即切分物和本体共享同一性质，无论怎么切，物体的本质是没有变化的。可表示为[+singular、+identity]，如例（9），“一大张纸”可以切分出某部分形成“一小张纸”，此处的切分并没有改变前者中“纸”的

(c) 2022 中国中文信息学会计算语言学专业委员会

类别	A类	B类	C类
名称	简单切分	独立型切分	不可切分
原始数量（条）	3683	1470	557
合并后数量（条）	1833	860	404
占比	64.50%	25.74%	9.76%
典型用例	一大张纸　一大块牛肉 一小块土地　一小块蛋糕 一大截路　一小块布 一小块玻璃　一小块石头 一小块金属　一小块黄油	一大束玫瑰　一大串钥匙 一大摞书　一大串名字 一大捆蜡烛　一厚沓报纸 一大串邮票　一大捆百合 一小堆黄豆　一小堆元件	一大间屋子　一大个苹果 一大幅油画　一大条鱼 一小条伤口　一大间办公室 一大棵树　一大扇玻璃窗 一大栋楼　一大朵荷花

表 2: 组词策略下的思维链

性质，“一小张纸”可以是前者“一大张纸”直接取出来的一部分；“一大块牛肉”再怎么切分也始终是“牛肉”。此类包含项目基本属于分类体系中的下级单位（subordinate unit）。

（9）a. 他从衣袋里拿出一大张纸来，上面尽是些日期和姓名。

b. 他正在切一大块牛肉下锅。

c. 结婚之后，岳家送给他一小块土地。

而B类中的短语虽然可以切分，但是切分后生成物却是由各自独立的个体组成的部分，符合[-singular、+identity]，如例（10）所示。“一大串玫瑰”本身就是由一枝枝玫瑰所构成的集合，从“一大束玫瑰”到“一小束玫瑰”，切分只带来了数量的差异，而并非对单一的玫瑰继续切割得到玫瑰花瓣、花茎、叶子等部分。B类短语的分割可以不是人为造成的，在多数情况下，个体之间是联系不紧密的关系。切分前后是全集（U）和子集（subset）或集体和个体之间的关系。

（10）a. 她弯腰，抱了一大束玫瑰塞进他手里。

b. 她穿着黑颜色的衣服，在她腰际还挂着一大串钥匙。

c. 因此他每次上课，走进教室里时总要夹着一大摞书。

而C类短语中的个体则不可以进行简单切割，这些短语更倾向于描述事物发展变化带来的结果或状态，可表示为[+singular、-identity]。例（11）中为C类短语部分用例，“一小栋楼”并不是简单切分“一大栋楼”得到的部分，“一小栋楼”和“一大栋楼”虽然都是“楼”，但前者并非是后者的一部分，两者不符合我们此处所指的“同一性质”。同样的还有“一大个苹果”“一小个苹果”等，它们的出现是事物自然发展成长而产生的不同的个体，此时所加入的形容词，如“大”和“小”，是一个相对说法，它们参照同类事物的平均水平（average value）。

（11）a. 省下的地皮又能盖一大栋楼。

b. 要是我每天只吃一大个苹果一定会瘦的。

c. 我要留一大间屋子放我的七七八八。

对统计数据进行进一步观察，我们发现“一形量名”构式所涉及名词出现集群分化现象，出现次数高的词汇比出现次数低词汇具备一些有趣的特点。例如，名词“路”和“距离”分别出现了145、112次，居于总排序的前两位，而像“人”“女人”这种词出现的数量则少得多，分别出现了27、2次，从词义角度看，高频名词大部分带有易计量、可拆分等含义。参见刘殊墨(2018)在其文中的观点，她认为形容词、数量结构和其修饰的名词成分所凸显的空间维度表现是否一致，是三者能否形成相互选择关系的关键，另外，名词物体所凸显的空间维度尺寸变化以及名词成分语义特征，也是影响到形容词能否插入数量结构的两个因素。罗旋(2007)则在其研究中表示，进入“数+形+量”结构中的量词集中包含[+数量评估性][+性状性]的语义特征。本文发现，物理抽象度越高的名词，其进入“一形量名”构式的可能性越低。所谓物理抽象度，即名词所指代事物的具体现实程度。物理抽象度高的对象具备更高的整体性、生命有机活性更高、形状不易概括，以及较低的拆分计量性；而物理抽象度较低的名词项目，则更加容易用科学直观手段进行定量统计，它们的性状更加方正清晰，因此容易拆分单说。像“人”这个概念，人是具备高有机活性的独立整体，一般难以用单一物理量进行概括或拆分计量，其符合高物理抽象度的典型概念；而像“距离”这种概念则恰恰相反，切分一段距离还可以称之为距离，并可以用完全不变的单一物理量进行测量，其具备低物理抽象度；“牛”和“牛肉”看似一样，其实不然，牛这一生命在经过屠宰成为牛肉之后，其生命活性大大降低。而牛肉要么成块，要么成团，其具体性状也变得更易描述，在进行拆分之后还可以称之为牛肉，而将牛拆分，则没有“一块牛”之类的用法，因此，我们说“牛肉”的物理抽象度要比“牛”低，其更容易进入“一形量名”构式。同样的还有“羊”“猪”等有生物体本身，它们的物理抽象度都要高于“羊肉”“猪肉”等动物产品，因为

(c) 2022 中国中文信息学会计算语言学专业委员会

它们都属于具有生命的完整个体，难以在不改变其性状的情况下简单拆分，形状也难以捉摸。动物自身和其经过屠宰形成的副产品在根本上是上下位概念之间的关系。物理抽象度及其细化计算不仅仅能帮助我们更好得把握进入“数形量名”结构词语的共同特点，而且也有助于汉语的教学与学习，通过参考名词物理抽象度对照表，教师和学生可以更加准确地判断一个“数形量名”结构是否成立，而且能更清晰地对抽象地语言现象进行解释，有利于学生（尤其是将汉语作为第二语言的学习者）更直观地认识及更准确地运用这一语言结构。为更加直观了解名词进入“数形量名”构式的能力与名词物理抽象度之间的联系，本文将物理抽象度拆分为以下四个要素进行定量计算，以检验两者相关性。这四个要素分别为整体性、有机活性、不规则形状、非拆分计量。我们认为物理抽象度高的词更倾向于拥有上述特征，因而不容易进入“数形量名”构式。将上述几个要素再分别分为三个层面测算名词对应强度等级：“非常符合”记为“**”，“基本符合”记为“*”，不符合计为“ ”，单个“*”为5。举例来说，我们描述“路”这个名词为：[+整体性]不符合；[+有机活性]不符合；[+不规则形状]基本符合；[+非拆分计量]不符合，物理抽象度等级为“*”，物理抽象度为5。图 4为结构中部分名词的物理抽象度。

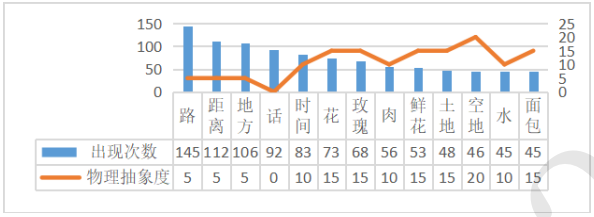


图 4: 物理抽象度

为证明物理抽象度确实与名词的构式进入能力相关，本文将物理抽象度区分为具体四个语义特征：[+整体][+有机活性][+低形状概括度][+非拆分计量]，并且从每一个方面对名词词频前50个项目进行标记测算，得出其物理抽象度。对物理抽象度及构式内词频进行Pearson相关性分析结果如下图。r=-0.591 ∈ [-1,-0.5]，认为两变量，即构式出现名词的物理抽象度与词频之间存在强负相关，证明名词的物理抽象度越高越不容易出现在“一形量名”构式。

		出现次数	物理抽象度
出现次数	皮尔逊相关性	1	-.591**
	显著性（双尾）		.000
	个案数	50	50
物理抽象度	皮尔逊相关性	-.591**	1
	显著性（双尾）	.000	
	个案数	50	50

**. 在0.01 级别（双尾），相关性显著。

表 3: 物理抽象度与构式名词词频皮尔逊相关性

5 量词抽象度考察

在前文讨论中，我们发现量词自身的属性在“一形量名”结构能否成立上发挥了重要作用，因此量词也成为形容词、量词、名词中受限最小的词类。另外，在对“一形量名”相关语料进行考察的过程中，我们还发现，量词中集合量词的出现可以显著降低名词物理抽象度，增强构式成立可能性。集合量词的语义特点能改变相关名词的物理抽象程度，从而对句子的构成产生促进作用。通过将个体量词替换为集合量词，原本不规范的句子便得以成为规范的句子，例如：

（11）a. *在县长周围聚集了一大个人。→ b. 在县长周围聚集了一大群人。

（12）a. *他提出的代价是一大头驴。→ b. 他提出的代价是一大群驴。

集合量词是相对个体量词来说的，属于名量词的一个子类。“个、只、句、本、点”等个体量词常表现量的单一性，如“一个苹果”、“一头牛”、“一匹马”等表现所指事物在整体数量上为

(c) 2022 中国中文信息学会计算语言学专业委员会

一。而集合量词本身就包含数量，用来描述成双或是成群的概念，常见的有“班、帮、堆、对、群”等。以上短语，用集合量词进行替换得“一堆苹果”、“一群牛”、“一群马”，在数词不变的情况下，都体现了数量的增加。赵元任(1979)划分出46个集合名词：“对、双、打、十、百、千、万、亿、行、身、槽、列、系列、排、副、套、堂、沓儿、串、挂、帮、房、批、组、窝、捆、群、胎、桌、进、部、种、类、样、派、流、路、号儿、师、旅、团、营、连、班、排、队”。黄伯荣、廖序东(2017)在《现代汉语》中划分集合量词为“对、双、副、班、堆、批、群、套、打”。本文另外采用赵元任和黄伯荣、廖序东(2017)对量词的分类，合并剔除相同的成分，共47项，对抽取语料进行分类。在上节名词词频与物理抽象度相关度分析中，本文剔除了集合量词所对应的名词项目，因为我们发现某些名词项目在纯个体名词数列中和整体（集合量词+个体量词）数列中的出现几率并不相同。例如，对于“一+形+量+人”这个结构，针对同一个名词“人”，添加集合量词后，结构出现率将大大增加。这就意味着，集合量词的使用会让本结构更易形成。为了更直观了解，我们不妨设置基于个体量词对应名词排序表的量z1，以及基于总名词（个体量词+集合量词）排序表的z2，两者计算的皆为某特定名词在总词频中所处位置，其计算公式及部分z值数据如下表 4所示。

z值/名词	路	话	时间	花	玫瑰
肉	鲜花	土地	面包	白色	人
...					
z1	0.02%	7.42%	9.30%	10.99%	12.48%
13.87%	15.01%	16.09%	18.01%	21.50%	25.27%
...					
z2	2.64%	0.02%	9.00%	10.45%	11.77%
14.06%	0.04%	17.00%	20.28%	0.05%	13.01%
...					

表 4: z值对应表

(注：z1(n)=个体量词n的位次/个体量词“一形量名”结构总个数; z2(n)=（个体+集合）量词n的位次/（个体+集合）量词“一形量名”结构总个数) z1（人）=25.27%，z2（人）=13.01%，z1（话）=7.42%，z2（话）=0.02% “人”在个体量词相关名词词频占比为前25.27%,而在加入集合量词相关名词项目后，位置则提前到了前13.01%，名词“话”在集合名词的影响之下，由前7.42%一举提升为前0.02%，跃居首位。我们至少可以认为，集合量词的出现能够降低某部分名词的物体抽象度，使其更和谐的进入到“数形量名”构式之中。集合量词增强了整个构式的数量性，因此使原本不能实现的句法构式得以实现。

6 结语

在“一形量名”结构中，形容词、名词成分的进入都受到语义的限制。本文在对BCC语料库抽取的几千条相关语料进行数据分析发现，可进入构式的形容词，都具备明显的“量化”特征，与量词语义成高度和谐关系。基于此，本文以进入构式名词词频以及其物理抽象度为对比数据，进行Pearson相关性分析，结果显示，随名词物理抽象度数值的升高，其进入“一形量名”的能力呈降低趋势，易进入结构的名词大部分具备“易量化”、“低有机活性”、“形状易概括”等特征。另外，集合量词对于研究构式的成立也有重要影响，集合量词本身就具备的数量性特征与构式本身具有的高描述性特征语义和谐，都降低了所在结构名词成分的物理抽象度，增强其进入构式的能力。下一阶段的研究将针对物理抽象度的特征进行更深层次的研究，并将从历时角度对本结构进行整体考量，服务于语言研究和汉语教学。

参考文献

刘殊墨. 2018. 名词对形容词插入数量名结构的影响. 新疆大学学报*(*哲学·人文社会科学版*)*, 46(06):134–

(c) 2022 中国中文信息学会计算语言学专业委员会

139.

朱德熙. 1982. 语法讲义. 商务印书馆.

杨烈祥. 2019. 量词短语的生成类型学研究. 硕士.

罗旋. 2007. “数+形+量”结构的构造. 硕士.

苏新红. 2013. 现代汉语分类词典. 商务印书馆.

贺颖. 2019. 汉语数量名短语的句法语义. 硕士.

赵元任. 1979. 汉语口语语法. 商务印书馆.

陆俭明. 1987. 数量词中间插入形容词情况考察. 语言教学与研究, 04:53–72.

陆俭明. 2010. 修辞的基础——语义和谐律. 当代修辞学, 01:13–20.

黄伯荣and 廖序东. 2017. 现代汉语. 高等教育出版社.

CCL 2022

基于熵的二语语音习得评价研究 —以日本学习者习得汉语声母为例

冯晓莉[1]，高迎明[1]，林炳怀[2]，张劲松[1]*
[1]北京语言大学 信息科学学院，北京市100083
[2]腾讯科技有限公司智能平台产品部
fengxiaoli314@163.com, gaoyingming1@sina.com,
binghuailin@tencent.com, jinsong.zhang@blcu.edu.cn

摘要

本文引入“熵”对学习者二语音素发音错误的分布情况进行了量化研究。通过对不同音素及不同二语水平学习者音素错误率和错误分散度的分析发现：1.错误率与错误分散度有较高的相关性，二者的差异反映出错误分布的差异性；2.错误率类似的音素中，与母语音素相似度越高的音素错误分散度越小；3.较初级水平，中级水平学习者音素错误率下降而错误分散度上升。由此可见，熵可以在错误率基础上可以进一步揭示学习者母语音系及二语水平对音素发音错误分散度的影响。

关键词： 二语语音习得；发音错误分散度；熵

An Entropy-based Evaluation of L2 Speech Acquisition: The Preliminary Report on Chinese Initials Produced by Japanese Learners

Xiaoli Feng[1], Yingming Gao[1], Binghuai Lin[2], Jinson Zhang[1]*
[1]School of Information Sciences, Beijing Language and Culture University, Beijing, 100083
[2]Smart Platform Product Department, Tencent Technology Co., Ltd, China
fengxiaoli314@163.com, gaoyingming1@sina.com,
binghuailin@tencent.com, jinsong.zhang@blcu.edu.cn

Abstract

This study introduced “entropy” to quantify the dispersion of second language (L2) learners’ phonetic errors. By comparing the error rates and error dispersion of different phones and different proficiency levels of learners（elementary level (EL), intermediate level (IL) and advanced level (AL)), we found that: (1) There is a high correlation between error rate and error dispersion, and the difference between them reflects the difference of error distribution; (2) The greater the difference between target and native phones, the higher the pronunciation error rate and the greater the error dispersion; (3) Compared with the EL speakers, the IL speakers’ phone error rate decreased while their error dispersion increased. In summary, entropy can reflect the differences of L2 pronunciation distribution resulting from the influence of learners’ native language and L2 proficiency.

Keywords: L2 speech acquisit ion , pronunciation error dispersion , entropy

1 引言

©2022 中国计算语言学大会

在第二语言语音教学研究中，评价指标是评估二语学习者语音习得效果不可或缺的工具。现有的评价指标主要从发音正确与否、发音质量好坏两个方面对学习者的发音进行评估。针对发音正确与否，研究者主要通过判断学习者产出的语音是否正确(Chen et al., 2016; Jia et al., 2006)及目标语音错成了什么(Jia et al., 2006; Jouvet et al., 2015)，以确定目标语音整体习得水平、语音错误类型及不同语音范畴间可能存在的混淆情况等。对发音质量好坏的评估则主要以知觉打分（如使用Likert量表）、声学测量等方式对学习者产出语音的自然度(Tsurutani and Luo, 2013)、口音度(Jesney, 2004)、类母语度(Sun and van Heuven, 2007)、可懂度(Crowther et al., 2015)、声学参数分散度(Xie and Jaeger, 2020; Smith et al., 2019)及不同语音范畴在声学上的交叠程度(Wang et al., 2006; Bohn and Flege, 2011)等进行评测，以判断学习者的发音偏离标准音的程度或距离。如果说发音质量好坏的评估是对学习者的发音偏离标准音“量”的测量，发音正确与否即是对学习者发音是否出现“质”的错误的评判。

针对学习者发音的正误，目前的研究主要采用正确率(Jia et al., 2006)、错误率(Chen et al., 2016)、混淆矩阵(Jia et al., 2006; Jouvet et al., 2015)等指标进行量化评估。如Chen et al. (2016)使用大规模中介语语音语料库考察了欧洲多母语背景学习者汉语声、韵、调的习得情况，结果显示学习者音段错误率为5%，声调错误率为32%。Flege et al. (1997)对学习者产出的英语元音[i]、[ɪ]、[ɛ]、[æ]进行了评估，研究发现不同母语背景学习者产出的元音的正确率存在较大差异，如德国学习者产出的元音[i]的正确率可以达到100%，而西班牙学习者只有57%。此外，混淆矩阵作为一种多维向量，同时呈现了目标语音的正确率及各种错误类型的比率分布，如Jia et al.(2006)以混淆矩阵的形式呈现了三组不同二语水平的中国学习者对英语单元音[i]、[ɪ]、[eɪ]、[ɛ]、[æ]、[ɑ]、[ʌ]、[u]的习得情况，结果发现与学习者母语音素发音相同的[i]、[u]的正确率最高，而学习者母语中不存在的[ɛ]、[æ]和[ɑ]、[ʌ]两组对立音位存在较大程度的互相混淆，正确率较低。在该类研究中，正误率（正确率和错误率）指标可以对不同语音单元的习得程度进行对比，混淆矩阵在正误率指标的基础上可以进一步揭示目标语音单元的发音错误分布及不同语音单元间互相混淆的程度和方向性。

然而，正误率指标并不能精准反映二语学习者语音习得状况的所有面貌。例如，Jouvet (2015)在考察法国学习者产出德语元音时发现，虽然[e:]和[a:]两个音素的正确率相同，都是83%，但是[e:]主要有[i:]、[ɛ:]、[ɪ]三种错误类型，对应百分占比分别为4%、2%和4%，而[a:]只有与母语音素发音相似的[a]一种主要错误形式，对应百分占比为16%。又有，Jia (2006)在研究中国学习者产出英语单元音[ʌ]时发现，随着学习者在目的语环境生活时间的延长，学习者在音素[ʌ]上的正确率始终在50%左右，其错误形式却由多种逐渐集中到[ɑ]一种主要错误上。因此，具有相似正误率的不同语音单元或者不同学习阶段的同一语音单元，其发音错误分布可能存在较大的差异，而这种差异往往可以反映出学习者产出的目标语音是否存在确定的错误类型。本文综合考虑学习者产出的目标语音的错误类型的数量及百分占比，提出发音错误分散度：发音错误分散度越小，目标语音实际产出的错误类型越少、百分占比越集中。发音错误分散度可以揭示出相似正误率的语音在发音错误对象上的确定性，从而为二语研究者及二语教学者在发音偏误研究及二语语音教学策略提供一定的参照。显然，正误率指标难以对学习者的发音错误分散度进行量化分析，混淆矩阵尽管可以详细呈现学习者在目标语音上发音错误的分布情况，却无法对发音错误分散的分散情况进行总体量化。

针对上述问题，本文引入信息论中的“熵”来对二语语音习得效果中的发音错误分散度进行量化研究。首先，我们将每个目标语音单元产出作为一个随机变量，每一种实际产出形式作为变量的取值，其概率倒数的对数对应着该取值的不确定性，所有可能取值不确定性的加权和反映该目标语音单元产出的平均不确定性，即“熵”。熵越小说明语音单元产出对象越集中（熵为0时即表示变量只存在一种产出形式），熵越大说明语音单元产出形式越分散。

为了验证熵在反映二语语音发音偏误分散度中的作用，本文使用日本学习者产出的汉语普通话音节首辅音（声母）为实验语料，从不同音素间的横向对比及不同二语水平学习者（EL、IL、AL）音素发音错误的纵向发展两方面开展了实验研究。下文研究方法部分对本研究使用的实验语料及评价指标进行了介绍；实验结果部分分别对不同音素及不同二语水平学习

根据《Creative Commons Attribution 4.0 International License》许可出版
*通讯作者：张劲松(jinsong.zhang@blcu.edu.cn)
致谢：本工作得到中央高校基本科研业务专项资金（20YJ040002）、北京语言大学梧桐创新平台(19PT04)、语言资源高精尖中心项目“面向智能语音教学的汉语中介语语音多模态语料库研究”(KYR17005)、教育部规划基金项目（18XJJA740001）及研究生创新基金项目（21YCX178）的资助。

者的音素错误率和发音错误分散度进行了对比分析；讨论和结论的部分分别对实验结果进行了讨论并得出本研究的结论。

2 研究方法

2.1 实验语料

2.1.1 发音人

本研究采用了BLCU-SAIT汉语中介语语音语料库中日本学习者的发音数据作为实验语料。发音人为55位日本学习者（男性17，女性38），年龄19-35周岁（平均年龄22.8周岁，标准差：4.2）。所有发音人均在日本出生长大，开始学习汉语的年龄均在18岁之后。根据发音人的HSK（汉语水平考试）等级，55位发音人被分为初、中、高三个等级，其中初级水平（HSK3级及以下）13人；中级水平（HSK4-5级）22人；高级水平（HSK6级）20人。

2.1.2 音素分布

每位发音人需完成BLCU-SAIT汉语中介语语音语料库中283个双音节词和103个句子的录音，双音节词和句子的设计均考虑了汉语拼音方案声韵组合的合法性及多样性，具体录音文本设计可参照王玮(2019; 2020)。本研究使用语音数据20921条，其中包含汉语声母93280个，录音文本中汉语声母频次分布如图1所示。

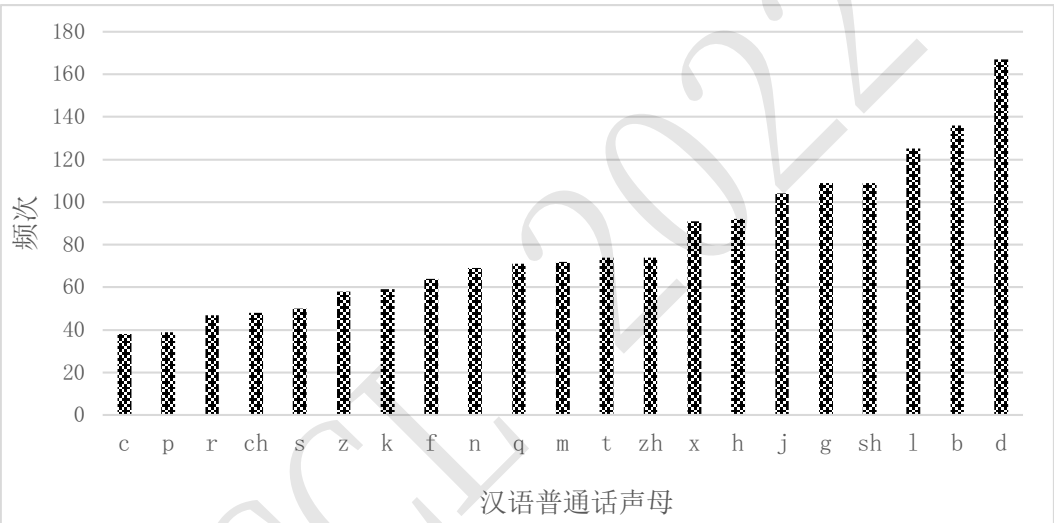


Figure 1: 录音文本声母频次分布

2.1.3 数据标注

本研究所使用的全部数据都进行了声韵母发音错误的人工标注。标注规范的设定参照了曹文、张劲松(2009)语料库标注方案中对音段发音偏误趋势标注的理念，标注内容不仅涵盖了传统语料库标注中替换式、删除式、插入式等错误种类，还对二语学习者中介语发音中“似A似B”的发音错误趋势进行了标注。标注符号为汉语拼音符号和附加符号（使用汉语拼音不足以描述清楚的发音错误，使用附加符号在中进行标记）。具体标注方案见王玮(2020)。

标注员为经过系统培训的语言学专业本科生和研究生，均来自中国北方方言区，普通话达到二级甲等及以上水平。所有标注员都可以对汉语普通话声母、韵母进行清晰的识读、辨别和区分，具备本标注任务所需语音学基本知识，如使用严式音标记音、正确描述语音发音属性等，并且可以熟练使用语音标注软件Praat(Boersma and Weenink, 2021)。标注过程由项目负责人将待标注数据进行等量分包然后随机分配给标注员，待标注完成后由三位有经验的质检员按照30%的比例对标注数据进行抽检，抽检正确率达到90%及以上方为合格（标注合格的标准为能够准确找出学习者的发音错误并使用正确的符号进行标注，漏标、错标均为标注不合格），否则需要将全部数据重新进行分配，进行再次标注，直至标注合格为止。

(c) 2022 中国中文信息学会计算语言学专业委员会

2.2 评价指标

2.2.1 错误率

在以往的研究中，错误率是衡量二语学习者的语音习得水平必不可少的指标，发音错误率的计算方式如公式(1)所示。

$$p = \frac{N_{err}}{N_{sum}} \times 100\% \tag{1}$$

其中，N_{sum}表示该目标语音样本的数量，N_{err}表示目标语音出现发音偏误的样本的数量，p表示目标语音的发音错误率.

2.2.2 发音偏误分散度

二语学习者在产出目标语音时实际发音可能包括正确发音和多种错误发音类型，并且每种发音类型所占百分比具有不确定性。为了对学习者的这种发音分散情况进行量化，本研究提出使用信息论中熵的计算方法对二语学习者发音分散度进行量化，量化方式如公式(2)所示。

$$H(x) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{2}$$

将每个目标语音单元的产出x作为一个随机变量，其实际产出形式P_i存在n种取值，每种取值对应的概率为p_i,其概率倒数的对数$-log_2 p_i$对应着该取值的不确定性。所有可能取值不确定性的概率加权和反映该目标语音单元产出的平均不确定性，即“熵”。熵越小说明语音单元产出对象越集中（熵为0时即表示变量只存在一种产出形式），熵越大说明语音单元产出形式越分散。

基于上述“熵”的计算方法，当对目标语音单元x的实际发音进行正误二元判断时，i仅存在两种取值，即正确和错误，对应的概率分别为p_{err}和p_{corr}，此时目标语音的产出形式的分布情况对应熵的基准值$H(x)_{base}$，量化方式如公式(3)所示。

$$H(x)_{base} = -\left(p_{err} log_2 p_{err} + p_{corr} log_2 p_{corr}\right) \tag{3}$$

熵$H(x)$与熵的基准值$H(x)_{base}$的差即为由发音错误的分散性造成的熵的增加，即由错误类型的增加或不同错误类型的百分占比均衡化导致的发音不确定性的增加。因此，通过计算$H(x)$与$H(x)_{base}$的差值我们可以得到发音错误分散度$\Delta H(x)$，计算公式如(4)所示。

$$\Delta H(x) = H(x) - H(x)_{base} \tag{4}$$

3 实验结果

本研究结合音素错误率和发音错误分散度两个指标，使用日本学习者产出的汉语中介语语音语料，对学习者习得汉语声母的发音错误情况进行了研究。首先，本研究所使用的55名日本学习者汉语中介语语音语料中声母的整体错误率约为10%。通过上述熵的量化公式，我们首先将21个汉语声母作为21个随机变量，每个声母在全部声母中的百分比为其对应的概率值为p_i，使用公式(2)可以得到正则语音的整体熵值为4.28。以同样的方式，我们将55名日本学习者在21个声母的每一个实际产出对象作为一个随机变量，每一个随机变量占全部语音产出形式的百分比为其对应的概率值p_i，通过(2)可以计算出学习者整体发音熵$H(x)$为4.88。以全部声母的错误率和正确率为随机变量，使用公式(3)可以计算出学习者整体发音基准熵$H(x)_{base}$为4.69，整体发音熵$H(x)$与整体发音基准熵$H(x)_{base}$的差值即为整体发音错误分散度$\Delta H(x)$，取值为0.19。

以下将使用上述量化方法分别对日本学习者21个汉语声母的错误率和发音错误分散度以及初、中、高三个水平发音人的平均音素发音错误率和发音错误分散度进行解析。

3.1 日本学习者汉语声母错误率和发音错误分散度对比研究

图2展示了日本学习者汉语21个声母的错误率和发音错误分散度的分布情况，柱状图对应声母发音错误率，曲线图对应各个声母的发音错误分散度。

(c) 2022 中国中文信息学会计算语言学专业委员会

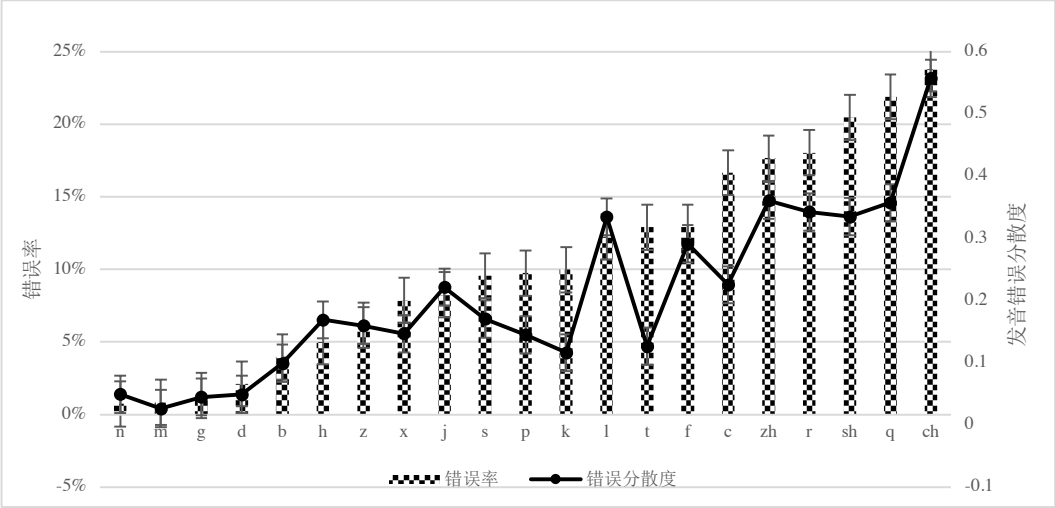


Figure 2: 日本学习者不同汉语声母发音错误率及发音错误分散度

根据图2首先可以看出，音素错误率和发音错误分散度整体分布趋势比较一致，即发音错误率较低的音素对应的音素发音错误分散度也较低，反之亦然。相关性检验结果显示，声母发音错误率与发音错误分散度的相关系数为0.91。同时，根据上述结果可以发现，与学习者母语发音相似度较高的m[m]、n[n]、b[p]、d[t]、g[k]的错误率和发音错误分散度最低，与学习者母语音素差异较大的zh[tʂ]、ch[tʂʰ]、sh[ʂ]、r[ʐ]等声母的发音错误率和发音分散度最高。

其次，对比不送气声母b[p]、d[t]、g[k]、j[tɕ]、z[ts]、zh[tʂ]与其对立的送气声母p[pʰ]、t[tʰ]、k[kʰ]、q[tʰ]、c[tsʰ]、ch[tʂʰ]可以发现，日本学习者产出不送气声母的错误率和发音错误分散度都明显小于与之对立的送气声母，这可能是由于日语的辅音仅存在清浊的对立而不存在送气与不送气对立(朱春跃, 2001)，日本学习者对送气这一声学线索不够敏感(Holt and Lotto, 2006)。

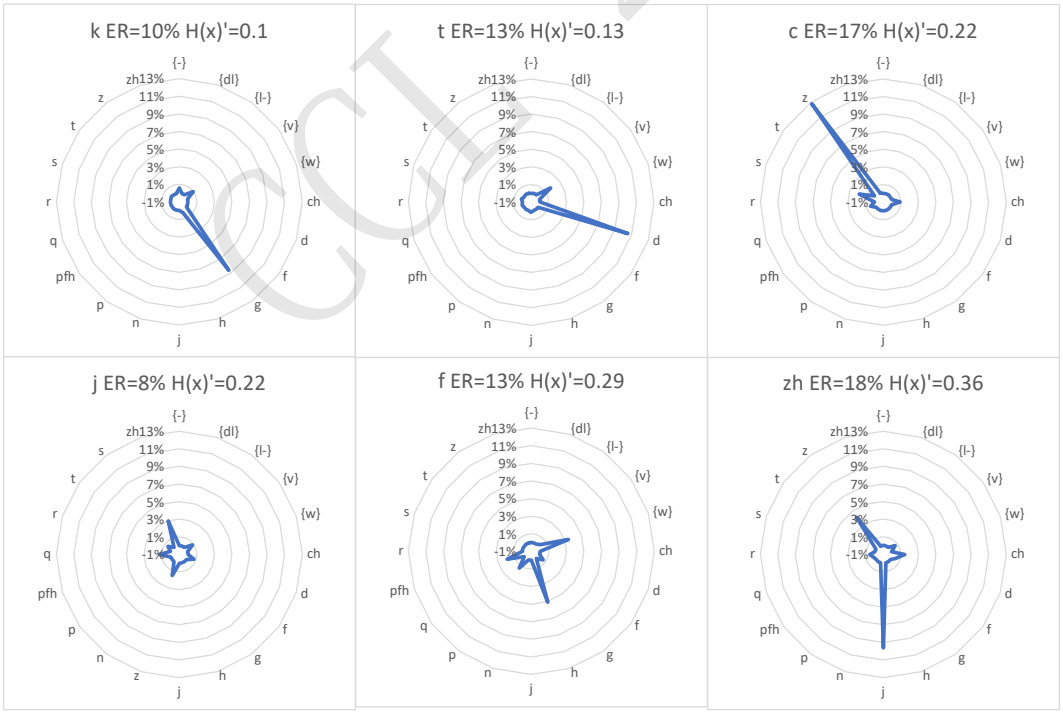


Figure 3: 错误率类似而错误分散度差异较大的音素错误分布情况示例

第三，学习者发音错误率和错误分散度的差异性主要通过发音错误率相似而错误分

(c) 2022 中国中文信息学会计算语言学专业委员会

散度差异较大的声母得以体现，如k[k^h]和j[tɕ],t[t^h]和f[f],c[ts^h]和zh[tʂ]等声母的错误率类似而错误分散度上却存在较大的差异。雷达图3详细成仙了上述音素的发音错误对象，可以看出k[k^h]、t[t^h]、c[ts^h]的错误对象主要是与之对立的不送气辅音g[k]、d[t]、z[ts]，发音错误分布非常集中；而与g[k]、d[t]、z[ts]在错误率上接近的j[tɕ]、f[f]、zh[tʂ]音素的发音错误类型明显更多也更分散，结合图2可以看出j[tɕ]、f[f]、zh[tʂ]的发音错误率明显高于g[k]、d[t]、z[ts]，由此可以看出，错误率类似的音素对应的发音错误的分布情况可以使用熵进行有效的量化。同时，由本研究中日本学习者产出汉语普通话声母的发音错误情况可以看出，当目标语音为学习者母语发音类似音素的送气对立音时，目标音素的发音错误分散度低于与其错误率接近的其它音素，并且错误类型主要为与目标语音对立的不送气音。

综上所述，本章节主要得到如下结果，第一，二语音素与学习者母语音素越接近时，学习者的发音错误率越低，发音错误分散度整体也越小。第二，日本学习者产出不送气声母的错误率和发音错误分散度都明显小于与之对立的送气声母。第三，当目标音为与学习者母语类似音素的送气对立音时，目标音素的发音错误分散度明显低于与其错误率接近的其它音素。由此可以得出，使用熵对发音错误分散度进行量化，可以在音素错误率的基础上进一步揭示出目标音素发音错误对象的分布情况：发音错误分散度越小，目标音素的错误类型越少、主要错误类型的百分占比越集中。

3.2 初、中、高水平日本学习者汉语声母错误率和发音错误分散度对比研究

本章节从发展的角度出发，分别对初、中、高三组不同二语水平的日本学习者产出汉语声母的错误率和发音错误分散度进行了考察。以下图表和公式中分别将初级、中级、高级三个等级简写为EL、IL、AL。首先根据分组计算每个音素的错误率和发音错误分散度，通过取平均得到每组音素的平均错误率和平均错误分散度，如图4所示。

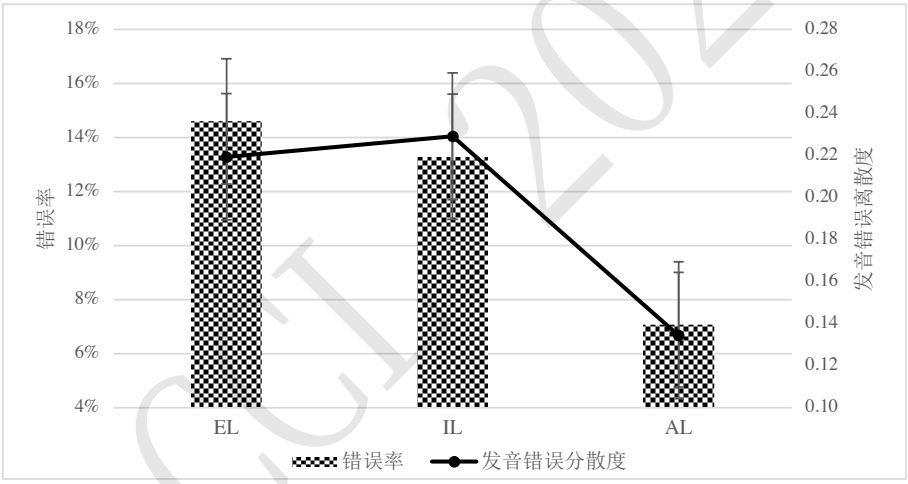


Figure 4: 初、中、高水平日本学习者平均声母错误率及评价发音错误分散度

根据图4可以看出，从初级水平到高级水平，学习者音素平均错误率整体逐步下降，而平均发音错误分散度却呈现先微弱上升后下降的发展趋势,相关性检验结果显示平均错误率与平均发音错误分散度的相关系数为0.966。运用Shapiro-wilk分别对每组数据进行正态性检验，结果显示，初、中、高三个水平的音素错误率均满足正态分布的条件（P_{EL}=0.115>0.05，P_{IL}=0.276>0.05，P_{AL}=0.419>0.05）；三组发音错误分散度中，除初级水平不满足正态分布的条件（P_{EL}=0.033<0.05），IL、AL均为正态分布（P_{IL}=0.287>0.05，P_{AL}=0.136>0.05）。针对满足正态分布和不满足正态分布的组别，我们分别使用配对样本T检验和Wilconxon检验考察不同组别错误率和不同组别发音错误分散度差异的显著性。检验结果显示，从发音错误率来看，与初级水平相比，中级水平学习者声母发音错误率下降不显著（t=-1.706，P=0.103>0.05），而中级水平到高级水平学习者发音错误率显著下降（t=4.38，P<0.001）；从发音错误分散度来看，与初级水平相比，中级水平学习者声母发音错误分散度上升不显著（z=-0.574，P=0.569>0.05），而中级水平到高级水平学习者发音错误率显著下降（t=4.38，P<0.01）。

第二十一届中国计算语言学大会论文集，第79页-第87页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

根据上述组间数据对比可以看出，从初级水平到中级水平，虽然学习者在发音错误率和发音错误分散度两个指标的差异均不显著，而两指标在变化趋势上却呈现出相反的方向，即从初级水平到中级水平，学习者音素错误率下降而错误分散度呈微弱上升的趋势；从中级水平到高级水平，发音错误率和发音错误分散度均显著下降，即从中级到高级水平，学习者的音素发音的准确性有显著进步且目标音素的发音不确定性显著下降。

4 讨论

本文针对第二语言语音习得效果的评价问题，在错误率这一指标的基础上提出了发音错误分散度，并且尝试使用“熵”的计算方法对二语学习者发音错误的分布情况进行量化。本章节我们将分别从不同音素错误率与发音错误分散度以及不同二语水平学习者音素错误率与发音错误分散度两方面对结果进行讨论。

4.1 不同音素错误率与发音错误分散度研究

第二语言语音习得相关理论和研究指出，学习者二语语音习得的效果往往受到到其母语经验的影响(Kuhl, 1993; Flege, 1995; Best and others, 1994; Best and Tyler, 2007)。Flege (1995)在SLM（speech learning model）中提出了一系列二语语音习得的假设，其中包括：1.学习者母语与目的语发音的相似度对二语语音习得有重要的影响；2.二语语音范畴的构建会受到等价归类（equivalent classification）机制的影响。以上假设分别从学习者一语对二语的影响和学习者二语语音的内在加工机制两方面对不同音素的习得结果进行了预测，根据预测学习者对不同的二语语音往往采用不同的学习策略：对与母语中发音一致的“相同音素”（identical phones）及发音类似的“相似音素”（similar phones），一般直接使用母语中的近似音代替二语中音素的发音；对于母语中不存在的“陌生音素”（new phones），只能预测该类音素可能存在学习困难，却无法预测这类音素可能会被发成什么音(鲁健骥, 1984)。由于“陌生音素”与学习者母语中的音素差异较大，学习者往往可以比较容易感知到其与母语音素的差异性，因此在初期学习时可能存在一定的困难，通过不断练习学习者有可能最终可以构建起新的语音范畴；对与母语发音接近的“相似音素”，由于难以感知其与母语音素的差异性，最终可能会被“等价分类”（category classification）到与之相似的母语的语音范畴中，形成复合型语音范畴（composite L1-L2 phonetic category）(Flege and Ocke-Schwen, 1997)。显然，上述预测不仅对不同二语语音的习得结果进行了预测，对不同的语音学习者可以出现的发音错误状态也进行了预测，如果学习者对不同音素的习得情况符合上述预测，那么”相同音素”的错误率最低，语音产出对象最集中，即直接由母语中对应的“相同音素”进行替换，此时发音错误分散度最小；“相似音素”的错误率较“相同音素”会有所上升，由于产出语音主要由母语中对应的“相似音素”进行替换，发音错误的分散度也继续保持较低的状态；“虽然在SLM中陌生音素可能不是最难习得的，其错误对象却存在最大的不确定性，因为在学习者母语中不存在与之对应的替换对象，在新的语音范畴建立之前，“陌生音素”可能存在多种不确定发音对象，发音错误率也保持在较高水平。

本研究使用发音错误率和发音错误分散度两个评价指标呈现出日本学习者产出汉语21个声母的情况。根据上述实验结果可以发现，与日语音素具有相同发音部位和发音方法的鼻音m[m]、n[n]，清塞音b[p]、d[t]、g[k]等音素可被归入“相同音素”，其对应错误率和发音错误分散度最低；日语中不存在的zh[tʂ]、ch[tʂʰ]、sh[ʂ]、r[ʐ]等音素可归入“陌生音素”(廖序东, 2017)(**?**)，其对应的发音错误率和发音错误分散度最高，其余“相似音素”的发音错误率和发音错误分散度整体居中。此外，“相似音素”中p[pʰ]、t[tʰ]、k[kʰ]与“相同音素”b[p]、d[t]、g[k]为送气对立音，它们虽然在错误率上与s[s]、l[l]、f[f]等音素接近，其发音错误分散度却明显低于s[s]、l[l]、f[f]等音素，根据发音错误分布雷达图3显示，p[pʰ]、t[tʰ]、k[kʰ]的错误对象主要为其不送气对立音b[p]、d[t]、g[k]，即日本学习者主要使用b[p]、d[t]、g[k]来替换p[pʰ]、t[tʰ]、k[kʰ]。“陌生音素”的错误分布有更大的不确定性，发音错误分散度表现为最高。上述实验结果不仅印证了SLM对二语语音习得结果的预测，也对学习者对不同音素的错误模式进行了探索，对音系对比、二语语音习得研究及二语语音教学策略的探索都能提供一定的参照。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 不同二语水平音素错误率与发音错误分散度研究

Slinker (2013)指出，中介语被认为是一种动态的、不稳定的语言系统，这往往是因为学习者在不同的学习阶段使用不同的学习策略来获得系统的规则和信息。Flege et al. (2021)在SLM-r（speech learning model-revised）也指出，当学习者初次接触目的语时往往使用母语中存在的语音进行替换发音，随着学习的不断深入，学习者可能会逐渐感知到二语语音与母语语音的差异，从而尝试构建新的语音范畴。

本文通过对比不同水平日本学习者汉语声母产出的情况可以看出，从初级水平到中级水平，学习者的音素错误率呈下降趋势而发音错误分散度呈现微弱上升趋势，虽然错误率和发音错误分散度的变化均没有呈现出显著性差异，我们可以尝试对错误率下降和发音错误分散度上升的现象进行解读：初级水平学习者二语发音错误率较高，在学习上主要使用母语中存在的语音替换二语中的语音，由于错误的对象主要为母语中的音素集合，因此错误的分散度较小；中级水平学习者，一方面音素错误率有一定的下降，同时学习者感知到二语语音与母语语音的差异性，尝试为二语语音构建新的语音范畴，在这一过程中可能会做更多更大胆的尝试，在尝试构建新的语音范畴的过程中也可能伴随更多类型的发音错误，从而造成发音错误分散度的增加。从中级水平到高级水平，学习者音素发音的错误率和发音错误分散度均呈现显著性下降，说明学习者的音素发音取得整体进步，不仅整体错误发音比例降低，而且发音的不确定性减小，错误的种类更集中，这在一定程度上可以反映出中介语语音系统逐渐趋于稳定的发展趋势(Selinker and Rutherford, 2013)。

上述解析只针对于本研究的实验结果，这种发音错误率于错误分散度发展的不同步性是否普遍存在于二语习得的发展过程中，以及不同音素的发展变化有怎样的具体表现有待于进一步探索。

5 总结与展望

本研究通过对日本学习者汉语不同声母发音错误率和发音错误分散度的分析与对比，主要有如下发现：第一，音素错误率与错误分散度两个参数存在较大的相关性，其差异性较好的反映出二语学习者发音错误的分布情况，即错误率相似的两音素，发音分散度越小错误类型越集中；第二，二语音素与学习者母语音素的差异性对学习者的音素发音错误率和发音错误分散度都有较大的影响：二语音素与学习者母语音素差异越大，发音错误率越高，错误分散度也越大；第三，送气音素p[p^h]、t[t^h]、k[k^h]的发音错误分散度明显低于与其具有类似错误率的其它音素。对比初、中、高三个水平学习者对不同音素的习得情况可以发现，与初级水平学习者相比，中级水平学习者的音素发音错误率下降而发音错误分散度微弱上升，中级水平到高级水平学习者的发音错误率和发音错误分散度都显著下降。上述发现证明了使用“熵”可以对二语学习者的音素发音错误分散度进行有效量化，同时发音错误分散度这一指标在错误率的基础上可以进一步揭示不同音素的发音错误分布情况及不同二语水平学习者音素的习得规律。

本文引入“熵”对二语语音发音错误的分散度进行量化，对学习者母语音系和二语学习经验对音素习得效果的影响进行了初步探索，为二语语音习得效果的研究提供了新的视角。然而当前的研究结果是否符合不同母语背景的学习者，学习者在感知上会有什么样的错误模式等问题则需要更深入且广泛的研究来探索。

参考文献

Catherine T Best et al. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The Development of Speech perception: The Transition from Speech Sounds to Spoken Words*, 167(224):233–277.

Catherine T Best and Michael D Tyler. 2007. Nonnative and second-language speech perception: Commonalities ans complementarities. *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, pages 13–34.

Paul Boersma and David Weenink. 2021. Praat: Doing phonetics by computer (version 6.2.04).

Ocke Schwen Bohn and James Emil Flege. 2011. Perception and production of a new vowel category by adult second language learners. *Second-Language Speech: Structure and Process*, pages 53–74.

(c) 2022 中国中文信息学会计算语言学专业委员会

Nancy F Chen, Darren Wee, Rong Tong, Bin Ma, and Haizhou Li. 2016. Large-scale characterization of non-native mandarin chinese spoken by speakers of european origin: Analysis on icall. *Speech Communication*, 84:46–56.

Dustin Crowther, Pavel Trofimovich, Kazuya Saito, and Talia Isaacs. 2015. Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL quarterly*, 49(4):814–837.

James Emil Flege and Ocke-Schwen Bohn. 2021. The revised speech learning model (slm-r). *Second Language Speech Learning: Theoretical and Empirical Progress*, pages 3–83.

James Emil Flege and Bohn Ocke-Schwen. 1997. Perpeption and production of a new vowel category. *Second-Language Speech: Structure and Process*, 13:53.

James Emil Flege, Ocke-Schwen Bohn, and Sunyoung Jang. 1997. Effects of experience on non-native speakers' production and perception of english vowels. *Journal of Phonetics*, 25(4):437–470.

James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, 92:233–277.

Lori L Holt and Andrew J Lotto. 2006. Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, 119(5):3059–3071.

Karen Jesney. 2004. The use of global foreign accent rating in studies of l2 acquisition. *Calgary, AB: University of Calgary Language Research Centre Reports*, pages 1–44.

Gisela Jia, Winifred Strange, Yanhong Wu, Julissa Collado, and Qi Guan. 2006. Perception and production of english vowels by mandarin speakers: Age-related differences vary with amount of l2 exposure. *The Journal of the Acoustical Society of America*, 119(2):1118–1130.

Denis Jouvet, Anne Bonneau, Jürgen Trouvain, Frank Zimmerer, Yves Laprie, and Bernd Möbius. 2015. Analysis of phone confusion matrices in a manually annotated french-german learner corpus. In *Workshop on Speech and Language Technology in Education*.

Patricia K Kuhl. 1993. Innate predispositions and the effects of experience in speech perception: The native language magnet theory. In *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*, pages 259–274. Springer.

Larry Selinker and William E Rutherford. 2013. *Rediscovering interlanguage*. Routledge.

Bruce L Smith, Eric Johnson, and Rachel Hayes-Harb. 2019. Esl learners' intra-speaker variability in producing american english tense and lax vowels. *Journal of Second Language Pronunciation*, 5(1):139–164.

Lei Sun and Vincent J van Heuven. 2007. Perceptual assimilation of english vowels by chinese listeners: Can native-language interference be predicted? *Linguistics in the Netherlands*, 24(1):150–161.

Chiharu Tsurutani and Dean Luo. 2013. Naturalness judgement of l2 mandarin chinese-does timing matter? In *INTERSPEECH*, pages 239–242.

Hongyan Wang, Vincent J Van Heuven, et al. 2006. Acoustical analysis of english vowels produced by chinese, dutch and american speakers. *Linguistics in the Netherlands 2006*, pages 237–248.

Xin Xie and T Florian Jaeger. 2020. Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America*, 147(5):3322–3347.

黄伯荣，廖序东. 2017. 现代汉语（增订六版）. 高等教育出版社.

曹文，张劲松. 2009. 面向计算机辅助正音的汉语中介语语音语料库的创制与标注. 语言文字应用, (4):10.

王玮，张劲松. 2019. 汉语中介语语音库的文本设计. 世界汉语教学, 33(1):13.

朱春跃. 2001. 语音详解. 外语教学与研究出版社.

王玮. 2020. 大规模汉语中介语语音库设计和标注研究. 博士学位论文, 北京语言大学.

鲁健骥. 1984. 中介语理论与外国人学习汉语的语音偏误分析. 语言教学与研究, (3):13.

(c) 2022 中国中文信息学会计算语言学专业委员会

儿童心理词汇输出策略及影响因素研究*

甘嘉铭，王治敏†
（北京语言大学汉语国际教育研究院，北京 100083）
ganjiaming723@163.com，wangzm000@qq.com

摘要

儿童心理词汇研究是儿童词汇研究中的重要部分。本文基于心理词表假设，对827位7-12岁汉语母语儿童展开调查，收集其脑内潜藏的心理词汇，并采用基础词汇定序模型，提取儿童心理词汇定序词表。通过分析词表发现，儿童词汇主要涵盖生活类词汇和以学习为核心的词汇。同时，儿童词汇输出存在思维链的现象，在输出思维链时儿童主要采用了场景策略、范畴策略以及组词策略。此外，通过探究儿童词汇输出影响因素，我们发现儿童输出的词汇量随年龄增长而不断增加，儿童词汇发展从低年龄组到高年龄组发生了显著变化，性别在儿童输出词数上无显著差异，但男孩、女孩关注的词汇类别有各自的倾向。

关键词： 儿童词汇；心理词汇；词汇输出策略；影响因素

A Study of Children's Mental Vocabulary Output Strategies and The Factors Influencing Them

Jiaming Gan, Zhimin Wang
(Research Institute of International Chinese Language Education,
Beijing Language and Culture University, Beijing 100083)
ganjiaming723@163.com, wangzm000@qq.com

Abstract

The study of children's mental vocabulary is an important part of children's vocabulary research. Based on the mental word list hypothesis, this paper investigates 827 native Chinese children aged 7-12 years old to collect their mental vocabulary in their brains, and uses the basic lexical sequencing model to extract a mental vocabulary sequencing word list for children. The analysis of the word list revealed that children's vocabulary mainly covered life words and learning-oriented words. At the same time, children's vocabulary output has a chain of thinking, and children mainly use scene strategy, category strategy and word formation strategy in outputting the chain of thinking. In addition, the investigation of the factors influencing children's vocabulary output revealed that children's vocabulary output increased with age, and children's vocabulary development changed significantly from the lower to the higher age groups, while gender did not differ significantly in the number of words children output, but boys and girls had their own tendencies in the categories of words they focused on.

* 基金项目：国家社科基金重大项目（18ZDA295）；国家语委科研项目（ZDI135-139）；中央高校基本科研业务费（19PT03）

† 通讯作者 corresponding author

第二十一届中国计算语言学大会论文集，第88页-第98页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

Keywords: Children's vocabulary , Mental vocabulary , Vocabulary output strategies , Influencing factors

1 引言

儿童的词汇输出是衡量儿童语言能力发展的重要指标，儿童输出的词汇越丰富，其语言运用越灵活。在前人研究中，主要集中于6岁以下儿童的词汇输出情况。潘伟斌(2017)以学龄前3-6岁幼儿为对象，基于儿童语料库（CHILDES）研究方法，对幼儿整体词汇系统发展进行研究，并对实词、虚词发展进行分类研究，发现幼儿词汇发展有共性顺序但存在个性差异，词频发展速率具有不均衡性。曾涛、邹晚珍(2012)考察了汉语儿童6岁前范畴层次词汇的发展情况，为佐证基本层次词语在汉语儿童早期语言发展中占据概念词汇的主导地位提供了经验支持。仅有少部分以小学低年级儿童为研究对象，程亚华等(2018)发现1-3年级学生口语词汇知识发展轨迹呈曲线形式，其中前两年呈线性发展，三年级时呈加速发展，发展速度是前期发展的两倍，起始水平和发展速度均存在显著的个体差异。

儿童词汇输出影响因素的研究也主要集中在儿童6岁以前。牛杰等(2015)发现，18-24月龄时，年龄与性别对儿童早期词汇发展有重要作用，但其影响力随月龄增长而减少。郭芙蓉(2017)表示，儿童5-6岁时，家庭经济资本、社会资本和文化资本均对儿童词汇水平具有显著影响，其中文化资本的影响最大，经济资本次之，社会资本最弱。马明明(2021)发现，在儿童6岁时，其词汇发展水平与使用语言类型、阅读书籍数量和阅读侧重有显著正相关。

前人的研究成果给予我们很大的启发，6岁以上的儿童的常用词是什么？儿童输出词汇时是采用什么策略组织词汇的？年龄与性别对6岁以上儿童词汇输出的影响是否仍然显著？这些都是值得深入研究的问题。因此，本文基于心理词汇理论，对7-12岁小学适龄儿童展开调查，收集儿童心中的基础词汇，提取儿童心理词汇定序词表，力求探究儿童词汇输出的策略，探索年龄与性别对儿童词汇输出的影响。研究可为小学儿童词汇联想提供策略参考，对儿童读物词汇选取具有重要借鉴意义。

2 儿童心理词汇定序词表

心理词汇（mental lexicon）的概念由19世纪60年代的认知心理学家Treisman(1960)提出，又可译为"大脑词库"、"心理词典"、"心理词库"、"内部词汇"、"内部词典"等。它是指保存在人脑内部的一个词表，这个词表存储了大量的词条，每个词条又包含词的写法、语音以及词义等信息。心理词表的词语按照一定的方式组织起来，如可以按词的使用频率来组织 (杨亦鸣et al., 2001)。

心理词汇的研究方法众多，其中广泛应用的是词汇联想测试法。词汇联想测试需要选择一定数量的刺激词，要求受试根据刺激词做出反应。但此方法需要选定刺激词，被试输出的词汇极易受刺激词影响。刺激词选取是否科学，对词汇联想测试的结果将产生直接影响。同时，刺激词的设置加入了调查者的主观因素，被试不可避免会局限于刺激词限定的范围，处于被动参与的局面，这对于探索儿童大脑内部词汇情况有不小的限制。

为了解决词汇输出受限的问题，在前人研究的基础上，本文采用儿童自主输出词汇的方法，不对儿童输出的词汇设置刺激词，最大限度体现了儿童思考，减少人工干预，为探索儿童心理词表提供更广阔的空间。

本文以儿童具有潜在的心理词表为前提，对7-12岁母语为汉语的儿童的心理词汇进行调查，提取儿童大脑内部的心理词汇，探究这个阶段儿童最常用的基础词汇，构建儿童心理词汇知识库。实际上，儿童12岁以前均处于语言发展的关键期，对7-12岁儿童词汇发展进行研究同样十分重要。12岁以前，儿童词汇发展皆十分迅速。相较于6岁以下儿童，7-12岁儿童有其自身的特点，该阶段儿童已经掌握了大量的基础词汇，各词类发展相对成熟，词语表达能力更强。在家庭、学校、社会共同影响下，儿童的次常用词汇也飞速发展，出现了一些更为专业性的、领域性的词汇。以7-12岁儿童为研究对象，有助于了解该阶段儿童的常用词汇，挖掘其词汇特点，探究其词汇输出规律。

本项研究采用开放式问卷的调查方式，让儿童在自然状态下输出心理词汇，从而考察儿童的心理词表。该问卷限时15分钟，让儿童尽可能多地说出其心目中的常用词，儿童可以拿着录

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

音设备进行头脑风暴，边想边说，不限制儿童的行动，让儿童在轻松快乐的氛围中完成挑战，以便较为自然地获取儿童心理词汇，体现了学生最大限度思考和减少人工干预两大特点。在儿童输出词汇的过程中，语料采集人员不能对儿童进行提示，同时，尽量避免儿童家长、同伴、老师对儿童的提示，让儿童自主完成挑战，确保语料真实体现儿童的认知水平、语言水平。在录音前，需确定儿童的基本信息，具体包括年级、性别、出生日期、年龄等。

研究对以汉语作为母语的一至六年级小学生发放问卷，共回收语料857份。去除部分录音时长不够、未按要求说词的语料，获得有效语料共827份。九成语料来自于北京儿童，北京儿童语料来自于北京市海淀区的小学。所采访的被试均无口吃等语言障碍且智力正常。研究通过对小学一到六年级学生进行录音调查，获得儿童词汇语料。具体学生年级分布，如图 1所示：

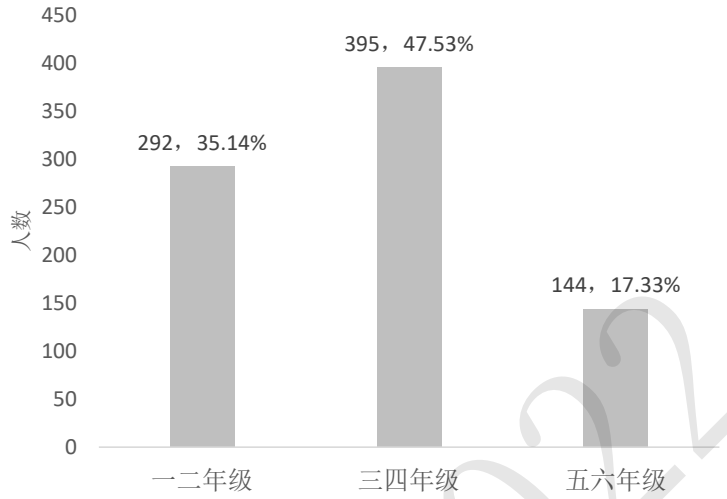


图 1: 学生年级分布

由图 1可知，一二年级、三四年级学生参与数量较多。三四年级学生人数最多，为395人，占总人数的47.53%。其次是一二年级学生，共292人，占总人数的35.14%。五六年级学生数量相对最少，为144人，占总数的17.33%。因此，在后续统计语料时，需要注意语料数量不均的问题，并进行等比换算。大部分而言，7-8岁为低年龄组，对应年级为一二年级；9-10岁为中年龄组，对应年级为三四年级；11-12岁为高年龄组，对应年级为五六年级，也存在少数年龄与年级不一致的情况。同时，我们对调查对象的男女比例进行考察，具体如图 2所示：

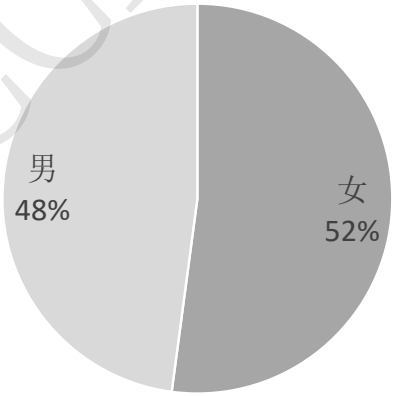


图 2: 儿童男女比例

根据图2我们可以看到，女生占比为52%，男生占比为48%，两者比例大体相当，可以忽略性别差异对整体语料的影响。

本文先对827份儿童心理词汇的基本情况进行统计，若不计单份语料内部的重复词汇，本次调查的7-12岁儿童出现词汇总数为225371个，去除不同语料之间的重复词汇，共有40200词。通过问卷调查，我们发现儿童对词的概念较为模糊，不仅包括传统语言学上界定的词，还包括一

(c) 2022 中国中文信息学会计算语言学专业委员会

些短语、短句。例如，“语文老师”“我的世界”“找不到”“做作业”“你好”“对不起”等。这些短语、短句在生活中使用频率非常高，因此，本文将这类短语、短句作为独立的语块，同样将其收录儿童心理词汇知识库。问卷结果绝大部分为常用词，仅存在少数的常用短语、短句，这说明儿童基本建立了词的概念，但与语言学中对词的理解还有一些出入，儿童对词和短语的界限还未建立起严格的区分意识。未来我们将对儿童心理词汇知识库中存在的语块展开深入研究。

Zhimin Wang et al.(2018)提出了提取基础词汇的定序方法与模型，通过设置位置区间、调整参数权重等手段不断改进，其常用度定序实验取得了很好的效果。根据该模型，本文以频次、词语输出位置、词语稳定性等参数对儿童心理词汇进行定序，提取了一份儿童心理词汇定序词表。

表格中显示的统计模型的参数包括Freq、Distinct_freq、Avg(Pos)、StDev(Pos)、distinct_freq*distinct_freq/StDev (Distinct Freq_Pos)。其中，freq为词语的频次，单份语料中的重复词既占位也计数。Distinct_freq为词语的确切频次，Avg(Pos)为词语的平均位置，StDev(Pos)为位置标准差，单份语料中重复词不计数但占位。distinct_freq*dist inct_freq /StDev(DistinctFreq_Pos)代表词语的常用度。

表中设定Freq、Distinct_freq两个与频次相关的量，其原因在于，在收集的问卷中单份语料，少数词语出现重复。例如，“英语”的Freq为545，但是存在单份问卷中“英语”出现两次或者多次的情况，因此有必要进行数据清洗。本文首先对所有单份语料进行去重处理，同时模型上也对重复词进行了过滤处理，只保留该词位置最小的记录。其他重复词不进入位置统计，但占位。经过处理，827份问卷中，去重后词语总数为40200条。其中“英语”的Distinct_freq为543。常用度定序模型使用参数Distinct_freq进行计算。我们截取儿童心理词汇定序词表中前20个词语示例如表 1：

ID	Word	Freq	Distinct_freq	Avg(Pos)	StDev(Pos)	$\frac{distinct\_freq*distinct\_freq}{StDev(DistinctFreq\_Pos)}$
1	电脑	708	708	43.83898305	54.71773583	9160.905371
2	衣服	553	553	86.27486438	78.10351989	3915.43173
3	英语	545	543	94.45488029	79.22256756	3721.78041
4	黑板	491	491	74.57637475	70.50554886	3419.319527
5	窗户	521	521	89.88291747	80.68282939	3364.296989
6	桌子	510	507	69.25246548	76.55734481	3357.600772
7	窗帘	522	522	94.12835249	83.23512012	3273.666207
8	头发	511	511	106.3189824	80.29588671	3251.984761
9	语文	510	510	99.18627451	81.39900638	3195.3707
10	手机	493	493	85.65720081	79.92925056	3040.801688
11	数学	508	508	103.3877953	85.39187546	3022.114207
12	耳机	476	476	91.70378151	75.44603137	3003.153326
13	鼠标	477	477	76.17610063	75.95060325	2995.749741
14	苹果	506	505	65.75247525	85.58667819	2979.72775
15	眼睛	501	501	109.2894212	84.85537147	2957.98599
16	空调	478	478	90.20920502	79.37889236	2878.397433
17	老师	492	492	98.06097561	87.96213047	2751.911518
18	椅子	482	481	81.1018711	84.29355956	2744.705541
19	裤子	459	458	94.97598253	77.67954054	2700.376425
20	键盘	465	465	79.69247312	80.23659819	2694.842564

表 1: 儿童心理词汇定序词表前20词

表 1的数据中，排序的前20位词语分别是“电脑、衣服、英语、黑板、窗户、桌子、窗帘、头发、语文、手机、数学、耳机、鼠标、苹果、眼睛、空调、老师、椅子、裤子、键盘”。这些词的频次和输出位置都比较平均，排序靠前合理，它们作为儿童的基础词汇完全符合人们的认知。

(c) 2022 中国中文信息学会计算语言学专业委员会

排名第一的词语是“电脑”，前20词中与之相关的还包括“手机、耳机、鼠标、键盘”，这些词大部分属于办公类的电子产品。随着全球信息化的冲击，电脑、手机等电子产品迅速普及，儿童对新鲜事物的接受度高，受到的影响也反映在输出的词汇上，也可看出“电脑”在儿童生活中的重要地位。1984年，邓小平同志提出“计算机普及，要从娃娃抓起。”多年过去，今天的电脑已成为儿童手里的“玻璃珠”，深入渗透到儿童的生活之中。电脑、网络成为了这个时代馈赠给儿童的宝贵礼物。

排名第二的词语是“衣服”，前20词中“裤子”一词与其同属服饰类，这些都是日常生活中与生活用品相关的基础词汇。前20 词的生活用品类词汇还包括家具、家电类，如“窗户、桌子、窗帘、空调、椅子”。生活用品类词汇在前20词中出现了7个，占整体的35%，占比很高，也凸显了它的重要性。

排名第三的词语是“英语”，前20词中的“语文、数学”与之对应。语数英是小学教学的主要科目，从词汇上体现了儿童主要的学习内容。我们还分别对7-8岁、9-10岁、11-12岁的儿童心理词汇定序词表进行统计，其中“英语”分别排名第八、第七、第二，排名越来越高。这说明随着年龄增长，儿童对英语更加关注。“英语”排名如此靠前，可能与中外交流日益密切相关，国家从政策层面重视英语。1964年，教育部将英语确定为第一外语，调整了中学的外语教学结构。目前国内小学三年级便开始在校学习英语。近些年，社会上掀起了“英语热”，英语教育机构层出不穷，国内也拥有了多种英语考试体系。

排名第四的词语是“黑板”，这与学生的学校生活息息相关。儿童在小学校园生活中的重要词汇既包括学校设施“黑板”，也包括学习科目“英语、语文、数学”，还包括教学的人“老师”。这也说明了儿童的生活通常围绕学校、学习，反映了学生的学习生活。

在前20词中，唯一出现的表人的词语为“老师”，也说明了“老师”是组成儿童社会关系的重要人物。需要思考的是，“妈妈”“爸爸”的排序在“老师”之后。儿童定序词表中，“妈妈”排在第22位，“爸爸”排在第24位，不在“老师”之前。这可能是由于采集语料的地点在学校及学校附近，近取诸身，学生更容易联想到“老师”。

此外，这20个词中还包括身体部位类词汇“头发、眼睛”、食品类词汇“苹果”，这些词语都是各自类别中的基础词。身体部位类词汇体现了儿童对自己身体的关注，从小父母就会对儿童进行身体部位指称的教学。“头发”的排序在第8位，“眼睛”的排序在第15位，“鼻子”在第21位，“嘴巴”在第33位，“耳朵”在第56位。身体部位词的输出排序，体现了儿童联想描述时可能存在先上再下、先中间后两边的联想顺序。

食品类词语第一名是“苹果”，也体现了“苹果”为人所熟知。通常说到水果，国人的第一反应便是苹果。这可能是因为苹果是现在市面上的主流水果，一年四季都能买到，且营养价值高。其寓意为“平安”，寓意美好。在各方面因素的综合影响下，苹果成为了家喻户晓、深受大家喜爱的食物。

根据对表 1的分析，我们可以初步判断儿童词汇的特点，儿童词汇主要涵盖生活类词汇和以学习为核心的词汇，此外，电脑、手机等电子产品类词汇在儿童生活中占据重要地位。

3 儿童词汇输出策略

儿童词汇输出存在一定的规律。通过观察语料，本文发现一些词语之间具有较强的关联，这些词可以形成连续的有关联的词串。这些词串如同儿童思维中的链条，将有相关特征的词语联结在一起，我们将之称为“思维链”。这些思维链是按照特定的方式组织起来的，其中的纽带就是儿童词汇输出策略。在这一章中，我们将以儿童心理词汇中的思维链为研究对象，分析、归纳儿童词汇输出策略，挖掘思维链背后反映的儿童认知发展情况，以期了解儿童的生活与心理。此外，本章还将对儿童词汇输出的影响因素进行研究。

上文提到了“思维链”的概念，但在对儿童词汇思维链进行研究之前，我们还需要对思维链进行进一步的界定。通过对儿童语料的充分研究，同时参考Xi Wang and Zhimin Wang(2020)关于词汇输出线索的分类，本文认为，2个及2个以上连续的存在明显关联的词语集合可以被视为一条思维链。通过观察语料，我们发现思维链背后的词汇输出策略可以分为三种：一是场景策略，连续输出的词属于一个场景，例如“熬粥——炒菜——煮饭”这一思维链，这三个词都处于“厨房”场景；二是范畴策略，连续词语拥有共同的上位词，例如“红色——黄色——蓝色——粉色”，这四个词均为颜色词，拥有共同上位词“颜色”；三是组词策略，包括同音组词、同字组词等。场景策略与范畴策略属于互斥关系。场景策略的词语需在常识上处于同一场景，

(c) 2022 中国中文信息学会计算语言学专业委员会

不限词性，范畴策略的词语需有共同的上位词，且词性一致。此外，儿童在词汇输出的过程中，存在词汇跳跃现象。例如，在“鳄鱼、恐龙、袜子、帽子、衣服、葡萄、香蕉、苹果”这组词中，儿童的思维由动物词跳跃至生活用品词，又跳跃至食品词，思维链中间断裂，则不可视为一条完整的思维链，而只能标为三条思维链。

我们提取了儿童语料中所包含的思维链，对每条思维链的词汇输出策略进行标注，并对思维链的长度（即单条思维链包含的词语数量）进行统计。下面将从场景策略、范畴策略、组词策略三个方面进行分析。

3.1 场景策略

通过观察儿童语料，我们发现一些连续输出的词可以归纳到同一个场景中，儿童可以通过联想与某个场景相关的词来输出思维链。我们将之称为“场景策略”，具体可分为“家、学校、游戏、交通、购物”五个场所场景和日常交际场景，详见表 2:

场景	思维链示例	最大长度
家	熬粥——炒菜——煮饭——厨房 窗户——床铺——卧室——阳台——家里	38
学校	上课——下课——玩耍——放学——做作业 班队——班会——戴口罩——复习——考试	34
游戏	滑滑梯——跷跷板——游乐园 王者荣耀——我的世界——和平精英	18
交通	电动车——小车——摩托车——运货车 大巴车——公交车——自行车	13
购物	囤货——退款——淘宝 购物——购买——超市——小卖铺	9
日常交际	对不起——请原谅—回头见 你好——再见——在吗——拜拜	5

表 2: 场景策略下的思维链

通过分析表 2，我们可以发现，儿童在输出与“家”“学校”相关场景的词汇时，输出思维链长度的最大值较大，分别为38和34。而且家校场景在语料中也出现得十分频繁。思维链的长度越长，表明儿童对该类场景的了解越深。可以看出，儿童主要接触的场景为“家”和“学校”，因此对这两个场景最为熟悉，输出的相关词汇也最多。

儿童对“游戏”场景较为热衷，词汇输出最大长度也较长。不论是高年级的儿童还是低年级的儿童都有自己感兴趣的游戏领域。年龄偏低的儿童对于运动类的游戏有较大的兴趣，如“滑滑梯”“跳皮筋”“放风筝”等。年龄稍大一些的儿童对于电子竞技类游戏表现出更多的兴趣，如“王者荣耀”“和平精英”等。

在上述表格中，可以看到，儿童熟知的词语基本囊括在家、校、游戏、交通等几大场景中，海外儿童汉语教材、汉语母语儿童绘本在编写时也可参考这些场景。此外，每个场景中儿童输出的词汇均为儿童心目中该场景的常用词汇，这对海外儿童汉语教材、汉语母语儿童绘本词汇的选取也具有较大的借鉴意义。

3.2 范畴策略

我们观察到，一些连续输出的词语可以归纳为同一范畴，这些词语具有共同的上位词，且词性相同。详见表 3:

根据表 3，我们可以发现“食品”“生活用品”“动植物”“文化用品”“亲属称谓”“颜色”“交通工具”“身体部位”等范畴是汉语母语儿童最常出现的心理词汇范畴。其中，食品类词汇出现的思维链最长，最大长度为57，即儿童一次能连续输出57个食品类词语，也表明了儿童对食品的喜爱。

在最常见的八个范畴中，连续产出词汇思维链较长的范畴为“食品”“生活用品”“动植物”和“文化用品”类词汇。这四类词汇从高到低反映了汉语母语儿童对于生存、生活、爱好、学习（发展）的重视程度，符合人类对世界的认知规律，也反映了儿童单纯健康的心理状态。

(c) 2022 中国中文信息学会计算语言学专业委员会

范畴类别	思维链示例	最大长度
食品	苹果——香蕉——桃子 汉堡——薯条——香肠——薯片——披萨	57
生活用品	袜子——帽子——衣服 项链——耳环——戒指——皮带	38
动植物	狮子——老虎——兔子 小树——玫瑰花——太阳花——向日葵	27
文化用品	卷笔刀——橡皮——书本——尺子 语文书——数学书——课堂作业本	24
亲属称谓	妹妹——妈妈——爸爸 姥姥——阿姨——爷爷——奶奶	20
颜色	黄色——粉色——蓝色 白色——黑色——绿色——紫色——灰色	19
交通工具	电动车——小车——摩托车——运货车 大巴车——公交车——自行车	13
身体部位	眼睛——眼——眉毛——鼻子——胡子 蛀牙——白牙——舌头——肠子	12

表 3: 范畴策略下的思维链

连续产出词汇思维链较短的范畴为“亲属称谓”“颜色”“交通工具”和“身体部位”类词汇。虽然这四类词汇的思维链较短，但其词类本身常用词语数量相对较少，而儿童输出词汇能够包含该词类的大部分词语。例如，在颜色类中，常见的颜色只有12种颜色（十二色相环），但汉语母语儿童颜色词输出最多为19个词，这说明汉语母语儿童在该词类掌握的词语较为丰富。

3.3 组词策略

通过观察语料，我们还发现一些连续输出的词语由同音、同字的语素组词而成，或根据其近义、反义组词，如表 4:

组词	思维链示例	最大长度
同音组词	保护——宝物——饱满——宝刀 羁绊——打扮——伴奏——花瓣	9
同字组词	国家——爱国——祖国 次年——次日——次品	6
近义词	各个——各位 侵犯——侵入	2
反义词	开始——结束 聪明——愚蠢	2

表 4: 组词策略下的思维链

从表 4中，我们观察到儿童使用组词策略时，思维链的长度均较短。同音组词和同字组词的思维链长度均不超过10，使用近义词、反义词策略的思维链长度仅为2。这意味着组词策略的联想效果较另两个策略低。

儿童在输出思维链的过程中使用了组词策略，也意味着儿童能够较为灵活地运用词汇中的语素进行新一轮的构词，表明儿童初步具有语素意识。但是部分儿童对词的概念不太明确，在输出词汇的过程中夹杂了少量的短语和句子，如“擦桌子”“牛吃草”等。

4 儿童心理词汇影响因素研究

儿童词汇发展受到多种因素的影响。目前研究表明，生理因素、语素意识、认知水平发展、家庭社会环境、母亲教育程度等因素对早期儿童语言发展发育都存在影响。其中，生理因素主要包括年龄和性别两个方面。

(c) 2022 中国中文信息学会计算语言学专业委员会

牛杰等(2015)指出生理因素（年龄、性别）对儿童早期词汇发展的影响力随月龄增长而减少。为了验证生理因素在小学阶段是否仍然存在影响，本研究将从年龄、性别两个方面对儿童词汇发展的影响进行探索。

4.1 年龄因素

章依文等(2002)指出年龄增长对2-3岁儿童语法发展起着促进作用。马明明(2021)发现4-6岁儿童的词汇发展水平存在显著的年龄差异。国外相关研究显示，儿童早期词汇发展存在词汇飞跃现象，表现为在独词句阶段词汇快速增长 (McShane, 1980; Dromi, 1987; Caselli and Casadio, 2001)。上述研究表明，在儿童早期词汇发展阶段，年龄是影响儿童语言发展的关键要素。

小学阶段儿童词汇发展是否仍然存在显著增长？我们以儿童输出词数作为衡量儿童词汇发展情况的主要参考，探究年龄因素对该阶段儿童词汇发展的影响。本文先对不同年龄组儿童词汇输出的词数进行描述性统计，如表 5所示：

年龄组	个案数	平均值	标准差	标准误	平均值的95% 置信区间	
					下限	上限
低年龄组（7-8岁）	391	246.86	76.42	3.865	239.26	254.46
中年龄组（9-10岁）	315	283.96	85.453	4.815	274.49	293.43
高年龄组（11-12岁）	121	326.11	118.296	10.754	304.81	347.4
总计	827	272.59	91.37	3.177	266.35	278.82

表 5: 不同年龄组词汇数量统计表

由表 5可知，低年龄组儿童输出词数的均值为246.86，中年龄组儿童均值为283.96，高年龄组儿童均值为326.11，即随着年龄的增长，儿童输出词数呈上升趋势。为了探究年龄因素对小学阶段儿童输出词数的影响是否显著，本文对不同年龄组儿童的输出词数进行了单因素方差分析，结果如表 6：

	平方和	自由度	均方	F	显著性
组间	646087.959	2	323043.98	42.592	0.000 ***
组内	6249740.609	824	7584.637		
总计	6895828.568	826			

表 6: 年龄在儿童输出词数上的ANOVA检验

由表 6可知，F（646087.959，2）= 42.592，p=0.000＜0.001，说明年龄因素显著影响到儿童的输出词数，上述描述性统计结果中的上升趋势具有统计学意义的显著性。为进一步探究各年龄组之间存在的差异，我们使用LSD法进行事后检验，结果见表 7：

(I) 年龄	(J) 年龄	平均值差值(I-J)	标准误	显著性	95% 置信区间	
					下限	上限
低年龄组（7-8岁）	中年龄组	-37.097***	6.594	0.000	-50.04	-24.15
	高年龄组	-79.246***	9.06	0.000	-97.03	-61.46
中年龄组（9-10岁）	低年龄组	37.097***	6.594	0.000	24.15	50.04
	高年龄组	-42.149***	9.315	0.000	-60.43	-23.87
高年龄组（11-12岁）	低年龄组	79.246***	9.06	0.000	61.46	97.03
	中年龄组	42.149***	9.315	0.000	23.87	60.43

表 7: LSD事后检验（不同年龄组两两比较）

由表 7可知，低年龄组与中年龄组之间（p＜0.001）、低年级组与高年级组之间（p＜0.001）和中年级组与高年级组之间（p＜0.001）均具有显著差异。可见，儿童词汇输出词数在各年龄组之间均具有显著差异，年龄因素对小学阶段儿童存在显著影响。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 性别因素

牛杰等(2015)指出，18-24月龄期间，性别对儿童早期词汇发展的影响力随月龄增长而减少。马明明(2021)则发现，4-6岁时，性别在儿童词汇能力发展中差异显著，女童的词汇能力要优于男童。以上研究均为儿童早期词汇发展性别影响的相关研究。

为探究性别因素对小学阶段儿童词汇发展的影响，本文对性别因素进行了独立样本T检验，详见表 8：

		莱文方差等同性检验		平均值等同性t 检验						
		F	显著性	t	df	Sig.（双尾）	平均值差值	标准误差值	差值95% 置信区间 下限	上限
词数	假定等方差	12.66	0.00	1.707	825	0.088	10.843	6.354	-1.628	23.314
	不假定等方差			1.69	751.965	0.092	10.843	6.418	-1.756	23.442

表 8: 性别在儿童输出词数上的差异性检验

由上表可知，莱文方差等同性检验的显著性p=0.00<0.05，说明两个独立样本的方差不相等，采用不假定等方差数据。t检验结果显示，t =1.69，p = 0.092>0.05，说明性别在儿童词汇输出上无显著性差异。可见，性别对7-12岁儿童词汇发展无显著影响。

一般而言，男孩与女孩的思维存在一定的差异性，两个群体感兴趣、关注的词汇可能存在不同。虽然不同性别儿童在词汇数量的表现上无显著差异，但不意味着在关注的词汇类型上不存在差异。

本文同样采用基础词汇的定序模型，对男孩、女孩的心理词汇进行定序排列，提取了男孩心理词表与女孩心理词表。研究选取了男孩、女孩心理词表中的前50个词进行对比，具体如表 9：

交集词汇（42）		电脑、衣服、头发、桌子、黑板、窗帘、裤子、眼睛、老师、英语、窗户、苹果、耳机、椅子、鼻子、鞋子、语文、鼠标、手机、空调、数学、妈妈、键盘、爸爸、红色、香蕉、嘴巴、蓝色、台灯、绿色、铅笔、袜子、学校、太阳、汽车、学生、足球、小学、教师、电视、摄像头、年级
非交集词汇（16）	男（8）	飞机、篮球、科学、中国、老虎、地球、电线、教室
	女（8）	口罩、红领巾、黄色、眼镜、白色、同学、耳朵、月亮

表 9: 男孩、女孩心理词表前50词对比分析表

由表9可知，在男孩、女孩心理词表前50个词中，交集词汇高达42个，占84%。可见，在男孩与女孩的心理词表中，两者常用度高的词语存在很大的相似性。

排序越靠前的词语，常用度越高，也表明了这些词在儿童心里的地位越高。我们可以通过观察这些相对另一性别排序更高的非交集词汇，初步判断男孩与女孩关注词汇类型的差异。男孩的非交集词汇为“飞机、篮球、科学、中国、老虎、地球、电线、教室”，从这些词语中可以看出，男孩对外界探索的欲望较强，喜欢凶猛的动物，热爱篮球这类的团队运动。女孩的非交集词汇为“口罩、红领巾、黄色、眼镜、白色、同学、耳朵、月亮”，可以看出女孩十分关注身边出现的人和事物，对颜色更为敏感，喜欢月亮这类寓意美好的事物。

5 结语

本文基于心理词汇理论，收集了827份7-12岁汉语母语儿童词汇有效语料，经过语料校对、清洗、去重等处理，共获得40200个词条。研究参考词汇定序模型，以频次、词语输出位置、词语稳定性等参数对儿童心理词汇进行定序，获得儿童心理词汇定序词表。

通过分析儿童心理词汇定序词表，本文对儿童词汇特点进行探究，发现儿童输出的词汇主要涵盖生活类词汇和以学习为核心的词汇，电脑、手机等电子产品类词汇在儿童生活中占据重要地位。

(c) 2022 中国中文信息学会计算语言学专业委员会

研究发现在儿童输出词汇的过程中，存在“思维链”的现象，儿童词汇输出策略是组织起“思维链”的纽带。儿童词汇输出策略可以具体分为场景策略、范畴策略和组词策略。通过分析、归纳儿童词汇输出策略，我们发现了思维链背后反映的儿童认知与心理，具体如下：

1. 在场景策略中，我们观察到“家”“学校”“游戏”“交通”“购物”五个场所场景和日常交际场景是儿童较为熟知的场景。其中，儿童最常接触的场景为“家”和“学校”，且对“游戏”场景十分热衷。

2. 在范畴策略中，我们发现儿童最常出现的心理词汇范畴是“食品”“生活用品”“动植物”“文化用品”“亲属称谓”“颜色”“交通工具”“身体部位”，其中，食品类词汇出现的思维链最长。通过对范畴策略思维链长度进行统计，我们还发现汉语母语儿童对于生存、生活、爱好、学习（发展）的重视程度呈现由高到低的排序。

3. 在组词策略中，我们可以看到儿童具有了初步的语素意识，但儿童对于词的概念界定尚不明确。

此外，我们还发现儿童输出的词汇量随着年龄的增长而增长，不同年龄组在儿童输出词数方面具有显著性差异，儿童词汇发展从低年龄组到高年龄组发生了显著变化。对于不同性别儿童的词汇输出情况，我们发现性别在儿童输出词数上无显著差异，但男孩、女孩的词汇输出有其倾向性。

通过对儿童词汇定序词表、输出策略以及影响因素的研究，本文得到以下启发：

1. 根据儿童词汇定序词表中的儿童常用词汇及词汇类型，儿童读物在选择词汇时可以进行参考，选取儿童需要掌握的词汇及感兴趣的词汇等。

2. 儿童词汇输出具有潜在的策略。小学儿童在学习词汇时可采用场景、范畴、组词联想等方法，建立词汇之间的联系，帮助词汇理解与记忆。

3. 随着年龄的增长，7-12岁儿童习得的词汇数量显著增多。在这个过程中，可以通过增加儿童阅读量，帮助儿童扩大词汇量，提高阅读理解能力。

本项研究通过构建儿童心理词汇知识库，提取儿童心理词汇定序词表，积累了丰富的儿童词汇资源，深入了解了儿童常用词及其类型。研究观察儿童词汇输出顺序，发现词汇“思维链”现象，总结儿童词汇输出策略，为儿童词汇联想记忆提供了方法支持。同时，验证了年龄与性别对7-12岁儿童词汇发展的影响，结合6岁以下儿童相关影响因素研究，可推出年龄与性别对12岁以前儿童影响的趋势。儿童心理词汇研究还有很多值得探索的空间，未来我们将对不同性别儿童输出词汇偏向、儿童词汇影响因素、儿童混合使用输出策略情况进行进一步研究。除此之外，7-12岁小学儿童心理词表相较于6岁以下儿童心理词表的词汇特点也是我们研究的重点。

参考文献

Maria Cristina Caselli and Paola Casadio. 2001. Lexical development in English and Italian. *Dissertations Theses Gradworks*.

Esther Dromi. 1987. Early lexical development. *Singular Pub Group*.

John McShane. 1980. *Learning to Talk*. Cambridge: Cambridge University Press.

Anne Treisman. 1960. Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*.

Xi Wang and Zhimin Wang. 2020. Study on the order of vocabulary output of international students. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 535–545.

Zhimin Wang, Huizhou Zhao, Junping Zhang, and Caihong Cao. 2018. Research on basic vocabulary extraction based on Chinese language learners. In *Paper presented at 19th Chinese Lexical Semantic Workshop*.

(c) 2022 中国中文信息学会计算语言学专业委员会

曾涛and 邹晚珍. 2012. 汉语儿童6岁前范畴层次词汇的发展研究. 心理科学.

杨亦鸣, 曹明, and 沈兴安. 2001. 国外大脑词库研究概观. 当代语言学.

潘伟斌. 2017. 基于语料库下的3-6岁儿童词汇发展研究. 北京印刷学院学报.

牛杰, 陈永香, and 朱莉琪. 2015. 生物和家庭因素对汉语儿童词汇和智能发展的影响. 中国当代儿科杂志报.

程亚华, 伍新春, 刘红云, and 李虹. 2018. 小学低年级儿童口语词汇知识的发展轨迹及其影响因素. 心理学报.

章依文, 金星明, 沈晓明, and 张锦明. 2002. 2～3岁儿童词汇和语法发展的多因素研究. 中华儿科杂志.

郭颖. 2017. 家庭背景对*5-6*岁儿童词汇水平的影响研究. 硕士.

马明明. 2021. *4-6*岁学前儿童词汇发展水平与家庭亲子阅读环境的相关研究. 硕士.

CCL 2022

汉语增强依存句法自动转换研究

余婧思[1,2,3]，师佳璐[1,2,3]，杨麟儿[1,2,3*]，肖丹[4]，杨尔弘[1,3]
[1]北京语言大学 国家语言资源监测与研究平面媒体中心
[2]北京语言大学 信息科学学院
[3]北京语言大学 语言资源高精尖创新中心
[4]信阳学院 文学院
yujingsi1107@gmail.com

摘要

自动句法分析是自然语言处理中的一项核心任务，受限于依存句法中每个节点只能有一条入弧的规则，基础依存句法中许多实词之间的关系无法用依存弧和依存标签直接标明；同时，已有的依存句法体系中的依存关系还有进一步细化、提升的空间，以便从中提取连贯的语义关系。面对这种情况，本文在斯坦福基础依存句法规范的基础上，研制了汉语增强依存句法规范，主要贡献在于：介词和连词的增强、并列项的传播、句式转换和特殊句式的增强。此外，本文提供了基于Python的汉语增强依存句法转换的转换器，以及一个基于Web的演示，该演示将句子从基础依存句法树通过本文的规范解析成依存图。最后，本文探索了增强依存句法的实际应用，并以搭配抽取和信息抽取为例进行相关讨论。

关键词： 依存句法；汉语增强依存句法；自动转换

Transformation of Enhanced Dependencies in Chinese

Jingsi Yu[1,2,3], Jialu Shi[1,2,3], Liner Yang[1,2,3], Dan Xiao[4], Erhong Yang[1,3]
[1]National Language Resources Monitoring and Research Center Print Media Language Branch, Beijing Language and Culture University
[2]School of Information Science, Beijing Language and Culture University
[3]Advanced Innovation Center for Language Resources, Beijing Language and Culture University
[4]College of Chinese Language and Literature, Xinyang University
yujingsi1107@gmail.com

Abstract

Syntactic analysis is a key step of the natural language understanding process. Affected by the rule that each word can only have one entered arc in basic dependent syntax, many functional and semantic relationship between content words cannot be indicated directly by the dependent arc and label. At the same time, there is still room for further refinement and improvement of the dependencies in the existing dependency syntax system in order to extract coherent semantic relations from them. In the face of this situation, this paper develops a guidelines of enhanced dependency syntax for Chinese based on the Stanford Dependency Syntax. The main contributions are: prepositions and conjunctions, parallel structures, syntactic alternations and special syntactics. In addition, the paper provides a converter for Chinese enhanced dependency transformation by python, as well as a web-based demo that parses sentences into

* 通讯作者
基金项目：国家语委项目（ZDI135-131）；中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

dependency graphs through universal syntactic dependencies and the specification of this paper. Finally, the paper explores practical applications of augmented dependency syntax such as collocation extraction and information extraction.

Keywords: dependency syntax , enhanced dependencies in Chinese , automatic transform

1 引言

句法分析是自然语言处理当中的关键技术之一，它是对输入文本的句子进行分析以得到其句法结构的过程。依存句法分析是其中的一种表示形式，它用于分析输入句子的句法结构，将词语序列转化为树状的依存结构(李正华, 2013)，来捕获句子内部词语之间的修饰或搭配关系，描述句法结构。依存句法分析广泛应用于自然语言处理的多个领域，如在搭配抽取中，通过大规模的语料进行依存句法分析，从中抽取想要的依存弧以获得具有句法关系的词对，再通过词对之间的共现频次、互信息、联合熵等统计方法来说明词对之间的相关性；再如在信息抽取中，利用依存句法分析来抽取关系三元组，进而达到信息抽取的目的。

依存句法分析在准确地反映句法关系、描述句法结构的同时，也带有一些浅层的语义表示，但语义关系还不够明确，一些实词之间的关系没有直接明确地表示出来，且缺乏对句法转换的抽象。此外，一些依存标签被用于多种情况，难以区分，在自然语言理解的下游任务，如信息抽取、文本挖掘、语义分析中，就需要投入许多工作来处理语法树。因此，研究人员在依存句法的基础上提出了增强依存句法，来满足依存句法反映语义信息的需求。目前，增强依存句法在英语上已获得了有益的探索，并在信息抽取、关系抽取上得到了应用，但在汉语中还未见相关研究。

本文在斯坦福依存句法规范的基础之上，制定了增强的依存句法规范，从利于搭配抽取和自然语言理解的角度重新构建依存图，将实词之间的语义关系显性地展示出来，并统一句式转换中的依存句法关系，以便于进一步的研究和应用。

2 相关研究

斯坦福依存句法框架中提出了几种对句法结构进行面向语义修改的方案，引入了Collapsed Dependencies 和 CCprocessed Dependencies 两种形式(de Marneffe and Manning, 2008)。Collapsed Dependencies 折叠了涉及介词（包括功能类似于介词的多词结构）、连词以及关系从句所指信息的依存关系，从而得到实词之间的直接依存关系，这对于关系抽取应用很有用。此外，该方案还考虑了其他依存关系，如关系子句及其先行词、xsubj 关系和 pobj 关系，甚至破坏了树结构，将依存关系结构转换为有向图。CCprocessed Dependencies 在 Collapsed Dependencies 的基础上，增加了并列词的传播，即当句中存在并列连词时，一个并列词的依存关系可以传播到其他并列词。这样，通过额外增添的和增强的关系，实词之间的关系更加明显，多数涉及实词之间关系的系统通常会采用这两种形式。

通用依存项目（Universal Dependences，简称UD）在第一个版本（UD v1）中(Nivre et al., 2016)同样提出了增强依存（Enhanced Dependencies）的概念，它增加额外的依存关系来表示先行词与关系从句中某个成分之间的主语关系，并在并列词之间传播关系。Schuster 和 Manning(2016)详细描述了增强英语UD（enhanced English UD），并介绍了更适用于自然语言理解任务的增强++表示（enhanced++ representation），对量名词短语和轻名词结构、多词介词、并列的介词或介词短语、关系代词的表示作了改进，并提供了转换器，实现了从基础依存句法（Basic Dependencies）到增强英语UD图和增强++英语UD图的转换。UD V2(Nivre et al., 2020)在先前研究的基础上，定义了五种增强类型：1. 省略谓语的空节点；2.并列项的传播；3. 控制和提升主语；4.关系代词；5. case 信息。

Candito 等人(2017)给出了更进一步的改进，他们沿着两个方向来丰富增强依存框架：扩展非限定性动词的论元依存类型（包括分词、控制名词和形容词、非限定动词以及更多不定式动词的情况）、中和句法转换（包括被动语态，中间被动语态，非人称和使役）。Nivre 等人(2018)评估了向UD现有树库添加增强依存句法的两种跨语言技术，分别是为英语开发的基于规则的系统和在芬兰语、瑞典语和意大利语上训练的数据驱动系统，结果表明，这两种系统都足够精确，可以在现有的 UD 树库中引入增强依存关系。

(c) 2022 中国中文信息学会计算语言学专业委员会

由于英语增强 UD 的转换不支持 Python，且覆盖范围有限，Aryeh 等人(2020)制定了 BART 表示，引入了覆盖范围广的、数据驱动的、语言学上合理的增强依存转化集，包括四种结构的增强：嵌套结构、并列结构、句式转换以及以事件为中心的表示，该转化集使事件结构和许多词汇关系更加明确。此外，他们提供了一个易于使用的开源 Python 库 pyBART[0]，用于将英语 UD 树转换为增强 UD 图或 BART 表示。该库可以作为一个独立的包工作，也可以集成在一个 spaCy 流水线中。当在信息抽取任务中进行评估时，使用增强依存分析结果，可以通过更少的训练样本得到更多的信息，因此BART 表示比增强 UD 产生更高的提取分数。

3 增强的依存句法规范

本文基于斯坦福依存句法，在借鉴英文增强依存句法思想的基础上，制定了增强依存句法标注规范。该规范通过修改依存标签、添加弧或节点的方式，将依存句法树转换为可以表示更多信息的依存句法图，显性地展示实词之间的语义关系，从而更有利于自然语言处理下游任务的应用。

3.1 介词和连词的增强

介词和连词是构造句子时较为常用的词类，对于句意的理解有很大的影响，当一句话中介词或连词发生改变时，句意可能会发生巨大的改变。例如在“我给小王讲了个故事”和“我替小王讲了个故事”这两句话中，只有介词“给”和“替”发生了改变，但句意却完全不同，在前一句话中，动作“讲”的对象是“小王”，而后一句话中，“讲”的对象并没有在句中出现。

在自然语言理解任务中，由于依存句法还带有一定的语义信息，因此常常通过依存句法来识别和提取所需信息，但是，当句中含有介词或连词时，基础依存句法不能完全满足自然语言理解任务中直接通过词之间的依存弧提取信息的需求，因此，需要对介词和连词来进行增强，以更好地适应自然语言理解及其下游任务。

介词的增强 在基础依存句法规范中，当一个介词短语修饰其他实词时，依存弧通常连接在介词短语中的实词和被修饰词上，增强依存句法规范要求把介词添加在该弧的依存标签上，原标签与添加的介词中间用“_”连接，如图1中将该依存弧的依存标签修改为“nmod:prep_向”。这有助于消除介词短语修饰时的歧义，促进实词之间关系的提取，特别是当只通过两个节点之间的依存弧来提取信息时，增强后的依存句法包含的信息更多，更有利于语义理解。

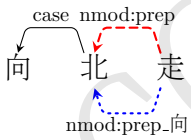


图 1: 介词的增强标注示例

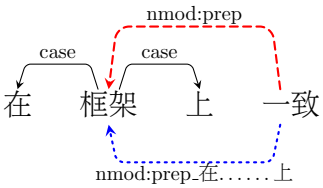


图 2: 框式介词的增强标注示例

除了此类单独出现的介词，汉语中还有一类特殊的介词，即框式介词。刘丹青在《汉语框式介词》一书中最早引入“框式介词”的概念，认为“框式介词是由前置词和后置词构成的使介词支配的成分夹在中间的一种介词类型”(刘丹青, 2002)。在增强依存句法中，用依存弧连接框式介词短语与被框式介词短语修饰的实词时，依存标签中也要把框式介词的两个部分都加上，两个部分中间用省略号连接。如图2中的依存标签“nmod:prep_ 在……上”。

除了 nmod:prep，在被分析为 advcl:loc 等的从句当中，如果从句中存在标签为 case 的依存弧，则在增强依存弧中也要将该弧指向的词添加在连接主句和从句的依存弧 advcl:loc 上。

连词的增强 并列结构是人类语言中最原始最普遍的一种结构式，并列连词可以连接词、短语或小句之间的并列。基础依存句法中用依存标签为 conj 的依存弧来连接句中并列的部分，用依存标签为cc的依存弧连接并列连词与并列项的其中一项。

在增强依存句法规范中，通过在依存标签 conj 上添加依存弧 cc 所指的并列连词，可以使得并列项之间的语义关系更加明晰，特别是当句中出现多个并列连词时，并列结构之间的并列

[0]https://pybart.apps.allenai.org/

(c) 2022 中国中文信息学会计算语言学专业委员会

类型就会更加明确，如图3，将依存标签修改为“conj_和”“conj_或者”，这三组并列结构中并列项之间的关系可以一目了然，计算机在提取并列项间的语义信息时也更加便利。

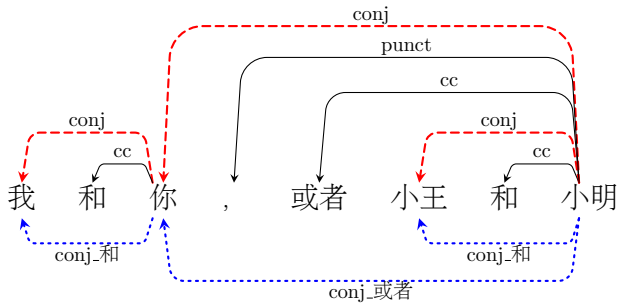


图 3: 连词的增强标注示例

3.2 并列结构的传播

在基础依存句法规范中，多个并列项之间由其中一个并列项作为父节点，来连接其他的句子成分，如主语、宾语。但从语义上来说，并列项之间通常是共享这些句子成分的。因此，在增强依存句法规范中，两个并列的结构共享其父节点和子节点，且依存标签相同。

并列成分的传播 主语、谓语、宾语、时间地点状语等成分在句中都可能由并列结构来承担，在基础依存句法中，只标出其中一个并列项与其支配词和从属词间的依存句法关系，在增强的依存句法图中，需要将并列结构中的其他项与支配词或从属词间的依存关系也表示出来，如图4为并列谓语的增强。

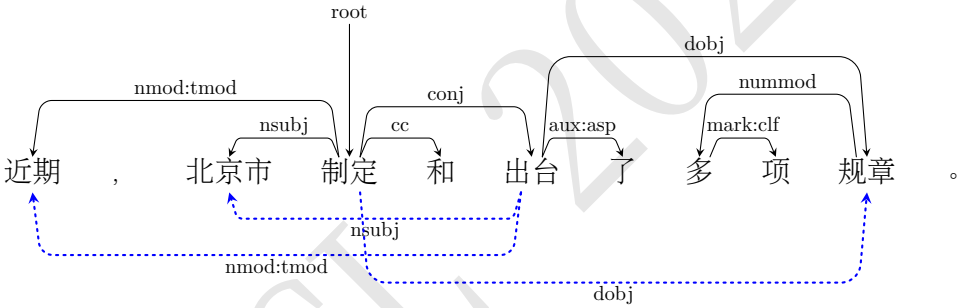


图 4: 并列谓语的传播标注示例

偏正短语中，并列修饰语、状语或中心语也需要传播其支配词或从属词。如图5为并列修饰语修饰中心语的情况，如图6为单个状语修饰并列中心语的情况。这时，在增强依存句法中，就需要补出未被标出修饰关系的修饰语、状语与中心语之间的依存弧。

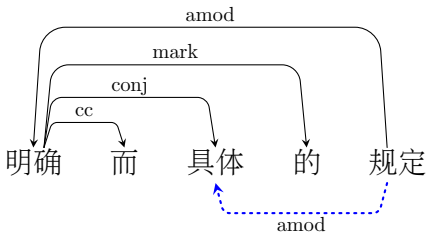


图 5: 并列修饰语的传播标注示例

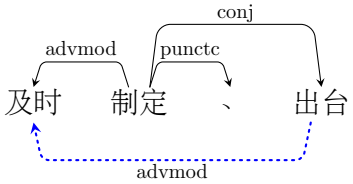


图 6: 并列中心语的传播标注示例

同位语的传播 由于同位语所指代内容相同，在句中承担的句子成分也相同，因此，本文把它看成是一种特殊的并列形式。在基础依存句法中，同位语之间用依存弧 appos 连接，其他句法成分连接在同位语的后一部分上。在增强依存句法中，需要将句中实词与同位语后一部分之间的依存关系，通过增加弧的方式添加在同位语的前一部分上，如图7。

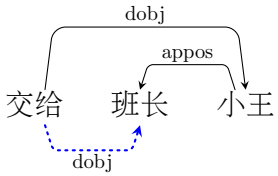


图 7: 同位语的传播标注示例

3.3 句式转换

依存句法是从句子的表层语法来进行分析的，缺乏对句式转换的抽象。同样的语义，采取不同的表述方式，实词之间的依存关系就可能会发生变化。如被动句“书被小王拿走了”和主动句“小王拿走了书”，这两句的句意是完全相同的，但由于句子形式改变，“书”和“小王”之间的依存关系也不同，被动句中，它们之间的关系用 nsubjpass 来表示，意为被动主语，而主动句中，他们之间则用表示宾语的 dobj 来连接。

上述情况对自然语言理解及其下游任务造成了一定的局限，为了使语义分析更简单，本文利用规则统一了句式转换，借助深层的语义关系将不同句式中相同实词间的关系用同样的依存关系来表示。

被动句的转换 在基础依存句法规范中，被动主语，一般为意义上的受事，通常用依存标签为 nsubjpass 的依存弧与谓语连接，意义上的施事主语仍用表示主语的 nsubj 标签与谓语连接，而在被动句转换后的主动句中，施事主语在主动句中形式上做主语，被动主语则作为主动句中的宾语。

为了将被动句与主动句中实词间的依存关系统一，本文采用更为常用的主动句中的依存关系作为标准，即被动主语与谓语之间的依存关系为 dobj。因此，在增强依存句法中，添加一条弧从句中的谓语指向被动主语，依存标签为 dobj，如图8。

此外，修饰成分是被动短语的偏正短语，在基础依存句法规范中，依存弧从中心语指向被动短语中动词，依存标签为 acl，这种表被动的短语在语义上，其中心语通常是被动短语中动词的受事，在转换后的主动句中，中心语是该动词的宾语。因此，在增强依存句法中，添加一条依存弧从被动短语的动词指向中心语，其依存标签为 dobj，如图9。

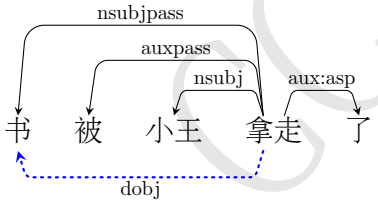


图 8: 被动句的转换标注示例

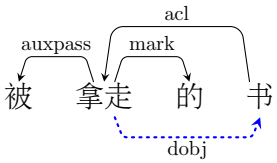


图 9: 被动短语的转换标注示例

有一种比较特殊的被动句，其动词是认作或任选义的动词，如句子“小王被选为班长”、“被誉为‘中国国酒’的茅台酒”，其转换为主动句式为“选小王为班长。”、“誉茅台酒为‘中国国酒’。”，将动词与其后的“为”拆分开来。但在基础依存句法当中，“选为”、“誉为”被当作一个词，难以拆开。面对这种情况，本文尊重了原本的分词及词性规范，在增强的依存句法当中，对此类动词不做特殊考虑。

“把”字句的转换 “把”字句是汉语特有的一种句式，其句式语义主要是主语对动词的受事作了某种处置。“把”是一个介词，它将原来充当动词宾语的受事成分提到动词之前，因此，“把”字句可以通过句式转换将“把”引导的宾语还原到动词宾语的位置。如图10中，“我把苹果吃了。”可以转换为“我吃苹果。”，因此在增强依存句法中增添了一条依存弧从该动词指向“把”引导的宾语，依存标签为 dobj。

(c) 2022 中国中文信息学会计算语言学专业委员会

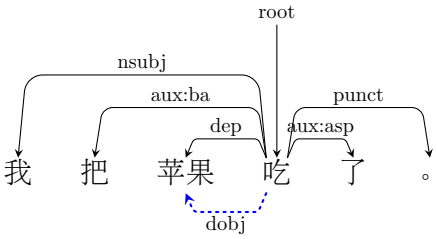


图 10: “把”字句的转换标注示例

形容词修饰语的转换 在偏正短语中，形容词短语来修饰名词中心语，那么这个偏正短语可以转换为以该中心语为主语、以该形容词为谓语的主谓短语，如图11中，“一个漂亮的女孩”可以转换为“女孩漂亮”。为了更好地捕获这些语义信息，在增强的依存句法中，为句子增添了一条从该形容词修饰语指向中心语的依存弧，依存标签为表示主语的 nsubj。

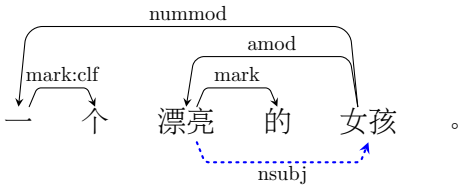


图 11: 形容词修饰的转换标注示例

动词短语修饰语的转换 一个动词短语来修饰名词中心语，如果在动词短语中，该动词不含宾语的话，那么中心语可能为该动词的受事。如图12，在语义上，“饭”是“做”的受事，那么该句可以转化为“妈妈做饭”，此时，“饭”是“做”的宾语。因此，在增强依存句法中，要增加一条依存弧由动词短语中的动词指向中心语，依存标签为 dobj。

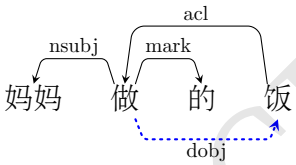


图 12: 动词短语修饰的转换标注示例

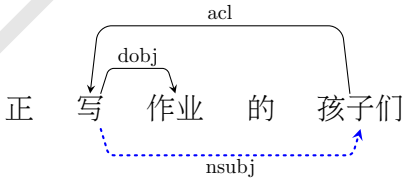


图 13: 动词短语修饰的转换标注示例2

如果修饰名词中心语的动词短语有宾语但不含主语的话，那么这种偏正短语也可能能够转化为一个中心语作主语、动宾短语作谓语和宾语的句子，如图13中“正写作业的孩子们”可以转化为“孩子们正写作业”，此时，“孩子们”为“写”的主语。那么，在增强依存句法中，需要增添从修饰语中的动词指向中心语的依存弧，其依存标签为 nsubj。

在基础依存句法当中，如果动词修饰语既不包含 nsubj 弧，也不包含 dobj 弧，那么其中心语可能是转化后句子的主语，例如“漂泊的游子”转化为“游子漂泊”，也可能是宾语，例如“设置好的页面”转化为“设置页面”，也可能存在修饰语中谓词是动宾结构，但在分词时未拆开的情况，例如“在外打工的父亲”中“打工”被看作是一个词，这些情况本文暂不予考虑。

3.4 特殊句式的增强

兼语句 兼语句是由兼语短语作谓语的句子，其谓语中第一个动词的宾语也是后一谓词的主语(年玉萍, 2003)，这个词就叫做“兼语”。例如在“老师通知我开会”一句中，“我”既是“通知”的宾语，也是“开会”的主语。在基础依存句法当中，受限于一个节点只能有一条入弧的规则，只标注了第一个动词和兼语之间的宾语关系，而没有标注出后一谓词与兼语之间的主语关系。因此，在增强依存句法中，需要增添一条依存弧由后一谓词指向兼语，依存标签为 nsubj，如图14。

(c) 2022 中国中文信息学会计算语言学专业委员会

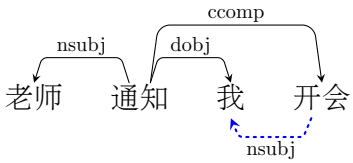


图 14: 兼语句的增强标注示例

连动句 连动句是现代汉语里一种特殊的句法结构，指的是谓语由两个或两个以上动词构成，在动词短语中间没有停顿，也没有关联词语，两个动词短语共用一个主语的句子(刘月华et al., 2001)。如在句子“外商来华投资。”中，“来华投资”是连动短语，它们的主语都为“外商”。但在基础依存句法中，只标注出第一个动词和主语之间的依存关系，因此，在增强依存句法中，应添加一条依存弧由连动短语中的其他动词指向主语，依存标签为 nsubj，如图15。

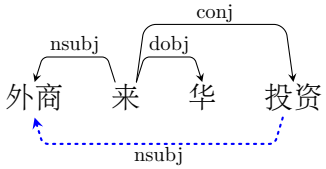


图 15: 连动句的增强标注示例

省略句 中文是一种话题驱动语言，为了表达的连贯性和简洁性，句子中常常省略某些语言成分，即句子存在缺省，本文讨论对句子中的主要结构即主语、宾语省略的增强。

含有动词性状语的句子中，存在状语中的动词和谓词共用一个主语的现象，由于汉语中的经济原则，那么状语或主句就可能省略主语。如图16 中，时间状语中省略了主语，但其实“吃饭”和“散步”的主语都为“他”。在基础依存句法中，只标出了“他”与“散步”之间的主语关系。那么在增强的依存句法中，还需要添加一条依存弧由“吃完”指向“他”，依存标签为 nsubj。

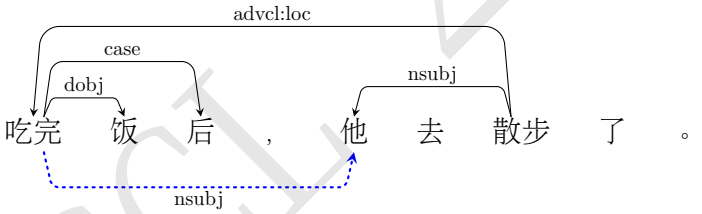


图 16: 省略主语的增强标注示例

在复句中，除了省略小句主语的情况以外，当几个小句的宾语相同时，也可能会省略小句中的宾语。如图17 中，第二个小句中没有宾语，但根据语义可知，其宾语仍为第一个小句中的宾语“小明”。因此在增强依存句法中，需要增添一条依存弧由省略宾语小句中的谓词“看见”指向其他小句中的宾语“小明”，依存关系标签为 dobj。

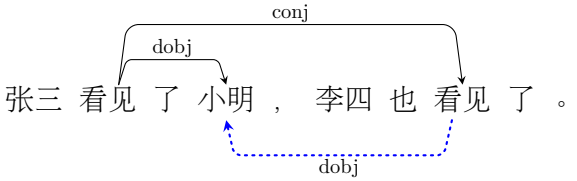


图 17: 省略宾语的增强标注示例

3.5 不确定情况的处理

在上述的规则中，也会产生一些不适应的情况，如句子“正睡觉的时候，妈妈回来了。”，如果按照增强依存句法规则，会把缺少主语小句中谓词“睡觉”的主语指向另一小句中的主语“妈

(c) 2022 中国中文信息学会计算语言学专业委员会

妈”，但是依照现实情况来看，“睡觉”的主语不可能是“妈妈”，其真正的主语需要联系上下文来确定。面对这些情况，本文并未放弃这几类增强规则，而是如图18，借用 Aryeh(2020)提出的 UNC=TRUE（不确定）这一概念，表示这条依存弧的正确性由用户来判断。

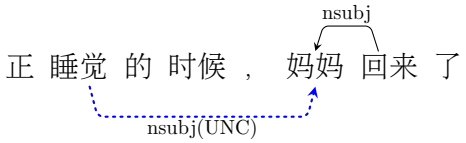


图 18: 不确定情况的处理示例1

同样的，汉语当中也存在复句中的某个小句省略了主语，但其省略的主语不是其他小句主语的情况，例如在“然而外祖母又怕都是孩子们，不可靠。”一句中，“不可靠”的主语是前一小句的宾语“孩子们”，而非前一小句的主语“外祖母”；再例如在“春游的时候，他告诉了我这件事。”一句中，“春游”的主语可能是“他”，也可能是“我”，也可能“他”和“我”都是主语，这需要根据句子的上下文来决定。此时，本文采取 Aryeh(2020)提出的概念 ALT=X，表示用户可以从中选择其一，如图19，其中 X 表示被省略主语或宾语的词在句子中的位置。

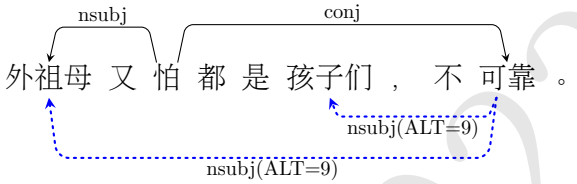


图 19: 不确定情况的处理示例2

3.6 依存句法增强器与演示平台

在斯坦福依存句法规范的基础上，本文提供了一个可以从基础依存句法分析到增强依存句法分析的转换器。在观察大量依存标注语料的基础上，寻找每类规则的规律，利用词性、依存弧的范围和指向、依存标签等约束实现了增强依存句法规范的规则转化。

此外，还提供了汉语依存句法增强转换在线平台[1]，如图20，可以将句子分析为基础依存句法和增强依存句法，并将它们可视化，便于比较和分析。

该界面分为四个部分，分别为输入句子搜索、选择示例搜索、基础句法依存演示、增强依存句法演示。用户可以在输入框中自主输入想要分析的句子，也可以在选择示例下拉框中选择，平台已经为17个汉语增强依存句法规则给出了示例演示。

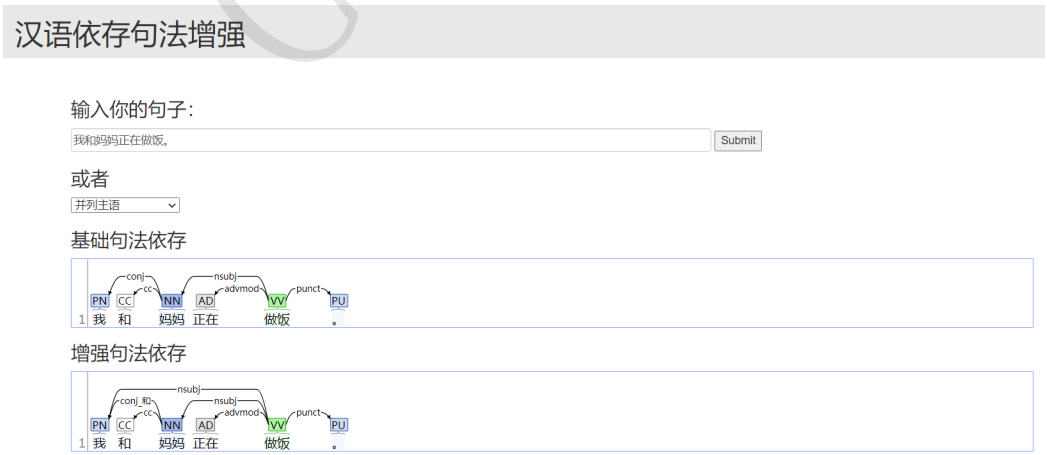


图 20: 汉语依存句法增强转换在线平台

[1]https://parser.litmind.ink

(c) 2022 中国中文信息学会计算语言学专业委员会

4 增强依存句法规范的实际应用

增强依存句法在基础依存句法的基础上扩充了实词间的依存关系，包含的句法和语义信息更多，在需要使用依存句法的任务中，就能更快速直接全面地获取所需要的信息。本小节从搭配检索、信息抽取两个方面来说明汉语增强依存句法规范在语料库检索中的实际应用。

4.1 搭配抽取中的应用

搭配通常是指两个或两个以上的词语所组成的一种语言表示，这种表示往往是某种语言习惯的表达(邵艳秋et al., 2019)。通过在语料库中抽取搭配，一方面便于汉语学习者检索自己所用搭配是否准确、常用，有利于学习者自学；另一方面也便于对外汉语教师和研究人员建立搭配库，通过检索某个词的常用搭配及其例句方便教学和语言本体的研究。此外，搭配也能支持自动翻译、信息检索、自动问答等应用研究。

依靠人工判断搭配费时费力，不仅主观性强，而且耗时巨大。随着计算机技术的发展，搭配抽取技术也有了长足的进步。目前，一种比较好的方法是基于依存句法分析的搭配自动抽取。通过依存弧来抽取搭配时，需要明确依存关系表示的搭配关系。例如，规定 nsubj 表示主谓搭配关系，dobj 表示动宾搭配关系，advmop:dvp 表示状中搭配关系，compound:nn 表示定中搭配关系，那么在图21所示句子中，通过基础依存句法抽取到的搭配如表1 所示。

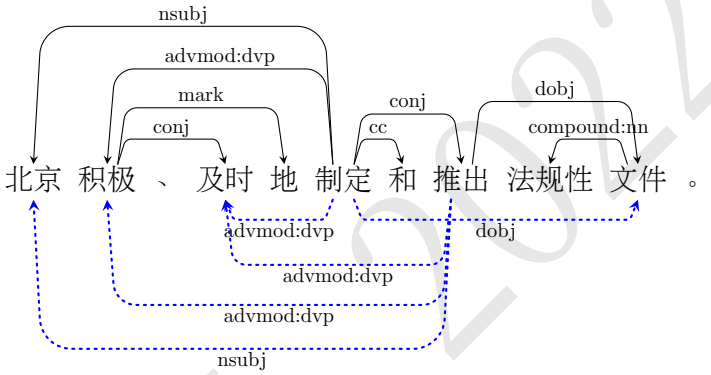


图 21: 依存句法标注示例

搭配类型	抽取结果
主谓搭配	北京制定
动宾搭配	推出文件
状中搭配	积极制定
定中搭配	法规性文件

表 1: 利用基础依存抽取到的搭配

但若对抽取到的搭配进行人工校对就会发现，由于并列情况的存在，通过基础依存句法只能抽取到并列项其中之一的搭配关系，而忽略了其他并列项的搭配。增强依存句法就能很好地解决这个问题，它把并列项之间的依存关系都通过添加依存弧的方式展现出来，用增强依存句法来抽取搭配就能找回那些被遗漏的搭配。这种全面的搭配抽取方式一方面能帮助学习者在用例句学习搭配时找到句中所有搭配，明确可使用的搭配；另一方面，可以扩大搭配库，便于后续的统计处理和研究工作，即使原始语料库较小，也能抽取出更多的搭配范式。如图21例句中，利用增强依存句法还能抽取出的搭配如表2。

(c) 2022 中国中文信息学会计算语言学专业委员会

搭配类型	抽取结果
主谓搭配	北京推出
动宾搭配	制定文件
状中搭配	及时制定 积极推出 及时推出

表 2: 利用依存句法重现的遗漏搭配

4.2 信息抽取中的应用

信息抽取的主要功能是从非结构化的文本中自动提取用户感兴趣的结构化事件信息，是各项自然语言处理任务例如知识图谱构建、翻译、篇章理解等应用的基石(项威and 王邦, 2020)。目前，信息抽取主要包括以下命名实体识别、指代消解、关系抽取以及事件抽取等几个方面的研究(张素香, 2007)。其中，比较常见的一种方法是利用依存句法来抽取信息。

用基础依存句法在检索平台中进行信息抽取时，如果句中存在大量信息嵌套和成分共享、抽取的信息之间有多层依存弧或存在句式转换的现象时，就需要对不同的情况建立多种抽取模式，甚至可能存在信息漏抽或抽取错误的情况。

例如“小王今年25岁，来自北京。”一句中，由于第二小句缺省主语，直接运用依存句法抽取主谓宾不能抽取到“小王来自北京”这一信息，必须对依存句法树进行一定的处理才能得到。运用增强依存句法之后，就能直接得到这些实词之间的语义关系，在信息抽取中无需花费大量的时间和精力处理句法树，这在句中存在并列结构、成分省略和转换句式时尤为明显。

5 总结

本文基于汉语基础依存句法制定了增强依存句法规范，使得句中尽可能多的实词间的语义关系更加清晰明确。此外，本文还提供了汉语增强依存句法转换的 Python 转换器以及方便进行可视化比较的 Web Demo，并给出了该规范在搭配抽取和信息抽取中的实际应用，以说明该规范在这些任务中的优势。

未来，还应进一步完善和补充汉语增强依存句法体系，以满足规模更大、句子更长、结构更复杂的语料。目前增强依存句法规范在汉语特殊句式中只考虑到了比较常见的一部分，之后还需要将判断句、倒装句等句式纳入到增强依存句法体系中来。此外，面对不确定的情况的处理，也可以更好地进行分类讨论，例如当复句中省略宾语时，如该谓语动词为不及物动词，那么不添加该谓语动词与其他小句成分间表示宾语的依存弧，因此，就需要对谓语动词进行及物和不及物的分类处理。最后，还应进一步探索其应用场景，找到更多适合其发挥的任务，挖掘其更大的优势。

参考文献

项威and 王邦. 2020. 中文事件抽取研究综述. 计算机技术与发展, 2(20):1–6.

刘月华, 潘文娱, and 故韡 . 2001. 实用现代汉语语法. 商务印书馆.

邵艳秋, 申资卓, and 刘世军. 2019. 基于依存搭配抽取技术的平面媒体语言监测研究. 山西大学学报:自然科学版, 3(42):526–533.

刘丹青. 2002. 汉语中的框式介词. 当代语言学, 4:241–253+316.

年玉萍. 2003. 谈谈兼语句. 延安教育学院学报, 1:40–42.

张素香. 2007. 信息抽取中关键技术的研究. Ph.D. thesis, 北京邮电大学.

李正华. 2013. 汉语依存句法分析关键技术研究. Ph.D. thesis, 哈尔滨工业大学.

Marie Candito, Bruno Guillaume, Guy Perrier, and DjaméSeddah. 2017. Enhanced ud dependencies with neutralized diathesis alternation. In *Proceedings of the Depling 2017-Fourth International Conference on Dependency Linguistics*.

(c) 2022 中国中文信息学会计算语言学专业委员会

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing universal dependency treebanks: A case study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis M. Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'20)*.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Aryeh Tiktinsky, Yoav Goldberg, and Reut Tsarfaty. 2020. pybart: Evidence-based syntactic transformations for ie. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

名动词多能性指数研究及词类标记的组合应用

周姣美
北京师范大学国际中文教育学院/
北京市海淀区新街口外大街19号
100875
zhoujiaomei@outlook.com

杨丽姣
北京师范大学国际中文教育学院/
北京市海淀区新街口外大街19号
100875
yanglijiao@bnu.edu.cn

肖航
教育部语言文字应用研究所/
北京市朝阳门内南小街51号
100010
exiaohang@163.com

摘要

名动词是汉语词类研究及词性标注的难点问题。过去五六十年来，有关名动词概念与体系类属、鉴定标准、标注方法等方面的争议不断，但基于语料库资源，以名动词的动态分布以及量化研究为支撑的研究较为缺乏。本文以现代汉语名动词作为主要考察对象，基于语言学理论方法的反思，将信息论与语料库方法相结合，引入香农-维纳指数作为量化指标，从多能性指数的研究视角对名动词进行考察，结合《信息处理用现代汉语词类标记规范》的修订研究，分析了名动词类别属性判断在现有印欧系语法词类体系框架下的困境，探讨了名动词跨类属性、词类标记的组合处理及其对于语料库建设、词典编纂等应用领域词类信息标注的探索意义。

关键词： 名动词；词语多能性指数；词类；标准修订

A study of nominal verb polyfunctionality index and the combined application of POS tag

Jiaomei Zhou
Chinese Language and Culture College
Beijing Normal University/
No. 19, Xinjiekou Outer Street
100875, Beijing
zhoujiaomei@outlook.com

Lijiao Yang
Chinese Language and Culture College
Beijing Normal University/
No. 19, Xinjiekou Outer Street
100875, Beijing
yanglijiao@bnu.edu.cn

Hang Xiao
Institute of Applied Linguistice Ministry of Education/
No.51, South Small Street, Chaoyangmennei
100010, Beijing
exiaohang@163.com

Abstract

Nominal verbs are a difficult problem in Chinese part-of-speech (POS) research and tagging. In the past fifty to sixty years, there have been many controversies about the concept and system, identification criteria and annotation methods of nominal verbs, but there is a lack of studies based on corpus resources, supported by the dynamic distribution of nominal verbs and quantitative studies. This paper takes modern Chinese nominal verbs as the main object of investigation, combines information theory with corpus methods based on the reflection of linguistic theoretical methods, introduces the Shannon-Wiener index as a quantitative index, and examines nominal verbs

第二十一届中国计算语言学大会论文集，第110页-第119页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

from the perspective of polyfunctionality index, combines with the study of Revision for Standard of POS Tag of Contemporary Chinese for CIP, analyzes the dilemma of nominal verb attribute judgment in the framework of the existing Indo-European grammatical POS system, and discusses the combined treatment of nominal verb co-category attributes and part-of-speech tags and their implications for the exploration of part-of-speech information tagging in the fields of corpus construction and lexicon compilation.

Keywords: nominal verb , word polyfunctionality index , part-of-speech , revision for standard of POS tag

1 引言

现代汉语中有些动词具有名词的语法性质，尽管在词典中被标记为动词，在具体语境中往往难以判断其究竟是名词还是动词，如“发展”“研究”等词。这类词并非个例，其语法功能与普通动词或普通名词相比均有较大区别，而规模却有扩大的趋势。从二十世纪五六十年代至今，学界对这一现象争议不断，出现了名动词、动词次类、动名兼类词、动词名物化、动词名用、动名漂移、动名跨类等诸多不同的观点。理论语言学上的争议与计算机中文信息处理的词性标注难点密切相关，尽管多个语料库及分词工具的词类标注体系均对这一词汇聚合规定了特定的标记，但并不能很好地解决实际操作中标注一致性差的难题。在本文中，为讨论方便，仍然把这类词汇聚合称为名动词。

《信息处理用现代汉语词类标记规范》是现代汉语信息处理系统中的一个重要参考词类标记集，它吸收了众多语言学研究成果，从现代汉语信息处理的实际需求出发，提供了一套统一的现代汉语词类标记体系。2020年发布的“《信息处理用现代汉语词类标记规范》修订稿”专门讨论了对名动词现象的处理，提出以组合标记“n_v”的形式标注名动词 (杨丽姣等, 2021)。

本文以现代汉语名动词作为主要研究对象，基于语言学理论方法的反思，将信息论方法、语料库语言学方法相结合，引入香农-维纳指数作为量化指标，衡量汉语名动词语法功能的灵活程度，从多能性指数这一新的视角对名动词传统难题进行考察。在此基础上，进一步讨论了面向信息处理用的词性标记组合设置以及名动词组合标记的可操作性问题。

2 相关研究

关于名动词现象，一直存在概念与体系类属、鉴定标准、标注方法等诸多争议。

黎锦熙(1960)提出动词名物化的说法，他认为一部分用作主宾语的动词失去动词的特点，获得了名词的语法特点，包括可以受定语修饰、可以用名词或代词复指、可以和名词组成联合结构等。朱德熙(1961; 1982)反对上述名物化说法，他认为动词作主语或宾语的时候仍然是动词。他提出了名动词的说法，将名动词看作动词的一个小类，定义为兼有名词性质的动词。陆俭明(1994)、俞士汶(2005)、陆丙甫(2009)赞同名动词的说法，不主张将其处理为动名兼类词，并对名动词的数量、语法特点作了进一步补充。胡明扬(1995)、郭锐(2002; 2011)、黄昌宁(2009)则认为这种现象可以处理成动名兼类，吕叔湘(1979)提出了动词名用的说法，认为这类现象中词义无明显变化，语法特点有所改变，与动名兼类不同。沈家煊(2007; 2009; 2012)认为对于名动词的词性区别应该淡化处理。在大的词类体系上，可以认为名词包含了动词。夏全胜等(2014)从心理语言学和神经语言学视角分析名动词语义加工机制，认为名动词与动词或名词的加工机制有所不同。

对于名动词的鉴定，概括起来主要有四种标准：语法功能标准、指称性标准、意义标准、数量标准，其中以功能标准为主，部分标准是否有效仍存在争议。朱德熙(1961; 1982; 1985)、胡明扬(1995)、俞士汶(2002)采用若干条功能标准判断名动词，如可作动词“有”的宾语、可充任“进行、加以”这一类动词的准谓词性宾语、可受名量词修饰等。郭锐(2002)结合数量标准和功能标准判断名动词，将这类词在所有动词中所占比例作为一个考量因素。陆丙甫(2009)根据名动词作宾语时的指称化程度判断其词性。安华林(2005)引入意义标准，如“研究”这类词没有语义转类，没有明确的名词性义项，可与“检查”等有明确的名词性义项的词区分开。

对于名动词的标注，朱德熙提出NV标记，并进一步将NV细分为NVt和NVi两个小类，NVt表示及物的名动词，NVi表示不及物的名动词。俞士汶(2002)在语料库标注规范中

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

将名动词标注为vn，黄昌宁(2009)认为应根据名动词所在语法位置标注其词性，出现在述语位置上标作动词（v），出现在名词的语法位置上标作名词（n）。杨丽姣等(2019; 2021)在《信息处理用现代汉语词类标记规范》修订方案提出了组合标记，用组合标记（n_v）来标注名动词，并将这类词处理为动名跨类词。

3 语料标注与词表提取

3.1 名动词标注语料库构建

本研究从国家语委现代汉语通用平衡语料库中选取600篇文本，包含散文、新闻、公文等多种语体，总规模达100万字。在对语料进行观察之后，本文认为单独为名动词设定组合标记存在其合理性。在实际使用时，名动词既有动词的功能，又有名词的功能，然而在词典中却没有名词的词性标记。标注语料时如果完全依据一个词的语境和功能进行标注，将引起词类系统的混乱，因为某些词的用法并未固定下来，因此将组合标记作为一种过渡用标记具有可行性和实用性。本文依照“《信息处理用现代汉语词类标记》修订稿”中的现代汉语词类标记设置，结合名动词的鉴定标准，组织语言学及应用语言学专业研究生10人，采用两两对照的方式对600篇文章中的名动词进行精标注，最终得到现代汉语精标注语料库。标注规则如下：

a.对于名物化的动词，应标注为名动词，若在句中没有被名物化，则仍按动词标注；根据名动词的定义，对于在句子中根据句法、语用功能都不能明确为指称（reference）还是述谓（predication）的词，则标注为名动词。

b.参考以下四条依据，符合其中任意一条标准则考虑该动词是否应标注为名动词：能受名量词修饰，充当动词“有”以及部分形式动词（“进行”、“作”、“加以”、“予以”、“给予”、“给以”等）的宾语，进入“有没有X”、“N的V”框架，充当体词性短语的中心语。

c.名动词的标注符号为n_v。

以下为名动词的标注示例：

• 东方/l 各国/n 随着/p 社会/n 经济/n 的/u 发展1/n_v ，/w 纷纷/d 探索/v 东方/l 文化/n 作为/v 独立/a 的/u 文化/n 生存/n_v 体/n_g 的/u 可能性/n 问题/n ，/w 本/rd 民族/n 文化/n 逐渐/d 成为/v 社会/n 发展2/n_v 的/u 精神支柱/n 。

• 我们/rp 一定/d 要/vu 根据/p 现在/t 的/u 有利/a 条件/n 加速/v 发展3/v 生产力/n。

发展1/n_v：“社会经济的发展”为“N的V”结构，此处“发展1”被名物化；

发展2/n_v：“社会发展”为“精神支柱”的定语，“发展2”是名词性短语的中心语；

发展3/v：“发展生产力”为动宾结构，此处“发展3”是动词；

《现代汉语词典》（第7版）中，“发展”词条如下：【发展】fāzhǎn动①事物由小到大、由简单到复杂、由低级到高级的变化：事态还在～，社会～规律。②扩大(组织、规模等)：～党的组织，～轻纺工业。③为扩大组织而吸收新的成员：～新党员，～工会会员。词典中“发展”只有动词词性，没有名词词性，但在实际使用中有名词的语义和语法特点。在标注过程中，我们发现类似“发展”的情况广泛存在于语料中。

3.2 名动词词表构建

首先提取语料中所有具有名动词标记（n_v）的词语，去重处理后得到初步词表；然后参考《现代汉语词典》（第7版）中的词性标记，删去已确定具有名词词性的词；最后，人工对筛选出的词语进行增删。

在词性标注的基础上，本研究对语料中的名动词进行初步提取，用Python和正则表达式提取出所有具有n_v、n和v标记的词，得到三个列表，去重并计数，取n_v列表和v列表的交集，此交集再与n列表取差集，得到初步的名动词表，共包含1002个词。

初步提取的词表中，有780词的n_v标记数量在3次及以下，在100万字的语料中出现频次过低，因此将这一部分词删去，剩余223个词；再查阅《现代汉语词典》（第7版），删去“全球化”、“用地”、“收费”3个未收录的词，对照词典中的词性标记，保留没有名词词性的词语，删去“认识”、“贷款”等37个动名兼类词；最后，人工筛查剩余词语，删去“国有”、“航天”2个人工标注错误的词、“惊喜”、“恐惧”、“烫伤”3个没有动词词性的词，所得词表如下表 1所示，共178个词。

(c) 2022 中国中文信息学会计算语言学专业委员会

发展	合作	展览	审美	休息	申请	交往	补充	启迪
研究	交接	治疗	死亡	预测	胜利	考察	成长	上市
服务	进步	刺激	探索	增长	实施	控制	促进	设立
医疗	提高	防雷	形成	转换	束缚	利用	发挥	探讨
学习	应用	观测	安排	表达	损害	流行	繁衍	体现
演出	重视	欢迎	护理	结合	体验	批评	繁殖	伪造
调查	出现	竞争	注意	伤害	威胁	破坏	分布	向往
发现	观察	理解	追求	生存	指导	认知	奉献	消化
消费	危害	使用	创造	宣传	制作	失败	革新	写作
实践	支持	处理	到来	转变	帮助	统治	关怀	信任
开发	关注	仲裁	发行	成功	保健	喜爱	观赏	选举
努力	污染	抽查	交流	出版	表意	想象	函授	循环
培训	准备	调整	介绍	诞生	尝试	预防	合成	延伸
笑	表现	发射	考验	独立	冲突	愈合	呼唤	运算
保护	冲击	分析	浪费	防治	搭配	运输	恢复	增加
表演	训练	改变	试验	分化	反射	运行	计算	直播
发生	进展	感悟	思考	解释	分裂	装扮	解剖	制约
管理	了解	革命	统一	经营	关心	综合	描述	贮存
改革	生长	记载	突破	精炼	监测	尊敬	配合	
生产	选择	旅游	享受	考试	检测	尊重	普及	

表 1: 现代汉语名动词表

4 名动词多能性指数计算方法

4.1 “香农-维纳”指数

Shannon和Weaver(1948; 1998)提出香农-维纳指数（Shannon–Wiener index），奠定了信息论的基础。香农-维纳指数也被称作香农物种多样性、香农熵、香农信息指数或H等，用于度量信息的不确定性，是基于现实世界的概率描述下的一种信息量的度量方式。香农-维纳指数的计算方法描述如下：对于一种信号源A，其发出的信号U有n种，每种信号对应出现的概率pi，那么信号源A的不确定性为单个信号出现的不确定性的统计平均值，该值就是A的香农-维纳指数。为了方便观察和计算，需要对香农-维纳指数进行标准化，计算公式如下所示：

$$H_{norm} = \frac{-\sum_{i=1}^{n}(p_i \cdot \ln p_i)}{\ln n} \tag{1}$$

标准化后的香农-维纳指数取值区间为[0，1]，H值越小，不确定性越小，H值越大，不确定性越大。

4.2 名动词多能性指数计算方法

4.2.1 词语多能性

Hieber(2020; 2021)提出用香农-维纳指数来计算英语和Nuuchahnulth语[1] 的词语多能性。词语多能性（lexical polyfunctionality）指一个词项（lexical item）具有不止一个话语功能或句法功能，如述谓（predication）、指称（reference）或修饰（modification），与之相对应的则是动词、名词、形容词。

如果一个词具有述谓、指称、修饰三个功能，并且在使用时的频率完全相等，那么可以称之为完全多能词（perfectly polyfunctional lexical item），例如，一个词在某个语料库中出现300次，三种功能各有100次；与之相反的是，如果这个词在某个语料库所出现的300次中，均作指称（reference）用法，则为完全单功能词（perfectly monofunctional lexical item）。对于在使用中具有多种句法功能的词，需要一个可以衡量一个词汇究竟有多“多能（polyfunctional）”的度量指标。

[1]加拿大西部的一种早期语言。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2.2 名动词的多能性

本研究3.2中提取的名动词在词典中大部分只有动词词性，小部分兼有动词和形容词词性，具有动词的典型语法特征。然而在实际使用中，基本都有名词的功能，体现出了跨类的特征。

在3.1的标注语料库中，本文分别统计出名动词作名词用法（标记为n_v）和作动词（标记为v）用法的次数。表 2为随机抽取的5个在《现代汉语词典》（第7版）中只有动词词性的名动词。按词频由高到低依次排列，“发展”是语料库中最高频的名动词，“合成”为低频名动词。5个词的n_v标记所占比例平均为32.3%，接近一半的“研究”被标注为名词词性，“发展”和“服务”两个词也有35%以上的名动词标记，词频较低的“分析”和“合成”二词的名动词标记比例相对高频词较低，分别为17.9%和21.1%。 在3.2的名动词表中，有10个词在《现代汉语词典》中标

	n_v数量	v数量	合计	n_v比例
发展	216	356	572	37.9%
研究	206	244	450	46.0%
服务	45	70	115	39.1%
分析	10	46	56	17.9%
合成	4	15	19	21.1%

表 2: 随机抽取的名动词词性标注情况

注有动词和形容词两种词性，这10个词分别为“努力、进步、应用、统一、成功、独立、精炼、失败、尊敬、尊重”。在语料中，这些词不仅有动词和形容词的功能，还有名词的功能，即动形兼类名动词。表 3为这些动形兼类词的标注情况。词频最低的词是“精炼”，在语料中出现了8次，其中2次被标注为名动词，1次被标注为动词，5次被标注为形容词；词频最高的词是“成功”，在语料中出现了92次，其中27次被标注为名动词，12次被标注为动词，53次被标注为形容词。从三种标记所占比例来看，10个词的n_v标记、a标记、v标记平均比例分别为27.3%、26.5%、46.2%。因此，在语料中这类词将近一半的用法仍为动词词性，名动词标记和形容词标记比例相当，前者略高于后者。

名动词	n_v数量	v数量	a数量	合计	n_v比例	a比例	v比例
努力	30	43	1	74	40.5%	1.4%	58.1%
进步	31	18	14	63	49.2%	22.2%	28.6%
应用	22	53	12	87	25.3%	13.8%	60.9%
统一	8	21	17	46	17.4%	37.0%	45.7%
成功	27	12	53	92	29.3%	57.6%	13.0%
独立	6	27	16	49	12.2%	32.7%	55.1%
精炼	2	1	5	8	25.0%	62.5%	12.5%
失败	6	14	1	21	28.6%	4.8%	66.7%
尊敬	5	9	7	21	23.8%	33.3%	42.9%
尊重	5	18	0	23	21.7%	0.0%	78.3%

表 3: 名动词表中动形兼类词的标注情况

4.2.3 名动词多能性指数计算方法

本文基于Hieber(2020; 2021)的词语多能性指数计算方法，运用香农-维纳指数来衡量现代汉语名动词的多能性，具体计算方法描述如下：

一个名动词在语料中有可能出现三种标记：n_v、v、a，对应该词语在不同语境中的不同词性：名动词、动词、形容词，每个词性标记的频率为：

$$p_i = \frac{n_i}{N} \tag{2}$$

其中，N为该词在语料库中的词频，ni为该词在语料中被标注为某一词类i的数量。

(c) 2022 中国中文信息学会计算语言学专业委员会

可以计算出该词的香农-维纳指数（H）如下式，H的取值区间为[0，1]，H值越小，不确定性越小，H值越大，不确定性越大。换言之，名动词作为一个具有名词性质的动词，该词的H值越小，它的多功能性就越小，H值越大，它的多功能性就越大。

$$H_{norm} = \frac{-\sum_{i=1}^{n}(p_i \cdot \ln p_i)}{\ln n} \tag{3}$$

其中n为名动词在语料库中被标注的词类种数，对于名动词而言，n=3。除了动形兼类词外的名动词没有a标记，这一部分词的概率等于0，然而ln0不存在。针对这种情况有两种常用方法可以解决 (Stefan Th. Gries, 2009)：一种是把三种标记的数量都加1，另一种方法是简单地认为ln0等于0。本研究采用前一种方法，在计算H时将除动形兼类词外的名动词的三种标记数量各加1，例如，表 2中的“发展”的n_v、v、a标记实际数量为215、355、0，经过处理后为216、356、1。

5 名动词多能性指数分析

名动词的多能性指数如下表 4所示，根据词频由高到低排列178个名动词，计算出的多能性指数值越大，表示这个词语的多能性越强，其取值区间为[0,1]。例如“发展”的多能性指数约为0.61，“形成”的多能性指数约为0.22，前者的多能性比后者更强。观察n_v和v标记数量，“发展”标注为n_v和v的次数分别为216次和356次，“形成”则是10次和186次，在数量和比例上来看，“发展”比“形成”更灵活更多能，与多能性指数得出的结论相符。 在名动词表中，有一部分

序号	词语	n_v数量	v数量	a数量	多能性指数（H）
1	发展	216	356	0	0.614171
2	研究	206	244	0	0.64062
3	发现	31	265	0	0.329753
4	学习	39	222	0	0.409249
5	出现	16	225	0	0.254059
6	使用	12	207	0	0.228963
7	发生	19	199	0	0.30238
8	形成	10	186	0	0.223093
9	生产	18	167	0	0.32769
.....					
169	繁衍	4	6	0	0.81752
170	分化	6	3	0	0.808014
171	表意	6	3	0	0.808014
172	关怀	4	4	0	0.850864
173	函授	4	4	0	0.850864
174	束缚	6	2	0	0.782776
175	信任	4	3	0	0.858673
176	直播	4	3	0	0.858673
177	启迪	4	2	0	0.852792
178	精炼	2	1	5	0.819448

表 4: 名动词多能性指数

词有动词、形容词、名词三种词性的功能，即动形兼类名动词。下表 5为10个动形兼类名动词的多能性指数，由表可知，这些词的H值均在0.6以上，其中“尊敬”一词的H值高达0.975，表明这是一个功能相当灵活的词语。从多能性指数的平均值来看（表 6），动形兼类词的H值平均为0.825，附录中所有名动词的多能性指数平均值为0.585，比前者低0.24。因此，动形兼类词的平均多能性比名动词强。

为更直观地观察多能性指数（H）在每个区间的分布情况，本文将计算所得的H值按照区间来统计其数量，如在区间[0,0.1]的H有0个，在区间(0.1,0.2]的H有2个，依次类推，将所有名动词与动形兼类的名动词的多能性指数分布情况统计如图 1和图 2所示，横坐标为10个多能

(c) 2022 中国中文信息学会计算语言学专业委员会

序号	词语	n_v数量	v数量	a数量	多能性指数（H）
1	努力	30	43	1	0.673248
2	进步	31	18	14	0.947666
3	应用	22	53	12	0.840001
4	统一	8	21	17	0.937593
5	成功	27	12	53	0.85852
6	独立	6	27	16	0.865649
7	精炼	2	1	5	0.819448
8	失败	6	14	1	0.703815
9	尊敬	5	9	7	0.974883
10	尊重	5	18	0	0.630712

表 5: 动形兼类名动词的多能性指数

	平均多能性指数
所有名动词	0.585
动形兼类的名动词	0.825

表 6: 词语多能性指数的平均值

性指数的区间，如0.1对应区间(0,0.1]，1对应区间(0.9,1]，纵坐标为在该区间H的数量。名动词作为一个具有名词性质的动词，该词的H值越小，它的多功能性就越小，H值越大，它的多功能性就越大。一般而言，名动词的H值越大，就说明其名词性程度越深。例如“利用”的H值约为0.16，“形成”的H值约为0.22，“生产”的H值约为0.33，“管理”的H值约为0.40，“演出”的H值约为0.59，“服务”的H值约为0.65，“实践”的H值约为0.70，“胜利”的H值约为0.81，以上8个词的H值分别位于8个不同的区间，它们的名词性逐渐增加，多能性越来越强。

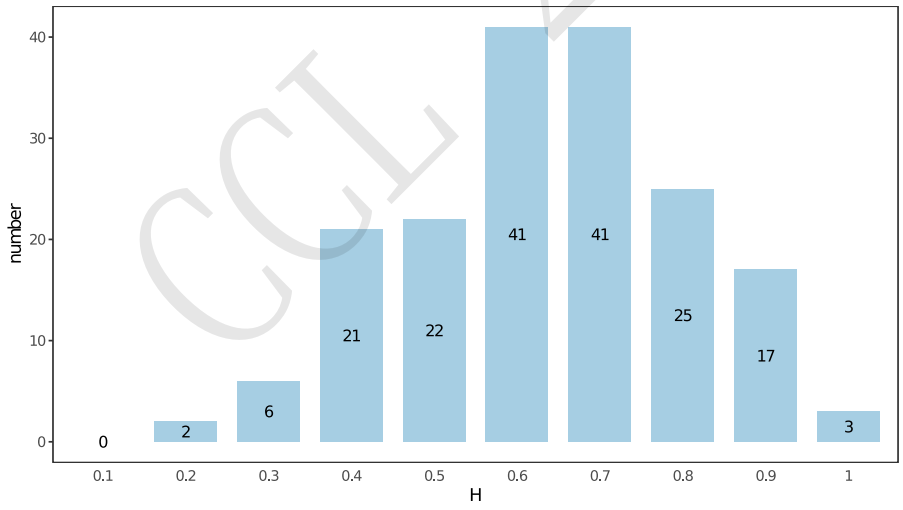


图 1: 名动词的多能性指数（H）数量分布

从图 1总体的分布趋势来看，H值在0.6以前时，名动词的数量随着H的增大而增加，H值到0.6、0.7时数量最多，0.7之后数量又逐渐减少；从数据的峰值来看，名动词的数量在(0.5,0.7]这个H值区间内最多，低于0.3或高于0.9的名动词数量都比较少，即多能性在中等偏高水平的名动词最多，多能性较差或较好的名动词相对更少。因此，名动词的总体多能性比较强，功能特别灵活和功能特别不灵活的名动词较少。

如图 2所示，动形兼类名动词的多能性指数数量分布与图 1不同，集中在强多能性区域。没有一个动形兼类名动词的H值低于0.6，H值最多分布在0.8及以上，这与表 6中计算的多能性

(c) 2022 中国中文信息学会计算语言学专业委员会

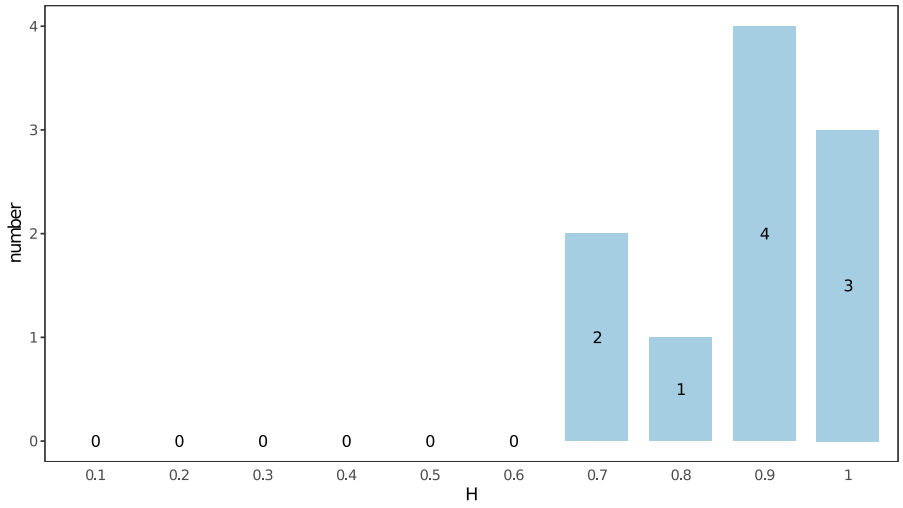


图 2: 动形兼类名动词的多能性指数（H）数量分布

指数平均值特点一致，动形兼类的名动词比普通名动词的功能更灵活，这主要是因为动形兼类名动词有动词、形容词和名词三种功能，所以这类词的多能性总体更强。

多能性指数值越大，表示该词语的多能性越强，人们在不同语境中使用该词语的不同词性就越灵活。结合多能性指数对词表中的名动词进行分析，本文发现这些名动词的性质存在差异，某些名动词的多能性更强，某些名动词的多能性较弱，多能性指数可以较好地反映名动词在语料中的使用情况。此外，多能性指数也可为名动词的辨别和处理提供参考。

6 临时活用、兼类、跨类的多能性指数对比分析

多能性指数可应用于其他现象的分析，如临时活用和兼类词等，为词类研究提供新视角。《现代汉语词典》（第7版）中存在大量动名兼类词，即在词典中标注有动词和名词两种词性的词，如“生活”、“工作”等。本小节计算前文精标注语料库动名兼类词n标记和v标记的数量，根据4.2中的词语多能性计算方法，计算语料中88个动名兼类词的多能性指数。表 7为词频最高的十个动名兼类词的多能性指数，H值最低约为0.07，最高最约0.64。

动名兼类词	n数量	v数量	合计	多能性指数
生活	338	55	393	0.383708
工作	242	88	330	0.544968
教育	72	173	245	0.573052
作用	217	2	219	0.073539
要求	66	102	168	0.63926
关系	153	13	166	0.281812
活动	103	39	142	0.594252
组织	92	50	142	0.624316
领导	92	16	108	0.425816
运动	77	9	86	0.358718

表 7: 动名兼类词的多能性指数

从平均多能性指数来看，名动词与动名兼类词分别约为0.585和0.584，动名兼类词总体上没有比名动词多能性更强。此外，两个词表都存在多能性很弱的词，例如动名兼类词“作用”的H值为0.07，语料库中标记为n的有217个，标记为v的仅有2个。名动词“利用”的H值为0.16，语料库中标注为v的有174个，而标注为n_v的仅有5个。因此“作用”的动词用法和“利用”的动词用法可看作是一种临时活用现象。

(c) 2022 中国中文信息学会计算语言学专业委员会

对于那些多能性比较弱的名动词，如“增加”、“利用”等，我们认为不应看作是典型的名动词，即不应处理为动名跨类词；对于多能性较强的动名兼类词，如“工作”、“活动”、“报告”等，其性质与名动词接近，可将其处理为跨类词。

7 总结及讨论

7.1 总结

名动词是汉语词类研究及词性标注的难点问题，目前基于语料库资源，以名动词的动态分布以及量化研究为支撑的研究较为缺乏。本文构建了一个规模较大的名动词词表，将信息论与语料库方法相结合，引入香农-维纳指数作为量化指标，从多能性指数的研究视角对名动词进行考察，并探讨动词的临时活用、兼类、跨类情况。

首先，计算每个名动词的三种标记的概率分布，从而判断该词语的用法情况。研究发现，大部分名动词仍然以动词用法为主，小部分名动词的用法已经偏向于名词词性。然而，仍缺乏一个统一的标准来衡量这些词语究竟有多灵活，即多功能性有多强。因此，本文选用基于“香农-维纳”指数的词语多能性指数来衡量名动词的多能性，多能性指数值越大（最高值为1），表示该词多能性越强，功能越灵活；多能性指数越小（最小值为0），表示该词多能性越弱，功能越不灵活。根据词语多能性指数计算结果，名动词的平均多能性指数为0.585，多能性较强。动形兼类的名动词属于名动词中功能更灵活的一类，平均多能性指数为0.825，在语境中兼有动词、名词、形容词的不同功能。

除名动词外，本文运用多能性指数对动词的临时活用、兼类、跨类情况进行探讨，对于那些多能性比较弱的名动词，如“增加”、“利用”等，不应将其看作是典型的名动词，即不应处理为动名跨类词；对于多能性较强的动名兼类词，如“工作”、“活动”、“报告”等，其性质与名动词接近，可将其处理为跨类词。

7.2 词类范畴、词类标记的组合应用及可操作性问题

从根本上说，名动词理论上的莫衷一是、词性标注上的困境与汉语自身的特点及汉语词类分析所依凭的印欧系语法框架有关。

沈家煊(2012)认为，“名动词”对汉语来说某种程度上是个伪概念，朱德熙先生因英语“V-ing”与名动词对当而引入的这一特定概念，从大量汉语的事实出发是难以成立的。更严重的是，名动词现象对于现代汉语词类体系的内部一致性构成了挑战。

汉语词类问题早在上世纪五十年代就有过几次大的讨论，核心成果是认为词类主要是依据语法功能原则进行划分的，这已经成为学界的基本共识。然而在印欧语系语法中动词与名词这一基本的对立性范畴，在汉语这种非形态语言中始终存在种种扭曲，从语义表达功能上看，名词的典型功能在于指称，动词的典型功能在于陈述，可是在具体语境中，名动词可视为指称性的抽象陈述，或者既表陈述也表指称，在现有词类概念体系或印欧系语法话语体系下，名动词在应用层面的混淆与歧义，恐怕难以消除。以《现代汉语词典》（第7版）为例，在词性上标为动词，而配比的例句中表示指称性意义的例句比比皆是。如，爱戴：动词。敬爱并且拥护：受到人民群众的爱戴。

杨丽姣、肖航等(2021)认为名动词不是兼类词，语境信息不能完全消除歧义；不是名词或动词次类，其语法语义功能与其有较大的差别。在“《信息处理用现代汉语词类标记规范》修订稿”中，杨丽姣、肖航等提出用组合标记n_v来标注名动词，实际上将这类词处理为跨类词或者是共类词，n_v是为提升汉语语言信息处理用词类标记体系的内部一致性，在处理名动词这类特殊语言现象时一种阶段性尝试。名动词问题的根本解决需要突破现有理论语言学框架，一旦系统性的汉语理论语言学话语范式得到确立和被广泛采用，汉语词词类体系以及词类概念必然进一步推陈出新。

多能性指数研究可为汉语词类研究提供新视角，为词类标记的组合应用提供量化指标。名动词多能性指数体现了词语功能的灵活性，在此基础上探讨名动词跨类属性和词类标记的组合处理具有可操作性，也有助于语料库和词表资源建设、词典编纂等应用领域的研究。

致谢

本文受到国家语委“十三五”科研规划2020年度省部级重点项目“面向国际中文教育的文本可读性智能评价方法研究及分析系统构建”(ZDI135-41)的资助。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Daniel W. Hieber. 2020. *Lexical flexibility: Expanding the empirical coverage.* https://files.danielhieber.com/publications/ucsb-doctoral-colloquium/slides.pdf.

Daniel W. Hieber. 2021. *Lexical polyfunctionality in discourse: A quantitative corpus-based approach.* University of California, Santa Barbara.

Stefan Th. Gries. 2009. *Statistics for Linguistics with R: A Practical Introduction.* De Gruyter Mouton, Berlin, New York.

Shannon C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379-423.

Shannon, C. E. and Weaver, W. 1998. *The mathematical theory of communication reprinted.* Urbana: University of Illinois Press.

2006. GB/T 20532-2006 信息处理用现代汉语词类标记规范. 中国标准出版社, 北京.

安华林. 2005. 从两种词表看名,动,形兼类的处理. 语言教学与研究, (4): 31-39.

郭锐. 2002. 《现代汉语词类研究》. 商务印书馆, 北京.

郭锐. 2011. 朱德熙先生的汉语词类研究. 汉语学习, (5): 14-26.

胡明扬. 1995. 动名兼类的计量考察. 语言研究, (2): 91-99.

胡明扬. 1996. 词类问题考察. 北京语言学院出版社, 北京.

黄昌宁, 李玉梅, 靳光瑾. 2009. 动名兼类词及其词性标注规则. 张普、王铁琨主编《中国语言资源论丛》（一）,商务印书馆, 北京.

黎锦熙, 刘世儒. 1960. 语法再研讨——词类区分和名词问题. 中国语文, (1): 5-8.

黎锦熙. 1992. 新著国语文法. 商务印书馆, 北京.

陆俭明. 1994. 关于词的兼类问题. 中国语文, (1): 28-34.

陆丙甫. 2009. 基于宾语指称性强弱的及物动词分类. 外国语*(*上海外国语大学学报*)*, (6): 18-26.

吕叔湘. 1979. 汉语语法分析问题. 商务印书馆, 北京.

沈家煊. 2007. 汉语里的名词和动词. 汉藏语学报, (1): 27-47.

沈家煊. 2009. 我看汉语的词类. 语言科学, 8(1): 1-12.

沈家煊. 2012. "名动词"的反思:问题和对策. 世界汉语教学, (1): 3-17.

夏全胜, 彭刚, 石锋. 2014. 汉语名词, 动词和动名兼类词语义加工的偏侧化现象——来自*ERP* 的研究. 心理科学, 37(6): 1333.

杨丽姣, 肖航, 刘智颖. 2019. 《信息处理用现代汉语词类标记规范》修订方案. 语言文字应用, (3): 87-95.

杨丽姣, 肖航, 刘智颖. 2021. 《信息处理用现代汉语词类标记规范》修订研究. 语言文字应用, 119(3): 111-120.

俞士汶, 段慧明, 朱学锋. 2002. 北京大学现代汉语语料库基本加工规范. 中文信息学报, 16(5): 51-66.

俞士汶, 段慧明, 朱学峰. 2005. 词语兼类暨动词向名词飘移现象的计量分析. 孙茂松、陈群秀主编《自然语言理解与大规模内容计算》, 清华大学出版社, 北京.

中国社会科学院语言研究所词典室. 2005. 现代汉语词典.第7版. 商务印书馆, 北京.

朱德熙, 甲文, 马真. 1961. 关于动词形容词"名物化"的问题. 北京大学学报*:* 哲学社会科学版, (4): 53-66.

朱德熙. 1982. 《语法讲义》. 商务印书馆, 北京.

朱德熙. 1985. 现代书面汉语里的虚化动词和名动词为第一届国际汉语教学讨论会而作. 北京大学学报*:* 哲学社会科学版, 世界汉语教学学会, (5): 3-8.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于新闻图式结构的篇章功能语用识别方法

杜梦琦[1]，蒋峰[1]，褚晓敏[1]，李培峰[1,2]
[1]苏州大学计算机科学与技术学院，苏州，中国
[2]苏州大学人工智能研究院，苏州，中国
{20205227068，20194027003}@stu.suda.edu.cn，{xmchu，pfli}@suda.edu.cn

摘要

篇章分析是自然语言处理领域的研究热点和重点，篇章功能语用研究旨在分析篇章单元在篇章中的功能和作用，有助于深入理解篇章的主题和内容。目前篇章分析研究以形式语法为主，而篇章作为一个整体的语义单位，其功能和语义却没有引起足够重视。已有功能语用研究以面向事件抽取任务为主，并未进行通用领域的功能语用研究。鉴于功能语用研究的重要性和研究现状，本文提出了基于新闻图式结构的篇章功能语用识别方法来识别篇章功能语用。该方法在获取段落交互信息的同时又融入了篇章的新闻图式结构信息，并结合段落所在篇章中的位置信息，从而有效地提高了篇章功能语用的识别能力。在汉语宏观篇章树库的实验结果证明，本文提出的方法优于所有基准系统。

关键词： 篇章分析；篇章功能语用；新闻图式结构

Discourse Functional Pragmatics Recognition Based on News Schemata

Mengqi Du[1], Feng Jiang[1], Xiaomin Chu[1], Peifeng Li[1,2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]AI Research Institute, Soochow University, Suzhou, China
{20205227068，20194027003}@stu.suda.edu.cn，{xmchu，pfli}@suda.edu.cn

Abstract

Discourse analysis is a hot topic in the field of natural language processing. The purpose of discourse functional pragmatics research is to analyze the function and role of discourse units, which is helpful to deeply understand the theme and content of discourse. At present, discourse analysis mainly focuses on formal grammar, but the function and semantics of discourse as a whole semantic unit have not attracted enough attention. The existing functional pragmatics researches are mainly oriented to event extraction task, but there is no general functional pragmatics research. In view of the importance and status of functional pragmatics research, this paper proposes a Functional Pragmatics Recognition Method Based on News Schemata(FPRNS). FPRNS not only obtains the interaction information of paragraphs, but also incorporates the information of news schemata and the location information of paragraphs, so as to effectively improve the recognition ability of discourse functional pragmatics. The experimental results in the Chinese macro discourse tree-bank show that the proposed method is superior to all baselines.

Keywords: Discourse Analysis , Discourse Functional Pragmatics , News Schemata

第二十一届中国计算语言学大会论文集，第120页-第131页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

当代语言学有两大流派，分别是形式主义和功能主义。目前自然语言处理领域的相关研究大都基于形式语法，这是由于形式语法的中心任务是研究语法成分之间的形式关系，运用鲜明的数理符号表示，促进了语言处理的可计算化。而功能语法以功能和语义为导向，将篇章作为一个语言使用单位，比较难以进行抽象化，因此目前篇章分析领域缺乏针对篇章整体的功能语用研究。

功能语用研究旨在分析篇章单元在篇章中所承担的角色和所起的作用，有助于挖掘篇章中具有价值的信息，深入理解篇章的主题和表达的含义，可以应用于自然语言处理中的其他任务，包括问答系统、信息抽取、作文自动评分等。

在功能语法的研究方面，Halliday (1994)创立了系统功能语言学，他明确地指出，系统功能语言学在本质上是“功能的”和“语义的”，而不是“形式的”和“句法的”，系统功能语言学的研究对象是“语篇”，而不是“句子”。Van Dijk根据新闻报道的结构和功能特点，提出了新闻图式理论(Van Dijk, 1988)。该理论将篇章研究和媒体研究有机结合起来，集中论述了新闻报道的语用结构，为篇章功能语用的研究奠定了基础。近些年来，研究者基于该理论标注了一系列语料资源。在英文方面，Yarlott et al. (2018)标注了来自ACE2语料库的50篇文章段落的功能语用，但该语料规模较小；Choubey et al. (2020)根据新闻图式理论中的功能语用类型做出调整，定义了8种功能语用类型，标注了802篇新闻中句子的功能语用，但其定义的功能语用类型只适用于事件抽取任务。在中文方面，Chu (2018)在新闻图式理论的基础上将细颗粒度的功能语用类型进行合并，补充了新闻图式理论中没有但在实际新闻报道中出现的功能语用类型，共定义了18种功能语用类型。在此基础上Jiang et al. (2018)标注了规模为720篇的宏观篇章语料库（MCDTB），为中文篇章功能语用研究奠定了基础。

P1:记者日前从云南省民政厅获悉，根据中国、老挝和联合国难民署三方达成的遣返在华老挝难民协议，云南顺利完成了十二批、二千九百一十七人的遣返任务，遣返人数占原在华老挝难民的百分之七十三以上。
P2:根据国务院的统一部署，自一九七八年以来，云南先后接收安置难民六万四千一百余人，其中老挝难民到一九九一年已达三千九百九十四人。依照中国政府缔结的有关国际公约，云南省对难民实行大量的国际主义和人道主义援助。十五年来，在全省尚有四十多个贫困县接受政府财政补助的情况下，已累计支付二点六亿元专为难民解决生产生活困难。
P3:国际社会公认，难民自愿遣返到原籍国是永久解决难民问题的最佳方案。随着中、老两国关系的不断改善，老挝政府主动表示愿意接收在华老挝难民回国，遣返难民的条件日趋成熟。一九九一年四月和七月，中老两国政府正式签订了《关于遣返在华老挝难民的议定书》和《备忘录》，分别委托中国云南省政府和老挝南塔省政府具体负责组织实施。
P4:在遣返过程中，云南省政府对难民的生活极为重视，并同老挝政府和联合国难民署密切配合，使遣返工作得以顺利进行。
P5:据悉，云南尚余一千零七十五名老挝难民，在难民完全自愿前提下，云南省政府将继续积极稳妥进行遣返。（完）

例1 chtb_0255

本文以中文篇章功能语用为研究对象，在MCDTB的基础上开展篇章功能语用的探索与研究。本文以MCDTB中的一篇文章(chtb_0255)来说明篇章的功能语用结构，文章内容如例1所示。其中，段落P1为全文“导语”，阐述了全文的主要内容“云南顺利完成了在华老挝难民的遣返任务”，P2阐述了近年来云南省对难民实行大量的人道主义援助，介绍了文中事件发生的“背景”，P1和P2两个段落形成文章的“总述”部分。P3、P4分别讲述了老挝政府愿意接受难民回国以及在遣返过程中各方密切配合使得遣返任务顺利进行，是文章的“情景”部分，两个段落组成全文的“故事”，同时P5补充了云南省将继续对剩余老挝难民的遣返工作，是全文内容的“补充”。

例1可用图1所示篇章功能语用结构树表示。其中，叶子节点表示新闻报道的段落P1到P5的

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家自然科学基金(61836007,61773276);江苏高校优势学科建设工程资助项目

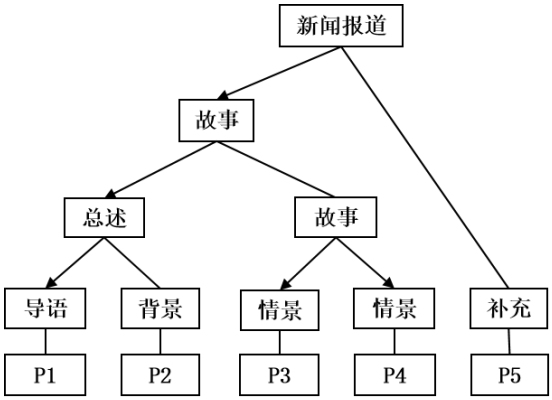


图 1. chtb_0255篇章功能语用结构树

功能语用；非叶子节点表示由其下层篇章单元组成的篇章单元在篇章中的功能语用；根节点表示整个篇章的功能语用。该结构树中的箭头指向重要的篇章单元。

新闻图式理论(Van Dijk, 1988)明确指出单个段落在整篇文章发挥着重要的作用，因此本文将篇章功能语用的识别建模成识别篇章中每个段落的功能语用的任务。结合新闻图式理论，本文分析了新华社的新闻报道，发现新闻中的每一个段落从属于一个更大的功能语用范畴，例如图1中“导语”、“背景”属于“总述”这个功能语用范畴，通过对文章进行范畴划分有助于读者先从整体上理解文章含义，进而更好地把握每一个范畴中段落的功能语用，这样的文章理解思路在Zhang (2014)的语言学研究中也有相应的论述。另一方面，我们研究了新闻报道本身的特点，比较规范的新闻报道文本，第一段功能语用常为“导语”，交代新闻事件的要素，最后一段则常常是对整篇新闻报道的“评论”或“补充”，帮助读者厘清事件的意义或补充事件后续发展。因此，段落在篇章中的位置信息对于识别篇章的功能语用具有重要的作用。

为了深入研究篇章功能语用，本文提出了一个基于新闻图式结构的篇章功能语用识别模型。该模型首先将段落通过XLNet获得段落初步编码，然后通过指针网络获得段落交互信息和篇章范畴划分信息，从而获得篇章组织结构信息。此外，该模型又结合段落在篇章中的位置信息，获得了更加丰富的段落表示，从而有效地提高了篇章功能语用的识别能力。在MCDTB的实验结果表明，本文提出的模型对篇章功能语用的识别有很好的效果。

2 相关工作

目前自然语言处理领域的相关研究大都基于形式语法，以研究语言的结构和形式为主要任务，这是由于形式语法运用鲜明的数理符号表示，而功能语法将篇章作为一个语言使用单位，比较难以进行抽象化和形式化，因此目前篇章分析领域缺乏针对篇章整体的功能语用研究。

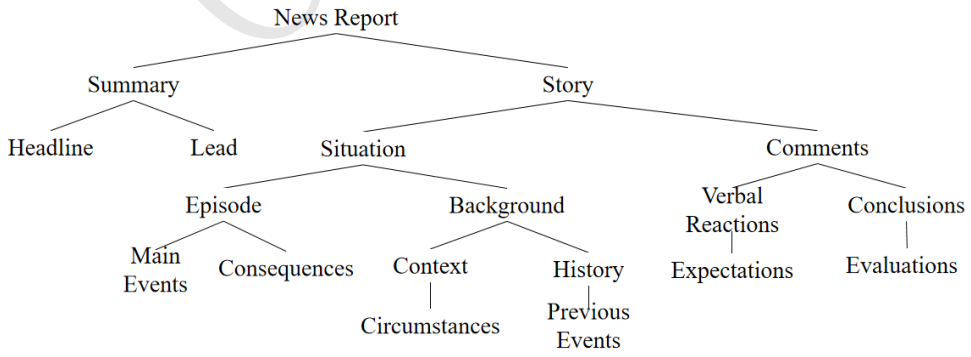


图 2. 假拟新闻图式结构

现有的涉及篇章功能语用的理论包括Van Dijk的新闻图式理论(Van Dijk, 1988)和Pan and Kosicki (1993) 提出的基于框架的方法。新闻图式理论(Van Dijk, 1988)将篇章研究和媒体研究有机结合起来，集中论述了新闻报道的语用结构。图2是一个假拟的新闻图式结构，该结构包含

(c) 2022 中国中文信息学会计算语言学专业委员会

总述和故事两个最主要的部分。其中，总述（Summary）由标题（Headline）和导语（Lead）两个部分组成，故事（Story）由情景（Situation）与评论（Comments）组成，新闻图式结构通过一个从上而下的层级顺序清晰地展示了新闻篇章的整体组织形式。Pan and Kosicki (1993)提出的基于框架的方法从四个维度分析新闻篇章：句法结构、脚本结构、主题结构和修辞结构，其中句法结构与新闻图式理论最为相似。

针对英文篇章功能语用的研究，Liddy (1991)、Kircz (1991)和Teufel et al. (1999)利用修辞结构和论元类型来定义功能类型并且为科学文章创建了语料库；Mizuta et al. (2006)、Wilbur et al. (2006)、de Waard et al. (2009)和Liakata et al. (2012)利用多种注释模式，对生物领域的功能语用进行了研究。但这些研究针对的是科学论文和生物领域，不适用于其他领域。一些研究者在新闻图式理论基础上，标注了事件领域的语料资源。例如，Yarlott et al. (2018)标注了来自ACE2（Automated Content Extraction Phase2）语料库的50篇文章，并且分别使用SVM、决策树和随机森林等传统机器学习方法做了验证性的实验。而Banisakher et al. (2020)在Yarlott的基础上，利用CRF模型对每个段落的功能语用进行预测，提升了段落功能语用的识别性能。Choubey et al. (2020)将新闻图式理论中的功能语用类型进行相应的调整，定义了8种功能语用类型，标注了802篇文章中句子的功能语用，并提出了一个两层的双向LSTM对文章中每个句子的功能语用进行预测。Choubey and Huang (2021)提出了一个演员-评论家框架来识别句子功能语用，该模型使用了多个评论家，评论家根据已知的子话题结构采取行动，而演员模型的目标是超越评论家，并且引入了一个分层神经网络建模句子、子话题和文档之间的交互。在上述的研究中，Yarlott et al. (2018)标注的语料规模较小，且标注者之间的Kappa值（55%）并不高；Choubey et al. (2020)定义的功能语用类型面向事件，不适用于其他领域。

而中文方面，Song et al. (2020)标注了1230篇作文中句子的功能语用，共定义了7种功能语用类型，并基于序列标注思想，结合句子在作文中的位置信息对句子功能进行预测，但是其标注的功能语用类型是面向学生议论文的，不适用于其他领域；Chu et al. (2018)和Chu (2018)参考宏观架构理论和新闻图式理论提出了一套宏观篇章结构表示体系，在此基础上Jiang et al. (2018)标注了720篇新闻报道，形成宏观汉语篇章树库（Macro Chinese Discourse Treebank, MCDTB）。在MCDTB语料库上，已有的工作都是针对篇章逻辑语义的研究(Jiang et al., 2021; Sun et al., 2020)，对篇章功能语用的研究只进行了初步尝试，研究较少。

3 任务介绍

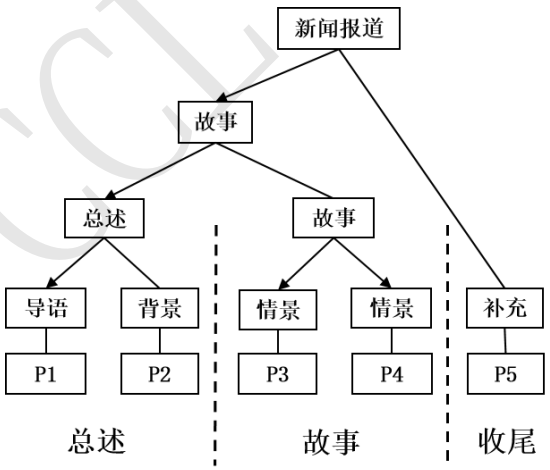


图 3. 新闻报道的范畴划分

一些语言学家在新闻图式理论的基础上又做了进一步阐述。例如，Norman (1995)将新闻图式结构中的“情节”和“评论”更名为“附属”和“收尾”，认为新闻应该由“总述-附属-收尾”三个范畴组成；Bell and Garrett (1998)认为新闻框架应该由“属性-概述-故事”三个范畴组成，有一些新闻还包含“补充”，如背景、评论和追踪报道等，其中“概述”包括“标题”和“导语”，相当于新闻图式结构中的“总述”；Zhang (2014)认为新闻篇章应该由“新闻摘要-新闻故事-新闻评议”三个范畴组成，其中“新闻摘要”对应新闻图式理论的“总述”，“新闻故事”对应“情节”，“新闻评

(c) 2022 中国中文信息学会计算语言学专业委员会

议”对应“评论”。综合上述理论体系研究和对MCDTB中样本的分析，本文认为“总述-故事-收尾”是新闻报道必备的功能语用结构，因此本文将新闻报道划分为“总述-故事-收尾”三个范畴。

基于上述理论研究和样本分析，本文在MCDTB语料上进行功能语用的范畴划分，形式如图3所示。具体划分方法：（1）将“导语”段以及对“导语”进行解释说明的段落（例如，背景段、补充段等）标注为“总述”；（2）将新闻篇章描述的主体事件标注为“故事”，该部分包括事件主要情节的详细阐述、原因分析、数据支撑等；（3）将对新闻报道事件进行评价、总结或者补充的段落标注为“收尾”。其中，“故事”是新闻的主体部分，是每篇新闻报道的必要部分，而“总述”和“收尾”则可以省略。经过语料处理，48.1%的新闻包含三部分，46.8%包含“总述-故事”两个部分，3.1%包含“故事-收尾”两个部分，还有2.1%仅由“故事”组成。

新闻图式理论(Van Dijk, 1988)指出单个段落在整篇文章发挥着重要的作用，因此本文将段落功能语用作为研究对象，将篇章功能语用识别任务专注于识别叶子节点的功能语用。具体而言，对于给定的一篇文章$T = \{P_1, P_2, ..., P_m\}$，篇章功能语用识别任务就是通过模型识别出段落的功能语用$T_{Fun} = \{Fun_1, Fun_2, ..., Fun_m\}$。本文使用准确率来评判模型对于段落功能语用的识别能力。以chtb_0255为例，正确的段落功能语用为{导语，背景，情景，情景，补充}，若模型的预测结果为{导语，补充，情景，情景，补充}，那么模型预测的准确率为4/5=80%。

4 FPRNS模型

本文提出了一种基于新闻图式结构的篇章功能语用识别模型（Functional Pragmatics Recognition Based on News Schemata，FPRNS），如图4所示。该模型主要由四部分组成：1）文本编码层（Text Encoding Layer）；2）范畴划分层（Category Segmentation Layer）；3）范畴识别层（Category Classifier Layer）；4）信息融合层（Information Fusion Layer）。

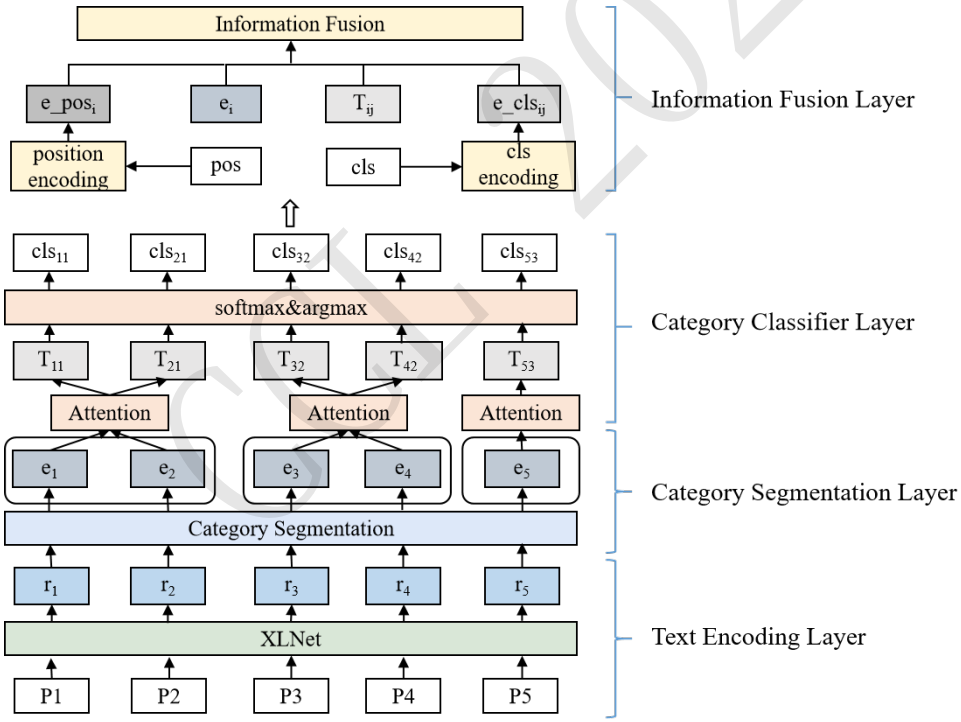


图 4. 篇章功能语用识别模型图

4.1 文本编码层（Text Encoding Layer）

文本编码层使用XLNet作为编码器对段落内容进行编码得到段落向量表示。假设一篇文章含有m个段落，段落序列为$T = \{P_1, P_2, ..., P_m\}$，将每段使用分隔符“SEP”将段落区分开，得到$T' = \{P_1, SEP, P_2, SEP, ..., P_m, SEP, CLS\}$。由于XLNet可处理的数据最大长度为1024，所以需要对段落序列T'进行处理。如果序列长度超过最大长度，则比较序列中每个段落的长度，将最长段落末尾的字符截掉，直至序列长度为最大长度。然后使用XLNet对段落序列T'进

行编码，取每个段落编码最后一个词的词向量作为该段落的语义表示，得到段落语义表示序列$T_r = \{r_1, r_2, ..., r_m\}$。

4.2 范畴划分层（Category Segmentation Layer）

范畴划分层采用指针网络对篇章中功能语用范畴进行划分。具体模型如图5所示。

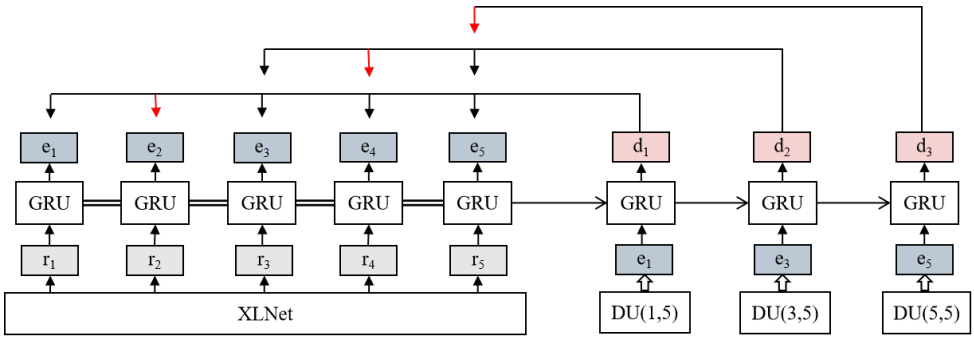


图 5. 范畴划分层模型图

4.2.1 编码层

Chung et al. (2014)的研究表明，GRU(Cho et al., 2014)和LSTM在很多任务上的性能不分伯仲，但是GRU拥有更少的参数，容易收敛，因此在编码层本文使用Bi-GRU进行编码。以chtb_0255为例，本文将段落序列T $= \{P_1, P_2, P_3, P_4, P_5\}$经过文本编码层得到将段落语义表示$T_r$，将$T_r$输入到Bi-GRU中，得到具有交互信息的段落语义表示序列$T_e = \{e_1, e_2, e_3, e_4, e_5\}$，其中$e_i = [e_i^f, e_i^b]$。$e_i^f$,$e_i^b$分别表示正向和反向的输出。

4.2.2 解码层

在解码层采用的也是一个GRU。本文将编码层的输出$T_e = \{e_1, e_2, e_3, e_4, e_5\}$作为解码层的输入。假设在第t步解码时，篇章单元队列为DU(l,5)，解码层会综合当前篇章单元的队头语义表示e_l和t步之前生成的篇章单元语义信息生成当前状态d_t。d_t和当前篇章单元语义信息$T_{e(l,5)} = \{e_l, e_{l+1}, ..., e_5\}$进行交互，通过softmax层得到关于$T_{e(l,5)}$的概率分布。其中$\sigma(.,.)$是融合当前状态表示和篇章单元交互信息，具体为点积运算；α_t为关于$T_{e(l,5)}$的概率分布。如公式（1）所示。

$$\begin{aligned} s_{t,i} &= \sigma(d_t, e_i)\ , i = l, ..., 5 \\ \alpha_t &= softmax(s_{t,i}) \end{aligned} \tag{1}$$

如果通过softmax层后e_i被分配的概率值越大，表明段落P_i和P_{i+1}之间的语义联系越松散，因此这两个段落应该分属于两个功能语用范畴中，从而将篇章划分为两个篇章单元DU(l,i)和DU(i+1,5)。每一步解码，将划分后的两个篇章单元的后者继续放入队列，递归地对篇章单元进行切分，直至队空，解码过程如图6所示。

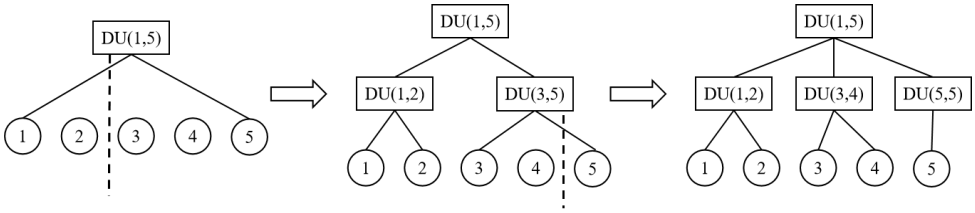


图 6. 解码过程

由章节3可知，每篇新闻报道最多只有三个范畴，但是在实际识别的时候可能会识别出三个以上的范畴。对于这种情况本文比较切分点的概率值，选取前两个概率值较大的切分点作为最终的切分点的位置。

(c) 2022 中国中文信息学会计算语言学专业委员会

本文采用负对数似然作为损失函数，记为L_1，如公式（2）所示，其中$y_{<t}$表示在第t步解码之前产生的篇章单元，θ为可训练的参数。

$$L_1(\theta) = -\sum_{i=1}^{batch}\sum_{t=1}^{T} logP(y_t|y_{<t}, X) \tag{2}$$

4.3 范畴识别层（Category Classifier Layer）

经过范畴划分层，得到三个范畴，每个范畴包含的段落分别为$\text{Topic}_1 = \{\text{P}_1, \text{P}_2\}$，$\text{Topic}_2 = \{\text{P}_3, \text{P}_4\}$，$\text{Topic}_3 = \{\text{P}_5\}$，范畴语义表示分别为$Topic_{e1} = \{e_1, e_2\}$，$Topic_{e2} = \{e_3, e_4\}$，$Topic_{e3} = \{e_5\}$。将具有上下文交互信息的段落语义表示$e_i$通过注意力机制（即范畴中每个段落语义信息的加权和）来获得范畴的语义表示，如公式（3）所示，将段落所在范畴语义表示序列记为$T_{topic} = \{T_{11}, T_{21}, T_{32}, T_{42}, T_{53}\}$，其中$T_{ij}$表示第i段对应第j个范畴。

$$T_{ij} = Attention(Topic_{ej}) \tag{3}$$

将T_{topic}送入到softmax层获取范畴所属类别的概率分布，从而得到每个范畴具体的类别信息，如公式（4）所示。将范畴类别序列记为$T_{topic_cls} = \{cls_{11}, cls_{21}, cls_{32}, cls_{42}, cls_{53}\}$,其中$cls_{ij}$表示第i段对应第j个范畴类别。

$$cls_{ij} = argmax(softmax(T_{ij})) \tag{4}$$

本文同样采用负对数似然作为范畴识别层的损失函数，记为L_2，如公式（5）所示。其中，θ为可训练的参数。

$$L_2(\theta) = -\frac{1}{N}\sum_{i=1}^{N} log(y\_cls_{ij}), y\_cls_{ij} = softmax(T_{ij}) \tag{5}$$

4.4 信息融合层（Information Fusion Layer）

信息融合层将段落语义表示、范畴语义信息、范畴分类信息以及段落位置信息进行融合，具体模型如图7所示。

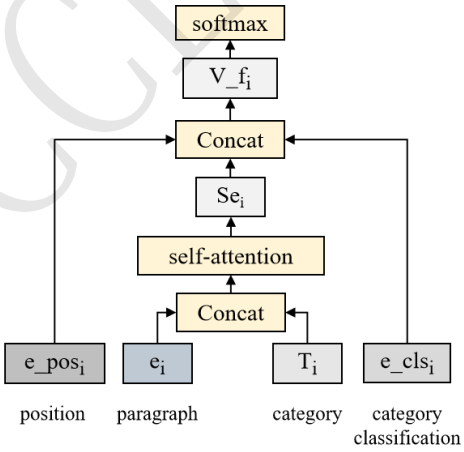


图 7. 信息融合层模型图

将段落交互信息T_e和范畴语义信息T_{topic}通过自注意力机制进行融合，获得更加丰富的段落语义信息表示序列$T_{Se} = \{Se_1, ..., Se_5\}$，其中$Se_i$表示段落交互信息和范畴语义信息融合后的段落表示，如公式（6）所示。注意力机制计算公式如（7）所示。

$$T_{Se} = Attention(concat(T_e, T_{topic})) \tag{6}$$

第二十一届中国计算语言学大会论文集，第120页-第131页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

$$Attention\,(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

将范畴类别信息$T_{topic_cls}=\{cls_{11},cls_{21},cls_{32},cls_{42},cls_{53}\}$当作特征进行编码，得到编码后的范畴类别表示序列$T_{e_cls}=\{e_cls_{11},e_cls_{21},e_cls_{32},e_cls_{42},e_cls_{53}\}$，其中$e_cls_{ij}$表示范畴类别编码后的结果。

本文认为篇章功能语用识别与段落所在篇章中的位置有关，因此引入了段落的位置信息，即段落在篇章中处于第几段。具体而言，本文对段落位置信息进行编码，得到段落位置信息表示序列$T_{position}=\{e_pos_1,e_pos_2,...,e_pos_5\}$。

将最终的段落语义表示T_{Se}、范畴类别表示T_{e_cls}、位置信息$T_{position}$进行拼接，得到段落的最终表示T_{V_f}，如公式（8）所示。将T_{V_f}送入到softmax层识别出其功能语用，使用$\hat{y}_i$表示段落功能语用的预测结果，如公式（9）所示。

$$T_{V\_f}=concat(T_{Se},T_{e\_cls},T_{position}) \tag{8}$$

$$\hat{y}_i=softmax(T_{V\_f}) \tag{9}$$

本文同样采用负对数似然作为损失函数，记为L_3，如公式（10）所示。其中θ为可训练的参数。

$$L_3(\theta)=-\frac{1}{N}\sum_{i=1}^{N}log(\ \hat{y}_i) \tag{10}$$

4.5 损失函数

本文共有三部分损失，分别为功能语用范畴划分层损失L_1，如公式（2）所示，范畴识别层的损失L_2，如公式（5）所示，和功能语用识别的损失L_3，如公式（10）所示。将最终的损失函数记为L，L计算方法如公式（11）所示：

$$L=L_1+L_2+L_3 \tag{11}$$

5 实验

5.1 实验设置

为了进一步扩大语料规模，依照MCDTB(Jiang et al., 2018)的标注方法，本文在MCDTB的基础上又新标注了480篇与新华社的新闻稿统一风格的新闻报道，形成了规模为1200篇的宏观篇章语料库MCDTB 2.0。本文在宏观篇章语料库MCDTB 2.0上对本文提出的篇章功能语用的识别方法进行了评估。宏观篇章语料库MCDTB 2.0共包含720篇来自宾州篇章树库(Chinese Treebank 8.0,简称CTB 8.0)的新闻报道和480篇来自Gigaword 2.0的新闻报道，共1200篇新闻报道，标注了6763个段落的功能语用，包含15种功能语用类型。

使用Pytorch作为深度学习框架，学习率为1e-4，训练轮数为20轮。文本编码层使用的XLNet版本为XLNet base，最大长度设置为1024，Bi-GRU编码器隐藏层维度设为512,位置信息以及范畴类别编码维度均设置为10，dropout设置为0.1，使用Micro-F1和Macro-F1来分析系统的性能。

5.2 实验结果

为了验证本文提出的模型的有效性，与基准系统进行了对比。各基准系统介绍如下：

（1）SVM：本文复现了Yarlott et al. (2018)的传统机器学习模型SVM来识别篇章功能语用。使用的特征分别为词袋模型、TF-IDF、段落语义特征和上一段的标签信息。

（2）特征+CRF：本文复现了Banisakher et al. (2020)的模型来识别篇章功能语用。除了SVM使用的特征外，还采用了词汇、位置和句法特征，并通过CRF模型来识别篇章功能语用。

(c) 2022 中国中文信息学会计算语言学专业委员会

（3）Song et al. (2020)：本文复现了Song的序列标注模型识别段落功能语用。此模型将段落语义信息与段落位置信息相结合，并通过自注意力机制获得更加丰富的段落表示，从而识别段落功能语用。

（4）Choubey et al. (2020)：本文复现了Choubey采用序列标注的模型识别段落功能语用。此模型使用分层的Bi-LSTM获得字符、段落和篇章之间的交互信息，并通过分类器识别段落的功能语用。

（5）XLNet(Yang et al., 2019):本文使用XLNet预训练模型获得段落语义表示，并通过分类器识别段落功能语用。

模型	Micro-F1(%)	Macro-F1(%)
SVM	53.23	16.72
特征+CRF	56.28	17.55
Song	64.35	18.27
Choubey	65.13	18.73
XLNet	62.26	15.02
FPRNS(Ours)	**68.19**	**22.63**

表 1. 不同模型的实验结果

实验结果如表1所示。从表1可以看出神经网络模型的识别性能在Micro-F1上均要优于传统机器学习模型，这说明相比传统机器学习模型，神经网络模型能够捕获到深层的语义信息。但是由于传统机器学习模型相比XLNet模型使用了除语义信息之外的结构特征，所以在Macro-F1上的性能高于XLNet模型。

本文提出的FPRNS模型在Micro-F1和Macro-F1的性能均取得最优值，分别达到了68.19%和22.63%，相较于表现最好的Choubey在Micro-F1和Macro-F1上分别取得了3.06%和3.9%的提升。相较于XLNet模型，FPRNS模型在Micro-F1提升了5.93%，在Macro-F1上提升了7.61%。这是由于相比XLNet模型，本文提出的FPRNS模型既捕获到篇章中段落之间交互信息，又融入了新闻图式结构信息，能够获得包含丰富信息的篇章段落表示，与此同时又结合了篇章位置信息，从语义和结构两个方面获得了更加准确的段落表示，因此FPRNS模型对于篇章功能语用的识别能力更强。

5.3 实验分析

5.3.1 消融实验

模型	Micro-F1(%)	Macro-F1(%)
base	65.61	19.00
+position	67.21	21.73
+TS	66.65	21.27
+TS+TSR	67.15	22.09
+TS+TSR+position	**68.19**	**22.63**

表 2. 消融实验结果

消融实验结果如表2所示。base表示先使用XLNet对篇章中的段落进行编码，然后将段落编码信息送入到Bi-GRU中获得具有上下文交互信息的段落语义表示；+position表示在based的基础上加入段落位置信息；+TS表示在base的基础上加上范畴划分信息；+TS+TSR表示在考虑范畴划分信息的同时添加范畴分类信息；+TS+TSR+position表示在融入范畴划分信息的同时结合段落在篇章中的位置信息。

从表2可以看出，加入位置之后，对于篇章功能语用的识别性能在Micro-F1上和Macro-F1上分别有1.6%和2.73%的提升。这是因为段落的位置信息对于位置比较固定的功能语用类型的识别性能影响较大。例如，位于第一段“导语”的识别性能由原来的83.9%提升到了92.31%，提升了8.41%；常位于最后一段对全文做出补充的“补充”类别的识别性能提升了3.8%。

(c) 2022 中国中文信息学会计算语言学专业委员会

范畴	模型	评价	结果	陈述	补充	次总述	情景	导语	背景	Mic-F1	Mac-F1
总述	base	0	0	0	26.67	19.05	63.55	92.14	21.05	72.98	14.83
	+TS	0	0	0	**30.77**	24.06	65.00	93.04	22.22	73.23	15.66
	+TS+TSR	0	0	0	30.12	**26.61**	**65.21**	**94.57**	**23.34**	**73.51**	**15.99**
故事	base	0	0	18.18	14.52	52.17	80.10	0	0	64.3	11.00
	+TS	0	7.12	19.38	27.74	56.49	80.69	0	8.70	65.87	13.28
	+TS+TSR	0	**7.74**	**20.44**	**30.43**	**58.77**	**80.83**	0	**15.11**	**66.93**	**14.91**
收尾	base	19.05	0	0	70.24	0	27.12	0	0	51.88	9.70
	+TS	27.27	0	0	71.35	0	**28.57**	0	0	53.69	10.70
	+TS+TSR	**34.78**	0	0	**75.35**	0	27.63	0	0	**56.89**	**11.46**

表 3. 各范畴下功能语用识别消融实验

从表2可以看到加入功能语用范畴划分信息后，模型对于功能语用的识别性能在Micro-F1和Macro-F1上均有提升。因为加入范畴划分信息相当于融入了篇章的组织结构信息，能够获得更加准确的段落表示。表3是各范畴下功能语用识别消融实验结果。从表3可以看出加入范畴划分信息之后，对于每个范畴下功能语用的识别性能均有所提升。其中“次总述”、“导语”、“背景”、“补充”、“评价”等功能语用性能的提升更大，因为每个范畴内部内容的组织与整篇文章的组织方式是类似的，即按照“事件总述(导语/次总述)—事件背景介绍(背景)——事件详细阐述(情景)——事件评议(补充/评价)”的形式进行组织。从表2可以看出，加入范畴类别信息之后，对于功能语用的识别性能有小幅提升，同时从表3可以看出每个范畴下功能语用的识别性能也都有所提升，这是由于“导语”仅存在“总述”下，“补充”、“评价”等通常在“收尾”处出现，所以当范畴类别信息作为特征加入之后，能够提升功能语用的识别性能。

5.3.2 错误分类样本分析

功能语用	评价	补充
补充	28.3	/
情景	54.3	39.0

表 4. 错误分类样本的比例(%)

为了分析FPRNS模型的混淆情况，本文统计了错误分类样本，其中混淆比较严重的类别如表4所示。从表4可以看出，因为“情景”类的数量较多，导致有54.3%的“评价”识别为“情景”类。而“补充”和“评价”在篇章位置相似，常位于全篇最后或每个范畴最后，和篇章主体部分语义联系相对松散，并且这两类功能语义相似，因此存在较大混淆。同时有39.0%的“补充”类识别为“情景”类：一方面是因为“情景”类数量较多，另外一方面由于“补充”类在语义上与“情景”类似，所以“补充”类与“情景”类存在混淆。

6 总结

现有的篇章分析研究主要是基于形式语法的，而篇章作为一个语义单位，其功能和语义却没有引起足够的重视，Van Dijk的新闻图式理论指出了段落在整篇文章发挥重要的作用，因此本文提出了一个基于新闻图式结构的篇章功能语用识别方法（FPRNS）识别段落的功能语用。该方法在获取到段落交互信息的同时又融入了篇章的新闻图式结构信息，并结合段落所在篇章中的位置信息，从而有效地提高了篇章功能语用的识别能力。在MCDTB 2.0的实验结果表明，本文提出的方法在Micro-F1和Macro-F1上均取得了最优性能，充分说明了本文提出方法的有效性。由于FPRNS在Macro-F1上的性能还有很大的提升空间，未来将挖掘更加丰富的篇章语义信息以识别出更多样本数量比较少的功能语用类型。本文主要针对新闻领域进行篇章功能语用的识别，此外，像新闻评论、行政裁定书和刑事裁定书等类型的篇章尽管有不同的组织形式，但是却有类似的结构和功能，未来将对这些类型的篇章进行标注并进一步深入研究其结构和功能语用。

第二十一届中国计算语言学大会论文集，第120页-第131页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Deya Banisakher, W Victor Yarlott, Mohammed Aldawsari, Naphtali Rishe, and Mark Finlayson. 2020. Improving the identification of the discourse function of news article paragraphs. In *1st Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE 2020)*.

Allan Bell and Peter Donald Garrett. 1998. *Approaches to media discourse*. Wiley-Blackwell.

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*.

Prafulla Kumar Choubey and Ruihong Huang. 2021. Profiling news discourse structure using explicit subtopic structures guided critics. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1594–1605.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xiaomin Chu. 2018. *Research on representation schema, resource construction and computational modeling of macro discourse structure.* Ph.D. thesis, Soochow University.

Xiaomin Chu, Feng Jiang, Sheng Xu, and Qiaoming Zhu. 2018. Building a macro chinese discourse treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Anita de Waard, Paul Buitelaar, and Thomas Eigner. 2009. Identifying the epistemic value of discourse segments in biology texts (project abstract). In *Proceedings of the Eight International Conference on Computational Semantics*, pages 351–354.

MA Halliday. 1994. An introduction to functional grammar 2nd edition, london: Arnold. *Halliday, Michael and Matthiessen, Christian (2004) An Introduction to Functional Grammar, London: Hodder*.

Feng Jiang, YX Fan, XM Chu, PF Li, QM Zhu, and Fang Kong. 2021. Hierarchical macro discourse parsing based on topic segmentation. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 13152–13160.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. Mcdtb: a macro-level chinese discourse treebank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504.

Joost G Kircz. 1991. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of documentation*.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487.

Fairclough Norman. 1995. Media discourse. *London: Edward Arnold*.

Zhongdang Pan and Gerald M Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75.

(c) 2022 中国中文信息学会计算语言学专业委员会

Wei Song, Ziyao Song, Ruiji Fu, Lizhen Liu, Miaomiao Cheng, and Ting Liu. 2020. Discourse self-attention for discourse element identification in argumentative student essays. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2820–2830.

Zhenhua Sun, Feng Jiang, Peifeng Li, and Qiaoming Zhu. 2020. Macro discourse relation recognition via discourse argument pair graph. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 108–119. Springer.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117.

Teun A Van Dijk. 1988. *News as discourse*. University of Groningen.

W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC bioinformatics*, 7(1):1–10.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

W Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. 2018. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33.

Xiaoxia Zhang. 2014. The enlightenment and application of macrostructure theory in teaching english newspaper reading. *Journal of Xi'an University of Arts and Science(Social Sciences Edition)*, 17(1):81–84.

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

融合知识的多目标词联合框架语义分析模型

陈旭东[1,2]，郑策[1]，常宝宝[1*]
1.北京大学计算语言学教育部重点实验室，北京100871
2.北京大学软件与微电子学院，北京102600
{xdc,zce1112zslx,chbb}@pku.edu.cn

摘要

框架语义分析任务是自然语言处理领域的一项基础性任务。先前的研究工作大多针对单目标词进行模型设计，无法一次性完成多个目标词的框架语义结构提取。本文提出了一个面向多目标的框架语义分析模型，实现对多目标词的联合预测。该模型对框架语义分析的各项子任务进行交互性建模，实现子任务间的双向交互。此外，本文利用关系图网络对框架关系信息进行编码，将其作为框架语义学知识融入模型中。实验表明，本文模型在不借助额外语料的情况下相比之前模型都有不同程度的提高。消融实验证明了本文模型设计的有效性。此外我们分析了模型目前存在的局限性以及未来的改进方向。

关键词： 框架语义分析；框架网络

Knowledge-integrated Joint Model For Multi-target Frame Semantic Parsing

Xudong Chen[1,2], Ce Zheng[1], Baobao Chang[1*]
1.The MOE Key Laboratory of Computational Linguistics, Peking University , Beijing 100871, China
2.School of Software and Microelectronics, Peking University , Beijing 100260, China
{xdc,zce1112zslx,chbb}@pku.edu.cn

Abstract

Frame semantic parsing is a fundamental task in natural language processing. Most of the previous research work focuses on the design of the single-target model. Therefore, these models can't extract frame semantic structures of multiple targets in one time. This paper designs a frame semantic parsing model for multiple targets which jointly predicts the results of multiple targets. We model and achieve the bidirectional interaction among the subtasks of frame semantic parsing. Moreover, Relational Graph Convolution Network (R-GCN) is utilized to encode the frame relation information, which is a way to exploit frame semantic knowledge into our model. The experiments shows that our model maintains good performance without extra training corpus. Ablation Study proves the effectiveness of our model.

Keywords: Frame semantic parsing , FrameNet

*通讯作者
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

框架语义分析任务是自然语言处理领域的基础性任务，被广泛应用于下游自然语言处理任务中，例如阅读理解(Guo et al., 2020)、问答系统(Shen and Lapata, 2007)等。

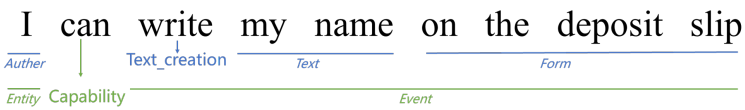


Figure 1: 框架语义结构的文本标注示例。

框架语义分析任务基于框架网络（FrameNet）标注体系，提取文本中的结构化的信息，包括语义框架以及语义框架下的一系列语义角色。Figure 1展示了一个包含框架语义结构标注的文本。在该文本中，can和write作为目标词分别触发了Capability和Text_creation两个框架，此外，图中还标注了跟框架有关的框架语义角色。这里以Text_creation为例，文本中的I、my name和on the deposit slip分别扮演了该框架下的作者（Author）、文本（Text）和载体（Form）三种角色。

因此，框架语义分析任务由多个子任务组成。在给定目标词的情况下，需要完成对目标词的框架识别以及文本中与框架有关的角色位置和类型的识别，后者统称为角色识别。先前的研究工作提出的模型绝大多数是针对单目标词的框架语义分析模型，无法一次性处理文本中包含多个目标词及其对应的框架语义结构的情况。它们通常会对多目标词文本采用了一个目标词一个文本的切分方式进行处理，让模型多次预测。此外，同个文本中的不同目标词在语义上会存在一定的相关性，独立多次的预测方式下模型无法学习到目标词之间的语义联系。我们统计了FrameNet训练集中不同目标词个数区间的句子数量分布情况，如Figure 2所示。可以看到绝大多数的句子都包含两个以上的目标词，且大部分句子的目标词数目集中在2到9之间。可以推断出，自然文本中存在多个目标词以及它们对应的框架语义结构是非常普遍的。因此，对多目标词进行联合框架语义分析具有一定的应用价值和研究意义。

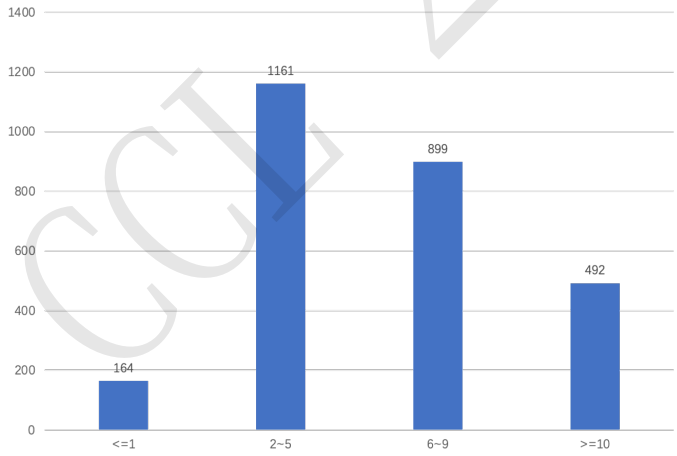


Figure 2: FN1.5训练集中不同目标词个数的句子数量分布情况。

本文提出了一个面向多目标词的框架语义分析模型，联合预测文本中所有目标词的框架语义结构。此外，我们在框架知识融入和子任务的交互性建模两个方面对模型进行改进。

在框架知识融入方面，FrameNet根据框架语言学理论定义了框架之间的多种关系，这些关系信息能够作为有价值的外部知识融入到框架识别任务中。对框架关系进行显式化建模，有利于模型学习到框架的语义。我们借助框架关系进行构图，并在此基础上进行扩展，将角色加入到框架语义关系中。我们利用图网络对扩展后的框架关系图进行编码，得到包含框架关系知识的框架和角色语义表示，将它们分别融入到框架语义分析的各项子任务中。

框架语义的各项子任务之间的关联性非常密切，加强子任务的交互能缓解误差传播，从而提升框架语义分析模型的表现。我们模型设计了一个两阶段的角色识别模块，分别采用序列标

(c) 2022 中国中文信息学会计算语言学专业委员会

注和生成式的方式进行角色位置识别，第一阶段的角色识别结果会传递到框架识别模块，框架识别模块的预测结果通过框架语义向量的方式传递到第二阶段的角色识别中，从而实现了子任务间的双向交互。在训练阶段，我们对所有子任务模块的参数进行联合优化。

我们模型在FN1.5数据集上进行实验，并且有了阶段性的结果。在不借助额外语料的情况下，模型在框架识别上有91.8%的准确率，在框架语义结构提取的F1值上达到了78.5%。消融实验证明了我们模型设计的有效性。此外，我们详细分析了模型目前存在的局限性，将在未来工作中作针对性的模型优化。

2 相关工作

框架语义分析由Daniel Gildea and Daniel Ju (2002)最先提出，并成为了2007年的SemEval会议的一项国际测评任务(Baker et al., 2007)。早期对于框架分析模型的研究主要基于传统的统计学模型，依赖人工构造一系列规则与特征作为模型的输入。Johansson (2007)利用支持向量机（Support Vector Machine）进行建模，构建出一系列分类器进行框架和角色的识别。选取的特征包括词形还原的目标谓词、目标的词形以及句子的依存句法特征等，该方法构建的语义分析系统成为SemEval 2007国际评测任务上最优模型。Das(2010)利用基于统计学的条件对数线性模型进行框架语义分析，以该模型为基础的语义分析器SEMAFOR(2010)的框架语义分析效果超越了Johansson提出的模型。

近些年，随着深度学习技术在自然语言处理领域的广泛应用，有研究者将深度学习语言模型应用到框架语义分析任务中。FitzGerald(2015a)针对框架角色识别任务，将框架、角色与文本特征信息通过全连接神经网络映射到特定维度的向量空间，将得到的特征向量作为输入对角色标注进行打分。，该模型可以看作是图网络模型，图节点是句子包含的所有文段以及框架中所有的角色，任意文段顶点与角色顶点存在一条边，边的值即为该文段被识别成对应框架角色的得分。Hermann(2014)最早提出了利用词向量进行框架识别任务，它通过挖掘上下文依存语法信息生成特征向量，并利用WSABIE算法(Weston et al., 2011)将特征向量映射到低维空间，同时训练得到低维的框架语义向量，通过计算特征向量与框架向量的相似性进行框架匹配。实验表明该方法相比之前的研究方法在结果上有显著提升。Hartmann(2017)针对框架语义分析中存在的领域依赖问题进行研究，提出了适用于跨领域的框架分析模型。该模型利用预训练词向量表示目标谓词与上下文信息，采用两层神经网络模型和基于WSABIE算法模型进行框架识别。其中上下文信息由SentBOW与DepBOW两种方式构建。其中，SentBOW是取句子中所有词的词向量均值作为上下文特征向量，DepBOW是取句法树中目标谓词依赖项的词向量均值作为上下文特征向量。在跨领域文本的框架识别任务中，相对于其他模型有更好的识别效果。Swayamdipta(2017) 将循环神经网络运用到框架语义角色识别任务中，提出了基于SegRNN的框架角色识别模型，该模型能够在不借助语法解析的情况下（仅在训练阶段进行语法标注）完成框架角色识别任务。SegRNN模型是将双向长短时记忆神经网络（LSTM）(Hochreiter and Schmidhuber, 1997) 模型与半马尔可夫随机条件场（Semi-CRF）(Sarawagi and Cohen, 2004)模型结合的联合模型。其中，Semi-CRF模型用于对文本进行切分，模型对文本的每个文段进行额外的编码，获得对应文段的特征向量。最后通过对文段特征向量进行分类，得到框架角色识别的结果。Yang(2017a)提出一种基于BIO标记方案的序列标注模型，该模型由多个LSTM网络堆叠，从而获得更深的上下文语义信息，此外该模型在输出层前添加了一层随机条件场（CRF）用于建模不同标签间的依赖关系。Peng(2018a)提出了一个基于文段的框架语义分析模型，该模型通过遍历句子中所有可能的文段并计算它们的匹配分数。此外，以上两个模型对框架语义分析的各项子任务采用联合训练策略，一定程度上加强了子任务的交互缓解误差传播问题。在此基础上，陈(Chen et al., 2021)通过在各任务模块间设计交互环节，进一步加强子任务之间的交互性。

预训练语言模型的提出，推动了自然语言技术的发展。以BERT(Devlin et al., 2019)为代表的预训练语言模型相比传统的深度学习语言模型，捕捉上下文语义信息的能力更强。目前有研究工作针对框架语义分析中的框架识别任务，提出了基于BERT的框架识别模型。Jiang(2021)利用框架网络语料库中的框架定义信息，将其与文本拼接，通过BERT模型进行编码，对编码后的特征向量进行二分类。该过程将框架识别任务建模成文本与框架的相似性判断任务。Su(2021)针对框架知识的融入进行研究，提出了融入框架知识的框架识别模型，证明了融入框架知识对框架识别任务的有效性。

(c) 2022 中国中文信息学会计算语言学专业委员会

以上的研究工作主要是面向单目标词的框架语义分析研究。本文模型实现了对多目标词的联合框架语义分析。此外，我们将框架关系知识融入到框架语义分析的各项子任务中，提升模型在整个框架语义分析任务上的表现。

3 模型介绍

本文模型主要由文本表示模块、框架关系图表示模块、角色识别模块和框架识别模块四部分组成。

文本表示模块用于编码文本的上下文信息。框架关系图表示模块基于R-GCN对框架关系进行建模，该模块的输出为包含框架关系信息的框架和角色向量表示。框架角色识别模块采用两阶段的识别方式。第一阶段为目标词感知的角色位置预测，该阶段仅利用目标词的语义信息（目标词所属的框架类别未知），采用序列标注的方式对角色位置进行识别，识别结果将送入框架识别模块中。第二阶段为框架感知的角色位置和类别预测，该阶段的输入除了目标词的语义信息外，还接收框架识别模块得到的预测框架所对应的框架向量表示（来自框架关系图表示模块），采用生成式的方式实现角色位置和类别的预测。该阶段的预测结果为最终的角色位置和类别的识别结果。框架识别模块包含一个角色位置感知的注意力机制和框架分类器，分别用于实现目标词的信息聚合和框架预测。

3.1 文本表示模块

文本表示模块用于将一段自然文本S进行编码表示，得到包含上下文语义表示的向量序列。该模块的输入为一段文本和一个目标词集合$T = \{t_1, t_2, ...t_k\}$，k 为该文本的目标词个数。

本文模型采用预训练好的英文版本的BERT语言模型作为编码器，从而得到包含上下文语义信息的表示序列。此外，常规的BERT语言模型的输入包含单词类型标记向量，用于区分不同的句子，这里我们针对框架识别任务的特点进行改进，将其作为识别某个单词是否属于目标词的标记向量。

此外，一些目标词可能是包含多个单词的词组。对于这种情况，我们模型采用平均所有单词的表示向量作为这一类目标词的语义表示向量。该过程用公式描述如下：

$$h_t = \text{Mean}(h_{w_i}), w_i \in t \tag{1}$$

最终得到所有目标词的语义表示$H_t = \{h_{t_1}, h_{t_2}, ...h_{t_k}\}$

3.2 框架关系图表示模块

框架关系图表示模块用于对框架关系进行建模，得到包含框架关系知识的框架和角色语义表示。由于框架网络定义了框架与框架之间的多种类型的语义关系，普通的GCN架构(Scarselli et al., 2008)只能建模单一类型边的图结构。因此我们采用R-GCN(Schlichtkrull et al., 2018)作为编码器，针对每个不同的关系，通过不同的参数进行图卷积计算。该模块的公式化描述如下：

$$g_i^l = \begin{cases} \text{ReLU}\left(\sum_{e\in E_f}\sum_{j\in N_i^e}\frac{1}{|N_i^e|}W_e^{(l)}g_j^{(l)} + \sum_{k\in N_i^r}\frac{1}{|N_i^r|}W_{r\_f}^{(l)}g_k^{(l)}\right) + g_i^{(l)}, i \in F \\ \text{ReLU}\left(W_{f\_r}^{(l)}g_k^{(l)} + \sum_{j\in N_i^r}\frac{1}{|N_i^r|}W_{r\_r}^{(l)}g_j^{(l)}\right) + g_i^{(l)}, i \in R \cap k \in f_i \end{cases} \tag{2}$$

其中F表示框架集合，l表示图卷积的层数。R表示角色集合，E_f表示框架间的关系类型集合，f_i为角色i对应的框架，W_{r_f}表示框架到角色的边的权重矩阵，W_{f_r}表示从角色到框架的边的权重矩阵，N_i^e表示框架i的邻居框架集合，N_i^r表示框架或角色i的邻居角色集合。

最终可以得到包含框架关系知识的框架语义表示G_f和角色语义表示G_r。

3.3 角色识别模块

角色识别模块包括角色位置的识别和角色类型的识别。其中角色位置识别采用两阶段的识别策略。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.3.1 目标词感知的角色位置识别

第一阶段的角色位置识别采用序列标注的方式，接收文本的上下文表示向量$H = \{h_1, h_2, ...h_n\}$和目标词向量$H_t = \{h_{t_1}, h_{t_2}, ...h_{t_k}\}$作为模块输入，这两部分向量都来源于文本表示模块。第一阶段的框架角色识别的目的是为了与框架识别模块中进行交互，指导目标词的上下文信息聚合，不作为最终的框架角色结果，所以这里从模型精简角度，并没有采用随机条件场（CRF）对标签约束进行建模。最终的预测标签通过以下公式得到：

$$y_i^t = \arg\max\left(\text{Linear}(\left[h_i; h_{t_i}\right])\right) \tag{3}$$

其中线性层Linear的输出维度为3，分别代表B（角色开始位置）、I（角色中间或结束位置）和O（非角色）三个标签。第一阶段的序列标注预测结果将传入框架识别模块用于目标词表示的信息聚合。

3.3.2 框架感知的角色识别

第二阶段的角色位置识别采用陈(Chen et al., 2021)提出的生成式的角色位置识别方式。该模块采用层次化的指针网络识别角色在文本中的边界位置，同时建模了不同角色间的信息交互。具体的设计细节见他们的论文。该模块的输入包括目标词的语义表示和预测框架的语义表示。输出为所有角色在文本中的边界位置索引集合。公式简化如下：

$$Span_t^{f_t} = \text{Generator}(h_t, g_{f_t}) \tag{4}$$

其中$Span_t^{f_t}$为模块预测的角色边界位置下标集合，可以表示成集合$\{(sta_1, end_1), ..., (sta_m, end_m)\}$,$m$为角色的个数。

此外，还需要对角色进行分类，我们将角色语义表示G_r融入到角色分类过程中，利用框架关系知识指导模型对框架角色的识别。该过程如下：

$$h_{span} = \text{MLP}([h_{i^end} + h_{i^sta}; h_{i^end} - h_{i^sta}; g_f]) \tag{5}$$

$$logit_r = \text{Classifier}(h_{span}) + G_r \cdot h_{span} \tag{6}$$

$$P(r|f_t, t, i^{sta}, i^{end}) = \text{Softmax}(logit_r) \tag{7}$$

其中，MLP为多层感知机网络，由Relu激活函数和两层线性层组成。Classifier是一个可训练的线性分类器，$P(r|f_t, t, i^{sta}, i^{end})$为该角色位置对应的角色类别预测概率分布。该阶段得到的框架位置和类别预测结果作为最终的角色识别结果输出。

3.4 框架识别模块

框架识别模块包括角色感知的注意力机制和一个基于多层感知机的分类器。角色感知注意力机制接收框架角色的序列标注信息，对标签为非O的词计算注意力系数，实现目标词的信息聚合。目标词t对其他非O标签的词的注意力系数计算如下，其他目标词的计算过程与之相同：

$$\alpha_i^t = \frac{\exp\left(h_i^\top W_1 h_t\right)}{\sum_{y_i^t \neq \text{"}O\text{"}} \exp(h_j^\top W_1 h_t)} \tag{8}$$

$$c_t = \sum_{y_i^t \neq \text{"}O\text{"}} \alpha_i^t h_i \tag{9}$$

其中α_i^t表示目标词t对单词i的注意力系数，c_t为信息聚合后的语义向量，W_1为权重矩阵，下同。

聚合后的语义向量c_t向量与目标词语义向量h_t得到最终的目标词语义表示r_t，并结合图网络编码得到的框架语义表示G_f进行框架识别，公式描述如下：

$$r_t = \text{Relu}\left(W_2 \cdot [h_t; c_t]\right) \tag{10}$$

$$P\left(f|t\right) = \text{softmax}\left(W_3 \cdot r_t + G_f \cdot r_t\right) \tag{11}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

3.5 联合训练设置

在训练阶段，我们对模型各部分模块参数进行联合优化，主要包括框架识别损失L_{frame}，第一、二阶段的角色位置识别损失L_{span}^{first}和L_{span}^{second}，和角色类别识别损失L_{role}，各部分的损失计算公式如下:

$$L_{frame} = \frac{1}{|k|}\sum_{i=1}^{k}\log\left(P(\hat{f}_i|S,t_i)\right) \tag{12}$$

$$L_{span}^{first} = \frac{1}{|k|}\frac{1}{|length|}\sum_{n=1}^{length}\sum_{i=1}^{k}\log(p(\hat{y}_n^{t_i}|t_i,S)) \tag{13}$$

$$L_{span}^{second} = \frac{1}{|k|}\sum_{i=1}^{k}\sum_{\tau=1}^{m}\log(P(\hat{sta}_\tau|S,t_i,f_i)) + \frac{1}{|k|}\sum_{i=1}^{k}\sum_{\tau=1}^{m}\log(P(\hat{end}_\tau|S,t_i,f_i)) \tag{14}$$

$$L_{role} = \frac{1}{|k|}\sum_{i=1}^{k}\sum_{\tau=1}^{m}\log(P(\hat{r}_\tau|S,t_i,f_i)) + \frac{1}{|k|}\sum_{i=1}^{k}\log\left(P(r_{None}|S,t_i,f_i)\right) \tag{15}$$

由于各部分模块的训练难度与参数收敛速度不同，我们加入损失调节系数用于调整模型训练的梯度优化方向，平衡各部分模块的训练收敛速度，公式描述如下：

$$L = \alpha L_{frame} + \beta L_{span}^{first} + \gamma L_{span}^{second} + \delta L_{role} \tag{16}$$

在本文模型实验中，调节系数α、β 、γ 和δ分别取0.1、0.1、0.3和0.3。

我们采用预训练的BERT语言模型作为文本的编码器，输出向量的维度为768。框架关系表示模块采用两层R-GCN作为图编码器。在训练集上训练100轮，并保存在验证集上取得最优结果的模型参数。训练的Batch大小为4，选用的优化器为BertAdam，学习率为3×10^{-5}。

4 数据集与评测指标

4.1 数据集

目前，针对框架语义分析任务有两个版本的数据集，分别为FN1.5和FN1.7，这两个数据集都来源于框架网络语料库中的篇章标注语料。之前的框架语义分析模型主要是在FN1.5版本的数据集上进行评测，因此我们对比我们模型与现有模型在FN1.5数据集上的结果。

FN1.5包含了1019种框架类型，9634种框架角色。在FN1.5篇章标注语料中，篇章中的每个句子可能包含多个框架以及它们对应的一些角色。之前的工作将篇章语料划分成单句形式，并按照一句话一个目标词进行实例的切分，它们的模型也可以看作是针对单目标词的框架语义分析模型。而我们模型能够同时处理一个句子中的所有目标词。此外FN还包含15万句的示例句子，每个示例句子只标注了单个目标词的框架语义结构。先前的一些工作会将这部分数据加入到模型的训练语料中，用于提升模型的表现。

4.2 评测指标

评测指标包括框架识别准确率和框架结构提取。分别用于评估模型在框架识别任务和框架语义分析整体任务的表现。

对于框架识别任务的评测，先前研究工作将能触发多个框架的目标词称为歧义（Ambiguous）目标词，并单独对歧义目标词这一集合进行评测。参照之前的相关研究，本文模型同样在所有目标词集合（All）和歧义目标词（Amb）集合上进行评测。

框架语义结构提取是一项评价框架语义分析整体性能的指标。该指标将框架本身视作特殊的框架角色，与其他框架角色统一计算精确率（Precision）、召回率（Recall）和F1值。此外该目标要求框架角色位置边界的精确匹配，但对于位置识别正确但类型识别错误的角色，会根据预测角色和正确角色在框架网络上的关联程度相应给与部分分数。该评测方式由Das[32]提出，并成为框架语义分析任务主流的评价指标。

(c) 2022 中国中文信息学会计算语言学专业委员会

5 实验结果与分析

我们在FN1.5测试集上对模型的框架识别准确率和整体框架分析表现进行评估。先前工作的模型指标数据来源于它们的论文。之前一些模型(Yang and Mitchell, 2017a; Peng et al., 2018a)采用集成学习策略，通过多个相同或者不同架构的模型进行专家投票的方式对框架和框架角色识别，从而进一步提升模型的整体表现。这里为公平衡量模型的真实性能，只报告单模型的指标表现。

Model	All	Ambiguous
SEMAFOR(2010)	83.6	69.2
Open-SESAME(2017)	87.0	-
Hartmann(2017)	87.6	73.8∗
Yang and Mitchell(2017a)	88.2	75.7∗
Hermann(2014)	88.4	73.1
Peng(2018a)(BASIC)	89.2	76.3
Chen(2021)	89.4	76.7
Chen (Bert)(2021)	90.5	79.1
Jiang and Riloff(2021)	91.3	81.0
Su et al(2021)	92.1	82.3
Our model	91.9	82.1

Table 1: 各模型在FN1.5上的框架识别准确率。其中∗表示它们模型的歧义目标词集合与其他模型不同

5.1 实验结果

框架识别的结果如Table 1所示。其中第一行表示的是现有的框架语义分析模型，能完整处理框架识别和角色识别任务。第二行表示的是单独的框架识别模型，它们只针对框架识别任务进行优化。可以看到我们的模型超过了之前的框架语义分析模型，接近目前在框架识别任务上表现最好的独立框架识别模型(Su et al., 2021)。我们分析认为，我们模型之所以与目前最优的框架识别模型存在差距，是因为我们模型在训练阶段对框架分析的各项子任务采用联合优化策略，保存在综合指标上表现最好的模型。然而，在训练阶段，在框架识别任务和角色识别任务的优化上呈现不同步的情况。由于角色识别任务相比于框架识别任务训练难度更大，因此需要更多的训练轮数。在整个训练过程中，框架识别的训练通常会提前收敛，最终导致过拟合的情况发生。虽然我们通过加入平衡系数在一定程度上缓解了收敛不同步问题，但没有从根本上解决。

框架语义结构提取的结果如Table 2 所示。其中，第一行和第二行分别表示基于流水线策略和基于联合训练策略的框架语义分析模型，同时也是针对单目标词的框架语义分析模型。†表示它们模型加入额外语料（例如15万句的示例句子）进行训练。除了陈的Bert版本的模型外，我们模型在不加入额外语料的情况下，显著高于其他所有框架语义分析模型。考虑到陈的Bert版本模型的效果提升一部分是通过加入15万句的示例句子得到的，为消除这部分因素带来的模型性能增益，将他们的模型仅在标准FN1.5训练集上进行训练，最终得到它们的F1值为76.4。说明我们模型在不无外部数据依赖的情况下提取的框架语言结构更优。

5.2 消融实验

我们设置了两组消融实验探究模型设计的有效性：一、框架关系知识的融入对框架语义分析任务的作用；二、子任务的双向交互对对框架语义分析任务的作用。相应实验组设置如下，用框架语义结构提取的指标在FN1.5测试集上进行评估。

(c) 2022 中国中文信息学会计算语言学专业委员会

Model	P	R	F1
SEMAFOR(2010)	69.2	65.1	67.1
Framat(2010)	71.1	63.7	67.2
Framat+context(2015)	71.1	64.8	67.8
Open-SESAME(2017)	71.0	67.8	69.4
FitzGerald et al(2015b)	74.8	65.5	69.9
Yang and Mitchell (2017b) (SEQ)†	69.6	70.9	70.2
Yang and Mitchell (2017b)(REL)†	77.1	68.7	72.7
Peng (2018b)(BASIC)†	79.2	71.7	75.3
Chen (2021)†	75.1	76.9	76.0
Chen (Bert) (2021)†	78.2	82.4	80.2
Chen (Bert,**w/o** exemplar)(2021)	75.1	77.6	76.4
Our Model	76.2	80.8	78.5

Table 2: 各模型在FN1.5上的框架结构提取结果。

实验组一（消去框架关系知识表示模块）：移除框架关系图表示模块，框架和框架角色的特征向量由随机初始化生成。

实验组二（消去角色-框架的信息传递）：去除第一阶段框架角色识别的损失计算，角色识别模块向框架识别模块的信息传递不受监督。

从Table 3中可以看到实验组一和实验组二相比原模型在各项指标上都有不同程度的下降，证明了扩展后的框架关系知识和框架与角色识别的交互对框架语义分析任务具有正面作用。此外发现，实验一在各项指标上相比原模型都有较为明显的下降，证明了扩展了框架角色后的框架关系信息对模型的预测具有较为关键的指导作用。

Model	P	R	F1
我们模型	76.2	80.8	78.5
实验组一	75.1	78.1	76.6
实验组二	75.5	80.5	77.9

Table 3: 消融实验结果.

5.3 模型存在的局限性

从模型的训练过程和实验结果中，我们发现并总结了了目前模型存在的一些问题。在未来工作中，我们将从这些问题入手进行研究，对模型作进一步的优化。

5.3.1 子任务训练不同步

从模型的训练过程中发现，框架语义分析中的各项子任务的训练难度是不同的，导致它们在相同学习率下收敛所需要的训练轮数不同，这对于采用联合训练策略的框架语义分析模型而言是一个挑战。其中，框架识别任务相对简单，所以框架识别模块的参数收敛较快，但此时角色位置和角色类别的识别能还远未达到最优。为缓解这一问题，我们在各部分模块的损失计算前加入平衡系数，用于调节子任务的收敛速度。然而这种策略仍存在两个问题。一是调节系数是固定的，无法随着模型的训练进行动态调节。二是调节系数的设置依赖人为经验，增加了模型调参的难度。因此，我们目前的模型仍无法在训练阶段同时在框架识别准确率和框架语义结构提取的综合指标上达到最优。

(c) 2022 中国中文信息学会计算语言学专业委员会

5.3.2 示例语料的利用问题

FN1.5数据集包含15万句的示例语料，这部分语料每个句子只包含一个目标词的框架语义结构信息。先前的研究工作(Yang and Mitchell, 2017a; Peng et al., 2018a; Chen et al., 2021)会将这部分数据作为额外的训练语料加入到模型的训练当中。它们在实验中发现，加入这部分数据会使模型的框架语义结构提取的表现提升3到4个百分点。然而，我们在实验中发现，我们的模型加入示例语料作为额外语料并未带来明显提升（在验证集上提升不到1个百分点）。经过分析认为，先前的模型都是针对单目标词的框架语义分析模型，示例句子的输入形式刚好符合模型的输入。然而我们模型采用多目标词的输入形式，标准训练集和示例语料在模型看来是两种异质的语料文本，在模型的训练过程中可能会产生冲突，从而导致对示例语料的利用不够充分。在未来工作中，我们将针对示例语料的处理方式进行研究，例如在模型训练前对示例语料进行自动标注，得到形式与标准训练集更为接近但有噪声的数据，标注后的数据更适合多目标词模型的训练。

6 总结

本文提出了一个面向多目标的框架语义分析模型，实现对多目标词的联合预测。该模型对框架语义分析的各项子任务进行交互性建模，实现子任务间的双向交互。此外，本文利用关系图网络对框架关系信息进行编码，将其作为框架语义学知识融入模型中。实验表明，本文模型在不借助额外语料的情况下仍具有良好的表现，消融实验证明了本文模型设计的有效性。此外本文分析了我们模型存在的一些问题，将在未来工作中针对这部分问题对模型作进一步的优化。

致谢

本文工作得到国家自然科学基金（61936012，61876004）支持，特此致谢。

参考文献

Collin F Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A Smith. 2010. Semafor: Frame argument resolution with log-linear models. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 264–267.

Xudong Chen, Ce Zheng, and Baobao Chang. 2021. Joint multi-decoder framework with hierarchical pointer network for frame semantic parsing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2570–2578, Online, August. Association for Computational Linguistics.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A Smith. 2010. Probabilistic frame-semantic parsing. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 948–956.

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.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015a. Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970, Lisbon, Portugal, September. Association for Computational Linguistics.

(c) 2022 中国中文信息学会计算语言学专业委员会

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015b. Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Comput. Linguistics*, 28(3):245–288.

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020. A frame-based sentence representation for machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 891–896, Online, July. Association for Computational Linguistics.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain, April. Association for Computational Linguistics.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, Maryland, June. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tianyu Jiang and Ellen Riloff. 2021. Exploiting definitions for frame identification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2429–2434, Online, April. Association for Computational Linguistics.

Richard Johansson and Pierre Nugues. 2007. LTH: Semantic structure extraction using nonprojective dependency trees. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 227–230, Prague, Czech Republic, June. Association for Computational Linguistics.

Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A. Smith. 2018a. Learning joint semantic parsers from disjoint data. 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 1492–1502, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hao Peng, Sam Thomson, Swabha Swayamdipta, and Noah A Smith. 2018b. Learning joint semantic parsers from disjoint data. 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 1492–1502.

Michael Roth and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Sunita Sarawagi and William W Cohen. 2004. Semi-markov conditional random fields for information extraction. *Advances in neural information processing systems*, 17.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

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.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 12–21.

Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z. Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 2021. A knowledge-guided framework for frame identification. 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 5230–5240, Online, August. Association for Computational Linguistics.

CCL 2022

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation.

Bishan Yang and Tom Mitchell. 2017a. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Copenhagen, Denmark, September. Association for Computational Linguistics.

Bishan Yang and Tom Mitchell. 2017b. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256.

CCL 2022

专业技术文本关键词抽取方法

宁祥东[1,2]，龚斌[1,2]，万林[1,2]，孙宇清[1,2,*]
1.山东大学，软件学院，济南，250101
2.教育部数字媒体技术工程研究中心，济南，250101
lcsxnxd@163.com，gb@sdu.edu.cn，wanlin@sdu.edu.cn，sun_yuqing@sdu.edu.cn

摘要

相关性和特异性对于专业技术文本关键词抽取问题至关重要，本文针对代码检索任务，综合语义信息、序列关系和句法结构提出了专业技术文本关键词抽取模型。采用预训练语言模型BERT提取文本抽象语义信息；采用序列关系和句法结构融合分析的方法构建语义关联图，以捕获词汇之间的长距离语义依赖关系；基于随机游走算法和词汇知识计算关键词权重，以兼顾关键词的相关性和特异性。在两个数据集和其他模型进行了性能比较，结果表明本模型抽取的关键词具有更好地相关性和特异性。

关键词： 关键词抽取 ； 句法结构 ； 语义信息 ； 专业文本

Keyword Extraction on Professional Technical Text

Xiangdong Ning [1,2], Bin Gong[1,2], Lin Wan[1,2], Yuqing Sun[1,2,*]
1.Shandong University,School of Software,Jinan,250101
2.Shandong University,ERC of Digital Media Technology, MoE,Jinan,250101
lcsxnxd@163.com，gb@sdu.edu.cn，wanlin@sdu.edu.cn，sun_yuqing@sdu.edu.cn

Abstract

For professional technical text keyword extraction problems, relevance and specificity are crucial, in order to achieve keyword extraction with relevance and specificity, we take semantic information, sequence relations and syntactic structure into account. Extraction of text semantic information using pre-trained language model BERT; We construct semantic association graph using sequence relation and syntactic structure, in order to capture long-distance semantic dependencies between words; We calculate keyword weights based on random walk algorithm and lexical knowledge, in order to take into account the relevance and specificity of keywords. Experimental results on professional text datasets show that keyword extracted by our model have better relevance and specificity.

Keywords: Keyword extraction , Syntactic structure , Semantic information , Profesional text

1 引言

开源代码平台为科研人员提供了分享和交流代码的环境，近几年，深度学习在自然语言处理、计算机视觉、生物计算等科研领域取得了很大成功。越来越多的深度学习模型和代码在开源平台分享，营造了可复用代码的生态环境，算法分享用户提供的代码描述文本包含功能和技术特征等专业词汇信息，以全球最大的开源代码存储库GitHub为例，其在2021年度报告中表明平台新增1600万个用户和6100万个新的代码库 (Liao et al., 2021)。专业文本的关键词抽取不仅要考虑查询关键词和代码描述文本的相关性，以提高代码检索的准确率，还要考虑关键词的特异性，以帮助用户探索新出现的代码。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：山东省自然科学基金(ZR2018ZB0420)

针对关键词抽取问题，现有方法主要包括三类：基于统计特征的文本关键词抽取方法主要依据词频、词长和词性等指标对候选关键词排序；基于图排序的关键词抽取方法将候选关键词视为节点，按照规则建立节点之间边，采用随机游走算法在词图上计算候选关键词的权重；基于主题模型的关键词抽取方法将候选关键词分配给文本包含的主题，选择每个主题下权重最大的词作为关键词。这些方法主要针对通用文本进行关键词抽取，主要关注关键词与文本的相关性，缺少考虑关键词的特异性，不适用于包含大量技术词汇的专业文本。

针对上述问题，本文主要贡献如下：

1）综合语义信息、序列关系和句法结构提出了专业技术文本关键词抽取模型。

2）采用预训练模型BERT作为文本编码器，提取文本抽象语义信息。

3）采用序列关系和句法结构融合分析的方法构建语义关联图，以捕获词汇之间的长距离语义依赖关系。

4）基于随机游走算法和词汇知识计算关键词权重，以兼顾关键词的相关性和特异性。

在两个数据集和其他模型进行了性能比较，结果表明本模型抽取的关键词具有更好地相关性；在关键词特异性分析中，基于随机游走算法和词汇知识的方法更好地提升了关键词的特异性；通过析构分析，验证了依存句法知识对模型性能带来的收益最大。

论文其他内容组织如下：相关工作中介绍了主流的关键词抽取方法，并分析了现有方法的优势和不足；对专业技术文本关键词抽取模型进行了细节介绍；在两个包含专业词汇的数据集上与现有方法进行了性能对比分析，讨论了本文方法的各部分对模型性能的影响，分析了关键词的重要性和特异性。

2 相关工作

针对关键词抽取问题，最具代表性的是基于图排序的关键词抽取方法，该方法主要思想是将描述文本中的候选关键词视为节点，然后按照一定的规则建立节点之间的边，采用随机游走算法 (Blanco et al., 2012)在词图上计算词汇权重。例如，Mihalcea (2004)等人提出了基于图排序的TextRank算法，使用文本中的单词作为节点，依据共现词汇构建边。目前已经提出了多种基于TextRank的方法，例如Wan (2008)等人提出的SingleRank方法将滑动窗口中单词的共现次数分配给词图中边的权重，该方法只能使用单个文本的信息来构建图。为了更好地表示图中节点之间的关系，越来越多的工作倾向于使用词之间的语义关系来计算图中边的权重。Tsatsaronis (2010)等人提出了SemanticRank的方法，该方法利用语义关系从文档中提取关键词和句子，在实验中证明该方法优于TextRank方法。一些工作将先验知识添加到图中的节点以强调单词的重要性，例如单词的位置、TFIDF值等。为了进一步提高TextRank算法的关键词抽取效果，Florescu (2017)等人提出了PositionRank算法，这是一种从学术文档中抽取关键词的无监督方法，该方法在词权值迭代的时候融入位置信息。Caragea (2014)等人比较了基于图的关键词抽取方法的各种中心性度量，结果表明，简单的中心性度量的结果与TextRank方法一样。这类方法能够融入深层次的文本语义信息和句法知识，但是受到分词结果的影响较大。

另一类代表性的方法是基于统计特征的关键词抽取方法和基于主题模型的关键词抽取方法。基于统计特征的方法依据统计指标对候选关键词排序 (Wang et al., 2020; Campos et al., 2020)，统计指标指标通常包括词频、词长和词性等，例如将候选关键词的TFIDF (Wang et al., 2020)值作为统计特征，依据TFIDF值的大小抽取出关键词集合，但是这类统计特征忽略了单词自身的属性，因此Campos (2020)等人提出了使用单词词性、在描述文本中出现的位置等指标为候选关键词设置不同的权重。这类方法运行速度快，但是不能提取深层次的文本语义信息。基于主题模型的方法一般是将候选关键词分配给文本包含的主题，选择每个主题下权重最大的词汇作为关键词，例如，Bougouin (2013)等人提出了一种基于主题的TopicRank方法，该方法将文档表示为一个完整的图，其中顶点不是单词而是主题。这类方法能够分析文档中的潜在主题，但是，不适用于频繁出现的关键词，难以适用于专业文本的关键词抽取。

部分工作将关键词抽取问题视为文本分类问题，将关键词库中的关键词作为类别标签，即对描述文本进行多标签分类。例如，基于循环神经网络的方法将文本中的词向量逐个输入到神经网络单元中，使用隐含层的最后一个输出来预测文本的标签 (Zheng et al., 2019)。基于卷积神经网络的方法将词向量拼接成矩阵，然后将其输入卷积神经网络后得到的文本向量，将文本向量输入到分类函数中来预测类别标签 (Wang et al., 2021; Jacovi et al., 2018)。以上两种方法由于隐藏数据的不可读性，导致了可解释性较差 (Sun et al., 2019)。Yang (2016)等人提出了包

(c) 2022 中国中文信息学会计算语言学专业委员会

括两个编码器和两个注意力层的HAN方法，该模型先将输入词汇聚合成句子向量，然后基于句子向量聚合成文本向量，通过注意力机制可以分析词和句子对类别的权重影响。这类方法一般需要依赖于大型训练数据集。

3 专业技术文本关键词抽取模型

3.1 问题描述

在基于关键词代码检索平台中，为了提高代码检索的准确率和帮助用户探索新出现的代码，专业技术文本的关键词抽取应满足以下三个性质：

- **相关性**：抽取出的技术特征关键词能够代表代码使用的技术和实现的功能。
- **重要性**：针对抽取的有限个关键词，要求按照重要程度排序。
- **特异性**：抽取出的技术特征关键词相对于代码检索平台中其他技术特征关键词的显著程度，有助于帮助用户探索新出现的代码。

3.2 整体框架

针对专业文本的关键词抽取问题，本文综合语义信息、序列关系和句法结构提出了专业技术文本关键词抽取模型。如图1所示，首先对代码描述文本进行删除停用词、保留相关词性的词和删除无意义的标点符号等预处理；为了提取文本的抽象语义信息，采用预训练语言模型BERT作为文本编码器，进而得到候选关键词向量列表；采用序列关系和句法结构融合分析的方法构建语义关联图，以捕获词汇之间的长距离语义依赖关系，图中的节点表示候选关键词，边的权重为候选关键词向量的余弦相似度值；基于随机游走算法和词汇知识计算关键词分数，以兼顾关键词的相关性和特异性；依据候选关键词的分数进行倒排序，使用语言模型得到TOP-K个关键词。第 3.3节至 3.5节对模型进行细节介绍。

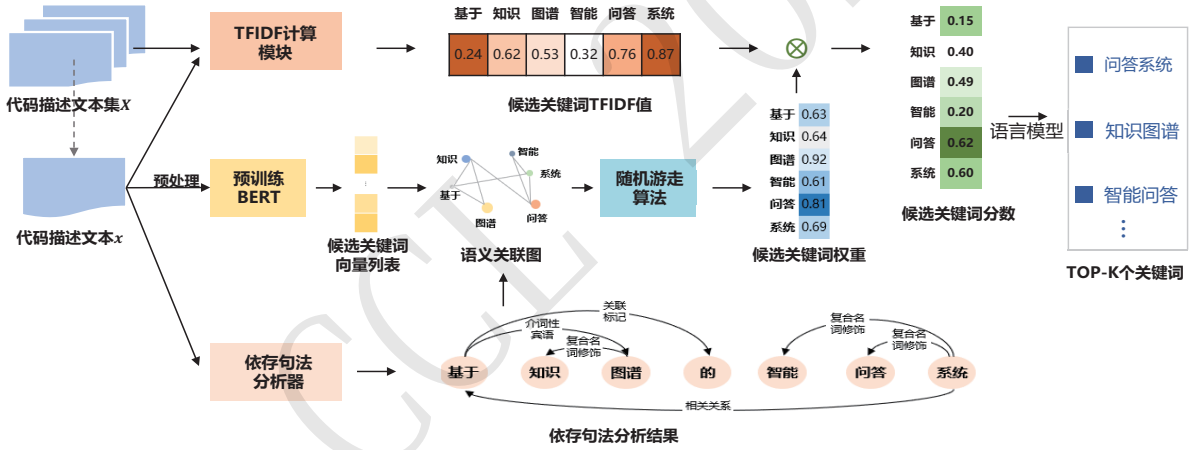


图 1. 专业技术文本关键词抽取模型

3.3 基于预训练语言模型的词汇编码

为了抽取出的关键词与文本更相关，我们对代码描述文本x进行删除停用词、保留相关词性的词和删除标点符号等预处理，经过预处理后得到候选关键词集合V_x。为了提取文本的抽象语义信息，本文采用预训练语言模型BERT (Devlin et al., 2019)对词汇上下文进行语义编码，依据编码结果得到候选关键词向量。

$$V_x = \{v_1, v_2, ..., v_n\} \tag{1}$$

$$[\boldsymbol{e}_{v_1}, \boldsymbol{e}_{v_2} ... \boldsymbol{e}_{v_n}] = BERT\,(v_1, v_2 ... v_n) \tag{2}$$

其中，v_i表示第i个候选关键词，n表示候选关键词数量，V_x表示候选关键词集合，$\boldsymbol{e}_{v_i}$表示第i个候选关键词向量。

3.4 融合序列关系和句法结构的语义关联图构建

本文基于共现词汇得到序列关系，融合序列关系和句法结构 (Chen et al., 2014)来构建建语义关联图的边E_x，语义关联图中的节点v_i表示候选关键词，边的权重为候选关键词向量的余弦相似度W_x，语义关联图是一个无向加权图。

$$E_x = \{(v_i, v_j) \,|v_i \in V_x, v_j \in V_x\} \tag{3}$$

$$w_{ij} = \begin{cases} cos\left(\boldsymbol{e}_{v_i}, \boldsymbol{e}_{v_j}\right), (v_i, v_j) \in E_x \\ 0, \text{其他} \end{cases} \tag{4}$$

$$W_x = \{w_{ij} | 1 \leq i \leq n, 1 \leq j \leq n\} \tag{5}$$

$$G_x = (V_x, E_x, W_x) \tag{6}$$

其中，E_x表示候选关键词存在的边集合，W_x表示边权重集合，w_{ij}表示v_i和v_j词向量的余弦相似度，G_x是语义关联图。

3.5 基于随机游走算法和词汇知识的关键词权重计算

本文模型综合考虑了关键词的相关性和特异性，采用随机游走算法 (Blanco et al., 2012)在语义关联图G_x上进行迭代计算后得到每个候选关键词的权重$WS_x\left(v_i\right)$，使得抽取出的关键词能够与文本更相关，具体计算公式如下所示。

$$WS_x\left(v_i\right) = (1-d) + d \times \sum_{v_j \in Nei(v_i)} \frac{w_{ij}}{\sum_{v_k \in Nei\left(v_j\right)} w_{jk}} WS_x\left(v_j\right) \tag{7}$$

其中，$WS_x\left(v_i\right)$为候选关键词v_i的权重，$WS_x\left(v_j\right)$表示上一次迭代后节点v_j的权重，$Nei(v)$表示v的邻节点集合，d为阻尼系数。

为了更好的解释基于随机游走算法计算候选关键词权重的过程，在此对计算过程进行详细说明：计算候选关键词权重的过程是一个马尔可夫过程，根据词向量的余弦相似度值可以得到词汇相似度矩阵$S_{n\times n}$，矩阵$S_{n\times n}$是一个对称矩阵，并且对角线上的元素全部取0，设定所有候选关键词的初始权重$\boldsymbol{B}_0$为该候选关键词的$tfidf$值，具体计算公式如下：

$$S_{n\times n} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \tag{8}$$

$$\boldsymbol{B}_i = S_{n\times n} \boldsymbol{B}_{i-1} \tag{9}$$

$$\boldsymbol{B}_0 = \left[tfidf_x\left(v_1\right), tfidf_x\left(v_2\right) ... tfidf_x\left(v_n\right)\right]^T \tag{10}$$

其中，$S_{n\times n}$表示候选关键词相似度矩阵，$\boldsymbol{B}_0$中的元素为所有候选关键词的初始值，$\boldsymbol{B}_i$表示第i轮计算后候选关键词的权重，$tfidf_x\left(v_i\right)$表示第i个候选关键词的$tfidf$值，只有当$\boldsymbol{B}_i$与$\boldsymbol{B}_{i-1}$的差值非常小且接近于零时达到算法收敛，算法收敛后可以得到候选词的权重。

$$tf_x\left(v_i\right) = \frac{count\left(v_i, x\right)}{size\left(x\right)} \tag{11}$$

其中，x表示代码描述文本，$size\left(x\right)$表示代码描述文本x中包含的候选关键词个数，$count\left(v_i, x\right)$表示代码描述文本x中包含第i个候选关键词的个数。

$$idf_x\left(v_i\right) = log\left(\frac{size(X)}{count(v_i, X)+1}\right) \tag{12}$$

第二十一届中国计算语言学大会论文集，第143页-第154页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

其中，X表示代码描述文本集，$size(X)$表示代码描述文本集中包含的代码描述文本数量，$count(v_i, X)$表示包含第i个候选关键词的代码描述文本的数量。

$$tfidf_x(v_i) = tf_x(v_i) \times idf_x(v_i) \tag{13}$$

tf_x表示第i个候选关键词v_i在代码描述文本x中的词频，idf_x表示第i个候选词v_i在整个代码描述文本集合X中的逆向文本频率。

本文模型采用词汇的$tfidf$值作为词汇知识，以兼顾关键词的相关性和特异性，将公式7得到的权重$WS_x(v_i)$与词汇知识进行融合，得到候选关键词分数$Score(v_i)$。为了更准确的抽取代码描述文本中的专业词汇，本文依据GitHub平台提供的代码主题，创建了一个专业词汇列表，如果候选关键词是专业词汇，那么该候选关键词的权重相对于其他候选关键词被设置为一个最大值，候选词分数的计算如公式14所示。

$$Score(v_i) = WS_x(v_i) \times tfidf_x(v_i) \tag{14}$$

$Score(v_i)$表示第i个候选词的分数，依据分数对词汇进行倒排序，使用语言模型 (Pauls et al., 2011)得到TOP-K个关键词作为技术特征关键词。

4 实验与结果分析

4.1 数据集

针对专业文本的关键词抽取问题，我们在实验中选择了两个公开且包含专业词汇的KDD、WWW数据集来验证模型的有效性。两个数据集均为ACM会议和万维网会议的研究论文，由Li (2021)等人的论文提供。

实验数据集的统计信息如表1所示，两个数据集均由论文摘要和关键词组成，数据集中的关键词均由论文作者给出，所以将作者给出的关键词视为参考关键词，将论文摘要视为描述文本。本文只保留至少包含两个句子和一个关键词的文档，KDD和WWW数据集分别包含704和1248个文档。分别在两个数据集上进行模型性能分析、关键词重要性分析、关键词特异性分析和模型析构分析。

数据集	文档总数	文档平均长度	文档平均关键词个数	关键词在文中存在比
KDD	704	204	4.16	68.12
WWW	1248	174	4.78	64.97

表 1. 实验数据集统计信息表

4.2 评价指标

（1）基于统计的精准率和召回率

本文使用精准率和召回率衡量关键词抽取算法的准确程度，命中集合指算法抽取出的关键词集合与参考关键词集合的交集，精准率$Precision$表示关键词抽取模型的准确程度，是命中集合与算法抽取出的关键词集合大小的比值。召回率$Recall$表示模型抽取的关键词对文本的覆盖程度，是命中集合与参考关键词集合大小的比值。为了避免精准率和召回率指标的冲突，我们使用精确率和召回率的调和平均数F_1分数来评估模型性能，公式如15所示：

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

（2）基于语义关系的精准率和召回率

基于统计的评价指标只能评估精准匹配的关键词，不能反映抽取关键词与参考关键词之间的语义关系，例如同义词。为此，我们设计了基于语义关系的评价指标。通过预训练语言模型BERT对词汇上下文进行编码得到词向量，通过计算参考关键词和抽取关键词向量的内积，得到语义相似性矩阵。语义精准率和召回率为参考关键词和抽取关键词最大相似性得分的累加，然后归一化，计算公式如16-18：

(c) 2022 中国中文信息学会计算语言学专业委员会

$$Precision_s = \frac{1}{|\hat{Y}|}\sum_{\hat{y}_i \in \hat{Y}} max_{y_j \in Y}\left(\hat{\boldsymbol{w}}_i{}^T \cdot \boldsymbol{w}_j\right) \tag{16}$$

$$Recall_s = \frac{1}{|Y|}\sum_{y_j \in Y} max_{\hat{y}_i \in \hat{Y}}\left(\hat{\boldsymbol{w}}_i{}^T \cdot \boldsymbol{w}_j\right) \tag{17}$$

$$F_1^s = 2 \times \frac{Precision_s \times Recall_s}{Precision_s + Recall_s} \tag{18}$$

其中，$\boldsymbol{w}_i$表示算法抽取出第i个关键词的词向量。$\boldsymbol{w}_j$表示第j个参考关键词的词向量，y_j表示第j个参考关键词，$\hat{y}_i$表示算法抽取出的第i个关键词。

（3）排名倒数指标

本文考虑了抽取关键词的排列顺序，使用排名倒数评估关键词排列的重要程度，排名倒数表示所有参考关键词在算法抽取的关键词集合中位置倒数的期望。如果参考关键词在抽取出的关键词集合中的位置越靠前，MRR值就会越大，公式如19：

$$MRR = \frac{1}{|Y|}\sum_{j=1}^{|Y|}\left(\frac{1}{Rank_{y_j}}\right) \tag{19}$$

其中，$Rank_{y_j}$表示参考关键词集合中第j个关键词在抽取的关键词序列中的序号。

（4）特异性指标

特异性表示抽取出关键词的显著程度，IDF值适用于评估算法抽取关键词的显著程度，TF值适用于评估关键词与代码描述文本的相关程度，为了兼顾关键词的显著程度和相关程度，本文使用TF和IDF的调和平均数$Specific$来评估关键词的特异性，公式如20-22：

$$IDF = \frac{1}{|\hat{Y}|}\sum_{i=1}^{|\hat{Y}|} log\frac{size(X)}{count(\hat{y}_i, X)} \tag{20}$$

$$TF = \frac{1}{|\hat{Y}|}\sum_{i=1}^{|\hat{Y}|}\frac{count(\hat{y}_i, x)}{size(x)} \tag{21}$$

$$Specific = 2 \times \frac{TF \times IDF}{TF + IDF} \tag{22}$$

其中，代码描述文本集中包含代码描述文本的数量为$size(X)$，包含算法抽取出的第i个关键词的文档数量为$count(\hat{y}_i, X)$。代码描述文本中词的数量为$size(x)$，算法抽取出的第i个关键词在代码描述文本x中出现的次数为$count(\hat{y}_i, x)$。

4.3 对比方法

对比模型如下：

- **TripleRank** (Li et al., 2021)：提出了一个无监督的关键词抽取TripleRank方法，该方法考虑了关键词位置、语义多样性和覆盖率的特征，依据这三种特征计算候选关键词的得分。
- **ISKE** (Chi et al., 2021)：提出了一种不依赖于外部资源的关键词抽取算法。使用迭代句子对单词进行排名，依据句子的语义信息生成候选关键词列表。使用加权信息初始化词的值，并使用这些值生成句子分数，依据句子的分数来更新候选关键词的值。
- **GTCRank** (Li et al., 2019)：提出了一种使用基于图排序和基于主题聚类的方法来提取关键词的无监督算法，使用基于图排序的方法来描述两个词之间的相关性，并使用基于主题聚类的方法将语义信息嵌入到词中。

CCL 2022

- **PositionRank** (Florescu et al., 2017)：提出了一个用于从学术文档中抽取关键词的无监督算法，该方法在迭代计算词权重的过程中融入了位置信息，融入方式有两种，一种是融入了该词出现的所有位置，另外一种是融入了该词出现的第一个位置。

- **YAKE** (Campos et al., 2020)：提出了一个无监督的关键词提取的YAKE算法，该算法依据从单个文档中提取的统计特征来选择文本中重要的关键词，统计特征主要包括候选关键词位置、词频等。

- **TPR** (Liu et al., 2010)：提出了将在单个词图上随机游走分解为在多个不同主题上随机游走的算法，在不同主题下分别计算候选关键词的权重，最后依据文档的主题分布来计算单词的最终排名分数。

- **RSKeyRank**：本文设计的专业技术文本关键词抽取模型，记为RSKeyRank算法，其输入是代码描述文本x，输出为RSKeyRank算法从代码描述文本中抽取到的关键词集合$\hat{Y}$，参考关键词集合为Y。

- **RSKeyRank-TFIDF**：表示RSKeyRank算法没有使用TFIDF计算模块。

- **RSKeyRank-BERT**：表示RSKeyRank算法没有使用预训练语言模型BERT。

- **RSKeyRank-Syntax**：表示RSKeyRank算法没有融入句法知识。

- **RSKeyRank-Syntax-BERT**：表示RSKeyRank算法既没有使用预训练语言模型BERT，也没有融入句法知识。此时，仅基于序列关系构建语义关联图中的边，边的权重为1。

4.4 模型性能分析

（1）基于统计的精准率和召回率

为了分析本文模型性能，我们使用精准率、召回率和F_1分数在KDD和WWW数据集上与上述模型进行了性能对比，如表2所示，表格上半部分是在KDD数据集上关键词抽取的效果，下半部分是在WWW数据集上关键词抽取的效果，TOP5和TOP10分别表示算法抽取5个和10个关键词。

从整体上来看，本文方法在两个数据集上都取得了最好的结果，说明本文方法抽取的关键词与文本更相关。在KDD数据集上抽取5个和10个关键词时，F_1分数分别达到了14.7%和15.6%，在WWW数据集上抽取5个和10个关键词时，F_1分数分别达到了16.5%和16.7%。在KDD数据集上抽取5个关键词时，RSKeyRank算法相较于YAKE算法和TripleRank算法的F_1分数提升了11.3%和2.2%，这表明采用图排序的关键词抽取算法相较于基于统计的关键词抽取算法有较大提升。RSKeyRank算法相较于ISKE算法的F_1分数提升了2.4%，这表明本文方法相较于使用迭代句子对单词进行排序的算法在关键词抽取任务上有较大提升。RSKeyRank算法相较于基于图排序的PositionRank、GTCRank和TPR算法的F_1分数分别提升了2.5%、5.3%和6.2%，这表明融入预训练语言模型BERT和句法知识可以提升关键词抽取模型的性能。

如图2(a)所示，横坐标表示RSKeyRank算法抽取关键词的个数，在KDD和WWW数据集上随着抽取关键词个数的增多，精准率一直在下降，召回率一直在上升，F_1分数则是先上升后保持不变。召回率是命中集合与参考关键词集合大小的比值，图2(a)中召回率一直在上升，表明随着RSKeyRank算法抽取出关键词个数的增多，抽取出关键词的覆盖程度就越高。图2(a)中精准率一直在下降，表明随着RSKeyRank算法抽取出关键词个数的增多，关键词抽取模型的准确程度在下降。F_1分数是精准率和召回率的调和平均，在图2(a)中F_1分数先上升后保持不变，RSKeyRank算法抽取出关键词的个数介于1和5之间时，召回率的上升速度比精准率的下降速度快，抽取出的关键词个数介于6和10之间时，精准率的下降速度与召回率的上升速度基本一致，这表明本文算法抽取的关键词个数大于等于5个时模型的性能趋于稳定。

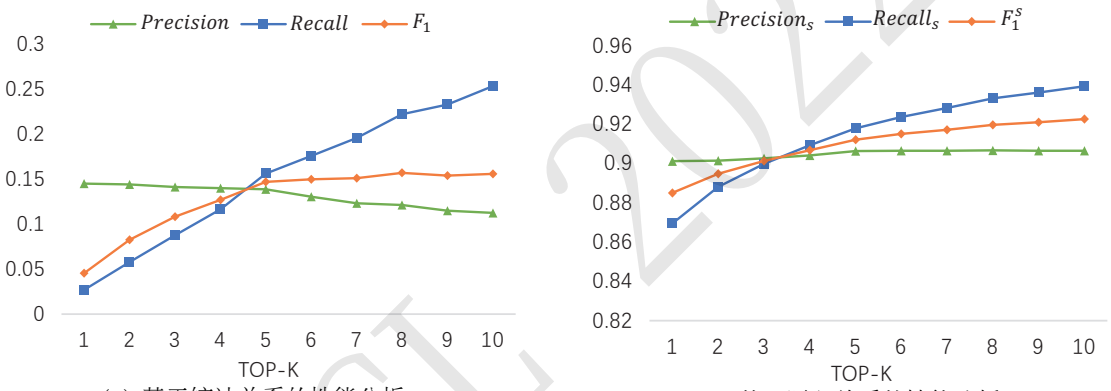
（2）基于语义关系的精准率和召回率

基于统计的评价指标只能评估精准匹配的关键词，不能反映抽取关键词与参考关键词之间的语义关系，例如同义词。为此，我们使用基于语义关系的精准率和召回

(c) 2022 中国中文信息学会计算语言学专业委员会

数据集	方法	TOP5			TOP10		
		Precision%	*Recall*%	F_1%	*Precision*%	*Recall*%	F_1%
KDD	TripleRank	11.9	14.5	12.5	9.2	19.8	11.8
	ISKE	12.0	14.3	12.3	9.1	22.0	12.2
	GTCRank	8.7	11.1	9.4	7.9	20.1	11.2
	PositionRank	11.7	14.2	12.2	9.0	19.9	11.7
	YAKE	3.1	4.0	3.4	3.5	8.8	4.8
	TPR	8.1	9.7	8.5	7.4	16.4	9.7
	RSKeyRank	**13.9**	**15.6**	**14.7**	**11.3**	**25.3**	**15.6**
WWW	TripleRank	12.9	14.2	12.9	10.1	19.6	12.5
	ISKE	12.8	13.9	12.7	10.2	19.8	12.6
	GTCRank	9.9	11.3	10.1	8.8	19.4	11.6
	PositionRank	12.4	13.6	12.3	9.9	19.7	12.3
	YAKE	4.4	5.0	4.5	3.9	8.6	5.1
	TPR	9.4	10.2	9.3	8.5	16.7	10.5
	RSKeyRank	**15.5**	**17.6**	**16.5**	**12.2**	**26.7**	**16.7**

表 2. 关键词抽取模型的性能对比



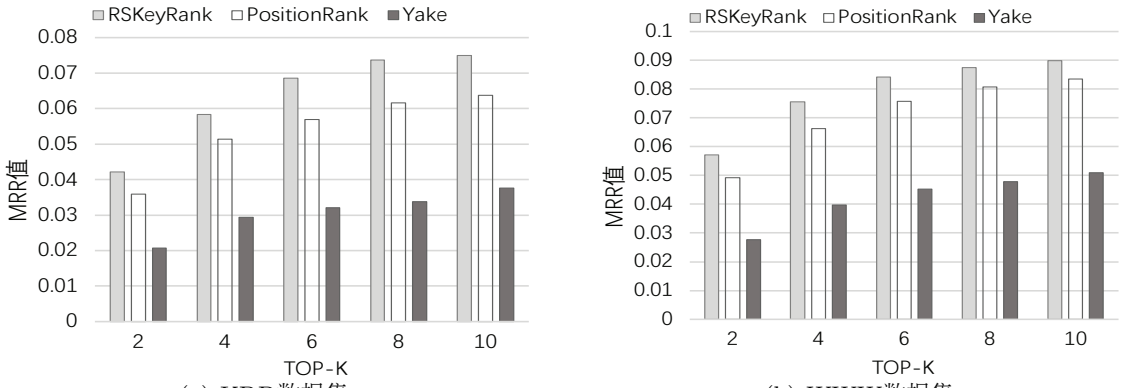
(a) 基于统计关系的性能分析　　(b) 基于语义关系的性能分析

图 2. KDD数据集上模型性能分析

率作为基于统计评价指标的补充。如图2(b)所示，横坐标表示RSKeyRank算法抽取的关键词个数，在KDD和WWW数据集上随着抽取的关键词个数的增多，$Precision_s$保持不变，$Recall_s$和F_1^s分数一直在上升，这表明随着RSKeyRank算法抽取出关键词个数的增多，算法抽取的关键词与参考关键词的语义关系就越强。图2(b)中精准率保持不变，因为RSKeyRank是无监督算法，不具备自学习的能力。

4.5 关键词重要性分析

为了评估指抽取出关键词排列顺序的重要程度，本文采用排名倒数评价关键词的重要性。如图3所示，横坐标表示各个模型抽取关键词的个数，纵坐标表示MRR值。只有参考关键词存在于抽取关键词集合时，参考关键词才有排名倒数，所以关键词的重要性和算法的性能保持一致。RSKeyRank算法的性能比Yake和PositionRank算法的性能要高，所以RSKeyRank算法相较于Yake和PositionRank算法的MRR值均达到了最高，说明了本文方法抽取的关键词排列顺序相较于其他方法更为合适。随着RSKeyRank算法抽取的关键词个数增多时，算法抽取的关键词与参考关键词相匹配的个数就越多，在两个数据集上MRR值一直在上升。



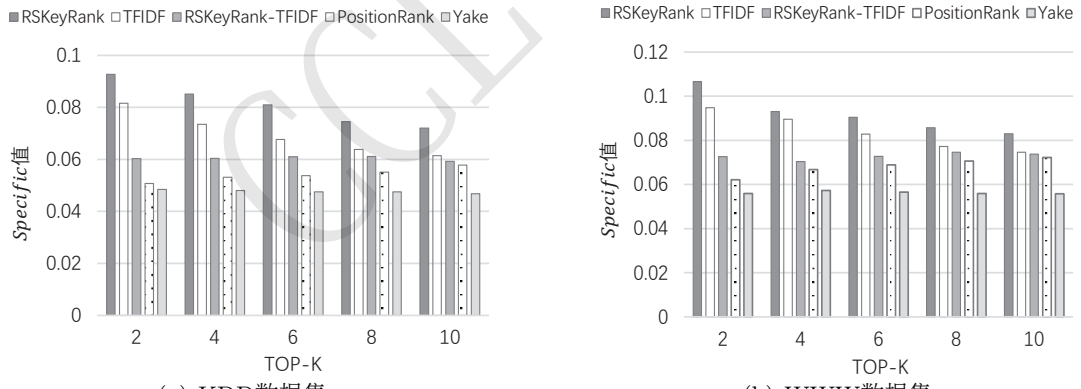
(a) KDD数据集　　(b) WWW数据集

图 3. 关键词的重要性分析

4.6 关键词特异性分析

为了评估算法抽取出关键词的显著程度，本文采用$Specific$评价指标评估算法抽取出关键词的特异性。如图4所示，横坐标表示各个模型抽取的关键词个数，纵坐标表示$Specific$值。

本文设计的算法相较于其他四种算法的$Specific$值均达到了最高，这表明RSKeyRank算法抽取到的关键词既与代码描述文本相关，又具有特异性。RSKeyRank算法相较于TFIDF方法在特异性评价指标的提升最小，因为RSKeyRank算法和TFIDF方法均使用了IDF计算模块。如果本文设计的方法没有使用TFIDF计算模块，那么RSKeyRank算法抽取出关键词的特异性值有较大的下降，但是会高于基于图排序的PositionRank算法和基于统计方法的Yake算法，因为根据 4.4节的性能分析，得知本文设计的算法抽取出的关键词与代码描述文本的相关性最高，所以特异性值高于PositionRank算法和Yake算法。这表明随机游走算法和词汇知识的融合更好地提升了关键词的特异性。RSKeyRank算法和TFIDF方法抽取出的关键词的特异性值明显高于PositionRank和Yake算法，因为PositionRank和Yake算法仅考虑了关键词与代码描述文本的相关性。PositionRank方法比Yake方法抽取出关键词的特异性值高，因为PositionRank算法抽取的关键词与代码描述文本的相关性高于Yake算法。



(a) KDD数据集　　(b) WWW数据集

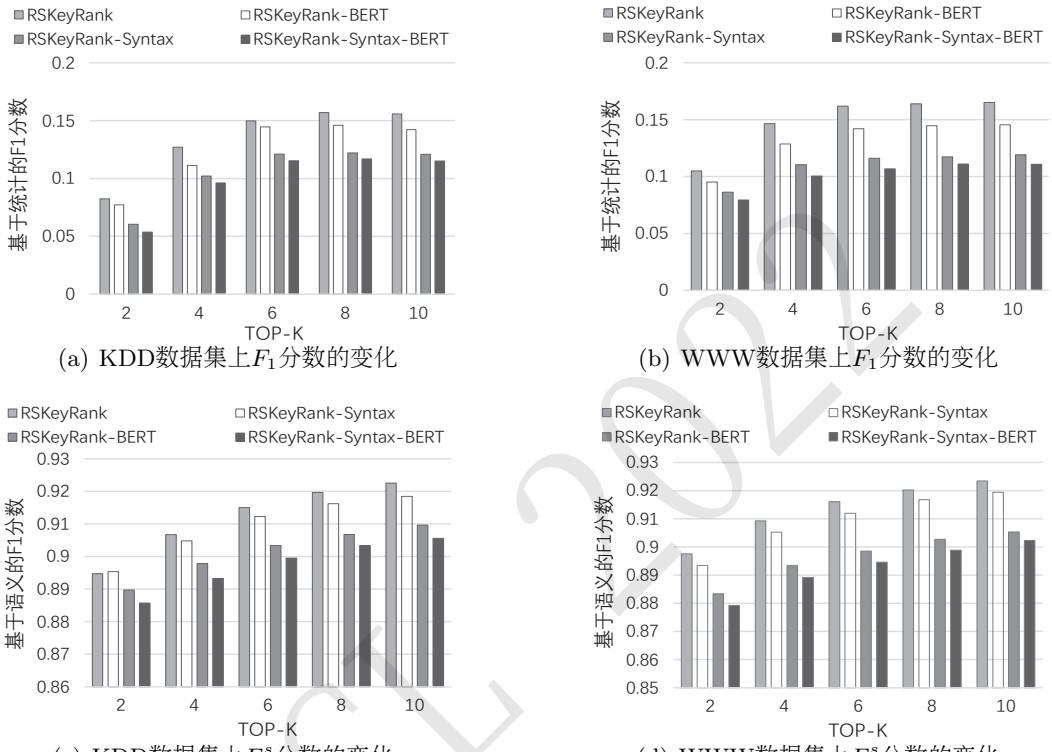
图 4. 关键词的特异性分析

4.7 模型析构分析

为了验证RSKeyRank算法各部分结构对模型带来的收益，本文对句法知识模块、预训练语言模型BERT进行有效拆分，通过基于统计和基于语义关系的评价指标在两个数据集上验证不同模块对RSKeyRank算法带来的收益，结果如图5所示。图5(a)和图5(b)表示各个模型在两个数据集上基于统计的F_1分数的变化，图5(c)和图5(d)表示各个模型在两个数据集上基于语义关系的F_1^s分数的变化。横坐标表示RSKeyRank算法抽取关键词的个数，纵坐标分别表示F_1分数

和F_1^s分数。

图5(a)和图5(b)中RSKeyRank相比于其他三个模型的F_1分数都要高，其他三个模型与RSKeyRank之间F_1分数的差距表示对模型性能带来的收益。这表明融入了句法知识对模型性能带来的收益最大，使用预训练语言模型BERT对模型性能带来的收益次之。本文设计的RSKeyRank算法是基于图排序的方法，图中的节点表示候选关键词，节点之间的边表示存在关系的两个关键词，因为句法知识可以捕获词汇的长距离语义依赖关系，使得词图更接近于真实分布，所以句法知识对模型带来了收益。图5(c)和图5(d)中RSKeyRank相较于其他三个模型的F_1^s分数最高,其他三个模型与RSKeyRank之间F_1^s分数的差距表示对模型带来语义相关性的收益，这表明使用预训练语言模型BERT对模型抽取出关键词与参考关键词之间的语义相关性带来的收益最大，融入句法知识对关键词的语义相关性带来的收益次之，因为预训练语言模型BERT可以捕获词汇之间的语义关系。



(a) KDD数据集上F_1分数的变化　(b) WWW数据集上F_1分数的变化

(c) KDD数据集上F_1^s分数的变化　(d) WWW数据集上F_1^s分数的变化

图 5. 模型析构分析实验结果

5 总结

针对专业技术文本的关键词抽取问题，本文综合语义信息、序列关系和句法结构提出了专业技术文本关键词抽取模型。采用预训练模型BERT作为文本编码器，提取文本抽象语义信息；采用序列关系和句法结构融合分析的方法构建语义关联图，以捕获词汇之间的长距离语义依赖关系；基于随机游走算法和词汇知识计算关键词权重，以兼顾关键词的相关性和特异性。在两个数据集和其他模型进行了性能比较，结果表明本模型抽取的关键词具有更好地相关性；在关键词特异性分析中，基于随机游走算法和词汇知识的方法更好地提升了关键词的特异性；通过析构分析，验证了依存句法知识对模型性能带来的收益最大。在今后工作中，计划进一步融合代码描述文本和代码结构进行关键词抽取，以提高模型性能。

参考文献

Adam Pauls, Dan Klein. 2011. Faster and Amaller N-gram language models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011: 258-267.

Adrien Bougouin, Florian Boudin and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. *International Joint Conference on Natural Language Processing*. 2013: 543-551.

Alon Jacovi, Oren Sar Shalom, Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018: 56-65.

Ammar Ismael Kadhim. 2019. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 2019, 52(1): 273-292.

Corina Florescu and Cornelia Caragea. 2017. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 2017: 1105-1115.

Cornelia Caragea, Florin Bulgarov, Andreea Godea et al. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014: 1435-1446.

Danqi Chen, Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014: 740-750.

George Tsatsaronis, Iraklis Varlamis and Kjetil Nørvåg. 2010. SemanticRank: ranking keywords and sentences using semantic graphs. *Proceedings of the 23rd International Conference on Computational Linguistics*. 2010: 1074-1082.

Haitao Wang, Keke Tian, Zhengjiang Wu et al. 2021. A short text classification method based on convolutional neural network and semantic extension. *International Journal of Computational Intelligence Systems*, 2021, 14(1): 367-375.

Jin Zheng, Limin Zheng. 2019. A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification. *IEEE Access*, 2019, 7: 106673-106685.

Jacob Devlin, Ming-Wei Chang, Kenton Lee et al. 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*. 2019: 4171-4186.

Ling Chi and Liang Hu. 2021. ISKE: An unsupervised automatic keyphrase extraction approach using the iterated sentences based on graph method. *Knowledge-Based Systems*, 2021, 223: 107014.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali et al. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 2020, 509: 257-289.

Roi Blanco, Christina Lioma. 2012. Graph-based term weighting for information retrieval. *Information Retrieval*, 2012, 15(1): 54-92.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 2004: 404-411.

Shengli Sun, Qingfeng Sun, Kevin Zhou et al. 2019. Hierarchical attention prototypical networks for few-shot text classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2019: 476-485.

Tuohang Li, Liang Hu, Hongtu Li et al. 2021. TripleRank: An unsupervised keyphrase extraction algorithm. *Knowledge-Based Systems*, 2021, 219:106846.

Tengfei Li, Liang Hu, Jianfeng Chu et al. 2019. An Unsupervised approach for keyphrase extraction using within-collection resources. *IEEE Access*, 2019, 7: 126088-126097.

Xinyun Wang and Hongyun Ning. 2020. TFIDF Keyword extraction method combining context and semantic classification. *Proceedings of the 3rd International Conference on Data Science and Information Technology*. 2020: 123-128.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*. 2008: 855-860.

(c) 2022 中国中文信息学会计算语言学专业委员会

Xiangke Mao, Shaobin Huang, Rongsheng Li et al. 2020. Automatic keywords extraction based on co-occurrence and semantic relationships between words. *IEEE Access*, 2020, 8: 117528-117538.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng et al. 2010. Automatic keyphrase extraction via topic decomposition. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 2010: 366-376.

Zhifang Liao, Yiqi Zhao and Shengzong Liu. 2021. The measurement of the software ecosystem's productivity with github. *Computer Systems Science and Engineering*, 36(1): 239-258.

Zichao Yang, Diyi Yang, Chris Dyer et al. 2016. A fusion model-based label embedding and self-interaction attention for text classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016: 1480-1489.

CCL 2022

基于实体信息增强及多粒度融合的多文档摘要

唐嘉蕊[1]，刘美玲[1,*]，赵铁军[2]，周继云[3]
1.东北林业大学信息与计算机工程学院，哈尔滨，150006
2.哈尔滨工业大学计算机科学系，哈尔滨，150001
3.约翰斯·霍普金斯大学利伯研究所，巴尔的摩，MD 21218，USA
{tjr,mlliu}@nefu.edu.cn,tjzhao@hit.edu.cn,zhoujiyun2010@gmail.com

摘要

神经网络模型的快速发展使得多文档摘要可以获得人类可读的流畅的摘要，对大规模的数据进行预训练可以更好的从自然语言文本中捕捉更丰富的语义信息，并更好的作用于下游任务。目前很多的多文档摘要的工作也应用了预训练模型(如BERT)并取得了一定的效果，但是这些预训练模型不能更好的从文本中捕获事实性知识，没有考虑到多文档文本的结构化的实体-关系信息，本文提出了基于实体信息增强和多粒度融合的多文档摘要模型MGNIE，将实体关系信息融入预训练模型ERNIE中，增强知识事实以获得多层语义信息，解决摘要生成的事实一致性问题。进而从多种粒度进行多文档层次结构的融合建模，以词信息、实体信息以及句子信息捕捉长文本信息摘要生成所需的关键信息点。本文设计的模型，在国际标准评测数据集MultiNews上对比强基线模型效果和竞争力获得较大提升。

关键词： 实体信息增强；预训练语言模型；多粒度融合；多文档摘要

Multi-Document Summarization Based on Entity Information Enhancement and Multi-Granularity Fusion

Jiarui Tang[1],Meiling Liu[1,*],Tiejun Zhao[2],and Jiyun Zhou[3]
1.School of Information and Computer Engineering, Northeast Forestry University, Harbin 150006, China
2.Department of Computer Science, Harbin Institute of Technology, Harbin 150001,China
3.Lieber Institute, Johns Hopkins University, Baltimore, MD 21218, USA
{tjr,mlliu}@nefu.edu.cn,tjzhao@hit.edu.cn,zhoujiyun2010@gmail.com

Abstract

The rapid development of neural network models enables multi-document summarization to obtain human-readable and fluent summaries, and pre-trained on large-scale data can better capture richer semantic information from natural language texts and better serve downstream tasks.Many current works on multi-document summarization also apply pre-trained models (such as BERT) with certain results, but these pre-trained models cannot better capture factual knowledge from texts and do not consider the structure of multi-document texts.This paper proposes a multi-document summarisation model MGNIE based on entity information enhancement and multi-granularity fusion, incorporating entity relationship information into the pre-trained model ERNIE, enhancing knowledge facts to obtain multi-layer semantic information and solving the factual consistency problem of summary generation. In turn, the multi-document hierarchy is fused and modelled at multiple granularities to capture the key information points required for summary generation of long text information in terms of word information, entity information and sentence information. The model designed in

第二十一届中国计算语言学大会论文集，第155页-第165页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

this paper achieves a significant improvement in effectiveness and competitiveness over the strong baseline model on the international standard evaluation dataset MultiNews.

Keywords: Entity Information Augmentation , Pre-trained Language Models , Multi-Granularity Fusion , Multi-document Summarization

1 引言

多文档摘要指的是在保留关键信息的情况下从同一主题相关的多个文档集合中生成简洁的摘要，其各个文档包含的信息虽属于同一个主题却并不相同。近年来，互联网科技迅速发展，使得我们在各种社交媒体上获得大量的数据信息，而随着新闻的快速传播，从同一主题的新闻中获取关键信息显得至关重要。随着深度学习技术在多文档摘要方面的广泛应用以及大规模数据集的发布，如WikiSum(Liu et al.,2018) ,MultiNews(Fabbri et al.,2019),生成式的多文档摘要取得了突破性进展。

最近，如BERT(Lee et al.,2018)等预训练语言模型的提出，将大规模语料库的训练好的语言模型应用于下游nlp任务，对BERT模型进行微调，使其能够更好的编码文本的上下文信息，捕捉到更深层的语义信息。最近在文本摘要方面，很多工作加入了预训练语言模型，(Liu et al.,2019)首先提出将BERT模型作为预训练模型应用于文本摘要任务，作者通过对BERT模型进行微调，通过将文档中的句子用[CLS]符号分割来学习句子表征，并且更改了区间分割嵌入来区分不同句子，作者还提出通过对编码器和解码器选取不同的优化器来解决预训练模型编码器和解码器不匹配的问题。目前在多文档摘要方面，虽然有加入预训练模型来提高模型性能的工作，但是并没有考虑带有事实信息的预训练模型来提升模型生成的事实一致性的工作。

对于的生成式多文档摘要，获取文本中丰富的语义信息对于生成连贯的摘要是非常重要的，以往的工作中，大部分生成式模型采用单词级语言生成，也有采用词级与句子级进行信息融合的摘要模型，以及应用段落级和篇章级的生成模型，能够充分获得丰富的文本信息。但在目前的工作中缺乏实体级的语义信息与其他语义单元的信息融合的生成式模型，从而丰富层次化的具有结构信息的自然语言文本。

在本文中，我们针对具有结构化的实体信息可以增强生成式摘要的事实一致性，并且融合了实体-关系信息的预训练语言模型能够使文本获得更高层级的语义表征。本文提出了基于实体信息增强以及多粒度融合的多文档摘要模型，它采用了融合了实体-关系结构化信息的ERNIE预训练模型(Zhang et al.,2019)来训练文本，来实现信息增强，将结构化的带有实体-关系的知识图通过tranE算法(Bordes et al.,2013)嵌入到预训练模型中，并实现了实体对齐，获得摘要所需的实体信息。同时我们还采用了多粒度信息融合，将词信息、实体信息和句子信息进行交互融合，从而获得多文档中更具层次化的文本语义信息。针对上文中提出的现有研究的问题，本文的贡献如下:

1）本文提出了一个基于实体信息增强的多文档摘要模型，通过采用具有结构化的实体-关系的知识图通过transE算法将结构化的图信息嵌入到ERNIE预训练模型中，使用实体链接工具TAGME来对文本中提及的实体进行提取，并进行训练从而在丰富上下文信息的基础上进一步加入实体信息实现信息增强。

2）文本提出多个粒度的信息来对原文本进行丰富的语义信息提取，我们将实体信息与词信息进行实体对齐，并通过句子信息和实体信息的融合对词token信息进行更新从而指导解码生成。

3）本文提出的模型在大规模数据集MultiNews上进行实验并取得了先进性结果表明了模型的有效性和可行性，并进行了对多粒度信息，以及是否加入融合实体信息的预训练模型进行了消融实验对比，来说明实体信息增强的有效性。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
*通讯作者
基金项目：国家自然科学基金(61702091)

2 相关工作

2.1 基于信息增强的多文档摘要

以往的多文档摘要是基于特征工程和主题模型的(Erkan et al., 2004;Christensen et al., 2013; Yasunaga et al., 2017)，通过特征增强和语义增强来提升模型性能。(Zheng et al.,2019)提出从文档视图和子主题视图中共同生成基础主题表示，通过考虑上下文信息，作者考虑了上下文信息、子主题显著性和相对句子显著性，并且以分级的方式来估计句子的显著性，从而抽取排名最高的句子作为摘要。(Alambo et al.,2020)提出基于中心性聚类的方法，使用相关参考分解从原文档中提取句子集合并且保持他们相互依赖，并且采用增强的多句压缩算法生成主题信息和摘要。其中还有依据数据增强的多文档摘要，(Pasunuru et al.,2021)提出了构建两个新的针对以查询为中心的多文档摘要数据集来实现数据增强，这两个数据集是互补的，并提出采用分层编码的方式来进行编码，同时对局部信息以及全局信息进行了编码，还加入了排序组件和查询组件。

BERT等预训练语言模型的提出，促进了多文档摘要任务的发展，(Li et al.,2020)提出利用图对文档进行编码，能更好的捕捉跨文档的关系，基于图来指导摘要生成，还提出了将BERT模型与作者提出的基于图指导的摘要模型结合起来，以更有效的处理长输入文本。针对事实一致性问题，提出的大多数方法是在评估指标方面对生成摘要的事实一致性进行评估，(Zhang et al.,2019)采用了一种弱监督的方法构造训练集，通过构造的句子文档对来判断是否具有事实一致性。近年来提出了通过外部知识库来生成文本的忠实性，(Dong et al.,2022)把不在原文本中但在与原文本链接的外部知识库中的实体视为对世界知识的忠实，原文本具有提取性，世界知识具有生成性，与之前通过过滤训练实例的只包含提取性的实体来提高事实一致性的工作相反，作者通过提供与来源相关的额外事实，以生成式的角度来提高生成实体的忠实性。

2.2 多粒度信息融合

对于的生成式多文档摘要，获取文本中丰富的语义信息对于生成连贯的摘要是非常重要的，以往的工作中，有采用单词级语言生成的，采用词级与句子级进行信息融合的摘要模型，以及应用段落级和篇章级的生成模型，从而获得丰富的文本信息。而在目前的工作中缺乏实体级的语义信息与其他语义单元的信息融合从而丰富层次化的具有结构信息的自然语言文本。

在近些年的工作中，随着深度学习的快速发展，对于多文档摘要的研究从多个粒度方面进行，大多数工作是采用单词级的文本嵌入表征来获得上下文信息，也将其他粒度的信息如段落、文档进行融合来输入表征。(Li et al.,2020) 提出了一种神经生成式多文档摘要(MDS)模型，该模型利用段落级和词级的图表示结构，如相似图和篇章图，来有效地处理多个输入文档并产生生成式摘要。transformer(Vaswani et al.,2017)的提出使得生成式文本摘要取得了突破性进展。(Zhao et al.,2020)提出SummPip模型是第一种结合语义知识和深度神经表示构造句子图的无监督摘要方法，(Jin et al.,2020)提出采用文档、句子、词多粒度信息交互网络，在不同语义粒度信息表征进行交互。(Yasunaga et al.,2017)提出在关系图上使用图卷积网络(GCN)，并将从递归神经网络获得的句子嵌入作为输入节点特征。通过多层分层传播，GCN生成高级隐藏句特征以进行显着性估计。

以上这些已提出的方法虽然在一定程度上解决了多文档摘要生成的事实一致性以及丰富的文本语义信息特征提取问题，但是针对通过预训练模型嵌入结构化的实体信息来进行信息增强的多文档摘要模型还很少，通过实验我们发现将实体信息融入文本单元中进行特征融合对于生成式文本摘要性能的提升具有有效性。

3 基于实体信息增强及多粒度融合的多文档摘要模型MGNIE

在这一节中，我们详细描述了我们提出的模型。模型的结构如图1所示。在本文中，首先使用transE算法将结构化的实体信息嵌入ERNIE预训练模型中，我们使用TAGME实体链接工具来提取文本中提及的实体，来对原文本进行实体信息融合的预训练，从而得到预训练后的词嵌入信息和实体嵌入信息，同时通过对句子进行编码获得句子嵌入信息，输入到Transformer编码层进行融合，最后通过句子信息和实体信息的融合对词token信息进行更新从而指导解码生成。

(c) 2022 中国中文信息学会计算语言学专业委员会

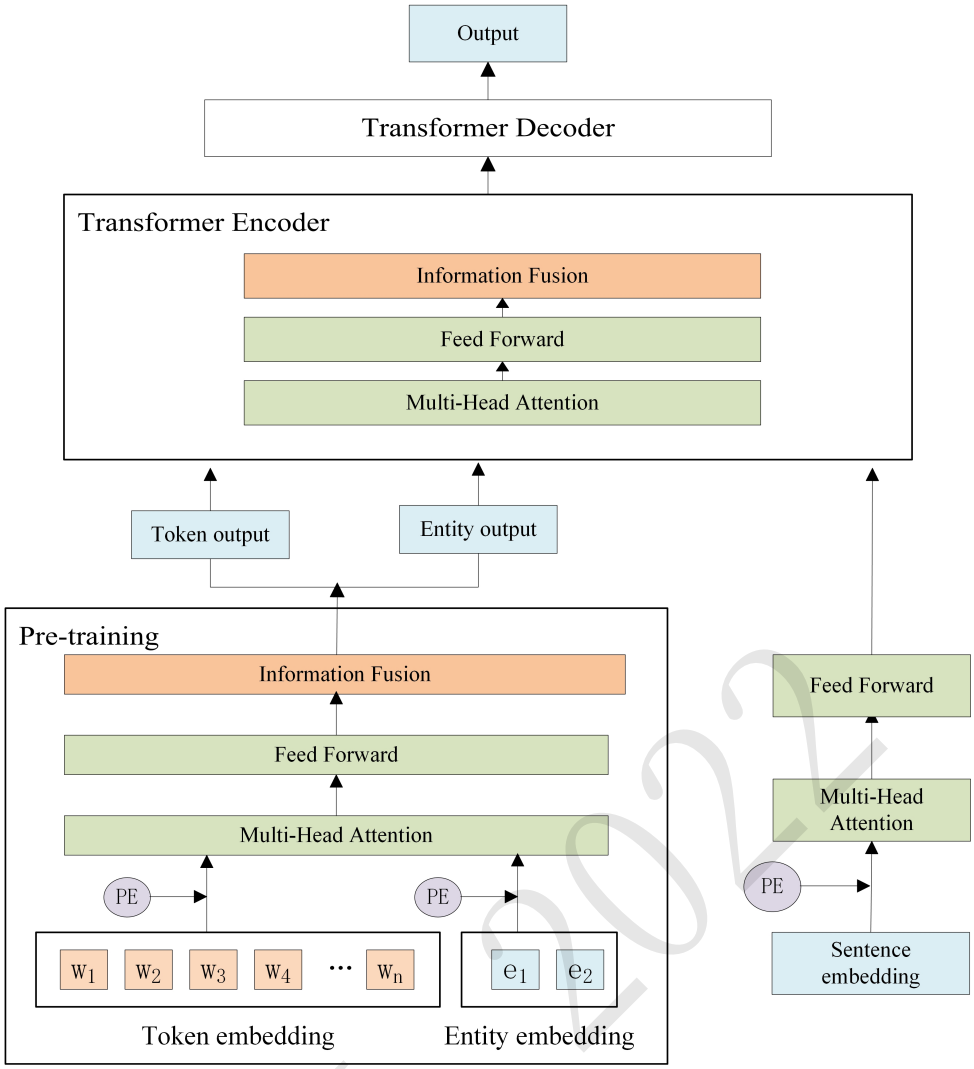


Figure 1: 基于实体信息增强及多粒度融合的多文档摘要模型MGNIE

3.1 嵌入知识图的预训练模型ERNIE

本文通过加入知识图结构来增强预训练模型的事实一致性，采用transE算法将Wikidata知识图的实体-关系信息输入到ERNIE模型进行训练，然后将带有实体信息的原文本输入到预训练模型中进行预训练从而得到词嵌入和实体嵌入。

我们将词序列集合定义为$W = \{w_1, w_2, \ldots, w_n\}$，其中n表示词序列的长度，将实体序列定义为$E = \{e_1, e_2, \ldots, e_m\}$,其中m为实体序列的长度，将句子序列表示为$S = \{s_1, s_2, \ldots, s_o\}$,其中o为句子序列的长度。知识图$KG_s$中的所有实体表示为E，我们将源文本中的实体与$KG_s$中的实体对齐。

在预训练模型中，我们将词信息与KG_s中的实体信息分别进行编码，然后输入到前馈神经网络层，进行异质信息融合。首先对词序列进行编码，将词嵌入e_w和段嵌入s_w以及位置嵌入p_w相加获得最终词嵌入:

$$h_w^0 = e_w + s_w + p_w \tag{1}$$

同样的，实体嵌入可计算为:

$$h_e^0 = e_e + s_e + p_e \tag{2}$$

则获得最终的词嵌入$h_w^{l-1} = \{h_{w_1}^{l-1}, h_{w_2}^{l-1}, \ldots, h_{w_n}^{l-1}\}$和实体嵌入$h_e^{l-1} = \{h_{e_1}^{l-1}, h_{e_2}^{l-1}, \ldots, h_{e_m}^{l-1}\}$,将它们作为输入，送入多头注意力中，

第二十一届中国计算语言学大会论文集，第155页-第165页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

$$h_w^l = MHAtt(h_{w_i}^{l-1}, h_{w_n}^{l-1}) \tag{3}$$

$$h_e^l = MHAtt(h_{e_j}^{l-1}, h_{e_m}^{l-1}) \tag{4}$$

原文本中的词token嵌入包含实体信息，与KG_s中的实体对齐，并将实体信息融入到词序列中进行异质信息融合，实现了外部实体嵌入的信息增强。源文本词序列token包含实体的表示为h_w，KG_s中的实体在源文本序列中有对应的表示为h_e，则可以表示为$h_e = f(h_w)$，融合后的表征为:

$$h_1 = \sigma(h_w^l, h_e^l) = (W_t w_i + W_e e_j + b) \tag{5}$$

$$h_w = \sigma(h_1 W_t + b_t) \tag{6}$$

$$h_e = \sigma(h_1 W_e + b_e) \tag{7}$$

其中W_t、W_e、b表示可训练的权重参数，h_1表示整合了token和实体信息的内部隐藏状态，本文使用非线性激活函数$GELU$对实体嵌入和词嵌入进行融合。对于没有相应实体的词嵌入，不进行信息融合而直接输出。为了简化，我们将经过预训练模型的词嵌入向量h_w仍表示为$h_w^l = \{h_{w_1} l, h_{w_2}^l, \ldots, h_{w_n}^l\}$，实体嵌入向量$h_e$表示为$h_e^l = \{h_{e_1}^l, h_{e_2}^l, \ldots, h_{e_m}^l\}$

本文将外部KG_s中的实体信息融入到源文本中，通过mask表示实体的词token，通过上下文对进行实体预测来预训练模型，使模型获得更丰富以及更高语义的信息，从而能够生成更好的表征。预训练模型mask实体自动编码过程的损失函数可以用下述公式来计算:

$$p(e_x|w_i) = \frac{exp(linear(w_i^0)e_x)}{\sum_{j=1}^{m} exp(linear(w_i^0)e_j)} \tag{8}$$

其中$linear()$表示一个线性层。

3.2 多粒度信息融合

本文采用多粒度信息，包括词嵌入、实体嵌入和句子嵌入，在预训练模型阶段，我们将外部的实体信息与原文本包含的实体信息进行融合从而获得实体级的信息增强，在摘要模型输入阶段，我们分别将源文本划分为词序列、实体序列和句子序列，并对不同粒度的信息进行信息融合，从而获得包含更加丰富语义的语言模型表征。

我们将经过预训练模型的源文本中的词向量嵌入表示为$h_w^{l-1} = \{h_{w_1}^{l-1}, h_{w_2}^{l-1}, \ldots, h_{w_n}^{l-1}\}$，将源文本中提取的实体向量嵌入表示为$h_e^{l-1} = \{h_{e_1}^{l-1}, h_{e_2}^{l-1}, \ldots, h_{e_m}^{l-1}\}$，采用同样的方法，我们可以获得句子序列表示$S = \{s_1, s_2, \ldots, s_o\}$进行编码获得句子嵌入:

$$h_s^0 = e_s + s_s + p_s \tag{9}$$

则句子嵌入表示为$h_s^{l-1} = \{h_{s_1}^{l-1}, h_{s_2}^{l-1}, \ldots, h_{s_o}^{l-1}\}$，送入多头注意力，可以得到句子的上下文信息:

$$h_s^l = MHAtt(h_{s_z}^{l-1}, h_{s_o}^{l-1}) \tag{10}$$

对多粒度信息进行融合以获得对源文本更加丰富的特征，实体信息的融合体现了生成摘要过程中对事实的准确性。我们加入的实体特征是在源文本中出现的，且能够链接到外部知识KG_S的结构化的实体关系。我们使用融合函数进行融合，首先获得词$token$与实体的融合信息h_w。

融合后的实体信息与词信息表示为部分词$token$融入了实体的嵌入信息，再将融合后的词序列信息与句子序列信息进行融合，得到融合后的词向量h_w':

$$h_2 = \sigma(h_1, h_s^l) = \sigma(\sigma(h_w^l, h_e^l), h_s^l) \tag{11}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

$$h_w^{'} = \sigma(h_2 W_t + b_t) \tag{12}$$

将获得的融合的词表征进一步送入前馈神经网络用来进一步的转化丰富的语义信息：

$$h = LayerNorm(h^{l-1} + h_w^{'}) \tag{13}$$

$$h^l = LayerNorm(h + FFN(h)) \tag{14}$$

FFN是两层前馈网络，采用ReLU隐藏激活函数，其中layerNorm是层规范化。h^l表示编码器的输出。将经过编码器融合的输入向量h^l以及隐藏状态输入到transformer解码器中进行逐词解码，编码器输出作为key和value，将输入嵌入和词位置编码输入到解码器中经过多头注意力机制以及前馈神经网络层，得到上下文表征作为query，输入多头注意力机制中，得到输出g^l，最后送入softmax，来计算目标词汇生成分布：

$$P_t = softmax(g^l W_g + b_g) \tag{15}$$

其中W_g、b_g为可训练的参数。本文使用的交叉熵损失函数为：

$$L = -\frac{1}{N}\sum_{n=1}^{N} logP(y_w^{(n)}) \tag{16}$$

其中$y_w^{(n)}$表示生成的真实摘要，N表示语料库的样本数。

3.3 Wikidata知识图实体嵌入

我们采用外部知识图来对文本的实体信息进行增强，采用transE算法将Wikidata知识图的实体-关系信息输入到ERNIE模型进行训练。Wikidata知识图是一个开放的多关系知识图谱，它包含了包括维基百科的结构化的数据，我们从Wikidata知识图抽取实体-关系三元组，并且通过transE算法学习实体嵌入。该算法将关系数据中的实体和关系嵌入低维向量空间。给定一个实体-关系三元组(h,l,t)，他们由h头实体，t尾实体，l关系组成，通过模型学习实体和关系的嵌入向量，算法的原理就是通过边所对应的关系对应于嵌入的转换，当(h,l,t)成立时，使得头实体向量和关系向量尽可能的靠近尾实体向量，并计算(h,l)和t之间的距离。

transE训练模型原理是从实体矩阵和关系矩阵中各自抽取一个向量，进行运算得到的结果近似等于实体矩阵中另一个实体的向量，从而达到通过词向量表示知识图中已存在的三元组。transE的损失函数为：

$$\mathcal{L} = \sum_{(h,l,t)\in S} \sum_{(h',l,t')\in S'_{(h,l,t)}} [\gamma + d(h+l,t) - d(h^{'}+l,t^{'})]_+ \tag{17}$$

$$S'_{(h,l,t)} = \{(h^{'},l,t) \in E\} \cup \{(h,l,t^{'}) | t^{'} \in E\} \tag{18}$$

其中公式中$S^{'}$表示头实体或尾实体被替换的负采样三元组。

4 实验

4.1 数据集与评价指标

MultiNews数据集由(Fabbri et al.,2019)提出，MultiNews数据集由新闻文章和人工撰写的摘要组成。该数据集来自不同的新闻来源(超过1500个网站)。MultiNews更类似于传统的多文档只摘要数据集，如DUC，但规模更大。正如Fabbri等人所述，数据集分为44，972个用于训练的实例，5622个用于验证的实例和5622个用于测试的实例。源文档和输出摘要的平均长度分别为2103.5个标记和263.7个标记。我们将源文档截断为句子S，并按照原始顺序将句子序列连成一个序列。我们使用Stanford coreNLP工具对数据集进行预处理，并采用TAGME实体链接工具提取源文档中的实体token。本文使用F1 ROUGE对生成摘要与标准摘要进行评估。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 基线和实施细节

在对比实验中，本文将提出的模型与现有几种先进的方法进行了比较：Lead是连接标题和排序的段落，并提取前k个标记；LexRank (Erkan et al., 2004)是一个广泛使用的基于图形的抽取式摘要，类似PageRank的算法来排列和选择段落；MMR (Carbonell et al.,1998)提取具有排序列表的句子基于相关性的候选句子和冗余；HIBERT (Zhang et al., 2019) 提出用BERT模型对句子进行预训练，然后对整个文档进行编码；PGN(See et al.,2017)是一个基于RNN的模型，具有注意力机制，允许系统通过指向从源文本复制单词进行抽象概括；Hierarchical Transformer (HT) (Liu et al.,2019)该模型将标题和段落作为输入来产生目标摘要；Hi-MAP(Fabbri et al.,2019)将指针生成器网络模型扩展为分层网络，并集成MMR模块来计算句子级得分；Flat Transformer (FT)是将基于转换器的编码器-解码器模型应用于平面令牌序列的基线；MGSum(Jin et al.,2020)是一个用于抽取式和生成式多文档摘要网络，联合学习了单词、句子和文档的语义表示。CTF+DPP(Perez-Beltrachini et al.,2021)提出基于DPP的注意力模型，将注意力权重用行列式点过程(DPP)给出的概率来计算，并将注意力机制与已有的模型相结合。

本文使用Pytorch来实现提出的模型，使用Stanford coreNLP工具对数据集进行预处理。优化器是Adam (Kingma et al.,2014)，学习率为$2e^{-5}$，$\beta_1 = 0.9$，$\beta_2 = 0.998$。所有模型都在1个GPU上进行500，000步的训练。模型中的所有线性层之前应用概率为0.1的下降。最大序列长度设置为256，batch size大小为32，模型中的隐藏单元的数量被设置为256，前馈隐藏大小为1024，头的数量为8。

4.3 MGNIE整体性能分析：

我们在MultiNews数据集上与先前的几种模型进行了对比，实验结果如下表，MGNIE表示本文提出的模型，它包含嵌入了实体信息的预训练模型，以及多粒度融合信息的摘要模型。MGNIE-BERT将预训练模型换成BERT模型，表明预训练模型中不包含实体关系。MGEIN模型在无论是在ROUGE-1，ROUGE-2还是ROUGE-L，较先前的工作都取得了优秀的性能，并有所提升。

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead	41.24	12.91	18.84
LexRank	38.27	12.70	13.20
MMR	38.77	11.98	12.91
TIBERT	43.86	14.62	18.34
MGSum-ext	44.75	15.75	19.30
MGNIE-BERT	45.75	17.01	20.39
MGNIE	**46.80**	**17.22**	**22.85**

Table 1: 在MultiNews上抽取式模型的对比实验

表1是本文提出的模型与抽取式模型在MultiNews数据集上的对比实验结果，根据结果我们可以发现，本文提出的模型较先前提出的基于抽取式的模型有较大提升。与TIBERT模型相比，我们采用BERT模型做为预训练模型，同时采用多粒度信息融合的方法，较TIBERT模型提升了1.89，表明我们的对多粒度信息进行融合方法对生成式模型的摘要生成有提升作用。与MGSum-ext模型相比，本文提出的模型在多粒度信息融合的基础上，加入了BERT预训练语言模型，较没有预训练模型的摘要模型提升了1.0，同时本文提出的模型在预训练模型中加入了来自知识图的实体信息，进一步取得了模型性能的提升。

Model	ROUGE-1	ROUGE-2	ROUGE-L
PGN	41.85	12.91	16.46
HT	42.36	15.27	22.08
Hi-MAP	43.47	14.89	17.41
FT	44.32	15.11	20.50
MGSum-abs	46.00	16.81	20.09
CTF+DPP	45.84	15.94	21.02
MGNIE-BERT	45.75	17.01	20.39
MGNIE	**46.80**	**17.22**	**22.85**

Table 2: 在MultiNews上生成式模型的对比实验

(c) 2022 中国中文信息学会计算语言学专业委员会

表2 是本文提出的模型与生成式模型MultiNews数据集上的对比实验结果，根据结果我们可以发现本文提出的模型较先前的具有先进性的生成式模型取得了更好的效果。本文的模型相比于MGSum-abs模型提升了0.8，这表明在多粒度信息融合方面，本文的模型加入了实体信息，并将知识图的实体信息与文本中的实体信息进行融合做预训练模型，更加丰富了文本的上下文表征，从而增强模型理解能力，对于模型性能的提升有更好的影响。可以观察到，在多粒度的基础上将预训练模型换成BERT模型，较于先前的工作有部分提升，而在预训练模型的基础上，不仅对原文本进行模型训练，还加入了外部知识的实体信息，使得预训练获得文本向量表示的效果更加好进一步展开分析，在ROUGE分数评估下，MGNIE以及MGNIE-BERT模型都优于现有的工作模型表现，说明我们采用多粒度的方法进行信息提取获得语义表征，能够挖掘到更多信息特征，从而获得更好的生成摘要效果。

通过观察表1和表2的数据我们可以发现，本文提出的模型对比先前的抽取式模型的提升比生成式模型的效果要好，因为我们使用了融合了丰富知识图信息的预训练模型对原文本进行训练，并且采用的多粒度信息融合来对文本的长距离信息进行了交互，实现了实体的信息增强，对于关键词的解码生成有重要的影响，本文提出的模型生成的摘要文本会相比于抽取式模型生成的摘要更具有连贯性，同时说明了实体信息对于摘要生成的有效性。

4.4 实体信息对摘要性能的影响：

本文通过引入实体信息增强来实现对摘要性能的提升，为了探究实体信息对摘要性能的影响，本文做了相关的消融对比实验。表3给出了不同粒度对实验结果的影响，其中without sent representation表示在多粒度融合中包含词信息和实体信息融合的模型，without entity representation表示在多粒度融合中包含词信息和句子信息融合的模型，其中MGNIE-BERT表示将预训练模型换成BERT模型，不加入外部实体信息的模型。MGNIE表示我们在文中提出的模型，加入外部实体信息增强的多粒度融合模型。

Model	ROUGE-1	ROUGE-2	ROUGE-L
without sent representation	45.51	16.32	19.57
without entity representation	44.02	15.98	19.23
MGNIE-BERT	45.75	16.01	20.39
MGNIE	**46.80**	**17.22**	**22.85**

Table 3: 消融实验对比

从表3的结果中我们可以发现without entity representation不加入实体信息表示表现不佳，表明加入实体表征的是非常有效的。并且MGNIE的表现比MGNIE-BERT要好，表明在预训练模型过程中进行外部实体信息嵌入式对预训练模型的效果有重要提升。通过在预训练模型中加入外部知识图谱的结构化的实体关系信息，以及在自然语言编码是加入实体表征可以充分的挖掘文本中的实体信息以及上下文语义信息，从而是模型获得更好的效果。

4.5 人工评测：

由于评估摘要生成的流畅性以及事实一致性在摘要生成中是十分重要的，所以进行人工评测是必不可少的。具体来说，本文选择5名研究生来对本文生成的摘要进行评估，在MultiNews数据集中随机选择50个样本，为了评估模型的质量，本文选择MGSum模型作为基线模型来进行对比，并从三个方面来进行评估：流畅性(fluency)，信息量(Informativeness)以及与原文本的忠实度(faithfulness)，流畅性是指文本的可读性，包括语法、名词短语和逻辑上的一致性。信息量表示摘要与原文包含的关键内容相关性的数量。忠实度是指摘要与原文的事实一致的相关性。本文选取的评分标准为1-5分，分数越大说明性能越好。

Model	Fluency	Informativeness	Faithfulness
MGSum	3.48	3.26	3.31
MGNIE	**3.83**	**3.58**	**3.96**

Table 4: 人工评测结果

表4是在MGSum模型与本文的模型的人工评估结果，从表中可以观察到本文的模型在流畅性、信息量以及忠实度方面较MGSum都有提升，尤其在忠实度方面的评估结果表明本文提出的基于实体信息增强来提升生成摘要的事实一致性是有效的。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.6 摘要生成实例分析：

Source Text
Document 1: san francisco (marketwatch)-trading in all nasdaq-listed stocks and options was halted on thursday due to technical problems on the bourse, according to nasdaq omx group (nasdaq:ndaq). **the exchange sent out a series of emails alerting investors that it was experiencing issues with " quote submissions. "** in response, the new york stock exchange has also stopped trading in all nasdaq securities at the request of nasdaq omx. " **all orders in those securities have been cancelled back to customers , " said nyse in a statement.** the nasdaq composite index (nasdaq:comp) was last at 3631.17, up 31.38 points, before trading was suspended. there was no immediate word on when transactions will resume. Document 2 : updated with nyse developments. **"a technical glitch knocked out trading in all nasdaq stock market securities for three hours thursday afternoon, an unprecedented meltdown for a u.s. exchange that paralyzed a broad swath of markets and highlighted the fragility of the financial world's electronic backbone. "nasdaq officials scrambled to figure out what happened and resume trading.** they shared few of their findings with trading firms or the public during regular trading hours, sowing confusion across wall street and leaving many investors frustrated. ", "the decision to reopen trading with about 35 minutes to go before the close came after exchange officials were sure that banks ... "
Gold
nasdaq is back in business after an apparent technical glitch brought the exchange to a rare halt this afternoon for more than three hours, reports the wall street journal. **the exchange hasn't fully explained what happened,** but trading of all nasdaq securities ground to a halt just after noon today, reports marketwatch. other exchanges quickly suspended trading of nasdaq stocks. " **all orders in those securities have been canceled back to customers, " says the new york stock exchange in a statement. nasdaq blamed " quote submissions " in an email to investors.**
Our Model
Nasdaq officials are scrambling to figure out what happened and resume trading in all nasdaq stocks and options, reports marketwatch. **the glitch knocked out of all nasdaq stock market securities for three hours thursday afternoon, an unprecedented meltdown for a us exchange that paralyzed a broad swath of markets and " highlighted the fragility of the financial world's electronic backbone, "** reports the wall street journal. the move came after officials were sure that banks would reopen trading with 35 minutes to go before the close.**"the exchange sent out a series of emails alerting investors that it was working. "** said nasdaq omx group.

Table 5: MGNIE模型摘要生成实例

表5展示了在MGNIE模型上摘要生成的实例与原文本以及标准摘要的对比，加粗的文字表示与原文的关键内容重合的部分。从表中我们可以观察到本文提出的模型在重合度方面与原文内容高度重合，在信息量方面也提取了大量关键信息，捕捉到了“three hours thursday afternoon”这个时间点，以及“the glitch”这个信息点，并且与标准摘要进行对比可以发现，本文模型生成的摘要在内容信息量以及内容重合度都很高。在事实一致性方面，可以发现生成的摘要无论是与原文的对比，还是对标准摘要的对比上都是保持事实一致的。在摘要的流畅度方面，我们可以发现生成的文本是可读的，并且句子之间的连接词使得文本承接上下文具有连贯性以及逻辑性。

5 结论

在本文中，我们针对生成式多文档摘要中存在的缺乏结构化信息的嵌入以及生成摘要的事实不一致性提出了基于实体信息增强以及多粒度信息融合的多文档摘要模型，具体来说，我们加入了预训练模型ERNIE并且将外部知识图中的实体-关系信息嵌入预训练模型以丰富语义信息，与此同时，我们还获取了词信息、实体信息以及句子信息层面的信息融合来编码语言表征，实现了更深层次的信息挖掘，实现了信息增强。最后进行了大量的对比实验表明，本文提出的方法在多文档摘要中取得了有效影响，在一定程度上实现了信息增加解决了事实一致性问题。

在未来的工作中，我们还将考虑将文本与外部知识图结构进行相互融合，将文本转换为结构化的图与外部知识图相连从而获得结构化的信息，来实现基于图结构的多文档摘要模型的性能提升。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Alambo A, Lohstroh C, Madaus E, et al. 2020. Alternation. *Topic-centric unsupervised multi-document summarization of scientific and news articles[C]//2020 IEEE International Conference on Big Data (Big Data). IEEE*, 2020: 591-596.

Bordes A, Usunier N, Garcia-Duran A, et al. 2013. Alternation. *Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems.*

Carbonell J, Goldstein J. 1998. *The use of MMR, diversity-based reranking for reordering documents and producing summaries[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998: 335-336.

Christensen J, Soderland S, Etzioni O. 2013. *Towards coherent multi-document summarization[C]//Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2013: 1163-1173.

Dong Y, Wieting J, Verga P. 2022. *Faithful to the Document or to the World? Mitigating Hallucinations via Entity-linked Knowledge in Abstractive Summarization[J]*, arXiv preprint arXiv:2204.13761, 2022.

Erkan G, Radev D R. 2004. *Lexrank: Graph-based lexical centrality as salience in text summarization[J]. Journal of artificial intelligence research*, 2004,22: 457-479.

Fabbri A R, Li I, She T, et al. 2019. *Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model[J]*, arXiv preprint arXiv:1906.01749, 2019.

Hendrycks D, Gimpel K. 2016. *Gaussian error linear units (gelus)[J]*, arXiv preprint arXiv:1606.08415.

Jin H, Wang T, Wan X. 2020. *Multi-granularity interaction network for extractive and abstractive multi-document summarization[C]//Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020: 6244-6254.

Kingma D P, Ba J. 2014. *Kingma D P, Ba J. Adam: A method for stochastic optimization[J]*, arXiv preprint arXiv:1412.6980.

Lee J D M C K, Toutanova K. 2018. *Pre-training of deep bidirectional transformers for language understanding[J]*, arXiv preprint arXiv:1810.04805.

Liu Y, Lapata M. 2019. *Hierarchical transformers for multi-document summarization[J]*, arXiv preprint arXiv:1905.13164.

Pasunuru R, Celikyilmaz A, Galley M, et al. 2021. *Data augmentation for abstractive query-focused multi-document summarization[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021)*, 2021: 13666-13674.

Perez-Beltrachini L, Lapata M. 2021. *Multi-document summarization with determinantal point process attention[J]. Journal of Artificial Intelligence Research*, 2021, 71: 371-399.

See A, Liu P J, Manning C D. 2017. *Get to the point: Summarization with pointer-generator networks[J]*, arXiv preprint arXiv:1704.04368.

Szegedy C, Vanhoucke V, Ioffe S, et al. 2016. *Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016: 2818-2826.

Vaswani A, Shazeer N, Parmar N, et al. 2017 *Attention is all you need[J]. Advances in neural information processing systems.*

Yasunaga M, Zhang R, Meelu K, et al. 2017. *Graph-based neural multi-document summarization[J]*, arXiv preprint arXiv:1706.06681.

Zhang X, Wei F, Zhou M. 2019. *HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization[J]*, arXiv preprint arXiv:1905.06566.

Wang Y, Sun Y, Ma Z, et al. 2019. *Zhang Y. Evaluating the factual correctness for abstractive summarization[J]. CS230 Project.*

(c) 2022 中国中文信息学会计算语言学专业委员会

Zhang Z, Han X, Liu Z, et al. 2019. *ERNIE: Enhanced language representation with informative entities[J]*, arXiv preprint arXiv:1905.07129

Zhao J, Liu M, Gao L, et al. 2019. *Summpip: Unsupervised multi-document summarization with sentence graph compression[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020: 1949-1952.

Zheng X, Sun A, Li J, et al. 2019. *Subtopic-driven multi-document summarization[C]//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: 3153-3162.

CCL 2022

融合提示学习的故事生成方法

倪宣凡　李丕绩*
计算机科学与技术学院/人工智能学院,
南京航空航天大学
江苏省，南京市，210016
xuanfanni@gmail.com, pjli@nuaa.edu.cn

摘要

开放式自动故事生成通过输入故事的开头、大纲、主线等，得到具有一致性、连贯性和逻辑性的故事。现有的方法想要提升生成故事的质量，往往需要大量训练数据和更多参数的模型。针对以上问题，该文利用提示学习在零样本与少样本场景下的优势，同时使用外部常识推理知识，提出了一种故事生成方法。该方法将故事生成分为三个阶段：输入故事的开头，常识推理模型生成可能的事件；根据类型不同，将事件填入问题模板中，构建引导模型生成合理回答的问题；问答模型产生对应问题的答案，并选择困惑度最小的作为故事下文。重复上述过程，最终生成完整的故事。自动评测与人工评测指标表明，与基线模型相比，该文提出的方法能够生成更连贯、具体和合乎逻辑的故事。

关键词： 故事生成 ； 预训练模型 ； 提示学习

A Story Generation Method Incorporating Prompt Learning

Xuanfan Ni, Piji Li
College of Computer Science and Technology/Artificial Intelligence,
Nanjing University of Aeronautics and Astronautics
Nanjing, Jiangsu 210016, China
xuanfanni@gmail.com, pjli@nuaa.edu.cn

Abstract

Open-ended automated story generation obtains a consistent, coherent and logical story by entering the beginning, out-line, or main line of story. To improve the quality of generated stories, existing methods often require a large amount of training data and models with more parameters. Aiming at the above problems, this paper proposes a novel story generation method. The method divides the story generation into three stages: input the beginning of story, and the common sense reasoning model generates possible events; according to different types, fill in the events into the question template to construct questions; the question answering model generates answers to corresponding questions, and selects the one with lowest PPL score as story below. Repeat above process to finally generate a complete story. Automatic evaluation metrics show that the proposed method is able to generate more coherent, specific, and logical stories than baseline models.

Keywords: story generation , pre-trained model , prompt learning

* 通讯作者

第二十一届中国计算语言学大会论文集，第166页-第177页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

开放式自动故事生成是自然语言处理(Natural Language Processing, NLP) 领域中一个非常经典的任务(Alabdulkarim et al., 2021)。在深度神经网络出现之后，特别是随着更大参数、更好架构的模型被提出，故事生成也取得了长足的发展。故事生成与其它的自然语言生成任务不同。机器翻译需要源语言与目标语言之间的匹配；文本摘要需要抽取输入文本的重要信息并进行填充，句子的结构、逻辑和语义大部分来自输入；而故事生成考验的是模型从训练故事中学到的知识，同时要兼顾连贯性和一致性。现在主流的做法都是探究故事的结构，由点到面，逐步生成(Ansag and Gonzalez, 2021)。但是考虑的角度不同、结构不同，性能会有很大的差异。

现有的故事生成模型或系统能够在连贯性和一致性上取得不错的效果，如采用Vaswani et al. (2017)提出的Transformer架构的GPT(Radford et al., 2018)，GPT2(Radford et al., 2019)，GPT3(Brown et al., 2020)等自回归语言模型；Liu et al. (2020)提出以角色为中心的神经故事模型；Tambwekar et al. (2018)训练实现给定故事目标或结局的神经语言模型；Fan et al. (2018)将故事生成分层，并训练模型定期给出指导。Yao et al. (2019)则在此基础上，使用高级故事生成计划来引导模型进行生成。这些工作往往需要大量的训练数据和结构复杂、参数量多的模型，在很多场景下是难以满足的。针对这一问题，使用预训练模型和注入外部知识来辅助生成是很好的解决方法：Ammanabrolu et al. (2021)通过常识推理、因果关系和情节顺序来构建故事生成系统，其中常识推理由COMET模型(Bosselut et al., 2019)给出；Guan et al. (2020)提出了知识增强的预训练模型，利用来自外部知识库的常识知识来生成合理的故事。但是，以前的研究更多是将这些外部知识作为数据，参与模型的训练。那么，是否有更好的外部知识使用方法?

最近，提示学习(Prompt Learning)的相关研究与应用发展的如火如荼(Liu et al., 2021)。大量工作都表明，提示学习在少样本和零样本场景下有着一般微调(Finetune)所不及的优势。例如，有一个文本情感分类的任务，对于输入“我爱这里的食物。”，去判断这句话的情感是积极的还是消极的。提示学习将输入重构成“我爱这里的食物。我觉得...”，然后交由预训练模型，如GPT-2等去生成。正面内容代表积极，负面内容代表消极。在这个例子中，重构后的输入被用来引导模型进行生成，原先的文本分类任务则被转化为文本生成任务。

仅仅更改输入的形式，在没有微调的情况下，预训练模型就可以执行不同于训练阶段输入的数据形式的下游任务。受提示学习的启发，我们将其应用到故事生成中，通过提示模板来重构任务形式，在少样本和零样本场景下，保证生成故事的质量。重构成何种任务形式是我们非常关注的一点。直觉上来讲，越相近的任务，重构的难度就越低，效果也越好。因此我们考虑同为文本生成任务的问答。相比较原先的故事生成任务与其他文本生成任务（文本摘要、机器翻译等），问答的优势有：

- 问答任务的输入通常是文档与问题，模型从文档与训练时学到的知识来对问题进行回答。将故事前文作为文档，通过模板构造合理的问题。生成的答案可以较好地保持一致性与连贯性；
- 在外部常识知识的帮助下得到的问题，可以看作是对上文的总结与后续发展的推测。使用这种优质的问题去引导问答模型进行回答，能够在保证逻辑性的情况下，推动故事发展；
- 应用于故事生成任务的自回归预训练模型，如果没有充足的训练数据，产生的内容很容易出现冗余、逻辑错误等情况；机器翻译、文本摘要任务都只能对输入进行总结与分析，其输出无法推动故事发展；而应用于问答任务的预训练模型，通过优质的问题进行引导，生成的内容是简短且与上文相关的答案，将这些答案作为故事下文能很大程度上避免这些不足之处。

通过以上思考，本文提出了一种故事生成方法。该方法将故事生成分为三个阶段：(1) 输入故事的开头，常识推理模型生成多个可能的事件；(2) 根据类型，将事件填入问题模板中，构建合理的问题；(3) 问答模型产生对应问题的答案，并选择困惑度最小的作为故事下文。重复上述过程，最终生成完整的故事。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

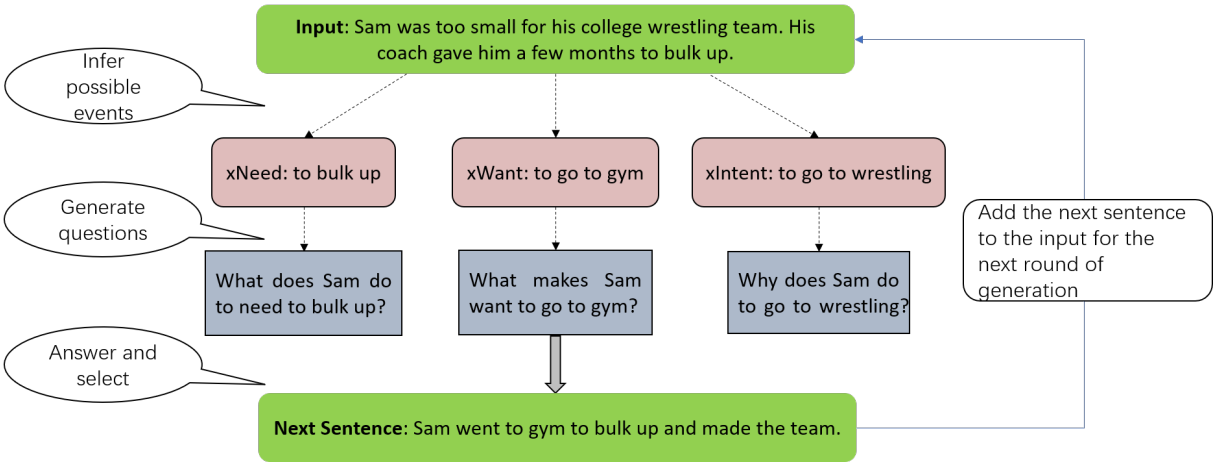


Figure 1: 融合提示学习的故事生成方法的示例

图1展示了本文提出的故事生成方法的一个样例。如图所示，对于输入的故事开头*Sam was too small for his college wrestling team. His coach gave him a few months to bulk up.* 常识推理模型生成多个可能的事件，如*to bulk up, to go to gym, to go to wrestling*等。这些事件有各自的类型，且与故事开头紧密相连。根据事件类型，构建一系列问题，对应上文三个事件，则为：*What does Sam do to need to bulk up? What makes Sam want to go to gym? Why does Sam do to go to wrestling?* 针对这些问题，模型生成答案，并选择句子*Sam went to gym to bulk up and made the team.*作为下文故事。通过优质的问题模板、合理的事件推理，这些问题能够很好地引导问答模型进行生成。最终，从生成答案中选择困惑度最低的作为故事下文。

我们选择Para-COMET(Gabriel et al., 2021)作为常识推理模型，为选定句子生成推理的同时结合故事中的其他句子；针对Para-COMET产生的事件，我们使用RoBERTa模型(Zhuang et al., 2021)找出与之对应的人物角色，并根据事件类型，构建提示问题模板，生成问题；我们使用ELI5QA模型(Fan et al., 2019a)和BART模型(Lewis et al., 2019)来对问题进行回答，并使用GPT2模型计算答案的困惑度。

本文的主要创新点有：

- 针对深度神经网络模型的训练缺少数据集的问题，本文提出的方法结合了新兴的提示学习思想，通过构建优质的提示模板，充分激发预训练模型的潜能，帮助模型回忆起训练时学到的知识，来较好地完成下游任务；
- 在故事生成过程中，通过常识推理构建的问题，可以看做是对前文内容的多角度、多方面的总结，并引导模型产生合理的下文句子；
- 本文在零样本与少样本场景下进行了实验，评测结果证明了该故事生成方法的有效性。进一步的消融实验突出了优质提示问题模板的重要性。

2 相关工作

2.1 故事生成方法

Liu et al. (2020)提出了以角色为中心的故事生成神经模型。其做法是，为一个故事分配一个角色，在上下文环境下生成该角色的一系列动作，最终生成完整的故事。由于在故事生成的每个阶段，给定的角色都参与选择动作，因此生成的故事具有很好的一致性。Fan et al. (2019b)为了解决长文本故事的一致性问题，提出了一种结构化故事生成模型。模型对动作序列、故事叙述以及命名实体进行建模。模型产生实体匿名故事，并用之前识别出来的实体去替换故事中的占位符。从不同角度来解构故事，在保证连贯性和一致性的前提下，将故事文本拆分细化，并对不同组成成分进行建模。本文提出的方法也将故事生成任务分解为多个阶段，每个阶段都使用不同的预训练模型，执行不同的任务。

(c) 2022 中国中文信息学会计算语言学专业委员会

Tambwekar et al. (2018)通过控制结局和事件顺序来控制故事情节。他们使用强化学习技术来优化预训练的序列到序列模型。他们的方法比单独的基础模型要好。但是，这种方法需要针对每个新的下游任务重新训练模型。Fan et al. (2018)将生成过程分为两个层次：前提和故事，来解决情节可控性问题。他们使用卷积网络首先生成一个写作提示，然后，该提示成为序列到序列模型的输入，并指导它生成以提示为条件的故事。这种方法通过直接编写故事提示，来适应不同的任务，但是编写的提示较为单调，这导致生成的故事缺乏趣味性。Yao et al. (2019)提出了计划写作(Plan-and-write)故事生成框架：该框架将故事的标题作为输入，然后生成故事情节。接着将故事情节和标题用作输入以控制在序列到序列模型中的故事生成。该模型存在几个主要问题：重复、偏离主题和逻辑不一致。这些模型采用分层故事生成，需要大量训练数据，且生成的内容往往会存在一定的缺陷。与之形成对比的，本文提出的方法在零样本场景下也能生成较好的故事。

Ammanabrolu et al. (2021)通过常识推理、因果关系和情节顺序来构建故事生成系统C2PO。他们将故事生成问题视为情节填充，从训练集中中提取情节点的轮廓，然后对其详细说明。在C2PO系统中使用软因果关系填充情节来生成叙事——创建一个可能的故事延续的分支空间，从COMET(Bosselut et al., 2019) 常识推理模型中迭代地提取常识因果推理。Guan et al. (2020)提出了知识增强预训练模型，利用来自外部知识库的常识知识来生成合理的故事。他们将常识知识编码，与大规模语料一起输入进Transformer模型中进行训练。还有很多类似的工作，都是将外部知识作为增强模型性能的手段。我们的工作则使用外部知识直接参与故事生成，同样为了提高生成故事的合理性与逻辑性。

2.2 提示学习

Liu et al. (2021)完成了一篇综述论文。论文总结了近几年提示学习的相关工作，并提出NLP中的新范式：预训练、提示、预测(Pre-train, Prompt, Predict)。他们从五个方面对提示学习方法做了一个介绍：预训练模型(Pre-trained Models)、提示工程(Prompt Engineering)、答案工程(Answering Engineering)、多提示学习(Multi-Prompt Learning)以及基于提示的训练策略(Prompt-based Training Strategies)。

在这篇综述之前，就有工作涉及到这种概念，也表现出提示学习的一些潜力。Li and Liang (2021)提出了前缀调优(Prefix-tuning, PT)。PT不改变模型参数，只是对不同的下游任务训练不同的连续向量，这个向量被称为前缀(Prefix)。在文本摘要任务中，前缀，输入，输出一起拼接，然后交由GPT2模型去训练。PT是连续提示的一种，这体现了提示模板形式的多样性：不一定是离散token，也可以是数字，符号，向量，词嵌入甚至是图片、音频、视频(已有工作将图片提示模板应用到计算机视觉中)。

在此之后，提示学习在文本生成领域的应用也被广泛探索。Castricato et al. (2021)通过模板构建提示，从后往前生成以目标事件为结尾的故事，这里的提示模板是对已生成内容的解释。而我们的工作构建的问题是对前文内容的总结，并进行推理生成。Lin et al. (2022)通过训练好的常识推理模型，生成对应输入事件的常识知识提示，来引导未来事件生成，提示形式是潜在的常识表征。我们的工作使用的提示是通过将推理事件填入问题模板中得到的，是具体的内容，相比之下具有更好的可解释性与可控性。

3 结合提示学习的故事生成

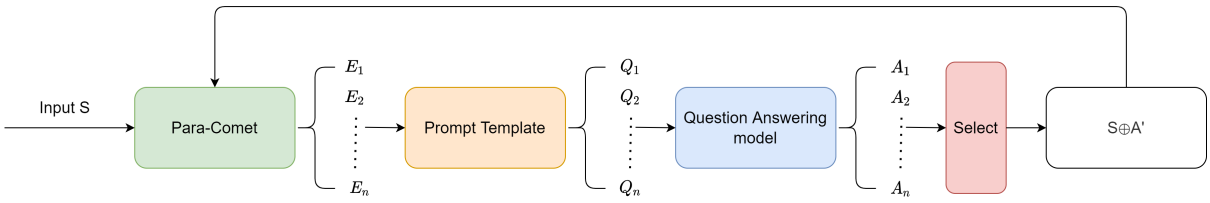


Figure 2: 故事生成方法的工作流程

本节介绍故事生成方法的工作流程。它由三个阶段组成，分别是事件推理、提示问题生成、答案生成与选择。其中每个阶段都使用了不同的预训练语言模型。整个流程通过优质的提

示问题模板，将不同模型结合起来，执行故事生成任务。如图2所示，对于输入S，常识推理模型生成n个可能的事件$E_1, E_2, \ldots, E_n$；接着用这些事件构造对应的问题$Q_1, Q_2, \ldots, Q_n$，并用问答模型进行回答，得到答案$A_1, A_2, \ldots, A_n$；选择困惑度最小的答案A'，作为故事下文。与开头合并为$S \oplus A'$，作为新一轮生成的开头。

3.1 事件推理

事件推理阶段使用Gabriel et al. (2021)提出的Para-COMET预训练模型，根据输入的故事开头，产生每一句对应的常识推理(Commonsense Inference)，这些常识推理表现为接下来可能发生的事件。常识推理一直被视为优质外部知识，来辅助文本生成。有了外部知识注入，模型便不会局限于输入的文本，而能从更符合人类社会常识的角度进行生成。

如图1 所示，对应输入的故事开头，模型生成如*to bulk up, to go to gym, to go to wrestling*等事件，这些事件有9种类型，如表1所示。本方法选择前六种用于后续生成。

Table 1: 事件的种类以及含义

Type	**Dimension**	**Template**
Causes	xIntent	PersonX wanted [ ]
	xNeed	PersonX needed [ ]
	xAttr	PersonX is seen as [ ]
Effects	xWant	PersonX wants [ ]
	xEffect	PersonX is likely [ ]
	xReact	PersonX then feels [ ]
	oWant	PersonY wants [ ]
	oEffect	PersonY is likely [ ]
	oReact	Others then feels [ ]

3.2 提示问题生成

由Para-COMET得到的事件不会显示与之对应的人物角色。*to bulk up, to go to gym, to go to wrestling*等事件的形式为to do结构。因此我们需要先通过预训练模型，得到对应事件的角色，然后将事件与角色填入提示问题模板中，生成最终问题。

3.2.1 链接角色与事件

我们初步的方案是使用BERT(Devlin et al., 2018)预训练模型。BERT是掩码语言模型，使用了Transformer模型的编码器层。BERT模型在完形填空任务上的表现非常优秀，因此我们将联系事件与角色的任务重构为完形填空的形式。以图1中的*to go to gym*为例，它的类型为**xWant**，那么我们将其重构为：

[MASK] wants to go to gym.

并将前文已生成的内容拼接在输入之前。模型生成的内容即为角色。但是，[MASK]的大小仅为一个单词。而很多故事中，角色的名字或代称会出现超过一个单词的情况，例如*Her mother, Daniel Wicky, My dog*等等。此时BERT会生成*She, He, It*等词，这些词虽然不会带来一致性与逻辑性的问题，但会大大降低故事的流畅性和多样性。

我们第二个考虑的模型是在Rajpurkar et al. (2016)提出的斯坦福大学问答数据集(The Stan-ford Question Answering Dataset, SQuAD)上训练的RoBERTa模型(Zhuang et al., 2021)。SQuAD是一个阅读理解数据集，给定一篇文章，准备相应问题，并给出问题的答案。相比填空，问答无疑更加自由，生成的内容也能应对大部分情况。依然以*to go to gym*为例，将其重构为：

Who wants to go to gym?

同样将前文以生成的内容拼接在问题之前。模型回答的内容经过筛选与清洗，得到与事件对应的角色。在本例中对应*to go to gym*的角色为*Sam*.

(c) 2022 中国中文信息学会计算语言学专业委员会

3.2.2 根据模板得到问题

有了事件与角色之后，根据不同的事件类型，我们将其填入问题模板之中以生成对应的问题。联系模板与生成模板如表2所示。

Table 2: 链接模板与问题模板

Event Type	Association Template	Question Template
xIntent	Who needs to [event]?	Why does [character] do [event]?
xNeed	Who needs [event]?	What does [character] do to need [event]?
xAttr	Who might be described [event]?	Why does [character] be [event]?
xEffect	Who [event]?	What makes [character] [event]?
xReact	Who feels [event]?	What makes [character] feel [event]?
xWant	Who may want [event]?	What makes [character] want [event]?

这样，由**xWant**类型的事件*to go to gym*，得到最终问题：

What makes Sam want to go to gym?

3.3 答案生成与故事选择

答案生成阶段，将问题生成模块中得到的问题输入进预训练的问答模型中，来让模型回答。本文选用了两种预训练模型：ELI5QA(Fan et al., 2019a)和BART(Lewis et al., 2019)。

3.3.1 使用ELI5QA模型进行回答

ELI5QA模型是一个在Fairseq-py框架下训练的、长文本形式的问答模型。它在ELI5数据集上进行训练。ELI5数据集全称是*Explain Like I'm Five*，从Reddit社区语料库收集。在这个数据集中，人们对开放式问题给出长而容易理解的答案，就像给五岁的孩子一样。

ELI5QA模型的输入是问题和文档，模型会从文档和训练得到的知识中生成对问题的回答。在本方法中，为了尽可能保证生成故事的一致性与连续性，我们将故事开头和已生成的内容作为文档，和问题一起输入进模型中。

问题模板的构建初衷是为了从ELI5QA模型中得到简短且相关的句子，作为故事的下文。为了保证这一点，ELI5QA模型采用Top-k采样算法进行生成。k设置过小则会容易生成更平淡或泛的句子，当k很大的时候，候选集合会包含一些不合适的token。我们取$k = 50$。

即使使用了Top-k 采样算法，生成的答案仍有一定概率出现无意义或重复的句子，因此我们需要对答案进行一定程度的清洗。采取的策略是：收集一些禁止短语，组成集合。我们舍弃掉那些包含禁止短语集合中的元素的句子以及它后面的所有句子。若首句长度不足6，我们也舍弃掉首句内容，并将剩余的句子添加到候选项中。这是因为首句有可能是对问题的*yes, no, Of course*这种回答，这些对故事生成没有帮助。这些禁止短语是生成样例中经常出现的无意义的词语、不友好的内容。通过这样的筛选，就能得到对应每个问题的答案集合。

3.3.2 使用微调的BART模型进行回答

除了ELI5QA模型外，我们还选用了BART模型作为生成答案的预训练模型。但是，原始的BART模型在执行QA任务时的效果不如人意，因为在训练时，数据的形式并非问答语料。因此我们将它在ROCStories数据集(Mostafazadeh et al., 2016)上进行微调，以提升生成效果。ROCStories数据集是常识性短篇小说的集合，包含100,000个五个句子的故事。每个故事都遵循一个日常主题。这些故事包含了日常事件之间的各种常识性因果关系和时间关系。我们将数据集按照70% : 15% : 15%的比例随机划分成训练集、验证集和测试集。对于训练集和验证集中的每一个故事，我们进行如下处理：

1. 对于每个故事中的一至四句，我们都遵循事件推理阶段和提示问题生成阶段，生成20个问题；

2. 当前句子以及它之前的所有句子作为Document，和问题以如下方式拼接起来：

   Question –T– Document

   拼接后的内容作为对应关键字Q的值；

第二十一届中国计算语言学大会论文集，第166页-第177页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

3. 当前句子后面的所有句子作为对应关键字A的值，和上文的关键字Q组成一个字典，存入jsonlines文件中。

这样就得到了训练集和验证集，设置学习率为$2e-5$，batch size为16，对BART进行训练。在解码时，将问题和前文内容拼接成训练集数据的形式，选用Top-k算法，设置$k=50$，进行生成。

3.3.3 选择答案作为故事下文

得到对于每个问题的答案集合之后，我们将这些集合合并，对其中的每个元素，我们都将它拼接在故事前文后，并使用在科幻摘要语料库上(Ammanabrolu et al., 2020)微调的GPT2模型来计算其困惑度（Perplexity, PPL），选择使拼接后困惑度最小的元素作为下文。该数据集由来自科幻电视和维基电影的2276个高质量情节摘要组成。这样做的目的是，选择那些更接近情节描述的答案作为故事下文。

4 实验设置

4.1 数据集

本文选取ROCStories数据集用于实验。训练集、验证集、测试集的划分与节3.3.2中一致。在训练时，使用ROCStories的训练集和验证集；在生成时，输入测试集中每个故事的前两句，模型或系统生成后三句。

4.2 指标

本文使用自动指标来评估实验结果，指标包括自动评测指标与人工评测指标。

4.2.1 自动评测指标

- 困惑度(**PPL**)：计算输入句子的指数平均负对数似然，评估文本流畅度和贴近人类语言的程度；
- 双语评估替补(Bilingual Evaluation Under-study，**BLEU** (Papineni et al., 2002))：计算生成句子和实际句子的N-grams，然后统计其匹配的个数；
- 基于召回率的主旨评估(Recall-Oriented Understudy for Gisting Evaluation, **ROUGE-L** (Lin, 2004))：计算最长公共子序列的重合率；
- 使用显式排序评估翻译的指标(Metric for Evaluation of Translation with Explicit Ordering, **METEOR** (Banerjee and Lavie, 2005))：基于单精度的加权调和平均数和单字召回率；
- 基于共识的图像描述评估(Consensus-based Image Description Evaluation, **CIDEr** (Vedantam et al., 2015))：利用TF-IDF来对不同N-gram赋予不同的权重；
- 使用BERT计算句子相似度得分的指标**BERTScore** (Zhang et al., 2019)
- 衡量文本多样性指标**Distinct-n**；(Li et al., 2015)：计算所有生成的文本中不同的n-gram的比率。

4.2.2 人工评测指标

- 连贯性得分(Coherence)：$0 \rightarrow 10$分，0分代表生成的文本完全无法阅读，10分代表生成的文本连贯清晰；
- 一致性得分(Consistency)：$0 \rightarrow 10$分，0分代表生成的文本和前文没有任何关联，10分代表生成的文本和前文叙事内容保持完全一致；
- 逻辑性得分(Logical)：$0 \rightarrow 10$分，0分代表生成的文本没有任何逻辑可言，10分代表生成的文本完全符合故事描述背景下的底层社会逻辑。

(c) 2022 中国中文信息学会计算语言学专业委员会

Table 3: 零样本场景实验结果

Models	PPL↓	BLEU-1↑	BLEU-4↑	DIST↑	METEOR↑	CIDEr↑	Rouge-L↑	BERTScore↑
GPT2	**10**	0.308	0.193	**0.950**	0.211	0.483	0.345	0.886
Ours wELI5	16	**0.386**	**0.282**	0.933	**0.230**	**1.151**	**0.384**	**0.891**

Table 4: 充足数据场景实验结果

Models	PPL↓	BLEU-1↑	BLEU-4↑	DIST↑	METEOR↑	CIDEr↑	Rouge-L↑	BERTScore↑
Finetune BART	11	0.474	0.358	0.918	0.251	2.459	0.469	0.912
Finetune GPT2	18	0.228	0.118	0.958	0.128	0.419	0.232	0.866
XLNET	14	0.266	0.144	**0.960**	0.190	0.056	0.299	0.878
HINT	11	0.440	0.283	0.916	0.232	1.566	0.425	0.919
Ours wBART	**10**	**0.493**	**0.376**	0.937	**0.262**	**3.058**	**0.475**	**0.923**

4.3 基线模型

本文实验选用的基线模型包括：

- GPT2：使用Transformer模型的解码器层
- BART：双向自回归的Transformer模型
- XLNET(Yang et al., 2019)：基于广义自回归预训练的双向模型
- HINT(Guan et al., 2021)：通过在解码过程中表示句子级别和语篇级别的前缀句子来生成连贯的文本

5 实验结果

5.1 零样本场景实验

本小节介绍在零样本场景下开展的实验，来说明本文提出的故事生成方法可以在没有训练数据的情况下生成不错的故事。实验选择预训练的GPT2作为基线模型，选择ELI5作为答案生成阶段的模型，比较二者生成的故事的各项指标得分。

如表3中的结果所示：

- 使用ELI5作为答案生成阶段的模型，在测试集上生成的故事的BLEU-1，BLEU-4，METEOR，CIDEr，Rouge-L，BERTScore指标得分要更高，说明其生成的故事比GPT2更贴近原故事。没有训练任何模型，仅仅通过构造合理提示问题模板，一个问答模型也能应用于故事生成任务中，并且有不错的表现；
- GPT2模型在DISTINCT指标上的得分保持领先，可能有两个原因：ELI5模型在处理冗余与重复时的能力不及GPT2模型；设计的问题模板没有达到最优；
- GPT2模型在PPL指标上的得分保持领先，这可能是由于预训练GPT2时，模型的参数数量与复杂度、训练数据等要超过ELI5QA模型，这使得GPT2模型生成的故事能够更为贴近人类语言。

5.2 充足数据场景实验

本小节介绍在充足数据场景下开展的实验，来说明本文提出的故事生成方法的性能也能随着数据量的增加而提升。实验选择在ROCStories的训练集和验证集上微调的BART、GPT2、XLNET作为基线模型，选择按照节3.3.2微调的BART作为答案生成阶段的模型，比较二者生成的故事的各项指标得分。

如表4中的结果所示，有了训练数据后，本文提出的故事生成方法的各项指标得分，相比零样本场景下的结果，都有了大幅提升，且能超过仅仅在数据集上微调的单个BART模型；相同训练数据下，本文提出的故事生成方法依然可以在多数指标上的得分保持对基线模型的优势。

第二十一届中国计算语言学大会论文集，第166页-第177页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

Table 5: 消融实验结果

Models	PPL↓	BLEU-1↑	BLEU-4↑	DIST↑	METEOR↑	CIDEr↑	Rouge-L↑	BERTScore↑
Ours w Prompt	**10**	**0.493**	**0.376**	**0.937**	**0.251**	**3.058**	**0.475**	**0.923**
Ours w/o Prompt	23	0.410	0.339	0.932	0.247	1.393	0.474	0.916

Table 6: 人工评测实验结果

Type	Models	Coherence↑	Consistency↑	Logical↑
Zero-shot	GPT2	5.43	5.20	5.33
	Ours wELI5	5.32	4.21	5.71
Enough Data	Finetune BART	5.76	5.45	5.80
	Ours wBART	**6.60**	**6.42**	**6.62**

5.3 消融实验

本小节介绍消融实验，来说明设置优质的提示模板能够为故事生成带来提升。实验分别选择按照节3.3.2微调的BART和训练过的不使用提示学习方法的BART模型作为答案生成阶段的模型。训练数据收集时，跳过提示问题生成阶段，直接将前文内容与事件拼接。

如表5中的结果所示，使用提示学习的故事生成方法的各项指标得分均超过不使用提示学习的故事生成方法。其中尤以PPL，BLEU-1，CIDEr指标提升幅度大，说明使用提示学习构建问题模板后，生成的句子比直接拼接要更贴近原文故事，且更为接近人类语言。具体原因有以下几点：

- 事件是短语结构，且没有对应的人物角色。直接使用时，对下文句子的影响很小，甚至会带来副作用。整个系统退化成了Encoder-Decoder框架，问题则变成输入故事前两句，BART模型输出后三句，且完全依赖模型本身的性能；
- 有了提示模板构建的问题，事件便不会独立于输入的故事前文，外部常识推理知识才能够帮助提升生成故事的逻辑性。

5.4 人工评估实验

本小节介绍人工评估实验，来更好地支撑通过自动评测指标得出的结论。实验从测试集中随机抽取100个故事开头，并由模型或系统生成后续故事。志愿者对这些故事从连贯性、一致性与逻辑性的角度进行打分，并计算平均得分。志愿者不会被告知故事的来源。

如表6中的结果所示，在零样本场景下，使用ELI5作为答案生成阶段模型的故事生成方法能在连贯性和逻辑性方面接近GPT2模型，但在一致性方面却有所不及。这可能是由于ELI5模型无法完全理解通过提示构建的问题，从而产生无意义、与前文无关或不连贯的回复。而在充足数据的场景下，使用微调BART作为答案生成阶段模型的故事生成方法则能在连贯性、一致性和逻辑性方面全面超越仅在数据集上训练的单个BART模型。

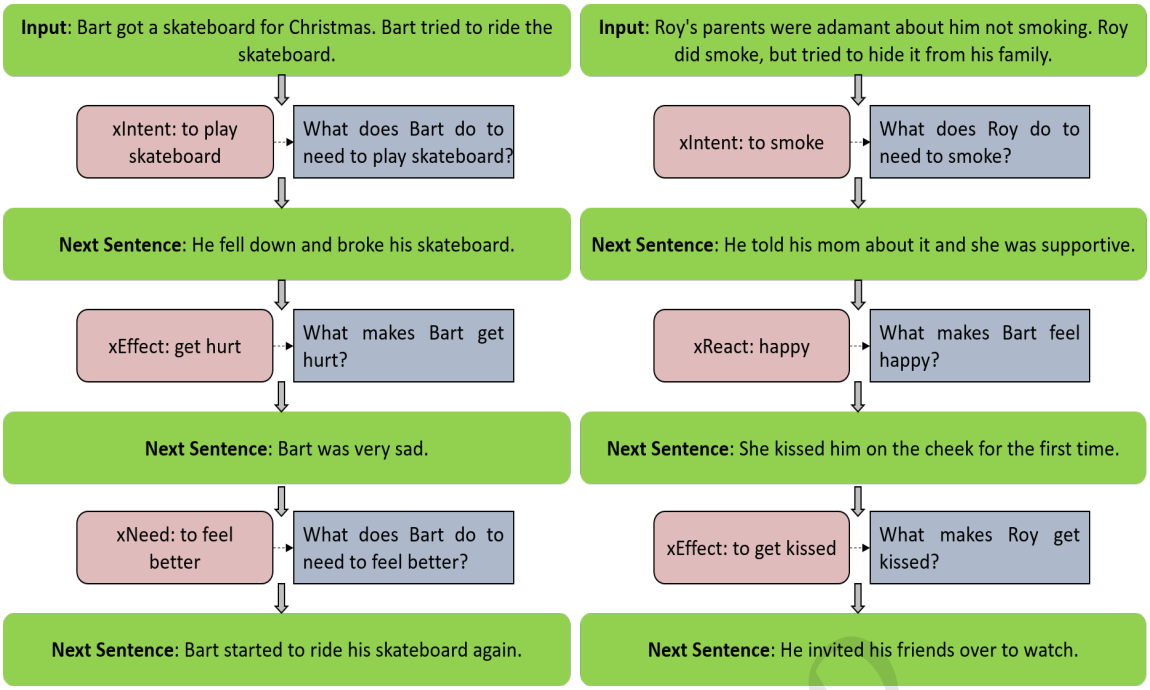
6 样例研究

图3展示了故事生成方法的两个样例的完整生成过程，包括：从给定故事开头，得到可能的事件；通过事件构建问题；根据问题，模型生成多个回答，并选择困惑度最小的那个作为下文句子。

6.1 验证生成流程

从图3(a)可以看出，在故事生成的第一轮，对于输入*Bart got a skateboard for Christmas. Bart tried to ride the skateboard.*，常识推理模型能够捕捉关键字，生成相关的事件*to play skateboard*。并且通过提示问题模板得到的问题*What does Bart do to need to play skateboard?*，也是对故事发展的一个很好的推测。有了优质的问题，预训练问答模型便能得到合乎逻辑的、保持一致性与连贯性的下文句子*He fell down and broke his skateboard.*。而随着生成故事的进行，这种逻辑性、一致性与连贯性并没有丢失。这说明我们的方法能够按照设计的思路进行故事生成。

(c) 2022 中国中文信息学会计算语言学专业委员会



(a) 样例1　　(b) 样例2

Figure 3: 故事生成样例

6.2 错误分析

但是，从图3(b)中，我们也能看到不符合逻辑的故事。如输入：*Roy's parents were adamant about him not smoking. Roy did smoke, but tried to hide it from his family.* 产生的下一句为：*He told his mom about it and she was supportive.* 故事开头已经说明Roy的家庭不允许Roy抽烟，且Roy也想方设法隐瞒，但第三句却变成了Roy告诉了他的妈妈（他抽烟的事实），结果他的妈妈还非常支持他，与前文逻辑不符合。导致这一原因是在生成第三句时，Para-COMET模型给出了事件推理*to smoke*，并且由该事件得到的问题*What does Roy do to need to smoke?*，其所生成的答案最终被选用。这也表明本文提出的故事生成方法主要受限于使用的预训练模型的性能。

7 总结与展望

本文针对故事生成任务需要大量数据的问题，提出了基于提示学习和预训练模型的故事生成方法，使用外部常识推理知识，充分发挥提示模板在少样本以及零样本场景下的优势。该方法将故事生成分为三个阶段：输入故事的开头，常识推理模型Para-COMET生成多个可能的事件，这些事件有六种类型，且是对故事发展方向的合乎社会逻辑的推测；根据类型，获取对应各个事件的人物角色，并将事件与角色填入问题模板中，构建总结上文并引导模型生成下文的问题；问答模型ELI5QA和BART产生对应的答案，并选择困惑度最小的作为故事下文。重复上述过程，最终生成完整的故事。实验表明，在提示学习与多个预训练模型的帮助下，无论是零样本场景还是充足训练数据场景，本文提出的故事生成方法都能在各项指标的得分上对基线模型保持优势。消融实验也突出了优质提示模板的重要性。

在未来，我们的工作主要可以分为三个方向：

- 选择更合适、性能更好的预训练模型，并设计与之匹配的提示问题模板，来提升生成故事的质量；
- 在选择答案作为下文句子时，尝试融合一些预训练模型，来避免选中逻辑上有问题、但困惑度得分低的句子，并且考虑在生成的各个阶段结束时，检查局部生成内容的正确性；
- 将提示学习的思想迁移到其他文本生成任务，如文本摘要，对话系统等。

8 致谢

我们感谢匿名审稿人，他们的建议有助于完善这项工作。本研究得到国家自然科学基金(No.62106105)、南京航空航天大学科研启动基金(No.YQR21022)和南京航空航天大学高性能计算平台的支持。

参考文献

Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic story generation: challenges and attempts. *arXiv preprint arXiv:2102.12634*.

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7375–7382.

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O Riedl. 2021. Automated storytelling via causal, commonsense plot ordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5859–5867.

Rebeca Amaya Ansag and Avelino J Gonzalez. 2021. State-of-the-art in automated story generation systems research. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–55.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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*.

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.

Louis Castricato, Spencer Frazier, Jonathan Balloch, Nitya Tarakad, and Mark Riedl. 2021. Automated story generation as question-answering. *arXiv preprint arXiv:2112.03808*.

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*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019a. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019b. Strategies for structuring story generation. *arXiv preprint arXiv:1902.01109*.

Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021. Paragraph-level commonsense transformers with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12857–12865.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. *arXiv preprint arXiv:2105.08963*.

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. *arXiv preprint arXiv:2101.00190*.

(c) 2022 中国中文信息学会计算语言学专业委员会

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*.

Li Lin, Yixin Cao, Lifu Huang, Shuang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. 2022. Inferring commonsense explanations as prompts for future event generation. *arXiv preprint arXiv:2201.07099*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. 2020. A character-centric neural model for automated story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1725–1732.

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. *arXiv preprint arXiv:2107.13586*.

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.

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, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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. *arXiv preprint arXiv:1606.05250*.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J Martin, Animesh Mehta, Brent Harrison, and Mark O Riedl. 2018. Controllable neural story plot generation via reinforcement learning. *arXiv preprint arXiv:1809.10736*.

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.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

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.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227.

(c) 2022 中国中文信息学会计算语言学专业委员会

生成，推理与排序：基于多任务架构的数学文字题生成

曹天旸[1,2*]，**许晓丹** [1,2*]，**常宝宝** [1†]
[1] 北京大学计算语言学教育部重点实验室，北京 100871
[2] 北京大学软件与微电子学院，北京 102600
{ctymy, diane1968, chbb}@pku.edu.cn

摘要

数学文字题是一段能反映数学等式潜在逻辑的叙述性文本。成功的数学问题生成在语言生成和教育领域都具有广阔的应用前景。前人的工作大多需要人工标注的模板或关键词作为输入，且未考虑数学表达式本身的特点。本文提出了一种多任务联合训练的问题文本生成模型。我们设计了三个辅助任务，包括数字间关系抽取、数值排序和片段替换预测。他们与生成目标联合训练，用以监督解码器的学习，增强模型对运算逻辑和问题条件的感知能力。实验证明所提方法能有效提升生成的数学文字题的质量。

关键词： 数学文字题生成；多任务学习

Generating, Reasoning & Ranking: Multitask Learning Framework for Math Word Problem Generation

Tianyang Cao[1,2*], **Xiaodan Xu**[1,2*], **Baobao Chang**[1†]
[1]Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Beijing 100871, China
[2] School of Software and Microelectronics, Peking University, Beijing 102600, China
{ctymy, diane1968, chbb}@pku.edu.cn

Abstract

A math word problem (MWP) is a narrative which reflects the underlying logic of math equations. Successful MWP generation has wide prospect in language generation and educational field. Previous works mostly require human-annotated templates or topic words, besides, they fail to consider the characteristics of MWP. This paper proposes a multitask learning based MWP generation framework. We devise three novel tasks, including number relation extraction, number ranking and sentence substitution prediction. These tasks are jointly trained with generation objective and supervise the learning of MWP decoder while enhancing the model' s comprehension of arithmetic logic and condition. Experiments demonstrate the effectiveness of our proposed method in equation consistency of generated MWPs.

Keywords: Math word problem generation , Multitask Learning

1 引言

[*] 共同一作
[†] 通讯作者
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第178页-第189页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

自然语言生成是自然语言处理中的一个重要领域，它的目的是生成流畅、可读性强且忠实于源输入的自然语言文本。在本论文中，我们主要聚焦于一种相对较新的数据文本生成任务—从等式生成数学文字题。公式到数学文本生成的目标是根据给定的等式，自动生成连贯且能够反映其运算逻辑的叙述性文本。如表 1所示是两个范例，其中包括了输入的数学表达式及其对应的数学问题文本。数学问题的生成涉及到内容规划等生成领域基础性课题，因而具有技术发展层面的意义；同时，数学问题的自动生成能够实现中小学数学问题的自动命题，有利于降低教师的教学负担，在计算机辅助教学领域有着广泛的应用前景。同时从学术的角度，该任务的定义还有进一步的拓展空间，例如如何处理符号更复杂的大学阶段数学问题，如何生成更加个性化、多样化的问题表达等等。

传统的数据-文本生成任务，往往是以作为结构化数据的表格记录或一系列三元组作为输入，因而输入的各部分之间是按照语言顺序或者时间顺序组织，而公式到问题文本的生成，其输入是由常量、未知变量、运算符三种符号组成的、不具有明显语义的抽象表达式，并且存在数学运算的逻辑。因此它与传统的生成任务有较大区别，模型也需要特殊的设计。针对该任务，前人的研究还存在很大不足。现有的一些工作 (Zhou and Huang, 2019; Wang et al., 2021; Liu et al., 2020) 在利用神经生成模型实现数学文字题自动生成方面取得了一定的效果。这些工作大多基于表达式（或表达式模板）和若干个话题词来生成数学文字题，他们更关注如何将关键词内容融入到解码过程中去，并反映数学问题发生的场景，而忽视对等式的结构特性及数字、变量间运算关系的理解。

数学表达式	$equ : (1 - 1/3 - 9/20) * x = 245$
数学文字题	At a local high school, ***1/3*** of the students are freshmen, ***9/20*** are juniors. And ***245*** are seniors. ***Find the total number of students.***
数学表达式	$equ : 45/(x - y) = 5$ $equ : 45/(x + y) = 3$
数学文字题	A boat travels ***45*** mi upstream (against the current) in ***5*** h . The boat travels the same distance downstream in ***3*** h. ***What is the rate of the boat in still water.***

Table 1: 数学文字题生成任务示例

首先，在数学文字题中，数字的角色非常重要，数字或者未知变量（x, y, z 等）通常代表现实场景中的某种物理量，例如物品的数量、种类、测度，交通工具的速度等属性。从直观上讲，数学文字题中的任意两个数字间都可能存在一定的逻辑关系，例如“…农场里饲养了 8 只动物，其中 3 只鸡，5 只兔子…”这段表述中，3 和 5 的关系是并列（分别表示鸡和兔子的数量），因而在叙述问题时，应该体现出这两种动物数目的求和；而在表 1的第一个例子中，x 和 1/3 的关系是相乘，因此在叙述问题时，应体现 x 是总体而 1/3 是比例系数。而现有的模型难以捕捉问题文本中数字所对应的物理量间的运算关系，因而对于结构较复杂，包含数字较多的公式，生成效果比较差。其次，在以往的工作中，数学文字题中的数字常常被替换成特殊符号（如 NUM0,NUM1）等，没有考虑数值隐含信息。而实际上数学文字题中，数字间的大小关系对于引导生成也有积极意义，具体有两方面：(1) 数字的大小能够反映一部分的语义信息，例如数字 a 比数字 b 更大，往往会出现数字 a 和 b 的差，或者从一堆数量为 a 的物品中拿走数量为 b 的物品这样的表达。(2) 数字间的大小关系蕴含着一些生活常识，例如电影院/公园所售门票中，成人票售价通常比儿童票更贵；船在航行时其顺水速度会大于逆水速度（顺水速度为船速度 + 水速度，逆水速度为船速度-水速度）；一个数值非常大的数字一般不太可能表达人的年龄，或者一支铅笔的价格等。最后，数学文字题一般由对背景的描述性句子、描述条件的语句和设问句组成，每个句子在文字题中都有其特定的功能。由于缺乏对句子结构的组织和规划，基线模型生成的语句常常逻辑比较混乱，如重复之前时刻的条件，产生和问题语境不相关的词汇或提问对象错误等。

针对以上问题，受前人在生成任务中引入多任务学习的启发 (Ge et al., 2021; Shen et al., 2021), 我们提出了多任务架构的数学表达式-文字题生成模型，将数学问题生成与数学问题理解融合到一个统一的框架中。我们基于 Transformer 构建我们的生成模型，并设计了数字关系抽取、数字排序和片段替换预测三个辅助任务。数字关系抽取中两个数字间的关系定义为他们在数学表达式树上的最近公共祖先（Least Common Ancestor），其中数学表达式树是等式所对应

第二十一届中国计算语言学大会论文集，第178页-第189页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

的后缀树，如图 1所示是找到数字间关系的例子，(a)(b) 都是方程组

$$\begin{cases} 2000 * (1 + 0.04)^5 = x \\ x - 2000 = y \end{cases} \tag{1}$$

对应的表达式树（pseudo root 是用于连接多棵树的虚拟节点），(c) 是方程 $51 * x + 8 * y = 510$ 对应的表达式树。绿色代表作为树的叶子的两个数字节点，而红色结点代表它们的最近公共祖先节点。数字关系抽取要求模型根据生成的问题文本预测其中两个数字间的运算关系标签，以此帮助解码器更好地组织物理量之间的运算逻辑，得到符合实际的表述。数字排序采用自回归的方式，对生成的数学文字题中的数字按照从小到大的顺序进行排序，以期使模型理解如何对公式中的数字进行组织，并感知到生成的问题语句是否合理，对更高质量的生成起到激励作用。片段替换预测则以句子为单位，将参考问题文本句中某一个句子按特定规律替换成另一个片段，在训练时提供给解码器，最终利用一个指针网络模块预测出被替换部分的左右边界。

上述三个辅助任务与标准生成任务进行联合训练。在基于 Dolphin_18K 拓展的数据集上的实验证明，所提出的方法在各项指标上均超越了基线模型，且能有效提升问题文本的逻辑描述与给定表达式之间的一致性。

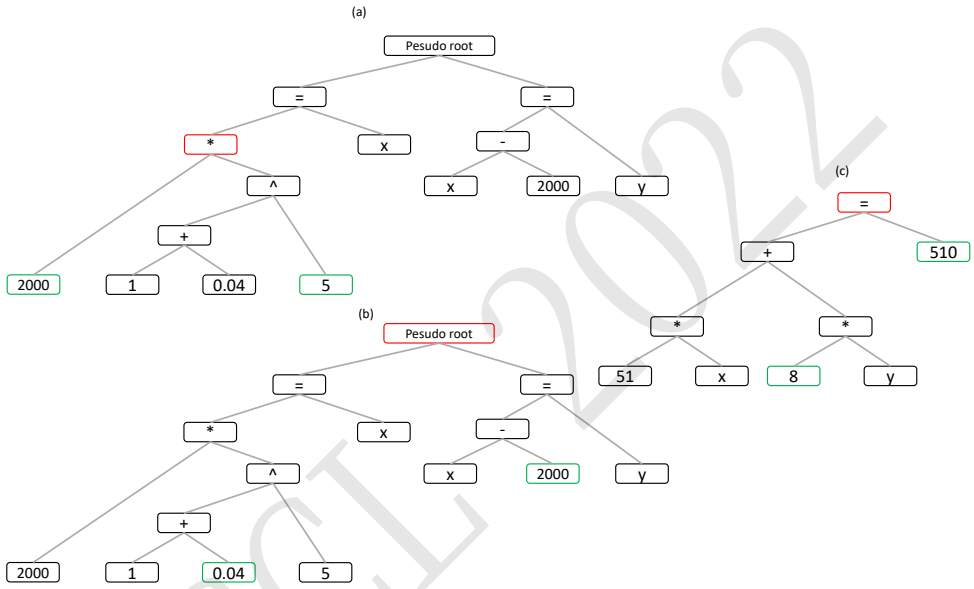


Figure 1: 表达式树上两数字节点的最近公共祖先

2 相关工作

数学文字题生成：早期的公式到文本生成工作主要是基于模板规则的方法，利用问答集编程技术（Answer Set Programming）(Polozov et al., 2015) 及框架语义技术 (Singley and Bennett, 2002; Deane, 2003) 等对模板中的插槽进行填充。利用深度学习架构，(Zhou and Huang, 2019; Wang et al., 2021) 的工作都是基于等式模板和关键词序列进行生成，他们的工作以等式模板和关键词作为输入，其中关键词是使用启发式规则直接从标准的问题文本中进行提取得到。模型使用端到端的方式进行训练，并在解码过程中融合模板和关键词两部分特征。然而这些方法在测试阶段也需要来自正确答案的关键词作为额外输入，这在现实场景中是不可用的。而本文工作采取了更符合实际的设定，即只根据数学等式进行问题文本的生成。(Liu et al., 2020) 的论文将输入的表达式抽象成莱文图，并用外部知识图谱子图引导生成和主题相关的句子。但该方法只能处理线性的表达式，且对于每个等式都需要额外的话题标注。

数据到文本生成：数据到文本生成是将结构化的数据转化为描述性文本 (Siddharthan, 2001; Gatt and Krahmer, 2018)。例如，(Puduppully et al., 2019a; Puduppully et al., 2019b; Gong et al., 2019; Wiseman et al., 2017) 关注于体育比赛新闻报道的生成，(Chisholm et al., 2017;

CCL 2022

Equation 1:

$$equ: x + y = 400 \; equ: 2 * x + 3 * y = 1050$$

Problem 1:
[1-11] The attendance at a baseball game was 400 people . [12-22] Student tickets cost $ 2 and adult tickets cost $ 3 . [23-29] Total ticket sales were $ 1050 . [30-38] How many tickets of each type were sold .
Relation Extraction:
r(2,3) = "+" r(400,1050) = "pseudo root"
Number Ranking:
2 < 3 < 400 < 1050
Problem After Replacement:
[1-11] The attendance at a baseball game was 400 people . [12-22] Student tickets cost $ 2 and adult tickets cost $ 3 . [23-38] *The red rose theater sells tickets for $ 4 . 50 and $ 6 . 00* [39-47] How many tickets of each type were sold .
Span Boundary:
[23,38]

Equation 2:

$$equ: (x + 150 * 0.8) / (x + 150) = 0.9$$

Problem 2:
[1-17]If x ounces of pure acid are added to 150 ounces of an 80% acid solution . [18-27]The concentration of the new mixture is 90% acid . [28-45]Find the number of ounces that were added to the original solution to produce the 90% solution .
Relation Extraction:
r(150,0.8) = "*" r(150,0.9) = "="
Number Ranking:
0.8 < 0.9 < 150
Problem After Replacement:
[1-15] *Results of a survey of fifty students indicate that 30 like red jelly beans .* [16-25] The concentration of the new mixture is 90% acid . [26-43] Find the number of ounces that were added to the original solution to produce the 90% solution .
Span Boundary:
[1,15]

Figure 2: 数据集中的两个文字题及其对应的三个辅助任务预测目标

Lebret et al., 2016) 等人的工作面向人物简历的生成，(Zhao et al., 2018; Gao et al., 2020) 等考虑结构化信息，从资源描述符三元组集合生成文本。此外，前人的工作在模型中设计了内容选择和文本规划机制以决定哪些内容应该被表述及按照怎样的顺序表述 (Puduppully et al., 2019a; Perez-Beltrachini and Lapata, 2018)。

多任务文本生成：近年来，多任务学习在文本生成和语言预训练领域得到了广泛应用。(Ge et al., 2021) 等人的工作基于上下文语句和被引用论文的摘要生成论文中的引用句。他们将作者写作引用句的原因分为 4 类，并联合训练生成器和分类判别器对引用的功能进行识别。(Shen et al., 2021) 在表达式解析任务中，联合训练了生成模型和排序打分模型以提升模型对错误表达式树的区分能力。

3 任务定义

我们的系统以一个或多个数学表达式 $\{E_1, E_2, ..., E_{|E|}\}$ 为输入，每个等式都由一系列的数学符号构成:$E_k = x_1 x_2 ... x_{|E_k|}$，其中 $|E_k|$ 是第 k 个公式的长度，由符号的数目来衡量。每个数学符号由以下三种符号构成：数学运算符，例如 "+，-，*，=，）" 等，数字，例如 "0.2，1,30" 等，变量，例如 "x, y, z" 等。任务的输出是一个数学问题文本：$\mathbf{y} = y_1 y_2 ... y_L$，该问题可以用输入的等式解决。$L$ 是问题文本的长度。我们的模型目标是根据输入的公式和之前时刻生成的词 $\boldsymbol{y}_{<t}$ 估计接下来生成的词的条件概率:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{L} P(y_t|\boldsymbol{y}_{<t}, E_1, E_2, ...) \tag{2}$$

4 模型结构

我们提出的多任务训练的问题文本生成模型如图 3所示。它包括一个标准的 Transformer 生成模型、一个用于预测生成的文字题中数字两两之间关系的**数字关系抽取**模块、一个对生成的文字题中的数字进行排序的**数字排序**模块以及一个对问题文本中被替换片段边界进行定位的**替换预测模块**。生成任务与所有辅助任务共享相同的编码器和解码器，且目标函数进行联合训练。辅助任务只在训练阶段有效。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.1 以 Transformer 为基础的编码器-解码器模型

序列到序列生成模型是目前生成任务的主流框架。输入的公式序列用 $E = (x_1, x_2, ..., x_n)$ 表示, 编码器由若干个 Transformer 层组成, 使用双向的自注意力机制将 E 映射到一系列连续的向量表示 $R = (\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n)$:

$$(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n) = Transformer_{ENC}(x_1, x_2, ..., x_n) \tag{3}$$

解码器同样也包括多个 Transformer 层, 其中增加了以编码器的输出为关注值的多头交叉注意力模块。解码器每次吸收一个词 s_i，利用前面时刻解码器的输出状态和编码器的输出表示 $\mathbf{R}$ 预测下一时刻词的分布:

$$P(*) = softmax(\mathbf{d}_i\mathbf{W} + \mathbf{b}) \tag{4}$$

$$\mathbf{d}_i = Transformer_{Dec}(\mathbf{R}, s_0, s_1, ..., s_{i-1}) \tag{5}$$

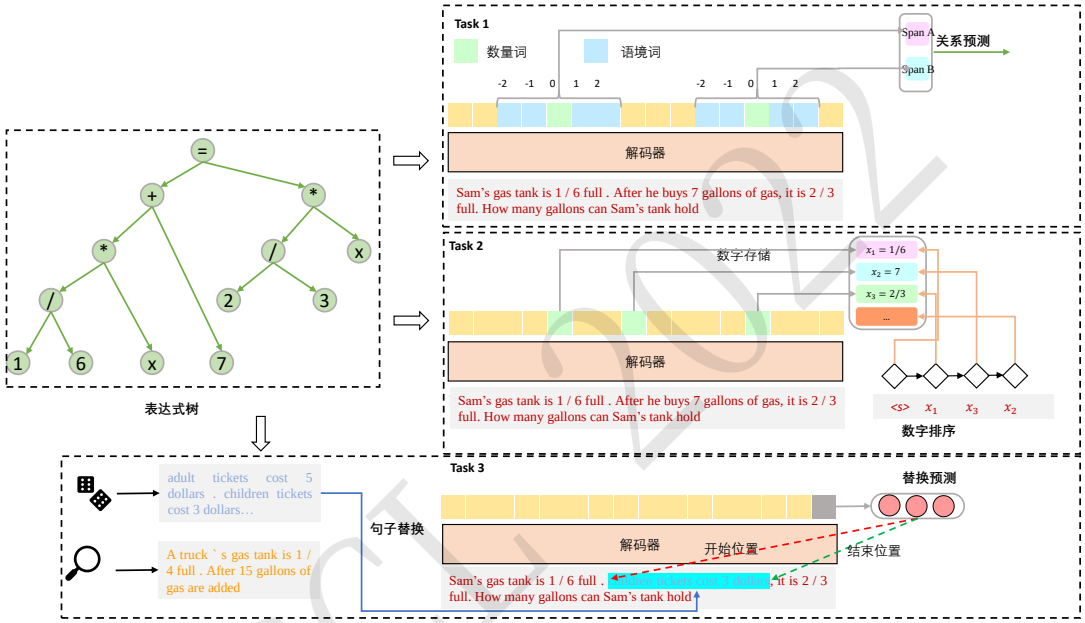


Figure 3: 基于多任务训练的数学问题生成模型

4.2 多任务训练

任务 # 1: 数字关系抽取: 数字关系抽取是利用生成的文字题中两数字的语境表示，对他们间的运算关系进行预测，从而帮助解码器在生成过程中感知到物理量之间的交互。数字间关系的标签类别包括 $\{+, -, *, /, \wedge\ , =, \surd, pseudo\ root\}$ 共 8 种，这些都是可能成为表达式后缀树上两叶子结点最近公共祖先的运算符或特殊标记。对于公式中的两个数字 x_a 和 x_b，他们在数学文字题中出现的位置分别为 p_a 和 p_b，我们将解码器视作对于生成的问题文本的编码器，并使用解码器输出状态序列中数字周围语境词的聚合信息作为该数字的表示。具体来说，对于 x_a，我们首先获得以 p_a 为中心，长度为 3,5,7 的片段（位置范围分别为 $p_a - 1 \sim p_a + 1, p_a - 2 \sim p_a + 2, p_a - 3 \sim p_q + 3$）的表示:

$$\mathbf{c}_{ak} = \text{MLP}([\mathbf{d}_{p_a-k}; \mathbf{d}_{p_a+k}; \mathbf{d}_{p_a-k} \odot \mathbf{d}_{p_a+k}])\ k \in \{1, 2, 3\} \tag{6}$$

其中 $\odot$ 代表逐元素乘积。随后我们学习一个参数 $\mathbf{u} \in \mathbb{R}^d$，通过分别比较 $\mathbf{u}$ 和 $\mathbf{c}_{a1}, \mathbf{c}_{a2}, \mathbf{c}_{a3}$ 来获得长度分别为 3，5，7 的片段的重要程度打分, 并利用得分作为注意力系数融合三个片段的表

示，得到数字 x_a 的最终表示：

$$att_i = \frac{\sigma(\mathbf{u}^T \mathbf{c}_{ai})}{\sum_{j\in\{1,2,3\}} \sigma(\mathbf{u}^T \mathbf{c}_{aj})} \tag{7}$$

$$\mathbf{c}_a^* = \sum_{i\in\{1,2,3\}} att_i \mathbf{c}_{ai} \tag{8}$$

其中 $\sigma(\cdot)$ 代表 sigmoid 函数。用同样的方式也可以得到 x_b 的表示 $\mathbf{c}_b^*$，于是分类的函数和数字关系抽取部分的损失函数可以定义为：

$$P(r(x_a, x_b)|\mathbf{c}_a^*, \mathbf{c}_b^*) \propto \exp(\mathbf{w}_1^T \sigma(\mathbf{W}_2 [\mathbf{c}_a^*; \mathbf{c}_b^*; |\mathbf{c}_a^* - \mathbf{c}_b^*|; \mathbf{c}_a^* \odot \mathbf{c}_b^*])) \tag{9}$$

$$\mathcal{L}_{relation} = \frac{1}{|\Omega|} \sum_{x_a, x_b \in \Omega, x_a \neq x_b} -\log P(r(x_a, x_b)|\mathbf{c}_a^*, \mathbf{c}_b^*) \tag{10}$$

其中 $\mathbf{w}_1$ 和 $\mathbf{W}_2$ 都是参数, 为了简便起见省略了偏置项。$\sigma(\cdot)$ 是一个激活函数，如 ReLU(·) 等。Ω 代表所有同时在数学公式和问题文本中出现的数字的集合，$r(x_a, x_b)$ 代表 x_a 和 x_b 在表达式树上的最近公共祖先。

任务 # 2: 数字排序：如前所述，数学文字题中数字的大小关系能够指示该数字可能代表的对象，同时也蕴含了现实场景中的一些隐含知识, 针对数字数值的比较和排序可以增强模型对于题目条件的理解能力。为此，我们提出数字排序模块用以监督解码器的学习。我们仿照任务 1 中的方法得到数字的特征表示，并将数学文字题中所出现的数字的表示构成一个存储，作为排序模块的输入。排序模块采用递归式生成，按从小到大的顺序，依次预测出最小数字在存储中的位置，第二小的数字在存储中的位置…直到最大的数字在存储中的位置。

具体地，将问题文本中的数字，按照下标从小到大依次记为 $z_1, z_2, ..., z_K$，其中 K 是数字数目，它们的特征表示分别为 $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K$。$H = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K]$ 构成了数字的存储。我们希望生成一个指针序列，按照这些数字升序的顺序依次指向 $[1, K]$ 中的某一个位置。假设把 $z_1, z_2, ..., z_K$ 升序排序后，按下标记作 $z_{p_1} < z_{p_2} < \ < z_{p_K}$，其中 $p_1, p_2, ..., p_K$ 是下标。生成器通过一个两层的 GRU 单元和指针网络实现，我们依次把排序过后数字的向量表示提供给 GRU，以自回归的形式生成下一个更大的数字在 H 中的位置。GRU 单元用零向量初始化，解码的过程可以形式化地写成：

$$\overline{\mathbf{h}}_t = GRU(\overline{\mathbf{h}}_{t-1}, \left[\mathbf{h}_{p_{t-1}}; \mathbf{r}^*\right]) \tag{11}$$

其中在 $t = 0$ 时，$\mathbf{h}_{p_{t-1}}$ 是开始标记 [sos] 的嵌入向量。$\mathbf{r}^*$ 代表对公式表示进行平均池化的全局向量：$\mathbf{r}^* = Meanpool([\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n])$。随后我们用指针网络预测 z_{p_t} 在 H 中的位置：

$$score(t, i) = \sigma(\mathbf{W}_3 \overline{\mathbf{h}}_t)(\sigma(\mathbf{W}_4 \mathbf{h}_i))^T \ 1 \leq i \leq K \tag{12}$$

$$P(q_t|q_1, q_2, ..., q_{t-1}, H) = softmax_i(score(t, i)) \tag{13}$$

其中 $\sigma(\cdot)$ 是激活函数,q_t 代表从小到大第 t 小的数（也就是 z_{p_t} 在 H 中的正确位置。最后，数字排序任务的损失可以写成：

$$\mathcal{L}_{rank} = -\frac{1}{K} \sum_{t=1}^{K} P(q_t|q_1, q_2, ..., q_{t-1}, H) \tag{14}$$

任务 # 3: 片段替换预测：前人的一些工作将随机选择的答案作为和输入无关的负样本，然后通过打分排序模型对每个输入-输出对赋予一个 $[0, 1]$ 之间的打分, 使模型学习鉴别低质量的回复。在本节中，我们对这种方法进行拓展，一方面，仅仅通过随机选择的句子作为和输入构成错误的公式-问题对可能对于模型效果提高有限，因为随机选择的数学文字题往往和真实的参考答案完全不相关，对于模型来说鉴别比较容易，因此我们希望模型判别更复杂的情形，即生成的句子能够在一定程度反映输入公式的逻辑，但是不完全正确。另一方面，我们希望模型在打分时能够进一步学习到问题文本中哪一段叙述不合理，或不符合场景。

为此，我们设计了一个片段替换的预测任务，它的目标是通过一定的随机策略，将数学文字题中的某一个句子替换成一个完全不相关或者有部分相关性的其他文字题中的句子，再将被替换后的问题文本提供给解码器作为输入，希望模型能够定位被替换的这个句子的范围。具体来说，对于由 N 个句子组成的参考数学问题文本 $P = \{S_0, S_1, ..., S_{N-1}\}$，我们按照如下规则进行替换：

- 在 1/3 的概率下，不对正确的问题文本进行替换。
- 在 1/3 的概率下，随机在训练数据集中选择一个问题文本 $P' = \left\{S'_0, S'_1, ..., S'_{N'-1}\right\}$，然后随机选择 $i \in [0, N-1]$,$j \in [0, N'-1]$，将原始题目中的第 i 个句子替换成 P' 中的第 j 个句子，替换后的数学文字题变为：$P = \left\{S_0, S_1, ..., S_{i-1}, S'_j, S_{i+1}, ..., S_{N-1}\right\}$。
- 在 1/3 的概率下，我们首先通过 BertScore (Zhang* et al., 2020) 工具，在训练数据集中检索一个和参考答案嵌入表示相似度最高的问题文本 $R' = \left\{S''_0, S''_1, ..., S''_{N'-1}\right\}$，然后随机选择 $i \in [0, N-1]$,$j \in [0, N''-1]$，将原始题目中的第 i 个句子替换成 R' 中的第 j 个句子，替换后的数学文字题变为：$P = \left\{S_0, S_1, ..., S_{i-1}, S''_j, S_{i+1}, ..., S_{N-1}\right\}$。

完成替换后，我们将新的数学问题文本 P 作为解码器的监督信号，在 P 的最后附加一个标记 [eos] 作为结束符，这样文本总长度记为 M。解码器输出的状态依次表示为 $\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_M$。我们用最后一个状态 $\mathbf{d}_M$ 和输入公式表示向量 $\mathbf{r}^*$ 拼接作为查询向量，再通过指针网络预测替换部分的范围边界。用 $start^*$ 和 end^* 分别表示被替换部分的开始位置和结束位置的下标。特别地对于上述（1）中原问题文本没有被替换的情形，我们增加一个可学习的向量 $\mathbf{v}$ 拼在 $\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_M$ 前，即 $D = [\mathbf{v}, \mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_M]$，此时 $start^* = end^* = 0$。指针网络预测开始和结束位置的表达式分别为：

$$P(start = i) = softmax(\mathbf{d}_i \mathbf{W}_4 [\mathbf{d}_M; \mathbf{r}^*]) \; 0 \leq i < M \tag{15}$$

$$P(end = j) = softmax(\mathbf{d}_j \mathbf{W}_5 [\mathbf{d}_M; \mathbf{r}^*]) \; i < j < M \tag{16}$$

我们把选择某一片段的概率定义为选中其开始位置和结束位置的概率乘积，这样替换任务的损失函数可以表示为：

$$P(start = i, end = j) = P(start = i) * P(end = j) \tag{17}$$

$$\mathcal{L}_{span} = -\log P(start = start^*, end = end^*) \tag{18}$$

在计算上述 $P(start = i, end = j)$ 时，会对所有可能的片段的得分进行归一化。

模型损失函数模型的损失函数由生成部分的对数似然（用 MLE 表示）和三个辅助学习任务的损失构成：

$$\mathcal{L}_{total} = MLE + \alpha(\mathcal{L}_{relation} + \mathcal{L}_{rank} + \mathcal{L}_{span}) \tag{19}$$

5 模型实验与分析

5.1 数据来源

我们的数据集基于 Dolphin_18K (Huang et al., 2017), 该数据集从 Yahoo!Answer 网站上爬取得到。由于 (Huang et al., 2017) 仅仅开源了 Dolphin_18K 的一个子集（3154 条样本），这对于生成模型的训练来说是不够的。因此我们复用 (Huang et al., 2017) 中给出的脚本从 Yahoo!Answer 上爬取并收集了额外的数据，将数据集的规模扩充到了 14943 个样例（代码和数据集将在论文录用后公开）。对于获得的数据我们进行了预处理，删除了问题文本长度超过 45 个词或低于 15 个词的公式-文本对，最终保留了 9643 个样例。表 2给出了数据集的统计信息。

5.2 基线模型和参数设置

我们将所提出的方法与以下基线模型对比：（1）**Seq2seq** (Bahdanau et al., 2014) 首先被用于机器翻译任务。在本工作中，我们实现了使用注意力机制和拷贝机制的 seq2seq 模型。（2）

(c) 2022 中国中文信息学会计算语言学专业委员会

	Train	Dev	Test
Size	7714	964	965
Equation Length (average)	16.69	16.23	16.63
Problem Length (average)	28.90	29.64	28.74
Tokens	7445	3065	2875

Table 2: 数据集统计信息

词向量维度	256	Transformer 层数	2
Transformer 隐层维度	256	Transformer 前馈网络中间层维度	512
GRU 隐层维度	256	GRU 层数	2
Adam β_1	0.99	Adam β_2	0.999
Batchsize 大小	32	学习率	e-4
Dropout Rate	0.2	(19) 式中的 α	1

Table 3: 模型超参数设置

SeqGAN (Yu et al., 2016) 基于生成式对抗网络，使用强化学习在每一步生成时评估所得到完整序列的得分期望。在文本生成和音乐生成等多个任务上都取得了提升。(3) **DeepGCN** (Guo et al., 2019) 是深度图卷积神经网络，由于数学公式可以转化成后缀表达式树，这样树上每个节点可以看作图的节点，同时在树上相邻的两节点间进行连边，于是公式文本生成问题可以转化为图到序列的生成的问题。(4) **Transformer** (Vaswani et al., 2017) 被广泛应用于生成任务的模型。(5) **DualCG** (Wei et al., 2019)，在本文中我们使用 DualCG 来把表达式解析和问题文本生成集成到一个统一的框架中。(6) **BART** (Lewis et al., 2020) 是使用标准 transformer 架构的强有力的预训练模型，我们在数学问题生成的数据集上对 BART 进行微调。

模型使用 Adam 优化器进行训练。生成任务和三个辅助任务共享相同的编码器和解码器。第三个任务中在获取训练数据时采取了类似于 Roberta 中动态遮蔽的动态策略，即替换和被替换的内容不是在预处理时决定，而是在执行每一轮次的训练时都运行一次随机替换，这种策略扩大了选取错误句子时的搜索空间，避免解码器反复看到相同的模式，实际上起到了数据扩充的作用。其余参数设置见表 3。

模型	**BLEU**	**ROUGE-L**	**BERTScore**	**METEOR**	**Dist1**(%)	**Dist2**(%)	**NR**(%)
Seq2seq (Bahdanau et al., 2014)	2.59	20.25	82.98	18.51	14.56	34.99	47.60
SeqGAN (Yu et al., 2016)	2.62	19.22	82.56	17.63	12.96	30.02	44.00
DeepGCN (Guo et al., 2019)	3.04	20.94	83.07	19.48	16.81	45.17	49.21
Transformer (Vaswani et al., 2017)	3.14	21.84	83.81	20.26	12.94	43.51	44.84
DualCG (Wei et al., 2019)	3.60	21.43	83.99	20.63	15.47	46.01	40.97
BART_{large} (Lewis et al., 2020)	4.15	22.26	**86.35**	22.30	12.77	46.76	43.47
Full Model	**4.20**	**23.13**	84.61	**22.32**	**19.03**	**53.89**	**71.06**
T1+T2	3.22	20.93	84.90	25.33	10.74	35.62	49.37
T1+T3	4.10	22.52	84.74	22.43	18.90	51.95	70.16
T2+T3	3.92	22.64	84.92	21.96	19.70	53.13	65.64
T1	3.37	21.87	84.57	20.70	16.60	47.63	70.52
T2	3.48	22.39	84.32	21.50	20.90	57.37	69.62
T3	3.71	21.68	84.25	21.37	20.67	57.42	67.81

Table 4: 本文方法和基线模型在主要评测指标上的性能对比。其中 NR 代表数字召回率，Trans 是 Transformer 的缩写

5.3 主要实验结果

自动评测：我们使用了以下自动评测指标：BLEU（BLEU-1 和 BLEU-2 值的平均）(Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang* et al., 2020)（一种基于嵌入表示相似度的文本生成评价指标），Dist-1，Dist-2（指示不同的一元组/二元组在所有一元组/二元组中的比例），数字召回率（衡量正确的问题文本中的数字有多大比例被正确拷贝了）。主要结果如表 4所示，其中也包含只保留部分辅助任务的消融实验。Full Model 表示完整的模型，T1,T2,T3 分别表示三个任务。

可以看到，使用完整的预训练任务设定，我们的模型在 BLEU、ROUGE-L、METEOR 指标上相比标准的 Transformer 模型分别取得了 33.7%、5.9% 和 10.2% 的提升，在 BERTScore 指标方面也取得了 0.8 个百分点的提升，这说明了所设计的学习任务增强了解码器对未来产生的语句的感知和搜索能力。由于针对数字的排序赋予了模型认知数值大小的能力，我们的模型在数字召回方面也表现更好。此外，我们发现完整的模型框架在自动评测指标方面取得了优于

模型	BLEU	ROUGE-L	BERTScore	METEOR	Dist1(%)	Dist2(%)	NR(%)
BART_{large}	4.15	22.26	86.35	22.30	12.77	46.76	43.47
BART_{large}+T1+T2+T3	**4.83**	**23.01**	86.24	**22.60**	**16.92**	**49.68**	**43.37**

Table 5: 在 BART 模型上增加辅助任务所获得的提升

BART 的效果，这说明针对数学公式的特点设计多任务学习目标能有效提升数学问题生成的语言质量。

为了探究多任务联合训练中每个任务所起的作用，我们进行了详细的消融实验，即三个辅助任务中任取两个或只选取一个，并报告了实验结果。如图所示，在只有任务一（关系抽取）和任务二（数值排序）的情况下，模型在 BLEU、ROUGE-L 和 METEOR 方面的表现分别下降 23.3%、5.42% 和 8.64%，在 BERTScore 得分上下降 0.06 个百分点；在只使用任务一（关系抽取）和任务三（片段替换预测）的情况下，模型得分相比完整设置下降不明显；在只使用任务二（数值排序）和任务三（片段替换预测）的情况下，模型在 BLEU、ROUGE-L 和 METEOR 方面的表现分别下降 6.67%、2.64% 和 1.61%。当只采用一种辅助训练任务时，自动评测指标得分均显著低于完整的模型，从而证明了联合训练的优势。

问题类别	测试集中占比%	本文模型	Transformer
1 数字基本运算，人口问题	9.02	6.40	7.19
2 几何类问题	17.72	5.59	4.99
3 概率问题，币值问题	6.84	**3.21**	2.41
4 数字基本运算	14.51	**4.64**	3.03
5 溶液类问题，物质类问题	9.33	**0.79**	0.24
6 金融类问题	8.08	**6.31**	2.47
7 百分比问题，几何类问题	8.08	**6.31**	2.47
8 销售问题	12.64	**4.27**	1.68
9 行程问题	10.36	**3.09**	1.79

Table 6: 在不同话题类型的测试集子集上，多任务架构和标准 Transformer 的 BLEU 值对比

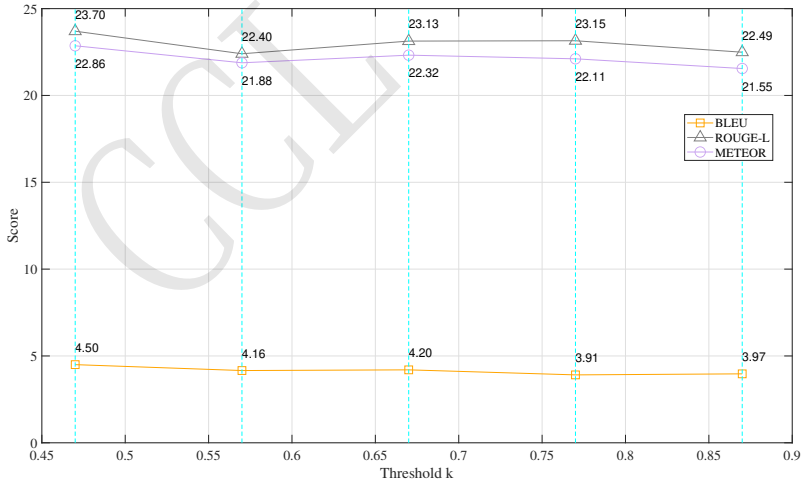


Figure 4: 任务三中 2、3 两种替换策略被采用的概率发生变化时，模型性能的变化

模型泛化能力: 为了验证所提出方法的泛化性能，我们将本文方法应用到 BART 生成模型上，即在微调 BART 的训练过程中加入三个辅助任务的目标函数。实验结果如表 5所示，可以发现在使用了多任务学习之后，BART 的效果获得了进一步提升。这验证了数字关系抽取和数学排序等任务能使生成模型更准确地表述物理量之间的逻辑关系，也说明针对任务特点设计的学习目标有助于增强通用生成模型的泛化能力。

不同类型问题上的实验分析: 为了进一步探究所提出的模型在不同子集上的性能，我们采

第二十一届中国计算语言学大会论文集，第178页-第189页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

用 (Shen et al., 2021) 中对问题文本话题的定义，根据所隶属的话题的不同将测试集划分为 9 个子集。属于不同话题的问题文本在文字表达方面有自身的特点，同时不同子集中高频的数学表达式也具备类别特征，例如存款、利息类问题常常涉及到增长率的幂次计算。相关结果如表 5.3 所示。表 5.3 中的第一列给出了不同问题类别的大致描述（根据该类别所包含的关键词人工归纳，其中数字基本运算是指不涉及具体生活场景，单纯描述几个数之间运算的问题），并给出了测试集中这 9 种类别占比。通过在细分子集上本文方法和标准 Transformer 模型的对比，可以发现：（1）我们的模型在第二个类别（几何类问题）上得分最高，在第六个类别（金融类问题）上得分最低，而基线模型也是如此。这可能是由于几何类问题涉及元素比较单一，遵循相似的模板，因而学习难度较低；而金融类问题包含较多的专用术语，给生成带来了一些困难。（2）本文的方法在 3∼9 类别上相比基线模型 BLEU 值均更高，而在类别 1、2 上 BLEU 值有所下降。尤其是我们的模型在销售、行程类问题子集上提升均比较明显，可能与这类问题需要较多数量关系的理解与推理有关，比如行程与速度的关系、单价与总价的关系、不同类别商品价格的关系等。

模型分析：进一步地，我们对于任务三中采取不同策略替换原问题中的片段对最终实验结果的影响进行探究，以验证所提出模型的鲁棒性。考虑阈值 k 和一个在 [0,1] 区间内符合均匀分布的数，当该数落在 $[0, 1/3]$ 时不进行替换，落在 $[1/3, k]$ 时以训练集中随机抽取问题文本作为不相关句的来源，落在 $[k, 1]$ 时以 BERTScore 检索出的问题文本作为不相关句的来源。默认情况下 k 取 2/3 即 0.67。当 k 分别取 0.47、0.57、0.67、0.77 和 0.87 时，分析模型的 BLEU、ROUGE-L 和 METEOR 指标的变化，并绘成折线图，如图 4所示，其中垂直的蓝色虚线代表了 k 的 5 个采样点。可以看到：（1）当 k 的取值发生变化时，模型的生成质量保持在较高水平，验证了所提出方法的稳定性。（2）总体而言当 k 的取值增加时，评测指标得分有小幅度的降低。这是由于 k 的取值越小，采用第三个策略的概率越大，就有更大的机会选取检索到的问题文本，即用于替换的错误片段和原文强相关，对模型来说造成的干扰较大，也更难区分。这样，在片段预测任务中，模型能够学习识别和真实句子语境类似，但不符合公式逻辑的错误片段，而不仅仅是能够对随机采样的句子进行定位。

	流畅度		**一致性**		**S1**(%)	**S2**(%)
	score	κ	score	κ		
本文模型	**3.97**	0.436	**4.06**	0.497	**32**	**57**
Seq2seq	3.78	0.256	3.48	0.483	23	34
SeqGAN	3.75	0.305	3.28	0.520	20	40
DeepGCN	3.61	0.295	3.55	0.494	29	52
Transformer	3.80	0.333	3.53	0.421	20	45
DualCG	3.88	0.346	3.66	0.455	28	53
BART_{large}	3.56	0.398	3.73	0.454	31	52

Table 7: 生成的数学文字题的人工评测结果

人工评价：为了更好的衡量所提出模型的实际生成质量，我们请了三位人工标注者来判断不同模型给出的结果的质量，其中采用了以下四个方面的评价。(1) 流畅度：流畅度主要衡量生成的数学文字题是否流畅，是否存在语法错误。(2) 一致性：一致性用于衡量数学文字题在文本层面是否连贯 (3) 可解决性 (S1)：由于生成的目标是数学文字题，我们需要考虑该问题是否能被解决，也就是有多大比例的生成的问题文本，可以根据它列出和原数学表达式相同（或者等效）的等式。(4) 可解决性 (S2)：是一个相对于可解决性 (S1) 更宽松的指标。它只要求列出的是一个合法的表达式，而不要求与给定等式相符。

我们随机挑选了 100 个生成的数学问题文本，并且按照五级打分制进行打分。我们把得分映射到 1∼5，而更高的分数代表更好的性能。为了说明不同评分者给出的结果间的一致性，我们使用 Kappa 系数 (κ 值) 来进行评估，κ 值越高说明打分可信度越高。平均后的得分如表 7所示。可以看到所提出的多任务模型的得分无论在流畅性、一致性还是在可解决性方面，得分都是最高的。其中在流畅度、一致性、S1、S2 上比 DualCG 分别提升了 2.32%、10.93%、14.28%、7.55%；比 BART_{large} 分别提升了 11.52%、8.84%、3.23%、9.61%。

6 总结

我们在数学文字题生成的研究中，提出了三个与生成任务联合训练的辅助任务，从物理量之间关系的预测、数值大小的比较以及无关片段的预测三个角度，使模型学习到数学应用题中的常见表述，进而提升了通用生成模型在该任务上的表现。

致谢

本文工作受到国家自然科学基金（61876004、61936012）支持，特此致谢。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv: Computation and Language*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from wikidata. 1:633–642.

Paul Deane. 2003. Automatic item generation via frame semantics: Natural language generation of math word problems. 12.

Hanning Gao, Lingfei Wu, Po Hu, and Fangli Xu. 2020. Rdf-to-text generation with graph-augmented structural neural encoders. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3030–3036. ijcai.org.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, pages 65–170.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO: A background knowledge- and content-based framework for citing sentence 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 1466–1478, Online. Association for Computational Linguistics.

Li Gong, Josep Crego, and Jean Senellart. 2019. Enhanced transformer model for data-to-text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 148–156. Association for Computational Linguistics, November.

Zhijiang Guo, Yan Zhang, Zhiyang Teng, and Wei Lu. 2019. Densely connected graph convolutional networks for graph-to-sequence learning, 08.

Danqing Huang, Shuming Shi, Chin-Yew Lin, and Jian Yin. 2017. Learning fine-grained expressions to solve math word problems. In *EMNLP*, pages 805–814. Association for Computational Linguistics, September.

Remi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. pages 1203–1213.

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. Association for Computational Linguistics, July.

Chinyew Lin. 2004. Rouge: A package for automatic evaluation of summaries. pages 74–81.

Tianqiao Liu, Qian Fang, Wenbiao Ding, Zhongqin Wu, and Zitao Liu. 2020. Mathematical word problem generation from commonsense knowledge graph and equations. *CoRR*, abs/2010.06196.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

(c) 2022 中国中文信息学会计算语言学专业委员会

Laura Perez-Beltrachini and Mirella Lapata. 2018. Bootstrapping generators from noisy data. 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 1516–1527. Association for Computational Linguistics, June.

Oleksandr Polozov, Eleanor O' Rourke, Adam M. Smith, Luke Zettlemoyer, Sumit Gulwani, and Zoran Popović. 2015. Personalized mathematical word problem generation. In *IJCAI 2015*, May.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019a. Data-to-text generation with content selection and planning. *AAAI 2019*, 33(01):6908–6915.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019b. Data-to-text generation with entity modeling. *ACL 2019*, pages 2023–2035.

Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021. Generate & rank: A multi-task framework for math word problems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2269–2279. Association for Computational Linguistics.

Advaith Siddharthan. 2001. Ehud reiter and robert dale. *Building Natural Language Generation Systems*. cambridge university press, 2000. $64.95/£37.50 (hardback), 234 pages. *Nat. Lang. Eng.*, (3):271–274.

Mark Singley and Randy Bennett. 2002. Item generation and beyond: Applications of schema theory to mathematics assessment. 01.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. pages 5998–6008.

Zichao Wang, Andrew Lan, and Richard Baraniuk. 2021. Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5986–5999. Association for Computational Linguistics, November.

Bolin Wei, Ge Li, Xin Xia, Zhiyi Fu, and Zhi Jin. 2019. Code generation as a dual task of code summarization. In *NeurIPS*, pages 6559–6569.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. pages 2253–2263.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2016. Seqgan: Sequence generative adversarial nets with policy gradient. *arXiv: Learning*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*. OpenReview.net.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *EMNLP*, pages 3901–3910. Association for Computational Linguistics, October-November.

Qingyu Zhou and Danqing Huang. 2019. Towards generating math word problems from equations and topics. *INLG 2019*, pages 494–503.

CCL 2022

基于SoftLexicon和注意力机制的中文因果关系抽取

崔仕林[1,2,3],闫蓉[1,2,3*]
[1]内蒙古大学计算机学院/内蒙古, 呼和浩特, 010021
[2]蒙古文智能信息处理技术国家地方联合工程研究中心/内蒙古, 呼和浩特, 010021
[3]内蒙古自治区蒙古文信息处理技术重点实验室/内蒙古, 呼和浩特, 010021
1437869230@qq.com, csyanr@imu.edu.cn

摘要

针对现有中文因果关系抽取方法对因果事件边界难以识别和文本特征表示不充分的问题，提出了一种基于外部词汇信息和注意力机制的中文因果关系抽取模型BiLSTM-TWAM+CRF。该模型首次使用SoftLexicon方法引入外部词汇信息构建词集，解决了因果事件边界难以识别的问题。通过构建的双路关注模块TWAM(Two Way Attention Module)，实现了从局部和全局两个角度充分刻画文本特征。实验结果表明，与当前中文因果关系抽取模型相比较，本文方法表现出更优的抽取效果。

关键词： 因果关系抽取 ； 序列标注 ； 外部词汇信息 ； 注意力机制

Chinese Causality Extraction Based on SoftLexicon and Attention Mechanism

Shilin Cui[1,2,3],Rong Yan[1,2,3*]
[1]College of Computer Science, Inner Mongolia University, Hohhot, 010021, China
[2]National and Local Joint Engineering Research Center of
Intelligent Information Processing Technology for Mongolian, Hohhot, 010021, China
[3] Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot, 010021, China
1437869230@qq.com, csyanr@imu.edu.cn

Abstract

Existing Chinese causality extraction methods have to face the problems of identifying causal event boundaries and inadequate text features representation, this paper proposes a Chinese causality extraction model BiLSTM-TWAM+CRF based on external lexical information and attention mechanism for addressing above issues. It is the first time that we introduce external lexical information by using the SoftLexicon method to construct word set for solving causal event boundaries problem. We construct a Two Way Attention Module (TWAM) and try to represent the text features as much as possible from both the local and global views. Experimental results show that our proposed method has better causality extraction performance than the existing Chinese causality extraction methods.

Keywords: Causal Relation Extraction , Sequence Labeling , External Lexical Information , Attention Mechanism

1 引言

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家自然科学基金(61866029)

随着海量数据的增长，如何快捷地从海量文本中寻找有用信息已经成为一项研究难题，因此信息抽取研究应运而生。因果关系抽取研究 (杨竣辉et al., 2016)作为其重要分支，多年来受到了学者们的广泛关注，在自然语言处理(Natural Language Processing,NLP)领域研究中占有重要地位。

因果关系抽取任务旨在从自然语言文本中自动抽取出文本中的因果关系。因果关系是文本中重要的一种语义关系，大量的存在于自然语言文本中。它常被用来描述‘原因’与‘结果’之间的前后联系 (冯冲et al., 2018)，在事件推理 (Radinsky et al., 2012)、未来场景生成 (Hashimoto et al., 2014)、问答 (Girju, 2003)和信息检索等任务中起着十分重要的作用。同时，根据文本中是否含有因果连接词，可以将因果关系分为显式因果关系和隐式因果关系。例如：①‘火灾导致两名儿童受伤。’中，‘导致’为因果连接词，‘火灾’和‘受伤’构成显式因果关系；②‘男子盗窃被抓获。’中，不含有因果连接词，‘ 盗窃’和‘被抓获’构成隐式因果关系。

目前，因果关系抽取任务的相关研究，主要从三个角度展开：①使用文本分类，判断句子中是否具有因果关系；②给定候选因果对，判断句子中是否包含因果关系；③通过序列标注方法，对句子抽取因果关系并确定因果关系的方向。这些研究方法虽然都能够对因果关系实现有效地抽取，但是依然存在以下问题。首先，中文文本与英文文本最大的区别在于没有明显的边界标示符，所以对中文文本处理需要先确定中文词汇的边界，即需要进行中文分词。同时，中文词语的一词多义等问题会导致词语边界模糊，很难通过分词工具得出准确的词语边界。其次，由于显式因果关系含有明显的因果连接词，大多数的模型只对显式因果关系做到了抽取，而隐式因果关系由于缺乏明显的词汇特征，使得大多数的模型对其抽取效果较差。最后，针对中文文本，单纯基于字符向量的方法无法利用中文词汇信息，致使现有的单纯基于字符向量的方法对文本的特征表示不充分。

最近的相关研究表明，引入外部词汇信息 (Zhang and Yang, 2018)和注意力机制 (Vaswani et al., 2017)，能够在一定程度上解决中文词汇边界难以识别和文本特征表示不充分的问题。本文延续这一思路，使用基于序列标注的方法，抽取中文文本中的因果关系。具体地，在字符向量的基础上，利用SoftLexicon方法 (Ma et al., 2020)引入外部词汇信息，通过构建词集的方式，将词典信息融合到字符表示层中，增强了字符的语义表达，从而解决中文词汇边界难以识别的问题。进一步地，本文结合注意力机制构建了双路关注模块TWAM(Two Way Attention Module)，该模块融合了通道注意力(Channel Attention) (Hu et al., 2020)和缩放点积注意力(Scaled Dot-Product Attention) (Vaswani et al., 2017)，能够从局部和全局角度捕获句子的语义特征，提取出深层次的语义信息，进而增强文本的特征表示。同时结合BiLSTM与CRF，本文构建了BiLSTM-TWAM+CRF模型，该模型不仅引入了外部词汇信息以解决无法利用词汇信息的问题，而且利用构建的TWAM，从局部和全局两个角度捕获了更多的文本特征，能够更加有效地抽取中文文本中的因果关系。

2 相关工作

2.1 因果关系抽取

现有抽取因果关系的方法主要分为三种：基于规则的方法 (Garcia, 1997; Ittoo and Bouma, 2011)、基于统计的方法 (Zhao et al., 2016; Nauta et al., 2019)和基于深度学习的方法 (Zhang et al., 2015; Jin et al., 2020)。基于规则的方法使用模式匹配抽取因果关系，根据文本结构特征，人为制定规则，虽然准确度高，但泛化能力和可移植性较差。基于统计的方法虽然通过抽象得到的文本特征克服了依赖领域规则的问题，但它需要复杂的特征工程，很大程度上依赖于标注语料的质量，并且人工特征工程会带来额外的噪声，从而影响抽取精度。

近些年来，由于基于深度学习的方法能够从自然语言文本中自动学习文本特征，有效地解决了跨领域抽取及人工干预等问题，涌现了大量利用深度学习技术来抽取文本中因果关系的研究。文献 (Zeng et al., 2014)使用卷积神经网络CNN(Convolutional Neural Networks)对词汇向量和位置向量提取特征信息，提高了关系抽取的性能。但是，由于CNN不适合处理长距离的上下文语义特征，进一步地，文献 (Zhang et al., 2015)提出使用双向长短期记忆网络BiLSTM(Bidirectional Long Short-term Memory Networks)提取词汇特征，利用BiLSTM网络中的LSTM(Long Short-Term Memory,长短期记忆网络)单元实现了对文本关系的抽取。虽然BiLSTM可以对句子的长距离依赖关系进行建模，但是单纯的BiLSTM提取到的语义特征并不充分。文献 (Zeng et al., 2016) 提出了卷积双向LSTM模型，该模型利用双向LSTM提取句子

(c) 2022 中国中文信息学会计算语言学专业委员会

级别特征，利用CNN抽取单词在句子中的局部上下文特征，既增强了句子中的语义信息表示，又避免了人工特征导致的噪声问题，实现了对因果关系的抽取。预训练模型BERT (Jacob et al., 2018)出现后，文献 (Gao et al., 2021)提出了一种基于BERT预训练模型的因果关系联合提取模型，该模型使用BERT增强句子的语义表示，利用迭代扩展卷积增强事件的因果关系，从而实现对事件内部因果关系的抽取。

另外，还有一些因果关系抽取研究是基于序列标注方法展开的。文献 (姜博et al., 2021b)基于BiLSTM+CRF (Huang et al., 2015)和BERT预训练模型提出了BERT+BiLSTM+CRF方法，使用BERT预训练模型增强句子中字符的特征信息，可以有效地抽取文本中的因果关系。文献 (郑巧夺et al., 2021c)联合了卷积神经网络与双向门控循环单元BiGRU(Bidirectional Gated Recurrent Unit)，使用两次序列标注任务实现了因果关系的抽取，并且结合BERT预训练模型，增强了文本特征的表达能力，提升了模型对语义特征的提取能力。文献 (Li et al., 2021)提出了SCIFI因果关系抽取器，将上下文字嵌入应用到因果关系抽取任务中，并且使用多头注意力机制学习因果关系词之间的依赖关系，实现了对因果关系的直接提取。与传统的因果关系抽取方法相比，基于深度学习方法的因果关系抽取模型效果有明显提升，而且采用序列标注的方法，能够实现真正的因果关系抽取。

2.2 SoftLexicon

近些年来，许多基于序列标注方法对中文因果关系抽取的研究，通常采用基于字符的方式。但是由于中文文本预处理需要分词，使得基于字符的方式不可避免地会引入错误的分词信息，导致词汇的边界模糊和标注错误。针对这一问题，很多研究开始在序列标注任务中引入外部词汇信息，以增强基于字符的特征表示。文献 (Zhang and Yang, 2018) 提出了Lattice-LSTM模型，将字符与字符匹配到的词汇融合，用以引入外部词汇信息。但是该模型结构复杂、训练和推理速度慢，并且不具备可迁移性。为了解决可迁移性和计算复杂问题，文献 (Ma et al., 2020)提出了一种引入词汇信息的简易方法SoftLexicon，将字符与词典进行匹配得到与字符相应的匹配词，然后按照字符在词语中的位置，将匹配得到的词语分别放置在四个词集$\{B, M, E, S\}$中，尽可能保留了所有的词典匹配结果。其中，四个词集$\{B, M, E, S\}$分别表示该字符在词语中的开头、中间、结尾和单独构成一个字。如果在词典匹配后，词集为空，则以'None'填充该词集。为了能解决中文因果关系抽取方法对因果事件边界难以识别的问题，本文首次在中文因果关系抽取任务中引入外部词汇信息，利用SoftLexicon方法将字符信息与词汇信息融合，不仅利用了词汇的边界信息，还利用了词汇的语义信息，用以提升中文因果关系抽取的能力。

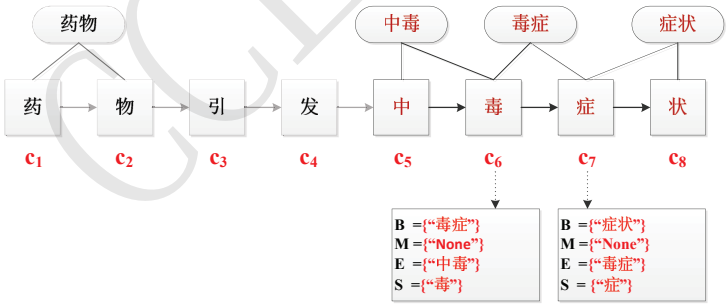


图 1: Softlexicon匹配示意图

如图 1所示，以'中毒症状'为例，字符'毒'与词典匹配可以得到三个词语'中毒'、'毒症'和'毒'。按照字符'毒'在这三个词语中的位置，于是，我们可以得到对应的四个词集，B={'毒症'}，M={'None'}，E={'中毒'}，S={'毒'}。

2.3 双向长短期记忆网络(BiLSTM)

循环神经网络RNN(Recurrent Neural Networks)自问世以来，由于其能够考虑上下文信息，被广泛用来处理时序性序列，但是该网络容易造成梯度消失和梯度爆炸。文献 (Hochreiter and Schmidhuber, 1997)在RNN基础上构建了长短期记忆神经网络(LSTM)，它既可以处理时序性序列，又可以缓解梯度消失问题。但LSTM只能根据前一个时刻预测下一个时刻，文

(c) 2022 中国中文信息学会计算语言学专业委员会

献 (Graves and Schmidhuber, 2005)在LSTM的基础上提出了双向长短期记忆网络(BiLSTM)，它由正向LSTM和反向LSTM组成，从两个方向建模句子的上下文信息，正向LSTM从前向后获取特征，后向LSTM从后向前获取特征。

2.4 注意力机制

近年来，许多研究在因果关系抽取任务中使用了注意力机制。文献 (Nauta et al., 2019)提出了一种基于注意力机制和卷积神经网络的因果关系抽取模型，通过注意力机制捕捉句子的上下文特征，完成了对时序性数据中因果关系的抽取。文献 (Jin et al., 2020)构建了级联结构神经网络CSNN，利用自注意力机制来挖掘句子内部的语义特征，实现了对文本中因果关系的抽取。文献 (Gao et al., 2021)建立了领域知识融合模型，利用注意力机制对因果知识建模，以捕获事件内部的语义特征，挖掘特征之间的内部因果关系。

3 BiLSTM-TWAM+CRF模型

本文提出的BiLSTM-TWAM+CRF模型结构主要由三部分组成，分别是嵌入层、BiLSTM-TWAM层和标签预测层，整体结构如图 2所示。首先，将输入的句子转换成基于字符的向量和基于SoftLexicon方法的向量，并融合两个向量。其次，将嵌入层得到的向量输入至BiLSTM-TWAM层中，使用BiLSTM 提取文本的长距离语义特征，并构建双路关注模块TWAM(Two Way Attention Module)学习字符之间的依赖关系，然后使用残差结构将BiLSTM和TWAM的输出融合，得到最终的语义特征。最后，将最终的语义特征输入到标签预测层中，计算序列的最优解，输出最优结果。接下来的部分将详细说明各个部分。

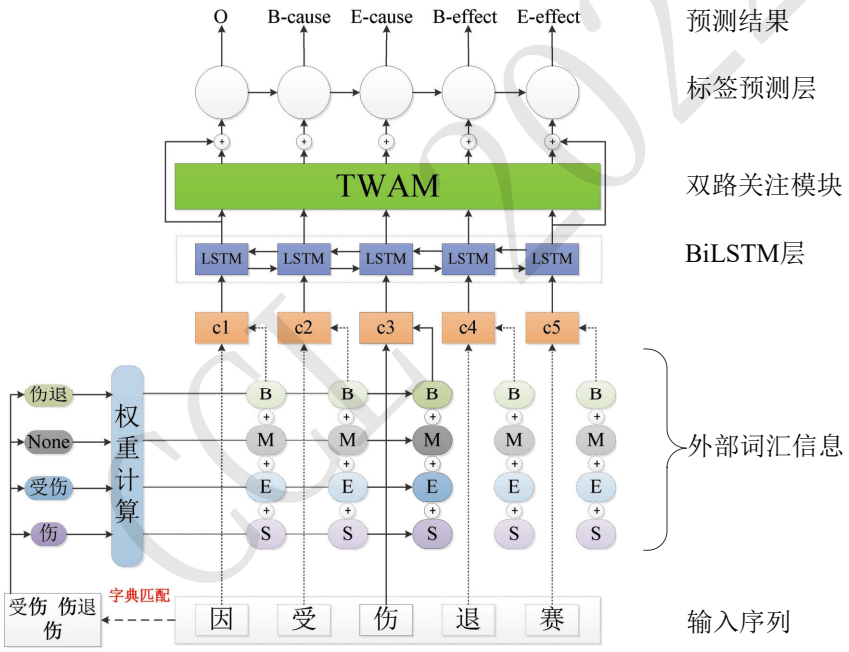


图 2: BiLSTM-TWAM+CRF模型结构

3.1 BiLSTM-TWAM层

本文提出的双路关注模块(TWAM)结构如图 3所示。BiLSTM-TWAM层由BiLSTM和TWAM组成，首先使用BiLSTM获得句子的长距离语义特征，然后利用TWAM提取深层次的语义信息，最后采用残差结构融合BiLSTM和TWAM的输出。

TWAM采用通道注意力和缩放点积注意力在局部和全局两个角度学习字符之间的依赖关系，使之能够更加充分地刻画文本特征。其中，通道注意力是按照通道对映射特征进行建模，将各通道的空间信息特征作为各通道的表示，使用池化操作聚合空间信息，提取出句子的局部特征。缩放点积注意力是对全部的映射特征进行建模，捕获到特征内部的相关性以及长距离依赖，提取出句子的全局特征。

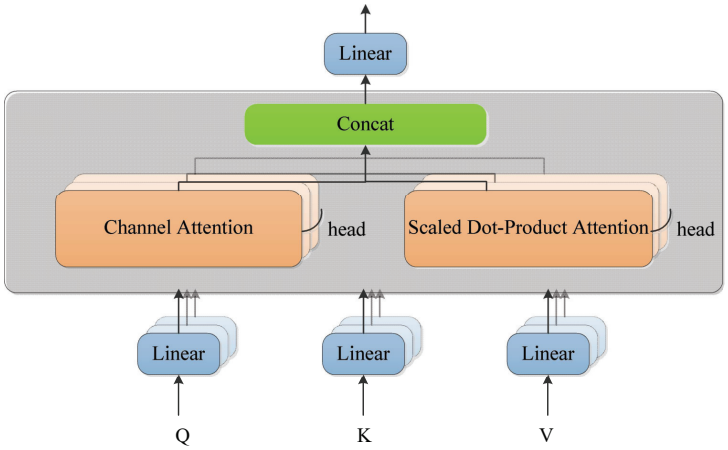


图 3: TWAM结构图

BiLSTM-TWAM层主要实现过程如下：

步骤一：将嵌入层得到的向量x^c输入BiLSTM中，对句子的长距离上下文语义特征进行提取得到语义特征表示H，然后将提取的特征表示H输入到TWAM 中。

步骤二：在TWAM得到来自BiLSTM的输出H后，建立查询矩阵Q、键矩阵K和值矩阵V，令Q=K=V=H。为获得更加丰富的语义信息，将Q矩阵、K矩阵和V矩阵分别映射至head个不同的子空间中，再输入至head个并行头中。

步骤三：每个并行头接收到不同子空间内的Q、K、V矩阵后，利用通道注意力和缩放点积注意力对不同子空间内的特征进行聚合，以达到关注不同子空间信息的目的。

首先，通道注意力使用平均池化和最大池化操作来聚合映射的空间信息特征，生成最大池化特征和平均池化特征，然后将两个特征送入多层感知器(MLP) 中，将MLP输出的特征进行融合，再经过sigmoid激活函数得到通道注意力结果，即文本的局部信息，计算如公式 1所示。

$$M_C(Q) = \sigma(MLP(AvgPool(Q) + MaxPool(Q))) \tag{1}$$

其中，σ为sigmoid函数，$AvgPool(Q)$和$MaxPool(Q)$表示平均池化特征和最大池化特征。

缩放点积注意力使用点积来计算Q矩阵与K矩阵的相似度，再除以$\sqrt{d_k}$(其中d_k为矩阵Q的维度)，并使用softmax函数计算权值，之后再乘以V矩阵得到缩放点积注意力结果，即文本的全局信息，计算如公式 2所示。

$$SDPA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

接着将每个并行头通过通道注意力得到的局部信息和缩放点积注意力得到的全局信息融合作为该并行头的结果，如公式 3所示。

$$\text{head}_\text{i} = (M_i(Q) \oplus SDPA(QW_i^Q, KW_i^K, VW_i^V)) \tag{3}$$

其中，W_i^Q、W_i^K和W_i^V分别为矩阵Q、K和V在第i个子空间内的权重矩阵，$\bigoplus$为拼接操作。

步骤四：将head个并行头得到的结果进行拼接后，再通过一个线性映射层得到TWAM的特征表示，计算如公式 4所示。

$$TWAM(Q, K, V) = Concat(head_1, \cdots, head_h)W^o \tag{4}$$

步骤五：最后本文使用残差结构将TWAM的特征表示与BiLSTM的特征表示融合，生成最终的语义特征表示。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.2 嵌入层

对于有n个字符的输入序列s=$\{c_1, c_2, \cdots, c_n\}$，每个字符$c_i$通过嵌入可以得到该字符的字符向量$x_{\mathrm{i}}^c$，如公式 5 所示。

$$x_{\mathrm{i}}^c = e^c(c_i) \tag{5}$$

其中e^c表示字符嵌入查找表。

利用SoftLexicon方法引入外部词汇信息，将输入序列s的每个字符与词典匹配，得到序列s所有的匹配词。然后将匹配词按照$\{B, M, E, S\}$划分为四个词集，$B(c_i)$表示以c_i开始的词集，$M(c_i)$表示c_i位于词语中间的词集，$E(c_i)$ 表示以c_i结尾的词集，$S(c_i)$表示c_i单独构成的词集，四个词集的表示如公式 6 所示。

$$\begin{cases} B(c_i) = \{\mathrm{w}_{i,k}, \forall w_{i,k} \in L, i < k \leq n\}, \\ M(c_i) = \{\mathrm{w}_{j,k}, \forall w_{j,k} \in L, 1 \leq j < i < k \leq n\}, \\ E(c_i) = \{\mathrm{w}_{j,i}, \forall w_{j,i} \in L, 1 < j \leq i\}, \\ S(c_i) = \{c_i, \exists c_i \in L\}. \end{cases} \tag{6}$$

其中，L表示本文使用的外部词典。

获得每个字符的$\{B, M, E, S\}$四个词集后，需要将每个词集压缩成一个固定的向量，在压缩过程中，我们用词频作为权重，进行动态加权处理，通过计算得到词集S的词集向量，计算如公式 7所示。

$$\begin{cases} v^s(S) = \frac{4}{Z} \sum\limits_{w \in S} z(w) e^w(w) \\ Z = \sum\limits_{\mathrm{w} \in B \cup M \cup E \cup S} z(w) \end{cases} \tag{7}$$

其中，$z(w)$表示词典中词w出现的频率，Z表示词集中所有词出现的频率之和，$v^s(S)$表示压缩后的集合向量。

完成词集的向量化后，采用向量拼接的方式，将四个词集添加到每个字符的表示中，每个字符的最终表示如公式 8所示。

$$\begin{cases} e^s(B, M, E, S) = [v^s(B), v^s(M), v^s(E), v^s(S)], \\ x^c \leftarrow [x^c; e^s(B, M, E, S)] \end{cases} \tag{8}$$

其中，x^c为融合词汇信息后的向量表示。

3.3 标签预测层

标签预测层采用的是条件随机场模型(Conditional Random Field,CRF)(Lafferty et al., 2001)。CRF是一种判别式的无向图模型，通过研究标签之间的关系，获得全局最优的标签序列。假设序列y=$\{y_1, y_2, \cdots, y_n\}$是给定输入句子$x$=$\{x_1, x_2, \cdots, x_n\}$的标签序列，则该序列的CRF分数计算如公式 9所示。

$$score(x, y) = \sum_{i=1}^{n} p_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} \tag{9}$$

其中，p_{i,y_i}表示句子中第i个字符为y_i标签的预测概率，A_{y_{i-1}, y_i}表示标签y_{i-1}到标签y_i的转移概率。

本文使用Viterbi(Forney et al., 1973)算法输出具有最大$score(x, y)$的标签序列，当损失函数最小时，表示该模型已经收敛，损失函数如公式 10所示。

$$E = log \sum_{y \in \mathrm{Y}} \exp^{s(y)} - score(x, y) \tag{10}$$

其中，Y表示句子x可能对应的标签序列的集合。

第二十一届中国计算语言学大会论文集，第190页-第200页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

4 实验与结果分析

4.1 实验数据集

本文实验选取中文突发事件数据集(Chinese Emergency Corpus,CEC)[0]和百度中文事件抽取数据集(DUEE)(Xinyu et al., 2020)作为语料。CEC数据集由上海大学语义智能实验室构建，包括地震、火灾、交通事故、恐怖袭击和食物中毒五个类别，借助互联网收集CEC生语料，按照XML语言对语料进行标记。DUEE数据集是百度发布的中文事件抽取数据集，从百度资讯信息文本中收集语料，共有六十五个事件类型。

本文采用'BMESO'序列标注方法对数据集中文本含有的因果事件进行标注。其中，'B-'表示该字在事件的开头，'M-' 表示该字在事件的内部，'E-'表示该字在事件的结尾，'S-' 表示该字本身为一个事件，'-cause'表示为原因事件，'-effect'表示为结果事件，'O' 表示无关字符。由于CEC数据集使用XML语言进行标记，首先对数据集去除HTML标签进行了格式处理，提取文本数据并获取其*Participant*、*Time*、*Denoter*和*Location*标签作为因果事件标注的依据。接着，用人工标注的方法，对获取到的文本数据标注出原因事件、结果事件和其他，最终从CEC 数据集提取出了1,026 条样本数据。对于DUEE数据集，首先获取其*Text*标签作为文本数据，同时将其*Event-Type*、*Trigger*和*Class* 标签作为因果事件标注的依据，再仿照CEC数据集，使用人工标注的方法对文本数据标注，最终提取出4,800条样本数据。两个数据集完成人工标注工作后，按照7:1:2的数量比例将两个数据集划分为训练集、验证集与测试集。CEC 数据集与DUEE数据集详细信息如表 1所示。

	CEC	DUEE
因果事件对	844	2,805
原因事件	898	2,938
结果事件	1,223	3,348

表 1: CEC和DUEE数据集描述

4.2 实验参数设置

本文用word2vec (Tomas et al., 2013)训练得到字符向量和词向量，选用Adam (Kingma and Ba, 2015)作为优化器，参数设置如表 2所示。

参数	参数设置
字向量维度	50
词向量维度	50
学习率	0.005
迭代次数	50
批大小	16
Dropout	0.5
BiLSTM隐藏单元数	100
TWAM中head	4

表 2: 实验参数

4.3 评价指标

本文实验根据句子中抽取得到的因果事件对是否正确来判定模型的抽取性能，即对于一组因果关系，若原因事件与结果事件均抽取正确，则因果关系抽取正确，否则抽取错误。本文将抽取因果事件对的准确率P、召回率R和$F1$值作为评价指标。

[0]https://github.com/shijiebei2009/CEC-Corpus

(c) 2022 中国中文信息学会计算语言学专业委员会

4.4 基线

为验证本文模型的有效性，本文与近几年提出的因果关系抽取模型和序列标注模型进行了对比实验。

(1) CNN-BiGRU (苗佳et al., 2021a)：用于事件触发词抽取，利用CNN提取词汇特征，使用BiGRU提取句子特征。

(2) CSNN (Jin et al., 2020)：级联结构模型，结合CNN和具有自注意力机制的LSTM对因果关系进行抽取。

(3) BERT+CNN-BiGRU (郑巧夺et al., 2021c)：基于残差思想的双层因果关系抽取方法，使用BERT提取语义特征，使用双层CNN-BiGRU模型增强语义表征能力。

(4) BiLSTM+CRF (Huang et al., 2015)：经典的序列标注模型，由BiLSTM与CRF分类器构成。

(5) SoftLexicon+BiLSTM+CRF (Ma et al., 2020)：字词联合的序列标注模型，利用SoftLexicon方法引入了词汇信息，在序列标注任务中取得了好的效果。

4.5 结果分析

4.5.1 不同方法的对比

本文提出的BiLSTM-TWAM+CRF模型采用 4.2节的参数设置，对比模型使用的参数设置都参考其原论文中的描述，对比实验结果如表 3 所示。

	CEC			DUEE		
模型	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$(%)
CNN-BiGRU	74.65	61.44	67.37	70.39	57.03	62.99
CSNN	70.55	54.87	61.72	71.99	62.10	66.56
BiLSTM+CRF	69.32	63.70	66.28	65.26	52.51	58.08
SoftLexicon+BiLSTM+CRF	74.00	65.33	69.33	70.48	**71.68**	71.04
本文模型	**76.43**	**71.90**	**74.04**	**76.40**	71.63	**73.80**
BERT+CNN-BiGRU	73.35	72.38	72.78	72.98	65.67	69.01
BERT+本文模型	**80.27**	**76.42**	**78.39**	**77.70**	**72.23**	**74.86**

表 3: 对比实验结果

从表 3可以看出，本文提出的BiLSTM-TWAM+CRF模型在CEC数据集和DUEE数据集上都取得了较好的效果。其中，对比没有引入外部词汇信息的模型，序列标注模型SoftLexicon+BiLSTM+CRF在三个评价指标上均表现突出，表明了利用SoftLexicon引入外部词汇信息对中文因果关系抽取任务的有效性。我们分析这主要是因为引入外部词汇信息后，与单纯基于字符向量的模型相比，基于字词联合的模型不仅能够利用词汇的边界信息，还可以利用词汇的语义信息，从而有效避免了单纯基于字符向量的方法不能够准确确定词语边界和标注错误的问题。

同时也可以观察到，在CEC数据集上，本文模型的效果优于序列标注模型SoftLexicon+BiLSTM+CRF，表明TWAM能够提升模型的中文因果关系抽取能力。TWAM中的通道注意力和缩放点积注意力，能够在局部和全局两个角度提取特征信息，可以更加充分的刻画句子的语义特征。TWAM中残差结构的引入也使得BiLSTM-TWAM层既能获取文本的上下文信息，又能对特征进行更深层次的特征提取。从表 3可以看到，在DUEE数据集上准确率P 和$F1$值分别提高了6.08%和2.76%，召回率R仅降低了0.05%，表明本文提出的TWAM在中文因果关系抽取任务中的有效性。

进一步地，从表 3中可知，本文所提模型效果均优于基线模型，尤其与现有因果关系抽取方法相比较，本文模型在准确率P、召回率R 和$F1$值上均有大幅提高，表明本文所提模型在中文因果关系抽取任务中，能够更加精确地抽取出文本中的因果事件对。另外，可以看到，本文

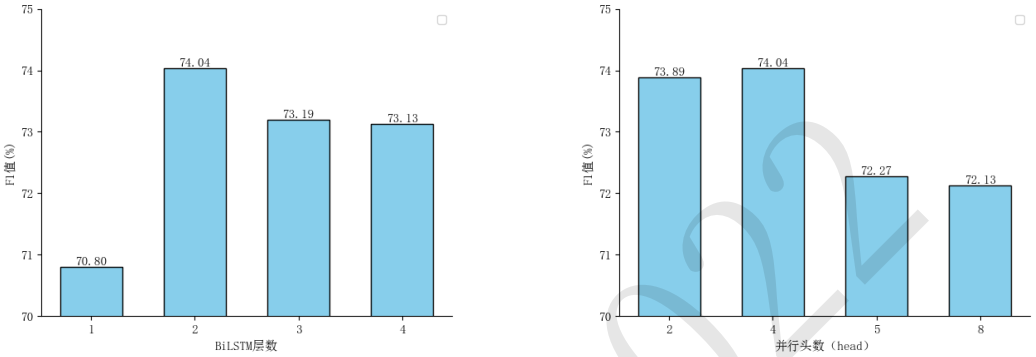
(c) 2022 中国中文信息学会计算语言学专业委员会

所提模型相较于基线模型在DUEE数据集上准确率的提升效果比CEC数据集显著。分析原因一是因为大规模数据能够减少噪声对因果关系抽取效果的影响，二是本文模型在大规模数据集上学习到的语义特征更为充分，能够更加准确地识别因果事件对。

此外，我们也发现加入预训练模型BERT后，使用BERT的模型效果比没有使用BERT的模型效果更好，表明预训练模型BERT丰富的语义知识能够提高因果关系抽取能力，并且BERT+本文模型的效果优于其他模型，表明本文提出的BiLSTM-TWAM+CRF模型能够有效地与预训练模型相结合，提高因果关系抽取性能。

4.5.2 超参数的选取

BiLSTM-TWAM层作为本文模型中的重要组成部分，为了评估BiLSTM的堆叠层数，以及TWAM中并行头(head)的个数对模型的影响，本文继续在CEC数据集上使用因果事件对作为评估标准进行了实验，实验结果如图 4所示。



(a) BiLSTM层数对$F1$值的影响　　(b) 并行头数(head)对$F1$值的影响

图 4: 模型参数设置对$F1$值的影响

从图 4可以看出，在其他参数相同的情况下，当BiLSTM层数与并行头数(head)分别取2和4时，模型抽取因果事件对的$F1$值最高。通过实验发现，将BiLSTM堆叠层数设置过大，会导致模型更加复杂和参数增多，学习到的特征会过于抽象而且会包含一些无用信息，设置过小又不能充分提取特征，使提取到的上下文特征缺乏语义信息表达。同时，当TWAM层中head数取5时，本文模型的$F1$值比head数取4时减少1.77%，表明随着head的数量增多，每个子空间的内含有的特征信息会逐渐减少，导致每个并行头无法提取到足够的特征信息，从而造成没有充足的语义特征表示。

4.5.3 消融实验

为了进一步验证本文所提BiLSTM-TWAM+CRF模型中每个组成部分的贡献，本文分别在CEC数据集和DUEE数据集上进行了消融实验，结果如表 4所示。

	CEC			DUEE		
模型	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$(%)
本文模型	76.42	**71.89**	**74.04**	**76.40**	**71.63**	**73.80**
-SoftLexicon	68.88	60.58	64.46	67.10	61.70	64.28
-TWAM	70.00	66.42	68.16	66.62	67.81	67.21
-残差融合	78.97	61.68	69.26	72.38	71.09	71.73

表 4: 消融实验结果

从表 4可以看出，本文所提模型的各个模块都发挥了一定的作用。模型在不使用SoftLexicon方法引入外部词汇信息时，CEC数据集与DUEE数据集的$F1$值分别下降了9.58%和9.52%，表明引入外部词汇信息能够增强字符向量的语义信息表示，使模型获取

(c) 2022 中国中文信息学会计算语言学专业委员会

到更多的文本特征信息。模型在不使用TWAM时，CEC数据集与DUEE数据集的$F1$值分别下降了5.88%和6.59%，表明引入TWAM能够在局部和全局两个角度学习文本序列中的特征信息，提取到更全面更深层次的语义特征。当没有采用残差结构融合TWAM和BiLSTM输出的特征时，CEC数据集与DUEE数据集的$F1$值分别下降了4.78% 和2.07%，表明该结构的应用，有效丰富了因果关系抽取任务中的语义特征，使得BiLSTM-TWAM层既能获取长距离的上下文特征表示，又能对特征在局部和全局角度进行更深层次的特征提取。实验结果表明，通过SoftLexicon方法引入的外部词汇信息对于模型的贡献最大，证明了在中文因果关系抽取任务中引入外部词汇信息的重要性，也表明基于字词联合的模型能够很大程度上提升中文因果关系抽取的能力。

5 结语

本文面向中文因果关系抽取，提出了一种基于外部词汇信息和注意力机制的因果关系抽取模型BiLSTM-TWAM+CRF，从一定程度上，实现了真正意义上的因果关系抽取。总体而言，本文提出的模型能够有效解决词语边界模糊和语义表征不充分的问题，具有较好的应用前景。后续工作将尝试从多特征融合角度来提升模型的多语种因果关系抽取效果。

参考文献

杨竣辉, 刘宗田, 刘炜, and 苏小英. 2016. 基于语义事件因果关系识别. 小型微型计算机系统, 37(3):433–437, January.

冯冲, 康丽琪, 石戈, and 黄河燕. 2018. 融合对抗学习的因果关系抽取. 自动化学报, 44(5):811–818, January.

苗佳, 段跃兴, 张月琴, and 张泽华. 2021a. 基于cnn-bigru模型的事件触发词抽取方法. 计算机工程, 47(9):69–74,83, September.

姜博, 左万利, and 王英. 2021b. 基于bert的因果关系抽取. 吉林大学学报*(*理学版*)*, 59(6):1439–1444, November.

郑巧夺, 吴贞东, and 邹俊颖. 2021c. 基于双层cnnbigrucrf的事件因果关系抽取. 计算机工程, 47(5):58–64,72, May.

Forney, G. D., and Jr. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Jianqi Gao, Xiangfeng Luo, Hao Wang, and Zijian Wang. 2021. Causal event extraction using iterated dilated convolutions with semantic convolutional filters. In *33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2021*, pages 619–623. IEEE, November.

Daniela Garcia. 1997. Coatis, an NLP system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management*, volume 1319 of *Lecture Notes in Computer Science*, pages 347–352. Springer, October.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, volume 12 of *MultiSumQA '03*, pages 76–83, USA. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*, volume 4, pages 2047–2052.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, pages 987–997. Association for Computational Linguistics, June.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023.

第二十一届中国计算语言学大会论文集，第190页-第200页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *CoRR*, abs/1508.01991.

IAshwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Natural Language Processing and Information Systems*, volume 6716 of *Lecture Notes in Computer Science*, pages 52–63. Springer, June.

Devlin Jacob, Changming Wei, Lee Kenton, and Toutanova Kristina. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xianxian Jin, Xinzhi Wang, Xiangfeng Luo, Subin Huang, and Shengwei Gu. 2020. Inter-sentence and implicit causality extraction from chinese corpus. In *Advances in Knowledge Discovery and Data Mining*, volume 12084 of *Lecture Notes in Computer Science*, pages 739–751, Cham. Springer.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, May.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001),*, pages 282–289, San Francisco, CA, USA, June. Morgan Kaufmann.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.

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, ACL 2020*, pages 5951–5960. Association for Computational Linguistics, July.

Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 909–918. Association for Computing Machinery, April.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg, and Dean Jeffrey. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119, December.

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*, volume 30, pages 5998–6008, December.

Li Xinyu, Li Fayuan, Pan Lu, Chen Yuguang, Peng Weihua, Wang Quan, Lyu Yajuan, and Zhu Yong. 2020. Duee: A large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing*, pages 534–545, Cham, October. Springer International Publishing.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344. ACL, August.

Ying Zeng, Honghui Yang, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2016. A convolution bilstm neural network model for chinese event extraction. In *Natural Language Understanding and Intelligent Applications*, volume 10102 of *Lecture Notes in Computer Science*, pages 275–287. Springer International Publishing.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 1554–1564, Melbourne, Australia, July. Association for Computational Linguistics.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China, October. ACL.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于GCN和门机制的汉语框架排歧方法*

游亚男[1,‡]，李茹[1,2,⋆,†]，苏雪峰[1,3,‡]，闫智超[1,‡]，孙民帅[1,‡]，王超[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
‡{2280493770，455375251，751824801，1070913573，1055342647}@qq.com
⋆{liru}@sxu.edu.cn

摘要

汉语框架排歧旨在候选框架中给句子中的目标词选择一个符合其语义场景的框架。目前研究方法存在隐层向量的计算与目标词无关，并且忽略了句法结构信息对框架排歧的影响等缺陷。针对上述问题，使用GCN对句法结构信息进行建模；引入门机制过滤隐层向量中与目标词无关的噪声信息；并在此基础上，提出一种约束机制来约束模型的学习，改进向量表示。该模型在CFN、FN1.5和FN1.7数据集上优于当前最好模型，证明了方法的有效性。

关键词： 汉语框架排歧；句法信息；GCN；门机制

Chinese Frame Disambiguation Method Based on GCN and Gate Mechanism

Yanan You[1,‡], Ru Li[1,2,⋆,†], Xuefeng Su[1,3,‡], Zhichao Yan[1,‡], Minshuai Sun[1,‡], Chao Wang[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

Chinese frame disambiguation aims to select a frame that matches its semantic scene for the target word in the sentence among the candidate frames.The current research methods have the defects that the calculation of the hidden layer vector has nothing to do with the target word, and ignores the influence of the syntactic structure information on the frame disambiguation. Aiming at the above problems, GCN is used to model the syntactic structure information; a gate mechanism is introduced to filter the noise information irrelevant to the target word in the hidden layer vector; and on this basis, a constraint mechanism is proposed to constrain the learning of the model and improve the representation vector . The model outperforms the current state-of-the-art models on the CFN, FN1.5 and FN1.7 datasets, proving the effectiveness of the method.

Keywords: Chinese frame disambiguation , Syntactic information , GCN , Gate mechanism

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）
† 通讯作者 Corresponding Author
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

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知识库（FrameNet，FN） (Baker et al., 1976)构建的汉语词汇语义知识库，包括框架库、句子库和词元库，其中相关术语如表1 (Li, 2012)所示。汉语框架语义分析是基于汉语框架网的标注资源提出的任务 (Shi et al., 2014)，而汉语框架排歧作为汉语框架语义分析中的重要研究任务，其正确与否直接关系到汉语框架语义分析的准确性。

汉语框架排歧任务即给定一个句子和一个目标词以及该目标词在CFN中能激起的所有框架，需根据目标词的上下文推断出该目标词在当前语境下所匹配的框架。如表2所示，在CFN中，目标词“炒”可以激起两个框架“解雇”、“烹饪”，“解雇”表示雇主终结与雇员之间的雇佣关系；“烹饪”表示烹调者对食物进行加工。在s1中，根据上下文信息可以判断“炒”激起框架为“解雇”，而s2中“炒”激起框架为“烹饪”。

术语	定义
目标词	在一个具体的句子中能够激起框架的词元
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
框架元素	框架的语义参与者，也称为框架语义角色

表 1: 框架语义分析术语表

词元	候选框架	例句
炒	烹饪、解雇	s1：如果再这么下去，老板<炒tgt “解雇”>你没商量。
		s2：舅妈特地<炒tgt “烹饪”>了过油肉，迎接我们的到来。

表 2: 语料示例

目前大多数研究将框架排歧看作多分类任务，主要有两类方法：一种是基于传统机器学习的方法，该类方法使用自然语言处理工具分析句子，人工抽取特征，通过机器学习方法训练模型。该类方法得到的特征矩阵维度大，特征稀疏，并且特征之间没有关联，模型难以在不同的树结构上并行计算，计算效率低下。二是基于深度学习的方法，该类方法实现了自动学习特征，避免了特征维度过高，有效的融合了上下文信息，一定程度上解决了特征之间无关联的问题，并且已被用于产生该任务的最先进性能。已有工作将依存信息集成到深度学习的模型中用来进一步提高模型性能，取得了一定的效果。但这些方法仍然存在以下不足：a)在框架排歧中不同目标词在上下文中关注的词应该有所不同，但现有方法在进行向量计算时并没有考虑这一点。如表2中，句子s2中，“炒”和“迎接”是两个目标词，在对它们进行框架排歧时希望模型得到的句子表示是不同的；b)已有方法只考虑句法上与目标词相邻的词，忽略了句法不相邻词对目标词的影响。

本文首先使用BERT (Devlin et al., 2018)进行上下文编码，利用哈工大语言技术分析平台（LTP） (Che et al., 2010)进行句法分析以此构建依存图。然后通过GCN对依存图中的信息交互进行建模，充分利用GCN对图的特征提取能力。针对a)问题，考虑到BERT得到的隐层向量并不是特定于目标词的，输入GCN更新节点的过程中会保留与目标词无关的冗余信息，因此本文引入门机制，计算一个特定于目标词的门向量，并将门向量应用于每个GCN层，改变上下文的表示，得到特定于目标词的隐层向量来过滤无关特征；针对b)问题，本文认为句法不相邻词对于目标词的表示学习具有指导作用，因此本文提出依据句法信息为句子中的每个词分配一个分数，明确量化其对目标词进行框架排歧的贡献，注入模型，以此来约束模型的学习，改进表示向量。

本文的贡献之处包括：1）提出一种基于GCN和门机制的向量调整方法，生成特定于目标词的向量表示，过滤与目标词无关的噪声信息。2）引入一种基于依存图的约束机制来获取句

(c) 2022 中国中文信息学会计算语言学专业委员会

子中每个词相对于目标词的重要性得分，注入模型，作为计算隐层向量的训练信号，改进向量表示。3）在数据集CFN、FrameNet1.5和FrameNet1.7上进行了详细的对比实验，实验结果表明，本文方法有效提高了框架排歧的准确率。

2 相关工作

著名语言学家Fillmore基于认知角度提出了框架语义学。此后，SemEval2007语义评测任务 (Baker et al., 2007)提出了框架语义结构抽取任务，包括目标词识别、框架识别和语义角色标注等任务。框架排歧作为框架识别的子任务也受到了广泛关注。

早期的框架排歧模型采用传统的机器学习方法，人工构建特征，使用条件随机场、最大熵、支持向量机等模型来建模。 (Li et al., 2011)使用开窗口技术和BOW策略抽取了词包等若干特征，用最大熵模型建模，特征信息稀疏。 (Li et al., 2013)提出了特征模板的自动选择算法，通过打分机制将得分高的特征加入特征模板，使用最大熵模型来进行框架排歧。这些传统的机器学习算法，人工选择了大量特征导致空间维度过高，特征稀疏，费时费力。

随着深度学习的发展，近年来有许多研究采用神经网络模型来进行框架排歧。 (Hermann et al., 2014)使用WSABIE算法将目标词以及框架表示学习映射到同一空间，计算它们之间的距离进行框架识别。 (Zhao et al., 2016)提出了一种通用的框架识别模型，通过使用DNN架构来学习目标词的上下文特征进行框架识别，对于未登录词元和歧义词元的框架识别有了较好的泛化能力。 (Zhang et al., 2017)针对人工抽取特征使得特征空间维度过高和特征之间缺乏关联性的问题，在词语分布式表征的基础上提出了基于距离和词语相似度矩阵的框架排歧模型，证明词语分布式表征对框架排歧的有效性。 (Botschen et al., 2017)使用Word2Vec训练词向量表征上下文来进行框架识别。 (Hou et al., 2020)提出了一种基于hinge-loss的框架表示学习算法，通过计算目标词表示和框架表示之间的相似度来进行框架排歧，相较之前的工作有了明显的提升。 (Guo, 2021)将BERT与Bi-GRU结合起来编码上下文信息，使用注意力机制融入局部和全局信息来进行框架识别，在CFN和FrameNet上提高了框架识别的准确率。 (Su et al., 2021)通过融入框架关系和框架定义来进行框架识别。但以上工作都不是特定于目标词生成的向量表示，会保留与目标词无关的噪声信息；并且未充分考虑句法信息的重要性。在框架语义角色标注中，大多句法信息是目标词的语义角色，对于目标词的所属框架选择有着重要作用。 (Li et al., 2010)使用层次条件随机场（T-CRF），将框架识别视为依赖树结构上的标注任务。 (Wang et al., 2013)使用T-CRF模型建模，选取词、词性和不同类型的依存特征进行框架语义角色自动标注。但以上模型难以在不同的树结构上并行计算，计算效率低下。近年来，图卷积网络（Graph Convolutional Networks，GCN） (Kipf and Welling, 2016)的兴起为依存树的构建提供了新的思路。GCN可以有效存储任何结构的依赖树信息，并且能够并行计算，计算效率得到了很大的提升。 (Zhang et al., 2018)将BiLSTM和GCN结合起来编码句子中的句法信息，用来解决关系抽取任务，证明了GCN编码依存树的有效性。

因此本文提出了基于GCN和门机制的框架排歧模型生成特定于目标词的向量表示，并利用句法信息构造依存图，使用一种基于依存图的约束机制来约束模型学习，改进向量表示。相比基于上下文生成的向量表示，使用目标词附近的局部信息，句法信息对目标词更为重要。实验结果表明，模型在CFN和FrameNet框架排歧数据集上取得了一定的提升。

3 基于GCN和门机制的框架排歧模型

在框架排歧中，给定句子$s:\{c_1,...,c_i,...c_n\}$和目标词$c^{'}$（可能由多个字组成），对于目标词$c^{'}$，它所能激起的框架$F=\{f_1,...,f_g\}$来自CFN的框架库，框架排歧任务就是在候选框架列表$\{f_1,...,f_g\}$中为$c^{'}$在当前语境句子s下找到最合适的框架，其形式化描述如公式（1）所示。

$$f=\arg\max_{f_i\in F} P\left(f_i|c^{'},s\right) \tag{1}$$

本文提出了一种基于GCN和门机制的框架排歧模型，模型整体架构如图1所示，该模型是针对给定目标词和包含该目标词的句子，通过模型训练得到目标词的表征，进行框架排歧。模型的整体包括编码层、依存图抽取模块、融合门机制的图卷积网络层（GGCN）、基于依存图的约束机制、分类层五个模块。本文通过预先训练的BERT来获取基于上下文的词表征，使用LTP进行句法分析，以此构建依存图，将图和词表征输入GCN来更新图节点的信息，为每

(c) 2022 中国中文信息学会计算语言学专业委员会

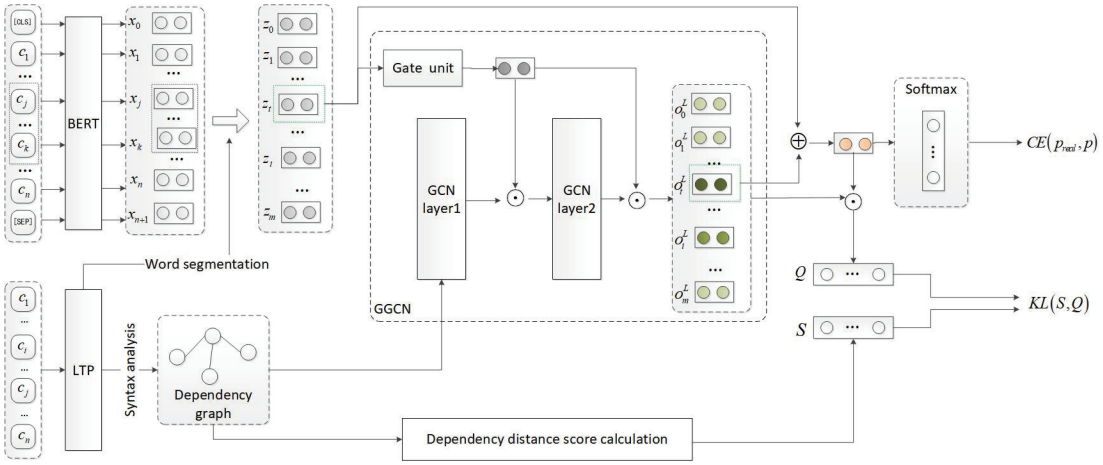


图 1: 基于GCN和门机制的框架排歧模型

个GCN层增加一个门，以过滤与目标词无关的信息；并依据依存图计算句子中每个词到目标词的距离，取其负值作为每个词基于句法的得分，输入模型，监督模型的训练，改进向量表示；将模型最终得到的目标词表征，通过softmax函数，得到一个向量表示，向量每个维度上的值为候选框架的概率分布，选择概率值最大的框架作为正确框架。

3.1 编码层

BERT是一个大规模的预训练模型，以无监督的方式对大规模未标记的语料库进行训练，能够大规模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Transformer encoder，相比传统的Transformer拥有双向编码能力，可以更彻底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。因此本文使用BERT作为编码层，对预训练的BERT进行了微调来适应框架排歧任务。

将“[CLS]+s:$\{c_1,\cdots,c_i,\cdots,c_n\}$+[SEP]”作为模型输入，编码层将输入中的每一个字符编码成字符嵌入$E_{token}(c_i)$、分段嵌入$E_{seg}(c_i)$和位置嵌入$E_{pos}(c_i)$三个向量，将三个向量相加输入BERT预训练模型得到输入的BERT向量$X \in R^{n\times d}$，如公式（2）、（3）所示。

$$E_i = E_{token}(c_i) + E_{seg}(c_i) + E_{pos}(c_i) \tag{2}$$

$$X = BERT(E_0,\cdots,E_i,\cdots,E_{n+1}) \tag{3}$$

Token Embedding：模型通过查询向量表将输入中的每个字符转换为一维向量。

Segment Embedding：框架排歧只输入模型一个句子，Segment Embedding全设为0。

Position Embedding：Transformers无法编码序列的顺序性，而文本中不同位置的字符携带的信息是不同的，应该用不同的向量表示，通过让BERT模型为每个位置学习一个位置嵌入来编码序列的顺序性。

本文采用对应于目标词token的BERT向量（如果有多个token，将其平均）作为目标词的表示z_t，如公式（4）所示。

$$z_t = avg(x_j,\cdots,x_k) \tag{4}$$

3.2 依存图抽取模块

从图2中可以看出，与目标词有依存句法关系的词往往是目标词的框架元素，对目标词所属框架的确定起重要作用。目标词“炒”的直接依存信息主语（SBV）、状语（ADV）和宾语（VOB）分别对应框架“烹饪”的框架元素“烹调者（cook）”、“方式（manr）”、“食物（food）”等，但如果只考虑直接依存关系很可能会忽略目标词的一些有用的信息，例如“迎接我们的到来”也是“烹饪”的框架元素。因此本文采用多层GCN来聚合直接或间接依存信息。

本文直接使用哈工大的LTP工具包来进行分词和句法分析，利用得到的分词信息将BERT得到的字向量做平均得到词的表示作为GCN的节点输入。

(c) 2022 中国中文信息学会计算语言学专业委员会

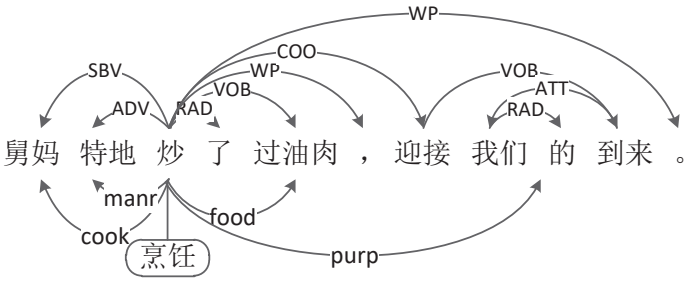


图 2: “炒”的依存关系和角色标注

3.3 融合门机制的图卷积网络层

从3.2中可知，框架排歧任务与目标词的依存句法信息密切相关。之前使用依存信息的框架排歧工作都是将依存特征通过拼接的方式来融入，特征不存在时用0向量表示，得到的特征向量维度高且稀疏。为了更好的融入目标词的依存信息，本文引入了图卷积网络（GCN），它是卷积神经网络（CNN）的一种改编，完善了CNN在非矩阵结构数据上的不适用性。GCN在进入下一层非线性变换之前，每个节点先将其自身的邻居节点的信息通过非线性方式聚合，相比BiLSTM和word2vec只利用了词周围的信息，GCN采用多层卷积，节点不仅利用了自身邻居节点的信息，同时也聚合了邻居的邻居信息，相比之下感受野更广，能够利用更大范围的信息。本文使用L层GCN来更新依存图的节点信息并抽出聚合依存句法信息的目标词表示。

首先定义无向图$G = (V, E)$作为句子s的依存图，其中$V = \{v_1, ..., v_m\}$是图节点的集合，是将句子s经过LTP分词得到句子的分词结果。E是图边的集合，$(v_i, v_j) \in E$表示第i个词和第j个词之间存在有向句法弧，为了实现信息的反向传播，向E中添加一条与有向句法弧方向相反的边(v_j, v_i)，同时为了利用节点自身的信息，对所有的节点向E中添加一个自循环，即(v_i, v_i)。完成构图之后使用神经网络模型$GCN\,(Z, A)$对图结构进行编码。首先获得节点的特征矩阵Z并计算图的邻接矩阵A。利用将经过BERT得到的字向量做平均作为图节点的特征矩阵$Z \in R^{m \times d}$；通过E构建邻接矩阵$A \in R^{m \times m}$，$A_{i,j} \in \{0, 1\}$表示第i个词和第j个词之间是否存在边。

GCN层的计算如公式（5）所示：

$$GCN\,(Z, A) = \hat{A}ReLU(\hat{A}ZW^{(l)})W^{(l+1)} \tag{5}$$

其中$\hat{A}$是基于对角矩阵D的正则化邻接矩阵，计算如公式（6）所示：

$$\hat{A} = \widetilde{D}^{-\frac{1}{2}}A\widetilde{D}^{-\frac{1}{2}} \tag{6}$$

$W^{(l)}$为第l层的权值矩阵，用于将节点的特征表示映射到相应的隐层状态。$W^{(l+1)}$为第$l+1$层的权值矩阵，用于将节点的隐层表示映射为相应的输出。

由于GCN生成的隐层向量不是特定于目标词产生的，会导致隐层向量融入与目标词无关的噪声信息。而在框架排歧中，更关注与目标词相关的信息，因此本文引入门机制来进行特征过滤，利用目标词的嵌入表示z_t计算得到门向量g^l，通过元素乘积将门向量应用到GCN相应的隐层向量，得到过滤向量o_i^l，计算公式如下：

$$g^l = \sigma\left(W_g^l z_t\right) \tag{7}$$

$$o_i^l = g^l \odot h_i^l \tag{8}$$

其中，W_g^l是GCN第l层的可学习参数，h_i^l是GCN第l层第i个节点对应的隐层向量。

3.4 基于依存图的约束机制

在使用GCN融入句法信息时，更关注依存图中邻居节点的信息，虽然非相邻词可能不会直接为目标词带来有用的上下文信息，但本文认为其仍然可以为隐层向量的计算提供有用的训练信号。所以本文提出一种基于依存图的约束机制，根据依存图计算句子中每个词到目标词的距离，取其负值作为句子中每个词基于句法的得分S，并且本文认为框架排歧句子中每个词的

(c) 2022 中国中文信息学会计算语言学专业委员会

重要程度可以用其携带的有用信息来衡量，如果词c_i的过滤向量o_i^L与目标词的最终表示z'更相似，那么词c_i对于目标词更重要，使用公式（9）计算每个词相对于目标词基于模型的重要性得分Q，其中W^z和W^o是可训练参数。将S和Q分别经过softmax进行归一化。本文认为两个得分之间应该具有一致性，而KL散度是用来度量两个概率分布相似度的指标，因此通过计算两个得分之间的KL散度衡量两者之间的差异，并加入损失函数来最小化两者之间的差异，计算公式如式（10）所示。

$$q_i = \sigma\left(W^z z'\right) \cdot \sigma\left(W^o o_i^L\right) \tag{9}$$

$$KL\left(S, Q\right) = -\sum_{i=1}^{n} s_i \frac{s_i}{q_i} \tag{10}$$

3.5 分类层

直接在GCN模块后计算损失、梯度下降，会导致梯度消失，所以本文使用线性插值法插值BERT得到的基于上下文的目标词表示z_t和GCN融合了句法信息的目标词表示o_i^L作为最终的目标词表征z'，如公式（11）所示，其中μ是权重系数。

$$z^{'} = \left(1-\mu\right) z_t + \mu o_t^L \tag{11}$$

将z'输入全连接层进行分类，通过softmax层计算各个候选框架概率值，如公式（12）所示。

$$p = soft\max\left(z^{'}\right) \tag{12}$$

框架排歧为多分类任务，所以模型采用交叉熵损失函数作为分类损失，如式（13）所示，其中p_{real}表示真实样本类别分布。

$$CE(p_{real}, p) = -\sum p_{real} \log\left(p\right) \tag{13}$$

最后，本文采用式（10）和式（13）组合作为模型整体的损失函数来训练模型，其中α为权重系数，计算公式如下：

$$loss = CE(p_{real}, p) + \alpha KL\left(S, Q\right) \tag{14}$$

4 实验设计与分析

4.1 实验数据

本文使用的框架排歧数据来源于CFN数据库（http://sccfn.sxu.edu.cn/portal-en/frame.aspx）中抽取出来的88个有歧义的词元，共10012条数据，涉及到90个框架，训练集和测试集按8:1:1的比例分配，如表3所示。

同时为了验证模型的可行性和有效性，本文在FrameNet1.5和FrameNet1.7数据集上进行了实验。FrameNet1.5和FrameNet1.7数据分布如表3所示。

数据集	CFN	FN1.5	FN1.7
训练集	8001	16092	19152
验证集	1004	2197	2263
测试集	1007	4320	6698
词元数	88	2931	3416
框架数	90	704	796

表 3: 数据集分布

4.2 实验指标

本文采用准确率作为评价指标，计算如公式（15）所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

分子为排歧正确的例句数，分母为排歧的总例句数。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.3 实验环境

本文实验环境如表4所示：

操作系统	Linux
CPU	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
GPU	Tesla P100-PCIE-16GB
Python	3.7.11
Pytorch	1.10.0

表 4: 实验环境

4.4 参数设置

在BERT中，隐层维度768维，最大序列长度采用512，batch_size设为4，learning_rate为1e-5，dropout设为0.1，GCN隐层维度为768。

4.5 实验结果与分析

本文分别在CFN和FrameNet数据集上进行了实验；并对GCN的层数、μ和α进行了分析。

首先在CFN数据集上进行了实验，本文设置了如下对比实验：（1）使用BERT模型作为基线模型；（2）BiLSTM框架排歧模型；（3）(Hermann et al., 2014)、(Botschen et al., 2017)、(Hou et al., 2020)和(Guo, 2021)四组对比实验；（4）BERT+GCN（BGCN）模型；（5）使用门机制的BERT+GCN+GATE（BGCNG）模型；（6）使用约束机制的BERT+GCN+CON（BGCNC）模型；（7）使用门机制和约束机制的BERT+GCN+GATE+CON（BGCNGC）模型。实验结果如表5所示。

Model	ACC
(Hermann et al., 2014)	63.48
(Botschen et al., 2017)	67.88
(Hou et al., 2020)	72.52
(Guo, 2021)	74.90
BiLSTM	71.70
BERT	73.98
BGCN	75.37
BGCNG	75.47
BGCNC	75.87
BGCNGC	**75.97**

表 5: CFN实验结果

表5中的实验结果显示，本文提出的方法显著优于之前的方法，相比(Hermann et al., 2014)和(Botschen et al., 2017)的方法分别提升了12.49%和8.09%，相比(Hou et al., 2020)提升了3.45%，上述方法都是通过现有数据集学习到框架表示，计算其与上下文表征的相似度，数据集不均衡，学到的框架表示较差，并且未利用大规模预训练模型，上下文特征抽取能力较弱。(Guo, 2021)的模型使用BERT作为编码器，BiGRU强化上下文语义表示，用全局和局部注意力机制抽取目标词的全局信息和局部信息，BGCN相比(Guo, 2021)提升了0.47%，证明句法信息相比目标词局部信息和全局信息对于框架排歧来说更重要。从BERT和BiLSTM模型的实验结果可知，BERT相比BiLSTM可以更充分的利用上下文信息，具有更强的信息表达能力。BGCNG通过增加门机制，强化相对于目标词重要的信息，降低冗余信息的影响，一定程度上提升了框架排歧的准确率；BGCNC模型加入约束机制来监督模型学习，充分利用目标词的句法信息，相比BGCN，准确率提升了0.5%；BGCNGC同时加入门机制和约束机制，相比基线取得了较大的提升。以上结果证明了本文提出模型对汉语框架排歧的有效性。

(c) 2022 中国中文信息学会计算语言学专业委员会

为了验证模型的有效性和通用性，本文还在英文数据集FrameNet1.5和FrameNet1.7上进行了实验，并与之前的方法进行了对比，实验结果如表6所示。KGFI是 (Su et al., 2021)提出的模型，将框架定义和框架元素通过框架关系构图，利用GCN融入到框架表示当中来进行框架排歧，并采取了框架过滤的规则，本文不考虑框架过滤规则。从表中可知，本文提出的模型在数据集FrameNet1.5和FrameNet1.7上都有所提升，并且在不利用外部知识的情况下在两个数据集上都超过了当前最新工作KGFI，在FrameNet1.5提升了0.33%，在FrameNet1.7提升了0.21%。

Model	FN1.5	FN1.7
(Hermann et al., 2014)	77.49	-
(Botschen et al., 2017)	81.21	-
KGFI	85.63	85.81
BERT	84.91	84.79
BGCN	85.32	85.01
BGCNG	85.63	85.45
BGCNC	85.77	85.85
BGCNGC	**85.96**	**86.02**

表 6: FrameNet实验结果

本文在数据集CFN、FrameNet1.5和FrameNet1.7探讨了GCN层数、μ、α对实验结果的影响。

依据图3可知，对于数据集CFN、FrameNet1.5和FrameNet1.7来说，GCN的层数为2时，取得的结果最好。经分析，当GCN层数为1时，目标词只聚合了自身邻居节点的信息，只利用了一阶依存信息，而对于目标词来说，有的二阶依存信息对其框架的选择也是必不可少的，相比之下两层GCN不仅聚合了自身的邻居信息，同时也融入了邻居的邻居信息，充分利用了二阶依存信息。但随着GCN层数的增加，在三个数据集上都表现出了下降的趋势，分析主要是因为GCN每次卷积都是节点和周围信息聚合的过程，随着GCN层数的增加，词节点聚合的信息越来越多，到最后会使得每个节点的嵌入表示变得非常相近，而框架排歧是针对句子中的某个词来排歧，并非在句子层面，给框架排歧带来了噪声干扰。

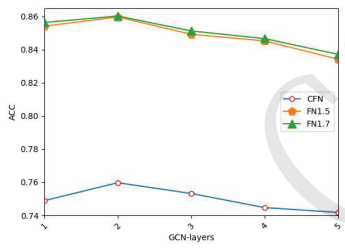


图 3: GCN层数的影响

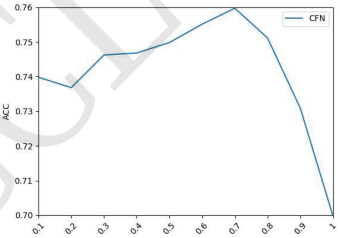


图 4: CFN上μ的影响

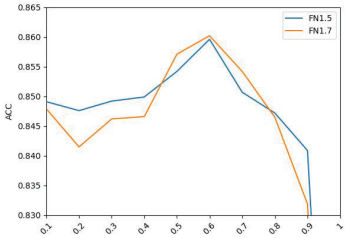


图 5: FrameNet上μ的影响

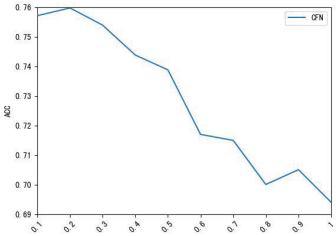


图 6: CFN上α的影响

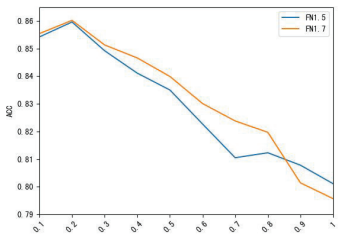


图 7: FrameNet上α的影响

(c) 2022 中国中文信息学会计算语言学专业委员会

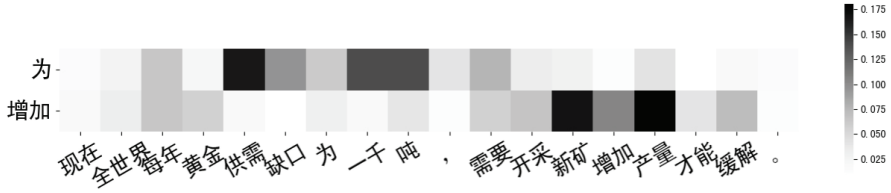


图 8: 门的可视化

本文设置了线性插值层来插值BERT和GCN的输出，插值参数μ控制BERT和GCN的权重，为了探究μ对整体模型的影响，在数据集CFN、FrameNet1.5和FrameNet1.7上对μ进行不同设置，结果如图4和图5所示。图4为10个epoch下CFN数据集上的μ参数分析，从图中可知，随着μ的增加，准确率越来越高，当μ为0.7时，模型效果达到最佳，表现效果好于仅使用BERT预测（$\mu = 0$）和GCN预测（$\mu = 1$）；$\mu > 0.7$时准确率开始下降，模型收敛相对较慢；当$\mu = 1$时，GCN后直接计算损失、梯度下降，会出现梯度消失的问题，模型收敛速度变慢。图5为15个epoch下FrameNet数据集上的μ参数分析，从图中可知，在FrameNet数据集上，当μ为0.6时，模型效果达到最佳。从实验数据中可知，不同的数据集，μ的最佳值不同。

图6、图7为α不同设置下的实验结果，从图中可知，当α取0.2时，准确率达到最优。当α取0.1到0.3时相对比较平稳，随着α的增大，准确率下降。

4.6 案例分析

本小节从框架排歧数据中选取了一条数据进行分析，通过对门权重进行可视化验证加入门机制的有效性。在标准化之后，绘制每个单词的值如图8所示。通过分析可知，目标词为“为”时，“供需”、“缺口”、“一千”、“吨”的权重较大，当目标词为“增加”时，“新矿”、“增加”、“产量”权重较大，由此可见门机制的加入，强化了对目标词重要的信息，降低了噪声信息的干扰。

5 总结

本文针对汉语进行了框架排歧研究，提出了一种基于GCN和门机制的框架排歧模型，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，通过GCN将句法信息融入目标词表示中，使用门机制过滤与目标词无关的噪声信息，并提出一种基于依存图的约束机制来监督模型学习，改进向量表示。目前汉语框架排歧还存在一定的挑战，如现有框架排歧数据规模有限，并且存在数据不平衡问题，对于出现频率低的框架很难通过有监督的模型训练得到好的性能，如何更好的解决数据不均衡性，可否将元学习 (Kumar et al., 2019; Holla et al., 2020; Du et al., 2021; Chen et al., 2021)方法应用到框架排歧上，以及模型对领域外的数据是否同样具有通用性还有待研究与解决。

参考文献

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.

Baker, Collin F and Ellsworth, Michael and Erk, Katrin. 2007. *SemEval-2007 Task 19: Frame semantic structure extraction.* Proceedings of the 4th International Workshop on Semantic Evaluations, pages:99-104.

Botschen, Teresa and Mousselly-Sergieh, Hatem and Gurevych, Iryna. 2017. *Prediction of frame-to-frame relations in the FrameNet hierarchy with frame embeddings.* Proceedings of the 2nd Workshop on Representation Learning for NLP, pages:146-156.

Che, Wanxiang and Li, Zhenghua and Liu, Ting. 2010. *Ltp: A chinese language technology platform.* Coling 2010: Demonstrations, pages:13-16. Beijing, China.

Chen, Howard and Xia, Mengzhou and Chen, Danqi. 2021. *Non-parametric few-shot learning for word sense disambiguation.* arXiv preprint arXiv:2104.12677.

(c) 2022 中国中文信息学会计算语言学专业委员会

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, pages:4171-4186.

Du, Yingjun and Holla, Nithin and Zhen, Xiantong and Snoek, Cees GM and Shutova, Ekaterina. 2021. *Meta-Learning with Variational Semantic Memory for Word Sense Disambiguation.* arXiv preprint arXiv:2106.02960.

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.

郭哲铭. 基于注意力机制的框架识别技术研究*[D]*. 山西大学, 2021.

Hermann, Karl Moritz and Das, Dipanjan and Weston, Jason and Ganchev, Kuzman. 2014. *Semantic frame identification with distributed word representations.* Meeting of the Association for Computational Linguistics, pages:1448-1458. Baltimore, USA.

侯运瑶, 曹学飞, 崔军, 王瑞波, 李济洪, 李茹. 2020. 基于框架表示学习的汉语框架排歧. 计算机应用研究, 37(12):5.

Holla, Nithin and Mishra, Pushkar and Yannakoudakis, Helen and Shutova, Ekaterina. 2020. *Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation.* arXiv preprint arXiv:2004.14355.

Kipf, Thomas N and Welling, Max. 2016. *Semi-supervised classification with graph convolutional network.* arXiv preprint arXiv:1609.02907.

Kumar, Sawan and Jat, Sharmistha and Saxena, Karan and Talukdar, Partha. 2019. *Zero-shot word sense disambiguation using sense definition embeddings.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistic，pages:5670-5681.

李茹. 汉语句子框架语义结构分析技术研究*[D]*. 山西大学, 2012.

Li, Qimai and Han, Zhichao and Wu, Xiao-Ming. 2018. *Deeper insights into graph convolutional networks for semi-supervised learning.* Thirty-Second AAAI conference on artificial intelligence.

李济洪, 高亚慧, 王瑞波, 李国臣. 2011. 汉语框架自动识别中的歧义消解. 中文信息学报, 25(03):38-44.

李国臣, 张立凡, 李茹, 刘海静, 石佼. 2013. 基于词元语义特征的汉语框架排歧研究. 中文信息学报, 27(4):44-52.

Li, Ru and Liu, Haijing and Li, Shuanghong. 2010. *Chinese frame identification using t-crf model.* Coling 2010: Posters, pages:674-682.

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.

Su, Xuefeng and Li, Ru and Li, Xiaoli and Pan, Jeff Z and Zhang, Hu and Chai, Qinghua and Han, Xiaoqi. 2021. *A Knowledge-Guided Framework for Frame Identification.* Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages:5230-5240.

王智强, 李茹, 阴志洲, 刘海静, 李双红. 2013. 基于依存特征的汉语框架语义角色自动标注. 中文信息学报, 27(2):34-41.

Wang, Xiaohui and Li, Ru and Wang, Zhiqiang and Chai, Qinghua and Han, Xiaoqi. 2020. 基于*Self-Attention* 的句法感知汉语框架语义角色标注. Proceedings of the 19th Chinese National Conference on Computational Linguistics, pages:616-623.

You, Liping and Liu, Kaiying. 2005. *Building chinese framenet database.* Natural Language Processing and Knowledge Engineering, pages:301-306.

赵红燕, 李茹, 张晟, 张力文. 2016. 基于*DNN*的汉语框架识别研究. 中文信息学报, 30(6):75-83.

张力文, 王瑞波, 李茹, 张晟. 2017. 基于词分布式表征的汉语框架排歧模型. 中文信息学报, 31(6):50-57.

Zhang, Yuhao and Qi, Peng and Manning, Christopher D. 2018. *Graph convolution over pruned dependency trees improves relation extractio.* arXiv preprint arXiv:1809.10185.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于中文电子病历知识图谱的实体对齐研究

李丽双*
大连理工大学
计算机科学与技术学院
辽宁，大连
lilishaung314@163.com

董姜媛
大连理工大学
计算机科学与技术学院
辽宁，大连
donjyuan@163.com

摘要

医疗知识图谱中知识重叠和互补的现象普遍存在，利用实体对齐进行医疗知识图谱融合成为迫切需要。然而据我们调研，目前医疗领域中的实体对齐尚没有一个完整的处理方案。因此本文提出了一个规范的基于中文电子病历的医疗知识图谱实体对齐流程，为医疗领域的实体对齐提供了一种可行的方案。同时针对基于中文电子病历医疗知识图谱之间结构异构性的特点，设计了一个双视角并行图神经网络(DuPNet)模型用于解决医疗领域实体对齐，并取得较好的效果。

关键词： 医疗知识图谱 ； 中文电子病历 ； 实体对齐 ； 结构异构性 ； 并行图神经网络

Research on Entity Alignment Based on Knowledge Graph of Chinese Electronic Medical Record

Lishuang Li*
School of Computer Science
and Technology
Dalian University of Technology
Dalian, China
lilishaung314@163.com

Jiangyuan Dong
School of Computer Science
and Technology
Dalian University of Technology
Dalian, China
donjyuan@163.com

Abstract

The phenomenon of knowledge overlap and complementarity is common in different medical knowledge graphs. It is urgent to use entity alignment to fuse the medical knowledge graphs. However, according to our research, there is not yet a complete solution for entity alignment in the medical field. Therefore, we propose a standardized entity alignment process based on the Chinese electronic medical record knowledge graph, which provides a feasible scheme for entity alignment in the medical field. Meanwhile, according to the characteristic of the structural heterogeneity of the medical knowledge graph, we design a Dual-view Parallel Graph Neural Network (DuPNet) to solve the problem of entity alignment in the medical field, which achieves good results.

Keywords: Medical knowledge graph , Chinese electronic medical record , Entity alignment , Structure heterogeneity , Parallel graph neural network

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：大连市科技创新基金(2020JJ26GX035)

1 引言

电子病历是信息化医疗健康服务的产物之一，它包含着大量的医学事实。随着国内电子病历的积累，利用自然语言处理技术从电子病历中自动化获取、整合医疗信息具有重要意义。构建电子病历相关的知识图谱是最有效的展示和利用电子病历中医疗信息的方法之一。然而，随着中文医疗知识图谱的广泛构建(奥德玛等, 2019; Xiu, 2020)，不同知识图谱之间存在着知识重叠和互补的现象，这就需要利用知识图谱融合来整合分散在各个知识图谱中的医疗知识，通过知识融合技术建立一个大规模的医疗知识图谱，可以为辅助决策、智能问答等下游应用提供技术支持(刘道文等, 2021; 刘勘和张雅荃, 2020)，从而促进智能医疗的发展。

知识融合中最关键的技术是实体对齐，其目的是判别不同知识图谱中的实体是否指向现实世界中的同一对象。据我们调研，目前中文医疗领域实体对齐的相关研究较少，大多数研究首先通过计算实体名称相似度生成候选实体对，然后再进一步利用结构、属性等信息判断候选实体对之间的相似性，这种方法虽然可以通过候选实体对降低模型复杂度，然而难以保证候选实体对的质量。在通用领域，实体对齐早期也是主要采用基于相似性度量的方法(Bhattacharya和Getoor, 2007; Jiang等, 2014)。随着知识图谱表示学习的兴起，越来越多的研究人员使用知识图谱表示学习解决实体对齐问题，最经典的是基于翻译的模型(Song等, 2021; Lu等, 2021),它们利用TransE(Bordes等, 2013)对三元组编码实现实体对齐。近年来，随着图神经网络的发展，一些研究使用图神经网络建模知识图谱的结构，用实体的邻域信息增强实体嵌入，即利用图卷积递归聚合邻居的嵌入表示来学习中心实体表示，通过计算实体间的嵌入距离实现实体对齐。

基于图神经网络的方法充分地利用了知识图谱的结构信息，提高了实体对齐模型的性能。然而，由于知识来源和构建目的不同，知识图谱之间存在着结构异构性，给此类方法带来了挑战。比如，由不同医院相同科室电子病历构建的两个知识图谱KG1和KG2存在的医疗知识的重叠与互补，造成了它们之间的结构异构性。Li等 (2019)利用知识嵌入和交叉图模型联合的半监督方法缓解结构异构性。Sun等 (2020)用实体邻域信息增强实体嵌入，并且使用图注意力机制为实体的每个邻居学习注意力分数来缓解结构异构性。Chen等 (2021)利用潜在的空间邻域聚合来处理结构异构性。然而这些研究过程仅考虑了结构异构性中的实体邻域异构性，如中心实体“艾滋病”与“AIDS”仅有实体“发烧”这一共同邻居，其余邻居均不同，该方法会为共同邻居“发烧”学习一个较高的权重。此类方法忽略了关系异构性对结构异构性的重要影响。事实上，来源不同的知识图谱往往具有关系独立性，例如，存在于KG1中的某一关系并不一定存在于KG2，导致了知识图谱之间的关系异构性，这是造成知识图谱结构异构性的重要原因。此外，现有的研究(Li等, 2019; Cao等, 2019)认为多层图卷积网络的输出层表示集成了实体的多跳邻域信息，因此他们将网络的输出层表示视为实体的嵌入表示。然而，我们发现随着卷积层数的增加，中心实体聚集的邻域信息呈指数级增长，因此给实体的表示带来了大量的噪声。

针对以上问题，本文设计了一个双视角并行图神经网络模型(DuPNet)用于中文电子病历的医疗知识图谱实体对齐。模型分别利用实体交互和关系交互缓解实体邻域异构性和关系异构性，以协同缓解医疗知识图谱的结构异构性。我们利用一个简洁有效的门控机制聚合网络层之间的输出，使得模型在捕获实体多跳邻域信息的同时，缓解由多层卷积引起的噪声问题。

此外，在医疗领域中，目前实体对齐相关研究相对较少，因此医疗领域中的实体对齐尚没有一个完整的处理流程，本文采用上述模型进行实体对齐，同时，针对医疗知识图谱的特点提出了一个规范的基于中文电子病历的医疗知识图谱实体对齐流程。主要贡献如下：

(1)提出了一个规范的基于中文电子病历的医疗知识图谱实体对齐流程，为医疗领域的实体对齐提供了一种可行的方案。

(2)针对基于中文电子病历医疗知识图谱之间结构异构性的特点，设计了一个双视角并行图神经网络(DuPNet)模型用于解决医疗领域实体对齐，并取得较好的效果。

2 相关工作

2.1 通用领域的实体对齐方法

在通用领域，JETEA(Song等, 2021)采用基于翻译的方法并且将实体类型匹配作为约束条件，使用一种迭代的方式将新检测到的对齐实体添加到训练数据中，以促进实体对齐。JTMEA(Lu等, 2021)也采用了基于翻译的方法，引入了一种具有属性增强的知识嵌入模型。然而基于翻译的方法无法充分利用知识图谱的结构信息。随着图神经网络的发展，越来越

(c) 2022 中国中文信息学会计算语言学专业委员会

多研究者使用基于图神经网络的方法解决实体对齐问题。GCN-Align(Wang等, 2018)首次尝试使用图神经网络进行实体对齐，将跨语言的实体嵌入到一个统一的向量空间中，并且将结构嵌入和属性嵌入相结合，以获得精确的对齐。KECG(Li等, 2019)提出一种基于联合知识嵌入模型和交叉图模型的半监督实体对齐方法，更好地利用种子对齐在整个图上传播。MUGNN(Cao等, 2019)提出了一种多通道的图神经网络框架处理实体对齐问题。AliNet(Sun等, 2020)通过使用门控策略和注意机制聚合多跳邻域，缓解实体邻域异构性。LatsEA(Chen等, 2021)利用潜在的空间邻域聚合来处理实体邻域异构性，并将实体对齐作为最大二部图匹配问题，采用匈牙利算法进行求解。AliNet和LatsEA在聚合邻域信息时认为实体的一跳邻居都同样重要。然而，并不是所有的一跳邻居都对中心实体有积极的贡献。上述基于图神经网络的模型虽然考虑了实体邻域异构性，但它们忽略了知识图谱之间关系异构性对结构异构性的重要影响。

2.2 中文医疗领域的实体对齐方法

目前中文医疗领域实体对齐的相关研究较少。宋文欣 (2018)分别用无监督和有监督的方法对医疗知识库进行实体对齐，首先计算实体指称项相似度生成候选实体对，然后在候选实体对之间得到最终的对齐实体对。蔡娇 (2020)采用基于网络语义标签的实体对齐算法用于遗传病领域的数据库，首先计算疾病名称相似度以生成候选实体对，然后用候选实体对计算多标签综合相似度，根据综合相似度判断实体对齐。这种方法虽然可以通过候选实体对降低模型复杂度，然而难以保证候选实体对的质量。

3 方法

3.1 双视角并行图神经网络实体对齐模型

为解决基于电子病历医疗知识图谱的结构异构性问题，本文设计与搭建了一个双视角并行图神经网络(DuPNet)实体对齐模型。

3.1.1 问题定义

本文将医疗知识图谱定义为$G = (E, R, T)$，其中E代表实体集，R代表关系集，T代表三元组集合，$e \in E$, $r \in R$, $t \in T$ 分别代表任一实体、关系、三元组。假设存在两个异构的医疗知识图谱G和$G' = (E', R', T')$，实体对齐最终目的是找出所有E和E'中指向同一对象的实体对。另外，$\mathbf{E}$和$\mathbf{E}'$分别代表实体特征矩阵，$\mathbf{R}$和$\mathbf{R}'$分别代表关系特征矩阵，均通过随机初始化的方式得到。

3.1.2 DuPNet模型架构

DuPNet从实体交互和关系交互的视角协同缓解医疗知识图谱的结构异构性。模型框架如图1所示。其中(1)和(2)代表由关系相似度矩阵得到的关系匹配度向量。从实体交互的视角来看，使用自注意力机制聚合实体的邻域信息，以缓解实体邻域异构性。从关系交互视角来看，由关系嵌入交互得到关系相似度矩阵，再由关系相似度矩阵得出关系匹配度作为跨图注意力分数聚合邻域信息，以缓解关系异构性。为得到更精确的实体表示，DuPNet利用门控机制聚合隐藏层和输出层的嵌入表示，从而缓解多层卷积引起的噪声问题。

3.1.3 实体交互视角

在实体交互视角中，通过自注意力机制迭代地为实体的每个邻居学习精确的自注意力分数，通过在训练的过程中，对重要的邻居赋予较高的权重，来缓解实体邻域的异构性。对于医疗知识图谱G中的任一实体e_i,自注意力分数由实体e_i和它的邻居实体的嵌入表示计算得到。自注意力分数$attn_{ij}^{e}$的计算公式如下:

$$attn_{ij}^{e} = \frac{exp(c_{ij}^{e})}{\sum\limits_{e_k \in \mathcal{N}_1(e_i) \cup \{e_i\}} exp(c_{ik}^{e})}. \tag{1}$$

$$c_{ij}^{e} = \sigma(\mathbf{q}[\mathbf{W}_1\mathbf{e}_i \| \mathbf{W}_2\mathbf{e}_j]). \tag{2}$$

其中c_{ij}^{e}是自注意力系数，代表实体e_j对e_i的重要程度。$e_k \in \mathcal{N}_1(e_i) \cup \{e_i\}$代表实体$e_i$包括自身在内的邻居，$\|$代表向量拼接，$\sigma(\cdot)$ 是激活函数，选择为$LeakyReLU(\cdot)$。$\mathbf{W}_1$,$\mathbf{W}_2$和$\mathbf{q}$是可训练参数。

第二十一届中国计算语言学大会论文集，第211页-第221页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

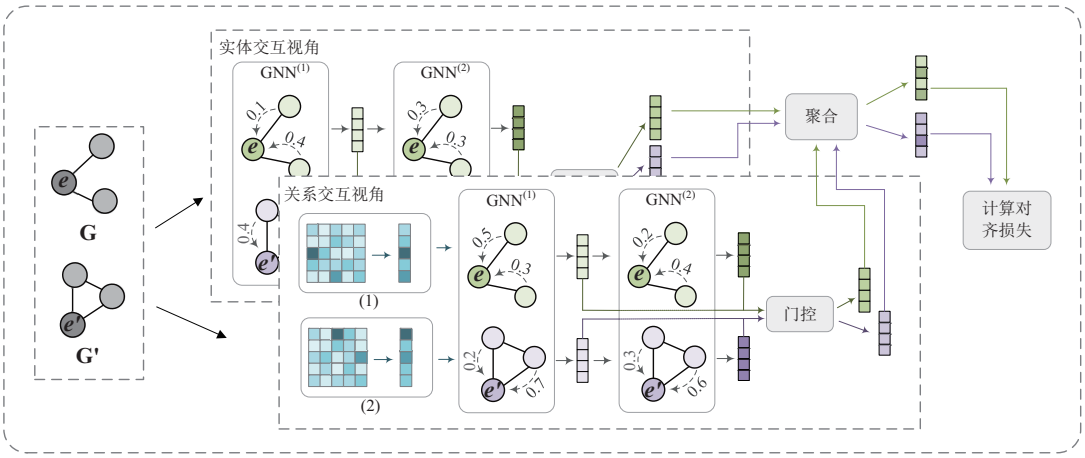


Figure 1: 双视角并行图神经网络实体对齐模型

在图神经网络中，节点的表示是通过递归聚合其邻居的特征向量来学习的。本文利用由公式(1)计算得到的自注意力分数$attn_{ij}^{e}$ 聚合邻居特征向量，得到实体e_i在实体交互视角中第l层的表示$\mathbf{h}_{i,e}^{(l)}$，计算公式如下：

$$\mathbf{h}_{i,e}^{(l)} = \sigma(\sum_{e_j \in \mathcal{N}_1(e_i) \cup \{e_i\}} attn_{ij}^{e} \mathbf{W}_3^{(l)} \mathbf{h}_{j,e}^{(l-1)}). \tag{3}$$

其中，$\mathbf{W}_3^{(l)}$是该视角网络中第l层的权重，$\sigma(\cdot)$为激活函数，选择为$ReLU(\cdot)$。

3.1.4 关系交互视角

在关系交互视角中，使医疗知识图谱G和G'的关系特征矩阵相互作用，得到关系相似度矩阵，然后进行最大池化操作，得到关系匹配向量**Match**。最后利用从关系匹配向量中得到的跨图匹配分数来聚合来自邻居的信息，关系匹配向量的计算公式为：

$$\mathbf{Match} = f_{max}(f_{sim}(\mathbf{R}, \mathbf{R}')). \tag{4}$$

其中，$f_{sim}(\cdot)$代表关系相似度计算函数，定义为$f_{sim}(\mathbf{R}, \mathbf{R}') = \mathbf{R}^T\mathbf{R}'$，$\mathbf{R}$和$\mathbf{R}'$分别代表待对齐的两个医疗知识图谱的关系特征矩阵，为可训练的参数。$f_{max}(\cdot)$代表最大池化操作函数。由关系匹配向量**Match**计算得到跨图匹配分数$attn_{ij}^{r}$，公式如下：

$$attn_{ij}^{r} = \frac{exp(c_{ij}^{r})}{\sum_{e_k \in \mathcal{N}_1(e_i) \cup \{e_i\}} exp(c_{ik}^{r})}. \tag{5}$$

$$c_{ij}^{r} = \underset{(e_i, r_{ij}, e_j) \in T}{\mathbf{Match}}[r_{ij}]. \tag{6}$$

其中**Match**$[\cdot]$代表关系匹配度索引操作。T代表知识图谱的三元组集合。利用跨图匹配分数计算实体在关系交互视角中第l层的表示$\mathbf{h}_{i,r}^{(l)}$，计算公式为：

$$\mathbf{h}_{i,r}^{(l)} = \sigma(\sum_{e_j \in \mathcal{N}_1(e_i) \cup \{e_i\}} attn_{ij}^{r} \mathbf{W}_4^{(l)} \mathbf{h}_{j,r}^{(l-1)}). \tag{7}$$

其中$\mathbf{W}_4^{(l)}$是该视角网络中第l层的权重，$\sigma(\cdot)$为激活函数，选择为$ReLU(\cdot)$。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.1.5 门控聚合

为了获得更准确的实体表示，利用门控机制来聚合网络中隐藏层和输出层的嵌入表示，将其应用于上述两个视角。门控机制在捕获实体的多跳邻域信息增强实体的嵌入的同时去除各层的冗余噪声，从而缓解多层卷积引起的噪声问题。

门控机制的实现细节如下：

$$Gate_l(\mathbf{input}_1, ..., \mathbf{input}_l) = \begin{cases} \mathbf{g}_l \cdot \mathbf{input}_{l-1} + (1 - \mathbf{g}_l) \cdot \mathbf{input}_l, & l = 2 \\ \mathbf{g}_l \cdot Gate_{l-1} + (1 - \mathbf{g}_l) \cdot \mathbf{input}_l, & l > 2 \end{cases} \tag{8}$$

其中，l代表网络层数，$\mathbf{input}_l$代表网络第l层输出，$\mathbf{g}_l$为一组可训练的参数。

以实体交互视角为例，任一实体e_i该视角下嵌入表示为$\mathbf{h}_{i,e}$，公式如下：

$$\mathbf{h}_{i,e} = Gate_l(\mathbf{h}_{i,e}^{(1)}, ..., \mathbf{h}_{i,e}^{(l)}). \tag{9}$$

其中，$\mathbf{h}_{i,e}^{(l)}$为网络第l层的输出表示。同理，任一实体e_i该关系视角下嵌入表示为$\mathbf{h}_{i,r}$，公式如下：

$$\mathbf{h}_{i,r} = Gate_l(\mathbf{h}_{i,r}^{(1)}, ..., \mathbf{h}_{i,r}^{(l)}). \tag{10}$$

实体e_i的最终嵌入表示$\mathbf{h}_i$由门控机制聚合两个视角的输出得到，具体计算公式如下：

$$\mathbf{h}_i = \mathbf{g}_a \cdot \mathbf{h}_{i,e} + (1 - \mathbf{g}_a) \cdot \mathbf{h}_{i,r}. \tag{11}$$

其中$\mathbf{g}_a$为一组可训练的参数，用来控制两个视角的聚合。

3.1.6 对齐损失函数

对齐损失函数由两部分构成，分别是实体对齐损失和三元组对齐损失。其中实体对齐损失函数如下：

$$L_{ent} = \sum_{(e,e') \in A_e^+} \sum_{(e_-,e'_-) \in A_e^-} max\{0, \gamma_1 + dis(\mathbf{e} - \mathbf{e}') - dis(\mathbf{e}_- - \mathbf{e}'_-)\}. \tag{12}$$

其中A_e^+代表实体对齐对正例集合，A_e^-代表实体对齐对负例集合，γ_1为边际超参数，$dis(\cdot)$代表L_2范数，用于计算实体间的距离。

此外，我们引入三元组损失建模实体和关系之间的联系，并将三元组损失函数定义如下：

$$L_{tri} = \sum_{(h,r,t) \in T} \sum_{(h_-,r_-,t_-) \in T_-} max\{0, \gamma_2 + dis(\mathbf{h} + \mathbf{r} - \mathbf{t}) - dis(\mathbf{h}_- + \mathbf{r}_- - \mathbf{t}_-)\}. \tag{13}$$

其中，T代表三元组正例集合，T_-代表三元组负例集合。

综上所述，DuPNet最终的损失函数如下：

$$L = L_{ent} + L_{tri}. \tag{14}$$

3.2 基于中文电子病历医疗知识图谱的实体对齐流程

医疗知识图谱融合的目的是通过整合各个医疗知识图谱中分散的知识来构建一个更加精确和完善的医疗知识库，实体对齐是其中最关键的一步。针对医疗实体对齐中的实际应用，本文提出了一个规范的基于中文电子病历医疗知识图谱的实体对齐流程，如图2所示。首先，由于中文电子病历中知识纷繁复杂，同一医学术语知识图谱中可能存在多个不标准的实体表述。针对这一问题，我们首先构建医学词根库对单个医疗知识图谱进行实体规范化。其次，对医疗知识图谱进行推理(Lan等, 2021)能够补全缺失的知识，基于电子病历的单个医疗知识图谱中的知识往往是不完整的，所以提出利用规则挖掘进行知识推理。经过上述处理，单个医疗知识图谱的知识精度和完整性得到了提升，同时也为后续的实体对齐提供了良好的基础。最后构建医疗实体种子对，用训练集训练上述模型DuPNet的网络参数，以实现医疗知识图谱的实体对齐。下面将以两个由不同医院相同科室的电子病历构建得到的医疗知识图谱为例进行阐述，多个医疗知识图谱对齐即以两个对齐为基础进行迭代处理。

CCL 2022

Figure 2: 基于电子病历医疗知识图谱的实体对齐流程

3.2.1 医疗知识图谱实体规范化

电子病历不同于医学书籍和文献，医生记录电子病历的习惯因人而异，导致在知识图谱中对于同一医学术语可能会有多个不同的医学实体表达，例如，对于医学术语“支气管炎”，医学实体“支气管炎”和“支气管炎症”可能同时存在于医疗知识图谱中。本文将这种具有相同词根的不同实体表达同一医学术语的情况称为“多词一义”问题，它使得知识图谱极度冗余。我们首先对每个医疗知识图谱进行实体规范化操作，提高知识图谱中实体的准确度，为后续实体对齐奠定良好的基础。

(1)医学词根库构建

在医疗领域中，医学词根可以代表医学实体中一个有意义的子串，且能够反应该医学实体的重要特征。由于医学术语的多个实体表达中大多包含相同词根，如上述例子中所示，“支气管炎”和“支气管炎症”中都含有相同词根“支气管”。 因此可以通过构造医学词根库推荐得到“多词

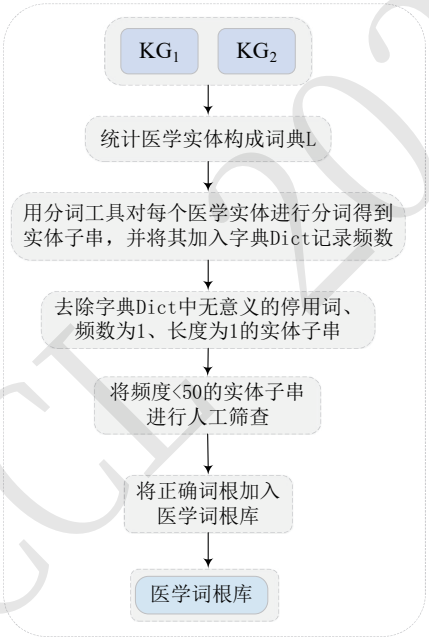


Figure 3: 医学词根库构建流程

一义”候选表，再由医学生对该表中“多词一义”实体进行标注并规范实体名称。医学词根库的构建方法如图3所示。

首先统计两个医疗知识图谱中的医学实体并构成词典L，然后利用北京大学开发的分词工具包pkuseg的医药领域对每个医学实体进行分词得到实体子串，并将其加入字典Dict，记录每个子串出现的频数。之后去除字典Dict中的无意义的停用词、频数为1以及长度为1的子串，因为这些子串对医学实体并没有很好的表征能力。例如子串“其他”、“任何”、“上”等。除此之外，子串频数过高可能会降低对医学实体的表征能力。例如子串“皮肤”出现的频数为743，包含该子串的实体“非黑色素瘤皮肤癌”、“亚急性皮肤型红斑狼疮”和“皮肤幼虫移行症”均不代表同一医学术语。因此我们将字典中频数小于50的子串加入医学词根库，在此之前为保证词根质量，我们先对其进行人工筛查。部分词根如图4所示。

(2)实体规范化

第二十一届中国计算语言学大会论文集，第211页-第221页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

胆固醇	细菌性	动脉硬化	杂音	颈动脉	狼疮性	霉菌	纤维瘤
血斑	粉碎性	并发症	腓肠	流产	结痂	胃肠道	化脓性
刺激	虹膜	心肌酶	湿疹	扭伤	肱骨	坏死性	回盲部
胃窦部	硬化	硫唑	肋骨	遗传性	尿路	硬化	斑丘
紫癜	粥样	体重	胰岛	阻滞	潮红	腋下	恶心
病理学	脑病	剧痛	挛缩	肠梗阻	贲门	继发	染色体

Figure 4: 医学词根示例

规范实体	**实体1**	**实体2**	**实体3**
不稳定型心绞痛	不稳定型心绞痛	不稳定性心绞痛	不稳定心绞痛
支气管炎	支气管炎	支气管炎症	
糖尿病肾病	糖尿病肾病	糖尿病性肾病	
双侧筛窦炎	双侧筛窦炎	双侧筛窦炎症	
原发性恶性肿瘤	原发性恶性肿瘤	原发恶性肿瘤	
急性胆囊炎	急性胆囊炎	胆囊急性炎症	
骨质疏松症	骨质疏松	骨质疏松症	
狼疮性肾炎	狼疮性肾炎	狼疮肾炎	
弓形虫病	弓形虫	弓形虫病	

Figure 5: “多词一义”规范表示例

根据上节得到的医学词根库，利用字符串索引算法推荐得到每个词根的“多词一义”候选表，再由医学生标注出其中正确的“多词一义”实体，并规范每一组“多词一义”实体的名称，由此得到“多词一义”规范表，该表的部分内容如图5所示。在知识图谱中，每一组“多词一义”实体被合并成同一规范实体，与“多词一义”实体相关的三元组中的实体也被其规范实体替代。

3.2.2 规则挖掘补全医疗知识图谱

现有医疗知识图谱通常由人工或半自动的方式构建，普遍存在不完备的问题。本文通过挖掘医疗知识图谱中潜在的规则来填补实体间缺失的关系从而达到补全的目的。首先，专家从现有的两个基于中文电子病历的异构医疗知识图谱中归纳出潜在的规则。之后，在每个医疗知识图谱中进行规则匹配，得到推理出的三元组。最后，为保证规则推理得到的三元组的质量，需要人工对推理出的三元组进行筛选。

(1)规则归纳

由于医学知识图谱存在精度要求高且复杂度高等特点，为保证补全三元组的正确性，我们请专家为每个医疗知识图谱归纳出规则集B。具体规则由前提三元组和结论三元组组成，其中，结论三元组可以由一系列的前提三元组推理得出。例如，[治疗改善疾病(x,y)]∧[疾病显示症状(y,z)]⇒[对症治疗(x,z)]。

(2) 规则落地

将由规则归纳得到的规则集合B应用于医疗知识图谱，给定一条规则$\beta \in B$，查找该知识图谱中满足该条规则的所有前提三元组，并依据规则推理出结论三元组，若结论三元组不存在于原来的知识图谱中则添加至原有知识图谱，即完整了一次三元组的补全操作。例如根据上述规则得到：[治疗改善疾病(泼尼松,肾病综合症)]∧[疾病显示症状('肾病综合症,蛋白尿)]⇒[对症治疗(泼尼松,蛋白尿)]。

(3)人工筛选

对于由规则落地推理出的结论三元组，虽然能够确保逻辑上的正确性，然而，有些医疗知识非常复杂，结论三元组仍然可能存在错误的情况，为进一步保证补全的结论三元组的准确性，专业的医学生对补全的结论三元组所表达的医疗知识进行确认，筛选出正确的结论三元组，将其补充到原医疗知识图谱中。

3.2.3 DuPNet用于电子病历医疗知识图谱实体对齐

(1)构建医疗实体种子对流程

基于图神经网络的实体对齐模型需要已知的实体种子对作为训练集和测试集，使得模型为待对齐实体学习到相近的嵌入表示。在医疗领域中，实体种子对主要由医学实体与其别名、简称等组成。例如，疾病实体“AIDS”与“艾滋病”为一组种子对。

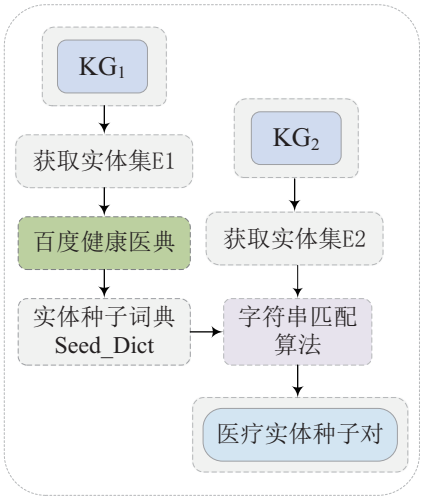


Figure 6: 医疗实体种子对构建流程

然而电子病历的内容中仅含有病人诊断治疗全过程的原始记录，并不直接保存有与疾病、治疗等实体相关的别名、简称等信息，因此给基于电子病历知识图谱的医疗实体种子对构建带来了困难。针对这一问题，本文提出的实体种子对标注流程如图6所示。

对于两个医疗知识图谱KG1和KG2，首先，获取KG1的实体集E1。然后从权威的健康知识科普平台“百度健康医典”利用网络爬虫技术为E1中的实体获得别名、简称等信息，构成实体种子词典Seed_Dict。最后将实体种子词典Seed_Dict与KG2实体集E2进行字符串匹配得到实体种子对。

(2)电子病历医疗知识图谱实体对齐

利用上述流程得到的医疗实体种子对作为训练集和测试集，在训练过程中通过最小化损失函数使得训练集中实体种子对的嵌入距离逐渐相近。训练完成后的模型具备了识别对齐实体对的能力，可以为待对齐实体学习到相近的嵌入表示，实现实体对齐。

4 实验

4.1 基于电子病历的医疗知识图谱数据详情

基于电子病历的医疗知识图谱KG_1和KG_2是对不同医院相同科室的电子病历进行三元组抽取得到的，其详细数据如表1中原始数据所示。经过构建“多词一义”规范表进行实体规范化后，KG_1和KG_2的实体数量分别减少494个和510个，修正后的医疗知识图谱为SKG_1和SKG_2。然后，在SKG_1和SKG_2基础上进行规则挖掘补全知识图谱，经由规则推理、人工筛选后得出的新三元组数量分别为11,639个和11,803个，补全后的医疗知识图谱为$CSKG_1$和$CSKG_2$。

	知识图谱	实体数量	关系数量	三元组数量
原始数据	KG_1	19,540	13	112,902
	KG_2	19,727	13	111,425
实体规范化	SKG_1	19,046	13	112,902
	SKG_2	19,217	13	111,425
规则挖掘补全	$CSKG_1$	19,046	13	124,541
	$CSKG_2$	19,217	13	123,228

Table 1: 基于电子病历的医疗知识图谱数据详情

(c) 2022 中国中文信息学会计算语言学专业委员会

	ZH_EN			JA_EN			FR_EN		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
JETEA(2021)	42.7	75.0	-	36.4	72.4	-	36.5	71.8	-
JTMEA(2021)	42.2	75.9	53.0	38.7	72.4	50.0	37.1	75.6	46.0
GCN-Align(2018)	41.3	74.4	54.9	39.9	74.5	54.6	37.3	74.5	53.2
KECG(2019)	47.8	83.5	59.8	49.0	84.4	61.0	48.6	85.1	61.0
MuGNN(2019)	49.4	**84.4**	61.1	50.1	85.7	62.1	49.5	87.0	62.1
LatsEA(2021)	52.2	76.3	61.3	53.9	77.2	62.5	53.8	78.7	63.2
Alinet(2020)	**53.9**	82.6	62.8	**54.9**	83.1	64.5	55.2	85.5	65.7
DuPNet	52.8	83.9	**63.2**	53.7	**85.8**	**64.8**	**55.7**	**87.7**	**66.8**

Table 2: DuPNet在标准数据集上的实验结果

4.2 实验结果与分析

4.2.1 评价指标

遵循前人工作(Chen等, 2017; Cao等, 2019; Sun等, 2020)，本文采用Hits@k和MRR作为模型的评价指标。其中，Hits@k代表前k个候选实体中正确对齐实体的百分比，MRR代表正确对齐实体排名倒数的平均值。

4.2.2 DuPNet处理结构异构性的能力测试

DuPNet旨在解决结构异构性这一重要问题，为验证DuPNet缓解结构异构性的能力，本文将DuPNet在标准数据集DBP15K(Sun等, 2017)上与解决结构异构性的模型进行对比，该类模型均未使用预训练模型。其中GCN-Align(Wang等, 2018)是利用图神经网络解决实体对齐的首次尝试，KECG(Li等, 2019)、MUGNN(Cao等, 2019)、LatsEA(Chen等, 2021)和Alinet(Sun等, 2020)均是解决结构异构性的经典模型。另外将DuPNet与最新的基于翻译的模型JETEA(Song等, 2021)和JTMEA(Lu等, 2021)进行对比。实验结果如表2所示，其中黑体代表最优结果，下划线代表次优结果。

从表2中可知，基于图神经网络的方法普遍优于基于翻译的方法，这是因为基于图神经网络的方法能够充分利用知识图谱的结构信息。DuPNet全面优于JETEA(Song等, 2021)和JTMEA(Lu等, 2021)。与解决结构异构性的模型相比，DuPNet除了在ZH_EN的Hits@1和Hits@10上是次优结果，在JA_EN的Hits@1上与次优结果持平，在其他所有数据集的所有指标上都是最优结果，证明了DuPNet在处理结构异构性方面的优越性。这是因为DuPNet提出的双视角交互和门控机制起了非常重要的作用。一方面，双视角交互综合考虑了实体邻域异构性和关系异构性，可以使得对齐实体学习到更加相似的表示。另一方面，DuPNet通过门控机制对网络的隐层和输出层进行聚合，可以学习到实体更精确的表示，在保留多跳邻域信息的同时有效地去除各层的噪声。

4.2.3 DuPNet在电子病历医疗知识图谱上的实验结果

	Hits@1	Hits@5	Hits@10	Hits@50	MRR
DuPNet(w/o ent)	73.9	82.7	86.5	92.6	78.1
DuPNet(w/o rel)	75.5	82.7	86.3	94.5	78.5
DuPNet(w/o gate)	72.5	82.9	86.1	94.0	77.1
DuPNet	76.1	84.1	86.8	94.5	79.4
DuPNet-Bert	84.3	92.3	94.5	99.2	87.6

Table 3: DuPNet在电子病历医疗知识图谱上的实验结果

(1)DuPNet总体结果

医疗知识图谱KG_1和KG_2经由实体规范化和规则挖掘补全后得到的知识图谱$CSKG_1$和$CSKG_2$。我们对$CSKG_1$和$CSKG_2$按照3.2.3节的方法构建医疗实体种子对，得到的实体种子对数量为910，为保证模型训练充分，将实体种子对的60%作为训练集，40%作为测试集。DuPNet在医疗知识图谱$CSKG_1$和$CSKG_2$上的实验结果如表3所示，Hits@1值达到76.1%，Hits@10达到86.5%。

(c) 2022 中国中文信息学会计算语言学专业委员会

(2)利用Bert提高DuPNet在实际应用中的性能

为提高模型在实际应用中的性能，我们在DuPNet的基础上引入Bert预训练模型。由于DuPNet模型中实体表示是通过随机初始化得到的，因此实体的嵌入表示中仅包含结构信息。在医疗领域，医学实体的名称蕴含着丰富的语义信息，能够反应医学实体的重要特征，因此我们用Bert对医学实体名称编码，用带有丰富语义信息的词嵌入初始化实体表示来增强实体对齐。我们将引入Bert后的模型命名为DuPNet-Bert。

DuPNet-Bert在医疗知识图谱$CSKG_1$和$CSKG_2$上的实验结果如表3所示，在Hits@1上达到84.3%，在Hits@10上达到94.5%。与DuPNet相比，DuPNet-Bert在Hits@1、Hits@10和MRR上分别提高了8.2%，8.0%和8.2%，充分验证了Bert能够给实体嵌入表示带来有意义的语义信息。

(3)消融实验

为了验证DuPNet中各个模块的有效性，我们进行了详细的消融实验，结果如表3所示。

首先，去除实体交互模块，并将该实验表示为DuPNet(w/o ent)。实体交互模块的去除导致DuPNet性能整体降低，在Hits@1、Hits@10和MRR上分别下降了2.2%、0.3%和1.3%，这是因为实体交互视角通过给邻居赋予精确的权重缓解实体邻域异构性。实验结果证明了实体交互视角的有效性。

然后，去除关系交互模块，并记为DuPNet(w/o rel)。与DuPNet相比，DuPNet(w/o rel)在Hits@1、Hits@10和MRR上下降了0.6%、0.5%和0.9%，原因在于关系交互视角通过使用关系匹配度能够充分缓解关系异构性，实验结果证明了关系交互视角的有效性。

最后，去除门控机制，用平均池化代替，并表示为DuPNet(w/o gate)。门控机制被去除后，DuPNet在Hits@1、Hits@10和MRR上分别下降了3.6%、0.7%和2.3%。这是因为门控机制在捕获实体多跳邻域信息的同时能够有效去除多层卷积带来的噪声。实验结果证实了门控机制的有效性。

5 结论

医疗知识图谱中知识重叠和互补的现象普遍存在，利用实体对齐进行医疗知识图谱融合成为迫切需要。然而据我们调研，目前在医疗领域的知识图谱实体对齐尚没有完整的处理方案。针对医疗知识图谱的特点提出了一种规范的电子病历医疗知识图谱实体对齐流程，为中文医疗领域的实体对齐提供了一种可行的方案。针对基于电子病历知识图谱结构异构性的特点，设计了一个双视角并行图神经网络模型并用于医疗知识图谱实体对齐，实验结果证明了该模型处理结构异构性的优越性，并且按照上述流程进行了实际的基于中文电子病历知识图谱实体对齐，取得了较好的效果。

参考文献

奥德玛, 杨云飞, 穗志方, 等. 2019. 中文医学知识图谱CMeKG构建初探. 中文信息学报, 33(10):1-9.

Bhattacharya Indrajit, Getoor Lise. 2007. Collective entity resolution in relational data. *Information Sciences*, 1(1):1-36.

Bordes Antoine, Usunier Nicolas, Garcia-Duran Alberto, et al. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 2787–2795.

蔡娇. 2020. 基于遗传病领域的实体对齐研究. 硕士学位论文. 苏州大学.

Cao Yixin, Liu Zhiyuan, Li Chengjiang, et al. 2019. Multi-Channel Graph Neural Network for Entity Alignment. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1452-1461.

Chen Muhao, Tian Yingtao, Yang Mohan, et al. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1511-1517.

Chen Wei, Chen Xiaoying, Xiong Shengwu. 2021. Global Entity Alignment with Gated Latent Space Neighborhood Aggregation. *China National Conference on Chinese Computational Linguistics*, 371-384.

(c) 2022 中国中文信息学会计算语言学专业委员会

Jiang Yong, Wang Xinmin, Zheng Haitao. 2014. A semantic similarity measure based on information distance for ontology alignment. *Information Sciences*, 278:76-87.

Lan Yinyu, He Shizhu, Liu Kang, et al. 2021. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. *BMC Medical Informatics Decis Mak*, 21-S(9): 335.

Li Chengjiang, Cao Yixin, Hou Lei, et al. 2019. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2723–2732.

刘勘, 张雅荃. 2020. 基于医疗知识图谱的并发症辅助诊断. 中文信息学报, 34(10):85-93,104.

刘道文, 阮彤, 张晨童, 等. 2021. 基于多源知识图谱融合的智能导诊算法. 中文信息学报, 35(01):125-134.

Lu Guoming, Zhang Lizong, Jin Minjie, et al. 2021. Entity alignment via knowledge embedding and type matching constraints for knowledge graph inference. *Journal of Ambient Intelligence and Humanized Computing*, 1-11.

宋文欣. 2018. 面向医疗领域的实体对齐研究. 硕士学位论文. 哈尔滨工业大学.

Song Xiuting, Zhang Han, Bai Luyi. 2021. Entity Alignment Between Knowledge Graphs Using Entity Type Matching. *International Conference on Knowledge Science, Engineering and Management*, 578-589.

Sun Zequn, Hu Wei, Li Chengkai. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. *International Semantic Web Conference*, 628-644.

Sun Zequn, Wang Chengming, Hu Wei, et al. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):222-229.

Wang Zhichun, Lv Qingsong, Lan Xiaohan, et al. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 349-357.

Xiu Xiaolei. 2020. Construction of a Digestive System Tumor Knowledge Graph Based on Chinese Electronic Medical Records: Development and Usability Study. *JMIR Med Inform*, 8(10):e18287.

Zhu Hao, Xie Ruobing, Liu Zhiyuan, et al. 2017. Iterative Entity Alignment Via Joint Knowledge Embeddings. *International Joint Conference on Artificial Intelligence*, 4258–4264.

CCL 2022

基于平行交互注意力网络的中文电子病历实体及关系联合抽取

李丽双*，王泽昊，秦雪洋，袁光辉
大连理工大学
计算机科学与技术学院
辽宁，大连
lils@dlut.edu.cn, dutzehao@mail.dlut.edu.cn
qinxueyang@snnu.edu.cn, 476708484@qq.com

摘要

基于电子病历构建医学知识图谱对医疗技术的发展具有重要意义，实体和关系抽取是构建知识图谱的关键技术。本文针对目前实体关系联合抽取中存在的特征交互不充分的问题，提出了一种平行交互注意力网络(PIAN)以充分挖掘实体与关系的相关性，在多个标准的医学和通用数据集上取得最优结果；当前中文医学实体及关系标注数据集较少，本文基于中文电子病历构建了实体和关系抽取数据集(CEMRIE)，与医学专家共同制定了语料标注规范，并基于所提出的模型实验得出基准结果。

关键词： 实体关系联合抽取 ； 双向特征交互模块 ； 自注意力机制 ； 中文电子病历 ； 数据集标注与构建

Parallel Interactive Attention Network for Joint Entity and Relation Extraction Based on Chinese Electronic Medical Record

LiShuang Li*, Zehao Wang, Xueyang Qin, Guanghui Yuan
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
lils@dlut.edu.cn, dutzehao@mail.dlut.edu.cn
qinxueyang@snnu.edu.cn, 476708484@qq.com

Abstract

The construction of medical knowledge graph based on electronic medical records is of great significance to the development of medical technology, where entity and relation extraction plays a pivotal role. To solve the issue of insufficient interaction in the current joint entity and relation extraction approaches, we propose a Parallel Interactive Attention Network (PIAN) which can fully exploit the correlation between entity and relation, achieving the state-of-the-art results on standard datasets. Since there are few Chinese medical entity and relation annotation datasets, we construct an entity and relation extraction dataset based on Chinese electronic medical records (CEMRIE), formulate the corpus annotation specification with medical experts, and give the benchmark results based on our proposed model.

Keywords: Joint Entity and Relation Extraction , Bidirectional Feature Interaction Module , Self-Attention Mechanism , Chinese Electronic Medical Record , Dataset Annotation and Construction

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：大连市科技创新基金(2020JJ26GX035)

1 引言

电子病历中记录了丰富的临床医学信息，例如疾病、症状等重要的医学实体，以及各类型医学实体之间的语义关系。随着医疗信息化的高速发展，各医院已经积累了海量的电子病历数据，但如何高效地从非结构化的电子病历文本中提取有价值的医疗信息仍是难题。目前，越来越多的研究者利用深度学习技术来自动抽取电子病历中的关键信息。其中知识图谱是一种能将信息要素结构化、规范化并以图的形式清晰展示的技术。构建基于电子病历的知识图谱可以对电子病历中的医学知识进行结构化描述，帮助医生和大众更方便的获取想要的知识，同时构建大规模知识图谱(Lan et al, 2021)也能为医疗问答、辅助决策等下游应用提供重要的技术支持。

其中，实体和关系抽取是构建知识图谱的关键技术之一，其主要目的是利用相关技术从结构化、半结构化或者自然语言中抽取得到实体关系三元组。实体和关系抽取主要分为流水线方法和联合抽取方法，目前在医学领域，实体及关系抽取主要基于流水线的方法，该方法将实体识别和关系抽取视为两个独立的任务。对于一段文本，首先使用实体识别模型抽取出所有实体，然后再利用关系抽取模型判断每个实体对的关系类别。命名实体识别作为关系抽取的研究基础和关键，二者之间联系密切。传统的基于流水线方法忽略了两个任务之间潜在的关联性，并且存在误差传播的问题，实体识别阶段的错误实体会严重影响关系抽取模型的结果。

为了解决流水线方法存在的问题，有研究考虑对实体识别和关系抽取进行联合学习。在通用领域，实体关系联合抽取主要分为统一编码和共享参数两种方法。统一编码方法(Zheng et al, 2017; Wang et al, 2020; Ren et al, 2021; Wang et al, 2021)将实体和关系编码到统一的标签空间，并学习统一的特征来同时表示实体与关系。然而，该类方法使用同一个模型对两个任务进行编码，一个任务的特征可能会与另一个任务特征产生冲突，导致特征混淆的问题，损害模型的整体性能。共享参数方法(Miwa and Bansal, 2016; Bekoulis et al, 2018; Wei et al, 2020; Yan et al, 2021) 通常采用相互独立的网络分别为实体与关系编码不同的特征表示，两个任务通过共享输入特征以及部分网络参数实现信息交互，可以避免特征混淆的问题。统一编码与共享参数各有其优势及局限性，共享参数的方法对实体和关系分别独立编码，克服了统一编码方法中的两个任务间特征冲突的问题，但两个任务间不能充分交互。

在医学领域，目前采用联合方法进行实体和关系抽取的相关研究较少，为了充分利用医学实体识别与关系抽取之间的密切联系，构建高质量的医学知识图谱，本文构建了一个平行交互注意力网络(Deep Interactive Attention Network, PIAN)，并应用于中文电子病历，进行中文医学实体和关系的联合抽取。模型采用两个平行的神经网络来分别编码医学实体和关系以抽取两种任务的特征，避免了两个任务的特征混淆问题。同时利用双向特征交互模块(Bidirectional Feature Interaction Module, BFIM)用于双向建模两个任务间深度的特征交互。具体地，BFIM可以使一个任务中每个字符的特征与来自另一任务对应的字符特征进行融合，并自适应地学习融合的比例。因此对于另一任务中有潜在价值的特征，BFIM可以提高这些特征的融合比例，来获取对当前任务有价值的信息。由于BFIM对称地建模了特征融合，并且使两个任务中每个字符的特征都参与了交互，因此其能够避免具有潜在价值的特征丢失并实现双向交互，从而实现医学实体和关系的深度融合。同时，为使实体识别与关系抽取两个任务能够更好的学习各自不同的任务特征，本文采用注意力机制(Self-Attention)(Guo et al, 2021)学习各自的任务特征表示。

目前，高质量的中文医学实体和关系联合抽取数据集较为匮乏，公开的主要有中文医学信息抽取数据集(Chinese Medical Information Extraction dataset, CMeIE)。中文电子病历中包含众多的生物医学实体以及实体间丰富的语义关系，本文基于真实的电子病历文本，首先通过分析大量中文电子病历语料的语言结构特点，与专家共同制定了实体及关系的标注规则，最终形成了一套标准的数据集标注流程与规范。然后构建了一个中文电子病历实体关系抽取数据集(Chinese Electronic Medical Record Information Extraction dataset, CEMRIE)，并利用所构建的平行交互注意力网络模型进行实体关系联合抽取，在实体和关系上的F1值分别达到了89.7%和80.4%的基准结果。

2 相关工作

2.1 流水线方法

医学领域的实体关系抽取主要是基于流水线的方法，其将实体识别与关系抽取视为两个独立的任务。对于医学实体识别，其模型大部分是基于LSTM-CRF结构(Rei et al, 2016)并

(c) 2022 中国中文信息学会计算语言学专业委员会

结合注意力机制(Luo et al, 2018)和预训练预训练语言模型(Lee et al, 2020)。与实体识别任务类似，当前主流的医学关系抽取模型主要采用CNN(Zeng et al, 2014)、注意力机制(Yi et al, 2017)、Transformer结构(Christopoulou et al, 2020)、预训练模型结合外部知识(Sun et al, 2020)、联邦学习(Sui et al, 2020)等结构。除此之外，图结构能够很好地建模层次结构复杂的生物医学长句，有利于关系特征的提取，因此图神经网络模型(Park et al, 2020)也被广泛应用于医学领域的关系抽取。Zheng等(Zheng et al, 2021)将实体的类别信息作为标签插入到抽取实体的两端，再通过关系抽取模型判断关系，取得了优异的效果。然而流水线方法忽略了两个任务的相关性以及依赖关系，并且其存在的错误传播问题也限制了流水线方法的性能，因此研究人员提出使用联合方法来同时抽取实体和关系。

2.2 联合抽取方法

2.2.1 通用领域的联合方法

在通用领域，当前主要有统一编码和共享参数两种方法。统一编码采用联合的标注策略，Zheng等(Zheng et al, 2017) 提出将命名实体识别和关系抽取联合抽象为一种序列标注任务，但是无法解决关系重叠问题。Wang等(Wang et al, 2020)提出单阶段抽取框架，以全新的字符对链接的角度解决了关系重叠的问题。随后Bekoulis等(Bekoulis et al, 2018)提出表填充的方法，即将实体和关系共同填入一个二维表中。由于二维表比一维序列具有更强的表示能力并且能够很好的表示嵌套实体以及重叠关系，因此越来越多的方法采用基于表填充的解码策略。Wang等(Wang et al, 2021)设计了三步走的近似解码策略同时解码出实体和关系，并采用了双仿射注意力机制来学习头实体和尾实体之间的相关性，然而，由于使用同一模型编码，两个任务的特征可能会彼此冲突，导致特征混淆的问题，从而降低模型性能。

基于共享参数的方法为两个任务分别学习独立的特征表示，避免了特征混淆，但是需要显式地建模任务间交互以利用其相关性。Miwa等(Miwa and Bansal, 2016)采取端到端神经网络模型，通过双向LSTM以及Tree-LSTM捕获单词序列信息和依存句法树结构信息。Bekoulis等(Bekoulis et al, 2018)提出了多头选择机制用于联合抽取。Wei等(Wei et al, 2020)提出了一种级联二进制标记框架，先抽取主体实体，再以关系作为条件抽取客体实体。然而这些方法仅通过共享输入或部分网络参数来实现信息共享，两个任务未能实现充分交互。Wang等(Wang and Lu, 2020)采用独立的序列编码器和表编码器来分别编码实体和关系任务，两个编码器之间存在显式交互，但是由于实体任务采用序列编码，其无法处理嵌套实体，表示能力有限。Yan 等(Yan et al, 2021)提出了分区过滤网络(Partition Filter Network)用于分别学习NER特征、RE特征以及共享特征，然而该模型在过滤阶段会裁剪掉部分特征，可能会丢失潜在有用信息。

2.2.2 医学领域的联合方法

目前，医学领域的联合抽取方法较少，Li等(Li et al, 2017) 构建了基于双向LSTM-RNN的联合学习模型，用于药物不良事件提取(Adverse Drug Extraction, ADE)，该模型采用共享参数的方式实现任务共享，但实际还是将两任务先后分开处理，仍然会产生误差传播。Lu等(Luo et al, 2020)提出一种基于标注策略的生物医学联合学习模型，将命名实体识别和关系抽取联合抽象为一种序列标注任务，通过合并两个任务的类型标签设计了一种新的标注方案和提取规则，但是无法解决实体关系重叠问题。Fei等(Fei et al, 2021)提出了基于Span的联合模型，采用双仿射注意力机制、语义依存分析以及图卷积网络，着重于解决实体关系重叠的问题，但是基于Span的方法会产生大量冗余跨度对，给模型引入噪声并且增加计算消耗。

3 模型

本文将医学实体关系联合抽取定义为$\text{Triple}\{\text{E}_1, \text{R}, \text{E}_2\} = \text{Model}(\text{S})$，其输入为句子序列S，输出为模型抽取出来的三元组Triple，其中包含两个实体E_1, E_2以及实体间的对应关系R。模型的主要结构如图1所示，实体识别和关系抽取两个任务分别采用相互平行的网络结构并结合特征交互模块。模型输入为字符序列，首先经过预训练语言模型编码，分别为实体识别和关系抽取生成初始特征向量，再通过BFIM进行特征交互，然后通过SAM学习各自任务的特征表示；两个任务通过“交互—学习—交互”的形式，逐步提高了两种特征的质量；最后将两个任务分别用表填充的方式进行解码并计算损失。

(c) 2022 中国中文信息学会计算语言学专业委员会

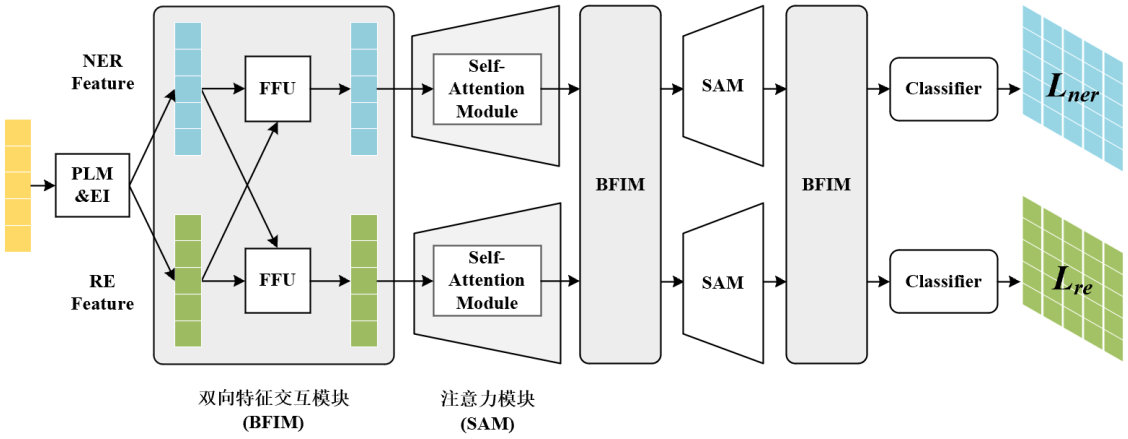


Figure 1: 平行交互注意力网络总体框架

3.1 初始特征生成

对于给定的输入句子$S = \{w_1, w_2, ..., w_n\}$，$n$表示输入序列长度，首先使用预训练语言模型例如BERT(Devlin et al, 2018)等，进行编码获得具有上下文信息的特征表示，然后分别为实体识别与关系抽取两个任务生成初始特征，考虑到实体识别和关系抽取既存在差异性又包含相关性。因此，模型采用两个独立的前馈神经网络(Feed Forward Network, FFN)并结合Dropout机制分别为两个任务生成相应的初始特征:

$$H_{bert} = \text{BERT(S)} = \{h_1, h_2, ..., h_n\}, \tag{1}$$

$$H_e = \text{Linear}_{Drop}(H_{bert}), \tag{2}$$

$$H_r = \text{Linear}_{Drop}(H_{bert}), \tag{3}$$

其中，H_e和H_r分别表示实体识别和关系抽取的初始特征表示。FFN的输入均为H_{bert}，保证了两种任务特征的相关性；并通过不同的FFN结合Dropout机制随机删除部分神经元，实现两种任务特征的差异化。

3.2 双向特征交互模块

为了建模任务间的信息交互并充分挖掘实体识别与关系抽取的相关性，本文提出了一个双向特征交互模块(BFIM)，如图1所示。该特征交互模块由两个独立的特征融合单元(Feature Fusion Unit, FFU)组成，用于融合两种任务特征，并为两个任务分别生成新的特征表示。对于来自两个任务同一位置的字符特征h_e^i, $h_r^i \in \mathcal{R}^{1\times D}$(D为字符特征维度)，首先使用一个线性层Linear将字符特征维度映射为1并拼接为$[\text{Linear}^{\text{e}}(h_e^i); \text{Linear}^{\text{e}}(h_r^i)] \in \mathcal{R}^{n\times 2}$(n为输入序列长度)，然后使用$Softmax$归一化获得融合分数(即两个任务的融合比例)，最后用融合分数与原始特征h_e^i, h_r^i做点乘操作并相加，获得某一任务的融合特征$h_{\triangle}^{*i}$:

$$(\gamma_{e,e}^i, \gamma_{e,r}^i) = Softmax([\text{Linear}^{\text{e}}(h_e^i); \text{Linear}^{\text{e}}(h_r^i)]), \tag{4}$$

$$(\gamma_{r,e}^i, \gamma_{r,r}^i) = Softmax([\text{Linear}^{\text{r}}(h_e^i); \text{Linear}^{\text{r}}(h_r^i)]), \tag{5}$$

$$\begin{bmatrix} h_e^{*i} \\ h_r^{*i} \end{bmatrix} = \begin{pmatrix} \gamma_{e,e}^i & \gamma_{e,r}^i \\ \gamma_{r,e}^i & \gamma_{r,r}^i \end{pmatrix} \begin{bmatrix} h_e^i \\ h_r^i \end{bmatrix}, \tag{6}$$

其中，Linear^{e}和Linear^{r}分别表示实体识别和关系抽取任务中FFU的线性层，$\gamma_{e,x}^i$和$\gamma_{r,x}^i$分别代表特征h_x^i在实体或关系任务新融合的特征中所占比例，[;]代表连接操作。FFU用于自适应地学习当前任务每个字符特征与另一任务对应的字符特征的最佳融合比例。

双向特征交互模块能够使每一个任务都能够自适应地融合来自另一任务的特征，同时实现更细粒度的字符级特征交互，即每个字符特征h_x^{*i}都会学习到不同的融合比例$\gamma_{x,e}^i$,$\gamma_{x,r}^i$，实现实体与关系特征之间的深度交互，从而充分利用任务间潜在有价值的特征。本文在5.4.2详细分析了每个融合单元FFU中两种任务的融合比例，验证了本文提出的BFIM的有效性。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.3 注意力模块

注意力机制广泛应用于计算机语言以及视觉领域，并且取得了显著的效果。其中最具代表性的是自注意力机制(Self-Attention)，其可以加强重要信息的权重，并减弱干扰信息的影响。本文为实体识别与关系抽取两个任务分别设置了独立的自注意力网络，用于学习两种不同的任务特征。自注意力的计算如下，并采用“多头”模式增强注意力特征的表示能力，使模型共同处理来自不同表示子空间的信息：

$$H_{\triangle}^{multi} = [head_0; ...; head_s]W^O, \tag{7}$$

$$head_i = Attention(H_{\triangle}^{*}W_i^Q, H_{\triangle}^{*}W_i^K, H_{\triangle}^{*}W_i^V), \tag{8}$$

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{9}$$

其中，W_i^Q，W_i^K，W_i^V，W^O为可训练参数，$\triangle$表示实体识别或关系抽取中的某个任务，s为注意力头数，[;]代表连接操作，经过注意力机制每个任务学习到各自具体的任务特征，之后再通过多层级的交互以及注意力网络的学习，不断提升任务特征的质量。然而随着模型深度的增加，网络可能会发生退化问题(Degradation problem)，即准确率会趋于饱和并极速下降。为了解决这一问题，在注意力模块后采用残差网络(He et al, 2016)结构并结合层归一化(Layer Normalization, LN)进行处理，得到最终输出$H_{\triangle}^{out}$：

$$H_{\triangle}^{out} = \mathrm{LN}(H_{\triangle}^{multi} + H_{\triangle}^{*}). \tag{10}$$

3.4 分类模块

本文采用基于表填充的方法对两个任务特征进行分类，将句中每个字符的特征分别与所有字符组成实体对特征，得到二维的特征表示。对于实体识别，使用实体的开始字符特征h_i与结束字符特征h_j拼接表示该实体，得到了该实体的特征表示$[h_i^e; h_j^e]$，然后通过线性层并使用ELU及$sigmoid$函数激活，得到该实体的概率分布：

$$h_{i,j}^e = \mathrm{ELU}(\mathrm{Linear}([h_i^e; h_j^e])), \tag{11}$$

$$e_{i,j}^k = \sigma(\mathrm{Linear}(h_{i,j}^e)), \tag{12}$$

与实体识别相同，关系抽取也采用基于表填充方式，使用头实体的开始字符与尾实体的开始字符来表示两个实体之间存在关系。通过线性层和激活函数得到实体对的关系特征$h_{i,j}^r$以及关系概率分布$r_{i,j}^\tau$：

$$h_{i,j}^r = \mathrm{ELU}(\mathrm{Linear}([h_i^r; h_j^r])), \tag{13}$$

$$r_{i,j}^\tau = \sigma(\mathrm{Linear}(h_{i,j}^r)). \tag{14}$$

3.5 任务训练

对于给定的训练数据集，模型的训练损失函数$\mathcal{L}$由实体识别任务的损失$\mathcal{L}_{ner}$和关系抽取任务的损失$\mathcal{L}_{re}$组成：

$$\mathcal{L}_{ner} = \Sigma_{i,j\in(1,n),\varepsilon\in\mathcal{E}}\mathrm{BCELoss}(e_{i,j}^{\varepsilon}, \bar{e}_{i,j}^{\varepsilon}), \tag{15}$$

$$\mathcal{L}_{re} = \Sigma_{i,j\in(1,n),\tau\in\mathcal{T}}\mathrm{BCELoss}(r_{i,j}^{\tau}, \bar{r}_{i,j}^{\tau}), \tag{16}$$

其中$\bar{e}_{i,j}^{\varepsilon}$与$\bar{r}_{i,j}^{\tau}$表示该实体和关系的真实标签，对两个任务均采用二分类交叉熵损失(Binary CrossEntropy Loss)函数计算各自任务损失，模型的总损失为$\mathcal{L} = \mathcal{L}_{ner} + \mathcal{L}_{re}$。

4 中文电子病历数据集构建

4.1 语料分析及预处理

本文的电子病历语料来源于某医院去隐私化后的电子病历，共包含四个科室：内科、外科、妇产科和皮肤性病科。病历中包含多个标签，例如“一般资料”、“主诉”、“现病史”、“查

(c) 2022 中国中文信息学会计算语言学专业委员会

体”等。每个病历中标签的类型不同，标签中的文本都包含着医疗信息，所对应的医疗信息类型区别也较大，如“查体”标签中包含较多的检查信息，“临床诊断”标签包含较多的治疗信息。对于不同科室的病历，由于医生之间的书写习惯、不同科室之间的规范不同，所包含的标签种类也不一致，这充分说明了中文电子病历领域不同样本之间的差异性，迫切需要一个统一的框架来兼容这种差异性。不仅是结构上的差异，文本语言也有很大不同，由于病历是由不同医生撰写，而医生的学历不同、籍贯不同、性别职位不同等原因会造成内容上的差异性。

除病历之间的差异性外，电子病历本身也有其特点。首先病历的组织形式属于半结构化文本，各个标签下的语言大都不是完整的一句话，语言较为简洁化、专业化；其次，病历文本中包含有大量的医学符号、数字，这些内容不同于在通用领域的含义，在医学上有其特殊意义；除此之外，病历中还包含有很多缩略词、同义词、医学专用术语等。这些病历特点给后期信息抽取带来了很大的困难。

本文将电子病历语料预处理的流程主要分为语料拆分、标签归类、标签数量统计和筛选。

语料拆分：电子病历是一种半结构化数据，包含标签及其对应的文本，对于一份完整的电子病历，本文需先根据其中的标签对病历进行拆分，每个标签下对应的为用自然语言形式描述的医学知识，方便模型处理，而且使每一段文本与对应的标签相关联，可以更精准的标识语料所包含的知识的类型。

标签归类：每个病历包含多个标签，标签表明该文本具有某种信息类型，标签之间有的差别较大，有的较为相近，统一处理会有差异性，例如标签“体格检查”和“诊疗经过”，前者包含更多的检查信息，后者包含更多的治疗信息，而“门诊化验”、“入院检查”等都与检查相关。本文先对标签进行手工归类，将包含相同方向医学知识的标签放在一个集合，后续再进行标签数量的统计和标签筛选。

标签数量统计和筛选：将标签按其所包含医学方向分类后，统计所有集合包含标签的数量，按数量多少对标签集合进行排序，从结果上看，标签“一般资料”、“主诉”、“现病史”、“既往史”数量较多，每个标签都包含着丰富的医疗信息，同时不同标签之间也会有差异性，这样分类排序之后可以更清晰的分析语料中所包含各类知识的数量比。依据统计的标签集合数量从中抽取了几个包含标签较多的集合作为本文构建数据集的语料。

4.2 数据标注及标注规则介绍

借鉴统一医学语言系统(Unified Medical Language System, UMLS)概念标准及各个中文医疗语料库的构建标准，结合抽取到的中文电子病历语料结构特点，与医学专家讨论制定了一套中文电子病历语料标注规范。

在标注规范中，本文将实体类型尽量泛化以适应语料中样本间的差异性，共定义了疾病、部位、症状、检查和治疗五种类型的实体，其详细定义如表1所示。命名实体的标注规则遵循实体间不重叠、不嵌套、实体内不含有标点符号的原则。根据确定的实体类型，本文将实体之间的关系类型分为七个类别：“疾病-疾病”、“疾病-部位”、“疾病-症状”、“治疗-疾病”、“治疗-症状”、“检查-疾病”和“检查-症状”，其详细定义如表2所示。

类别	定义
疾病	指导致病人身体或心理上出现的非正常现象，或者由医生对病人做出的诊断，并且是可以治疗的。
症状	泛指由疾病或其它突出状况导致身体不适或异常的表现，通常是病人主观感觉的不适或病理改变。
部位	指人的身体部位，包括器官或器官组成、身体系统以及身体位置或区域等。
检查	为了确认疾病是否存在，或是为了解疾病的病因而进行的检查项目、查体以及实施的检查设备。
治疗	针对疾病或症状而采取的药物、手术或医疗设备等治疗方法。

Table 1: 实体各类别定义

4.3 数据标注过程

为了获取高质量标注电子病历数据，在与医学专家讨论并制定中文电子病历实体及关系标注规范的基础上，选取了10,000份电子病历文档进行人工标注。标注过程可以分为两步：预标注和正式标注。预标注使用了20%的数据，每份电子病历数据采取两人同时标注的策略，标注结束后双方交互验证，对于标注相同的结果初步认定是准确的，反之交由医学专家进一步分析

关系类别	定义
疾病-疾病	泛指疾病之间的相关并发症、疾病表明疾病或者疾病的别名等。
疾病-部位	泛指疾病体现的部位，一般指发病部位，也有转移部位等情况。
疾病-症状	泛指疾病的一种体现形式，一般指疾病导致的某种症状。
治疗-疾病	泛指治疗应用于疾病，使疾病好转或恶化，或是没有提及治疗效果。
治疗-症状	泛指针对某些症状采取的治疗手段，或是因治疗所产生的症状。
检查-疾病	泛指检查确认了疾病的发生，或为证实疾病而采取某种检查手段。
检查-症状	泛指检查显示正常或者异常症状，或者检查确认是否存在症状。

Table 2: 关系各类别定义

评判，并找出标注结果不同的原因，以此进一步修订、细化电子病历标注规范。预标注结束后，正式标注严格按照制定的电子病历标注规范进行执行，同样地，每一份电子病历数据由两人同时标注，对于标注不同的结果，则分配给另外的标注者进行标注，直至标注结果相同。最终经过筛选去重后共获得实体30,058个，关系17,904对，各类别标注数量如表3、4所示。为了方便模型训练，标注完成后对较长的电子病历文本进行拆分，最终得到12,219条可训练样例。

4.4 数据集质量评估

为了保证数据集的质量，随机抽取了2,000个标注实体及2,500个标注关系进行质量评估，在评估阶段有医学专家全程参与，评估结果如表3、4所示，可以看出，实体和实体关系的准确率分别为92.9%和95.6%。另外，在评估过程中可以发现，实体的错误主要集中在实体的类型错误，即实体本身是正确的，但实体的类型是错误的，比如：“头痛”既可以充当疾病实体也可以充当症状实体，在不同的电子病历文本中可能表现出不同的实体类型；类似的，关系类别的错误也集中在实体类型的错误上，如“疾病-疾病”与“疾病-症状”。

实体类别	总数	抽样数	错误数	准确率(%)
疾病	12542	600	45	92.5
症状	7767	400	28	93.0
部位	843	100	4	96.0
检查	4646	400	31	92.3
治疗	4260	500	35	93.0
All	30058	2000	143	92.9

Table 3: 实体数量统计及质量评估

实体类别	总数	抽样数	错误数	准确率(%)
疾病-疾病	2935	200	8	96.0
疾病-症状	4221	200	13	93.5
疾病-部位	855	100	3	97.0
治疗-疾病	3388	500	15	97.0
治疗-症状	1938	500	27	94.6
检查-症状	1947	500	20	96.0
检查-疾病	2620	500	25	95.0
All	17904	2500	111	95.6

Table 4: 关系数量统计及质量评估

5 实验与结果分析

5.1 数据集与评价指标

基于中文电子病历的实体及关系抽取数据集CEMRIE总共包含12,219条样本，数据集按照80%:20%的比例划分为训练集和测试集，其详细数据如表5所示。采用严格匹配的方式来评估抽取三元组的效果，即当且仅当预测的三元组中，实体边界、实体类型以及关系类型完全正确时，预测的三元组才被视为正确；实验使用准确率P、召回率R以及F1值，并采用微平均(Micro-F1)的方式对结果进行评估。

CEMRIE	样本数	实体数	关系数
训练集	9773	24077	14349
测试集	2446	5981	3555

Table 5: CEMRIE数据集详情

5.2 超参数设置及实验环境

采用RoBERTa$_{\text{base}}$预训练模型(Liu et al, 2019)，隐藏层向量维度设为768。采用Adam优化器，初始学习率设为2e-5，并且当训练轮次到达20、50和70时，学习衰减为原来的一半，训练100轮。批处理大小设置为4，随机种子设置为99。

实验环境：操作系统Ubuntu20.04LTS，显卡NVIDIA GeForce RTX3090GPU，PyTorch版本为1.7.1，Python版本为3.7.11。

5.3 PIAN联合抽取实体关系性能测试

5.3.1 标准数据集实验结果

Dataset	Model	PLM	NER			RE		
			P	R	F1	P	R	F1
ACE05	Tab2Seq(Wang and Lu, 2020)	$\text{ALBERT}_{\text{large}}$	-	-	89.5	-	-	64.3
	PURE(Zheng et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	89.7	-	-	65.5
	PFN(Yan et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	89.5	-	-	66.8
	UNIRE▲(Wang et al, 2021)	$\text{ALBERT}_{\text{large}}$	**89.9**	**90.5**	**90.2**	**72.3**	60.7	66.0
	PIAN(Ours)	$\text{ALBERT}_{\text{large}}$	89.5	90.1	89.8	69.9	**66.5**	**68.1**
	PIAN(Ours)	$\text{BERT}_{\text{base}}$	88.6	88.5	88.5	67.7	61.6	64.5
	-w/o BFIM	$\text{BERT}_{\text{base}}$	87.6	88.3	87.9	70.1	57.5	63.2
	-w/o SAM	$\text{BERT}_{\text{base}}$	88.3	88.4	88.3	68.1	60.4	63.9
	-w/o BFIM&SAM	$\text{BERT}_{\text{base}}$	87.5	88.3	87.9	69.1	59.0	63.6
ACE04	Tab2Seq(Wang and Lu, 2020)	$\text{ALBERT}_{\text{large}}$	-	-	88.6	-	-	59.6
	PURE(Zheng et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	88.8	-	-	60.2
	PFN(Yan et al, 2021)	$\text{ALBERT}_{\text{large}}$	-	-	89.3	-	-	62.5
	UNIRE▲(Wang et al, 2021)	$\text{ALBERT}_{\text{large}}$	88.9	**90.0**	89.5	67.3	59.3	63.0
	PIAN(Ours)	$\text{ALBERT}_{\text{large}}$	**89.6**	89.9	**89.8**	**70.2**	**58.8**	**64.0**
CMeIE	Baseline(Zhang et al, 2021)	$\text{RoBERTa}_{\text{large}}$	-	-	-	-	-	55.9
	CopyR_{RL}(Zeng et al, 2019)	$\text{BERT}_{\text{base}}$	-	-	-	54.0	55.7	54.6
	CasRel(Wei et al, 2020)	$\text{BERT}_{\text{base}}$	-	-	-	58.4	58.0	58.1
	NPCTS(王泽儒和柳先辉, 2022)	$\text{BERT}_{\text{base}}$	-	-	-	59.3	57.6	58.4
	PIAN(Ours)	$\text{RoBERTa}_{\text{base}}$	-	-	-	**63.8**	**58.8**	**61.2**

Table 6: PIAN在标准数据集上的实验结果，其中▲表示该方法利用了跨句信息；由于CMeIE为线上测评，无法提供实体结果

为了验证PIAN在实体关系联合抽取上的有效性，首先在标准数据集ACE04[0] (Doddington et al, 2004)，ACE05(Christopher et al, 2005)和中文医学数据集CMeIE[1](Zhang et al, 2021)上，与当前较为先进的模型进行比较，如表6所示。在数据集ACE05上，PIAN在关系结果上取了较大提升，比当前结果最好的共享参数模型PFN提高了1.3%；实体识别也取得了相当的结果，F1值仅比UNIRE低0.4%，这是因为UNIRE利用了额外的跨句子信息进行训练，其更有利于实体的预测。而本文方法没有使用跨句子信息，因此实体效果提升不明显。在数据集ACE04上，PIAN超越了当前效果最好的统一编码模型UNIRE，实体结果提升了0.3%，关系结果提升了1.0%。在中文医学数据集CMeIE上，PIAN结果显著高于NPCTS模型，NPCTS是对CasRel模型的改进，是一种基于指针级联标注策略的联合抽取模型，其中CasRel与CopyR_{RL}的实验结果为(王泽儒和柳先辉, 2022)复现的结果。本文模型在不借助医学预训练语言模型的情况下，达到了较好的结果，验证了模型在中文医学语料上的有效性。

5.3.2 消融实验

为了验证本文提出的BFIM以及注意力模块的有效性，基于$\text{BERT}_{\text{base}}$预训练模型在ACE05测试集上进行消融实验，结果如表6所示。当模型移除了交互模块后，由于两个任务无法进行有效的交互，实体和关系的效果均有下降。当移除注意力模块后，尽管两个任务的网络可以交互，但缺少学习自身任务特征的神经网络，与任务特征相关性较强的特征无法通过注意力网络进行强化，进而影响两个任务特征的质量。当两个模块全部移除后，模型性能下降最明显。综上，消融实验表明了本文提出的BFIM以及注意力模块能够显著提升模型效果，验证了PIAN模型的有效性。

5.4 PIAN在中文电子病历上的实验结果

5.4.1 CEMRIE数据集基准结果

将PIAN模型应用于中文电子病历上进行实体关系联合抽取，实体识别结果如表7所示，

[0]ACE04与ACE05语料来自于新闻文章、网上论坛等多种资源，共定义7种实体类型以及6种关系类型。
[1]CMeIE语料源于儿科及百种常见疾病训练语料，共定义11种医学实体类型以及44种关系类型。

实体类别	P(%)	R(%)	F1(%)
疾病	91.6	91.0	91.3
症状	81.5	82.7	82.1
部位	76.9	80.5	78.7
检查	94.5	93.1	93.8
治疗	86.9	84.3	85.6
All	**90.4**	**90.1**	**90.2**

Table 7: 实体各类别效果

实体类别	P(%)	R(%)	F1(%)
疾病-疾病	74.9	73.0	74.0
疾病-症状	77.5	77.0	77.3
疾病-部位	74.2	74.3	74.2
治疗-疾病	76.2	76.0	76.1
治疗-症状	70.3	72.4	71.3
检查-疾病	86.9	84.3	85.6
检查-症状	82.8	81.3	82.0
All	**80.8**	**80.6**	**80.7**

Table 8: 关系各类别效果

实体识别各个类别均取得了较好的效果，总体达到了90.2%的F1值。从结果可以看出，“症状”和“部位”类型的结果较差，其原因之一是“症状”类型实体与“疾病”类型实体在语义上较为相近，例如“偏头痛”属于疾病，但是容易与症状“头痛”混淆。“部位”类型的实体占总实体比例较少，模型无法充分学习其特征，若遇到复杂“部位”实体，如“输尿管跨越髂血管”、“肾盂输尿管连接部”等，则无法正确抽取。“疾病”、“治疗”和“检查”类型的实体由于特征较为明显且样本充足，抽取效果较好。

关系各类别的抽取结果如表8所示，可以看出关系抽取总体达到了80.7%的F1值。其中“疾病-症状”、“治疗-症状”以及“疾病-疾病”类别的效果较差，其原因是“症状”类型实体在语义上与“疾病”类型实体类似，模型不易区分，如果“疾病”(或“症状”)类型实体被错误预测为“症状”(或“疾病”)类型，关系抽取模型就会使用这些错误信息导致预测出错误的关系类型，所以与“症状”有关的关系类型效果较差。由于一部分“疾病”类型实体被预测为“症状”实体，因此会影响“疾病-疾病”类型关系效果。

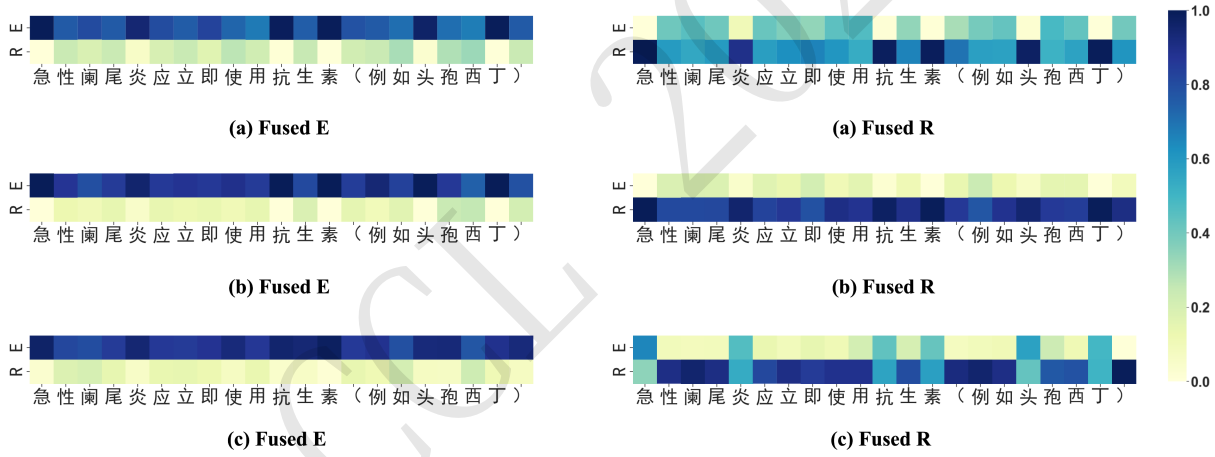


Figure 2: 双向特征交互模块中各个融合单元融合比例

5.4.2 双向特征交互模块分析

为了进一步探究BFIM的作用原理，将交互模块中的融合分数可视化，以更加直观地研究两个任务交互的具体信息。对于例句“急性阑尾炎应立即使用抗生素（例如头孢西丁）”，其中包含疾病实体“急性阑尾炎”、治疗实体“抗生素”和“头孢西丁”，其中疾病与两个药物实体存在“治疗-疾病”关系。

图2展示了模型中三个BFIM在生成融合特征时，两个任务特征融合的比例，其中Fused E/R表示融合后的特征，E/R表示分别来自实体任务和关系任务的特征比例。如图2.(a)所示，在第一个BFIM中，两个任务都捕获到了实体的头尾信息。如图2.(b)所示，第二个BFIM执行了较少的特征交互，这是因为在经过第一次交互后，每个任务都通过各自的注意力网络学习到了具有自身任务特点的特征，为了避免两种不同的特征混淆，该交互模块最大程度的保留了其原始的输入特征。如图2.(c)所示，在生成关系任务的融合特征时，来自实体识别任务的实体开始字符特征以较大比例融合到了对应的关系任务字符中，这表明实体的开始字符特征对关系

抽取任务具有重要作用；同时实体任务特征也融合了来自关系任务的特征。通过以上分析可知，BFIM能够从另一任务发现对自身任务有价值的信息，并通过自适应地学习一个最佳的融合特征比例实现充分的信息交互。

6 结论

本文提出了一种基于中文电子病历的平行交互注意力网络(PIAN)用于联合抽取实体及关系。为了充分利用实体识别和关系抽取两个任务的相关性，提出一个双向特征交互模块，其可以自适应地使一个任务中每个字符的特征与另一任务对应的特征动态融合，实现了双向细粒度的交互方式。模型在多个标准数据集上达到了最优结果，验证了本文方法的有效性。鉴于当前中文医学实体关系标注语料十分稀缺，与医学专家研讨并制定了标准的数据集标注规范，构建了中文电子病历实体关系抽取数据集CEMRIE并检验了数据集的质量，提供了基准结果。

参考文献

Bekoulis Giannis, Deleu Johannes, Demeester Thomas and Develder Chris. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34-45.

Christopoulou Fenia, Tran Thy Thy, Sahu Sunil Kumar, Miwa Makoto and Ananiadou Sophia. 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39-46.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. The automatic content extraction (ace) program-tasks, data, and evaluation. *Journal of biomedical informatics*, 45(5):57-45.

Devlin Jacob, Chang Ming-Wei, Lee Kenton and Toutanova Kristina. 2018. 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*, 4171-4186.

Doddington George R, Mitchell Alexis, Przybocki Mark A, Ramshaw Lance A, Strassel Stephanie M and Weischedel Ralph M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. *Lrec*, 2(1):837-840.

Fei Hao, Zhang Yue, Ren Yafeng and Ji Donghong. 2021. A span-graph neural model for overlapping entity relation extraction in biomedical texts. *Bioinformatics*, 37(11):1581-1589.

Guo Menghao, Liu Zhengning, Mu Taijiang and Hu Shimin. 2021. Beyond self-attention: External attention using two linear layers for visual tasks. *arXiv preprin:2105.02358*.

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

Lan Yinyu, He Shizhu, Liu Kang, Zeng Xiangrong, Liu Shengping and Zhao Jun. 2021. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. *BMC Medical Informatics and Decision Making*, 21(9):1-12.

Lan Zhenzhong, Chen Mingda, Goodman Sebastian, Gimpel Kevin, Sharma Piyush and Soricut Radu. 2019. Albert: a lite bert for self-supervised learning of language representations. *International Conference on Learning Representations*.

Lee Jinhyuk, Yoon Wonjin, Kim Sungdong, Kim Donghyeon, Kim Sunkyu, So Chan Ho and Kang Jaewoo. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234-1240.

Li Fei, Zhang Meishan, Fu Guohong and Ji Donghong. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):1-11.

Liu Shengyu, Tang Buzhou, Chen Qingcai and Wang Xiaolong. 2016. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016.

(c) 2022 中国中文信息学会计算语言学专业委员会

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

Luo Ling, Yang Zhihao, Yang Pei, Zhang Yin, Wang Lei, Lin Hongfei and Wang Jian. 2018. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381-1388.

Luo Ling, Yang Zhihao, Cao Mingyu, Wang Lei, Zhang Yin and Lin Hongfei. 2020. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of biomedical informatics*, 103:103384.

Miwa Makoto and Bansal Mohit. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1105-1116.

Park Chanhee, Park Jinuk and Park Sanghyun. 2020. Agcn: attention-based graph convolutional networks for drug-drug interaction extraction. *Expert Systems with Applications*, 159:113538.

Rei Marek, Crichton Gamal KO and Pyysalo Sampo. 2016. Attending to characters in neural sequence labeling models. *arXiv preprint arXiv:1611.04361*.

Ren Feiliang, Zhang Longhui, Yin Shujuan, Zhao Xiaofeng, Liu Shilei, Li Bochao and Liu Yaduo 2021. A novel global feature-oriented relational triple extraction model based on table filling. *Proceedings of the Empirical Methods in Natural Language Processing*, 2646-2656.

Sui Dianbo, Chen Yubo, Zhao Jun, Jia Yantao, Xie Yuantao and Sun Weijian. 2020. Feded: federated learning via ensemble distillation for medical relation extraction. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2118-2128.

Sun Cong, Yang Zhihao, Su Leilei, Wang Lei, Zhang Yin, Lin Hongfei and Wang Jian. 2020. Chemical–protein interaction extraction via Gaussian probability distribution and external biomedical knowledge. *Bioinformatics*, 36(15):4323-4330.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Lukasz and Polosukhin Illia. 2017. Attention is all you need. *Proceedings of the Conference and Workshop on Neural Information Processing Systems*, 5998-6008.

Wang Jue and Lu Wei. 2020. Two are better than one: joint entity and relation extraction with table-sequence encoders. *Proceedings of the Empirical Methods in Natural Language Processing*, 1706-1721.

Wang Yucheng, Yu Bowen, Zhang Yueyang, Liu Tingwen, Zhu Hongsong and Sun Limin. 2020. Tplinker: single-stage joint extraction of entities and relations through token pair linking. *Proceedings of the International Conference on Computational Linguistics*, 1572-1582.

Wang Yijun, Sun Changzhi, Wu Yuanbin, Zhou Hao, Li Lei and Yan Junchi. 2021. Unire: a unified label space for entity relation extraction. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 220-231.

王泽儒, 柳先辉. 2022. 基于指针级联标注的中文实体关系联合抽取方法. 武汉大学学报(理学版), 1-7.

Wei Zhepei, Su Jianlin, Wang Yue, Tian Yuan and Chang Yi. 2020. A novel cascade binary tagging framework for relational triple extraction. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1476-1488.

Yan Zhiheng, Zhang Chong, Fu Jinlan, Zhang Qi and Wei Zhongyu. 2021. A partition filter network for joint entity and relation extraction. *Proceedings of the Empirical Methods in Natural Language Processing*, 185-197.

Yi Zibo, Li Shasha, Yu Jie, Tan Yusong, Wu Qingbo, Yuan Hong and Wang Ting. 2017. Drug-drug interaction extraction via recurrent neural network with multiple attention layers. *International Conference on Advanced Data Mining and Applications*, 554-566.

Zeng Daojian, Liu Kang, Lai Siwei, Zhou Guangyou and Zhao Jun. 2014. Relation classification via convolutional deep neural network. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2335-2344.

(c) 2022 中国中文信息学会计算语言学专业委员会

Zeng Xiangrong, He Shizhu, Zeng Daojian, Liu Kang, Liu Shengping and Zhao Jun. 2019. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning. *Proceedings of the Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, 367-377.

Zhang Ningyu, Chen Mosha, Bi Zhen, Liang Xiaozhuan, Li Lei, Shang Xin, Yin Kangping, Tan Chuanqi, Xu Jian, Huang Fei and others. 2021. Cblue: a Chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Zheng Suncong, Wang Feng, Bao Hongyun, Hao Yuexing, Zhou Peng and Xu Bo. 2017. Joint extraction of entities and relations based on a novel tagging scheme. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1227-1236.

Zhong Zexuan and Chen Danqi. 2021. A frustratingly easy approach for entity and relation extraction. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 50-61.

CCL 2022

基于框架语义映射和类型感知的篇章事件抽取*

卢江[1]，李茹[1,2,*†]，苏雪峰[1,3]，闫智超[1]，陈加兴[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
{lujiangsxu, suexf, 15735104675, 18536201921}@163.com
{liru}@sxu.edu.cn

摘要

篇章事件抽取是从给定的文本中识别其事件类型和事件论元。目前篇章事件普遍存在数据稀疏和多值论元耦合的问题。基于此，本文将汉语框架网（CFN）与中文篇章事件建立映射，同时引入滑窗机制和触发词释义改善了事件检测的数据稀疏问题；使用基于类型感知标签的多事件分离策略缓解了论元耦合问题。为了提升模型的鲁棒性，进一步引入对抗训练。本文提出的方法在DuEE-Fin和CCKS2021数据集上实验结果显著优于现有方法。

关键词： 汉语框架网；框架语义映射；类型感知；事件抽取

Document-Level Event Extraction Based on Frame Semantic Mapping and Type Awareness

Jiang Lu[1], Ru Li[1,2,*†], Xuefeng Su[1,3], Zhichao Yan[1], Jiaxing Chen[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
{lujiangsxu, suexf, 15735104675, 18536201921}@163.com
{liru}@sxu.edu.cn

Abstract

Document-Level event extraction is the identification of its event type and event arguments from a given text. At present, the problems of sparse data and multi-event argument coupling are most in Document-Level events. Based on this, We map Chinese FrameNet with Chinese text events, and the sliding window mechanism and trigger word Paraphrase are introduced to improve the data sparseness problem of event detection. The use of a multi-event separation strategy based on type-aware labels alleviates the problem of meta-coupling. In order to improve the robustness of the model, adversarial training is further introduced. The experimental results of our approach on the DuEE-Fin and CCKS2021 datasets are significantly better than the existing methods.

Keywords: Chinese FrameNet , Frame Semantic Mapping , Type Awareness , Event Extraction

1 引言

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
*基金项目:山西省重点研发计划项目(NO.202102020101008);国家自然科学基金重点项目(NO.61936012) This work was supported by the Key Research and Development Proiect of Shanxi Pcovince (NO.202102020101008) and Key Natural Science Foundation of China (NO.61936012)
† 通讯作者Corresponding Author

事件抽取旨在从非结构化的文本中抽取结构化的事件信息，包含事件检测 (Li et al., 2020; Liao et al., 2021; Lin et al., 2019; Cao et al., 2021)和论元识别 (Aly et al., 2019; Chalkidis et al., 2019; Chang et al., 2020)两部分。根据事件抽取的粒度不同，可分为：句子级事件抽取（Sentence-level Event Extraction，SEE）和篇章级事件抽取（Document-level Event Extraction，DEE）。DEE任务难点主要在于文本的复杂和事件结构的复杂。文本的复杂体现在输入文本为篇章，这使得输入文本受限的模型需要在考虑篇章全局信息的情况下进行数据分割；事件结构复杂性体现在篇章中包含多事件，不同事件之间互有重叠，较单一事件抽取难度更大。

篇章事件抽取是从给定的篇章中完成对事件类型的识别以及论元实体的抽取。如图1所示，该篇章由多个句子组成，包含“质押”和“解除质押”两个事件，根据预先定义的事件类型和论元角色，完成事件论元表中的触发词以及论元实体的抽取。

触发词
论元

篇章

【S1】原标题：浙江康恩贝(6.500,0.01,0.15%)制药股份有限公司关于控股股东所持公司部分股份解除质押的公告。

【S2】2019年7月4日，浙江康恩贝制药股份有限公司（以下简称：本公司）接控股股东康恩贝集团有限公司书面通知，2019年7月2日康恩贝集团在中国证券登记结算有限责任公司上海分公司办理了股权质押解除手续，将其质押在杭州联合农村商业银行股份有限公司武林支行的本公司16，300，000股无限售流通股（占本公司总股本的0.61%）解除质押。

……

【S6】公司控股股东的一致行动人胡季强先生和浙江博康医药投资有限公司合计持有本公司股份240，438，148股，占本公司总股本的9.01%，其中已质押股份199，250，000股，占合计所持本公司股份的82.87%，占本公司总股本的7.47%。

【S7】综上，公司控股股东康恩贝集团和一致行动人合计持有本公司股份947，686，559股，占本公司总股本的35.53%，其中已累计质押所持本公司股份数量464，480，000股，占合计持有本公司股份的49.01%，占本公司总股份的17.41%。

【S8】浙江康恩贝制药股份有限公司，2019年7月5日。

事件论元表

触发词	事件类型	质押方	质押物	披露时间	质押股票/股份数量
质押	质押	胡季强	股份	2019年7月5日	464，480．000
触发词	事件类型	质押方	质押物	披露时间	质押股票/股份数量
解除质押	解除质押	康恩贝集团有限公司	无限售流通股	2019年7月2日	16，300，300

图 1: 篇章事件抽取示例

分析主流的篇章事件抽取数据集发现，DEE任务面临的挑战主要体现在两个方面：1）篇章数据稀疏。现有数据集人工标注难度较大，导致篇章中各事件类型呈现长尾分布2）篇章包含多事件。同一论元分布在不同事件中，存在多值论元耦合现象，也即论元重叠。因此本文探索引入汉语框架网（Chinese FrameNet，CFN）(Li et al., 2013)与事件建立一定的映射关系以改善数据稀疏问题；融合事件类型标签和篇章全局信息以缓解篇章多值论元耦合问题。

CFN是由山西大学以Fillmore的框架语义学为理论基础、同时参照FrameNet开发的汉语框架语义知识库。CFN包括框架库、词元库和例句库。现有1313个框架，21123个词元，94353条例句，这些句子是CCL（Center for Chinese Linguistics PKU）和BCC（BLCU Corpus Center）的真实语料 (Zhao et al., 2016)。

针对DEE中数据稀疏的问题。以往大多通过引入字典例句或百度百科等语料的方式进行简单的数据扩充，并没有针对性地考虑数据本身包含的事件和论元信息，而CFN中的例句来自真实语料，符合现实语义场景。CFN中定义的框架与中文事件抽取中的事件具有天然相似的结构：1）事件由事件触发词和一组论元组成。类似地，CFN中的框架由激活该框架的词元和一组框架元素构成。它们分别扮演着类似事件中的触发词和论元这一角色。2）中文篇章事

(c) 2022 中国中文信息学会计算语言学专业委员会

件通常包含多个触发词对应的多个事件，CFN中的“事件”也包含多个词元对应多个框架。此外，CFN中许多框架表达了某些具体类型的事件，表1中CFN框架与事件存在对应关系，分别为：[（‘出售’– >‘质押’），（‘盈亏’– >‘亏损’），（‘倒闭’– >‘破产’）]，同时CFN中的例句与篇章子句也有着相似的描述，表1中的相关事件句即为框架语义映射下的CFN例句。这促使我们探索是否存在CFN中的框架到事件的某种映射，以及是否可以通过使用CFN来改进事件检测。因此，本文在事件检测阶段根据事件和CFN相似的结构这一特点，建立两级框架语义映射[一级：‘框架’– >‘事件’；二级：‘词元’– >‘触发词’）]，同时通过对既有知识CFN框架-词元的引入，将CFN中与事件映射后的例句作为外部数据以此增强事件检测任务的性能。

框架-词元	事件-触发词	CFN中相关例句
出售-出让	质押-质押	作为交换条件，一是史玉柱**出让**80%的股份； 二是收购方出资5000～6000万元
盈亏-赔	亏损-损失	会计师事务所即使被处罚，也不过承担有限责任， 只要将公司注册时的资本**赔**完就行了
倒闭-关闭	企业破产-倒闭	企业亏损或不能维持下去而**关闭**，全体职工利益都要受到影响

表 1: CFN与中文事件的映射

针对DEE论元识别阶段多值论元耦合的问题，目前大多数模型采用联合抽取方法以便于更好地捕获事件检测和论元之间的交互，但却无法针对性的解决篇章中论元耦合现象。本文兼顾事件类型与论元的交互，在论元识别阶段把事件类型作为类别标签与篇章文本拼接共同作为输入特征，从而将篇章多事件转化为多个单事件，同时整合篇章全局信息增强模型建模表示。

本文的贡献主要有以下几点：1)将CFN与中文篇章事件抽取任务相结合，探索框架和事件之间的联系。2)在事件检测阶段，通过引入CFN这一外部知识，构造CFN框架与事件的两级映射，将映射后的CFN相关事件句作为外部数据改善数据稀疏问题。3)在论元识别阶段，通过融合事件类型标签以及整合全局上下文信息，将篇章多事件转化为多个单事件，在一定程度上缓解了多值论元耦合问题。

2 相关工作

现阶段根据抽取方式的不同，DEE任务分为管道式抽取（Pipeline extraction）(Chen et al., 2015)和联合(Joint extraction)抽取 (Cui et al., 2020)。Pipeline抽取将触发词作为事件的核心，(Yang et al., 2019; Liu et al., 2016)把事件检测和论元识别视作独立的多阶段分类任务。(Chen et al., 2015)提出基于深度学习的事件抽取模型DMCNN，该模型使用两个动态多池卷积神经网络进行触发词分类和论元分类。(Liu et al., 2016)针对英文事件抽取数据稀疏问题，第一次将FrameNet应用于事件检测任务取得了明显效果。以上方法通过引入预训练模型，在一定程度上丰富了文本语义，但针对中文篇章事件抽取任务，并未考虑引入外部知识，将事件检测和论元识别任务孤立地完成，缺少信息间的交互。联合抽取使用深度学习和联合学习进行特征交互。(Yang et al., 2018)针对篇章事件数据人工标注复杂度高、难度大的问题，提出了一个事件自动抽取框架DCFEE。(Zheng et al., 2019)针对篇章事件抽取中论元分散问题提出了Doc2EDAG模型，将篇章级的事件表填充任务转化为基于实体的有向无环图的路径扩展任务。(Xu et al., 2021)分别基于篇章级和多粒度的解码，提出了Git模型。大多数联合抽取采用图构建 (Yao et al., 2019)的方式捕获事件类型与论元之间的交互，避免了错误传播，但是针对篇章事件中多值论元耦合的问题并不能有效的解决。

事实上，事件检测和论元识别两阶段的上下文表示本质上捕获了不同的信息，然而联合抽取方式共享两者的模型结构和参数，会影响事件抽取整体性能。通过将CFN外部数据用于事件检测任务，可针对性的提升其实验效果，事件检测性能的提升也有助于论元识别的准确率，因此本文采用Pipeline抽取，相比于大多数Pipeline模型，更加注重事件类型与论元的交互。在事件检测阶段，引入框架语义映射，将CFN相关事件句作为外部数据在一定程度上改善了事件稀疏问题；在论元识别阶段，提出了基于类型感知的多事件分离策略，在一定程度上缓解了多值论元耦合问题。

(c) 2022 中国中文信息学会计算语言学专业委员会

3 篇章事件抽取模型

3.1 任务定义

篇章事件抽取任务可形式化地描述为：从给定的篇章中识别事件（事件检测任务）和事件中的论元（论元识别任务）。本文采用滑动窗口的方法将长篇章分割为多个片段，每个分段$s = \{c_1, c_2, c_3, ..., c_i\}$（$c_i$表示分段中第$i$个字符）包含多个事件$e$（$e \in E$，$E$表示事件类型总数），每个事件由一个触发词$t$和一组论元$\{a_1, a_2, ..., a_j\}$组成，$a_j$表示该事件中的第$j$个论元。

3.2 模型结构

针对篇章事件中存在的数据稀疏和多值论元耦合问题。本文从事件检测和论元识别两个方面解决事件抽取上述难点，模型如图2所示，主要包含以下4个模块。1）文本输入：事件检测阶段，将CFN映射后的事件句、触发词汉语释义信息以及滑窗切分的文本共同作为输入；论元识别阶段，将滑窗切分的文本作为输入。2）事件检测：对CFN事件句、片段文本、触发词释义进行编码，并根据事件中触发词对事件类型识别。3）论元识别：对输入的片段文本进行编码，并根据已知的事件类型对相应事件中的论元进行识别。4）文本输出：事件检测阶段输出事件类型，论元识别阶段输出各事件论元，将事件类型与论元进行组合形成完整的事件。

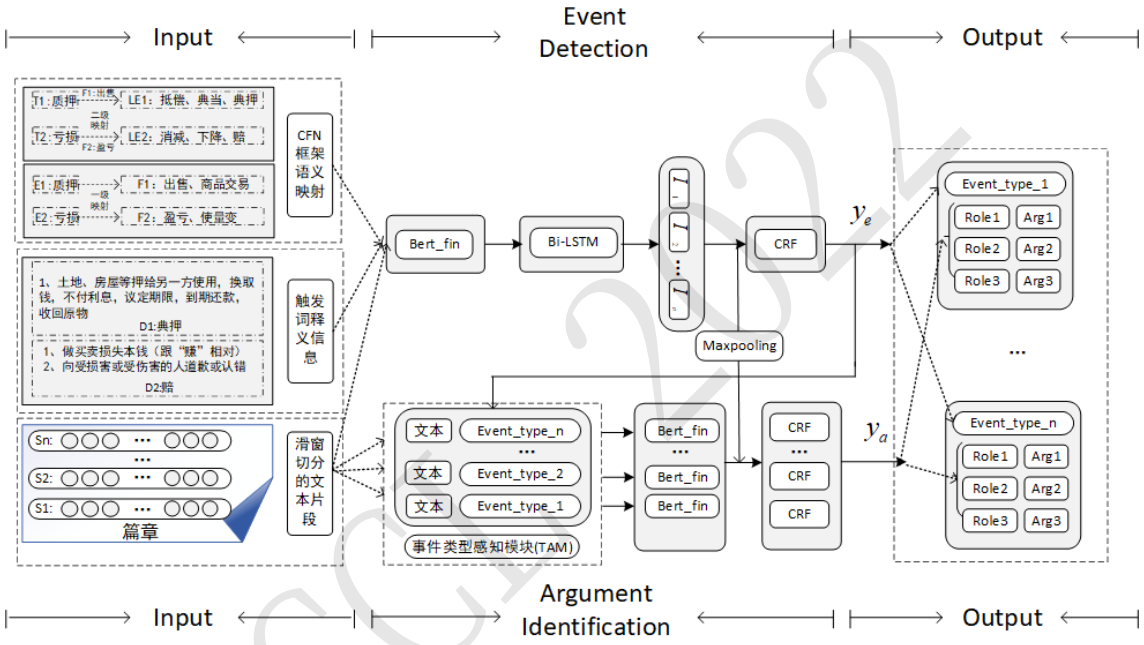


图 2: 篇章事件抽取模型总体图

3.3 事件检测

3.3.1 CFN与中文事件的映射

由于CFN中框架与事件结构高度相似，同时CFN中框架与词元有着天然的关联关系，因此本文在引入CFN这一外部知识时需遵循以下两点原则：

1）框架中的词元与中文事件中的触发词匹配一致时，倾向于表达相同类型的事件。

2）隶属于同一框架下的句子倾向于表达相同类型的事件。

基于以上两点原则，本文制定了框架与事件的两级映射，具体映射形式如下：

1）框架与事件的一级映射。如图3，CFN中的“出售”框架与“质押”事件表达的意义相近。

2）框架中的词元与事件触发词的二级映射。如图3，“出售”框架中的‘典押、典当、抵偿’词元对应于“质押”事件中的‘质押、出质、延期’触发词。

基于以上规则，CFN中的框架与事件形成二级映射，而后通过树查询的方式进行（框架-词元）与（事件-触发词）的相似度计算，取相似度前10的词元作为与触发词匹配的目标词元，并将中该词元下的事件句作为外部数据用以缓解数据稀疏问题，提升事件检测任务的性能。

第二十一届中国计算语言学大会论文集，第234页-第245页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

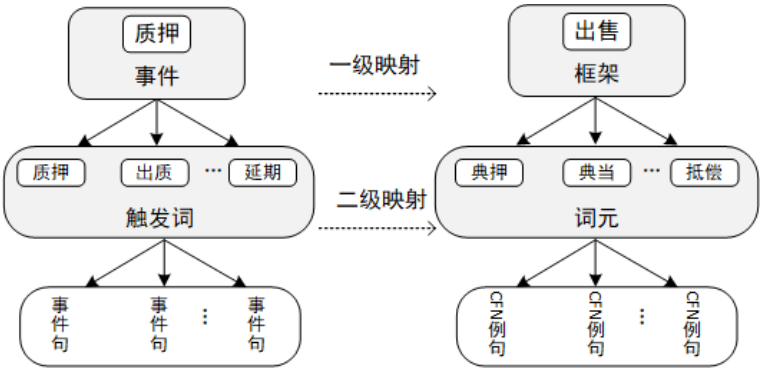


图 3: CFN框架在事件下的两级映射

3.3.2 序列建模层

为了充分感知上下文信息，本文采用滑窗机制将文本按字符分成多片段，每个分段$s = \{c_1, c_2, c_3, ..., c_i\}$作为事件检测任务的一次输入。同时为了增强触发词对事件类型的语义理解，将其汉语释义信息$t = \{t_1, t_2, ..., t_j\}$（$t_j$表示触发词释义的第$j$个字符）拼接到触发词之后，共同作为文本的输入。编码层采用BERT对文本编码，如公式1、2所示。

$$input = [CLS] + s + [SEP] + t + [SEP] \tag{1}$$

$$s' = BertEncoder(input) \tag{2}$$

其中，$s' = \{c_1, c_2, ..., c_{len}\}$，$s' \in R^{len \times d}$，$len$表示文本片段与释义信息的总长度，$d$为隐层维度。为了进一步增强篇章中的长实体信息的表示，在预训练模型之后加入bi_lstm进一步增强其语义表示，将前向$\vec{h}$和后向$\overleftarrow{h}$隐层向量拼接作为下一层的输入I，如公式3、4所示。

$$h = bi\_lstm(s') \tag{3}$$

$$I = concat([\vec{h}, \overleftarrow{h}]) \tag{4}$$

在解码阶段，使用CRF输出其最优得分序列，对于已知的输入序列$s' = (c_1, c_2, ..., c_n)$对应的输出标签结果为$y_e = (y_1, y_2, ..., y_n)$，定义当前序列得分如公式5所示，其中，$I_{i,yi}$表示第$i$个位置$soft\max$输出的概率。每个位置得分由隐层输出向量$I$和CRF转移矩阵$A$共同组成，最后利用$soft\max$计算归一化后的概率，采用最大对数似然函数优化目标函数，如公式6、7所示。

$$score(x, y_e) = \sum_{i=1}^{n} I_{i,yi} + \sum_{i=1}^{n+1} A_{yi-1,yi} \tag{5}$$

$$P(y_e|x) = \frac{\exp(score(x, y_e))}{\sum_{y'_e} \exp(score(x, y'_e))} \tag{6}$$

$$\log P(y_e|x) = score(x, y_e) - \ln(\sum_{y'_e} \exp(score(x, y'_e))) \tag{7}$$

3.4 论元识别

3.4.1 多值论元耦合

论元识别难点在于多事件混淆，从而导致不同事件之间论元耦合。如图4所示，篇章中包含三个事件，分别是：事件1：‘高管变动’；事件2：’回购’；事件3：‘高管变动’。事件1和事件3触发词‘离职’相同，但是论元不尽相同，事件2中的论元“宁波美诺华药业股份”同时属于事件1和事件3的论元。因此由于篇章多事件，导致同一论元出现在不同事件中。

第二十一届中国计算语言学大会论文集，第234页-第245页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

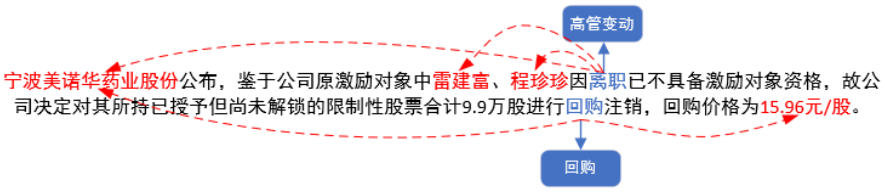


图 4: 多值论元耦合

3.4.2 多类型感知层

针对多值论元耦合这一难点，本阶段设置了一个类型感知模块，其核心在于将事件检测阶段感知的每个事件类别标签$e = \{e_1, e_2, ..., e_j\}$（$e_j$表示当前事件类型的第$j$个字符）分别与原始文本片段$s = \{c_1, c_2, c_3, ..., c_i\}$拼接共同作为模型输入，依据不同的类型标签对文本中多事件进行分离，从而将多事件论元识别转化为多个单事件论元识别，编码层输入如公式8、9所示:

$$input = [CLS] + s + [SEP] + e + [SEP] \tag{8}$$

$$s' = BertEncoder(input) \tag{9}$$

其中，$s' = \{c_1, c_2, ..., c_{len}\}$，$s' \in R^{len \times d}$，$len$表示文本片段与类型标签的总长度，$d$为隐层维度。为了进一步增强事件检测与论元识别的交互，将事件检测中篇章全局信息经过$\max pooling$得到表示$\hat{I}$，之后与文本片段的编码s'进行融合，进一步增强文本信息表示。如公式10、11所示，I_n表示滑窗切分后的第n个片段的编码，s''表示融合篇章全局信息的隐层表示。

$$\hat{I} = \max pooling\{I_1, I_2, ..., I_n\} \tag{10}$$

$$s'' = \hat{I} \oplus s' \tag{11}$$

3.4.3 多标签解码层

本阶段将类型标签分离后的单事件经过各自的CRF解码层，由原先对多事件论元角色分类任务转化为多个单事件论元角色分类任务。CRF解码与事件检测阶段一致，如公式12、13所示，$y^*_{a_i}$表示第i个事件经过CRF解码之后的标签序列，y^*_a表示所有事件论元标签序列的集合。

$$y^*_{a_i} = CRF(s'') \tag{12}$$

$$y^*_a = \{y^*_{a_1}, y^*_{a_2}, ..., y^*_{a_n}\} \tag{13}$$

为解决正负样本不均衡现象，采用$Focalloss$损失函数，如公式14所示，P表示当前序列得分经过softmax后的归一化概率值，α为正负样本比例平衡因子，γ为样本难度平衡因子。

$$L_{fl} = \begin{cases} -\alpha(1-p)^\gamma \log p, p^* = 1 \\ -(1-\alpha)p^\gamma \log(1-p), p^* = 0 \end{cases} \tag{14}$$

3.5 文本输出

由于本文将事件检测预测的事件类型作为类别标签融入论元识别文本输入中，因此本阶段只需将事件检测预测全部事件类型与论元识别预测的所有论元标签进行拼接。公式如下所示，y^*_e表示事件检测模型的预测标签，y_e表示事件检测模型的真实标签。y^*_a表示论元识别模型的预测标签，y_a表示论元识别模型的真实标签。

$$y^*_e = \arg\max_{y_e} f(y_e|d, t) \tag{15}$$

$$y^*_a = \arg\max_{y_a} f(y_a|d, e, r) \tag{16}$$

$$y^* = y^*_e \oplus y^*_a \tag{17}$$

第二十一届中国计算语言学大会论文集，第234页-第245页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

3.6 对抗训练

为增强模型的鲁棒性，在论元识别训练阶段加入对抗训练 (Aleksander et al., 2018)（adversarial training）对编码层的Embeedding参数矩阵混合一些微小扰动，让模型自适应这种改变，公式如下，D代表训练数据，x表示数据输入序列，y表示样本的预测标签，θ表示模型参数，Δx表示在训练样本中增加的扰动，$L(x,y;\theta)$表示样本的$loss$，$\nabla xL(x,y;\theta)$表示$loss$对x的梯度，ζ表示常量，Ω是扰动空间。$L(x+\Delta x,y;\theta)$表示在训练数据x上增加扰动Δx，内层公式在于寻找损失函数最大的扰动，外层公式是最优化求解的最小化公式。

$$\min_{\theta} E_{(x,y)-D}[\max_{\Delta x\in\Omega} L(x+\Delta x,y;\theta)] \tag{18}$$

$$\Delta x=\zeta sign(\nabla xL(x,y;\theta)) \tag{19}$$

4 实验

4.1 数据集

4.1.1 数据集介绍

本文使用DuEE-fin以及CCKS2021篇章事件抽取数据集进行实验。1)DuEE-fin数据集 (Liu et al., 2020)。百度最新发布的首个包含触发词信息的金融篇章事件抽取数据集，包含13种事件类型和92种论元角色，共1.17万个篇章；2) CCKS2021篇章事件抽取数据集。包含13种论元角色的5000个篇章。DuEE-fin和CCKS-2021训练集、验证集、测试集统计如表2所示。

数据集	训练集	验证集	测试集
DuEE-fin	7047	1174	60000
CCKS-2021	4000	500	500

表 2: 数据集分布

4.1.2 数据稀疏分析

DuEE-fin数据集未公开测试集真实标签。因此本文对训练集和测试集中8221篇金融篇章进行分析，从图5、图6中可以看出，Duee-Fin数据集存在数据稀疏问题，13种不同的事件类型文本呈现长尾分布，以训练集为例，‘股份回购’事件占比13.8%，‘被约谈’事件仅占比1.8%，因此各事件类型分布稀疏且极不均匀。

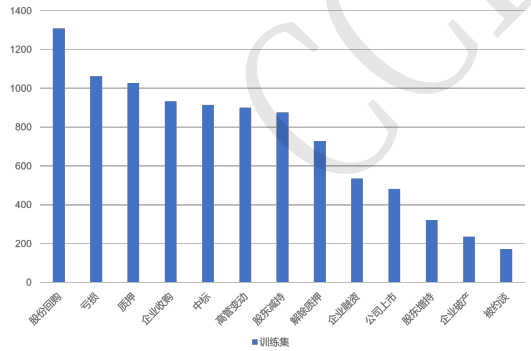


图 5: DuEE-fin训练集各事件类型分布

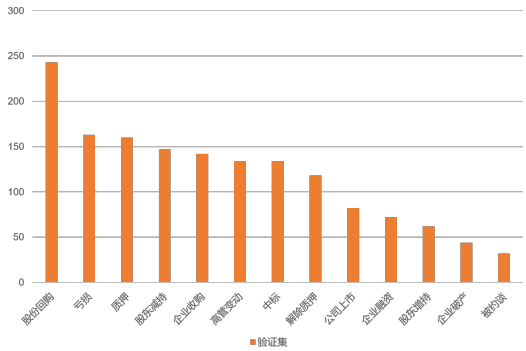


图 6: DuEE-fin验证集各事件类型分布

4.2 参数设置及评价指标

本文使用金融领域预训练模型FinBERT。由于文本篇幅较长，采用滑窗机制分片段输入。为了综合考虑模型的训练效率，滑窗大小为400，步长为200，与 (Liu et al., 2021)实验保持一致，实验参数如下，epoch为50，learning_rate为5e-5，使用实验的环境为32G显存的Tesla V100，采用Adam优化器对模型进行优化。实验采用准确率(Accuracy)、召回率（Recall）、F1值评价指标衡量模型性能。

第二十一届中国计算语言学大会论文集，第234页-第245页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

4.3 实验结果

4.3.1 事件检测对比实验

为了验证本文提出方法的有效性，在DuEE-fin公开数据集上将本文的方法与以下五种方法进行了对比。模型具体如下：（1）BERT (Devlin et al., 2019)；（2）BERT-wwm (Y.Cui et al., 2018)；（3）Ernie (Yu et al., 2019)；（4）FinBERT (Liu et al., 2020)；（5）BERT-BiLSTM-CRF：主流事件抽取模型。由于官方未给出测试集的真实标签，只允许在线提交事件抽取最终结果，因此本文事件检测在1174条验证集篇章上进行实验。

Model	Dev
BERT	90.15
BERT-wwm	90.19
Ernie(baseline)	**89.94**
FinBERT	90.42
BERT-BiLSTM-CRF	**91.32**
Ours	**93.83**

表 3: 事件检测在DuEE-fin上实验结果

事件类型	触发词/CFN词元/CFN例句	Ernie(baseline)	Ours	Δ
公司上市	15/14/383	87.50	93.43	+5.93
股东减持	7/18/183	90.35	91.86	+1.15
股东增持	6/12/229	86.42	88.50	+2.08
企业收购	16/35/675	91.26	94.83	+3.57
企业融资	18/2/25	87.15	88.95	+1.80
股份回购	6/4/38	94.62	96.24	+1.62
质押	6/4/58	93.04	94.66	+1.62
解除质押	15/4/58	89.42	89.52	+0.10
企业破产	3/25/269	84.60	89.46	+4.86
亏损	11/13/165	93.21	94.32	+1.11
被约谈	3/48/950	83.65	93.85	**+10.20**
中标	5/2/8	91.84	93.25	+1.41
高管变动	74/29/1003	89.52	96.20	+6.68
overall	**185/206/3986**	89.94	93.83	**+3.89**

表 4: DuEE-fin上各事件类型实验结果

从表3中可以看出，在验证集上较基线提高了3.89%，较当前主流模型提高了2.51%，从表4中可以看出，13种事件类型的触发词共计185个，经过CFN与DuEE-fin事件映射后引入CFN词元206个，词元下的CFN例句3986条，证明了事件与框架存在较强的相关性；各事件类型识别效果均有一定的提升，其中“被约谈”事件提升了10.2%，证明了通过框架语义映射引入CFN这一外部知识对事件检测任务性能提升有积极作用。

4.3.2 论元识别对比实验

为了验证本文提出方法的有效性，在DuEE-fin公开数据集上将本文的方法与以下8种方法进行了对比。模型具体如下：（1）BERT (Devlin et al., 2019)；（2）Ernie (Yu et al., 2019)；(3) BERT-BiLSTM-CRF，主流事件抽取模型：（4）DCFEE (Yang et al., 2018)，一种篇章级中文金融事件抽取模型：（5）Doc2EDAG (Zheng et al., 2019)，一种端到端的基于有向无环图的预训练模型；（6）Greedy-Dec，Doc2EDAG模型提出的基线；（7）GIT (Xu et al., 2021)，一种基于transformer结构的篇章级和多粒度的事件抽取模型；（8）PTPCG (Tong et al., 2021)，采用非自回归解码算法进行图剪枝的快速轻量级模型。由于未给出测试集的真实

标签，因此本文通过在线提交预测文件得到测试集上的实验结果。在论元识别阶段，本文的模型与以下9种模型进行对比,分别对应3种Pipeline抽取模型和6种当前最新的联合抽取模型。

	Model	Dev	Online Test
pipeline抽取	BERT	55.64	44.93
	Erine(baseline)	55.13	45.60
	BERT-BiLSTM-CRF	**65.46**	**54.20**
联合抽取	DCFEE-O	44.52	41.33
	DCFEE-M	37.63	36.12
	Greedy-Dec	47.35	40.71
	Doc2EDAG	66.10	58.13
	Git	**67.71**	55.62
	PTPCG	66.42	**60.15**
	Ours	**69.82**	**64.50**

表 5: 论元识别在DuEE-fin上实验结果

从表5的数据中可以看出，在Duee-fin验证集和在线测试集上实验效果与其他Pipeline模型相比F1值分别提升了4.36%和10.3%，与当前最新的联合模型相比F1值分别提升了2.11%和4.35%，验证了本文实验方法的有效性。证明了基于事件检测的类型感知策略和融合篇章全局信息对解决多值论元耦合有明显效果。

4.4 消融实验

为了验证不同模块的有效性，本文分别对CFN框架映射模块、类型感知模块（Type Aware Module）以及对抗训练（FGM）消融去掉，来检测模型性能的变化，具体结果如表6所示。

Model	Online Test
Ours	64.50
-CFN	62.35
-TAM	63.26
-FGM	63.72
-ALL	60.33

表 6: 消融实验的结果

Model	CCKS-2021 Test
BERT	60.58
ERNIE	**60.63**
Doc2EDAG	68.35
Git	68.74
Ours	**70.25**

表 7: 模型在CCKS-2021数据集实验结果

表6的数据可以看出:

（1）CFN框架映射模块、类型感知模块以及对抗训练对事件抽取任务均有明显效果提升。其中，CFN框架映射模块消融去掉后，模型性能下降较大，我们认为：1）CFN有助于提升事件检测子任务性能；2）事件检测性能的提升有助于增强后续论元识别子任务的准确率。这充分说明了CFN框架语义映射对模型效果有积极作用。

（2）相比于模型-CFN，模型-TAM性能下降较大，但也取得有效的提升，这说明类型感知的多事件分离方法对于解决篇章多值论元耦合也有一定的效果。

（3）模型-FGM性能有所下降，这说明对抗训练之后，增强了模型的鲁棒性。

4.5 模型泛化能力实验

为了验证本文方法的泛化性，选取了CCKS-2021面向金融领域篇章要素抽取评测数据集进行实验。由于CCKS评测任务主办方未公开测试集，本文只能在训练集与验证集上进行测试。本文将原始验证集按照1:1的比例进行随机等比例切分，形成新的验证集与测试集,数据划分比例见表2所示，并选取了目前主流的事件抽取模型进行对比实验，实验结果如表7所示，可以看出：本文方法相比于基线提升了9.62%。这充分说明了本文提出的模型在篇章事件抽取任务中有较好的泛化能力。

4.6 案例分析

为了验证本文方法的有效性，进一步对论元识别模型的预测数据进行了分析，具体细节如图7所示。文本中包含同一事件类型“高管变动”的两个事件。左侧是本文模型的预测结果。

(c) 2022 中国中文信息学会计算语言学专业委员会

右侧是ERNIE基线预测结果。可以看到“高管变动”这一事件类型对应‘杜秋龙’和‘陈体引’两个事件，这两个事件类型下的‘任职公司’、‘变动类型’论元角色分别对应‘卧龙地产集团股份有限公司’和‘辞职’两个论元发生了论元耦合。本文模型通过类型感知标签将两个事件分别进行论元识别，避免了两个事件之间相同论元重叠的问题。而基线模型由于没有对多事件进行区分，因此将两个事件预测为一个事件，这大大降低其论元识别的准确率。因此本文模型经过类型感知的事件分离方法可以较好的区分同一论元在不同事件中的耦合问题。

篇章

[S1]卧龙地产集团股份有限公司（以下简称“公司”）近期收到公司董事杜秋龙先生、监事陈体引先生提交的书面辞职报告

[S2]杜秋龙先生因个人原因，申请辞去公司董事、董事会提名委员会委员及副总经理的职务；陈体引先生因年龄原因，申请辞去公司监事会主席的职务

事件类型	变动类型	高管姓名	高管职位	高管职位	任职公司
高管变动	辞职	杜秋龙	公司董事	董事会提名委员会及副总经理	卧龙地产集团股份有限公司
事件类型	变动类型	高管姓名	高管职位	高管职位	任职公司
高管变动	辞职	陈体引	监事	公司监事会主席	卧龙地产集团股份有限公司

本文模型预测事件结果

事件类型	变动类型	高管姓名	高管姓名	高管职位	高管职位	任职公司
高管变动	辞职	杜秋龙	陈体引	公司董事会主席	董事会提名委员会及副总经理	卧龙地产集团股份有限公司

ERNIE 模型预测事件结果

图 7: 不同事件之间论元耦合案例分析对比

5 结论

本文围绕篇章事件数据稀疏和事件多值论元耦合的问题展开研究。提出了基于框架语义映射和类型感知的篇章事件抽取模型。该模型在事件检测阶段，通过引入CFN框架语义并制定了三类映射约束，建立了框架与事件的两级映射，将CFN相关事件句作为外部数据,改善了数据稀疏问题。在论元识别阶段，通过设计类型感知模块将多事件分离成单事件进行论元识别，缓解了多值论元耦合问题，为了增强事件检测与论元抽取的交互，将事件检测的全局向量信息与论元识别隐层向量融合，进一步丰富了模型的语义表示。在提升篇章多事件抽取性能的同时，对大量实验结果分析发现，不同的窗口大小和步长对实验也有一些影响，后续可以对滑动窗口参数做进一步分析,此外篇章中仍存在很多跨句论元分散的问题，如何设计跨句论元抽取框架是后续的研究重点。

参考文献

Li, Zhihui and Chang, Xiaojun and Yao, Lina and Pan, Shirui and Zongyuan, Ge and Zhang, Huaxiang. 2020. *Grounding Visual Concepts for Zero-Shot Event Detection and Event Captioning.* Association for Computing Machinery, pages:297–305.

Liao, Jinzhi and Zhao, Xiang and Li, Xinyi and Zhang, Lingling and Tang, Jiuyang. 2021. *Learning Discriminative Neural Representations for Event Detection.* Association for Computing Machinery, pages:644–653.

Lin, Hongyu and Lu, Yaojie and Han, Xianpei and Sun, Le 2019. *Cost-sensitive Regularization for Label Confusion-aware Event Detection.* Association for Computational Linguistics, pages:5278–5283.

(c) 2022 中国中文信息学会计算语言学专业委员会

Cao, Yuwei and Peng, Hao and Wu, Jia and Dou, Yingtong and Li, Jianxin and Yu, Philip S. 2021. *Knowledge-Preserving Incremental Social Event Detection via Heterogeneous GNNs.* Association for Computing Machinery, pages:3383–3395.

Aly, Rami and Remus, Steffen and Biemann, Chris. 2019. *Hierarchical Multi-label Classification of Text with Capsule Networks.* Association for Computational Linguistics, pages:323–330.

Chalkidis, Ilias and Fergadiotis, Emmanouil and Malakasiotis, Prodromos and Androutsopoulos, Ion 2019. *Large-Scale Multi-Label Text Classification on EU Legislation.* Association for Computational Linguistics, pages:6314–6322.

Chang, Wei-Cheng and Yu, Hsiang-Fu and Zhong, Kai and Yang, Yiming and Dhillon, Inderjit S. 2020. *Taming Pretrained Transformers for Extreme Multi-Label Text Classification.* Association for Computing Machinery, pages:3163–3171

李茹, 王智强, 李双红, 梁吉业, Collin Baker. 2013. 基于框架语义分析的汉语句子相似度计算. 计算机研究与发展, pages:1728–1736.

赵红燕, 李茹, 张晟, 张力文. 2016. 基于*DNN*的汉语框架识别研究. 中文信息学报, 30(6):75-83.

Chen, Yubo and Xu, Liheng and Liu, Kang and Zeng, Daojian and Zhao, Jun. 2015. *Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks.* Association for Computational Linguistics, pages:167–176.

Cui, Shiyao and Yu, Bowen and Liu, Tingwen and Zhang, Zhenyu and Wang, Xuebin and Shi, Jinqiao. 2020. *Edge-Enhanced Graph Convolution Networks for Event Detection with Syntactic Relation.* Association for Computational Linguistics: EMNLP 2020, pages:2329–2339.

Yang, Sen and Feng, Dawei and Qiao, Linbo and Kan, Zhigang and Li, Dongsheng. 2019. *Exploring Pre-trained Language Models for Event Extraction and Generation.* Association for Computational Linguistics, pages:5284–5294.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina 2019. *Pre-training of Deep Bidirectional Transformers for Language Understanding.* Association for Computational Linguistics, pages:4171–4186.

Liu, Shulin and Chen, Yubo and He, Shizhu and Liu, Kang and Zhao, Jun 2016. *Leveraging FrameNet to Improve Automatic Event Detection.* Association for Computational Linguistics", pages:2134–2143.

Yao, Liang and Mao, Chengsheng and Luo, Yuan. 2019. *Graph Convolutional Networks for Text Classification.* AAAI Press, numpages:8.

Yang, Hang and Chen, Yubo and Liu, Kang and Xiao, Yang and Zhao, Jun. 2018. *A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data.* Association for Computational Linguistics, pages:50–55.

Zheng, Shun and Cao, Wei and Xu, Wei and Bian, Jiang. 2019. *Doc2EDAG:An End-to-End Document-level Framework for Chinese Financial Event Extraction.* 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.

Xu, Runxin and Liu, Tianyu and Li, Lei and Chang, Baobao. 2021. *Document-level Event Extraction via Heterogeneous Graph-based Interaction Model with a Tracker.* Association for Computational Linguistics, pages:3533–3546.

Aleksander Madry and Aleksandar Makelov and Ludwig Schmidt and Dimitris Tsipras and Adrian Vladu. 2018. *Towards Deep Learning Models Resistant to Adversarial Attacks.* ArXiv, volume:abs/1706.06083.

Liu, Z. and Huang, D. and Huang, K. and Li, Z. and Zhao, J. 2020. *FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining.* International Joint Conference on Artificial Intelligence,

Li, and X, et al. 2021. *DuEE-fin: a document-level event extraction dataset in the financial domain released by Baidu.* Retrieved from https://aistudio.baidu.com/aistudio/competition/detail/46,

(c) 2022 中国中文信息学会计算语言学专业委员会

Y. Cui, W. Che, T. Liu, B. Qin and Z. Yang. 2018. *Pre-Training With Whole Word Masking for Chinese BERT*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pages:3504-3514

Yu Sun and Shuohuan Wang and Yukun Li and Shikun Feng and Xuyi Chen and Han Zhang and Xin Tian and Danxiang Zhu and Hao Tian and Hua Wu. 2019. *ERNIE: Enhanced Representation through Knowledge Integration*. ArXiv, volume:abs/1904.09223

Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, Min Zhang. 2021. *Efficient Document-level Event Extraction via Pseudo-Trigger-aware Pruned Complete Graph*. https://arxiv.org/pdf/2112.06013.pdf.

Liu, Yaduo and Zhang, Longhui and Yin, Shujuan and Zhao, Xiaofeng and Ren, Feiliang. 2021. *An Effective System for Multi-Format Information Extraction*. Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II pages = 460–471.

CCL 2022

期货领域知识图谱构建*

李雯昕[1,2]，眷红英[1]，关同峰[1,3]，韩英杰[1]
1.郑州大学计算机与人工智能学院，河南郑州
2.中移在线服务有限公司，河南郑州　3.中原银行，河南郑州
wency.li@foxmail.com; iehyzan@zzu.edu.cn;
guantf.gtf@foxmail.com; ieyjhan@zzu.edu.cn

摘要

期货领域是数据最丰富的领域之一，本文以商品期货的研究报告为数据来源构建了期货领域知识图谱（Commodity Futures Knowledge Graph，CFKG）。以期货产品为核心，确立了概念分类体系及关系描述体系，形成图谱的概念层；在MHS-BIA与GPN模型的基础上，通过领域专家指导对242万字的研报文本进行标注与校对，形成了CFKG数据层，并设计了可视化查询系统。所构建的CFKG包含17,003个农产品期货关系三元组、13,703种非农产品期货关系三元组，为期货领域文本分析、舆情监控和推理决策等应用提供知识支持。

关键词： 知识图谱；命名实体识别；实体关系抽取；期货文本

Construction of Knowledge Graph in Futures Field

Wenxin Li [1,2], Hongying Zan[1], Tongfeng Guan[1,3], Yingjie Han[1]
1.School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, Henan, China
2.China Mobile Online Services, Zhengzhou, Henan, China　3.Zhongyuan Bank, Zhengzhou, Henan, China
wency.li@foxmail.com; iehyzan@zzu.edu.cn;
guantf.gtf@foxmail.com; ieyjhan@zzu.edu.cn

Abstract

The futures field is one of the fields with the most abundant data.This paper used the Research Report of commodity futures as the data source to construct the Commodity Futures Knowledge Graph (CFKG). Taking futures products as the core, the concept classification system and the relationship description system were established, thus the CFKG conceptual layer was also built.Based on MHS-BIA and GPN model, more than 2.42 million words of research reports were annotated and proofread under the guidance of field experts, forming the CFKG data layer,and designed a visual query system of the knowledge graph. The futures domain knowledge graph constructed in this thesis containing 17,003 agricultural product relation triples and 13,703 non-agricultural product relation triples, which provided structured knowledge support for text analysis, public opinion monitoring, reasoning and decision-making in the futures field.

Keywords: Knowledge Graph , Named Entity Recognition , Relation Extraction , Futures Text

期货市场智能化舆情分析研究项目（20200464A）

第二十一届中国计算语言学大会论文集，第246页-第256页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

期货行业是个数据驱动的行业，随着期货市场逐渐扩大，期货领域相关数据和类型不断增多，从大数据中挖掘有价值的信息并应用于期货领域是必然趋势。该领域数据具有数据信息密集、数据量庞大、数据种类多样等特征，传统业务对行情、咨讯数据已经形成了高度依赖。但是，也存在数据利用效率低，对现有的数据资源价值的挖掘、分析和利用能力较弱的问题。因此，使用实体识别和关系抽取技术将实体、属性、关系等从非结构化、半结构化数据中抽取出来并建立关联，从内容分散、多元异构的期货文本中挖掘重要信息形成期货知识图谱，有助于金融从业者高效获取信息，为期货领域文本分析、舆情监控等关键技术提供坚实的基础，促进期货领域智能化发展。

知识图谱可提供一种更好的组织和理解信息的能力，2012年5月Google正式提出知识图谱的概念。发布基于维基百科(Vrandečić and Krötzsch, 2014)、Freebase(Bollacker et al., 2008)的知识图谱。根据数据源可将知识图谱划分为通用和领域知识图谱，通用知识图谱以百科类网站为数据源，面向公共领域，如跨语言百科知识图谱XLore(Wang et al., 2013)，中文模式知识库Zhishi.me(Niu et al., 2011)等。领域知识图谱依托特定专业领域的数据进行构建，在知识图谱构建过程中依赖领域专家或工程师依据项目的具体应用背景和需求进行图谱规则的制定，追求领域知识的专业性和准确性，如医学领域(奥德玛et al., 2019)、电商领域(Xu et al., 2021)和旅游领域(Kärle et al., 2018)等。将知识图谱应用于期货领域，可链接多数据源，形成商品和用户的知识描述体系，从而让商家和用户更直观地了解、认识、分析用户群体和商品。虽然目前学术界对于该领域知识图谱的研究还不多，就相关领域而言，爱智慧科技有限公司构建产业链图谱(Chen et al., 2021)，图谱包括有色金属和非金属材料等产品的行业结构与产品上下游信息。

期货行业除了对期货涉及的交易品种、合约、期货公司等实体外，还需对期货相关实体更大粒度的结构化标签如宏观经济主题、交易品种类别等进行识别，对期货相关实体细粒度的结构化标签等也需要进行识别。因此，以上相关研究无法满足期货领域知识图谱分析需求。由于期货领域相关企业内部的信息属于企业机密，可供研究的高质量公开数据较少，并且期货数据更新速度快、信息密集，无法及时进行知识更新及挖掘补充。期货市场尚未有成熟的信息检索产品进行大规模投放，相关研究正处于探索阶段，说明研究期货领域知识图谱的构建方法的重要性和迫切性。针对以上问题，本文针对期货领域的特点，针对商品期货展开研究，采用实体识别和关系抽取技术从非结构化期货文本中抽取出实体和关系，构建期货领域知识图谱（Commodity Futures Knowledge Graph，CFKG）。

2 CFKG构建流程

知识图谱从逻辑结构上可划分为概念层和数据层两个部分。概念层对实体概念和关系分类体系进行建模，对数据层进行约束。数据层通过实体关系三元组对概念层各类知识的定义进行表达。

CFKG构建过程如图1所示，首先设计概念层，即制定相应的知识描述体系。根据期货公司发布的期货研报中收集期货领域语料与术语集合，通过示例标注与分析设计实体概念体系和关系分类体系，经领域专家评估后形成知识图谱描述体系。以概念层为基础，获取多来源期货领域文本，采用半自动的方式对实体及实体关系进行标注构建语料库（Corpus for Entity and Relation annotation in Futures domain，CERF）。并展开期货实体关系联合抽取模型的研究，采用自动方式实现知识的自动更新，CERF语料库与模型自动抽取的实体及关系三元组经数据整合后形成CFKG的数据层。

3 CFKG概念层构建

本文根据期货产品的特点将其划分为农产品期货和非农产品期货两种，分别构建概念分类描述体系和关系分类描述体系。CFKG以期货产品作为描述主体，通过建立与产品相关的价格、地名、质量指标等概念之间的联系，形成各类概念之间的网状关系。CFKG中定义了产品、企业、价格、地名、价格指标等20类农产品期货实体与14类非农产品期货实体，对期货产

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

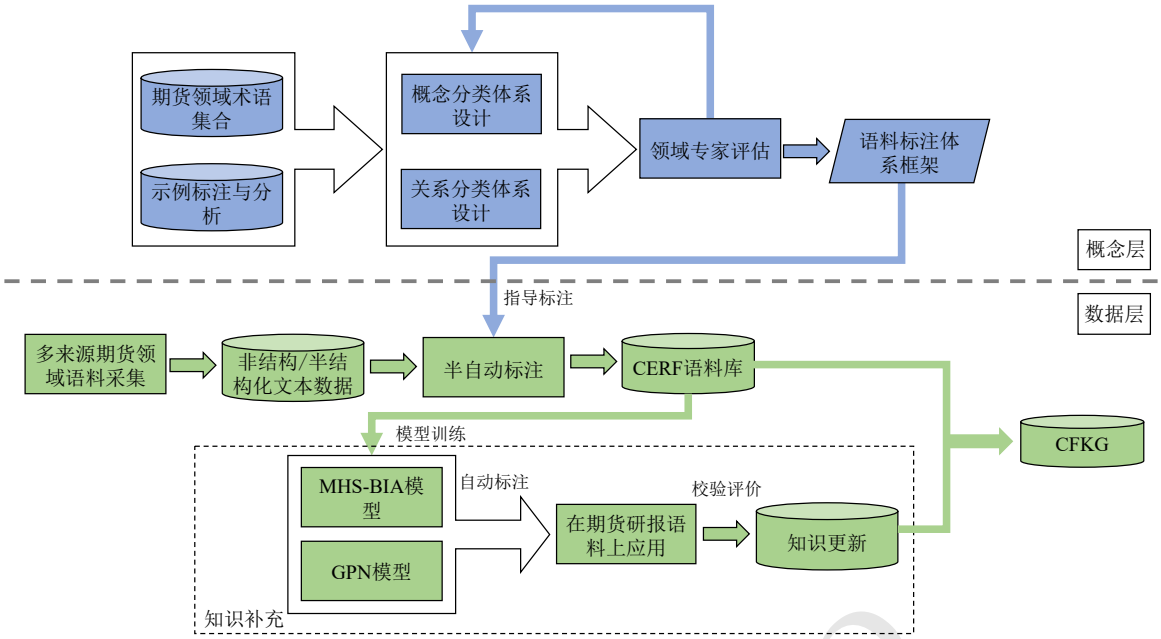


图 1. CFKG构建流程图

品研究报告中各类产品描述进行考察，在概念分类体系基础上进一步设计实体间关系描述体系，以产品作为关系描述中心拓展与其他期货实体间的关系。农产品期货包含9类关系类型，包括42种子关系；非农产品期货包含11类关系类型，包含36种子关系。由于篇幅所限，表1与表2仅展示了CFKG中农产品期货概念分类体系及关系分类体系，其余子关系描述体系不再一一详述。

关系类型	子关系	关系定义
产品-产品	同义	包括别名、简称、英文缩写以及同义词
	替代	在生产、流通或消费环节的可代替品
	下游	指以该产品为原材料生产或加工的产品
	AKO	A kind of：泛指一个实体是另一个实体中的一类

表 1. CFKG部分标注关系定义

4 CFKG数据层构建

CFKG数据层构建主要包括期货领域语料采集、半自动标注及模型自动抽取三部分。

4.1 领域数据采集

期货领域缺少专用的行业规范资源，如专用术语标准、行业指南等可供参考，因此需要从多方面获取期货文本资源，形成期货领域知识语料库。中文期货领域文本包括期货品种简介、期货新闻、期货研报等网络资源，不同来源的数据在文本内容上存在差异，如期货研报是我国期货业内专家针对不同期货品种和行业动态做出的分析和研究；期货品种简介以归纳方式总结各类期货品种的特点；期货新闻侧重于期货关联商品行业动态和发展状况。CFKG以产品为核心，抽取多来源文本中的相关内容，期货领域数据来源如表2所示，其中“半”表示半结构化数据，“非”表示非结构化数据。

通过对比分析数据收集阶段获得的语料，期货研报相比新闻文本数据中包含对不同期货品种的行业动态、产业链信息做出分析和研究，且由专业的期货分析师撰写，可信度高，专业性强，因此选取研报数据作为人工标注的语料。并取期货文本较为丰富的六个期货品种：棉花、

名称	描述	形式	语料规模/篇
研究报告	期货领域行业研究人员针对不同的期货品种和行业动态做出分析和研究;	半/非	634
品种简介	郑商所官网提供的已上市的期货品种简介，包含期货领域知识、术语和规范;	半	195
新闻文本	行业内期货分析师团队编写，提供每日国内外期货行情、期货报价等;	半	1,182

表 2. 期货领域多来源数据采集

苹果、白糖、红枣、玻璃、PTA作为研究对象。从Wind金融终端获取的研报为PDF格式，原始语料经预处理后的待标注语料规模总计174万字，包含74,437个句子。

4.2 半自动标注

半自动标注指根据期货产品的实体和关系分类体系，使用基于词典库的双向最大匹配算法对语料进行预标注，在此基础上进行人工标注和校对。为了提升标注效率，借鉴医学关系标注平台(张坤丽et al., 2020)，结合期货领域知识进行重新配置，形成了面向期货领域的实体和关系标注平台。其中词典库根据表2采集的数据经半自动标注-专家确认形成。预先在标注平台配置期货产品的实体和关系分类体系，在标注关系时仅可选择预定义的实体和关系类型，使用不同颜色表示不同的实体概念，使用连线和类型标签表示实体间关系类型。具体的标注流程如图2所示。 标注一致性用来描述两份标注结果的一致程度，一般使用Kappa值(Carletta,

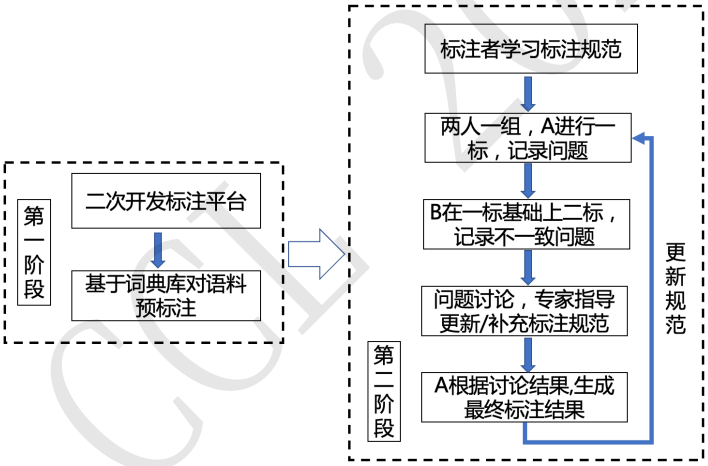


图 2. 半自动标注过程

1996)和F1值(Hripcsak and Rothschild, 2005)进行评价。文献(Artstein and Poesio, 2008)指出，标注一致性高于80%，可认为标注语料是可靠的，统计结果显示CERF语料库中实体和关系标注一致率F1值分别为85.70%和81.80%，均高于80%，说明本文构建的标注语料库可靠。

4.3 CERF语料库构建结果

期货领域文本包含丰富的实体关系三元组，因此单个实体可能同时涉及多个三元组中，进而导致实体出现重叠。实体重叠问题可分为普通型（NOR）、单个实体重叠型（SEO）和实体对重叠型（EPO）。普通型指无实体重叠问题；当单句中存在某个实体与其他多个实体存在关系，属于单个实体重叠型；实体对重叠型指句子中的相同实体对存在多种语义关系。CERF标注语料库规模如表3、表4所示，其中，实体重叠类型总数与语料总数不一致，原因为同一条语料中可能同时出现实体对重叠类型和单个实体重叠型两种类型的三元组。

(c) 2022 中国中文信息学会计算语言学专业委员会

产品类别	农产品			非农产品		
语料类别	训练集	验证集	测试集	训练集	验证集	测试集
语料数目	9,652	2,758	1,388	2,765	790	401

表 3. CERF实体语料库规模

产品类别	农产品			非农产品		
语料类别	训练集	验证集	测试集	训练集	验证集	测试集
子关系数目	42	42	42	36	36	36
语料数目	3,216	402	402	2,526	316	316
三元组数目	8,976	1,110	1,223	8,723	1,014	1,095
#单个句子中三元组数目						
1	1,173	114	136	758	105	97
2	741	104	94	488	63	65
3	488	54	51	373	48	44
4	292	31	35	282	32	25
≥ 5	522	69	86	625	68	85
#实体重叠类型						
NOR	1,398	171	164	873	123	115
SEO	1,818	231	238	1,653	193	201
EPO	666	33	64	167	11	9

表 4. CERF关系语料库规模

4.4 实体及关系抽取算法研究

在CERF标注语料库基础上，展开期货领域实体识别和关系抽取模型研究，将模型应用于期货研报文本进行实体识别及关系抽取，将数据整合后得到结构化三元组作为CFKG数据层的补充。（1）MHS-BIA模型

针对三元组重叠问题，本节在Bekoulis et al. (2018a)和Zhang et al. (2016)等研究基础上，提出基于Biaffine注意力的多头选择模型（Multi-Head Selection based on BIAffine attention，MHS-BIA）的改进算法，将信息抽取问题定义为一个多头选择问题(Bekoulis et al., 2018b)。模型能够同时识别实体，包括实体类型和实体边界，以及实体对之间所有可能存在的关系，使用Sigmoid损失获得实体间多个语义关系，以此能够独立预测不互斥的类。具体做法为：将模型识别的实体中最后一个字符称为实体“头”，如产品实体“烟台苹果”的尾字符“果”。模型认为该实体与本句中其他任意实体均可能存在语义关系。依次判别实体尾字符x_i其他实体尾字符y_l之间是否存在语义关系$\hat{c}_l$,若存在语义关系则将判断结果记作元组$(\hat{y}_l, \hat{c}_l)$。对于没有语义关系的元组，引入“N”标记无关系。模型框架如图3所示，包括实体识别模块和关系分类模块。

实体识别模块中的多头选择网络思想是将输入序列$W = \{w_1, w_2, ..., w_n\}$中每个字符$w_j, j \in \{1, 2, ..., n\}$组合，进而判断这两个字符是不是某个实体的头尾字符，并且将其归属于预定义实体类型e_k。通过将两个编码特征矩阵拼接，再通过线性变换层。计算w_i和w_j之间组成实体且实体关系为e_k的分数公式如（1）和（2）所示:

$$d = \dot{c}_i + \dot{c}_j \tag{1}$$

$$s^{(e)}(i, j, k) = \delta(Wd + b) \tag{2}$$

其中，d为两个编码特征矩阵拼接后的结果，c_i 和c_j表示BERT编码层的输出序列经过一层线性层变换和激活函数后得到的编码表示，$\dot{c}_i$ 和$\dot{c}_j$表示复制n份后得到的编码特征表示。W和b分别表示参数权重和偏差，δ为Relu激活函数。

(c) 2022 中国中文信息学会计算语言学专业委员会

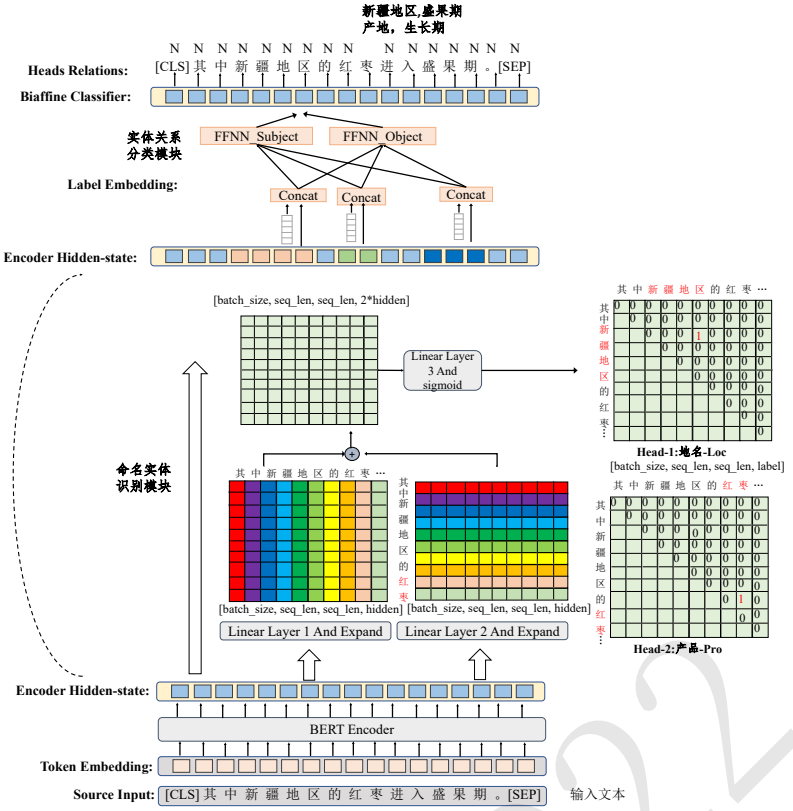


图 3. MHS-BIA模型

实体识别的交叉熵损失函数，与字符w_i和w_j之间存在关系e_k的可能性计算公式分别如（3）和（4）所示:

$$L_{ent} = \sum_{i=0}^{n} \sum_{j=0}^{m} -\log Pr\left(head = y_{i,j}, ent\_type = e_{i,j} \mid w_i\right) \tag{3}$$

$$Pr\left(head = w_j, ent\_type = e_k \mid w_i\right) = \sigma\left(s^{(e)}(i,j,k)\right) \tag{4}$$

$\sigma(\cdot)$为sigmoid函数，n为输入序列长度，m包括字符w_i的实体数量，$y_i \subseteq W$为除了w_i之外的其他字符，e_k表示实体关系。若两个字符是某实体类型下实体的头尾字符则为1，反之为0。

关系分类模块和实体识别模块共享同一个特征抽取层，将关系模块学习的特征与实体软标签进行向量拼接，结合两个子任务之间的联系。模型的任务为，对于给定的输入文本序列W和预定义好的关系集合$\mathcal{R}$，正确预测实体尾字符w_i,$i \in \{0,1,...,n\}$与其他实体尾字符所存在的语义关系$\hat{r}_i \subseteq \mathcal{R}$ 计算两个实体尾字符之间存在关系r_k的得分公式如（5）至（7）所示:

$$g_i = \frac{\sum soft\max\left(s_i\right) \cdot \mathrm{M}}{N} \tag{5}$$

$$z_i = \left[h_i; g_i\right], i = 0, \ldots, n \tag{6}$$

$$s\left(z_j, z_i, r_k\right) = V \cdot \mathrm{f}\left(U z_j + W z_i + b\right) \tag{7}$$

式（5）表示学习实体标签向量表示g_i，其中s_i代表输入序列第i个字符的状态分数向量，N代表预定义的实体标签种类数目，M为标签向量矩阵。式（6）表示向量拼接过程，z_i为特征抽取层的输出表示h_i与实体标签向量表示g_i两者拼接后的向量表示。式（7）表示计算关系系数的过程，其中$f(\cdot)$为relu激活函数，$V, b \in \mathbb{R}^l$,b代表向量维度，l代表当前层隐藏单元数。$U, W \in \mathbb{R}^{l \times (2d+b)}$，d为编码层隐藏单元数量。

(c) 2022 中国中文信息学会计算语言学专业委员会

传统浅层双线性分类器使用特征提取层的输入直接参与下一步计算，缺点在于模型任意时刻的输出均包含其他时刻的信息，本文在关系分类模块引入Biaffine注意力机制取代浅层双线性分类器，计算实体头尾字符的向量表示。深层Biaffine注意力机制模型将抽取的特征传入前馈神经网络，增加了本身的偏差项。Biaffine计算如公式（8）-（10）所示:

$$z_i' = FFNN_{Subject}\left(z_i\right) \tag{8}$$

$$z_j' = FFNN_{Object}\left(z_j\right) \tag{9}$$

$$s_m\left(z_i', z_j'\right) = z_i' U_m z_j' + W_m\left(z_i' \oplus z_j'\right) + b_m \tag{10}$$

其中z_i和z_i为式（6）中特征抽取的输出与实体软标签向量的拼接层，$FFNN_{Subject}$和$FFNN_{Object}$为前馈神经网络，z_i'和z_j'表示降维后的结果。实体对在所有关系类型上的得分计算如式（10）所示，b_m为偏差，$U_m \in \mathbb{R}^{d \times c \times d}$，$W_m \in \mathbb{R}^{(2d \times c)}$，d代表前馈神经网络隐藏单元数，c为关系类型数量。预测实体尾字符之间所有关系类型的概率值如式（11）所示，交叉熵损失函数如式（12）所示。其中n为序列长度，m为实体尾字符所涉及三元组数目，$r_i \in \mathcal{R}$为实体对的语义关系。

$$Pr\left(head = w_j \mid w_i\right) = softmax\left(s_m\left(z_i', z_j'\right)\right) \tag{11}$$

$$L_{rel} = \sum_{i=0}^{n}\sum_{j=0}^{m} -\log Pr\left(head = y_{i,j}, relation = r_{i,j} \mid w_i\right) \tag{12}$$

（2）GPN模型

基于全局指针网络的实体关系联合抽取模型（Joint extraction model based on Global Pointer Networks, GPN）在TPLinker模型(Wang et al., 2020)基础上，引入全局指针思想联合解码，将标注抽取框架设定为字符对链接问题，识别实体时将实体首位字符视作一个整体进行判别，最后使用条件层归一化的方式取代句子编码和主实体信息简单相加的方式，解决暴露偏差问题。

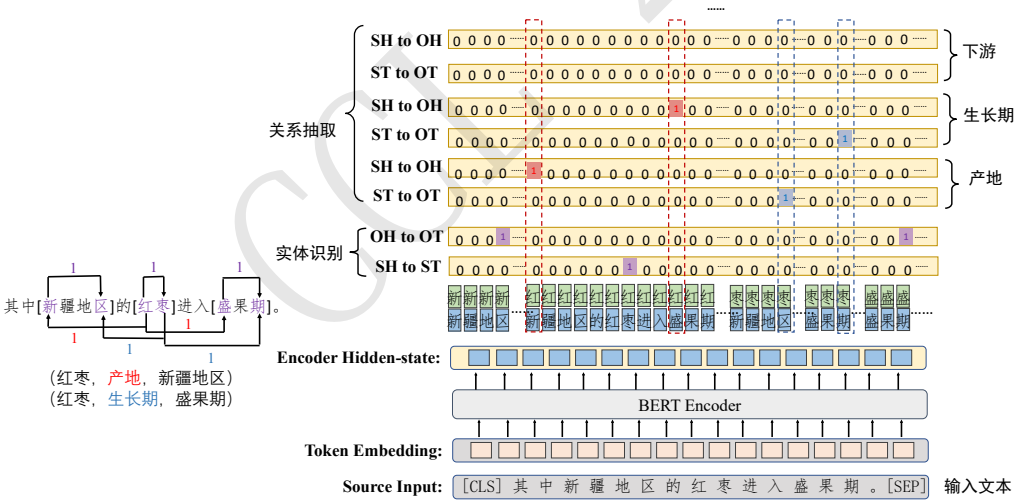


图 4. GPN模型

关系抽取的结果通常由三元组（Subject, Predicate, Object）表示，但模型在抽取过程实际为“五元组”（S_h,S_t,P,O_h,O_t）的抽取，其中S_h,S_t分别表示主实体的首尾位置，O_h,O_t分别表示客实体首尾位置。GPN模型使用单阶段实体关系联合抽取的标注方案，通过对主、客实体的头字符和尾字符进行标记，将实体识别和关系抽取描述为token对链接问题。模型标注框架如图4所示，分为实体识别和关系抽取两个模块，通过链接三种类型的矩阵实现标注过程。

(c) 2022 中国中文信息学会计算语言学专业委员会

编码阶段：1.主、客实体识别：SH to ST（Subject Head to Subject tail）表示识别主实体首尾字符；OH to OT（Object Head to Object tail）表示识别客实体的首尾字符；2.主、客实体头token识别：SH to OH，表示识别主、客实体首字符间的关系，如三元组（红枣，产地，新疆地区）：（红，新）=1；3.主、客实体尾token识别：ST to OT，表示识别主、客实体尾字符间的关系，如三元组（红枣，产地，新疆地区）：（枣，区）=1。

给定长度为n的输入序列$W = (w_1, w_2, \ldots, w_n)$，通过BERT等预训练模型将字符序列编码为向量序列h_N表示，其中每个字符w_i,$i \in \{1, 2, ..., n\}$映射到低维的语义向量$x_i = h_N[i]$,$i \in \{1, 2, \ldots, n\}$。GPN模型采用缩放点积型注意力机制，其字符对（$w_i, w_j$）的分类特征$x_(i,j)$计算公式如式（13）所示，其中$x_i, x_j \in {}_k$为字符的语义向量表示，$d_k$为向量维度，$x_{i,j}^{(\cdot)}$为不同类型的字符矩阵中值的分数表示。

$$x_{i,j}^{(\cdot)} = \frac{x_i^T x_j}{\sqrt{d_k}} \tag{13}$$

GPN模型为主实体的首尾（SH-ST）、客实体的首尾（OH-OT）、主实体和客实体的首字符对间的关系（SH-OH）、主实体和客实体的尾字符对间的关系（ST-OT）标注设计统一的框架。给定一个字符对（w_i, w_j）的特征表示$x_{i,j}^{(\cdot)}$，其链接标签计算公式如（14）和（15），其中$P\left(y_{i,j} = l\right)$表示（$w_i, w_j$）之间链接被识别为l的概率。

$$P\left(y_{i,j}\right) = Softmax\left(W \cdot x_{i,j}^{(\cdot)} + b\right) \tag{14}$$

$$link\left(w_i, w_j\right) = P_{argmax}\left(y_{i,j} = l\right) \tag{15}$$

解码阶段：1.解码SH-ST、OH-OT可得到句子中所有的实体，将主实体首字符索引作为关键字，整个实体作为值存储在字典D中；2.对于关系分类，解码SH-OH后，得到主实体和客实体的首字符token对，并在字典D中关联首字符索引链接的实体值。解码ST-OT后，得到主、客实体尾字符token对存储于集合T中；3.对第2步获取的主、客实体头字符token对链接到实体对，遍历集合T查询是否存在其尾字符token，若存在则输出实体关系三元组。

（3）实验结果

使用MHS-BIA模型与GPN模型在CERF语料库上进行实体关系抽取实验，并使用BiLSTM-CRF模型(Huang et al., 2015)和CASREL模型(Wei et al., 2019)作为对比模型。设置词向量维度为300，位置向量维度为300，Dropout 为0.5，GPN实体关系抽取结果见表6和表7。针对CERF语料库中部分实体类别标注错误，非农产品数据标记样本少等问题，对实体语料库训练集做了数据增强处理(Wei and Zou, 2019)，包括根据频次修正实体类型，简单数据增加等。

模型	农产品			非农产品		
	Micro-P (%)	Micro-R (%)	Micro-F1 (%)	Micro-P (%)	Micro-R (%)	Micro-F1 (%)
BiLSTM-CRF	75.00	76.67	75.82	70.70	76.98	73.71
CASREL	79.54	72.84	76.04	75.29	74.59	74.93
MHS-BIA	76.92	76.65	76.79	71.43	76.56	73.91
MHS-BIA-EDA	77.24	73.85	76.52	74.67	76.93	75.78
GPN	79.44	74.59	**76.94**	77.35	72.16	74.66
GPN-EDA	77.55	74.76	76.13	79.03	73.72	**76.28**

表 5. 模型命名实体识别实验结果

命名实体识别实验结果表明数据增强的方法更适用于训练数据少的任务，比如非农产品语料库中，数据增强的方法获得了1.62%的提升。而在农产品语料中，数据增强的方法并未获得提升。可能的原因为本章采用增加简单句的方法扩充训练数据，农产品的训练数据较非农产品更充足，增加数据样本的方法并不能提升效果。因此过多的数据增强对模型提升有限，该方法对训练数据越小的任务提升效果越明显。

实体关系抽取实验结果表明GPN模型结果优于其他模型，而MHS-BIA模型相比GPN的不足，可能原因在于实体和关系类别数目的限制：当实体或关系类别数目过多时，产生信号稀疏

模型	农产品			非农产品		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
CASREL	53.49	52.26	52.87	54.38	49.84	52.01
MHS-BIA	57.43	52.31	54.75	56.00	50.08	52.87
GPN	59.92	52.79	**56.13**	60.64	48.13	**53.67**

表 6. 模型实体关系抽取实验结果

问题，导致模型监督信号减弱，训练难度增大；此外，实体和关系任务抽取时共用一个编码器，使得关系分类判别依然依赖实体识别结果，存在暴露偏差，最终导致误差积累。本文未在关系抽取实验中使用数据增强技术，因为相较于实体识别，关系抽取语料的句子包含特定的关系三元组，多数语料文本长度更长。而自动构造的相似句长度较短并且生成的三元组质量不高，无法满足关系抽取任务，后续可考虑通过远程监督扩充关系抽取的语料。

5 构建结果及展示

CFKG的数据层由半自动标注构建的CERF语料库和模型自动标注结果两部分组成。本文抽取年份为2020.10-2022.1的六种期货产品的研报作为数据补充，采用实验效果较好的GPN模型进行实体关系联合抽取，经数据整合后得到结构化的实体关系三元组，作为CFKG数据层的补充。图谱中实体关系数量如表7、表8所示，共包含17,003个农产品关系三元组、13,703种非农产品关系三元组。

关系类型	半自动标注	自动抽取	合计
产品-属性	2,680	873	3,553
产品-地名	1,995	1,081	3,076
产品属性-变化	2,195	1,744	3,939
价格-因素	1,033	275	1,308
期货术语-属性	963	589	1,552
产品-价格	787	226	1,013
产品-产品	661	102	763
产品-因素	595	731	1,326
产品-其他	400	73	473

表 7. CFKG中农产品实体关系三元组统计

关系类型	半自动标注	自动抽取	合计
产品-属性	3,520	1,205	4,725
产业链条件-因素	3,358	777	4,135
价格-因素	805	180	985
期货术语-属性	681	199	880
行业-因素	649	60	709
生产设施-产业链条件	564	92	656
产品-地名	478	96	574
产品-企业	322	223	545
产品-产品	219	28	247
期货品种-期货术语	141	11	152
其他指标-属性	95	0	95

表 8. CFKG中非农产品实体关系三元组统计

为了直观反应CFKG中实体之间的关系，设计了期货领域知识图谱可视化查询系统。系统通过问句解析模块与知识图谱检索模块，实现对CFKG中节点进行查询和检索，展示界面如

(c) 2022 中国中文信息学会计算语言学专业委员会

图5所示。

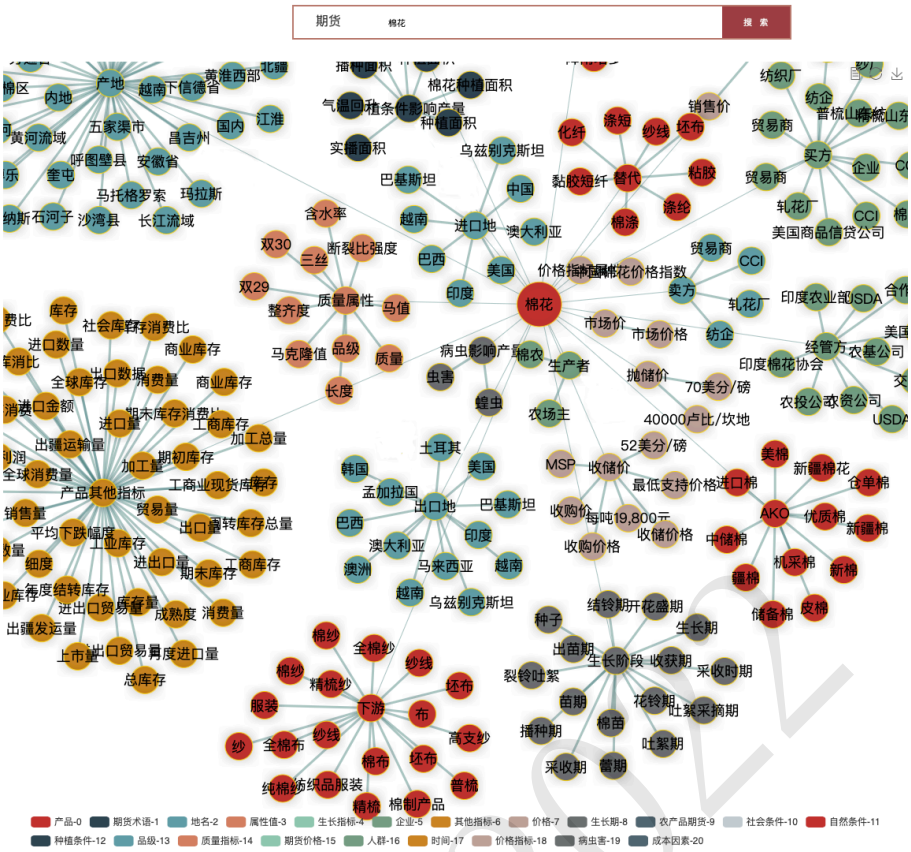


图 5. 棉花产品知识图谱展示

6 结语

本文描述了期货领域知识图谱CFKG的构建过程。首先在概念层整合多来源的期货文本，经领域专家指导下设计了知识图谱描述体系。在数据层采用半自动标注的方法构建了期货领域实体及关系标注语料库，使用自主构建的GPN模型进行自动标注，实现CFKG的知识更新。最后设计了期货领域知识图谱可视化查询系统对图谱进行可视化展示。未来的研究工作将在提升CFKG数据质量的同时，展开基于小规模语料的模型研究，或使用现有的信息抽取模型抽取新的语料作为训练语料补充，促使模型抽取效果迭代提升。其次本文构建的知识图谱属于静态的知识表示，但期货领域相比于传统领域信息更新快，故对知识更新速度具有较高的需求，未来工作可针对期货领域事理图谱展开研究，以及在静态知识图谱中融入动态知识实现图谱的动态更新。

参考文献

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018a. An attentive neural architecture for joint segmentation and parsing and its application to real estate ads. *Expert Systems with Applications*, 102:100–112.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018b. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

(c) 2022 中国中文信息学会计算语言学专业委员会

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.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *arXiv preprint cmp-lg/9602004*.

Huajun Chen, Ning Hu, Guilin Qi, Haofen Wang, Zhen Bi, Jie Li, and Fan Yang. 2021. Openkg chain: A blockchain infrastructure for open knowledge graphs. *Data Intelligence*, 3(2):205–227.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Elias Kärle, Umutcan Şimşek, Oleksandra Panasiuk, and Dieter Fensel. 2018. Building an ecosystem for the tyrolean tourism knowledge graph. In *International Conference on Web Engineering*, pages 260–267. Springer.

Xing Niu, Xinruo Sun, Haofen Wang, Shu Rong, Guilin Qi, and Yong Yu. 2011. Zhishi. me-weaving chinese linking open data. In *International Semantic Web Conference*, pages 205–220. Springer.

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 *International semantic web conference (Posters & Demos)*, volume 1035, pages 121–124.

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*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2019. A novel cascade binary tagging framework for relational triple extraction. *arXiv preprint arXiv:1909.03227*.

Guohai Xu, Hehong Chen, Feng-Lin Li, Fu Sun, Yunzhou Shi, Zhixiong Zeng, Wei Zhou, Zhongzhou Zhao, and Ji Zhang. 2021. Alime mkg: A multi-modal knowledge graph for live-streaming e-commerce. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4808–4812.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2016. Dependency parsing as head selection. *arXiv preprint arXiv:1606.01280*.

奥德玛, 杨云飞, 穗志方, 代达劢, 常宝宝, 李素建, and 昝红英. 2019. 中文医学知识图谱cmekg 构建初探. 中文信息学报, 33(10):1–7.

张坤丽, 赵旭, 关同峰, 尚柏羽, 李羽蒙, and 昝红英. 2020. 面向医疗文本的实体及关系标注平台的构建及应用. 中文信息学报, 34(6):36–44.

(c) 2022 中国中文信息学会计算语言学专业委员会

近四十年湘方言语音研究的回顾与展望 ——基于知识图谱绘制和文献计量分析

杨玉婷
暨南大学文学院
暨南大学汉语方言研究中心
131871934510163.com

刘新中*
暨南大学文学院
暨南大学汉语方言研究中心
tlxzh@jnu.edu.cn

彭志峰*
暨南大学文学院
暨南大学汉语方言研究中心
pengzhifeng@jnu.edu.cn

摘要

湘方言语音研究已取得丰硕的研究成果，本文以中国知网中的“学术期刊库”为数据来源，采用文献计量分析的方法，通过CiteSpace等工具，从发文信息、聚类分析及演进趋势等维度对相关文献进行统计分析和可视化知识图谱绘制，全方位考察近四十年的研究概貌，提出“类型学”、“语音层次”将会是湘方言语音研究较新的、有待进一步开拓的领域，为今后开拓新的研究方向提供理论依据，为湖南语保工程的资源进行深度开发利用提供数据支撑。

关键词： 湘方言 ； 语音 ； CiteSpace ； 计量分析 ； 知识图谱

Review and Prospect of the Phonetic Research of Xiang Dialects in Recent Forty Years:Based on Knowledge Mapping and Bibliometric Analysis

Yuting Yang
Jinan University
131871934510163.com

Xinzhong Liu*
Jinan University
tlxzh@jnu.edu.cn

Zhifeng Peng*
Jinan University
pengzhifeng@jnu.edu.cn

Abstract

Great achievements have been made in the study of Xiang dialects phonetics, this paper takes ‘academic Journal Database’ in CNKI as the data source, using the method of bibliometric analysis, using tools such as CiteSpace, related literatures were statistically analyzed and visualized knowledge maps were drawn from the aspects of publication information, cluster analysis and evolution trend. Taking a full look at the last 40 years of research, it is proposed that ‘typology’ and ‘phonetic hierarchy’ will be a relatively new field in the study of Xiang dialect phonetics, which needs to be further explored. It provides theoretical basis for exploring new research directions in the future and data support for the in-depth development and utilization of Hunan Language protection Project resources.

Keywords: Xiang dialects , Phonetics , CiteSpace , Econometric analysis , Mapping knowledge domain

1 引言

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
*共同通讯作者：刘新中（tlxzh@jnu.edu.cn），彭志峰（pengzhifeng@jnu.edu.cn）
基金项目：国家社科基金重大项目（汉语方言学大型辞书编纂的理论研究与数字化建设）

湘方言又称湘语，历史源远流长。1900年，章太炎作为最早给现代汉语方言分区的人，将汉语方言分为九种，湖南、湖北、江西被视作同一种方言。但直到1937年李方桂在《中国年报》上发表《语言和方言》，正式提出湘语作为八大汉语方言之一。湘方言的几代学者在研究范围的拓展、研究方法的创新与多样、具体语言学问题探讨等方面开展了诸多研究。

2015年，国家正式启动“中国语言资源保护工程”，同年12月，湖南语保工程正式启动。2019年底，湖南已完成省内100个调查点的调查摄录工作。进入语保工程二期，语保工作者积极投身于湘方言资源调查、保存、展示和开发利用等为核心的各项工作中。因此，全面统计和概览湘方言语音研究的全貌，对已取得的成果进行系统性的梳理和总结，十分迫切也很有必要。

从已有的论文来看，对湘方言语音研究做文献综述的研究以描述和梳理等方法为主，鲜见采用计量分析和图谱绘制的方法对该领域发展趋势进行科学系统的研究。本文选取与湘方言语音研究有关的主题词，在中国知网（以下简称CNKI）中的“学术期刊”数据库进行检索，全面梳理1980年至2021年的相关论文，进行计量分析和知识图谱绘制，厘清国内湘方言语音研究的历史脉络和演进趋势，为今后的理论研究和湖南语保工程的开展提供数据参考。

2 研究方法与数据来源

本研究采用的文献计量分析方法，是通过数据挖掘、计量分析、可视化图谱绘制等手段，对学科主题进行全面分析。文献计量分析工具是采用美国德雷塞尔大学陈超美教授所开发的CiteSpace。该软件的原理是基于共引分析理论和寻径网络算法，对某个领域的文献资料进行计量分析，探寻该学科领域演化的关键路径及其知识拐点，绘制可视化图谱来分析该学科领域研究热点具体的变化过程、学科演化的潜在动力机制，并探测学科发展前沿，作为科学计量学普遍使用的新工具来有效帮助读者充分了解所从事的研究领域。

本文以CNKI中的“学术期刊”数据库为来源库，时间跨度为1980至2021年，根据专家建议和文献阅读，确定检索式为（湘语+湘方言+新湘语+老湘语+湖南话+湖南方言+长益片+衡州片+娄邵片+辰溆片+永州片）*（语音+声母+韵母+声调+元音+辅音），检索时间为2022年4月10日，对检索到的数据进行去重、勘误以及选择等清洗后，筛选出有效数据383条，输出Excel数据表和Refworks文献数据表。然后对有效数据进行标准化处理，主要包括同义词和缩写词的规范统一，机构名称的合并（同一个机构前后名称不一致，改成最新的名称），缺少关键词的文献进行关键词提取，不规范关键词的删除等，从而保证软件运行结果的客观性和准确性。

本文通过Excel对年度发文量、发文作者、研究机构和刊发载体进行统计并绘制图表，再通过CiteSpace对Refworks文献数据表进行可视化分析，分析类型依次为作者、机构和关键词，生成年度发文量统计图、作者共现网络、研究机构共现网络、关键词聚类图谱、关键词突现趋势图，分析湘方言语音研究的发文量、核心作者、合作网络、研究热点和演进趋势等。

3 发文信息分析

发文信息包含发文时间、核心作者、研究机构、刊发载体等，用于分析湘方言语音研究论文发表的基本信息。

3.1 发文时间分布

学术论文数量的时序变化是衡量某领域发展的重要指标。绘制分布曲线对文献分布做历史的、全面的统计，对评价该领域所处的阶段、预测发展趋势和动态具有重要意义（邱均平et al.,2012）。湘方言语音研究的年度发文量分布如图1所示,可大致分为三个阶段：（1）改革开放后至1984年为起步阶段，开展湘方言语音研究的高校与机构还较少，研究规模不大，涉及的主题也不多。2）1985-2001的平稳发展阶段，发文量缓慢上升，以湖南师范大学为主，其他高校的成果数量较少。（3）2002年至今的高速发展阶段，2006年、2009年、2010年分别出现3次发文高峰，其中2010年达到峰值（27篇）。21世纪初为湘方言语音研究发展的重要节点，此前的方言普查为之后的语音研究积累了丰富的方言材料，同时湖南省内高校的学科建设取得初步成果，如湖南师范大学分别于1985年、1998年获得汉语言文字学硕士和博士学位授予权，吉首大学于2006年获得汉语言文字学硕士学位授予权。一批高校将方言研究与学科建设、人才培养有机结合，培养了一批具备调查能力的方言工作者，他们中的许多人逐渐成为湘方言研究领域

的骨干力量。这一时期，湘方言工作者在田野调查的基础上，尝试发掘出语音研究的相关规律，归纳出具有普遍指导意义的理论性结论。同时，从原来单一的学科内部研究向跨学科和社会应用发展，取得一批优秀的研究成果。其中2005-2013的发文总量为198篇，约占四十年发文总量的二分之一。总体而言，湘方言语音研究文献量呈较为稳定的上升趋势。这表明湘方言语音研究的热度和受关注程度在不断上升，专家学者在个别年份进行了集中、大范围的研究。

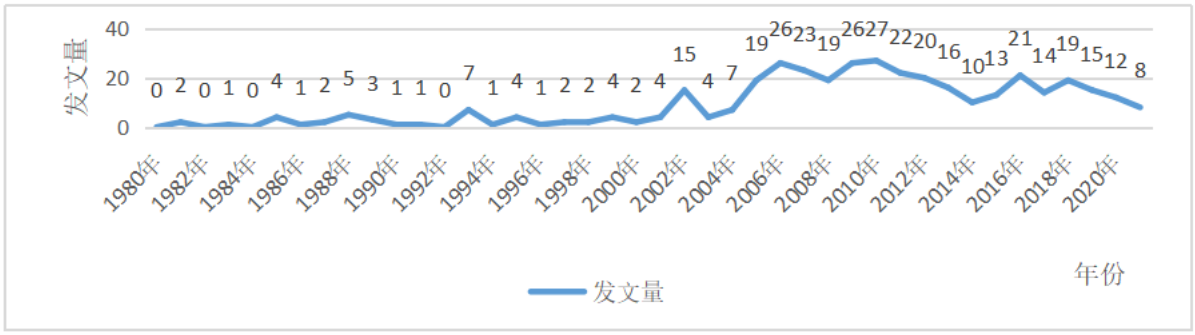


Figure 1: 文献年度发文量分布

3.2 核心作者分布

核心作者概念由德瑞克·约翰·德索拉·普赖斯提出：该领域论文的高产作者，具有一定的学术影响力。宗淑萍（2016）的研究表明：发文量与被引量分别从“量”和“质”方面反映了核心作者的学术水平。发文量代表了作者对期刊的重要性，被引量代表了作者的学术影响力。

根据普赖斯定律，核心著者至少发表论文数为m_p篇，计算公式：$m_p = 0.749\sqrt{n_{pmax}}$，其中$m_p$为统计时段内作为核心作者至少发表的论文数量，$n_{pmax}$是统计时段内发表论文最多的作者发表的论文数量。计算出湘方言语音研究领域核心作者至少发表的论文数$m_p \approx 3.18$，按照取整原则即发表3篇或3篇以上论文的作者入选为核心作者候选人。

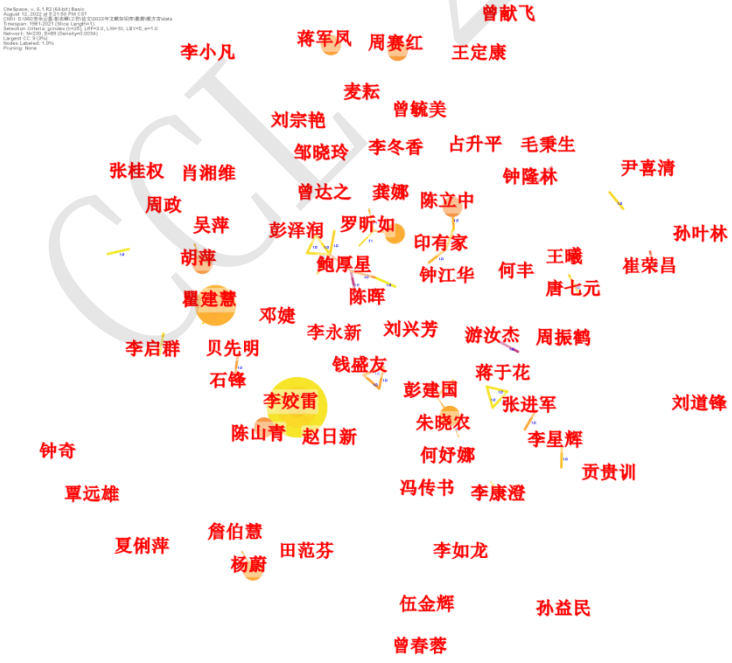


Figure 2: 作者共现网络

本文根据综合指数对核心作者候选人进行测评。对于发文量和被引量2个指标权重值大小，不同的学者有不同的观点，其中宗淑萍（2016）将权重值分别确定为0.5和0.5。本文认为，重要性和学术影响力具有相同的重要程度，因此对核心作者的发文量和被引量的权重值均设定

为0.5。按照公式$z_i = x_i/\bar{x} \times 100 \times 0.5 + y_i/\bar{y} \times 100 \times 0.5$计算得出综合指数值：$z_i$为第i位候选人综合指数，$\bar{x}$为核心作者候选人平均发文量，$\bar{y}$为核心作者候选人发文平均被引量，$x_i$为第i 位候选人累计发文量，$y_i$为第i位候选人累计被引量。综合指数统计表见附录表。

本文基于图2与核心作者分布情况表（见附录），将排名靠前的作者大致分为四类研究方向：（1）湘方言的区划，鲍厚星、陈立中等对湘方言的区划问题作深入研究，对湘方言内部分片研究有重要影响，帮助正视湖南方言内部的差异性；（2）语言接触与影响，瞿建慧、李冬香、李启群等关注湘方言、乡话、赣方言等不同方言接触影响后的方言面貌；（3）语音学本体研究，李姣雷、夏俐萍等对语音特征有深入研究，在语音本体研究领域贡献了大量文章；（4）实验语音学研究，彭建国、贝先明等对湘方言的实验语音学研究有较为深入的研究。

图2可看出：湘方言语音研究形成了数个规模较小的作者合作群体，展开合作研究的核心作者大多为师生或同事关系，如鲍厚星与陈晖、彭泽润（师生与同事），詹伯慧与杨蔚（师生），朱晓农与彭建国（师生）。结合综合指数表，可见排名靠前的核心作者的互动和合作较少。

总体而言，湘方言语音研究已形成小规模的作者合作群体，高等院校、科研院所、学术团体等机构之间通过互相合作而产生学术增量的现象并不多，主要依靠部分高产作者、部分高产机构进行研究。通过对作者的籍贯和年龄进行统计，核心作者以湘籍学者居多，中青年群体研究人员出现一定程度的断层，说明湘方言语音研究的梯队建设任重道远。

3.3 主要研究机构的分布与合作

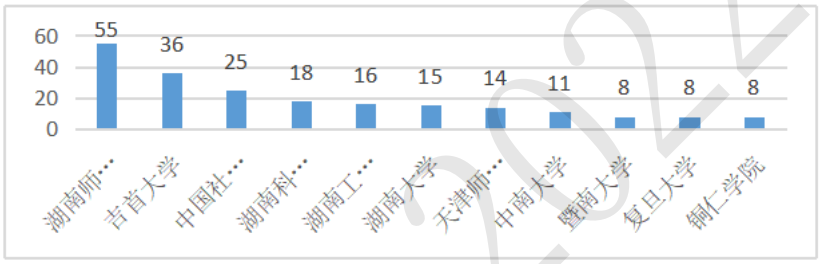


Figure 3: 湘方言语音研究高产研究机构发文量

图3为湘方言语音研究高产研究机构发文量图。可看出：湖南师范大学作为主要的湘方言语音研究机构，排名第一（55篇），其中文学院（49篇）、外国语学院（3篇）、物理与信息科学学院（2篇）、新闻与传播学院（1篇）。吉首大学排名第二，发文36篇。中国社会科学院语言研究所排名第三，发文25篇。这些机构产出活跃，组成了湘方言语音研究的学术重镇。

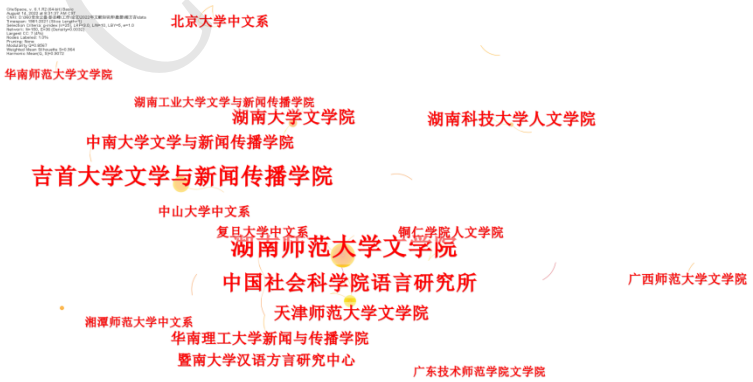


Figure 4: 研究机构共现网络

由研究机构共现网络（图4）和Excel表中的研究机构统计相结合分析，较为明显的合作机构有：吉首大学文学与新闻传播学院和中山大学中文系，华南理工大学新闻与传播学院和暨南大学汉语方言研究中心。

(c) 2022 中国中文信息学会计算语言学专业委员会

总体分析：1.发文机构以湖南省内的高校为主；2.各研究机构在发文数量上存在一定差距，前三名机构发文量明显超过其他机构；3.研究机构之间的合作并不紧密。

3.4 主要刊发载体分析

对于刊发载体的分析应结合学科领域核心期刊和期刊发文量两个指标来进行综合评测。按邱均平（2007），学科领域核心期刊遴选的依据是布拉德福定律，布拉德福定律是对出版该文献的专业期刊中专业文献数量分布规律的总结，是文献计量学中的基本定律之一。

根据布拉德福定律，按学科文献载文量的多少，将期刊划分为三个区域，每一个区域的载文量相等，三个区域的期刊数量之比为$1:n:n^2$。其中，第一个区域为该领域的核心期刊区，第二个区域为相关期刊区，第三个区域为边缘期刊区。湘方言语音研究的论文总数为383篇，刊文的期刊为129种，每一个区域约为128篇。其中，核心期刊区有5种（126篇），相关期刊区有26种（129篇），边缘期刊区有98种（128篇）。三个区域的期刊数量之比为5∶26∶98，n的比值接近5，在各分区论文占比大致相等的情况下，期刊数量呈现$1:n:n^2$数量的区域分布,基本符合布拉德福定律的定义。从湘方言语音研究发文的核心期刊区来看，《方言》（64篇）、《语言研究》（21篇）、《中国语文》（16篇）、《语言科学》（15篇）是国内语言学科的专业刊物，《湖南师范大学社会科学学报》（11篇）是湘方言研究重镇湖南师范大学的学报。上述5种期刊也均为CSSCI 中文社会科学引文索引(2021-2022)来源期刊和北京大学《中文核心期刊要目总览》来源期刊，学科权威性较强。从总体分布来看来看，排名靠前的刊物以湖南省内高校学报居多，约占总发文量的30%，核心刊物和CSSCI期刊数量约占总发文量的40%。

尽管跨学科协作研究是方言研究的趋势，但从刊发载体分布来看，并未在其他相关学科中形成核心期刊群，学界应大力推动湘方言语音的跨学科研究，将“湘方言语音”放在更多元的学科背景中进行分析和探讨。

4 湘方言语音的研究热点及演进趋势分析

关键词作为文章的主题提炼，能体现文章的核心内容，有助于深入挖掘该领域的研究内涵。通过关键词的出现频率、聚类分析和共现图谱，可以分析研究热点和预测演进趋势。

4.1 关键词共现图谱分析

Figure 5: 关键词聚类图谱

(c) 2022 中国中文信息学会计算语言学专业委员会

关键词表达了论文的研究主题，是论文的核心与精髓。通过CiteSpace绘制出现频次≥2的关键词聚类图谱（图5）。除了“湘语”（81）、“湘方言”（28）、“汉语方言”（17）以外，“同音字汇”（24）、“声调”（19）、“赣语”（18）、“入声”（15）、“西南官话”（14）、“乡话”（14）、“演变”（13）、“文白异读”（13）、“湘西乡话”（12）、“声母”（11）、“塞擦音”（9）、“层次”（8）、“比较”（8）、“全浊声母”（8）等关键词共现频次均大于等于8，它们反映了1980-2021湘方言语音研究的热点，建构了该领域的知识网络。由关键词共现图可以看出，整个高频关键词网络是比较紧密的，在一定程度上说明湘方言语音的研究都是围绕这些核心展开的。

4.2 关键词聚类分析

关键词聚类分析用于将相似的关键词进行归类，在关键词共现的基础上进行，其原理是将检索时间范围内的关键词按照对数似然率算法（LLR）进行提取，运行后得到关键词聚类，每个聚类由多个紧密相关的词组成。表中知识图谱模块值（Q值）均值为0.784，轮廓值（S值）均值为0.9255。Q值>0.3，说明聚类结构显著，S值>0.5，说明聚类合理。本次聚类效果显著。

聚类号	聚类成员数量	轮廓值	年份中位数	聚类标签（LLR）
0	24	0.966	2009	赣语
1	21	0.938	2009	方言调查
2	21	0.969	1996	粤北土话
3	21	0.978	2013	同音字汇
4	20	0.951	2008	不送气
5	20	0.987	1997	比较
6	16	0.9	1994	演变

Table 1: 聚类表（本文选取前7个聚类）

表1显示了关键词聚类情况，0、1、2、5、6号聚类着重考察湘方言与相关方言的接触，内部片区划分。3、4号聚类着重考察音韵研究。结合图5，我们将意义相近的关键词再次进行合并与归类，着重阐释各关键词聚类标签下关注的研究重点，得出两个热点领域。

4.2.1 语言接触和片区划分

（1）语言接触

“语言接触”是湘方言语音研究的重要方向，湘方言地区诸方言相互影响和渗透。关注度较高的关键词有“赣语”“西南官话”“乡话”。瞿建慧、李冬香、罗昕如、杨蔚等以湘方言复杂的方言现象为依托，关注湘方言、乡话、土话、赣方言等不同方言接触影响后的语言面貌，对方言分区和归属进行探讨，分析语言接触现象，探索语言互动的规律并构建其理论。瞿建慧从声调、声母、文白异读和语音比较及演变等角度，研究湘西汉语方言、少数民族语言的语音特征。瞿建慧（2012）指出，湘西州土家语、苗语与汉语方言都有一套完整的浊声母系统，大量浊声母汉借词的进入强化了湘西土家语和苗语浊声母的语音特征，延缓了其清化进程，而浊声母清化和浊塞擦音擦音化则是各自演变的产物。李冬香关注与比较湖南、江西境内赣语、湘语的音韵特点，把动态演变和静态结构联系起来，从共时表现和历时演变的角度对不同方言相互接触后的层次关系与渗透关系进行探讨分析。李冬香（2003）对比湖南、江西、粤北等地的方言，发现湘语、赣语、平话、乡话及土话中普遍存在表处置结果和动作完成的“咖”；通过考求其语音和语义，提出它们都来源于古代汉语的“过”，由此分析湘语、赣语及客家话的关系密切程度。罗昕如关注湖南省内的湘语、湘南土话及广西省内的湘语，揭示在方言间的接触和频繁影响下，其语音面貌与特征的演变，总结方言的发展趋势。罗昕如（2002）从声母、韵母、声调及部分常用词语等方面对蓝山境内四个土话片的代表点及其中一片内部八个姓氏的土话（共11个点）进行比较，兼同蓝山官话作对比，对其存在的主要差异及差异程度进行分析。杨蔚从语言接触和影响的角度，结合历史地理背景，探讨湘西乡话的分布与分片、语音特点和内部关系等情况，分析湘西乡话共时体系的形成原因，把握其层次关系与动态演变，与湘方言、赣方言和吴方言等周边汉语方言进行横向比较，考察方言之间的密切联系。杨蔚（2002）对沅陵乡话、东安土话、宜章土话、江永土话等集中在湖南境内尚未分区定性的土话韵母进行共时

(c) 2022 中国中文信息学会计算语言学专业委员会

描写，将其与吴语、湘语进行比较，研究其韵母演变格局，指出沅陵、东安、宜章、江永与吴语、湘语韵母的相同之处应该是原来相同或相近的方言因素的共同留存。

（2）省内方言分布和湘方言片区划分

湖南是多方言地区，除了湘方言，还包含西南官话、赣方言、客家话及系属未定的湘南土话和乡话。湘方言不仅仅分布于湖南省内，还分布于广西东北部、陕西南部、四川及贵州等地。鲍厚星、颜森、李蓝、陈晖等对湖南省内方言分布和湘方言的分区分片问题作了深入思考，研究湖南方言内部的差异性。鲍厚星在《汉语方言学大词典》中指出：根据湘语确认的语音标准，并结合考察人文历史地理等因素，湘语可划分为五片：长益片、娄邵片、衡州片、辰溆片、永全片。鲍厚星、颜森（1986）将湖南省汉语方言分为湘语、赣语、客家话、江淮官话、西南官话、乡话区六个区。李蓝（1994）对当时几种湖南方言分区说法作了分析和评论，提出按照声韵调系统三重投影的方法重新给湖南方言分区，并用湖南花鼓戏的流行区域来验证其分区结果。鲍厚星、陈晖（2005）讨论湘语的分区问题，对《中国语言地图集》中的湘语分区进行介绍，提出对湘语分区的重新思考与湘语内部分片结果。学者们基于对湖南省内汉语方言的实际调查，结合之前专家学者的研究成果，对湘方言进行分区并简述各方言区的语音共同特征及主要差异。学界对湘方言内部具体方言归属问题也有着较高的关注。陈立中（2002）根据古全浊声母的今读、知组与端组的分混等语音演变关系证明湖南汝城话属于湘南土话。瞿建慧、谢玲（2011）经过对比分析，认为主要分布在湖南省西南地区会同、靖州、通道等地的湘西南酸汤话是当地苗族居民向汉人学习湘语过程中形成的中介语，在演变过程中受到了官话和赣语的影响，属于湘语。覃远雄（2021）以归属仍有争议的广西全灌话为研究对象，将其语音特点与湘语、西南官话分别对比分析，提出将全灌话视作跟湖南永州等湘南官话相同方言的结论。

4.2.2 方言音韵研究

方言音韵研究是语音研究的重点和根本，学界重点关注声调、声母、文白异读等方向，学者们从声调、声母、韵母等方面对湘方言语音面貌不断深入研究并形成了丰富的成果。

（1）声调

声调研究是湘方言语音研究的主要领域，众多学者运用传统音韵研究和实验语音学的方法对其有较为深入的研究。李星辉（2005）认为古入声韵尾在湘语中发展渲变的大致轮廓是从有塞音韵尾演变到塞音韵尾发生合并，再到韵尾塞音进一步弱化为只剩下音节后的紧喉动作，最后入声韵特色逐渐消失，发音同于舒声韵。21世纪初开始，实验语音学在湘方言声调研究中得到了越来越多的应用和发展。朱晓农在《语言语音学和音法学：理论新框架》《全浊弛声论——兼论全浊清化（消弛）》等文中提出完整的理论框架和工作取向，强调要加强关于发声态的研究，以实验的手段研究跟语言有关的发音性质、语音的组织、分布和演化问题，解决了“湖南岳阳话中的假声”等具有语言学意义的问题。彭建国、朱晓农（2010）指出湖南岳阳话中的假声是具有语言学意义的一种发声态，涉及阴去、次阴去、阴入三个调类，作声调描写时应分中声域和高声域两个声域，认为湘语中普遍存在高域调。

（2）声母

早在1987年，湘方言语音的计算机实验研究就进入了学界的研究视野，陆致极（1987）以《汉语方音字汇》为基本材料，利用计算机整理出包括长沙、双峰在内的汉语十七个方言调查点在声母和韵母分布方面的数据，并采用聚类分析程序对数据进行统计，对各方言之间的联系密切程度以及归类状况作出计量描写，根据统计结果讨论各方言之间的亲疏关系以及分区的问题。蒋军凤（2008）发现湘乡市现辖的5乡13镇4个办事处中，只有金薮乡存在部分端组字今读舌尖后塞擦音[t]组声母的现象，根据该方言的新老差异及类似材料作为旁证，推测出金薮乡方言端见组今读舌尖后塞擦音的演变过程。陈晖（2008）详细考察古全浊声母在湘方言中的今读音情况，揭示了仍保留浊音的湘语点与吴语，浊音清化的湘语点与闽语在古全浊声母演变上的差异，总结了湘方言古全浊声母演变的特点和规律，指出古全浊声母在湘语部分方言中仍保留浊音，在部分方言中浊音清化，并对古全浊声母的今读在方言分区中的局限性进行了探讨。

（3）韵母

学者们从韵母入手对湘方言语音深入研究并形成了丰富的成果。李姣雷、赵日新（2016）对学界有关湘语蟹假果遇四摄元音推链音变的观点进行讨论，具体分析湘方言中果摄一等与遇摄模韵、蟹开二与假开二、假开二与果摄一等的关系，认为湘方言中不存在蟹假果遇摄的元音推链现象，果假摄元音的高化是一种自主的后高化演变。贝先明（2008）运用元音格局的方法

(c) 2022 中国中文信息学会计算语言学专业委员会

分析长沙、萍乡、浏阳三地方言的一级元音格局，从中古韵摄的角度比较三地方言一级元音格局的异同，发现了始发方言、混合方言、目的方言的元音格局分布模式。

4.3 关键词突现趋势分析

关键词突现能够显示出该领域研究热点的演化动态，预测其发展趋势。突现趋势图中Keywords为关键词，Year为检索数据的年份，Strength为突显强度，突显度越大，说明研究前沿越明显，Begin为某一关键词研究热点的起始年份，End为终止年份。关键词的突显年份即该研究领域在某一较短时间段内追踪的热点。

由图6可知，“方言点”、“方言音系”和“全浊声母”的突现时区相对一致，都在1986-1998，其中“方言点”的突现年限（1987-1998）稍长于“方言音系”（1988-1995）和“全浊声母”（1988-1996）。随着湘方言语音研究的进一步深入，方言工作者增加了对湘方言语音的全面认识，对个别地点的深入调查也取得了重要成绩。这一时段有关语言本体的研究体现了湘方言不同方言点的语音基本面貌，主要围绕方言音系、全浊声母等语音特点展开，发文量也不断上升。

2002-2014湘方言内部区划与归属成为研究的热点，“分区”（2005-2007）、“内部差异”（2010-2014）等成为突现热词。同时，有关湘方言语音本体的研究持续且集中，具体体现在“声韵调”（2002-2007）、“语音特点”（2006-2007）、“语音”（2010-2012）等突现热词。

2010年开始，湘方言研究呈现出多元化的新气象。实验语音学在湘方言语音研究中得到了全方位的应用，围绕湘方言语音的声学特性而展开研究，强调量化分析和实证研究。语言接触和影响也成为学界普遍关注的重点，学界围绕湘方言、客家话、赣语、湘南土话和乡话之间的基础关系与各方言点的语音面貌展开论述。如“乡话”（2013-2018）、“湘西乡话”（2015-2017）等突现热词。这两个关键词突现时间虽然较短，但以其为研究主题涌现了一批学术成果。进入新时代，国务院办公厅印发了《关于全面加强新时代语言文字工作的意见》，强调推进新时代语言文字事业改革发展，学界围绕湘方言区的教学与社会应用展开了一系列研究。

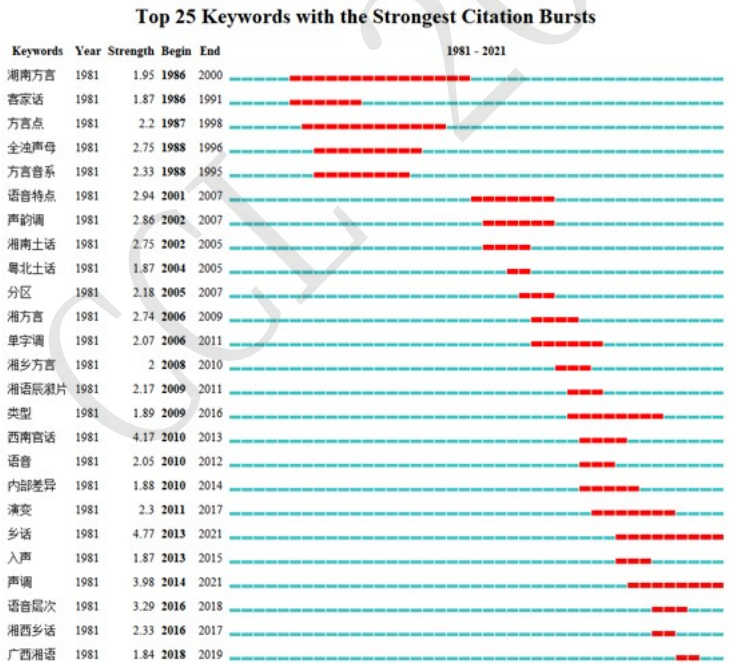


Figure 6: 关键词突现趋势图

通过上述分析可以发现，近四十年来湘方言语音研究热点交叠更新，受国家语言政策和学界认识不断深入的影响，研究热点的演进在不同的时间呈现出不同的特点。从关键词突现终止的时间可以看出，近几年新颖度较高的研究焦点是“类型学”、“语音层次”，显示出在近些年湘方言语音研究中，它们是在国内目前较新的、有待进一步开拓的领域。目前学界在湘方言语音研究领域尚未形成新的研究热点。

5 湖南语保工程和方言文化数字化传承

5.1 方言文化的保护与传承

湖南省作为国家首批中国语言资源保护工程试点省份之一，于2015年正式启动语言资源保护工程，从湖南省语保调查点一览图（图7）可以看出，按照“一县一点”的原则在全省铺开，对全省100个方言点进行调查和数据处理，分别于2016年、2017年、2018年获得国家立项21项、28项、25项，均位居全国第一，积累了大量的音视频、图片、文化资源。同时，在国家语保工程的总体框架内，在国家语保工程的总体框架内，湖南省整合了政府、民间、企业的力量，开展了省级语保工程，全方位采录了省内的语言材料，初步建设了湖南语言资源数据库，有力推动了湘方言文化的保护与传承。

Figure 7: 湖南省语保调查点一览图

5.2 湖南方言文化资源数字化传承

2022年5月，中共中央办公厅、国务院办公厅印发《关于推进实施国家文化数字化战略的意见》，提出“统筹利用文化领域已建或在建数字化工程和数据库所形成的成果，关联形成中华文化数据库”的重点任务[0] 。本文提出：在语保工程二期启动建设之际，有必要在已有语料数据库的基础上，对湘方言文化资源进行全方位的数字化处理，建设湖南方言数字博物馆，实现数字化成果的全民共享。

语言数字博物馆是集语言社会生态数据采集、语料数字化采录和存储、语料数据分类存储和转写标注、语料资源专业检索和大众化服务为一体的网络空间平台，数据类型包括数字化文献资料、数字多媒体基础语料、数字多媒体转写资料、社会生态数据四类，其中的语料库和文献库是博物馆最重要的基础数据资料。语料库包括编码、条目、音标注音、拼音、注释、媒体、语言、音韵、方言或民族文字、分级分类、普通话词对译、字、词汇、句子、语篇等主要字段。文献库包括书本编码、书名、地点、地点编码、语言、作者编码、作者、出版社编码、出版社、出版年份、出版语言、目录页数、内容页数、阅读量、下载量、文件、上传时间、引用、在线阅读、本地下载等主要字段。

6 结语

随着研究的深入，学界积累了丰富的湘方言材料，对湘方言语音的基本面貌有了整体把握，对湘方言语音的认识逐渐清晰，对湘方言个别地点的深入调查也取得了重要成绩。重视差异性的同时也重视共同性，内容从调查语音铺开，从单点调查到成片地区的比较，从共时描写到历时研究，从纯语音研究到与地域文化的结合，探讨了入声的演变、泥来母的分混和古全浊声母的演变等一系列重要的语音现象。但由于湘方言内部情况复杂，在发展过程中受到来自其他语言或方言的影响，形成了多样的方言现象，因此仍然有颇多问题有待进一步探讨分析。

在新时代背景下，湘方言语音研究议题亟待深化。一方面，借助相关语言政策，聚焦重点与创新机制，紧扣新时代，提出新目标，持续深入对湘方言语音方面的研究。同时，还应该将

[0]http://www.gov.cn/xinwen/2022-05/22/content_5691759.htm 中共中央办公厅国务院办公厅印发《关于推进实施国家文化数字化战略的意见》

研究焦点定位于国家和地区的发展需求上，在科学研究的基础上不断拓展和深化，多做应用研究，如在已有成果的基础上建设相关数据库，充分利用方言文化资源，大力传承弘扬中华优秀语言文化，做到学术价值与社会价值并重。

本研究为今后湘方言语音研究的提供了理论依据和数据支持，但需要说明的是，可视化软件和文献计量分析相结合的方法，虽然可以直观、全面地揭示湘方言语音研究领域的研究现状、热点和趋势，但其中可能存在一些局限，诸如“标引者效应”的存在、数据规范化处理和相关参数与阀值的设定等，这可能对分析结果有微弱影响，但不会影响基本结论。同时，由于受CNKI文献来源的局限，尚有会议论文和未入库的论文未能考虑到评价体系中，因此在核心作者分析等方面可能不够全面和准确。这些问题的改进有待于在后续研究中不断探索。

参考文献

Beibei Hu, Yang Ding, Xianlei Dong, Yi Bu, and Ying Ding. 2021. On the relationship between download and citation counts: An introduction of granger-causality inference. *Journal of Informetrics*, 15(2):101125.

Jinhyuk Yun, Sejung Ahn, and June Young Lee. 2020. Return to basics: Clustering of scientific literature using structural information. *Journal of Informetrics*, 14(4):101099.

Tibor Braun and András Schubert. 1988. Scientometric versus socio-economic indicators. scatter plots for 51 countries. 1978–1980. *Scientometrics*, 13(1):3–9.

Tibor Braun and András Schubert. 1997. Dimensions of scientometric indicator datafiles: World science in 1990–1994. *Scientometrics*, 38(1):175–204.

陈超美, 陈悦, and 侯剑华. 2009. CiteSpace II:科学文献中新趋势与新动态的识别与可视化. 情报学报, 28(3):401-421.

陈晖. 2006. 湘方言语音研究. 湖南师范大学出版社.

陈悦. 2014. 引文空间分析原理与应用: CiteSpace 实用指南. 科学出版社.

陈悦, 陈超美, 刘则渊, 胡志刚, and 王贤文. 2015. CiteSpace 知识图谱的方法论功能. 科学学研究, (2):244.

范波. 2020. 21 世纪以来我国民族语言政策与规划研究文献的计量分析. 民族学刊.

侯剑华, and 胡志刚. 2013. CiteSpace 软件应用研究的回顾与展望. 现代情报, 33(4):99-103.

孔繁秀, and 孙瑶. 2020. 《中央民族大学学报》学术影响力的文献计量研究——基于CNKI数据分析. 中央民族大学学报*(*哲学社会科学版*)*, (4):168.

李杰. 2016. CiteSpace: 科技文本挖掘及可视化. 首都经济贸易大学出版社.

邱均平, 杨思洛, and 宋艳辉. 2012. 知识交流研究现状可视化分析. 中国图书馆学报, (2):78-89.

田立新. 2015. 中国语言资源保护工程的缘起及意义. 语言文字应用, (4):2-9.

詹伯慧. 2018. 《汉语方言学大词典》说略. 方言, 40(1):1-4.

周春雷, 王伟军, and 成江东. 2007. CNKI 输出文件在文献计量中的应用研究. 图书情报工作, 51(7):124.

周庆生. 2019. 中国语言政策研究七十年. 新疆师范大学学报*(*哲学社会科学版*)*, 6:60-71.

赵蓉英, and 许丽敏. 2010. 文献计量学发展演进与研究前沿的知识图谱探析. 中国图书馆学报, (5):60-68.

宗淑萍. 2016. 基于普赖斯定律和综合指数法的核心著者测评———以《中国科技期刊研究》为例. 中国科技期刊研究, 27(12):1310.

邹微. 2022. 中国彝语文研究的回顾与展望(1992-2021)——基于CiteSpace 的文献计量分析. 民族学刊, 12(10):106-114.

朱亚丽. 2004. 《现代图书情报技术》核心著者测评. 数据分析与知识发现, 20(12):83-84.

附录A.湘方言语音研究核心作者分布情况表

(c) 2022 中国中文信息学会计算语言学专业委员会

排名	综合指数	作者	发文量	被引频次
1	680.80	鲍厚星	8	915
2	240.57	瞿建慧	18	141
3	192.69	彭建国	11	154
4	154.98	李冬香	11	98
5	146.89	李星辉	10	98
6	116.62	李姣雷	13	17
7	113.85	李启群	6	97
8	112.57	蒋军凤	12	23
9	105.76	陈立中	5	97
10	103.77	夏俐萍	8	58
11	102.44	陈山青	9	44
12	91.62	曾毓美	5	76
13	90.26	朱晓农	3	98
14	84.88	陈晖	4	78
15	82.21	彭泽润	6	50

CCL2022

(c) 2022 中国中文信息学会计算语言学专业委员会

基于知识监督的标签降噪实体对齐

苏丰龙，景宁
国防科技大学，长沙，410073
xueshu2021@qq.com

摘要

大多数现有的实体对齐解决方案都依赖于干净的标记数据来训练模型，很少关注种子噪声。为了解决实体对齐中的噪声问题，本文提出了一个标签降噪框架，在实体对齐中注入辅助知识和附带监督，以纠正标记和引导过程中的种子错误。特别是，考虑到以前基于邻域嵌入方法的弱点，本文应用了一种新的对偶关系注意力匹配编码器来加速知识图谱的结构学习，同时使用辅助知识来弥补结构表征的不足。然后，通过对抗训练来执行弱监督标签降噪。对于误差累积的问题，本文进一步使用对齐精化模块来提高模型的性能。实验结果表明，所提的框架能够轻松应对含噪声环境下的实体对齐问题，在多个真实数据集上的对齐准确性和噪声辨别能力始终优于其他基线方法。

Keywords: 图对齐，标签精化，降噪算法，多模态监督

Refined De-noising for Labeled Entity Alignment from Auxiliary Evidence Knowledge

Fenglong Su，and Ning Jing
National University of Defense Technology, Changsha, 410073, China
xueshu2021@qq.com

Abstract

Most existing entity alignment solutions rely on clean labeled data to train models, with little attention to seed noise. To address the noise problem in entity alignment, this paper proposes a labeling noise reduction framework that injects auxiliary knowledge and incidental supervision in entity alignment to correct the seed errors in the labeling and bootstrapping process. In particular, considering the weaknesses of previous neighborhood-based embedding approaches, this paper applies a new dual relational attention-matching encoder to accelerate the structure learning of the knowledge graph while using auxiliary knowledge to compensate for the lack of structural representations. Then, weakly supervised label noise reduction is performed by adversarial training. For the problem of error accumulation, this paper further uses the label refinement module to improve the performance of the model. Experimental results show that the proposed framework can easily cope with the entity alignment problem in noise-laden environments and consistently outperforms other baseline methods in terms of alignment accuracy and noise discrimination on multiple real datasets.

Keywords: Graph Alignment , Label Refinement , De-noising Algorithm , Multi-modal Enhanced

第二十一届中国计算语言学大会论文集，第268页-第280页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

实体对齐(Entity Alignment, EA)旨在识别不同知识图谱中的等效实体，是知识图谱融合的关键步骤，并且在过去几年中得到了深入研究。最近，越来越多的人开始关注利用知识图谱（Knowledge Graph，KG）嵌入技术来解决EA问题，其关键思想是学习KG的向量表示并根据嵌入之间的相似性找到等效实体。具体来说，他们应用TransE(Bordes et al., 2013)或GCN(Kipf and Welling, 2016)来获得每个实体的密集嵌入，然后通过对齐模块将这些嵌入映射到统一的向量空间，最后，实体之间的成对距离决定它们是否对齐。总的来说，这些方法通常包括以下三个步骤：（i）将预先标注的实体种子对指定为锚；(ii) 训练由锚种子和伪标签引导的EA模型；(iii) 根据训练好EA模型去对齐剩余实体对。尽管过去的工作取得了一定的成绩，但是本文仍然从当前的EA研究中观察到了以下几个问题：

(1).对干净种子的依赖。大多数方法依靠预先对齐的锚来连接两个KG，并使用统一的KG结构嵌入来对齐实体。在实践中，这种干净的标记数据的获取成本非常高。如果仅仅依靠结构学习的低质量引导也可能会给模型带来额外的错误，从而促使半监督的方法无法获得准确的嵌入来进行对齐。更重要的是，在现实世界的数据标注过程中，噪声的参与是一个常见的且不可避免的问题。因此，训练数据中存在的标签噪声不应该忽视。

(2).低效的拓扑编码器。以前基于邻域嵌入的EA方法其拓扑编码模块并不稳健，例如，RDGCN(Wu et al., 2019)，NMN(Wu et al., 2020)，HMAN(Yang et al., 2019)和RNM(Zhu et al., 2021)]，当在没有名称嵌入的情况下初始化这些网络时，它们在EA任务中几乎没法工作。GMNN(Xu et al., 2019)的运行效率低下，甚至需要5天才能生成结果(Zhao et al., 2020)。这些方法计算复杂度高，一般为$\mathcal{O}(|E_1||E_2|)$，在大型数据集上更是无法运行。因此，当前EA任务中的拓扑编码效率需要提高。

(3).不灵活的降噪技术。当前EA主要基于等效实体具有相似邻居结构的假设。然而，现实生活中的KG只有少数实体与其他实体紧密相连，其余大多数实体只拥有相当稀疏的邻域结构。也就是说，通过拓扑连接来描述实体的特征是非常片面的，仅仅依靠结构信息来进行实体对齐也是不合理的，因此EA降噪技术不应当局限于三元组噪声(Chen et al., 2019; Huang et al., 2022; Jia et al., 2019)。此外，现有的标记降噪方法并不能很好地解决误差累积的问题。但是，如果从其他模态中学习节点的表示信息却能够使用更多对齐证据来自动纠正错误从而进一步精化标签。

为了克服以上问题，本文探讨了如何将噪声检测与EA模型结合起来训练，提出了一个知识辅助的EA降噪框架以纠正标签错误。特别是考虑到以前基于邻域嵌入方法的弱点，本文应用了一种新的对偶关系匹配编码器来加速KG的结构学习。然后通过对抗强化学习来执行辅助知识的弱监督降噪，接着进一步地使用标签精化模块来改善误差累积的问题。最后，本文在抗噪声的环境下实现了EA任务。本文工作的主要贡献可以总结如下：

- 据本文所知，这是第一个将辅助知识引入标签降噪EA的工作。此外，本文还利用对齐精化模块来解决误差累积的问题。
- 针对以往图编码器效率低下的问题，本文通过新的对偶匹配图编码以更加有效的方式充分考虑了关系信息，同时降低了计算复杂度。
- 通过进行全面的实验评估，本文所提的附带监督的噪声感知EA框架在对齐精度和噪声辨别能力方面始终优于其他最先进的方法。

本文的组织结构如下：第 2节介绍了相关工作。在第 3节中，本文正式定义了降噪EA的任务。第 4节详细阐述了本文的LDEA的框架。在第 5节中，本文介绍了实验结果并进行了详细的分析。第 6节总结了这篇文章。

2 相关工作

关于EA任务，现有的大部分研究都集中在研究如何高效地构建嵌入模型，这些方法试图将KG的实体嵌入到潜在空间中，并计算实体向量之间的距离来作为对齐的证据。这些技术使用的种子都是无噪声的标记数据，很少关注EA中标记和引导过程中的种子错误。本节将介绍与之相关的一些研究。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第268页-第280页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

知识图谱对齐。 知识图谱对齐是扩增知识图谱覆盖范围的最基本任务之一。早期技术利用手工制作特征、众包和基于OWL的等价推理来解决EA问题。由于耗费大量的人力和时间，逐渐被基于嵌入的模型所取代。这些方法应用结构学习来获得每个实体的密集嵌入，然后通过对齐模块将这些嵌入映射到统一的向量空间，最后，实体之间的成对距离决定它们是否对齐。

现有工作主要研究如何学习高质量的实体嵌入表示。根据使用的实体特征，可以分为两类：基于关系的方法和属性增强方法。前者使用知识图谱的关系结构进行表示学习，主流技术主要包括翻译模型(Zhu et al., 2017; Chen et al., 2016; Sun et al., 2019a)、循环神经网络模型(Guo et al., 2019)和图神经网络模型(Wang et al., 2018; Sun et al., 2020; Mao et al., 2020; Wu et al., 2020)。除了基于关系结构的表示学习外，属性增强的方法通过引入实体额外的属性信息来帮助对齐，如实体摘要(Chen et al., 2018)、实体名称向量(Zeng et al., 2020)、图像(Liu et al., 2021)、时间知识(Xu et al., 2021)、实体属性知识(Trisedya et al., 2019; Sun et al., 2017)等。最近也有一些工作有考虑将本体技术与传统实体匹配技术(Xiang et al., 2021)相结合，通过优化表示学习效率(Mao et al., 2021)，并引入主动学习技术(Nayyeri et al., 2021)来辅助对齐等。

知识图谱错误检测。 知识图谱中的错误会对相关应用和知识获取产生负面影响。为了解决这个问题，错误检测最近引起了广泛关注。更具体地说，相关工作包括数值错误检测(Li et al., 2015)，通过来自外部KG的三元组查询进行验证(Wang et al., 2020)，研究KG嵌入学习稀疏性和不可靠性(Pujara et al., 2017)，错误感知的少样本知识图谱补全(Wang et al., 2021)等。但是，以上这些现有方法都是对单个KG开展的研究，在多个KG的融合方面并没有涉及，只涵盖了一些不确定性KG的补全(Chen et al., 2019; Huang et al., 2022)和包含三元组噪声的KG补全(Jia et al., 2019)，并且很少有工作专注于KG对齐过程中的标签错误。

在同质图的相关研究中，错误检测也被广泛关注。如，对稀疏和含噪同质图的分类(Dai et al., 2021; Dai et al., 2022)，基于含噪数据的网络对齐(Huynh et al., 2020)，部分正确种子的子图匹配(Yu et al., 2021)，分布式渗入图匹配(Davalas et al., 2019)，不确定度感知的同质图网络对齐(Zhou et al., 2021)等。值得注意的是，同质图中的结构噪声是通过随机删除边来模拟的，例如，常见的社交网络和人工合成图，这与KG的应用和背景有很大的不同。此外，还有一些工作专注于图上的标签精化，例如，图主动学习(Zhang et al., 2022)、精化邻域的一致性(Heimann et al., 2021)、伪标签精化(Li and Song, 2022)和软标签编辑(Xin et al., 2022)等。

3 问题定义

形式上，一个KG可以表示为$\mathcal{G} = (E, R, T, M)$，其中，$E$和$R$分别是实体和关系的集合。$T \subset E \times R \times E$是事实三元组的集合。$M$是实体附加的模态信息，如实体的图片，描述和属性等。假设$\mathcal{G}_1 = (E_1, R_1, T_1, M_1)$和$\mathcal{G}_2 = (E_2, R_2, T_2, M_2)$是两个KG，并且$\mathcal{LS} = \{(e_i^1, e_j^2), AS | e_i^1 \in E_1, e_j^2 \in E_2\}$是标记种子的集合。因此，常规EA的任务旨在基于预先对齐的标记种子$\mathcal{LS}$和已知的多模态知识M找到新的对齐实体对。在本文中，本文将$AS(e_1, e_2) \in \{0, 0.5, 1\}$用来指示实体对能否对齐的概率。

而降噪EA是给定有噪声的实体对$\mathcal{LS}^U$和可信的实体对$\mathcal{LS}^T$，模型通过学习能够意识到$\mathcal{LS}^U$中的噪声存在（AS = 0），然后找到可信的样本（AS = 1）并对齐剩余的实体对。本文的研究目标是验证所提的框架能够很好地检测标签错误并在抗噪声的环境下对齐实体。

4 本文的框架

本文框架提取了实体的拓扑、视觉、文字和属性信息，来相互补充实体特征的缺失并且辅助实体对齐。在本节中，首先讨论了如何为这四个方面构建原始特征，然后将其输入到框架以促进KG结构学习。

4.1 多模态知识

不同类型的特征从各个方面表征了实体身份。直观地说，融合多模态知识的EA模型优于单模态模型，因为它们聚合了更多信息。(1).对于文字知识，本文使用多语言Bert(Devlin et al., 2019)来获得文字嵌入。(2).对于视觉知识，本文采用ResNet50(He et al., 2016)学习图像嵌入。(3).在这项工作中，本文认为属性信息也可以为判断两个实体是否相同提供重要线索。根据之前的工作(Sun et al., 2017)，本文直接使用属性三元组作为支撑知识。

(c) 2022 中国中文信息学会计算语言学专业委员会

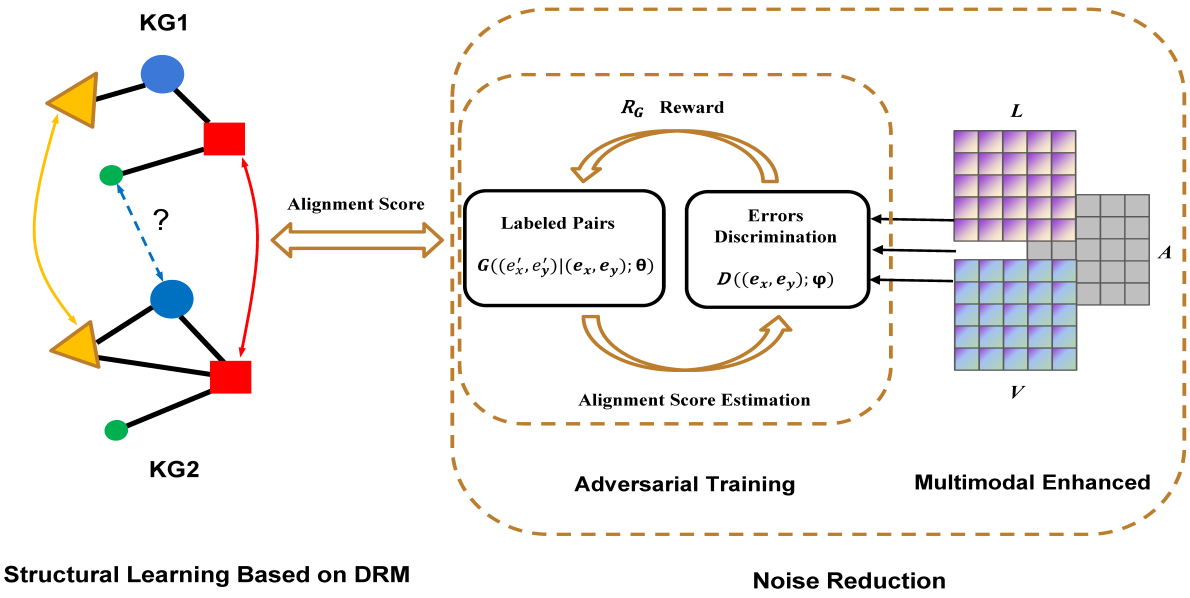


Figure 1: 本文的LDEA 框架。左边是基于DRM 的结构学习模块，右边是带有对抗训练的知识监督降噪模块。

形式上，给定文本向量表示$\boldsymbol{L}$，视觉特征表示$\boldsymbol{V}$和属性特征表示$\boldsymbol{A}$，计算实体对(e_1, e_2)之间的相似度得分，然后利用该相似度得分来预测潜在的对齐实体对。为了计算总体相似度，本文首先计算实体对之间的特定特征相似度得分，即$Sim_{\boldsymbol{A}}(e_1, e_2)$，$Sim_{\boldsymbol{L}}(e_1, e_2)$和$Sim_{\boldsymbol{V}}(e_1, e_2)$，最后来获得实体对$(e_1, e_2)$的融合相似度:

$$Sim_{\boldsymbol{fused}}(e_1, e_2) = \beta_1 Sim_{\boldsymbol{L}}(e_1, e_2) + \beta_2 Sim_{\boldsymbol{V}}(e_1, e_2) + \beta_3 Sim_{\boldsymbol{A}}(e_1, e_2) \tag{1}$$

其中，β_1，β_2和β_3分别代表文本，视觉和属性信息的相似度权重。

4.2 图编码器

本文采用对偶关系匹配编码器(Dual Relational Matching Encoder, DRM)有效地关注它们的邻居，从而进一步地计算每个实体在KG中的隐表示。输入是两个矩阵：$\boldsymbol{H}_e \in \mathcal{Z}^{|E|\times d}$ 表示实体嵌入，$\boldsymbol{H}_r \in \mathcal{Z}^{|R|\times d}$ 表示关系嵌入。实体e_i 在第ℓ层的输出嵌入由以下公式获得:

$$\boldsymbol{h}_{e_i}^{\ell+1} = \tanh\left(\sum_{e_j\in\mathcal{N}_{e_i}}\sum_{r_k\in\mathcal{R}_{ij}} \alpha_{ijk}^{\ell}\left(\boldsymbol{h}_{e_j}^{\ell} - 2\boldsymbol{h}_{r_k}^{T}\boldsymbol{h}_{e_j}^{\ell}\boldsymbol{h}_{r_k}\right)\right) \tag{2}$$

其中，对于α_{ijk}^{ℓ}，本文指定权重:

$$\alpha_{ijk}^{\ell} = \frac{\exp\left(\boldsymbol{v}^T\boldsymbol{h}_{r_k}\right)}{\sum_{e'_j\in\mathcal{N}_{e_i}}\sum_{r_{k'}\in\mathcal{R}_{ij'}}\exp\left(\boldsymbol{v}^T\boldsymbol{h}_{r_{k'}}\right)} \tag{3}$$

其中，$\boldsymbol{v}^T$ 是一个注意力向量。然后，本文通过堆叠更多层来创建全局感知的图表示。所以，实体e_i 的最终嵌入是:

$$\boldsymbol{h}_{e_i}^{combined} = \left[\boldsymbol{h}_{e_i}^{0}\left\|\boldsymbol{h}_{e_i}^{1}\right\|\ldots\|\boldsymbol{h}_{e_i}^{\ell}\right] \tag{4}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

其中，|| 代表级联操作。

为了降低计算成本，本文采用一个代理匹配注意力层来捕获跨图对齐信息，专注于计算所有实体和有限锚点之间的相似度，而不需要计算所有节点到节点的交互。这种交互方式可以将计算复杂度从$\mathcal{O}(|E_1||E_2|)$ 大大降低到$\mathcal{O}(|E_1|+|E_2|)$。该代理匹配注意力层的输入是两个矩阵：$\boldsymbol{H}^{combined} \in \mathcal{Z}^{|E|\times\ell\times d}$ 表示通过简化关系注意层获得的实体嵌入，$\boldsymbol{Q} \in \mathcal{Z}^{n\times\ell\times d}$ 表示具有随机初始化的代理向量，其中n 表示代理向量的数量。

$$\beta_{ij} = \frac{\exp\left(\cos\left(\boldsymbol{h}_{e_i}^{combined}, \boldsymbol{q}_j\right)\right)}{\sum_{k\in\boldsymbol{S}_p}\exp\left(\cos\left(\boldsymbol{h}_{e_i}, \boldsymbol{q}_k\right)\right)} \tag{5}$$

其中，β_{ij} 是每个实体与所有代理向量之间的相似度，$\boldsymbol{S}_p$ 表示代理向量的集合。然后，实体e_i的交叉图嵌入可以计算为：

$$\boldsymbol{h}_{e_i}^{p} = \sum_{j\in\boldsymbol{S}_p}\beta_{ij}\left(\boldsymbol{h}_{e_i}^{combined} - \boldsymbol{q}_j\right) \tag{6}$$

最后，本文采用门控机制将$\boldsymbol{h}_{e_i}^{p}$ 和$\boldsymbol{h}_{e_i}^{combined}$ 结合起来，控制单图和多图之间的信息流：

$$\eta_{e_i} = \text{sigmoid}\left(\boldsymbol{M}\boldsymbol{h}_{e_i}^{p} + \boldsymbol{b}\right) \tag{7}$$

$$\boldsymbol{h}_{e_i}^{final} = \eta_{e_i}\cdot\boldsymbol{h}_{e_i}^{p} + (1-\eta_{e_i})\cdot\boldsymbol{h}_{e_i}^{p} \tag{8}$$

其中，$\boldsymbol{M}$ 和$\boldsymbol{b}$ 是门权重矩阵和门偏置向量。因此，上述编码器可将知识图谱$\mathcal{G}_1$ 和$\mathcal{G}_2$ 的实体映射到相同的嵌入空间中，得到实体的最终嵌入表征：

$$\boldsymbol{H}_{e_1}^{final}, \boldsymbol{H}_{e_2}^{final} = f_{\text{DRM}}\left(\mathcal{G}_1, \mathcal{G}_2\right) \tag{9}$$

附带证据降噪对齐。 值得注意的是，本文采用带有对齐概率的边际损失函数用作降噪对齐优化目标，即：

$$\mathcal{L}_{EA} = \sum_{(e_i,e_j)\in\mathcal{LS}}\sum_{(e_i',e_j')\in\mathcal{LS}'} AS(e_i,e_j)[sim(e_i,e_j) - sim(e_i',e_j') + \lambda]_+ \tag{10}$$

4.3 降噪训练

本文在对抗性训练的启发下(Goodfellow et al., 2014)，通过生成对抗网络来建模降噪过程。包含2个模块：噪声生成器和噪声鉴别器，相互迭代优化。

噪声生成器。 更具体地说，利用上述编码器f_{DRM}中学习到的(e_x, e_y) 的嵌入来生成噪声对(e_x', e_y')，使它们在潜在分布上尽可能地接近真实分布，以便在$D(;\theta)$固定的情况下，使(e_x', e_y')被识别为噪声的可能性最小。因此，本文定义生成噪声实体对(e_x', e_y') 的概率如下：

$$G((e_x', e_y')|(e_x, e_y);\theta) = \frac{exp(f_\theta(e_x', e_y'))}{\sum exp(f_\theta(e_x^*, e_y^*))} \tag{11}$$

$$(e_x^*, e_y^*) \in \mathcal{N}(e_x, e_y) \subset \mathcal{LS}'_{(e_x,e_y)} \tag{12}$$

由于生成器生成的实体对(e_x, e_y) 是离散的，本文使用基于策略梯度的强化学习算法对优化器对其进行优化。对于实体对(e_x, e_y)，其梯度$\mathcal{L}_G$ 可以推导出如下公式：

(c) 2022 中国中文信息学会计算语言学专业委员会

$$
\begin{aligned}
&\nabla_\theta \mathcal{L}_G(e_x, e_y) \\
&= \nabla_\theta \mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[log(1 - D((e'_x, e'_y); \phi))] \\
&= \mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[\nabla_\theta log G(\cdot|(e_x, e_y); \theta) log(1 - D((e'_x, e'_y); \phi))] \\
&\simeq \frac{1}{\eta}\sum_{\eta}^{1} \nabla_\theta log G(\cdot|(e_x, e_y); \theta) log(1 - D((e'_x, e'_y); \phi))
\end{aligned} \tag{13}
$$

其中，最后一个近似相等意味着从当前生成器中采样η个负实体对(e'_x, e'_y)来使用采样近似。更具体地说，(e_x, e_y) 可以看作是状态，$G(\cdot|(e_x, e_y); \theta)$ 是策略，(e'_x, e'_y) 是动作，$log(1 - D((e'_x, e'_y); \phi))$是奖励。因此，生成器（代理）可以通过根据当前状态和策略执行操作来与判别器（环境）交互，然后通过最大化来自环境的奖励作为执行操作的响应来更新自身。此外，本文在奖励函数中引入了一个广泛使用的基线函数$\mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[log(1 - D((e'_x, e'_y); \phi))]$来减少方差，并更新奖励如下：

$$
\mathcal{R}_G = log(1 - D((e'_x, e'_y); \phi)) - \mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[log(1 - D((e'_x, e'_y); \phi))] \tag{14}
$$

因此，噪声生成器使用如下优化目标函数：

$$
\theta^* = \underset{\theta}{argmin} \sum_{(e_i,e_j)\in\mathcal{LS}^T} \mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[log(1 \quad - D((e'_x, e'_y); \phi))] \tag{15}
$$

其中，本文假设$\mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[log(1-D((e'_x, e'_y); \phi))] = \mathcal{L}_G(e_x, e_y)$是一对样本$(e_x, e_y)$的损失函数。

噪声鉴别器。 本文将噪声鉴别器定义为二元分类器，其目标函数是在$G(\cdot|(e_x, e_y); \theta)$ 固定的情况下最大化正确区分正样本与生成负样本的对数似然，定义如下：

$$
\phi^* = \underset{\phi}{argmax} \sum_{(e_i,e_j)\in\mathcal{LS}^T} \mathbb{E}_{(e_x,e_y)\sim\mathcal{LS}^T}[log D((e_x, e_y); \phi)] \quad + \mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[log(1 - D((e'_x, e'_y); \phi))] \tag{16}
$$

$$
D((e_x, e_y); \phi) = \sigma(f_\phi(e_x, e_y)) = \frac{exp(f_\phi(e_x, e_y))}{exp(f_\phi(e_x, e_y)) + 1} \tag{17}
$$

其中，$f_\phi(x, y)$是以ReLU为激活函数的两层神经网络，其输入为$\|x - y\|_1$；$\sigma(x)$ 是Sigmoid 函数。因为f_ϕ 对ϕ 是可微的，所以可以通过随机梯度下降来更新目标函数。

值得注意的是，本文将公式 10 中提到的对齐分数定义如下：

$$
AS(e_x, e_y) = \begin{cases} 1, & \sigma(f_\phi(e_x, e_y)) \geq \delta \\ 0.5, & \sigma(f_\phi(e_x, e_y)) < \delta \in Sim_{\boldsymbol{fused}}(e_x, e_y) > \tau \\ 0, & \sigma(f_\phi(e_x, e_y)) < \delta \notin Sim_{\boldsymbol{fused}}(e_x, e_y) > \tau \end{cases} \tag{18}
$$

其中，$\sigma(f_\phi(e_x, e_y))$是判别器的输出，δ和τ是阈值，用来区分真假实体对。本文将$AS(e_x, e_y)$设置为$\{0, 0.5, 1\}$，分别表示某一对实体不能对齐，不确定和能对齐。此外，集合$\mathcal{LS}^T$ 可以根据$\mathcal{LS}^U$中实体对的对齐分数的变化进行扩充，即在每一轮对抗训练后，从集合$\mathcal{LS}^U$中选择$AS = 1$的实体对加入$\mathcal{LS}^T$集合。

降噪损失。最后本文将整体的降噪损失函数定义如下：

$$
\mathcal{L}_{NR} = \max_{\phi} \min_{\theta} \sum_{(e_i,e_j)\in\mathcal{LS}^T} \mathbb{E}_{(e_x,e_y)\sim\mathcal{LS}^T}[log D((e_x, e_y); \phi)] \quad + \mathbb{E}_{(e'_x,e'_y)\sim G(\cdot|(e_x,e_y);\theta)}[log(1 - D((e'_x, e'_y); \phi))] \tag{19}
$$

详细的算法流程如算法表 1所示。

Algorithm 1 Refined De-noising for Labeled Entity Alignment(LDEA)

Input:
知识图谱$\mathcal{G}_1$和$\mathcal{G}_2$，标记种子集合$\mathcal{LS}^U$ ，可信种子集合$\mathcal{LS}^T$ 。
Output:
$\mathcal{G}_1$和$\mathcal{G}_2$的实体嵌入$\boldsymbol{H}_{e_1}$、$\boldsymbol{H}_{e_2}$
初始化嵌入$\boldsymbol{H}_{e_1}$、$\boldsymbol{H}_{e_2}$；生成器G、鉴别器D和编码器f_{DRM} 的参数;
初始化$\mathcal{LS}^U$ 和$\mathcal{LS}^T$ 的信任分数;
训练文字和视觉模块，得到对应实体的嵌入e_1、e_2;
while 不收敛 **do**
  //基于标记种子集合$\mathcal{LS}^U$和可信种子集合$\mathcal{LS}^T$训练噪声感知的拓扑模型;
  for m=0; $m < m_{EN}$ **do**
    从$\mathcal{LS}^T$ 中采样一批实体对，根据公式10更新$\mathcal{L}_{EA}$
  end for
  从$\mathcal{LS}^T$ 中采样一批实体对，为每个(e_1, e_2) 生成噪声实体对$(e_1', e_2') \sim G(\cdot \mid (e_1, e_2))$;
  for $m = 0$; $m < m_D$ **do**
    根据公式16更新ϕ
  end for
  for $m = 0$; $m < m_G$ **do**
    根据公式15更新θ
  end for
  //标签精化;
  for each $e_1 \in \mathcal{LS}^U$ **do**
    $e_1 \leftarrow NearestNeighbor(e_1, E_2)$
    if $NearestNeighbor(e_2, E_1) = e_2$ and $Sim_{\boldsymbol{fused}}(e_1, e_2) > \tau$ **then**
      $AS(e_1, e_2) = 1$;
    end if
  end for
  更新$\mathcal{LS}^U$ 中实体对的信任分数，将$AS = 1$ 的$\mathcal{LS}^U$中的实体对添加到$\mathcal{LS}^T$ 中;
end while

5 实验

本文使用Keras框架，实验是在配备GeForce GTX 1080Ti GPU 和128 GB 内存的工作站上进行的。本文采用了两个评估指标：Hits@k和MRR，较高的Hits@k 和MRR 分数表示更好的对齐性能。

5.1 数据集

本文在实验中使用了五个数据集。(1) DBP15K：该数据集由DBpedia 构建的三个跨语言子集组成。每个子集包含15,000个用于训练和测试的预对齐锚种子。它们都包含实体图像和摘要。(2) DW100K：该数据集包含实体属性，每个数据集包含两个单语子集，每个子集包含100,000个预对齐的锚种子。本文随机拆分30% 的标记节点对进行训练，其余70% 用于测试。由于上述数据集中给定的标记节点对都是干净的，因此本文需要生成一些噪声数据来替换部分干净数据，以模拟标注噪声。根据最近的工作(Pei et al., 2020)，本文随机破坏40%的训练集作为噪声样本，并将其余60%作为正确样本。然后本文随机选择50%的正确样本作为可信的正标签$\mathcal{LS}^T$，并将剩下50%的正确样本与噪声样本混合作为不可信的标签$\mathcal{LS}^U$。本文的研究就是使用含噪声的训练集$\mathcal{LS}^U$去训练本文的抗噪模型，使其能够从不可信的标签$\mathcal{LS}^U$中尽可能检测出噪声。表 1 描述了数据集的统计信息。

5.2 实验设置

超参数。在这项工作中，本文最佳配置为：嵌入维度d设定为200，间隔参数γ为3.0，GCN层数ℓ为2，用于生成器和判别器的MLP网络有两层，分别有100和30个隐藏单元。本文设定m_{EN}为1500，m_D和m_G都为500，负采样的数量η设定为10。此外，本文将信任分数设定为

(c) 2022 中国中文信息学会计算语言学专业委员会

Table 1: 数据集统计信息

Datasets	DBP15K$_{ZH-EN}$		DBP15K$_{JA-EN}$		DBP15K$_{FR-EN}$		DWY100K$_{W}$		DWY100K$_{Y}$	
	ZH	EN	JA	EN	FR	EN	DBP	Wikidata	DBP	YAGO3
Entities	19388	19572	19814	19780	19661	19993	100000	100000	100000	100000
Relations	1701	1323	1299	1153	903	1208	330	220	302	31
Triples	70414	95142	77214	93484	105998	115722	463294	448774	428952	502563
Images covered	15912	14125	12739	13741	14174	13858	-	-	-	-
Abstracts covered	14840	14954	14946	14946	14923	14926	-	-	-	-
Attributes covered	248035	343218	248991	320616	273825	351094	341770	779402	383757	98028
Anchors	15000		15000		15000		100000		100000	
$\mathcal{LS}^T$+$\mathcal{LS}^U$	1350+3150		1350+3150		1350+3150		30000+70000		30000+70000	

Table 2: 跨语言数据集和单语言数据集的主要结果（40%噪声）

Models	DBP15K$_{ZH-EN}$			DBP15K$_{JA-EN}$			DBP15K$_{FR-EN}$			DWY100K$_{W}$			DWY100K$_{Y}$		
	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR
MTransE	.169	.362	.216	.148	.345	.198	.143	.338	.192	.154	.325	.203	.137	.318	.186
IPTransE	.185	.394	.258	.174	.386	.242	.181	.402	.269	.193	.414	.296	.158	.346	.223
GCN-Align	.223	.424	.316	.223	.439	.321	.231	.462	.337	.331	.487	.392	.376	.525	.448
AlignEA	.263	.457	.342	.254	.451	.338	.278	.471	.357	.331	.487	.392	.376	.525	.448
MuGNN	.274	.471	.361	.279	.481	.368	.284	.485	.372	.348	.503	.417	.401	.554	.475
AliNet	.286	.468	.365	.295	.470	.379	.298	.486	.384	.372	.514	.437	.420	.563	.490
REA-KE	.235	.437	.319	.236	.451	.334	.229	.456	.332	.312	.468	.379	.352	.513	.432
REA	.289	.486	.380	.293	.498	.388	.304	.539	.403	.368	.547	.444	.426	.577	.494
CPUGA-KE	.228	.426	.316	.230	.446	.323	.228	.457	.334	.298	.462	.375	.356	.509	.427
CPUGA	.306	.506	.397	.312	.521	.406	.321	.556	.424	.390	.568	.467	.449	.603	.524
LDEA	**.568**	**.790**	**.668**	**.542**	**.772**	**.645**	**.583**	**.823**	**.689**	**.507**	**.724**	**.607**	**.712**	**.873**	**.784**
Improv.	85.6%	56.1%	68.3%	23.5%	73.7%	48.2%	81.6%	48.0%	62.5%	30.0%	27.5%	30.0%	58.6%	44.8%	49.6%

”Improv.” 表示与SOTA 相比增加的百分比。

离散概率形式$\{0, 0.5, 1\}$，阈值δ为0.01，相似度阈值τ为区间$[0.5, 0.95]$，以便在实体对齐的性能和噪声的影响之间取得更好的平衡。本文使用Adam优化器对公式 16和公式 15中的损失函数进行优化，学习率为0.01，并应用SGD对公式 10中的损失函数进行优化。每个评估重复5次后报告平均结果。其余基线，本文遵循原始论文中的设置来实现最佳性能。

对比模型。为了验证本文提出的方法的有效性，本文将LDEA与几种基于嵌入的方法进行了比较：如MTransE(Chen et al., 2016)、IPTransE(Zhu et al., 2017)、AlignEA(Sun et al., 2018)、GCN-Align(Wang et al., 2018)、MuGCN(Cao et al., 2019) ，AliNet(Sun et al., 2020)，REA-KE(Pei et al., 2020)和CPUGA-KE(Pei et al., 2022)，值得注意的是，这些方法都没有噪声检测模块。另外，本文还将提出的方法与带有降噪模块的方法进行了比较：如REA(Pei et al., 2020)和CPUGA(Pei et al., 2022)。

5.3 主要结果

在表 2 中，列出了所有被评估模型的对齐结果。本文的框架在所有数据集上始终保持最佳性能。在跨语言数据集（DBP15K）上，LDEA 在Hits@1和MRR方面均优于其他方法约23%-85%。在单语言数据集（DW100K）上，与之前的SOTA 相比，性能提升了27%-49% 以上。尽管一些基于嵌入的方法，如GCN-Align、MuGNN和AliNet，使用了先进的图编码技术，但它们仍然会受到给定的标记噪声的影响。由于这些方法没有任何抗噪机制，噪声标签对模型的影响比较大。此外本文的方法在和含降噪模块的方法REA、CPUGA相比也显示出了优越性。实验结果表明，本文的LDEA框架设计在多个真实世界数据集的对齐精度和噪声辨别能力始终优于其他基线方法。

5.4 错误检测定量分析

本文在DBP15K$_{JA-EN}$上测试了$\mathcal{LS}^U$ 的10% - 60%不同强度的噪声。如图 2 所示，所有方法的性能都随着噪声比例的增加而下降。值得注意的是，本文的LDEA曲线的斜率比较平坦，这意味着LDEA模型对于任何比例的噪声数据都具有一定的鲁棒性，应该标签细化模块作用导致了上述结果。实验结果表明，与其他基线相比，LDEA具有很大优势，可以有效地减轻噪声的影响。

本文先后改变$\mathcal{LS}$ 中可信实体的比率来评估不同比率$\mathcal{LS}^T$ 在LDEA上的效果。图 3 表明，

(c) 2022 中国中文信息学会计算语言学专业委员会

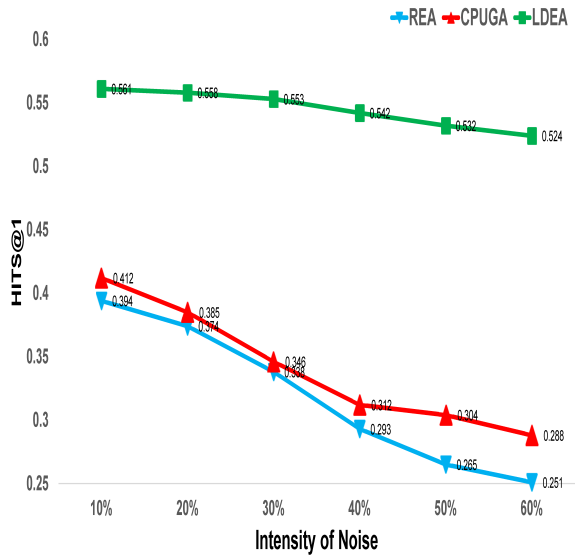


Figure 2: 不同强度噪声在DBP15K$_{JA-EN}$上的结果

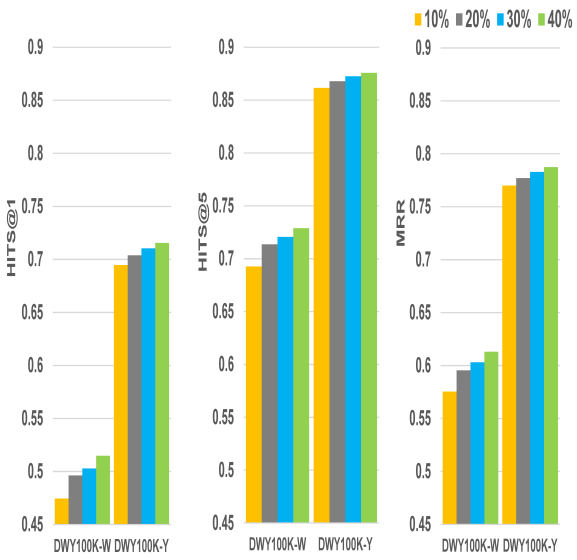


Figure 3: 不同比例正样本在DW100K上的结果

随着$\mathcal{LS}^T$ 在$\mathcal{LS}$ 中的百分比增加，DW100K的性能有所提高。特别是，模型将40% 的$\mathcal{LS}$ 作为可信实体对实现了最佳性能。原因是较大的$\mathcal{LS}^T$ 可以提供更多关于真实实体对的信息，以帮助噪声检测模块将真实实体对与$\mathcal{LS}^U$ 区分开来。相比之下，噪声检测模块却会与较小比例的$\mathcal{LS}^T$ (10%) 形成欠拟合，并导致错误地将一些真实样本识别为噪声。

表 3报告了在20%和40%两种不同强度的噪声中，本文模型的错误检测结果。使用了含标签精化模块的融合判别器，针对$\mathcal{LS}^U$中的实体对进行二元分类，用来区分真实实体对和噪声实体对。高准确率表明大部分的噪声被正确地检测出来。表中的召回率意味着本文的判别器能够在不同的数据集上将$\mathcal{LS}^U$中83%-88%的真实实体对判别为阳性。总的来说，这些结果表明，本文的噪声检测模块表现稳健，对涉及标注噪声的实体对齐任务很有帮助。值得注意的是，由于LDEA使用了标签细化模块，原本在对抗性训练中动摇的标签被辅助知识监督后修正，这大大改善了模型的噪声检测性能。

5.5 消融实验

本小节通过消融实验来检验所提出的模块的有用性。我们在DBP15K$_{JA-EN}$上使用了三种变体：（1）LDEA-w/o-CNS使用随机负采样判别器+标签精化模块，为每个可信的正样本对随机生成负样本对；（2）LDEA-w/o-SNS使用距离最近的负样本进行负采样+标签精化模块。(3) LDEA-w/o-LR，使用随机负采样，不包括标签降噪模块。从表 4，可以看到，本文提出的降噪模块和标签精化模块都是有效的。此外，LDEA-w/o-LR的表现不如LDEA，因为在没有标签精化模块的情况下，一些标签被误判为不可信任的样本。此外，LDEA-w/o-SNS比LDEA-w/o-CNS取得了更好的性能，因为当噪声实体更接近真实数据（干净的标签）时，模型可以工作得更好，而LDEA-w/o-CNS使用简单随机负采样不能为判别器选择利于模型学习的高效负样本。

Table 3: 错误检测性能

Noise	20%			40%		
	Pre.	Rec.	F1	Pre.	Rec.	F1
DBP15K$_{ZH-EN}$	0.946	0.852	0.899	0.962	0.856	0.909
DBP15K$_{JA-EN}$	0.931	0.836	0.884	0.963	0.843	0.903
DBP15K$_{FR-EN}$	0.943	0.835	0.889	0.956	0.868	0.912
DWY100K$_{W}$	0.953	0.861	0.907	0.948	0.880	0.914
DWY100K$_{Y}$	0.941	0.859	0.900	0.937	0.887	0.912

Table 4: 消融实验

Methods	DBP15K$_{JA-EN}$		
	Hits@1	Hits@5	MRR
LDEA	.542	.772	.645
-w/o-CNS.	.526	.737	.613
-w/o-SNS.	.539	.744	.615
-w/o-LR.	.540	.751	.627

5.6 标签精化效果分析

为了验证标签精化的效果，本文使用假阳性率R_{FP}和假阴性率R_{FN}(公式 20)来衡量引导过程中迭代样本的质量。从图 4可以看出，随着epoch的增长，LDEA，CPUGA和REA的假阳性

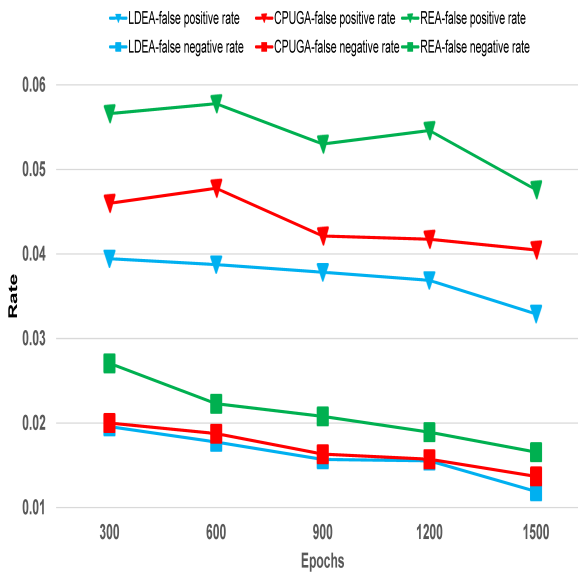


Figure 4: 标签精化效果影响分析

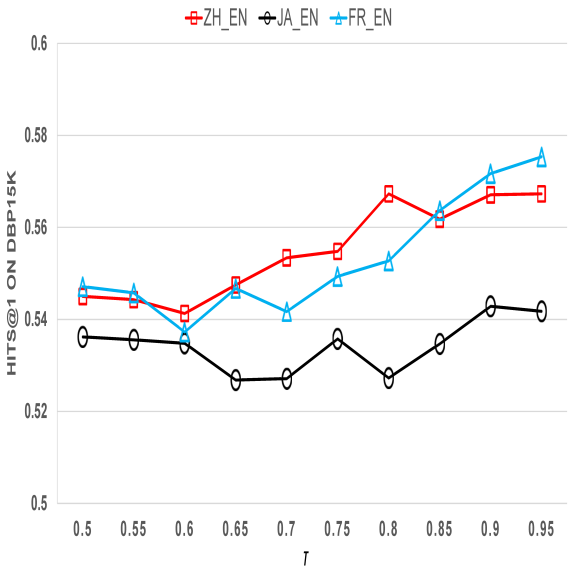


Figure 5: τ的不同取值(x轴)对EA结果的影响

率R_{FP}和假阴性率R_{FN}都在下降，说明三个模型的降噪模块都在发挥作用。然而值得注意的是，两条蓝色曲线始终在最下方，说明本文的LDEA模型整体效果明显好于CPUGA和REA，可能的原因是标签精化模块在降噪对抗训练之后把误判的标签进行了纠正，近一步剔除了负样本，对于整个降噪EA起到了附带的监督作用。

此外，本文还在DBP15K上验证τ的不同取值对EA结果的影响。从图 5中可以看出，随着τ的取值越来越大，对于相似实体对的标签验证越来越严格，从而导致结果稍微提升。本文认为如果τ拥有较低的阈值会给整体的除噪框架带来更多新的不相似实体对(错误的伪标签)从而影响模型的效率，比如部分折线下沉的拐点。总之，从整体上看，τ的不同取值确实会给模型带来一定的影响，但是HITS@1影响效果甚微,尤其在$\text{DBP15K}_{\text{JA-EN}}$数据集上表现一般。最后，本文根据各个数据的多次实践，一般τ的取值大概在0.8至0.9左右。

$$R_{FP} = \frac{\left|\mathcal{LS}^{+}_{\text{iter}} - \mathcal{LS}_{\text{test}}\right|}{\left|\mathcal{LS}^{+}_{\text{iter}}\right|}, R_{FN} = \frac{\left|\mathcal{LS}^{-}_{\text{iter}} \cap \mathcal{LS}_{\text{test}}\right|}{\left|\mathcal{LS}^{-}_{\text{iter}}\right|} \tag{20}$$

6 结论

为了应对EA中标记噪声的挑战，本文提出了一个改进的EA降噪框架，以纠正标记和引导过程中的种子错误。特别是，考虑到以前基于邻域的嵌入方法的弱点，本文应用了一种新的对偶关系匹配编码器来加速KG的结构学习。然后，通过对抗强化学习来执行具有知识弱监督降噪模块。对于误差累积的问题，本文进一步使用对齐细化模块来改进本文的模型。最后，本文在抗噪声的环境下实现了对齐实体的任务。在多个数据集上的实验结果证明了本文降噪框架的优越性。对于未来的工作，本文将尝试在更加复杂的场景下，比如在稀疏和不完备的KG中实现EA抗噪任务。

参考文献

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS 2013*.

Yixin Cao, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li, and Tat-Seng Chua. 2019. Multi-channel graph neural network for entity alignment. In *Proc. of ACL*.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2016. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment.

(c) 2022 中国中文信息学会计算语言学专业委员会

Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2018. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In *Proc. of IJCAI*.

Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. 2019. Embedding uncertain knowledge graphs. In *Proc. of AAAI*.

Enyan Dai, Charu Aggarwal, and Suhang Wang. 2021. Nrgnn: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs. In *Proc. of KDD*.

Enyan Dai, Wei Jin, Hui Liu, and Suhang Wang. 2022. Towards robust graph neural networks for noisy graphs with sparse labels. In *Proc. of WSDM*.

Charalampos Davalas, Dimitrios Michail, and Iraklis Varlamis. 2019. Graph matching on social networks without any side information. In *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proc. of AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL 2019*.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

Lingbing Guo, Zequn Sun, and Wei Hu. 2019. Learning to exploit long-term relational dependencies in knowledge graphs. In *ICML 2019*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR 2016*.

Mark Heimann, Xiyuan Chen, Fatemeh Vahedian, and Danai Koutra. 2021. Refining network alignment to improve matched neighborhood consistency. In *Proc. of SDM*.

Jiacheng Huang, Yao Zhao, Wei Hu, Zhen Ning, Qijin Chen, Xiaoxia Qiu, Chengfu Huo, and Weijun Ren. 2022. Trustworthy knowledge graph completion based on multi-sourced noisy data. In *Proceedings of the ACM Web Conference 2022*.

Thanh Trung Huynh, Van Vinh Tong, Thanh Tam Nguyen, Hongzhi Yin, Matthias Weidlich, and Nguyen Quoc Viet Hung. 2020. Adaptive network alignment with unsupervised and multi-order convolutional networks. In *Proc. of ICDE*.

Shengbin Jia, Yang Xiang, Xiaojun Chen, Kun Wang, and Shijia E. 2019. Triple trustworthiness measurement for knowledge graph. In *Proc. of WWW*.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *NIPS*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks.

Jia Li and Dandan Song. 2022. Uncertainty-aware pseudo label refinery for entity alignment. In *Proceedings of the ACM Web Conference 2022*.

Huiying Li, Yuanyuan Li, Feifei Xu, and Xinyu Zhong. 2015. Probabilistic error detecting in numerical linked data. In *Database and Expert Systems Applications - 26th International Conference, DEXA 2015, Valencia, Spain, September 1-4, 2015, Proceedings, Part I*.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proc. of AAAI*.

Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020. Mraea: an efficient and robust entity alignment approach for cross-lingual knowledge graph. In *WSDM 2020*.

Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. Boosting the speed of entity alignment 10 ×: Dual attention matching network with normalized hard sample mining. In *WWW 2021*.

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

Mojtaba Nayyeri, Sahar Vahdati, Emanuel Sallinger, Mirza Mohtashim Alam, Hamed Shariat Yazdi, and Jens Lehmann. 2021. Pattern-aware and noise-resilient embedding models. In *Proc. of ECIR*.

Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang. 2020. REA: robust cross-lingual entity alignment between knowledge graphs. In *Proc. of KDD*.

Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang. 2022. Graph alignment with noisy supervision. In *Proceedings of the ACM Web Conference 2022*.

Jay Pujara, Eriq Augustine, and Lise Getoor. 2017. Sparsity and noise: Where knowledge graph embeddings fall short. In *EMNLP*.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *ISWC 2017*.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Proc. of IJCAI*.

Zequn Sun, Jiacheng Huang, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019a. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *ISWC 2019*.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019b. Rotate: Knowledge graph embedding by relational rotation in complex space. In *Proc. of ICLR*.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proc. of AAAI*.

Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. 2019. Entity alignment between knowledge graphs using attribute embeddings. In *Proc. of AAAI*.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proc. of AAAI*.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *EMNLP 2018*.

Yaqing Wang, Fenglong Ma, and Jing Gao. 2020. Efficient knowledge graph validation via cross-graph representation learning. In *Proc. of CIKM*.

Song Wang, Xiao Huang, Chen Chen, Liang Wu, and Jundong Li. 2021. REFORM: error-aware few-shot knowledge graph completion. In *Proc. of CIKM*.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proc. of IJCAI*.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2020. Neighborhood matching network for entity alignment. In *ACL 2020*.

Yuejia Xiang, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Zhenxi Lin, and Yefeng Zheng. 2021. Ontoea: Ontology-guided entity alignment via joint knowledge graph embedding.

Kexuan Xin, Zequn Sun, Wen Hua, Wei Hu, and Xiaofang Zhou. 2022. Informed multi-context entity alignment. In *Proc. of WSDM*.

Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. 2019. Cross-lingual knowledge graph alignment via graph matching neural network. In *Proc. of ACL*.

Chengjin Xu, Fenglong Su, and Jens Lehmann. 2021. Time-aware graph neural networks for entity alignment between temporal knowledge graphs. In *EMNLP 2021*.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *EMNLP 2019*.

Liren Yu, Jiaming Xu, and Xiaojun Lin. 2021. Graph matching with partially-correct seeds. *J. Mach. Learn. Res.*

(c) 2022 中国中文信息学会计算语言学专业委员会

Weixin Zeng, Xiang Zhao, Wei Wang, Jiuyang Tang, and Zhen Tan. 2020. Degree-aware alignment for entities in tail. In *Proc. of SIGIR*.

Wentao Zhang, Yexin Wang, Zhenbang You, Meng Cao, Ping Huang, Jiulong Shan, Zhi Yang, and Bin Cui. 2022. Information gain propagation: a new way to graph active learning with soft labels.

Xiang Zhao, Weixin Zeng, Jiuyang Tang, Wei Wang, and Fabian Suchanek. 2020. An experimental study of state-of-the-art entity alignment approaches. *IEEE Transactions on Knowledge and Data Engineering*.

Fan Zhou, Ce Li, Zijing Wen, Ting Zhong, Goce Trajcevski, and Ashfaq A. Khokhar. 2021. Uncertainty-aware network alignment. *Int. J. Intell. Syst.*

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative entity alignment via joint knowledge embeddings. In *Proc. of IJCAI*.

Yao Zhu, Hongzhi Liu, Zhonghai Wu, and Yingpeng Du. 2021. Relation-aware neighborhood matching model for entity alignment. In *Proc. of AAAI*.

CCL 2022

第二十一届中国计算语言学大会论文集，第268页-第280页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

基于图文细粒度对齐语义引导的多模态神经机器翻译方法

叶俊杰[1,2]，郭军军[1,2,*]，谭凯文[1,2]，相艳[1,2]，余正涛[1,2]
1.昆明理工大学信息工程与自动化学院，云南昆明650500
2.昆明理工大学云南省人工智能重点实验室，云南昆明650500
junjieye.cdx@qq.com,guojjgb@163.com, kwtan0909@qq.com,
sharonxiang@126.com,ztyu@hotmail.com

摘要

多模态神经机器翻译旨在利用视觉信息来提高文本翻译质量。传统多模态机器翻译将图像的全局语义信息融入到翻译模型，而忽略了图像的细粒度信息对翻译质量的影响。对此，该文提出一种基于图文细粒度对齐语义引导的多模态神经机器翻译方法，该方法首先跨模态交互图文信息，以提取图文细粒度对齐语义信息，然后以图文细粒度对齐语义信息为枢纽，采用门控机制将多模态细粒度信息对齐到文本信息上，实现图文多模态特征融合。在多模态机器翻译基准数据集Multi30K 英语→德语、英语→法语以及英语→捷克语翻译任务上的实验结果表明，论文提出方法的有效性，并且优于大多数最先进的多模态机器翻译方法。

关键词： 多模态神经机器翻译；图文细粒度；语义交互；对齐语义；Multi30K

Based on Semantic Guidance of Fine-grained Alignment of Image-Text for Multi-modal Neural Machine Translation

Junjie Ye[1,2], Junjun Guo[1,2,*], Kaiwen Tan[1,2], Yan Xiang[1,2], Zhengtao Yu[1,2]
1.Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2.Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China
junjieye.cdx@qq.com,guojjgb@163.com, kwtan0909@qq.com,
sharonxiang@126.com,ztyu@hotmail.com

Abstract

Multi-modal neural machine translation aims to use visual information to improve the quality of only-text translation. Traditional multi-modal machine translation incorporates the global semantic information of images into the translation model, while ignoring the influence of fine-grained information of images on translation quality. In this regard, this paper proposes a multi-modal neural machine translation method based on fine-grained alignment of image-text with semantic guidance. The method firstly interacts image and text information across modalities to extract fine-grained aligned semantic information of image-text. Fine-grained alignment of semantic information is used as a pivot, and multi-modal fine-grained information is aligned to textual information using a gating mechanism to achieve multi-modal feature fusion of image and text. The experimental results on the multi-modal machine translation benchmark dataset Multi30K English → German, English → French and English → Czech translation tasks show that the proposed method is effective and outperforms large Most state-of-the-art multi-modal machine translation methods.

Keywords: Multi-modal neural machine translation , Fine-grained , Semantic interaction , Alignment semantics , Multi30K

第二十一届中国计算语言学大会论文集，第281页-第292页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

多模态神经机器翻译（multi-modal neural machine translation, MNMT）(Caglayan et al., 2019; Yin et al., 2020; Li et al., 2021) 旨在利用额外模态信息（如图像、视频、声音）优化传统的文本机器翻译模型，通过融合图像等多模态特征提升机器翻译的性能(Caglayan et al., 2019; Yao and Wan, 2020; Ye and Guo, 2022)。近年来多模态神经机器翻译受到国内外研究者的广泛关注，相较于传统纯文本机器翻译系统，多模态神经机器翻译通过融合图像模态的信息，不仅可以提高翻译性能，还可以补全文本语义信息以及解决歧义词翻译问题(Huang et al., 2020; Li et al., 2021)。

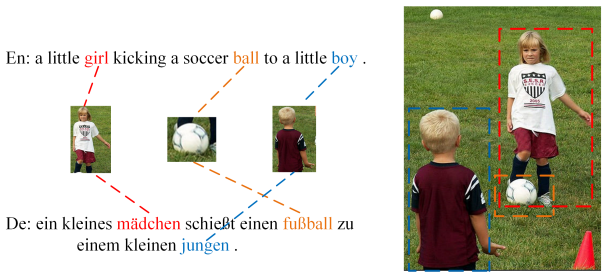


Figure 1: 图像模态和文本模态的语义表示

图像是一种语种无关的模态信息，可以同时被不同语言的人理解，因此可以利用视觉信息为枢轴跨越语言障碍，如上图1 中，不同语言对可以对齐在同一个图像区域，例如：女孩、男孩、足球等。然而，视觉和文本两种模态信息之间存在较大的语义鸿沟，跨模态表示学习及语义对齐通常较难。视觉和文本（源语言）的语义对齐存在两个层级，1）文本实体与图像全局语义信息对齐，图像全局特征可以提供有效的场景信息，例如：图1全局信息为草地上两个小孩子在踢球，2）文本实体与图像对象通过细粒度对齐，图像的细粒度特征信息可以增强文本中的重要信息，例如：图1 中的对象“女孩”、“足球”、“男孩”对齐到文本实体“girl”、“ball”、“boy”。

目前，已有多模态神经机器翻译模型主要通过设计合理的多模态融合模型，实现图文信息的有效整合，主要包含三种图文融合策略：1）跨模态注意力机制(Kwon et al., 2020; Song et al., 2021; Zhao et al., 2021; Li et al., 2021)，将图文模态映射到同一空间向量，然后采用注意力机制抽取与文本信息相关的视觉区域。2）多模态Transformer 融合方法，使用Transformer 分别编码文本特征和视觉特征(Takushima et al., 2019; Nishihara et al., 2020)，然后采用多头注意力机制或者拼接方法将它们融合作为编码器的输出(Yao and Wan, 2020; Gain et al., 2021; Li et al., 2021)。3）门控融合方法(Yin et al., 2020; Lin et al., 2020; Li et al., 2021)，基于多模态门控机制过滤图像中与文本关联性不强的信息，实现图文对齐融合。通过上述方式，文本语义信息和图像语义信息有一个简单的对齐。

然而图像和文本之间存在较大的语义鸿沟，仅仅采用上述图文融合策略，很难实现图像和文本语义的细粒度对齐和融合。为了提升图文对齐融合的能力，本文提出了一种基于图文细粒度对齐语义引导的多模态神经机器翻译方法，采用软对齐的方式实现图文特征的层级融合，图文细粒度对齐语义为枢纽，采用多模态门控机制比对图文两种模态信息，实现图文特征对齐融合，提升了多模态神经机器翻译的性能。与以前的工作相比，本文的主要贡献是两方面的：

- 提出一种图文细粒度对齐语义引导的多模态神经机器翻译方法，采用跨模态注意力机制，以图文细粒度对齐语义为引导，实现了融合图文信息的多模态神经机器翻译。

- 基于多模态机器翻译公共数据Multi30k的实验结果表明，本文提出的模型优于其它多模态机器翻译方法，并显著提高了英德、英法和英捷克语机器翻译的性能。

* 通讯作者：郭军军email地址：guojjgb@163.com
项目基金：国家重点研发计划(2020AAA0107904)；国家自然科学基金(61866020, 61762056)；云南省科技厅自然科学基金项目(2019FB082, 2019QY1801)
©2022 中国计算语言学大会根据《Creative Commons Attribution 4.0 International License》许可出版

2 相关工作

近年来，多模态神经机器翻译受到了广泛的关注，特别是图文多模态融合方法在许多任务中都显示出巨大的潜力。国内外研究学者针对图文多模态机器翻译融合方法开展了许多研究，并取得了一定进展。目前，多模态神经机器翻译主要有基于循环神经网络（recurrent neural network, RNN）的机器翻译模型，与基于Transformer 的机器翻译模型。

2.1 基于RNN的多模态神经机器翻译模型

早期的多模态融合方法主要是基于循坏神经网络（RNN）的seq2seq 框架，利用全局视觉特征初始化RNN 编码器解码器的隐藏状态(Calixto et al., 2017b; Caglayan et al., 2017; Huang et al., 2016)，或利用视觉特征增强文本语义表征能力，提升机器翻译的性能(Huang et al., 2016)。尽管这些方法提高了机器翻译的性能，但视觉特征实际上并没有与文本特征对齐。为了更好地对齐视觉和文本语义特征，Caglayan et al. (2016b; Caglayan et al. (2016a)利用多模态注意力机制同时关注图像及其对应的文本，以对齐视觉和文本语义特征；Calixto et al. (2017a) 分别对源句子单词和图像采用了两种特定于模态的注意机制，以更好地对齐视觉和文本特征。Delbrouck and Dupont (2017) 提出了一种局部视觉注意机制，将局部视觉特征与相应的文本特征对齐。李志峰et al. (2020) 提出一种融合覆盖机制双注意力解码方法，借助覆盖机制分别作用于源语言和源图像，以解决模型过翻译及欠翻译问题。

2.2 基于Transformer的多模态神经机器翻译模型

随着机器翻译技术的发展，基于Transformer结构的多模态神经机器翻译方法被提出。我们将现有的多模态融合策略从三个方面总结如下：**1）**跨模态交互注意机制，Zhao et al. (2022) 利用对象检测特征和额外的区域相关注意机制来融合视觉区域特征和文本特征；Nishihara et al. (2020) 提出了一个有监督的跨模态注意模块，用于对齐文本特征和视觉特征；Song et al. (2021) 在每个Transformer 编码器层采用了一个共同注意图更新模块来对齐多模态特征。**2）**特征连接方法，Yao and Wan (2020) 使用多模态Transformer 来对齐视觉特征和文本特征；Takushima et al. (2019) 拼接视觉全局特征和文本特征作为多模态特征；Takushima et al. (2019) 直接连接文本表示和视觉表示作为多模态表示，以保留细粒度特征并避免模态特定特征的混淆。**3）** 门控融合方法，Yin et al. (2020) 提出了一种基于图的多模态神经机器翻译方法，通过文本图像门控注意力机制提取多模型特征；Lin et al. (2020) 采用门控机制来融合动态上下文引导胶囊网络提取的视觉特征；Li et al. (2021) 使用门控融合方法解决歧义词翻译问题；Zhao et al. (2021) 基于多模态Transformer，提出了一种词域对齐引导的方法来建立文本和视觉特征之间的语义相关性。

3 方法

论文提出了一种图文细粒度对齐语义引导的多模态神经机器翻译模型，模型总体架构如下图2 所示。它主要包含四个网络：图文编码器、跨模态语义交互模块、多模态语义融合模块及解码器。

3.1 图文编码器

3.1.1 图像和文本模态信息表征

不失一般性，将$x_k = \{x_1^k, \ldots, x_n^k\}$ 和z_k 分别表示为源句输入及其对应图像，其中k 表示图文对的序号，n 是x_k 的序列长度。文本序列通过带有位置嵌入的传统嵌入层嵌入，图像特征由预训练的ResNet-101 模型(He et al., 2016) 提取。文本表征向量E_k^x 和视觉表征向量E_k^z 计算如下：

$$\mathrm{E}_k^x = \mathrm{Emb}_x(x_k) + \mathrm{PE}_x(x_k) \quad (1)$$

$$\mathrm{E}_k^z = \mathrm{Emb}_z(z_k) \quad (2)$$

其中，Emb_x 是文本序列词嵌入层，PE_x 是位置嵌入层，Emb_z 是基于ResNet-101 的视觉特征提取层，$\mathrm{E}_k^x \in R^{n \times d_1}$ 和$\mathrm{E}_k^z \in R^{7 \times 7 \times d_2}$，论文中$d_1$=128，$d_2$=2048。

(c) 2022 中国中文信息学会计算语言学专业委员会

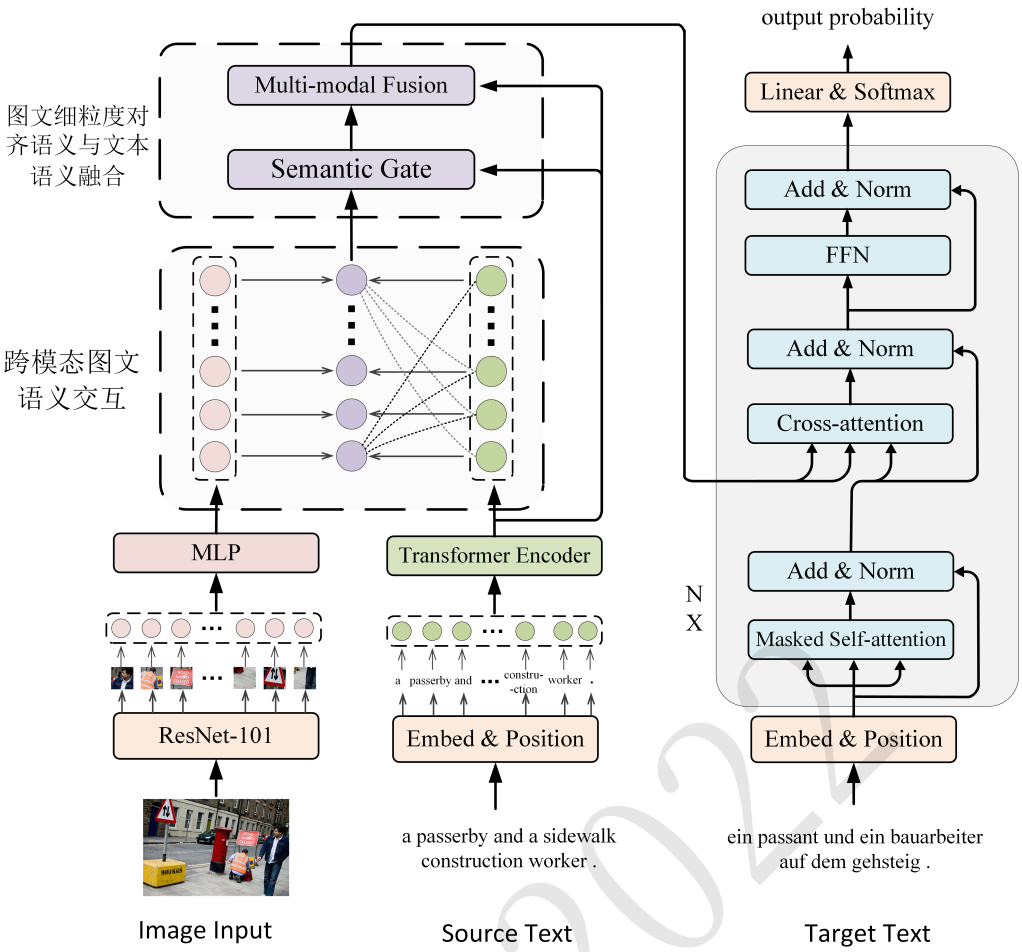


Figure 2: 基于图文细粒度对齐语义引导的多模态神经机器翻译模型

3.1.2 源语言文本编码器

编码器采用传统多头Transformer 编码器，具体框架如图3 所示，每个编码器层由两个子层组成：多头自注意力层和位置前馈网络（FFN）层。首先使用多头自注意力模块，将源文本表示作为查询/键/值矩阵来建立单词到单词的相互连接，可以表示为，

$$\mathrm{H}_{x_k}^{l} = \mathrm{Multihead}(\mathrm{E}_k^x, \mathrm{E}_k^x, \mathrm{E}_k^x) \tag{3}$$

$$= \mathrm{Concat}(head_k^1, \cdots, head_k^M) \tag{4}$$

其中，M 表示头数，Multihead$(\cdot)$ 是多头注意力层，$l = \{0, \cdots, 3\}$ 是Transformer 层索引。多头注意力的输出计算如下：

$$\mathrm{head}_k^{c\in[1,M]} = \sum_{j=1}^{n} \alpha_{ij}(\mathrm{E}_{k_j}^x \mathbf{W}_{k,c}^V) \tag{5}$$

其中n 表示源语言序列x_k的长度，α_{ij}为自注意力权重系数，且为：

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{(\mathrm{E}_{k_i}^x \mathbf{W}_{k,c}^Q)(\mathrm{E}_{k_j}^x \mathbf{W}_{k,c}^K)^{\mathrm{T}}}{\sqrt{d}}\right) \tag{6}$$

其中α_{ij} 是文本特征和文本特征的点积注意力矩阵，$\mathbf{W}_{k,c}^V$, $\mathbf{W}_{k,c}^Q$, $\mathbf{W}_{k,c}^K$ 是参数矩阵。

然后使用FFN 神经网络更新序列每个位置的状态，并得到F_{x_k}，如下所示：

$$\mathrm{F}_{x_k} = \mathrm{FFN}(\mathrm{H}_{x_k}^{l}) \tag{7}$$

CCL 2022

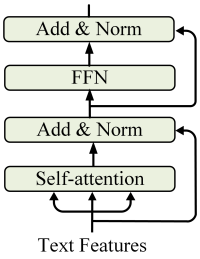


Figure 3: 传统多头Transformer 编码器模型

3.1.3 图像编码器层

图像由预训练的ResNet-101 模型抽取为$7 \times 7 \times 2048$ 维的特征矩阵，并把每个区域特征映射到与文本同一空间，将图像特征转化为一个49×128 维的特征矩阵。如下所示，

$$\mathrm{F}_{z_k} = \mathrm{MLP}(\mathrm{E}_k^z) \tag{8}$$

其中，MLP 是多层感知器。

3.2 跨模态语义交互模块

类似于Nishihara et al. (2020)，论文采用跨模态注意力机制实现文本语义和图像语义特征交互，将源文本语义表征作为查询矩阵，图像语义表征作为键/值矩阵，构建图文细粒度对齐语义表征向量，

$$\mathrm{H}_k = \text{Multihead-Im2te}(\mathrm{F}_{x_k}, \mathrm{F}_{z_k}, \mathrm{F}_{z_k}) \tag{9}$$

$$= \sum_{j=1}^{m} \widehat{\alpha}_{ij} (\mathrm{F}_{z_{k,j}} \mathbf{W}_1^V) \tag{10}$$

$$\widehat{\alpha}_{ij} = \mathrm{softmax}\left(\frac{(\mathrm{F}_{x_{k,i}} \mathbf{W}_2^Q)(\mathrm{F}_{z_{k,j}} \mathbf{W}_3^K)^{\mathrm{T}}}{\sqrt{d}} \right) \tag{11}$$

其中，Multihead-Im2te表示文本语义和图像语义的跨模态注意力机制，$\widehat{\alpha}_{ij}$是文本语义和图像语义的相似度权重，表示第i 个词与第j 个图像区域的相似度，$i \in (1, \cdots, n)$，m 是图像划分区域的个数，论文中m 为49。

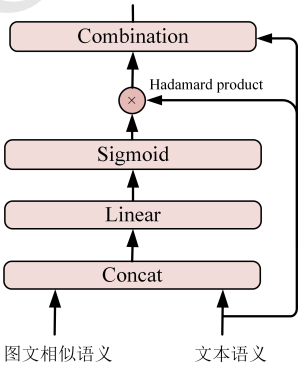


Figure 4: 多模态语义融合模块

3.3 多模态语义融合模块

为了充分的交互多模态语义信息，该文首先以图文细粒度对齐语义信息为枢纽，基于图文细粒度对齐语义信息来进一步交互文本模态和图像模态信息。为此，类似于Yin et al. (2020)，

本文采用门控策略来实现进一步的交互，如图4 所示，

$$\Omega = \text{Sigmoid}(\text{W}_\Omega(\text{H}_k \,||\, \text{F}_{x_k})) \tag{12}$$

$$\widehat{\text{H}}_k = \text{F}_{x_k} \otimes \Omega \tag{13}$$

其中，$\otimes$ 是哈达玛乘积（对应位置元素相乘），W_Ω是模型参数矩阵，|| 表示拼接操作，论文将图像和文本特征在最后一个维度进行拼接，$\widehat{\text{H}}_k$是有用的多模态信息。

然后采用相加的方式实现多模态特征的融合，具体为：

$$\text{O}_k = \widehat{\text{H}}_k + \text{F}_{x_k} \tag{14}$$

最终把编码器的输出O_k 输入到解码器进行解码。

3.4 目标语言解码器

类似地，$t_k = \{t_1^k, \ldots, t_s^k\}$ 表示源语言x_k 对应的目标句子序列，其中s 是t_k 的句子长度，目标句子表征为$\text{E}_k^t = \text{Emb}_t(t_k) + \text{PE}_t(t_k)$。如图2 右图所示，解码器采用传统的多头Transformer解码框架，每个解码器层由三个子层组成：1）掩码多头自注意层；2）跨语言多头注意层；3) 前馈网络层。

首先使用多头自注意力机制对目标句子特征进行提取，可以表示为，

$$\text{Q}_k = \text{Multihead}(\text{E}_k^t, \text{E}_k^t, \text{E}_k^t) \tag{15}$$

然后采用跨语言多头注意力机制实现图文多模态特征O_k 和目标序列特征Q_k 的交互，如下所示，

$$\text{Y}_k = \text{Cross-att}(\text{Q}_k, \text{O}_k, \text{O}_k) \tag{16}$$

然后使用FFN 神经网络更新序列每个位置的状态，并得到F_{d_k}，如下所示：

$$\text{F}_{d_k} = \text{FFN}(\text{Y}_k) \tag{17}$$

最后将解码器最后一层的输出作为softmax输入，通过softmax层预测目标句子的概率分布，可以表示为

$$\text{P} = \text{Softmax}(\text{W}_p\text{F}_{d_k} + b) \tag{18}$$

其中b和W_p是参数，F_{d_k}代表解码器最后一个隐藏状态的输出。

4 实验

数据集： 论文基于多模态神经机器翻译公共数据集Multi30K [-1] 基准数据集的英语→德语、英语→法语和英语→捷克语多模态翻译任务进行实验，其中训练、验证和测试集分别包含29k、1014 和1000 个文本图像对。每张图像都包含一个英文描述句子以及由专业翻译者翻译成的德语、法语和捷克语。论文采用四个测试集来评估提出的的多模态神经机器翻译模型，1) Test2016 测试集[0]，Multi30K 中划分的包含1,000 个示例图文句子对；2）Test2017测试集[1]，WMT2017中包含的1,000个测试图文句子对例子，包含更难翻译和理解的源句；3）我们还使用带有歧义COCO数据集作为域外测试数据，其中包含461个含歧义动词的图文句子对示例，并鼓励使用图像进行消歧；4）Test2018测试集[2]包含1,071个图文句子对实例，该测试集实体词多，低频词多。

数据预处理： 论文采用bpe分词对源语言、目标语言文本进行切分，bpe切割的粒度为6k，每个语言对的词表大小分别为英语→德语（En→De）的5,644→5,876，英语→法语（En→Fr）

[-1]https://github.com/multi30k/dataset
[0]https://www.statmt.org/wmt16/multimodal-task.html
[1]https://www.statmt.org/wmt17/multimodal-task.html
[2]https://www.statmt.org/wmt18/multimodal-task.html

(c) 2022 中国中文信息学会计算语言学专业委员会

的5,644→5,684，英语→捷克语（En→Cs）的5,644→5,972。采用Resnet-101模型对图像特征进行提取得到具有49 个局部空间区域特征的7x7x2048 维向量。
评估指标: 使用广泛用于评估机器翻译质量的BLEU和METEOR两个指标来评估翻译质量，1) 4-gram BLEU 指标(Papineni et al., 2002)，它在准确性和流畅度方面衡量翻译的质量，2) METEOR [3] 指标(Denkowski and Lavie, 2014)，它考虑了翻译质量的精度和召回率。

4.1 实验设置

论文基于Transformer (Vaswani et al., 2017) 搭建机器翻译框架，类似于Wu et al. (2021)，我们的模型堆叠4 层编码器-解码器。本文将编码器和解码器隐藏状态的维度设置为d_{model}=128，前馈网络的内层设置为d_{ffn}=256。学习率设置为0.005。max-tokens 设置为4096，warmup 更新步数设置为2000，标签平滑值设置为0.2。模型采用β_1, β_2 = (0.9, 0.98)的Adam 优化器。模型头数为4，并将dropout 设置为0.3 以避免过度拟合，beam size 设置为5。当BLEU 分数在验证集上的10 个epoch 内没有提高时，模型停止训练。我们采用单个GTX 3090 GPU 训练模型。

4.2 比较模型

为了直观验证本文提出的多模态神经机器翻译模型的优势，该文和以下最近最先进的多模态神经机器翻译模型进行比较，

- VAG-NMT (Zhou et al., 2018): 采用背景注意力机制来利用视觉信息增强模型翻译性能。
- DCCN (Lin et al., 2020): 提出了一种动态上下文引导胶囊网络（DCCN）来引导视觉特征提取以提高机器翻译性能。
- MNMT+SVA (Nishihara et al., 2020): 一种有监督的视觉注意机制，用于捕获与文本相关的视觉区域以进行机器翻译。
- OVC+L_v (Wang and Xiong, 2021): 构建了一个对象级的视觉上下文语义框架，以有效地探索和捕获视觉信息以指导机器翻译。
- WRA-guided (Zhao et al., 2021): 基于多模态Transformer，提出了一种词域对齐引导的方法来建立文本和视觉特征之间的语义相关性。
- IO-MMT (Song et al., 2021): 搭建了一个关系感知图编码器，以充分利用图像和源语句内部的关系，并在目标端提出一个有效的多模态奖励函数，以提高翻译视觉一致性。
- DLMulMix (Ye and Guo, 2022): 提出了一种新型双级交互式多模态混合编码器(DLMulMix)，提取有用的视觉特征来增强文本级机器翻译。

进一步的，为了更公平地证明本文提出的模型的优越性和有效性，在相同的参数设置和训练设备的基础上本文复现了三个最受欢迎的多模态融合方法，

- Gated Fusion MNMT (Li et al., 2021): 一种有效的多模态融合方法，通过增强文本中的重要信息来提升机器翻译性能。该方法广泛应用于多模态神经机器翻译和自然语言处理领域的其他多模态任务中。
- Multimodal self-att (Yao and Wan, 2020): 提出了一种图像感知多模态Transformer 模型来提取有用图像信息以提高机器翻译性能。该方法主要是将文本特征和视觉特征连接起来进行多模态交叉注意。
- Doubly-ATT (Arslan et al., 2018): 在解码器的源-目标交叉注意子层和自注意子层之间使用了一个额外的视觉注意子层，视觉诱发的注意力权重和源语言的注意力权重相加作为双重注意力权重。

[3]http://www.cs.cmu.edu/ alavie/METEOR/

(c) 2022 中国中文信息学会计算语言学专业委员会

Model	Multi30K En→De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
已有的多模态机器翻译模型						
VAG-NMT (Zhou et al., 2018)	-	-	31.6	52.2	28.3	48.0
DCCN (Lin et al., 2020)	39.7	56.8	31.0	49.9	26.7	45.7
MNMT+SVA (Nishihara et al., 2020)	39.9	58.1	-	-	-	-
OVC+L_v (Wang and Xiong, 2021)	-	-	32.4	52.3	28.6	48.0
WRA-guided (Zhao et al., 2021)	39.3	58.3	32.3	52.8	28.5	48.5
IO-MMT (Song et al., 2021)	41.3	59.2	33.5	52.8	-	-
DLMulMix (Ye and Guo, 2022)	41.77	58.93	33.07	51.85	29.90	49.09
基于Fairseq复现的翻译模型						
Transformer (NMT) (Vaswani et al., 2017)	40.96	58.35	32.59	51.21	29.16	48.37
Doubly-ATT (Arslan et al., 2018) †	41.44	59.08	33.15	52.34	29.22	48.41
Multimodal self-att (Yao and Wan, 2020) †	41.50	58.52	32.51	51.33	29.10	48.48
Gated Fusion MNMT (Li et al., 2021) †	41.58	58.88	33.01	51.90	30.04	48.95
Our model	**42.37**	**59.67**	**34.78**	**54.06**	**31.02**	**50.64**

Table 1: Multi30k 英语→德语（En→De）翻译任务在BLEU 和METEOR 指标上的比较结果。† 表示基于我们的Transformer 模型复现以前的多模态融合方法。最佳结果以粗体突出显示。Transformer (NMT) 表示使用纯文本数据进行机器翻译。

4.3 在英语→德语多模态翻译任务上的实验结果

英语→德语多模态翻译任务的实验结果如下表1 所示。本文从三个方面对现有模型进行总结和比较:

1) 与现有的多模态机器翻译模型比较: 实验结果表明，本文提出的模型优于现有的多模态翻译模型，并且在大多数测试集上BLEU 和METEOR 评估指标提高了1∼2 个点。并且，该文提出的模型只需少量参数即可获得出色的结果。根本原因是本文提出的方法可以有效地交互细粒度的多模态语义信息，而现有模型在进行多模态融合时只是简单的整合多模态信息。

2) 与纯文本机器翻译比较: 本文提出的多模态翻译模型在BLEU 和METEOR 指标上显著优于纯文本机器翻译基线，并在所有测试集上提高了大约2 个点。这表明本文提出的多模态机器翻译模型可以有效利用图像信息来增强机器翻译。

3) 与复现的多模态方法比较: 为了更公平地比较本文提出的模型的有效性，基于相同的训练环境，本文复现了最近的三种多模态融合方法。结果被展现在表1，本文的方法在所有评估指标上都比最近的多模态融合方法有了显着改进，这表明深度的交互视觉细粒度语义信息有助于提高翻译性能。

4.4 在英语→法语多模态翻译任务上的实验结果

为了探索所提出模型的稳健性，该文在英语→法语多模态翻译任务上进行实验，结果如表2所示。与英语→德语任务类似，该文提出的模型在英语→法语任务上与现有的多模态翻译模型、纯文本翻译模型和复现的多模态融合方法进行比较，得出以下有趣的结论:

首先，与现有模型相比，本文提出的模型在两个评价指标上仍然取得了显著的提高，这与英语→德语翻译任务的结果是一致的。另外，与纯文本机器翻译基线模型相比，具有图像信息的多模态机器翻译模型取得了优异的结果，这表明本文提出的多模态翻译模型可以有效的与视觉信息交互以增强机器翻译。

其次，在英语→法语任务上复现近期有竞争的多模态融合方法，结果表明本文提出的方法优于复现的多模态融合方法。相比较于现有的多模态翻译模型，本文的模型取得了较好的翻译结果。英语→法语翻译任务的结果再次证明了所提出方法的有效性和普遍性。

4.5 消融实验

为了进一步验证本文提出方法的有效性，我们移除了模型的不同组件，以进行消融研究，结果被报道在表3。分析英语→德语和英语→法语翻译结果可以总结为两点：1）移除图文细粒

(c) 2022 中国中文信息学会计算语言学专业委员会

Model	Multi30K En→Fr					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
已有的多模态机器翻译模型						
VAG-NMT (Zhou et al., 2018)	-	-	53.8	70.3	45.0	64.7
DCCN (Lin et al., 2020)	61.2	76.4	54.3	70.3	45.4	65.0
OVC+L_v (Wang and Xiong, 2021)	-	-	54.2	70.5	45.2	64.6
WRA-guided (Zhao et al., 2021)	61.8	76.3	54.1	70.6	43.4	63.8
IO-MMT (Song et al., 2021)	62.5	76.9	54.9	71.7	-	-
DLMulMix (Ye and Guo, 2022)	62.23	76.85	55.18	**73.37**	44.42	66.41
基于Fairseq复现的翻译模型						
Transformer (NMT) (Vaswani et al., 2017)	60.33	75.64	53.45	71.57	43.61	65.72
Doubly-ATT (Arslan et al., 2018) †	60.94	75.99	53.63	71.56	44.78	65.35
Multimodal self-att (Yao and Wan, 2020) †	61.44	75.77	54.56	71.62	44.59	65.08
Gated Fusion MNMT (Li et al., 2021) †	61.24	76.26	54.15	71.77	44.29	64.91
Our model	**62.73**	**77.34**	**55.56**	73.14	**46.59**	**67.68**

Table 2: Multi30k 数据集上英语→法语（En→Fr）翻译任务的比较结果。

Model	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
En→De 多模态翻译任务						
MNMT_*gat*	41.32	59.16	34.22	53.89	29.69	49.91
MNMT_*att*	41.65	59.36	34.27	53.74	30.64	50.21
MNMT	42.37	59.67	34.78	54.06	31.02	50.64
En→Fr 多模态翻译任务						
MNMT_*gat*	62.46	77.12	55.48	73.04	45.65	67.06
MNMT_*att*	61.98	76.86	54.86	72.57	45.59	66.83
MNMT	62.73	77.34	55.56	73.14	46.59	67.68

Table 3: 模型不同组件的消融实验，MNMT_*att*是移除跨模态语义交互模块的模型，MNMT_*gat*是移除图文细粒度对齐语义与文本语义融合模块的模型，MNMT是本文完整的模型。

度对齐语义与文本语义融合模块，相比于完整的模型，在三个测试集上的两个评估指标都有所下降，这表明细粒度的图文语义交互可以为文本机器翻译提供更精确的信息，特别在MSCOCO测试集上模型的性能有重大的衰退，这表明该模块能有效的利用细粒度图像信息来解决翻译中的歧义性单词。2）移除跨模态图文语义交互模块，相比于完整的模型，模型整体翻译效果都有所下降，特别是在英语→法语多模态翻译任务上BLEU 和METEOR 评估指标都下降超过0.5个分数，这表明，跨模态交互图文语义建立图文细粒度对齐语义信息有助于指导图文多模态对齐融合，提升机器翻译的性能。通过以上的消融实验对比分析，验证了本文模型不同组件的有效性。

4.6 在英语→捷克语多模态翻译任务上的实验结果

为了进一步验证本文方法的有效性和鲁棒性，我们在英语→捷克语多模态翻译任务上评估模型。表4 给出了复现的多模态融合方法和本文的多模态融合方法的BLEU值和METEOR值。可以看到，本文的方法取得了最好的效果，在基线模型（NMT）的基础上取得了+2.01、+3.38的BLEU值提升及+1.11、+2.05 的METEOR值提升。相比于复现的方法，在两个评估指标上本文的方法取得了超过+1 点的BLEU值和METEOR值提升，翻译结果显著提升。这证明了本文方法对于不同语言对是有效且通用的。

4.7 翻译实例分析

为了验证本文方法在翻译过程中确实有效地指导了目标序列的生成，我们通过一些具体

(c) 2022 中国中文信息学会计算语言学专业委员会

En→Cs				
	Test2016		Test2018	
Model	BLEU	METEOR	BLEU	METEOR
Transformer (NMT)	32.70	32.34	27.62	29.03
Doubly-ATT (Arslan et al., 2018) †	33.25	32.28	29.12	29.87
Multimodal self-att (Yao and Wan, 2020) †	33.12	32.01	28.75	29.51
Gated Fusion MNMT (Li et al., 2021) †	33.77	32.24	29.43	29.41
Our model	**34.71**	**33.45**	**31.00**	**31.08**

Table 4: 实验结果在英语→捷克语（En→Cs）多模态翻译任务。


Src： a man *urinating* on a street corner .
MNMT_att： ein mann *urcht* an einer straßenecke .
MNMT_gat： ein mann *uriert* an einer straßenecke .
MNMT： ein mann *uriniert* an einer straßenecke .
Tgt： ein mann *uriniert* an einer straßenecke .


Src： a *bicycle rider* is going down *a long stair* way .
MNMT_att： ein *radfahrer* fährt *eine lange treppe* hinunter .
MNMT_gat： ein *fahrradfahrer* fährt *einen langen treppenweg* hinunter .
MNMT： ein *fahrradfahrer* fährt *eine lange treppe* hinunter .
Tgt： ein *fahrradfahrer* fährt *eine lange treppe* hinunter .


Src： two *men dressed in green* are preparing food in a restaurant .
MNMT_att： zwei *männer in grüner kleidung* bereiten in einem restaurant essen zu .
MNMT_gat： zwei *männer in grün* bereiten in einem restaurant essen zu .
MNMT： zwei *grün gekleidete männer* bereiten in einem restaurant essen zu .
Tgt： zwei *grün gekleidete männer* bereiten in einem restaurant essen zu .


Src： a bride and groom stand together with a *bouquet* in the sunlight .
MNMT_att： eine braut und ein bräutigam stehen zusammen mit einem *strauß* im sonnenlicht .
MNMT_gat： eine braut und ein bräutigam stehen zusammen mit einem *blumenstrauß* im sonnenlicht .
MNMT： eine braut und ein bräutigam stehen zusammen mit einem *bukett* im sonnenlicht .
Tgt： eine braut und ein bräutigam stehen zusammen mit einem *bukett* im sonnenlicht .

Figure 5: En→De 测试集翻译实例，下划线标记表示提升的翻译.

的翻译实例对本文方法的有效性进行验证。图5 是本文选取的一些英语→德语测试集翻译实例。从图中可以看出，本文提出的完整翻译模型（MNMT）翻译效果最好、翻译结果与参考答案基本对齐，有效的提升了预测句子的质量。例如，第一个翻译实例中，本文方法成功的翻译“urinating”到“uriniert”，而缺乏细粒度语义交互的两个模型翻译错误。第二个翻译实例中，MNMT模型准确翻译源语言句，MNMT_*att* 模型翻译文本实体“bicycle rider”失败，验证了跨模态图文语义交互帮助对齐文本实体与图像对象，MNMT_*gat* 模型翻译“a long stair”有轻微错误，验证了图文细粒度对齐语义与文本语义交互帮助对齐图文细粒度信息。第三个翻译实例中，MNMT 模型翻译结果与参考答案相同，MNMT_*att* 和MNMT_*gat* 模型正确翻译单词“men”、“green”，但翻译语序有误，没有对齐目标序列。第四个翻译实例中，MNMT 模型正确翻译“bouquet”为“bukett”，而没有充分语义交互的MNMT_*att* 和MNMT_*gat*模型翻译错误该词。上述实例分析表明，本文方法通过交互文本语义和图像语义细粒度信息，可以显著提高翻译的译文质量。

5 总结

本文提出了一种新颖的基于图文细粒度对齐语义引导的多模态神经机器翻译方法，该方法首先跨模态交互图文信息，以提取图文细粒度对齐语义信息，然后以图文细粒度对齐语义信息为枢纽，采用门控机制将多模态细粒度信息对齐到文本信息上，实现图文多模态特征融合。三个基准多模态翻译任务的实验结果证明本文提出的方法的有效性和优越性，并在三个基准任务上取得了强有竞争力的结果。进一步的消融实验分析表明，本文提出的方法可以深度交互图文

(c) 2022 中国中文信息学会计算语言学专业委员会

多模态语义信息，提取有用的模态细粒度信息以提高机器翻译的性能。在未来的工作中，我们会探索从多模态神经机器翻译模型的解码器方面进行改进，进一步提升多模态神经机器翻译的性能。

参考文献

Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. 2018. Doubly attentive transformer machine translation. *arXiv:1807.11605*.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost Van de Weijer. 2016a. Does multimodality help human and machine for translation and image captioning? *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 627–633, Association for Computational Linguistics*.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016b. Multimodal attention for neural machine translation. *arXiv:1609.03976, http://arxiv.org/abs/1609.03976*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. 2017. Lium-cvc submissions for wmt17 multimodal translation task. *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7–8, 2017, pp. 432–439. doi:10.18653/v1/w17-4746*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. 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 4159–4170, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017a. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada, July. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017b. Incorporating global visual features into attention-based neural machine translation. *In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp. 992–1003. doi:10.18653/v1/d17-1105*.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark, September. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44.

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.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online, July. Association for Computational Linguistics.

(c) 2022 中国中文信息学会计算语言学专业委员会

Soonmo Kwon, Byung-Hyun Go, and Jong-Hyeok Lee. 2020. A text-based visual context modulation neural model for multimodal machine translation. *Pattern Recognition Letters*, 136:212–218.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. *arXiv preprint arXiv:2203.09173*.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4304–4314.

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.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Enhancing neural machine translation with dual-side multimodal awareness. *IEEE Transactions on Multimedia*.

Hiroki Takushima, Akihiro Tamura, Takashi Ninomiya, and Hideki Nakayama. 2019. Multimodal neural machine translation using cnn and transformer encoder. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*.

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.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 2–9.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal 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 6153–6166, Online, August. Association for Computational Linguistics.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Junjie Ye and Junjun Guo. 2022. Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation. *Applied Intelligence*, pages 1–10.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3025–3035*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Region-attentive multimodal neural machine translation. *Neurocomputing*.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium, October-November. Association for Computational Linguistics.

李志峰, 张家硕, 洪宇, 尉桢楷, and 姚建民. 2020. 融合覆盖机制的多模态神经机器翻译. 中文信息学报, 34(3):12.

(c) 2022 中国中文信息学会计算语言学专业委员会

多特征融合的越英端到端语音翻译方法

马候丽[1,2]，董凌[1,2]，王文君[1,2]，王剑*[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
1341584939@qq.com,46761956@qq.com,175360805@qq.com,
1528906057@qq.com,gaoshengxiang.yn@foxmail.com,ztyu@hotmail.com

摘要

语音翻译的编码器需要同时编码语音中的声学和语义信息，单一的Fbank或Wav2vec2语音特征表征能力存在不足。本文通过分析人工的Fbank特征与自监督的Wav2vec2特征间的差异性，提出基于交叉注意力机制的声学特征融合方法，并探究了不同的自监督特征和融合方式，加强模型对语音中声学和语义信息的学习。结合越南语语音特点，以Fbank特征为主、Pitch特征为辅混合编码Fbank表征，构建多特征融合的越-英语音翻译模型。实验表明，使用多特征的语音翻译模型相比单特征翻译效果更优，与简单的特征拼接方法相比更有效，所提的多特征融合方法在越-英语音翻译任务上提升了1.97个BLEU值。

关键词： 端到端语音翻译；特征融合；越南语；语音表征；音高特征

A Vietnamese-English end-to-end speech translation method based on multi-feature fusion

Houli Ma[1,2],Ling Dong[1,2],Wenjun Wang[1,2],Jian Wang[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
1341584939@qq.com,46761956@qq.com,175360805@qq.com,
1528906057@qq.com,gaoshengxiang.yn@foxmail.com,ztyu@hotmail.com

Abstract

Speech translation encoder requires to represent both acoustic and semantic information, a single speech feature whether Fbank or Wav2vec2 feature is insufficient. Based on the difference between hand-crafted Fbank feature and self-supervised wav2vec2 feature, this paper proposes a representation fusion method by cross-attention mechanism, and explores different self-supervised features and fusion method. Multiple features complement each other, strengthen the modelling ability of acoustic and semantic information in speech. In addition, combined with Vietnamese language features, the Fbank feature is used as the main, and the Pitch feature is used as the supplementary, hybrid encoding representation to construct a Vietnamese-English speech translation model. Experiments show that the proposed framework outperforms baselines with

*王剑（通信作者）：1528906057@qq.com
基金项目：国家自然科学基金（61732005, 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, 202002AD080001-5）；云南省基础研究计划（202001AS070014）；云南省学术和技术带头人后备人才（202105AC160018）
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

gains of up to 1.97 BLEU, the speech translation effect of using multi-feature is better than that of single-feature, and the proposed multi-feature fusion method is more effective for performance improvement than the simple feature concatenation.

Keywords: end-to-end speech translation , feature fusion , Vietnamese , speech representation , pitch feature

1 引言

语音到文本的翻译旨在将源语言的语音翻译为目标语言的文本 (Stentiford and Steer, 1990)，广泛应用于会议演讲、跨境旅游、同声传译等各个领域。然而构建语音翻译数据集对标注者和成本要求较高，目前仅有少数语言对有公共语料库，如英语到德语、法语等语言以及少数欧洲语言 (Di Gangi et al., 2019; Iranzo-Sánchez et al., 2020)。大多数语言对缺少语音翻译标注语料，如越南语-英语尚无公开的语音翻译数据集，研究工作相对匮乏，迫切需要开展相关研究。

端到端语音翻译使用一个模型直接将源语言语音映射到目标语言文本，避免了级联方式固有的错误累积、高延迟等缺陷 (Bérard et al., 2016)，因此备受研究者关注。端到端语音翻译模型同时进行跨模态跨语言的映射，且训练数据较稀缺，翻译性能与级联模型仍存在较大差距。此外，语音数据受说话人情绪、音量、口音和外界噪声等因素产生多变性 (Han et al., 2021; Liu et al., 2020)，限制了端到端语音翻译模型的性能。而特征提取是语音翻译的重要步骤,特征的好坏直接影响翻译的效果。因此，探索有效的语音表征对于语音翻译任务至关重要。

语音翻译中常使用人工设计的Fbank特征或基于自监督的Wav2vec2特征 (Baevski et al., 2020)作为模型输入。如图 1左侧为Fbank特征的提取过程，在一次分帧的基础上，进行逐帧变换。在采样率为16K、帧长为25ms、帧移为10ms的设置下，每帧Fbank特征覆盖400个采样点，能够表示语音声学信息中的局部特征，但Fbank特征提取方法根据人声预先设置，不可动态学习，具有一定的局限性。图 1右侧为Wav2vec2特征提取过程，Wav2vec2模型通过堆叠的7层不同步长和卷积核大小的卷积神经网络，进行逐层循序计算提取特征。每帧Wav2vec2特征覆盖3240个音频采样点，堆叠的CNN增大了语音声学信息的覆盖范围，有利于对语音中语义信息进行表征和学习。但由于Wav2vec2表征在大规模无标注语音上进行自监督预训练，表征在目标任务上的表现性能高度依赖训练域和目标域的相关性 (Berrebbi et al., 2022)。单一的Fbank或Wav2vec2特征作为模型的输入时，不能满足模型同时对语音中声学信息和语义信息建模的需求。

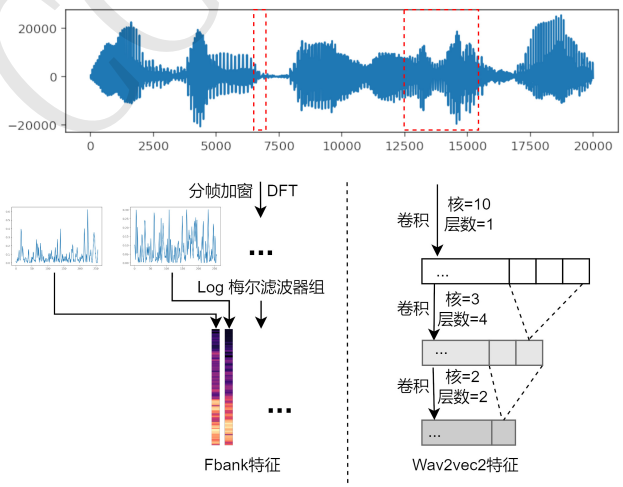


图 1. Fbank和Wav2vec特征提取过程比较

越南语属音调语言，以单音节为主，每个音节的元音带有6种音调之一。音调不同则词义不同，音调错误会产生句子歧义 (Nga et al., 2021)，音调信息通常使用音高特征（Pitch）

(c) 2022 中国中文信息学会计算语言学专业委员会

表示。对于越南语而言，音调信息在一定程度上有助于表征语音中的声学信息和语义信息。 Huy Nguyen (2019)在越南语语音识别的实验上也表明，融合音调信息的语音特征能够提高识别的准确性。因此，音调信息对于构建越南语-英语语音翻译模型至关重要。

为此，本文对Fbank以及预训练Wav2vec2特征进行实验，分析其在越南语-英语语音翻译和越南语语音识别任务上的表现性能。实验表明不同特征表现效果及其可视化结果均存在差异性。基于两种特征间的差异，提出通过对Wav2vec2特征和Fbank特征分别进行编码后，使用交叉注意力机制进行特征融合的语音翻译模型，帮助编码器同时编码语音中的声学信息和语义信息，弥补单一特征表征能力不足的缺陷。并使用混合Pitch的编码块和Fbank编码块交替编码的方式，在Fbank特征编码的过程中显式的加入越南语音调信息，提高编码器对于音调信息的敏感性，进一步辅助编码器对语义信息的建模。此外，本文还探究了不同的特征融合方式对于语音翻译性能的影响, 为语音翻译任务选择最优融合方式。

本文贡献如下：（1）探究和分析了Fbank特征和Wav2vec2特征在越南语语音识别和越-英语音翻译任务上的表现。（2）基于Wav2vec2特征和Fbank特征间的差异，提出Wav2vec2和Fbank特征相融合的方法进行不同特征间的相互补充，加强编码端的表征能力，并比较了不同融合方法和特征对语音翻译效果的影响。（3）结合越南语语言特点，使用不同编码块交替编码的方式在Fbank特征中加入越南语音高特征，增强语音翻译模型对声学信息和语义信息的编码能力。（4）提出多特征融合的语音翻译框架，将Fbank、Wav2vec2、Pitch三种特征进行有效融合构建编码端语音表征，提升了越南语-英语语音翻译质量。

2 相关工作

2016年， Bérard et al. (2016)提出端到端的语音到文本的翻译的设想。随着深度学习的发展，端到端的语音翻译逐渐得以实现， Duong et al. (2016)建立直接的语音翻译模型，在不使用任何源语言文本前提下，使用单个模型建立源语言语音到目标语言文本的映射。但端到端的语音翻译模型面临着严重的资源不足问题，为此研究者们分别从数据层面和方法层面来缓解资源不足和映射困难。数据层面主要从音频数据增强和数据合成两方面弥补资源不足问题，其中语音增强是语音类任务普遍使用的方法，包含SpecAugment对语音的语谱图在时域和频域进行掩蔽防止模型过拟合 (Park et al., 2019)，对原始语音进行变速等操作。数据合成分别通过为文本翻译语料生成音频数据、为语音识别语料生成目标本文数据，以及为语音翻译语料生成多说话人音频数据等方式 (Post et al., 2013)。但合成大规模语料需耗费大量时间和成本，同时由于机器合成数据分布单一，需要设计合适的比例将合成数据与真实数据混合，且对模型性能的提升仍非常有限。方法层面的探索主要通过不同方法引入额外数据、训练子模块来提高语音翻译的性能。 Anastasopoulos and Chiang (2018)和 Liu et al. (2020) 采用多任务和预训练-微调的方法，通过整合额外的ASR和MT数据来提升语音翻译的性能。此外，研究者们还在课程学习、元学习、知识蒸馏等深度学习方法上进行探索 (Kano et al., 2017; Liu et al., 2019; Indurthi et al., 2020)。这些方法借鉴了相关领域的方法，都能在一定程度上提高语音翻译的性能。越南语-英语的语音翻译语料和文本翻译的训练语料有限，但现有多任务和预训练的方法不能直接迁移，且使用单一的语音特征，对语音中的声学和语义信息的表征存在一定局限性。因此本文提出多特征融合的越-英语音翻译方法，旨在通过融合自监督语音表征，结合越南语语音的音高特征，对传统Fbank特征进行补充，在不使用额外数据以及额外训练步骤的情况下，最大程度提高越英语音翻译的性能，降低训练成本的同时满足低资源语言语音翻译的需求。

语音特征表示的好坏会直接影响语音翻译的效果，语音信号的不确定性以及噪声增加了语音特征提取的难度。传统的语音特征提取采用信号处理的方法对语音信号进行频域分析，目前使用较多的是Fbank和MFCC两种特征。MFCC特征由于DCT变换造成一定程度的语音信息丢失，因此基于深度神经网络的语音识别和语音翻译模型中更多使用的是Fbank特征 (Mohamed, 2014) 。但由于实际的语音数据以及场景的复杂性，人工设计方式提取的语音特征在一定程度具有局限性。近几年在语音识别和说话人识别领域出现使用基于深度学习的方式提取语音特征 (Tüske et al., 2014)。由于神经网络可以从原始波形中提取出更合适的语音表征，研究者们基于CNN构建声学模型，直接使用原始波形作为输入 (Ravanelli and Bengio, 2018)。由于原始波形是长序列数据，训练时对内存资源和硬件要求高，使得训练过程更加具有挑战性。 Baevski et al. (2020)采用自监督的方法在大量无标注的原始音频上学习Wav2vec2语音表征，通过多层卷积神经网络和Transformer模型提取语音表征，使用该表征在下游语音识别任务

(c) 2022 中国中文信息学会计算语言学专业委员会

中经少量标注数据上进行微调，在Librispeech的测试集上实现了4.2%的词错率。

与以往使用单一语音特征的语音翻译模型相比，本文根据Fbank特征和Wav2vec2特征的差异性，探索两种特征的融合方法，并结合越南语语音特点有效融入语音音高特征，提升越南语-英语语音翻译模型效果。

3 方法

本文提出一种有效融合多特征的越-英语音翻译模型，模型由编码器和解码器组成。对越南语音频分别提取Fbank特征、Wav2vec2特征和Pitch特征，作为编码器输入。编码器由交替特征编码层、Wav2vec2编码层和表征融合层组成，使用交替特征编码层对Fbank和Pitch两种频谱特征进行混合编码输出频谱表征，通过加入越南语的音调信息增强模型对语义信息的表征能力；同时使用Wav2vec2编码层对Wav2vec2特征进行编码输出自监督表征；将频谱表征和自监督表征输入到表征融合层，通过交叉注意力机制学习不同类型表征间的对齐和融合，得到最终的编码器输出表征。最终，将编码器输出表征输入解码器，输出目标语言文本词序列，具体过程如下所述。

3.1 越南语音频多特征提取

音频特征提取是语音翻译模型构建的重要基础环节，模型需要根据输入的语音特征同时对声学信息和语义信息建模，使用单一特征同时对两类信息进行建模存在较大挑战。本文对音频序列分别提取Fbank特征、Pitch特征以及Wav2vec2三种特征，作为模型的输入。其中，Fbank特征和Pitch特征为人工特征，Wav2vec2特征为自监督方法所提取的特征。训练语料为$S = \{(x, y)\}$，其中$x = (x_1, \ldots, x_m)$为音频序列，$y = (y_1, ..., y_n)$为目标语言文本序列，m和n分别为源音频序列和目标文本序列的长度。特征提取过程如式 (1)所示：

$$i = \text{Extractor}_i(x) \in \mathbb{R}^{d_i}, i \in \{FilterBank, Pitch, Wav2vec2\}, \tag{1}$$

Fbank特征：使用torchaudio包[1]，设置帧移为10ms，帧窗口大小为25ms，提取80维的Fbank特征序列为$f = (f_1, f_2, ...f_{l_f})$，其中$f \in \mathbb{R}^{d_f}$，$d_f$为Fbank特征维度，$l_f$为序列长度；

Pitch特征：使用pySPTK工具[2]中的SWIPE算法进行提取，搜索频率范围设置为50Hz至400Hz，提取的Pitch特征序列为$p = (p_1, p_2, ...p_{l_p})$，其中$p \in \mathbb{R}^{d_p}$，$d_p$为Pitch特征维度，$l_p$为序列长度；

Wav2vec2特征：开源的w2v2-vi模型[3] 在100小时的越南语有声读物进行预训练，使用该模型的第7层CNN输出的512维向量进行实验，特征序列为$w = (w_1, w_2, ...w_{l_w})$，其中$w \in \mathbb{R}^{d_w}$，$d_w$为Wav2vec2特征维度$d_w$，$l_w$为序列长度。

3.2 多特征融合编码器

与以往对单一特征进行编码的语音翻译模型不同，本文提出在编码端对Fbank、Wav2vec2以及Pitch三种语音特征进行编码，模型编码器由Wav2vec2特征编码层、Fbank-Pitch交替特征编码层和表征融合层三部分组成。其中，利用Fbank-Pitch交替特征编码层显式加入Pitch特征，进一步辅助编码器对语义信息的建模。多特征融合编码器如图 2所示。

Fbank-Pitch交替特征编码层：Fbank特征序列f经下采样$D(\cdot)$后，叠加位置编码pos_f，与下采样后的Pitch特征序列$D(p)$共同作为交替特征编码层的输入，编码输出隐层状态序列h_1，如式 (2)，下文简称该序列为频谱表征；

$$h_1 = \text{AlternatedEncoder}(D(f) + pos_f, D(p)) \tag{2}$$

Wav2vec2特征编码层：对于Wav2vec2特征序列w，使用CNN作为编码器，并通过维度转换，得到隐层状态序列h_2，下文简称自监督表征，编码过程如式 (3)所示；

$$h_2 = \text{Wav2vec2Encoder}(w) \tag{3}$$

[1] https://pytorch.org/audio

[2] https://github.com/r9y9/pysptk

[3] https://huggingface.co/dragonSwing/wav2vec2-base-pretrain-vietnamese

(c) 2022 中国中文信息学会计算语言学专业委员会

表征融合层：本模块输入为频谱表征h_1和自监督表征h_2，使用交叉注意力机制进行表征间的融合和对齐，输出融合表征向量。在不增加表征长度和特征维度的情形下，通过交叉注意力机制自动学习两种表征间的对齐，进行相互补充和增强。特征融合过程如式 (4)所示。

$$h_x^A = \text{FusionLayer}(h_1, h_2) \tag{4}$$

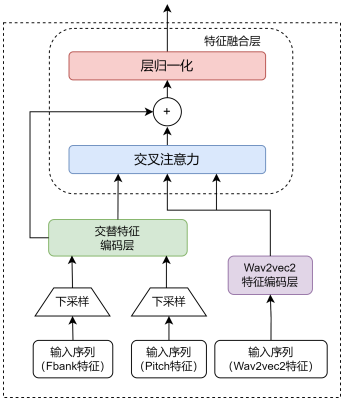
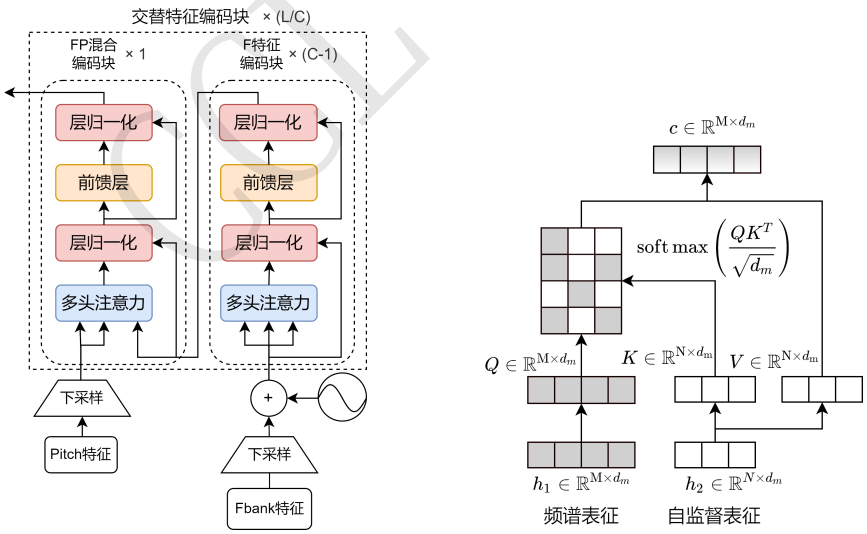


图 2. 多特征融合的编码器

3.2.1 Fbank-Pitch交替特征编码层

越南语中，音调可用于区分词义，语音中的音调信息通过音高特征（Pitch）表示。在语音特征中显式加入音调信息可增强模型对于语义信息的建模能力。不同于以往使用单一的Fbank特征作为Transformer编码块输入的工作，本文根据越南语的语音及音调特点，以Fbank特征为主，Pitch特征为辅，使用两种编码块交替编码的方式进行特征编码。两种编码块包括以Fbank作为输入，基于自注意力的Transformer编码块，和以Fbank和Pitch作为输入，基于交叉注意力的Transformer编码块，下文简称F特征编码块（F-Block）和FP混合编码块（FP-Block）。交替编码的方式在不增加编码块个数和模型复杂度的基础上，融合Pitch信息进行更有效的编码。



(a) Fbank-Pitch交替特征编码层　　(b) 特征融合层中的交叉注意力机制

图 3. 多特征融合编码器子模块

如图 3（a）所示，交替特征编码层共包含L个编码块，交替周期为C，则包含(L/C)个交替周期。每个交替周期内有$(C-1)$个F-Block和1个FP-Block, 其中i为当前编码块的块数，编码块的设置如式 (5)。F-Block专注对Fbank特征进行编码，而FP-Block采用交叉注意力机制同时

(c) 2022 中国中文信息学会计算语言学专业委员会

对Pitch特征和Fbank特征进行混合编码。F-Block和FP-Block均对Fbank特征进行编码，间隔多个块对Pitch特征进行编码，该设计与实际发音相符，即区分词义和句义主要通过不同音素的发音，辅以不同的音高。本文设置$C=3$，$L=12$，具体说明见4.1.3。

$$Block_i = \begin{cases} F-Block, & i\%C \neq 0 \\ FP-Block, & i\%C = 0 \end{cases} \tag{5}$$

3.2.2 基于多头交叉注意力的表征融合层

Fbank特征根据人耳对声学信号的感知，手工设计结构来提取特征，对复杂的音频的特征提取具有局限性；基于自监督-预训练方式得到的自监督表征缺乏对具体任务和数据的适应性。为更好的对音频进行表征，在编码器中，将带有音调信息的频谱表征和自监督表征使用交叉注意力机制进行融合，使得不同类型特征间相互补充，满足语音翻译任务需要同时对声学信息和语义信息建模的要求。

将Fbank-Pitch交替特征编码层输出的频谱表征h_1，和Wav2Vec2特征编码层输出的自监督表征h_2，通过多头交叉注意力机制进行特征融合。多头交叉注意力计算过程如图 3（b）所示，将频谱表征h_1作为q，自监督表征h_2作为k和v，首先通过式 (6)的线性变换分别得到向量Q,K,V，其中W_i^Q, W_i^K, W_i^V均为随机初始化的参数矩阵；

$$Q = qW_i^Q, K = kW_i^K, V = vW_i^V \tag{6}$$

然后经过式 (7)计算单头注意力得到向量序列$head_i$，其中d_m模型的隐层维度，与向量Q,K,V的维度相等；

$$\begin{aligned} head_i &= \text{Attention}(Q,K,V) \\ &= \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \end{aligned} \tag{7}$$

再将各个头的向量序列经过式 (8)运算进行拼接，输出音频序列的向量表征c，其中h为多头注意力的头数，W_i^O为随机初始化的参数矩阵；

$$\begin{aligned} c &= \text{MultiHead}(Q,K,V) \\ &= \text{Concat}(head_1, head_2, \cdots, head_h)W^O \end{aligned} \tag{8}$$

最后，通过残差网络将c与h_1相加后经层归一化得到最终的编码器输出表征h_x^A，如式 (9)所示。

$$h_x^A = \text{LayerNorm}(c + h_2) \tag{9}$$

3.3 解码器

本文使用的解码器遵循Transformer 解码器的模型架构。首先将目标语言的文本序列映射为词嵌入向量，继而通过自注意力网络和交叉注意力网络关注到编码器的输出，经过前馈神经网络映射得到解码器的输出，如公式 (10)所示:

$$\begin{aligned} h_e &= \text{Emb}(y) \\ h_y &= \text{Decoder}(h_x, h_e) \end{aligned} \tag{10}$$

通过softmax函数将输出映射到目标语言的词表上，得到目标语言文本序列的预测概率，如公式 (11)所示:

$$p(y|x) = \text{softmax}(h_y) \tag{11}$$

最终，整个语音翻译模型的损失函数如式 (12)所示。

$$L_{st} = -\sum_{(x,y)\in S} \log p(y|x) \tag{12}$$

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

4 实验设置与结果分析

4.1 实验设置

4.1.1 数据集

实验所采用的数据集来自VLSP2019[4]中的越南语语音识别数据集，该语料包含约416小时的有声小说音频以及人工校对的越南语文本，具体设置见表 1。为进行越南语到英语的语音翻译，调用Google机器翻译服务将越南语文本翻译为英语文本。

数据集	时长/h	句数/k
训练集	395.5	300.0
开发集	13.1	10.0
测试集	7.2	5.5

表 1. 实验数据集

4.1.2 数据预处理

对于语音的Fbank特征，在训练集上使用SpecAugment的LB策略增强语音数据 (Park et al., 2019)，以保证更好的泛化性和鲁棒性。其中，语音的Fbank特征和Pitch特征序列使用均值和方差归一化处理。对于语音的Wav2vec2特征，使用开源的w2v2-vi模型提取512维的语音特征进行实验，下文称w2v2-vi特征。由于普通Transformer模型自注意层的计算复杂度为输入长度的平方，为了对输入数据更有效的计算，实验中采用卷积神经网络对输入序列进行下采样，通过设置不同的层数合理控制输入模型的序列长度，使不同特征的序列长度基本保持一致。所有的卷积层均使用相同配置，步长为2，卷积核大小为5，Wav2vec2、Fbank和Pitch特征下采样的输出维度分别为256、256和32。过滤了小于5帧大于3000帧的音频。

对于目标语言文本，区分大小写同时保留标点。句子使用词表大小为4k的Unigram SentencesPiece模型 (Sennrich et al., 2015)进行分词，采用256维的词嵌入并叠加了位置嵌入。

4.1.3 模型配置与评价指标

为保证实验的公平性，本文所进行的实验均基于Fairseq的Transformer-S2T-S框架[5]，所提方法基于该框架实现。模型的基本配置中，编码器有12层，解码器有6层，多头注意力头数为4，隐层变量维度为256，前馈网络的维度为2048，dropout为0.1。所有实验的训练配置参数均为：使用Adam优化器 (Kingma and Ba, 2014)，其中$\beta_1 = 0.9, \beta_2 = 0.997$；使用标签平滑率为0.1的交叉熵损失作为目标函数 (Müller et al., 2019)；学习率最大阈值为1e-3，学习率预热为10000，使用inverse sqrt 动态调整学习率。整个训练过程在1张Tesla T4 GPU上进行。解码使用大小为5的束搜索算法，使用区分大小写的ScareBLEU[6]作为模型性能的评价指标。

为选定最优的交替周期，选择训练集的20%，将交替特征编码器作为编码器进行初步的语音翻译实验。交替特征编码块中有12个编码块组成，为保证两种编码块均匀分布，交替期周期C分别在2、3、4、6中选择，其中Pitch特征编码比例依次减少。实验结果如表 2所示，其中FP/F表示编码器中FP-Block与F-Block总个数比例。由表可知，在交替周期为3时，编码器获得最佳翻译效果，下文实验均采用该设置。

C	2	3	4	6
P/FP	6/6	4/8	3/9	2/10
BLEU	7.32	8.30	7.96	8.09

表 2. 不同交替周期C在测试集上的BLEU值

[4] https://vlsp.org.vn/
[5] https://github.com/pytorch/fairseq
[6] https://github.com/mjpost/sacrebleu

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 实验结果

4.2.1 Fbank特征与Wav2vec2特征在ASR和ST的比较

遵循fairseq的端到端语音到文本 (Wang et al., 2020)的模型设置，先在该编码器-解码器模型上，分别使用Fbank特征和w2v2-vi特征进行语音翻译和语音识别任务，对两种特征均采用2层卷积网络进行下采样。

特征	ASR(WER)	ST(BLEU)
Fbank特征	3.98	37.59
w2v2-vi特征	4.13	36.89

表 3. 特征比较实验

在测试集上的实验结果如表 3所示，Fbank特征在ASR任务上的词错率较w2v2-vi特征低0.15，在ST任务上的BLEU值高0.98。在高资源设置下，使用越南语语音的Fbank特征在ASR任务和ST任务上的性能略优于w2v2-vi特征。

Nguyen et al. (2020)验证了自监督表征在低资源设置下(小于100小时)明显优于fbank表征，在中等资源设置下两种特征的效果接近（100小时至300小时）。在附录 A中，进一步对Wav2vec2特征和Fbank特征，在MuST-C的英语-越南语数据集在不同资源设置下进行比较，实验结果在中等资源设置下两种特征在语音翻译上的性能差异较小，同时表明在高资源设置下（大于300小时），Fbank特征优于Wav2vec2特征。

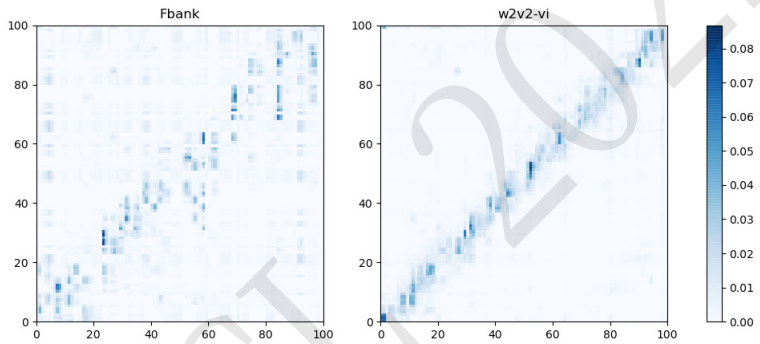


图 4. 两种语音特征编码端的注意力可视化图

Fbank特征在语音翻译任务和语音识别的表现更优，这是因为Fbank特征的计算过程是使用人工设计的结构对语音信号进行降噪和时频分析的过程，能达到较好的效果。而w2v2-vi特征取得的效果相对较差，这是因为该特征编码器预训练阶段在大规模的无标签语音数据上进行自监督的学习，所提取的特征对于语音数据有较好的泛化性，但缺少对目标任务和数据集的适应性。两种特征分别输入到语音翻译模型编码器，其注意力向量序列压缩后的可视化图如图 4所示，由图可知，Fbank特征的注意力更加分散，w2v2-vi特征在编码端学习到的注意力更加集中，这表明不同的特征提取方式所提取的特征存在差异性，促使编码器在训练过程中关注不同的信息。相比于使用人工结构提取的Fbank特征，w2v2-vi在自监督-预训练阶段，从原始波形中动态学习特征提取。从最终输入编码器的序列长度看，经下采样后的Fbank特征序列长度是Wav2vec2特征序列的两倍，而过长的序列长度容易分散注意力。观察w2v2-vi的特征也可发现其注意力更加聚集，说明从原始波形中提取的Wav2vec2特征更加有助于对语音信号进行建模，可以潜在地帮助神经网络发现更好的、更易适应于语音翻译模型的语音表示。

4.2.2 不同融合方法及不同自监督特征的对比实验

为验证融合的有效性，设置单一特征和融合特征两组实验，同时对常用的融合方法和不同的Wav2vec2特征进行比较，选择最优的特征融合方式。使用w2v2-vi和XLSR-53 [7] 两

[7] https://github.com/pytorch/fairseq/blob/main/examples/wav2vec

(c) 2022 中国中文信息学会计算语言学专业委员会

种Wav2vec2模型提取的特征作为融合特征，分别采用拼接融合-L、拼接融合-F和注意力融合三种融合方式进行实验，其中拼接融合-L将频谱表征h_1和自监督表征h_2在长度维度进行拼接，如式 (13)所示，两种特征的特征维度相同，即$d_w = d_z$，通过在长度维度进行拼接最终编码得到的向量$h_x^L \in \mathbb{R}^{d_w \times (M+N)}$ 作为编码器输出。拼接融合-F将h_1和h_2在特征维度进行拼接，如式 (14)，对Fbank特征和Wav2vec2特征分别应用2层和1层卷积下采样，使得二者长度维度相近在相同数量级，长度差异主要在于卷积时的填充。在特征维度进行拼接时，选择长度最大的特征长度作为表征最终输出的长度。经拼接后输出向量在特征维度是原始向量的两倍，再通过线形层变换为原始表征维度。在该融合方式下最终编码得到的向量$h_x^F \in \mathbb{R}^{d_w \times \max(M,N)}$。

$$h_x^L = \text{ConcatL}(h_1, h_2), h_x^L \in \mathbb{R}^{d_w \times (M+N)} \tag{13}$$

$$h_x^F = \text{Linear}(\text{ConcatF}(\text{h}_1, \text{h}_2)), h_x^F \in \mathbb{R}^{d_w \times \max(M,N)} \tag{14}$$

有无融合	方法	w2v2-vi特征	XLSR-53特征
w/o融合	单一特征	36.89	36.37
w/融合	拼接融合-L	38.78	38.44
	拼接融合-F	37.62	38.17
	注意力融合	38.86	38.63

表 4. 融合方法比较比较实验，w/o融合表示无融合，w/融合表示有融合

由表 3和 4可知，w/融合方法比w/o融合方法提升0.73+BLEU值，表明经两类特征融合得到的编码表征，相比于单一的Fbank、w2v2-vi和XLSR-53特征编码的表征能提升语音翻译的性能，通过卷积网络提取的自监督表征和基于频谱的Fbank特征相融合后，有利于编码语音中的局部和全局信息，来更好的表征语音中的声学信息和语音信息，从而提升翻译性能。从表 4的融合方法看，基于注意力融合方法的BLEU值高于拼接融合-L方法和拼接融合-F方法，通过交叉注意力机制学习两种特征之间的对齐关系，编码输出的序列在序列长度和隐层维度保持不变，解码时不增加额外的计算开销，实验结果表明注意力融合方式是最佳的融合方法。从特征类型看，w2v2-vi特征在注意力融合方式下和拼接融合-L方式下略优于XLSR-53的特征，分析可能原因是w2v2-vi的预训练语料为100小时的越南语有声读物，对越南语音频特征提取有优势，而XLSR-53的训练数据为53k小时的多语言音频，其中越南语占比较少，故对越南语音频的特征提取可能产生干扰。而拼接融合-F方式较其他两种方式性能差距较大，其原因可能是在特征维度进行拼接后使用线性层对特征维度进行降维，而使用单层线性层对两种拼接特征进行降维并非最优的降维方式。

4.2.3 不同模型的对比实验

为验证所提方法的有效性，分别使用Fairseq S2T模型、编码器经ASR预训练的ST模型以及MT和ST多任务联合训练的模型作为基线模型在数据集上进行实验，下文简称Fairseq ST基线，ST+ASR PT基线和MTL ST基线。为公平比较，所有模型均不采用额外数据进行预训练或训练。其中，MTL ST基线中MT和ST的损失均为NLL损失，比重分配为4:6，模型基于Tranformer架构。由表 5中Fairseq S2T基线的实验知，在资源充足的情况下Fbank特征的翻译效果优于Wav2vec2特征，故其余基线模型均使用Fbank特征作为输入特征。

实验结果如表 5所示，相比于Fairseq ST基线模型，ST+ASR PT 基线采用经过ASR预训练后的编码器参数来初始化ST模型的编码器，充分利用单语数据来学习语音中的声学信息，提升了1.18个BLEU值。MTL ST基线通过共享解码器参数来进行文本翻译任务和语音翻译任务的联合训练，由于训练过程不采用额外数据,且联合训练过程的损失分配导致单个任务的性能非最优的结果，故而相比与Fairseq ST基线下降3.64个BLEU值。所提模型相比于最优的ST+ASR PT基线提升了0.79个BLEU值，相比于使用Fbank特征的Fairseq ST基线提升了1.97个BLEU值。使用Fbank-Pitch交替特征编码层和表征融合层来融合w2v2-vi特征和Pitch特征，不同特征间的差异性使其相互补充得到更丰富的编码表征，因此翻译质量得到进一步的提升。

(c) 2022 中国中文信息学会计算语言学专业委员会

方法	特征	BLEU	参数量/M
Fairseq ST	w2v2-vi	36.89	47.0
Fairseq ST	Fbank	37.59	47.5
MTL ST	Fbank	33.95	84.0
ST+ASR PT	Fbank	38.77	47.5
所提方法	Fbank+w2v2-vi+Pitch	39.56	45.4

表 5. 四种不同模型的BLEU值

此外，所提模型同时对输入音频的三种特征进行编码，但参数量略小于Fairseq ST基线，这是因为所提方法在交替特征编码器中，使用F特征编码块和FP特征编码块交替编码，其中FP特征编码块参数量小于F特征编码块，且对Wav2vec2编码层和特征融合层层数少，具体设置见4.1节。所提模型相比于ST+ASR PT基线和MTL ST基线,不需要额外对模型的部分模块进行预训练或联合训练步骤，训练效率更高。

4.2.4 消融实验

相比表 5所列的三类端到端基线模型，本文提出交替特征编码器和表征融合层来融合额外的Pitch特征和Wav2vec2特征。本节对所提模型进行了消融实验，评估所提方法中不同模块及额外特征对模型性能的贡献。

模型	BLEU
所提方法	39.56
- 交替特征编码块	38.97
- Pitch特征	38.86
- 特征融合层	37.95
- Wav2vec2特征	37.52

表 6. 消融实验结果

由表 6可知，所提出的不同特征编码模块及融合不同特征对模型性能均能带来正向增益，全部使用可以达到最优结果。交替融合模块按实际区分语义的成分的比重将Pitch特征和Fbank特征进行有效融合，与直接采用Fbank特征和Pitch特征在特征维度进行拼接的方式相比，可以带来0.59个BLEU的提升。进一步去掉Pitch特征，直接使用Fbank的自注意编码表征，翻译性能继续下降0.11个BLEU值。这意味着，越南语的Pitch特征中可能包含能提高语音翻译效果的语义信息，证明了Pitch特征对于有音调语言语音建模的必要性。表征融合层将自监督表征和频谱表征使用交叉注意力机制进行深度融合，来增强编码端输出的表征，去掉该层直接将两种表征用式 (14)在特征维度拼接，性能会下降1.97个BLEU值。进一步去掉Wav2vec2特征，即只使用交替融合模块混合Fbank特征和Pitch特征，翻译性能会继续下降0.43，这表明Wav2vec2特征可以对Fbank特征进行补充，该补充对提升语音翻译的性能是有益的。

5 结论

针对单一特征对复杂语音表征能力不足的问题，本文根据Fbank特征和Wav2vec2特征之间的差异性以及越南语语音特点，使用多特征融合的语音翻译框架将Fbank特征、Wav2vec2特征及Pitch特征进行有效融合，使不同特征间相互补充，增强编码端输出的表征对声学信息和语义信息的表征能力，从而提高越英语音翻译的性能。未来的工作将探索其他自监督表征与传统特征的深度融合，在真实的噪音数据集及低资源场景下的表现，并验证所提方法在其他音调语言上的有效性。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. 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 82–91, New Orleans, Louisiana, June. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain, December.

Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel Lopez-Francisco, Jonathan D. Amith, and Shinji Watanabe. 2022. Combining spectral and self-supervised features for low resource speech recognition and translation.

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.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California, June. Association for Computational Linguistics.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225, Online, August. Association for Computational Linguistics.

Van Huy Nguyen. 2019. An end-to-end model for vietnamese speech recognition. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. End-end speech-to-text translation with modality agnostic meta-learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7904–7908.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Proc. Interspeech 2017*, pages 2630–2634.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proc. Interspeech 2019*, pages 1128–1132.

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

Abdel-rahman Mohamed. 2014. *Deep Neural Network Acoustic Models for ASR*. Thesis. 1 online resource.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? 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.

(c) 2022 中国中文信息学会计算语言学专业委员会

Cao Hong Nga, Chung-Ting Li, Yung-Hui Li, and Jia-Ching Wang. 2021. A survey of vietnamese automatic speech recognition. In *2021 9th International Conference on Orange Technology (ICOT)*, pages 1–4.

Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Yannick Estève, and Laurent Besacier. 2020. Investigating Self-supervised Pre-training for End-to-end Speech Translation. In *Interspeech 2020*, Shangai (Virtual Conf), China, October.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany, December 5-6.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

F. W. M. Stentiford and M. G. Steer, 1990. *Machine Translation of Speech*, page 183–196. Chapman & Hall, Ltd., GBR.

Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney. 2014. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Proc. Interspeech 2014*, pages 890–894.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Miguel Pino. 2020. fairseq S2T: fast speech-to-text modeling with fairseq. *CoRR*, abs/2010.05171.

A Fbank与Wav2vec2特征比较

为进一步比较Wav2vec2特征和Fbank特征，使用MuST-C的英语-越南语数据集上基于Tranformer进行语音翻译实验，表 7为原始数据集的划分。

数据集	时长/h	句数/k
train	432.9	230.9
dev	2.5	1.3
tst-HE	1.2	0.6

表 7. MuCT-C 英语-越南语数据集

本文对该数据集的训练集进行进一步划分，分别抽取训练集时长的100%、75%、50%划分为新的训练集，分别记作train-1.0、train-0.75、train-0.5如表 8所示。对于英语音频的Wav2vec2使用开源的Wav2Vec 2.0 Base [7]预训练模型进行提取，为加速模型收敛，均采用ASR预训练，由于两种特征序列长度的差异，实验结果采用在相同步数下进行比较。其余实验设置与 4.2.1中相同。

训练集	train-1.0	train-0.75	train-0.5
时长/h	432.9	324.4	216.5
Fbank	22.93	21.09	18.01
Wav2vec2	21.56	20.08	17.19

表 8. 不同资源设置下，两种特征在tst-HE测试集上的BLEU

实验结果如表 8所示，由表可知，在train-1上，Fbank特征的翻译效果要略优于Wav2vec2特征；在train-0.5和train-0.75上，两种特征的翻译性能基本持平，且随着训练集的减少，两种特征的性能差距逐渐减小。

(c) 2022 中国中文信息学会计算语言学专业委员会

融入音素特征的英-泰-老多语言神经机器翻译方法

沈政[1,2]，毛存礼[*1,2]，余正涛[1,2]，高盛祥[1,2]，王琳钦[1,2]，黄于欣[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

1591744723@qq.com, maocunli@163.com, ztyu@hotmail.com

gaoshengxiang.yn@foxmail.com, 2424172505@qq.com, huangyuxin2004@163.com

摘要

多语言神经机器翻译是提升低资源语言翻译质量的有效手段。由于不同语言之间字符差异较大，现有方法难以得到统一的词表征形式。泰语和老挝语属于具有音素相似性的低资源语言，考虑到利用语言相似性能够拉近语义距离，提出一种融入音素特征的多语言词表征学习方法：（1）设计音素特征表示模块和泰老文本表示模块，基于交叉注意力机制得到融合音素特征后的泰老文本表示，拉近泰老之间的语义距离；（2）在微调阶段，基于参数分化得到不同语言对特定的训练参数，缓解联合训练造成模型过度泛化的问题。实验结果表明在ALT数据集上，提出方法在泰-英和老-英两个翻译方向上，相比基线模型提升0.97和0.99个BLEU值。

关键词： 多语言神经机器翻译；泰语；老挝语；低资源语言；音素；参数分化

English-Thai-Lao multilingual neural machine translation fused with phonemic features

Zheng Shen[1,2], Cunli Mao[*1,2], Zhengtao Yu[1,2], Shengxiang Gao[1,2], Linqin Wang[1,2], Yuxin Huang[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

1591744723@qq.com, maocunli@163.com, ztyu@hotmail.com

gaoshengxiang.yn@foxmail.com, 2424172505@qq.com, huangyuxin2004@163.com

Abstract

Multilingual neural machine translation is an effective methods to improve the performance of low-resource language translation. However, characters in different languages are significantly different, and existing methods are difficult to obtain a unified word representation form. Thai and Lao are low-resource languages with phonemic similarity. Considering that the use of language similarity can shorten the semantic distance, in this article, a multilingual word representation learning method incorporating phonemic features is proposed: (1) Design the phoneme feature representation module and the Thai-Lao text representation module, and then obtain the Thai-Lao text representation after fused phoneme features based on the cross-attention mechanism to shorten the semantic distance between the Thai-Lao and Laotian; (2) In the fine-tuning stage, specific training parameters for different language pairs are obtained based on parameter differentiation, which alleviates the problem of over-generalization of the model

*毛存礼(通信作者):maocunli@163.com

国家自然科学基金重点项目（61732005,U21B2027）；国家自然科学基金（62166023，61866019）；云南省自然科学基金重点项目(2019FA023);云南省重大科技专项计划项目(202103AA080015，202002AD080001)

(c) 2022 中国中文信息学会计算语言学专业委员会

caused by joint training. Experimental results on ALT datasets show that the proposed method improves the BLEU values by 0.97 and 0.99 in Thai-English and Lao-English translation tasks, respectively, compared with the benchmark model.

Keywords: multilingual neural machine translation , Thai , Lao , low-resource languages , phoneme , parameter differentiation

1 引言

近年来，神经机器翻译(NMT)(Sutskever et al., 2014; Bahdanau et al., 2014; Wang et al., 2022)因其优越的性能引起了广泛的关注，成为机器翻译领域的主流方法。随之而来，基于统一的翻译模型实现多种语言对联合训练的框架成为了研究热点(Xu et al., 2021)，目前，MNMT(Wang et al., 2019; Bapna et al., 2022)在低资源语言(Man et al., 2020)翻译上取得了较好的效果，相比单独训练双语翻译模型，MNMT能够通过共享跨语言知识来提升资源稀缺语言的机器翻译性能。然而，在如何利用语言之间特有的知识上仍有较大的研究空间。

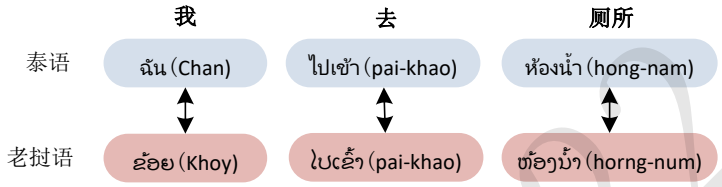


Figure 1: 泰语-老挝语音素相似性示例

现有方法在进行多语言词表征时，由于不同语言之间字符差异性较大难以得到统一的词表征形式，例如，泰语和老挝语属于孤立型语言，不具备天然分词，在机器翻译模型训练的过程中，泰语、老挝语和英语之间的语言差异性极大，仅仅通过联合训练或参数共享的方式无法得到准确的语义表征。泰语和老挝语都属于汉藏语系壮侗语族的壮傣语支，在构词特点、词语音素以及句法结构上都有相同或相似的地方，特别是在音素层面上，大部分具有相同含义的泰语、老挝语音素相同(Ding et al., 2016; Yu et al., 2020)。如图1所示，泰语和老挝语句法结构基本一致，都属于主语-谓语-宾语(Subject-Verb-Object, SVO)的结构，音素也有较高的相似度，如汉语“去”对应的老挝语音素 ***pai-khao*** 和泰语音素 ***pai-khao*** 相同，并且，汉语“我”和“厕所”对应的泰语、老挝语音素也具备一定的相似性，这说明泰语、老挝语两种语言在音素层面上存在大量的一致性。Tan et al. (2019)已经证明，相似性高的语言进行多语言联合训练时，该特性有助于提高翻译模型性能，这是因为模型在训练过程中能自动学习到语言在句法、词法等层面上的相似特征。

因此，鉴于泰语、老挝语之间的语言相似性，利用多语言机器翻译模型能够有效地学习到泰语和老挝语的相似特征，提升模型翻译性能。然而我们观察到泰语和老挝语的相似性并不体现在字符层面上，单纯利用联合训练的方式无法学习到音素层面的相似性特征。针对以上问题，本文提出了融入音素特征的多语言词表征学习方法，在Transformer的框架下，将音素和文本分别进行词向量表示，并基于交叉注意力机制将二者融合，最后，基于参数分化策略对模型进行微调。

本文的贡献主要有以下三点：

（1）为了进一步拉近泰语、老挝语之间的语义表征距离，提出联合泰语、老挝语音素特征和文本表示方法，基于交叉注意力机制进一步学习融合音素特征后的文本表示。

（2）在我们的工作中，由于泰-老之间具有语言相似性，为了更有效的对泰-老语言特定参数进行分化，我们在参数分化方法的基础上对模型进行微调，以此缓解联合训练造成的模型过度泛化问题。

（3）在公开数据集ALT上，实验结果表明所提方法优于多个基线模型，在泰-英翻译方向上BLEU值达到17.99，在老-英翻译方向上的BLEU值达到15.40，表明翻译性能提升得益于提出的融入音素特征的多语言词表征学习方法。

第二十一届中国计算语言学大会论文集，第305页-第316页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

2 相关工作

泰语和老挝语是典型的低资源语言，其双语平行语料稀缺，所以相关机器翻译研究较少。早期主要利用基于统计规则的机器翻译方法实现对泰语、老挝语的翻译(Phitakwinai et al., 2008; Asawavichienjinda et al., 2005)，但是统计机器翻译需要大量双语语料，用于泰语和老挝语效果不佳。随着机器翻译技术的发展，神经机器翻译方法逐渐被应用到泰语和老挝语的翻译任务上(Saengthongpattana et al., 2019; Poncelas et al., 2020)，但是当前研究多集中在现有方法的简单应用，没有有效利用泰语和老挝语的语言特点，因此翻译效果不佳。

MNMT可以通过不同语种之间的知识迁移提升资源稀缺语言的翻译效果，目前已经成为解决低资源机器翻译的主流方法之一。近年来，研究人员对MNMT模型结构进行了很多探索。主要有三类：(1)对所有源语言使用相同的编码器，对目标语言使用不同的解码器。Dong et al. (2015)在一对多的翻译场景下，提出所有源语言共享编码器，为每个目标语言分配不同解码器的方法。(2)对所有源语言和目标语言都使用不同的编码器和解码器。Zoph and Knight (2016)提出多种语言联合训练注意力机制的多对一多语言机器翻译方法。上述方法由于都要为每种语言单独训练编码器或解码器，极大地限制了模型在语言数量上的可扩展性。(3)对所有的源语言和目标语言均使用相同的编码器的解码器。Johnson et al. (2017)提出了一种多语言联合训练的单编码器-单解码器模型，并在源语言首字符前增加目标语言标志位以指导目标语言的生成。该方法虽然实现了利用单一模型翻译多个语种，但是忽略了语种之间的差异性。最新的多语言机器翻译方法多倾向于在单一模型上设计语言特定的子模块。Wang and Zhang (2021)训练过程中利用梯度差异逐步分离语言特定参数。Xie et al. (2021)根据各个神经元在该语言对上的重要性划分子网络。Khusainova et al. (2021)通过显示不同语言之间相关程度的语言树控制共享参数的数量。Zhu et al. (2021)引入一个轻量级的适配器，通过该适配器学习各个语种的特有信息。Zhang et al. (2020a)利用门控机制训练多语言机器翻译模型，实现模型动态选择是否共享参数。Zhang et al. (2020b)通过语言感知的归一化层将不同语言映射到不同的高斯空间中，并利用语言感知的线性层对不同语言之间的关系进行建模。

上述方法为本文提供了较好的思路，本文在Johnson et al. (2017)提出方法的基础上进行改进，针对现有方法难以得到统一的泰老词表征形式问题，提出融入音素特征的英-泰-老多语言神经机器翻译方法，利用泰语和老挝语的音素相似性，拉近其语义距离。Wang and Zhang (2021)首次提出利用参数分化的思想训练多语言神经机器翻译模型，取得了不错的效果，而本文主要研究泰语-老挝语这种具有语言相似性的语种，这样的语种更适合在模型充分学到到语言相似特征后再进行参数分化，因此，我们在微调阶段再利用参数分化学习语言特定参数。

3 研究背景

在本节中，我们主要介绍基于注意力机制的Transformer框架(Vaswani et al., 2017)和泰语、老挝语的语言相似性。

3.1 基于Transformer的神经机器翻译模型

Transformer是基于序列到序列框架(Sutskever et al., 2014)实现的，由多层编码器和多层解码器堆叠而成，每层编码器包含一个多头自注意力层和一个前馈神经网络层，每层解码器除了上述两个模块，在多头自注意力层后是一个多头交叉注意力层，每个模块由残差连接和归一化进行关联。对于给定源语言句子$x = (x_1, x_2, ..., x_n)$，首先通过编码器将其编码为一个稠密的隐向量表示，然后利用解码器将其解码成目标语言句子$y = (y_1, y_2, \ldots, y_z)$。

多头注意力机制是Transformer中的一个重要模块,可以表示为：

$$MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_M)W \tag{1}$$

其中，$Q(query)$，$K(key)$，$V(value)$是输入句子的隐向量表示，W为参数矩阵，M为多头注意力机制头数，每个头计算如下：

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$= softmax\left(\frac{(QW_i^Q)(KW_i^K)^T}{\sqrt{d_k}}\right)(VW_i^V) \tag{3}$$

第二十一届中国计算语言学大会论文集，第305页-第316页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

其中，W_i^Q，W_i^K，W_i^V是参数矩阵，d_k是隐藏层的维度。

多头自注意力机制能建立句子中词与词之间的相互连接，得到融合上下文信息的隐向量表示，多头交叉注意力机制主要用于连接源语言与目标语言向量表示。由于Transformer网络不使用递归方式编码，因此在模型中使用位置嵌入来利用序列位置信息。

3.2 多语言神经机器翻译模型

与NMT模型相比，MNMT对多个语言对进行联合训练，实现了多任务参数共享。我们选择Johnson et al. (2017)提出的方法作为实验的基准模型，该方法训练了一个单编码器-单解码器的MNMT模型，并通过在源语言句子首位增加目标语言标志位指导目标语言的生成，其目标函数为所有语言对下每个词生成概率的乘积:

$$\mathcal{L}(D;\theta) = \sum_{l=1}^{L}\sum_{d=1}^{|D_l|}\sum_{t=1}^{N} \log P(y_t^l|x^l; y_{<t}^l; \theta_{enc}; \theta_{dec}; \theta_{attn}) \tag{4}$$

其中，D是训练语料中所有平行句对的集合，θ是模型中所有参数的集合，L表示模型联合训练的语言对总数，N表示目标语言句子长度，$|D_l|$表示训练语料中属于第l个语言对的平行句对数量，$P(y_t^l|x^l; y_{<t}^l)$表示第l语言对中第d个句子的第t个单词的翻译概率，θ_{enc}表示模型中编码器的参数，θ_{dec}表示模型中解码器的参数，θ_{attn}表示模型中注意力机制的参数。

3.3 泰语、老挝语语言相似性

本小节我们主要回答以下两个问题:

（1）泰语、老挝语语言相似性如何体现?

（2）语言相似性如何影响机器翻译性能?

3.3.1 泰语、老挝语语言相似性如何体现

泰语和老挝语都属于汉藏语系壮侗语族的壮傣语支，两国文字的发音方面较为相似(Ding et al., 2016)，泰老两种语言的音素都是由元音、辅音、尾辅音、声调符号等构成的，并且大多数情况下一一对应。为了更准确地利用泰老的音素相似特征，我们利用编辑距离对语料中泰老平行句对的音素相似性进行统计分析。

音素相似度	<0.6	0.6-0.7	0.7-0.8	>0.8
语料占比(%)	1.58	61.93	26.08	10.41

Table 1: 泰-英、老-英数据集统计信息

由表1可得，语料中音素相似度在0.6以上的平行句对占比达到98.42%，这说明泰老在音素层面有较高的相似性。

3.3.2 语言相似性如何影响机器翻译性能

受人文地理等的影响，现有的很多语言都是由同一古语言发展而来，从而在发音、书写、句法等方面有较高的相似性。NMT属于跨语言任务，其关键在于如何根据源语言的语义表征解码出对应目标语言，当两种语言属于相似语言时，其语义空间更加接近，翻译效果更佳，例如，两种语言可通过共享词表共享其同源词向量表征，WMT相似语言翻译任务对此已进行了大量的研究(Ojha et al., 2019; Baquero-Arnal et al., 2019)，证明了利用语言相似性可以有效提升翻译模型性能。由3.3.1的分析可以得出，泰语和老挝语具有较高的音素相似性，但传统NMT模型无法充分利用该特性。因此本文考虑基于交叉注意力机制通过泰老音素相似性拉近其语义距离，提升翻译模型性能。

4 融合音素特征的多语言神经机器翻译模型

本文的目标是构建一个源端为泰语和老挝语，目标端为英语的MNMT模型。总体框架如图2所示，主要包括音素特征表示模块、泰老文本表示模块、基于交叉注意力机制的音素-文本表示模块、目标语言解码器。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.1 泰老文本表示模块

给定一个泰语或老挝语句子为$x = (x_1, x_2, ..., x_n)$，其中n为文本x的序列长度，文本序列通过带有位置嵌入的传统嵌入层得到其词向量表征E_t，计算如下：

$$E_t = Emb_t(x) + PE_t(x) \tag{5}$$

其中，Emb_t为文本序列词嵌入层，PE_t为文本位置嵌入层，$E_t \in R^{n \times d_k}$。

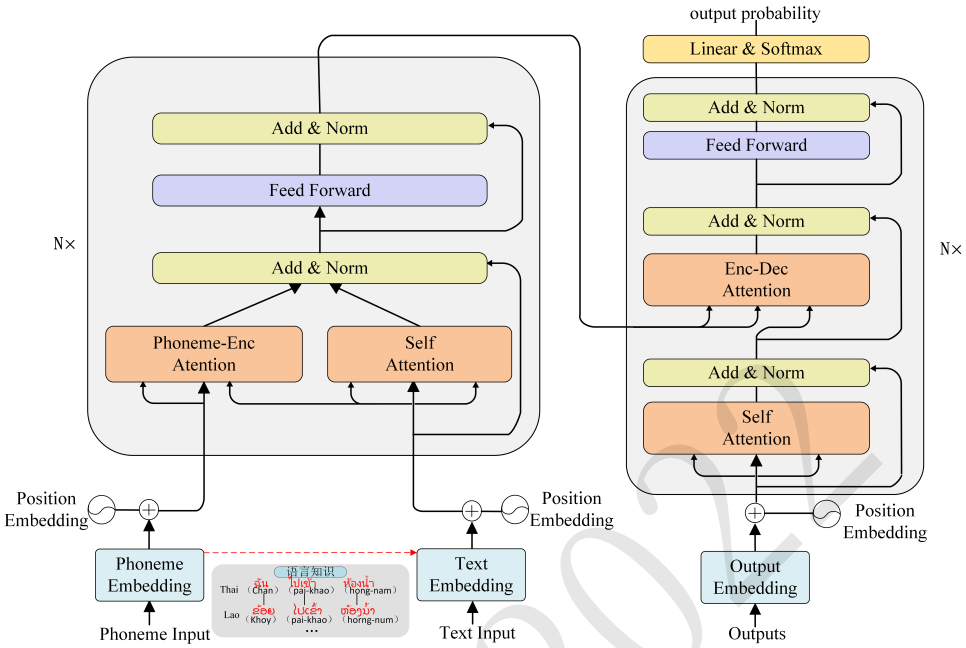


Figure 2: 融合音素特征的多语言神经机器翻译模型框架

4.2 音素特征表示模块

对于文本序列x，通过G2P（字符转音素）[0]工具将其转化成对应的音素序列$x_p = (x_{p1}, x_{p2}, ..., x_{pm})$，其中$m$为音素$x_p$的序列长度，音素序列通过带有位置嵌入的传统嵌入层得到其词向量表征E_p，计算如下：

$$E_p = Emb_p(x_p) + PE_p(x_p) \tag{6}$$

其中，Emb_p为音素序列词嵌入层，PE_p为音素位置嵌入层，$E_p \in R^{m \times d_k}$。

4.3 基于交叉注意力机制的音素-文本表示模块

为了拉近老挝语和泰语的语义距离，本文通过交叉注意力机制将音素特征融入泰老文本表示，如图3所示。首先，文本词向量表征E_t经过自注意力层计算源语言序列上下文向量H_t：

$$H_t = MultiHead(E_t, E_t, E_t) \tag{7}$$

同时，文本词向量表征E_t为查询向量，音素词向量表征E_p为键向量和值向量，经过音素-文本交叉注意力机制得到融入音素特征的文本表示H_p：

$$H_p = MultiHead(E_t, E_p, E_p) \tag{8}$$

然后，采用加权的方式将H_t和H_p进行融合：

$$H = \alpha * H_t + (1 - \alpha) * H_p \tag{9}$$

[0]https://github.com/dmort27/epitran

(c) 2022 中国中文信息学会计算语言学专业委员会

其中α是超参数。

最后，使用位置前馈网络（FFN）更新序列每个位置的状态，得到H_{enc}：

$$H_{enc} = FFN(H) \tag{10}$$

经过多层编码后，将编码器的输出H_{enc}输入解码器进行解码。

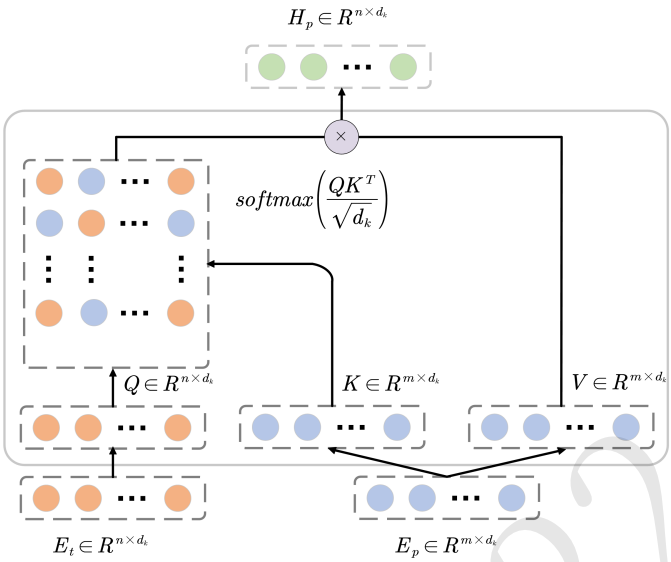


Figure 3: 基于交叉注意力机制的音素-文本表示模块

4.4 目标语言解码器

与泰老文本表示模块类似，首先将泰语或老挝语句子x对应的英语句子$y = (y_1, y_2, \ldots, y_z)$进行词向量表征得到$E_y$，其中$z$为目标语言序列长度。如图2右侧所示，本文解码器采用传统的Transformer框架，每层解码器由多头自注意力层、多头交叉注意力层、前馈神经网络层三个子层组成。

首先利用多头自注意力机制提取目标句子特征：

$$H_y = MultiHead(E_y, E_y, E_y) \tag{11}$$

然后使用多头交叉注意力机制实现融合音素特征的源语言上下文向量H_{enc}和目标句子特征H_y的交互：

$$H_c = MultiHead(E_y, E_{enc}, E_{enc}) \tag{12}$$

然后，使用FFN更新序列每个位置的状态，得到H_{dec}：

$$H_{dec} = FFN(H_c) \tag{13}$$

最后将解码器最后一层的输出作为softmax层的输入，并预测目标句子的概率分布：

$$P = Softmax(W_p H_{dec} + b) \tag{14}$$

其中W_p和b是模型参数。

4.5 微调

在微调阶段，考虑到不同语言之间的参数干扰问题，我们基于参数分化思想(Wang and Zhang, 2021)，与之不同的是我们没有在训练阶段使用该思想，主要是因为训练阶段分离参数会导致模型无法充分学习到语言相似特征。因此，我们基于该思想对模型进行微调，即针对训练好的模型，分别利用泰语-英语和老挝语-英语的验证集获取两个语言对在各个参数上的梯度，并依此计算各个参数上两个语言对梯度的余弦相似度：

$$sim(\theta_i) = \frac{g_i^{t_1} \cdot g_i^{t_2}}{||g_i^{t_1}|| \cdot ||g_i^{t_2}||} \tag{15}$$

第二十一届中国计算语言学大会论文集，第305页-第316页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

其中，θ_i是模型第i个参数，t_1指老挝语到英语的翻译任务，t_2指泰语到英语的翻译任务，$g_i^{t_1}$是任务t_1在$_i$上的梯度。

模型每微调一定步数计算一次梯度，并对t_1和t_2梯度相似度较低的参数进行分离，即t_1和t_2的该参数不再共享，两个任务分别针对该参数微调，直到模型再次收敛。

5 实验

为了在实验中公平的比较模型性能，并且同时验证所提泰-老音素相似性特征的有效性，在本文中我们的主要研究对象是泰语和老挝语这两种音素相似的语言，具体实验如下。

5.1 实验数据

本文的泰-英、老-英的语料直接来源于公共数据集亚洲语言树库（ALT）[1]，泰-英和老-英分别有20106条平行语料。由于该数据集没有划分训练集、验证集和测试集，本文选取泰-英和老-英数据各1000条作为验证集，取1106条作为测试集，剩余18000条作为训练集，如表2所示：

数据集	训练集（句对）	验证集（句对）	测试集（句对）
泰-英	18000	1000	1106
老-英	18000	1000	1106

Table 2: 泰-英、老-英数据集统计信息

5.2 实验环境及配置

本文实验的神经网络模型是基于Torch1.8实现的，编译语言为Python 3.8，在单个NVIDIA Tesla T4 GPU上进行实验。在实验中，我们使用BPE对所有源语言和目标语言进行联合子词切分，词表大小为4k。本文选择Transformer模型作为基础模型，模型的编解码器分别设置为3层。在编码器和解码器中的词向量和隐藏层的维度设置为128维。优化器选择参数设置为β_1 = 0.9，β_2 = 0.98的Adam优化器优化模型。我们参照Vaswani et al. (2017)使用warm_steps = 4000的warm-up策略来调整学习率，每个批次包含大约4096个词。我们训练模型直到连续10次验证集的BLEU值没有提升，则认为模型收敛并停止训练，该方法可以有效防止模型过拟合。在解码过程中，beam search设置为5,并采用BLEU(Papineni et al., 2002)指标来评估模型性能。

5.3 实验结果及分析

本文初步探索了泰老音素级别的相似性，并在编码端利用该特征参与模型训练，解码端统一设置为英语，这是因为在自回归的框架下逐词解码并转音素再参与模型解码会导致翻译错误的累积。因此，我们主要考虑在泰-英和老-英方向上，探究音素相似性对实验结果的影响。

5.3.1 一对一及多对一的翻译场景下不同模型实验结果对比分析

在实验中，我们基于fairseq[2]框架与其他模型对比，并按照原论文参数复现，在达到最好结果时进行比较分析。设置对比试验如下：

(1)Bilingual：Vaswani et al. (2017)为每个语言对分别训练一个Transformer神经机器翻译模型，其参数设置与本文提出的方法一致。

(2)Multi-Source：Zoph and Knight (2016)在多对一的翻译场景下为每个源语言分配不同的编码器，目标语言共享解码器。

(3)Adapter：Bapna et al. (2019)在Transformer每一层的顶端为每一个语言对增加一个额外的适配器，每个适配器分别学习各个语言地特定知识。

(4)PD：Wang and Zhang (2021)提出在训练阶段利用不同语言对参数梯度的相关性分离语言特定参数。

(5)LaSS：Lin et al. (2021)通过判断神经元重要性为各语言对裁剪冗余神经元，以此使对所有语言对都重要的神经元学习通用知识，对单个语言对重要的神经元学习语言特定的知识。

[1]http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html
[2]https://github.com/facebookresearch/fairseq

(c) 2022 中国中文信息学会计算语言学专业委员会

(6)基线模型（Baseline）：基线模型是指基于Transformer框架，不使用音素特征和参数分化策略下进行的翻译实验。

翻译场景	方法	老-英	△	泰-英	△
一对一	Transformer	9.72	-	14.70	-
多对一	Multi-Source	12.75	+3.03	16.13	+1.43
	Adapter	14.53	+4.81	16.79	+2.09
	PD	14.04	+4.32	16.36	+1.66
	LaSS	12.54	+2.82	15.24	+0.54
	Baseline	14.43	+4.71	17.00	+2.30
	本文方法（Transformer-Base）	7.12	-2.60	8.44	-6.26
	本文方法	**15.40**	**+5.68**	**17.99**	**+3.29**

Table 3: 一对一及多对一翻译场景下的实验结果

如表3所示，在一对一的翻译场景下，基于Transformer框架在老-英和泰-英翻译方向上BLEU值分别达到了9.72和14.70。在多对一的翻译场景下，所有模型相比一对一场景下的BLEU值均有明显提升，其中，本文提出的方法在老-英和泰-英翻译方向上BLEU值分别达到了15.40和17.99，取得了最高水平，在老-英和泰-英翻译方向上BLEU值分别提升了5.68和3.29，这说明利用MNMT方法将老挝语-英语和泰语-英语联合训练，可以通过知识迁移有效缓解老挝语和泰语数据稀缺导致的模型翻译性能不佳的问题。

此外，本文方法相比Multi-Source在老-英和泰-英翻译方向上BLEU值分别提升了2.65和1.86，这说明共享编码器可以有效利用泰老语言相似性提升模型翻译效果。相比Adapter，本文方法在老-英和泰-英翻译方向上BLEU值分别提升了0.87和1.20，这说明低资源情况下单独训练额外参数效果不佳。相比PD，本文方法在老-英和泰-英翻译方向上BLEU值分别提升了1.36和1.63，这说明该方法会过早分离模型参数从而导致模型知识迁移不充分。相比Lass，本文方法在老-英和泰-英翻译方向上BLEU值分别提升了2.86和2.75，这说明该方法依赖大规模的模型参数和训练数据，在低资源情况下会出现过度裁剪而丢失部分共有参数的问题。相比Baseline，本文方法在老-英和泰-英翻译方向上BLEU值分别提升了0.97和0.99，说明本文方法可以有效拉近泰老之间的语义距离并缓解联合训练造成的模型过度泛化的问题，提升翻译模型性能。为了进一步证明本文参数设置的合理性，本文设置了在Transformer-Base参数下的对比实验，该参数下的的BLEU值远小于本文参数下的BLEU值，这说明参数过大会导致模型过拟合，从而使测试集上的翻译效果较差。

5.3.2 消融实验

为了探究融入音素特征和基于参数分化的微调策略的有效性，本文设置了消融实验，如表4所示。

方法	老-英	△	泰-英	△
Baseline	14.43	-	17.00	-
Baseline+音素	15.13	+0.70	17.74	+0.74
Baseline+音素（拼接）	6.50	-7.93	9.77	-7.23
Baseline+参数分化	14.64	+0.21	17.23	+0.23
Baseline+音素+参数分化	**15.40**	**+0.97**	**17.99**	**+0.99**

Table 4: 消融实验

实验结果表明，融入音素特征使模型在老-英和泰-英翻译方向上BLEU值分别提升了0.70和0.74，说明该方法可以有效拉近泰老之间的语义距离，缓解泰老字符差异较大导致的词表征形式不统一的问题。基于参数分化思想的微调策略使模型在老-英和泰-英翻译方向上BLEU值分别提升了0.21和0.23，说明该方法可以学习到语言特定知识，缓解联合训练造成模

(c) 2022 中国中文信息学会计算语言学专业委员会

型过度泛化的问题。基线模型+音素的方式相比基线模型+参数分化的方式，在老-英和泰-英翻译方向上的BLEU值提升更为明显，说明本文提出的方法对翻译性能带来的提升更依赖于泰语和老挝语之间的音素相似性。两种方法可同时使用，此时模型效果达到最佳，在老-英和泰-英翻译方向上BLEU值分别提升了0.97和0.99。为了进一步证明本文方法的有效性，本文利用Koehn (2004)提出的重采样方法进行了显著性检验($p < 0.05$)。

为了探讨音素融合方式对实验结果的影响，我们设计了利用拼接方式融合音素特征的实验，即模型输入为拼接了音素的文本，实验结果表明，拼接方式会使模型BLEU值远低于基线模型，这说明直接拼接会使输入序列过长，使得模型学习困难较大，从而导致模型性能下降。

5.3.3 音素特征融合层数对翻译效果的影响

为了探究编码器不同层数融合音素特征对实验结果的影响，我们设计对比试验，对比三层，结果如表5所示。

融合层数	老-英	泰-英
第一层	14.05	16.89
第二层	14.74	16.95
第三层	14.64	16.78
第一层+第二层	15.00	17.38
第一层+第三层	14.90	16.95
第二层+第三层	14.52	16.93
第一层+第二层+第三层	**15.13**	**17.74**

Table 5: 音素特征融合层数对翻译效果的影响

表5实验结果表明，音素特征融合层数对模型翻译效果有着重要的影响，当只进行单层融合时，第二层效果最佳，当进行两层融合时，第一层+第二层效果最佳，当三层都融入音素特征时模型效果达到最优。当只进行单层融合时，模型无法从音素特征中学习到有效的信息，甚至会引入噪声信息，从而影响模型性能。当模型进行两层融合时，音素特征低层融合比高层融合效果更佳，这表明，模型低层更容易学习到泰语和老挝语的音素相似特征。当模型进行三层融合时，模型能充分利用泰老音素相似性拉近语义距离，提升翻译效果。

5.3.4 音素特征和文本特征融合比例对翻译效果的影响

本文利用超参数α对模型学到的文本特征和音素特征比例进行平衡，并针对该超参数的设置进行了探究，讨论α取值对实验结果的影响。

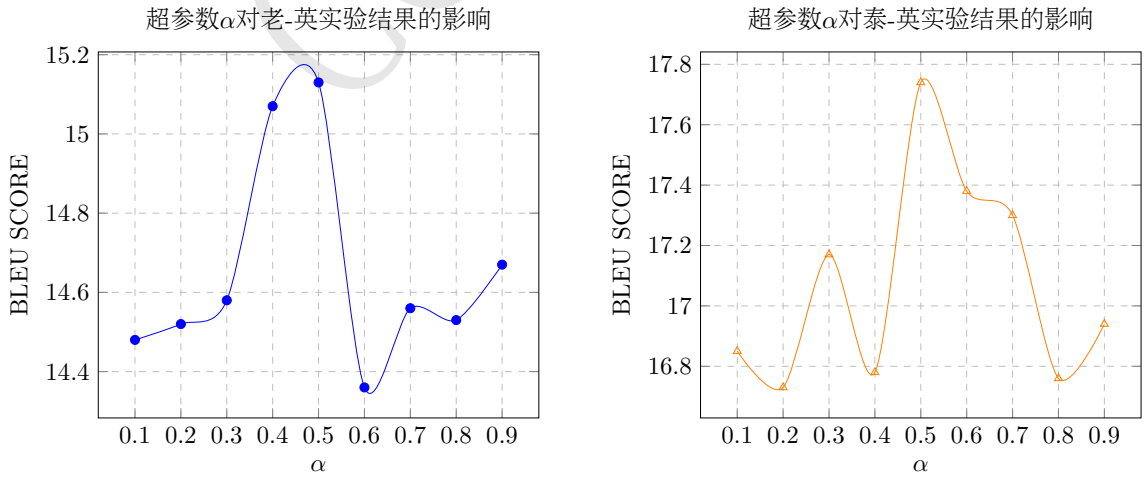


Figure 4: 超参数α对老-英、泰-英实验结果的影响

如图4左图所示，在老-英翻译方向上，当文本占比小于音素占比时，α从0.1变化到0.4，实验性能逐步上升，在0.5时逐渐达到最优，此时文本特征和音素特征比例达到平衡，当α达

(c) 2022 中国中文信息学会计算语言学专业委员会

到0.6时，意味着文本占比大于音素占比，实验性能降到最低，这说明音素特征在占比0.4时对文本特征干扰最大，当α达到0.9时，此时，文本占比达到最高，性能相较α在0.6到0.8有所提升，说明此时的文本特征相较音素特征对翻译性能影响更大。

如图4右图所示，在泰-英翻译方向上，当音素占比小于0.5时，此时的文本占比大于0.5，音素特征占比较小在一定程度上为模型引入了噪声，因此模型翻译效果不佳。当α达到0.9时，音素特征占比极小，对文本特征的干扰也达到最小，性能相较α在0.6到0.8有所提升。

我们得到结论，在泰-英和老-英翻译方向上，模型对于α的取值较为敏感，当音素特征和文本特征占比趋近0.5时文本特征和音素特征比例达到平衡状态，可使模型更好地学习到相似特征，从而性能逐渐达到最优。

5.3.5 翻译实例分析

本文分别以泰语-英语和老挝语-英语的翻译结果为例，分析融入音素特征并基于参数分化微调对模型生成译文质量的影响。

老-英翻译示例	
源语言	**ລັດຖະບານທະຫານ**ບໍ່ຍອມຮັບ**ສຽງສ່ວນໃຫຍ່ຂອງNLD** ໃນ**ການເລືອກຕັ້ງ**ຄັ້ງນັ້ນ.
Baseline	The **NLD government** has not accepted **the majority of the polls**.
本文方法	**military government** does not accept **the majority of the NLD** in the **election**.
参考译文	The **military junta** did not accept **NLD 's majority** in that **election**.
泰-英翻译示例	
源语言	Hans ผู้เป็น**สามี**เคยเป็น**นักโทษ**สมัยสงครามโลกครั้งที่สอง และไปทำงานเป็นชาวนาให้ครอบครัวของ Josie
Baseline	Hans was the second World War II and to work as a procedure for Josie family.
本文方法	Hans, a **husband**, who had been the second World War **prisoner**, and worked as Josie 's family.
参考译文	**Husband** Hans was a **prisoner** of war in World War II, and went to work as a farmer for Josie 's family.

Table 6: 老-英、泰-英翻译示例

如图6所示，基线模型会出现部分关键词漏译、错译的现象，例如，在老挝语-英语的翻译中，基线模型输出的英语句子中漏译了“election”，错译了“military junta”和“NLD 's majority”；泰语-英语翻译结果类似，基线模型输出的英语句子也漏译了“husband”和“prisoner”，这是由于基线模型泰老词表征没有统一，知识迁移效果不佳。而本文方法有效缓解了基线模型中的错译、漏译问题，这充分证明了在模型中融入额外音素相似特征可以有效拉近泰老之间的语义距离，同时通过基于参数分化的微调策略也可以有效缓解模型过度泛化的问题。

6 结论

针对现有多语言神经机器翻译方法难以得到统一词表征形式的问题，本文利用泰语和老挝语音素相似性拉近泰老之间的语义距离，并在微调阶段基于参数分化缓解联合训练造成模型过度泛化的问题。实验结果证明了本文方法的有效性和优越性，在老挝语-英语和泰语-英语翻译任务的BLEU值分别达到了15.40和17.99，比基线模型均有明显提升。我们的工作不仅利用多语言模型提升了低资源语言泰语、老挝语到英语的翻译性能，我们还探究了泰语、老挝语之间的音素相似性特征，并将其融合到模型中，有效地改善了在多语言翻译过程中低资源语言由于数据稀缺以及语言特性导致的共享语言知识困难的问题。此外，我们的方法进一步为更多低资源的相似性语言机器翻译任务提供了较好的思路。

第二十一届中国计算语言学大会论文集，第305页-第316页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Thanin Asawavichienjinda, Kammant Phanthumchinda, Chitr Sitthi-Amorn, and Edgar J Love. 2005. The thai version of the quality-of-life in epilepsy inventory (qolie-31-thai version): translation, validity and reliability. *JOURNAL-MEDICAL ASSOCIATION OF THAILAND*, 88(12):1782.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*.

Ankur Bapna, Orhan Firat, Pidong Wang, Wolfgang Macherey, Yong Cheng, and Yuan Cao. 2022. Multilingual mix: Example interpolation improves multilingual neural machine translation.

Pau Baquero-Arnal, Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. 2019. The mllp-upv spanish-portuguese and portuguese-spanish machine translation systems for wmt19 similar language translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 179–184.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2016. Similar southeast asian languages: Corpus-based case study on thai-laotian and malay-indonesian. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 149–156.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. 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 1723–1732.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Albina Khusainova, Adil Khan, Adín Ramírez Rivera, and Vitaly Romanov. 2021. Hierarchical transformer for multilingual machine translation. *arXiv preprint arXiv:2103.03589*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. *arXiv preprint arXiv:2105.09259*.

Zhibo Man, Cunli Mao, Zhengtao Yu, Xunyu Li, Shengxiang Gao, and Junguo Zhu. 2020. 基于多语言联合训练的汉-英-缅神经机器翻译方法(chinese-english-burmese neural machine translation method based on multilingual joint training). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 446–456.

Atul Kr Ojha, Ritesh Kumar, Akanksha Bansal, and Priya Rani. 2019. Panlingua-kmi mt system for similar language translation task at wmt 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 213–218.

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.

Suwannee Phitakwinai, Sansanee Auephanwiriyakul, and Nipon Theera-Umpon. 2008. Thai sign language translation using fuzzy c-means and scale invariant feature transform. In *International Conference on Computational Science and Its Applications*, pages 1107–1119. Springer.

Alberto Poncelas, Wichaya Pidchamook, Chao-Hong Liu, James Hadley, and Andy Way. 2020. Multiple segmentations of thai sentences for neural machine translation. *arXiv preprint arXiv:2004.11472*.

Kanchana Saengthongpattana, Kanyanut Kriengket, Peerachet Porkaew, and Thepchai Supnithi. 2019. Thai-english and english-thai translation performance of transformer machine translation. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–5. IEEE.

(c) 2022 中国中文信息学会计算语言学专业委员会

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. *arXiv preprint arXiv:1908.09324*.

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.

Qian Wang and Jiajun Zhang. 2021. Parameter differentiation based multilingual neural machine translation. *arXiv preprint arXiv:2112.13619*.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223.

Shuo Wang, Zhixing Tan, and Yang Liu. 2022. Integrating vectorized lexical constraints for neural machine translation. *arXiv preprint arXiv:2203.12210*.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. *arXiv preprint arXiv:2107.06569*.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, and Deyi Xiong. 2021. Modeling task-aware mimo cardinality for efficient multilingual 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 361–367.

Zhiqiang Yu, Zhengtao Yu, Yuxin Huang, Junjun Guo, Zhenhan Wang, and Zhibo Man. 2020. Transfer learning for chinese-lao neural machine translation with linguistic similarity. In *China Conference on Machine Translation*, pages 1–10. Springer.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2020a. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020b. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. *arXiv preprint arXiv:2104.08154*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.

CCL 2022

机器音译研究综述

李卓[1,2]　王志娟[1,2,*]　赵小兵[1,2]
[1]中央民族大学　信息工程学院，北京　100081
[2]国家语言资源监测与研究少数民族语言中心
*Corresponding author: Zhijuan Wang
li_zhuo98@163.com, wangzj.muc@gmail.com, nmzxb_cn@163.com

摘要

机器音译是基于语音相似性自动将文本从一种语言转换为另一种语言的过程，它是机器翻译的一个子任务，侧重于语音信息的翻译。音译后可知道源单词在另一种语言中的发音，使不熟悉源语言的人更容易理解该语言，有益于消除语言和拼写障碍。机器音译在多语言文本处理、语料库对齐、信息抽取等自然语言应用中发挥着重要作用。本文阐述了目前机器音译任务中存在的挑战，对主要的音译方法进行了剖析、分类和整理，对音译数据集进行了罗列汇总，并列出了常用的音译效果评价指标，最后对该领域目前存在的问题进行了说明并对音译学的未来进行了展望。本文以期对进入该领域的新人提供快速的入门指南，或供其他研究者参考。

关键词： 音译；综述；语料库；评价指标

Survey on Machine Transliteration

Zhuo Li[1,2]　**Zhijuan Wang**[1,2,*]　**Xiaobing Zhao**[1,2]
[1]School of Information Engineering, Minzu University of China
[2]Natural Language Resource Monitoring and Research Center of Minority Languages
*Corresponding author: Zhijuan Wang
li_zhuo98@163.com, wangzj.muc@gmail.com, nmzxb_cn@163.com

Abstract

Machine transliteration, the process of automatically converting text from one language to another based on phonetic similarity, is a subtask of machine translation that focuses on the translation of phonetic information. After transliteration, you can know the pronunciation of the source word in another language, making it easier for people who are not familiar with the source language to understand the language, and it is beneficial to eliminate language and spelling barriers. Machine transliteration plays an important role in natural language applications such as multilingual text processing, corpus alignment, and information extraction. This paper expounds the challenges existing in the current machine transliteration tasks, analyzes, categorizes and organizes the main transliteration methods, summarizes the transliteration data sets, and lists the commonly used evaluation indicators of transliteration effects. The existing problems are explained and the future of transliteration is prospected. This article is intended to provide a quick introductory guide for newcomers to the field, or as a reference for other researchers.

Keywords: Transliteration , Research Status , Corpus , Evaluation Metrics

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

机器音译是指利用计算机将源语言中的给定名称(源书写系统或拼写体系中的文本字符串)自动转换为目标语言中的名称(目标书写系统或拼写体系中的另一文本字符串)(Wei, 2004)。关于目标语言中名称表示的具体要求如下：它符合目标语言的音系，在语音上等同于源名称，并且与源语言名称的对等上符合用户的直觉。例如炒面在伦敦的中餐馆菜单里常被写作Chow Mein。机器翻译、数据挖掘以及跨语言信息检索和抽取等系统的性能极大依赖于命名实体(人名、地名、机构名、专有名词等)的音译准确性，尤其在涉及到人名、专有名称、技术术语时。因此，研究机器音译有重要的意义。

机器音译按照源语言(音译输入语言)与起源语言(来源于何种语言)是否一致可分为正向音译与反向音译(Mammadzada, 2021)。将单词从其起源语言音译为外语称之为正向音译。例如将张三(汉语)音译为Zhang San(英语)。而将用本语言拼写的外语词音译回起源语言称之为反向音译。例如将Zhang San(英语)音译回张三(汉语)。反向音译相比于正向音译来说更加困难。这是因为反向音译需要消除在正向音译中引入的噪声，正向音译的过程中往往会过滤掉不发音的音节，例如De Sciglio(意大利语)音译为德西利奥(汉语)，其中的字母g不发音。此外反向音译往往不允许有音译变体，它应该尽可能的接近原词也是反向音译更困难的一个重要的原因。比如说雷欧、李傲(汉语)音译为英语只有一个正确结果Leo。

音译与翻译和转写/转录都有所不同(Zepedda, 2020)。翻译在于使用不同语言传达语句的原始意义，其并不知道单词在原始语言中的发音。翻译与音译相反，它更注重单词的意义而不是发音。而转写是将一种字母表中的字符转换为另一种字母表中字符的过程[0]。转写字符之间是一一对应转换的，即被转换字母表中的每一个字符只能转换为另一个字母表中的一个字符，才能保证两个字母表能够完全的、无歧义的转换(冯志伟, 2012)。例如阿拉伯语单词كتب，其英语翻译为book，英语音译结果为kataba，而拉丁转写结果为ktb。

由于不同语言之间的较大差异性，音译任务存在着诸多困难与挑战。

一是源语言与目标语言使用的是不同的字母体系。例如拉丁/罗马字母源于希腊字母，它作为罗马文明的成果之一，随着征服推广到西欧地区。西里尔/斯拉夫字母是通行于斯拉夫语族部分民族中的字母书写系统。而阿拉伯/天方字母则在伊斯兰教兴盛的地区使用。音译处理的过程中需要了解不同字母体系中的字符编码。此外字母体系的书写方向也是必须要考虑的一点。例如阿拉伯字母、希伯来字母、波斯字母、乌尔都字母遵循从右到左的书写原则，而罗马字母、西里尔字母、婆罗米字母遵循从左到右的书写原则(Prabhakar and Pal, 2018)。

二是音译变体的存在。由于音译是一个基于个人认知的创造性过程，导致不同的专业音译者也有不同的观点。此外，同一种语言存在的不同方言也会导致音译变体的存在。而在音译语料的搜集过程中很难捕获到所有的变体。这种情况会让音译的质量评估变得很困难，因此很难建立起让所有人都信服的音译评估标准。

三是不同字母体系中涵盖音的范围不同，会导致发音缺失的问题。这与春秋时期创立的音阶——宫商角徵羽只能对应于现代音阶的do、re、mi、sol、la相类似(Jacques, 2017)。这将导致目标字母体系中缺少某些发音就必须使用多个字母来近似表示其发音，甚至会出现字母组合后仍无法找到类似发音的情况。因此需要让音译模型学习如何“创造”出缺失的相似发音，以保持发音的完整。

四是很难让音译模型学会“察言观色”。音译通常是对命名实体进行的。但如何让系统判断不同词采用音译还是翻译，需要模型通过从大量的训练语料或上下文中意识到这一点。例如Kunlun Mountains(英语)，第一个单词应该音译为昆仑(汉语)，而第二个单词应该翻译为山(汉语)。这对于传统的音译方法来说有着巨大的挑战，而基于深度学习的音译方法通过大量语料的学习和在注意力机制的帮助下相对来说能较容易的学习到这一点。

本文的组织方式如下。第二节描述了音译涉及到的主要语言。第三节综合阐述了具有代表性的音译方法，并对它们进行了分类整理。第四节罗列了音译的相关语料库资源。第五节介绍了音译质量/性能评估中常使用的指标。第六节对整个音译学的未来进行了展望，讨论了未来的工作方向。第七节对全文进行了总结。

[0]在实践中，如果两种语言中字母和声音之间的关系相似，则音译与转写十分接近。

第二十一届中国计算语言学大会论文集，第317页-第332页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

2 音译相关语言

我们统计了来自于谷歌学术的400篇与音译相关的论文，整理了其中涉及到的36种主要语言对，如表1所示。可以看到英语作为世界上使用最广泛的语言，对其相关语言对的研究占比最高。其次是语言使用人数/使用地区较多的汉语、阿拉伯语、日语、印地语等。而对于使用人数/使用地区较少的语言的研究并不多。从语系的角度来看，印欧语系(梵语、英语、旁遮普语、孟加拉语、波斯语等所属的印度-伊朗语族，法语、西班牙语所属的罗曼语族，英语、瑞典语所属的日耳曼语族，以及以俄语为代表的的斯拉夫语族)是主要的研究对象，而汉藏语系(汉语、藏语、泰语)、闪含语系(阿拉伯语、希伯来语)、南亚语系(越南语)、希腊语族(希腊语)和其他语系的研究相对较少[1]。

英语⟷汉语	英语⟷朝鲜语	英语⟷希伯来语	英语⟷阿拉伯语
英语⟷日语	英语⟷泰语	英语⟷波斯语	英语⟷孟加拉语
英语⟷俄语	英语⟷印地语	英语⟷旁遮普语	英语⟷西班牙语
英语⟷越南语	汉语⟷朝鲜语	英语⟷马拉地语	英语⟷泰卢固语
汉语⟷日语	日语⟷朝鲜语	印地语⟷乌尔都语	印地语⟷旁遮普语
英语⟶希腊语	英语⟶法语	英语⟶泰米尔语	英语⟶坎纳达语
英语⟶奥里亚语	梵文⟶英语	马拉地语⟶英语	古吉拉特语⟶英语
瑞典语⟶芬兰语	法语⟶日语	法语⟶汉语	法语⟶阿拉伯语
印地语⟶坎纳达语	藏语⟶汉语	西班牙语⟶汉语	阿拉伯语⟶印地语

Table 1: 音译研究涉及的主要语言对。日语包括日语汉字和日语片假名。朝鲜语包括朝鲜语汉字和朝鲜语谚文。朝鲜语同韩语，下同。A⟶B表示源语言A到目标语言B的音译方法。A⟷B表示A、B均可作为源语言或目标语言，下同。

3 方法

获取音译的方法主要有两类：音译生成和音译挖掘。音译生成是将一种语言中的给定单词自动生成为另一种语言表示的对应音译的过程。对于新的词语(不来源于训练语料)，生成性音译系统也可以自动生成目标语言音译词。音译挖掘是从不同资源中提取/挖掘音译对的过程，资源可以是平行语料库或可比语料库，也可以是两种语言之间的Web资源。音译对的自动提取可以用获取的新的音译对来丰富现有的音译语料库，减轻建设音译生成所需语料库的人力劳动。除这两类方法之外，还有一些方法将生成和挖掘结合起来进行音译，可以称它们为融合方法或者混合方法。

3.1 音译生成

用于音译生成的模型包括基于信道的模型、支持向量机、最大熵模型、决策树、隐马尔科夫模型、条件随机场、循环神经网络和Transformer等。**噪声信道模型(NCM)**以假设目标语言的文本T(信道意义上的输入)经过噪声信道变为源语言的文本S(信道意义上的输出)。音译模型将观察到的源语言文本S，转换成最有可能产生S的目标语言文本T'，即利用贝叶斯规则将$p(y|x)$重写为$p(x|y) * p(y)/p(x)$。由于NCM有两个组件模型($p(x|y)$和$p(y)$)意味着可以把整体的问题单独的解决，但它也受限于过高的解码成本。**信源信道模型(SCM)**是基于贝叶斯定理的混合模型，它借鉴了基于规则和统计方法的思想。其优点在于考虑了表示源语言单词的语音属性的字素，但其切分产生的错误会传播到后面的步骤中，会导致生成错误的音译结果。此外，由于两种语言都要生成可能的字素，因此时间复杂度较高。**联合信源信道模型(JSCM)**通过n-gram音译模型在不同语言之间进行直接拼写映射，JSCM与SCM相比，它在不分解联合概率的情况下估计了最优的音译结果字符，但JSCM并未使用语言学知识，且用概率模型识别源语言中的音译单元准确性仍有待提高。**支持向量机(SVM)**是一种用于二分类的机器学习算法，通过在特征空间上找到最佳的分隔超平面使得正负样本间隔最大。由于音译是一个多分类问题，因此需要先对问题进行二值化。在训练阶段，为两个不同字母体系的语言中每个字母训

[1]朝鲜语和日语所属语系现在仍存在着争议。

(c) 2022 中国中文信息学会计算语言学专业委员会

练一个SVM，就可在给定源语言字母序列下预测所有可能的类标签，并选择最可能的类标签。当观测样本较多时效率会明显下降。此外，核函数和参数的选择对结音译结果的影响很大。**决策树(DT)**是在已知各种情况发生概率的基础上，通过构建决策树学习如何将每个源字素转换为目标字素，从而生成音译结果。DT的优点在于它充分的考虑了广泛的上下文信息，但缺少语音信息的考虑。**最大熵模型(MEM)**是由最大熵原理推导实现的。最大熵原理可以表述为在满足约束条件的模型集合中选取熵最大的模型，即不确定性最大的模型。在选定特征作为约束且分配权重后，其余对音译有影响的特征将同等对待。MEM的优势在于它可以灵活地选择特征，鲁棒性强。但当样本量大时，对偶函数优化求解的迭代过程缓慢，开销较大。**隐马尔科夫模型(HMM)**是一种有限状态自动机，通过定义观察序列和标记序列的联合概率对生成过程进行建模。它由一个隐藏的马尔科夫链根据状态转移概率，随机生成一个状态随机序列，然后再由每个状态根据观测概率生成各自对应的一个观测，由此构成可观测的随机序列。因此输出标签的概率取决于当前输入标签和之前的输出。HMM的缺点在于它只依赖每一个状态和它对应的观察对象，忽略了观察序列的长度和上下文信息。**条件随机场(CRF)**是条件概率分布模型，定义了给定特定观察序列X(一串源语言音译单元)的标签序列Y(一串目标语言音译单元)上的条件概率$P(Y|X)$。条件随机场的优势在于避免了标签偏置的问题，所有特征能进行全局归一化，它相比HMM来说不需要独立性假设条件，因此可以容纳任意的上下文信息，但其训练代价大、复杂度高。**循环神经网络(RNN)**是处理序列数据的神经网络。初始化的字符向量表征传输到RNN中，以获取语义向量表征，并构建一个全连接网络来预测序列此时的隐藏状态，再根据隐藏状态预测出标签。但其不具备长期记忆，且会造成梯度消失的问题。而LSTM模型在此基础上加入了遗忘机制，选择性的保留或遗忘前期的某些数据，并用加法代替乘法解决了梯度爆炸的问题，但仍然难以捕获长距离的依赖关系，且RNN的序列递归结构使得难以并行化计算，效率过低。**Transformer**是一种全新架构的神经网络，它利用自注意机制解决了RNN和CNN中难以充分利用上下文信息的缺点。对齐后的双语预料的每个字符/音节被表征成向量，输入到编码器后经过自注意力层和全连接层进行残差连接和正则化，输出结果作为下一个编码器的输入，通过多次重复后又经过解码器解码得到音译结果。但它对于数据规模、计算成本要求较高。

音译生成的方法根据音译过程中信息的来源(发音或拼写)，可分为基于字素的方法(Θ_G)、基于音素的方法(Θ_P)、基于混合的方法(Θ_H)和基于组合的方法(Θ_C)，如表2所示，四类模型的示意图如图1所示。

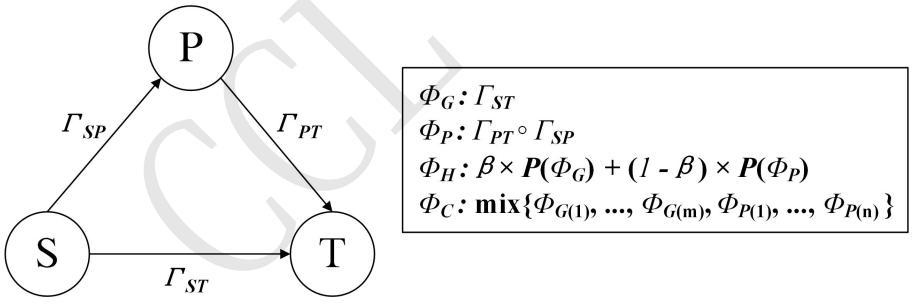


图 1. 四类音译生成模型。其中S表示源字素，P表示源音素，T表示目标字素，Γ_{ST}、Γ_{SP}、Γ_{PT}表示三个转换过程，Φ_G直接将源字素转换为目标字素，Φ_P将源字素转换为源音素，再生成目标字素，有时需要源音素到目标音素多一步的转换，Φ_H由Φ_G和Φ_P的线性插值组成，β是权重参数($0 \leq \beta \leq 1$)，Φ_C是多个字素、音素方法的混合，$(g \circ f)(x) = g(f(x))$。

基于拼写/字素的方法(Θ_G)：基于拼写的方法依赖于从单词的字符中获得的统计信息。它直接将源语言字素/字符/多个字符转换为目标语言字素/字符/多个字符。基于字素的方法旨在建立从源单词中的一组字符到目标单词中字符的直接映射模型。与基于音素的方法相比，基于字素的方法减少了音译过程中涉及的步骤，从而消除整个过程中的一些潜在错误源，实现起来也较为容易。但因缺少语音形式的信息，因此也很难获得精准的音译结果。尽管基于字素的方法总体上比优于基于音素的方法，但其在处理发音与拼写有很大差异的单词时性能较弱。Li等(2004)提出了一个音译框架，该框架通过联合信源信道模型实现了从英语到汉语的直接正字法映射。它通过n-gram模型，让正字法对齐实现自动化，能直

(c) 2022 中国中文信息学会计算语言学专业委员会

接从双语词典中得到对齐的音译单元。Malik(2008)提出了一个基于有限状态转换器的印地语-乌尔都语的音译模型，利用了转换器的特点，采用灵活通用的中间转录方案获得音译结果。Dhore等(2012b)使用条件随机场作为概率统计工具和n-gram作为特征集，实现了印地语-英语的命名实体音译。Wang等(2015)使用不同的字母分割方法来对英语到朝鲜语和英语到汉语的命名实体进行音译。Merhav和Ash(2018)比较了基于长短期记忆人工神经网络(LSTM)的编码器-解码器模型、Transformer和传统的加权有限状态转换器三者的性能，实验结果显示了基于Transfromer的方法表现最佳，和其余两个模型相比更适合音译。

基于音素的方法(Θ_P)：它将音译视为一个语音的过程，而不是一个正字法(字母映射)的过程。音素是声音的最小单位，该方法通常从源语言字素生成源语言音素，再从源语言音素映射到目标语言音素，最后生成目标语言字素。音素表征成为源语言与目标语言之间的中间形式/支点，所以也称为支点法。一般来说，基于音素的方法的主要优势是提升了发音在音译过程中的作用。然而，整个过程需要多个步骤，包括从源字素到源音素、源音素到目标字素、有时还需要源音素到目标音素的转换，这增加了错误传播的机会。音译系统可能在每一步生成很多个备选方案，多个步骤会导致早期的错误在后期显著增加。而这些错误直接影响候选音译目标词的最终排名，导致正确的音译排在错误的音译之后，或者可能生成完全错误的音译结果。基于音素的方法的另一个缺点是，这类方法依赖于双语发音资源，但并非所有语言都能轻易获得发音资源。Knight和Graehl(1998)利用加权有限状态变换器结合romaji(拼写日语的罗马字母)到音素、音素到对应的英语以及英语单词的概率实现了日语到英语音译建模。Jung等(2000)使用了扩展马尔可夫窗口实现了英语到朝鲜语的音译。该方法通过发音词典将英语单词转换为英语发音，再分割为音素块。每个块通过扩展马尔科夫窗口对应于手工规则定义下的朝鲜语字母。Oh和Choi(2002)提出了一个利用发音和上下文规则的英语到朝鲜语音译模型。它通过发音词典中提取英语发音单元并与对应音素对齐。对于字典中没有的单词则被拆分成两个单词继续在字典中进行搜索。若仍未找到。则用英语的希腊语起源单词。对齐后使用英语到朝鲜语标准转换规则(EKSCR)生成音译。Dhore等(2012a)提出了一个将印地语和马拉地语的命名实体音译为英语的模型。该方法基于发音统计法把音素和命名实体长度作为监督学习的特征，并将Unicode编码的印地语或马拉地语送给音节化模块，该模块将命名实体分割成若干音译单元。音译模块再通过语音图谱将梵文字母中的每个音节单元转换为英语。

基于混合的方法(Θ_H)：它结合了基于音素和基于字素的方法。通常，基于音素的方法比基于字素的方法更容易出错，它的成功率通常低于基于字素的方法。然而，基于字素的方法不能很好的处理发音与拼写有很大差异的单词。混合方法旨在结合这两类方法的优势，以提高整体精度，但实施起来较为困难。混合音译方法将基于字素和基于音素的信息结合到一个系统中，进而生成候选音译目标词。Oh和Choi(2005)提出了一种字素和音素相结合的音译模型，并且在英语到朝鲜语和英语到日语的数据集上进行了实验。AI-Onaizan和Knight(2002)基于有限状态机提出了一种结合音素和字素的混合音译模型，实现了阿拉伯语到英语的音译。

基于组合的方法(Θ_C)：能将多个基于字素或音素的方法(不局限于两者)组合起来。音译的组合方法把多个独立的系统输出汇集在一起，再将这些输出组合成一个候选音译列表。基于组合的方法可以将每个系统的独特优势相互结合，减少音译系统的错误率，给出更优的音译结果，同样它实施起来也较为复杂。Oh和Isahara(2007)使用音译系统的组合来研究英语到朝鲜语和英语到日语的音译。他们提出了一个基于支持向量机和最大熵模型的方法来对单个音译系统的输出进行重排序。每个系统生成的候选音译列表中给出不同生成目标词的置信度。Karimi等(2008)提出了基于辅音-元音的CV-Model方法。该方法在训练过程中将单词对齐后，通过替换元音和辅音生成辅音-元音的序列，序列与原字符一同分解为特定的模式。根据特定模式生成转换规则，从而形成音译模型。Najafi等(2018)的音译系统是基于多个系统的组合，他们使用了五种音译方法：DirecTL+(Jiampojamarn et al., 2010)、Sequitur(Bisani and Ney, 2008)、OpenNMT (Klein et al., 2018)、BaseNMT(Sutskever et al., 2014)和RL-NMT(Najafi et al., 2019)，取得了一定的效果。

根据使用的技术，音译生成可以分为**基于规则的方法**和**基于统计方法**(隐马尔科夫模型、条件随机场、决策树、支持向量机等)。基于规则的方法依赖大量人为规则及音译名称，系统易于实施且在缜密设计的规则下能提供良好的性能。但需要丰富的经验和大量的语言专业知识，根据一种语言设计的规则也不能转移到另一种语言，其对语言的依赖程度较高，整个工程的成本昂贵。而统计方法无需语言模型的专业知识，在大量数据和良好的算法下能让训练结果达到

(c) 2022 中国中文信息学会计算语言学专业委员会

满意的效果。统计学习的方法来源于机器学习和数据挖掘，都是从数据中学习。因此适应性、扩展性和维护成本都优于基于规则的方法。但其仍存在着不足，训练数据的规模对音译结果质量影响较大，有些语言甚至没有可供使用的语料库资源，且需要上下文信息来引导结果的生成。深度学习(RNN、Transformer等)最初是机器学习中的一个小分支，近年来随着算力资源的发展已经开始成为新的研究方向，它相比传统的机器学习需要更庞大的数据量和运算资源，通过一种端到端的模式解决整个问题代替传统机器学习的多个子问题逐个解决。还有一个显著差别在于深度学习提取的是单词的深层次特征，这些特征虽然目前很难解释，但从结果上看却能准确的表示出单词的某些特性。因此，也可以将基于深度学习的方法单独作为一类。

从总体上看，近年来基于规则的、基于传统机器学习的模型使用频率相对较少。这是因为基于规则的方法成本昂贵，而基于传统机器学习的模型在数据量较大的情况下，效果不如基于深度学习的模型。而基于深度学习的模型是目前机器音译研究的主流。在最近一次的命名实体研讨会(NEWS 2018)的音译评测任务中，除了新加坡国立大学和新加坡科技设计大学组成的团队提供了使用Sequitur和Moses的基线系统，其余团队则都使用了基于神经网络的模型。英国的爱丁堡大学参加了15个音译任务的评测，并在11个任务评测中取得了第一的成绩，该团队使用的是带有注意力机制的RNN编码器-解码器模型(Chen et al., 2018a)。目前为止，关于音译生成的研究涉及的语言对比较多，音译研究使用的方法也各有不同。语言的不同特点、训练语料库的大小，以及是否加入语言学家的知识等，这些都对选择适合语言的的研究方法有着影响。而且对于同一语言对之间的音译，研究者们使用的数据集也不尽相同，所以很难确切地比较出不同方法的优劣。

分类	作者及年份	语言对	模型	数据集(规模)	评价指标(分数)
基于音素	(Knight and Graehl, 1998)	日语→英语	有限状态转换器	N/A(1,549)	准确率(64%)
	(Jung et al., 2000)	英语→朝鲜语	扩展马尔科夫窗口	N/A(8,368)	召回率(87.5%)
	(Oh and Choi, 2002)	英语→朝鲜语	直接映射模型	N/A(7,185)	准确率(67.83%),字符准确率(93.49%)
	(Surana and Singh, 2008)	英语→印地语① 英语→泰卢固语②	识别适应性音译机制	N/A(2,000)	①:MRR(87%),精确率(80%); ②:MRR(82%),精确率(71%);
	(Dhore et al., 2012a)	印地语→英语 马拉地语→英语	统计模型	N/A(15,224)	准确率(97.3%),F值(94.2%),MRR(96.8%), 召回率(93.2%),精确率(95.7%)
基于字素	(Li et al., 2004)	英语↔汉语	联合信源信道模型	N/A(37,694)	正向准确率(70.10%),反向准确率(37.9%)
	(Rama and Gali, 2009)	英语→印地语	噪声信道模型	NEWS 2009 (10,949)	准确率(46.3%),F值(87.6%), MRR(57.3%),MAP(45.4%)
	(Dhore et al., 2012b)	印地语→英语	条件随机场模型	N/A(7,251)	准确率(65.01%)@1-gram
	(Grundkiewicz and Heafield, 2018)	泰语→英语① 英语→泰语② 英语→希伯来语③ 英语→孟加拉语④ 英语→坎那达语⑤ 英语→泰米尔语⑥ 英语→印地语⑦ 汉语→英语⑧ 英语→汉语⑨ 英语→越南语⑩ 希伯来语→英语⑪	带注意力机制的RNN编码器-解码器模型	NEWS 2018▲	平均F值{①(0.8454); ②(0.7760); ③(0.8042); ④(0.9006); ⑤(0.8673); ⑥(0.8405); ⑦(0.8515); ⑧(0.8300); ⑨(0.6791); ⑩(0.8893); ⑪(0.7532);}
	(Merhav and Ash, 2018)	英语→阿拉伯语① 英语→希伯来语② 英语→片假名③ 英语→俄语④ 阿拉伯语→英语⑤	Transformer I 基于LSTM的编码器-解码器模型 II	SubWikiLang▲	错误率{①I:(0.45),II(0.53); ②:I(0.44),II(0.49); ③I:(0.51),II(0.60); ④:I(0.35),II(0.40); ⑤I:(0.75),II(0.81);}
	(Chatterjee and Sarkar, 2021)	孟加拉语↔英语	支持向量机模型 I 隐马尔可夫模型 II	N/A(1,000)	准确率{正向I(53.9%),正向II(51.9%); 反向I(45.2%),反向II(43.1%);}
基于混合	(Al-Onaizan and Knight, 2002)	阿拉伯语→英语	信源信道模型 加权有限状态转换器	N/A(—)	准确率(49.08%)
	(Oh and Choi, 2005)	英语→朝鲜语① 英语→日语②	决策树模型 I 最大熵模型 II 基于记忆学习 III	EKSet(7,185) EJSet(10,398)	准确率{ ①I(62.0%),II(63.3%),III(66.9%); ②I(66.8%),II(67.0%),III(72.2%);}
基于组合	(Karimi, 2008)	英语→波斯语① 波斯语→英语②	信源信道模型 和投票法	①:N/A(1,500) ②:N/A(2,010)	①:正确率(74%) ②:正确率(53%)
	(Najafi et al., 2018)	阿拉伯语→英语① 英语→日文汉字② 英语→片假名③ 英语→朝鲜语④ 波斯语→英语⑤ 英语→波斯语⑥	DirecTL+ Sequitur OpenNMT BaseNMT RL-NMT	NEWS 2018▲	平均F值{①(0.9087); ②(0.7678); ③(0.8098); ④(0.7113); ⑤(0.9515); ⑥(0.9373);}

Table 2: 音译生成经典技术。N/A表示相关数据集并未公开，——表示数据集未注明大小，▲表示涉及到的数据集详见第四节。EKSet和EJSet数据集现已无法访问。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.2 音译挖掘

音译对通常是从平行语料库或可比语料库或Web中挖掘出来的。平行语料库是两种或多种语言的对齐文本集合。对齐文本是一种语言到一种或多种语言的精确翻译。可比语料库也是两种或多种语言的文本集合，各种语言的文本是相似的，但不是彼此的精确翻译。

迄今为止，音译领域已经涌现出许多音译挖掘技术。这些技术可以分为三类：基于语音相似度、基于机器学习技术和基于词共现。

基于语音相似度：基于语音相似度的方法测量平行或可比语料库中词1和词2之间的相似性，并提取出词2'作为词1'最接近的候选音译词。可以用许多方法计算相似性，如莱文斯坦距离算法、最长公共子序列(LCS)算法和Jaro-Winkler距离(Jaro, 1989)算法。Udupa等(2008)提出了一种命名实体等价物/对等物挖掘方法，该方法通过跨语言文档相似度和音译相似度模型，能够有效地从可比语料库中挖掘命名实体音译的等价物。

基于机器学习：El-Kahki等(2011)提出了一种增强的音译挖掘技术，该技术使用生成图强化模型来推断源字符序列和目标字符序列之间的映射。Fukunishi等(2013)提出了一种挖掘音译对的技术，该方法在对齐过程中使用非参数化的贝叶斯方法。

基于词共现：Karimi(2011)已经对基于词共现的部分提取模型进行过讨论。Wu等(2012)为NEWS 2012提出了一种英韩音译系统。他们训练了多个音译模型，使用两种重排序方法从不同模型的预测结果中选择最佳的音译模型。其中一种重排序方法是基于网络语料库中音译对的共现。另一种是基于对齐结果特征的间接监督联合学习重排序的方法。实验结果表明使用基于网络重排序方法的音译模型可以在英语-朝鲜语音译中获得更优的结果。

3.3 融合方法

有些方法同时使用音译生成和音译挖掘技术。Zhao等(2007)提出了一种基于隐马尔可夫模型的框架来音译命名实体，此外，通过与从网络搜索引擎收集的统计数据生成的自动拼写检查器相结合，进一步提高了音译准确性。Chinnakotla等(2008)开发了一个印地语和马拉地语到英语的跨语言信息检索系统，将音译生成、音译提取等多种技术有效地融合在了一起。

4 音译语料库资源

我们对可用于音译任务的数据集进行了搜集和整理，如表3所示。列出的数据集仅包含目前可以获取(免费/收费)的，对于现无法访问的数据集未列出，数据集的详细介绍见附录A部分。

- NEWS 2018[2]：继NEWS 2009(ACL-IJCNLP 2009)、NEWS 2010(ACL 2010)、NEWS 2011(IJCNLP 2011)、NEWS 2012(ACL 2012)、NEWS 2015(ACL 2015)和NEWS 2016(ACL 2016)之后举办的连同音译共享任务的命名实体研讨会，旨在提供一个通用平台，用于对跨多种语言的不同音译方法和系统进行基准测试。

- 中日朝鲜词典研究所(CJKI)为汉语、日语、朝鲜语、阿拉伯语、西班牙语等语言编制了一系列综合字典数据集[3]，这些数据集包含大量通用词汇、专有名词和技术术语的语法、语音和语义属性。其中可用于音译任务的包括汉语拼音-汉语数据集(CHD)、汉英人名数据集(CEN)、中英地名数据集(CEP)、中日人名数据集(CJN)、中日地名数据集(CJP)、日语-多语言地名数据集(JMP)、日本公司数据集(JCD)、日英人名数据集(JEN)、朝鲜英人名数据集(KEN)、朝鲜英地名数据集(KEP)、朝鲜日人名数据集(KJN)、朝鲜日地名数据集(KJP)、朝鲜汉人名数据集(KCN)、朝鲜中地名数据集(KCP)、阿拉伯语外国名字数据集(DAFNA)、阿拉伯地名数据集(DAPNA)、美国财政部外国资产控制办公室名单扩展(XOFAC)和中越人名数据集(CVP)。

- 英语-乌克兰语数据集[4]：从维基百科中提取的英语-乌克兰语命名实体数据集，该项目由欧盟委员会赞助。

[2] http://workshop.colips.org/news2018/dataset.html
[3] https://www.cjk.org/data/all
[4] http://catalog.elra.info/en-us/repository/browse/ELRA-M0104

(c) 2022 中国中文信息学会计算语言学专业委员会

数据集	语言对/数据量	总规模	数据集	语言对/数据量	总规模
NEWS 2018	英语→泰语	32,781	CHD$	汉语拼音→汉语	超过60万
	泰语→英语	29,273	CEN$	汉语→英语	超过200万人名
	英语→波斯语	15,386	CEP$	汉语→英语	超过百万地
	波斯语→英语	17,677	CJN$	汉语→日语	超过200万人名
	英语→汉语	43,318	CJP$	汉语→日语	超过10万地名
	汉语→英语	34,002	JMP$	日语→13种	超过310万地名
	英语→越南语	5,256	JCD$	日语→英语	超过60万公司名
	英语→印地语	14,937	JEN$	日语→-英语	超过55万人名
	英语→泰米尔语	12,957	KEN$	朝鲜语→英语	超过200万人名
	英语→坎那达语	12,955	KEP$	朝鲜语→英语	约9万地名
	英语→孟加拉语	15,623	KJN$	朝鲜语→日语	超过200万人名
	英语→希伯来语	12,501	KJP$	朝鲜语→日语	约9万地名
	希伯来语→英语	11,447	KCN$	朝鲜语→汉语	超过200万人名
	英语→片假名	30,828	KCP$	朝鲜语→汉语	约9万地名
	英语→日文汉字	12,514	DAFNA$	英语→阿拉伯语	超过24万人名
	英语→朝鲜语谚文	9,387	DAPNA$	阿拉伯语→英语	超过1万地名
	阿拉伯语→英语	33,354	XOFAC$	英语→阿拉伯语	超过2500万人名
	波斯语→英语	8,000	CVP$	汉语-越南语	超过4.3万人名
	英语→波斯语	13,204	英语-乌克兰语数据集	英语→乌克兰语	624,168
SubWikiLang	英语→俄语	164,640	中英命名	汉语→英文	797,452
	英语→片假名	98,820	实体列表$	英文→汉语	869,136
	英语→阿拉伯语	74,973	TRANSLIT	超过180种	3,008,239实体
	英语→希伯来语	50,049	ParaNames	超过400种	14,017,168实体
	阿拉伯语→英语	15,898	Trabina	591种	1,129
	英语→Arpabet	126,191	BanglaNLP⋆	孟加拉语→英语	13,214
Xlit-Transliteration	印地语→英语	32,508	Xlit-Crowd	印地语→英语	14,919
	迈蒂利语→英语	28,88	Xlit-IITB-Par	印地语→英语	68,922
	孔卡尼语→英语	21,342	TfNSW	英语→12种	1538个车站名
Bittlingmayer⋆	英语→亚美尼亚语	39,707	Aksharantar	21种印度语言→英语	2600万
	英语→希腊语	37,505	FIRE 2013	印地语→英语	1,462
	英语→波斯语	78,663	Praneeth⋆	英语→泰卢固语	38,568
	英语→俄语	179,853	ANETAC	英语→阿拉伯语	79,924
EnToFrNE$	英语→法语	1,167,263	EnToSSLNE$	英语→6种南斯拉夫语言	26,155

Table 3: 音译数据集资源。注意：每届NEWS研究会的数据集在上一届的基础上进行扩展，在此只列出了最近举办的一届情况。词的发音模拟为表示音子和语段的符号串，音子是言语的发音，用语音符号表示。音子包含三套不同的字母符号——IPA，Arpabet，SAMPA，其中Arpabet 是高级研究计划署(ARPA)开发的语音转录代码。⋆表示数据集的作者未对数据集命名，这里用作者(团队)名称代替数据集名称。$表示数据集需要付费才可获取。

- SubWikiLang[5](Merhav and Ash, 2018)：它包括从维基数据(Wikidata)搜集后过滤的以英语为源语言的4组人名数据集、Rosca和Breuel(2016)从维基百科(Wikipedia)标题中提取的阿拉伯语到英语的数据集和CMU发音字典共同组成。维基数据是维基媒体基金会主持的一个自由的协作式多语言辅助知识库，旨在为维基百科、维基共享资源以及其他的维基媒体(Wikimedia)项目提供支持(是它们的超集)，其中的每个文档都有一个主题或一个管理页面，且被唯一的数字标识。CMU发音字典是由卡内基梅隆大学创建的一个开源发音词典，它为北美发音中的英语单词提供映射拼写/语音。

- 中英命名实体列表(LDC2005T34)[6]：由语言数据联盟(Linguistic Data Consortium, LDC)提供，包含九对从新华社通讯社文本中汇编的汉语-英文双向名称实体列表。LDC是由大学、图书馆、公司和政府研究实验室组成的语言公开联盟，隶属于宾夕法尼亚大学文理学院。

- TRANSLIT[7]：由数据集JRCNames(Ehrmann et al., 2017)、SubWikiLang[5]、Geonames[8]、

[5] https://github.com/steveash/NETransliteration-COLING2018
[6] https://catalog.ldc.upenn.edu/LDC2005T34
[7] https://github.com/fbenites/TRANSLIT
[8] https://download.geonames.org/export/dump/alternateNamesV2.zip

(c) 2022 中国中文信息学会计算语言学专业委员会

谷歌英语-阿拉伯语音译数据集[9]及其维基百科中所有语言的命名实体转储数据共同统一格式后合并而成。产生的数据集包含180多种语言的约160万条词条，以及约300万个名称变体(Benites et al., 2020)。

- ParaNames[10](Sälevä and Lignos, 2022)：它采用与SubWikiLang[5]中维基数据基于相同的获取方式。所不同的是，SubWikiLang只包含几种人名数据，而ParaNames包含了维基数据中的所有语言对，包含人名、地名、组织机构名三种实体。
- Trabina[11]：Wu等(2018)利用《圣经》的广泛传播性，把其中1129个英文姓名翻译成了591种语言，平均一个姓名有其他语言的52%覆盖率，超越了维基百科的覆盖率。
- BanglaNLP团队和Bengali.AI社区成员发布了一个从维基百科中爬取整理后的孟加拉语-英语的音译数据集[12]。
- Xlit-Transliteration[13]：作者在母语使用者的帮助下创建了三种印度语言(印地语、迈蒂利语、孔卡尼语)-英语的数据集。
- Xlit-Crowd[14]：由众包/外包网站Amazon Mechanical Turk获得，包含14,919对印地语-英语的音译对(Khapra et al., 2014)。
- Xlit-IITB-Par[15]：由Moses音译模块从印度理工学院和印度语言技术中心提供的英语-印地语平行语料库中自动挖掘出来，包含68,922对音译对(Kunchukuttan et al., 2018)。
- TfNSW[16]：由澳大利亚新南威尔士州交通网络中每个车站和码头的英语名称到12种语言(阿拉伯语、法语、德语、希腊语、印地语、意大利语、日语、朝鲜语、简体中文、繁体中文、西班牙语、越南语)的音译结果组成。
- Bittlingmayer从维基百科中下载了英语到亚美尼亚语、希腊语、波斯语、俄语四种语言的文章并把单词转换它们为单一字符，从而生成了四个音译数据集[17]。
- Aksharantar[18]：由21种印度语言-英语的音译对组成，是目前印度语言最大的公开音译数据集(Madhani et al., 2022)。
- FIRE 2013[19]：第五届信息检索论坛会议(FIRE 2013)提供了一个较小的印地语-英语的音译数据集，该数据集是通过使用宝莱坞歌词对齐后收集得到的(Roy et al., 2013)。
- Praneeth从维基百科平行语料的文章标题中，以及从美国社交新闻聚合、网络内容讨论网站Reddit的社区将讨论文本音译创建了英语-泰卢固语的数据集[20]。
- ANETAC[21]：一个英语到阿拉伯语的命名实体音译和分类数据集，该数据集由公开获得的平行翻译语料库建立(Ameur et al., 2019)。
- EnToFrNE[22]：由1,167,263个英语和法语平行命名实体组成，实体包括任务、组织、地点、产品和杂项五大类组成。

[9] https://github.com/google/transliteration
[10] https://github.com/bltlab/paranames
[11] https://github.com/wswu/trabina
[12] https://github.com/arijitx/BanglaNLP
[13] http://transliteration.ai4bharat.org/#/resources
[14] https://github.com/anoopkunchukuttan/crowd-indic-transliteration-data
[15] https://www.cfilt.iitb.ac.in/iitb_parallel
[16] https://opendata.transport.nsw.gov.au/dataset/tfnsw-station-names-other-languages
[17] https://github.com/deepchar/deepchar
[18] https://huggingface.co/datasets/ai4bharat/Aksharantar
[19] http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/index.html
[20] https://github.com/notAI-tech/Datasets/tree/master/En-Te_Transliteration
[21] https://github.com/MohamedHadjAmeur/ANETAC
[22] http://catalog.elra.info/en-us/repository/browse/ELRA-M0052

第二十一届中国计算语言学大会论文集，第317页-第332页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

- EnToSSLNE[23]：由七种语言的26,155个平行命名实体组成，包括英语和六种南斯拉夫语：波斯尼亚语、保加利亚语、克罗地亚语、马其顿语、塞尔维亚语和斯洛文尼亚语。

5 音译质量评价指标

在介绍具体指标之前，我们首先给出相关符号的定义：

- N：测试集中实体(样本/名称)的数量。
- J：测试集中实体的参考音译结果(标签/正确音译结果)的数量。
- K：音译系统输出的候选音译(音译预测/输出结果)的数量。
- n_i $(1 \leq i \leq N)$：测试集中第i个实体的参考音译结果数量，音译任务中存在一个名称有多个对应的正确音译结果的情况。
- $r_{i,j}$ $(1 \leq i \leq N,\ 1 \leq j \leq J)$：测试集中第$i$个实体的第$j$个参考音译。
- $c_{i,l}$ $(1 \leq i \leq N,\ 1 \leq l \leq K)$：测试集中第$i$个实体的第$l$个候选音译。

我们根据现有论文对音译性能的主要评价指标进行了整理，如表4所示。

定义	公式
前k预测准确率($Top-k\ ACC$，简称ACC)表示所有预测结果中概率最大的前k个结果中包含正确音译的比例，是最常用的音译评价指标。值为1表明每个单词的k个结果中至少有1个是正确的，值为0表明所有单词的音译结果都与候选词不匹配。通常k取1。有时候也用单词错误率(WER)来替代，准确率和单词错误率之和为1(Chen et al., 2018b)。	$Top-k\ ACC$ $= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{l=1}^{k(\leq K)} \begin{cases} 1, & r_{i,j}=c_{i,l}\ (\exists\ r_{i,j}) \\ 0, & \text{otherwise} \end{cases}$
召回率(R)、精确率(P)是两个衡量第i个单词音译输出结果性能的指标。它们根据候选词(c)与最佳匹配参考词(r)的最长公共子序列(LCS)长度计算得到，LCS由c、r和它们之间的编辑距离(ED)共同决定，最佳匹配由二者的最小编辑距离决定，$\|x\|$表示x的长度(Chen et al., 2018b)。	$LCS\left(c,r\right)=\frac{1}{2}\left(\|c\|+\|r\|-ED\left(c,r\right)\right)$ $r_{i,m}=\arg\min_j\left(ED\left(c_{i,k},\ r_{i,j}\right)\right)$ $R_i=\frac{LCS\left(c_{i,l},\ r_{i,m}\right)}{\|r_{i,m}\|}$ $P_i=\frac{LCS\left(c_{i,l},\ r_{i,m}\right)}{\|c_{i,l}\|}$
平均F值(简称F值)衡量了音译候选词与最接近的参考词之间的差异，它同时考虑了R和P。对于每个源词，当候选词与任意参考词的LCS为0时，F值为0。当第一个候选词匹配其中一个参考词时，F值为1(Chen et al., 2018b)。	$F_i=2\frac{R_i\times P_i}{R_i+P_i}$
平均倒数排名(MRR)是给定一组实体，其所有音译结果排序列表中第1个正确音译的排名倒数的平均值，$1/MRR$表示样本对应的正确音译结果的平均排名。MRR接近1表明正确音译结果靠近n-best列表顶部(Chen et al., 2018b)。	$MRR=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{Rank}_i}$ $=\frac{1}{N}\sum_{i=1}^{N}\begin{cases}\min_j\ \frac{1}{j}, r_{i,j}=c_{i,l}(\exists\ r_{i,j},c_{i,l}) \\ 0,\ \text{otherwise}\end{cases}$
平均精度均值(MAP)与MRR所不同，MAP考虑了所有参考音译，它衡量了第i个源名称的n-best的精准度，$num(i,k)$表示k-best列表中第i个源词的正确候选数量(Chen et al., 2018b)。	$MAP=\frac{1}{N}\sum_{i}^{N}\frac{1}{n_i}\left(\sum_{k=1}^{n_i}num\left(i,k\right)\right)$
字符准确率(CA)衡量了音译对中匹配字符的比例。有时候也用字符错误率(CER)替代，它们相加和为1(Nießen et al., 2000) 。	$CA=\frac{\|r\|-ED\left(c,r\right)}{\|r\|}$

Table 4: 音译任务常用的评价指标。

[23] http://catalog.elra.info/en-us/repository/browse/ELRA-M0051

6 未来研究方法和开放问题

6.1 构建公开、权威的音译数据集

目前虽然一些会议已经组织过音译相关的评测任务，并提供了一些公开的音译数据集。但是整体上来说仍然存在着语言对不够丰富、数据集规模不够大、数据集质量不高等问题，并没有专门针对不同语言对的统一音译任务数据集。音译数据集不同于翻译数据集，其更难获取。目前的音译数据集往往来源于各国的政要、名人、地名、组织以及外来词。若直接从翻译数据集中筛选出用于音译任务的数据集则比较困难(其在翻译数据集中占比很低、通晓不同语言的专家很少)。在不违背各国法律的基础上，急需各国各自贡献一部分标准实体名称用于全球音译数据库的构建，这对于音译的研究将会起很大的推进作用。后续的研究者也将会有统一的参考依据为方法之间的性能比较。但也要注意，这必将花费大量时间、金钱和人力资源成本。

6.2 低资源语言音译研究

从表1中看到目前音译的研究集中于世界上7000种语言中的30多种，而绝大多数语言还没有被研究。这些语言的使用者较少，其数据集资源量也较低，常称它们为低资源语言(Magueresse et al., 2020)。低资源语言缺乏有用的训练属性，如受监督的数据、母语使用者以及专家的数量等，但不应忽略其存在。仅非洲和印度就大约有2000种低资源语言，有超过25亿的人使用它们(Tsvetkov, 2017)，为这些语言开发音译技术具有相当可观的经济前景。此外，研究音译可以支持某种语言，防止其灭绝并提升其影响力。中国的“一带一路”政策涉及全球数十个国家、数十亿人口和上百种民族语言，利用技术手段突破语言障碍问题的重要性日益凸显。针对低资源语言匮乏的特点，一些学者也开始尝试了一些新的方法。Le等(2019)提出了一种利用基于RNN的模型建立一个低资源机器音译系统的方法。Upadhyay等(2018)提出了一种使用约束发现来挖掘名称对以重新训练生成模型的Bootstrapping算法。但这些方法的效果还离实际落地使用差得较远。未来针对低资源语言，我们建议从两个方面入手。一是采用多种数据增强的方式扩大其数据量并提高数据的质量。二是利用语言模式/相似性、设计更健壮的学习模型以提高音译的准确性和鲁棒性。

6.3 深度学习模型的可解释性

现有的基于深度学习的音译方法研究还较少，但它的出现可以说重振了整个音译学(Santos et al., 2018; Shillingford and Parker Jones, 2018; Khare et al., 2021)。其在音译准确性、泛化性、鲁棒性方面的表现均优于传统方法，有时甚至优于人类基线(Alam and ul Hussain, 2021)，展现出了令人难以置信的性能。传统方法是基于人工认知所驱动的，需要语言学家针对语言及其结构(语法、句法、语音学等)进行研究，设计出音译模型，人工干预性很强。但是无论是基于统计的方法、决策树、支持向量机或更复杂的模型，仍然可以理解模型的内部结构。但是基于数据驱动的深度学习的音译方法存在着乃至整个深度学习社区目前都难以解答的问题，也就是说深度学习模型是从中怎么学习到不同语言各自的特性以及他们之间的对应关系。深度学习的模型不需要手动设计的特征，但由于它的黑盒性质导致我们不能对其进行分析和理论上的改进。而透明度是模型可解释性的关键所在，有助于我们理解模型中的每一层的学习内容以及它们是如何交互的，甚至能让我们能够找到模型所存在的漏洞，防止恶意攻击和非法侵入。虽然说现在已经有一些像注意力机制(Alam and ul Hussain, 2021)、知识图谱(Moussallem et al., 2020)、模拟模型方法(Hou and Zhou, 2020)等来对其作出了一定的解释，但是其解释并不能得到整个深度学习社区的一致同意，仍然存在着分歧(Cremer, 2021)。为了让基于深度学习的音译模型不超出人类设计它时的范围(道德伦理、预设规则、潜在偏见、种族歧视等)(Muscat, 2011)，在未来将音译模型应用于现实世界迫切的需要完备的可解释性，以使系统更加透明、更具解释性、可控性、安全性和健壮性，这是接下来需要思考和研究的方向。

6.4 迁移学习方法的使用

近年来，基于深度学习的音译取得显著的进展，这是由于如NEWS研讨会所提供的大规模音译平行预料。但这些预料仅关注了高资源语言，整个音译社区极度缺少低资源语言(Khare et al., 2021)。尽管基于深度学习的音译模型已经取得了较好的效果，但支持其训练需要足够大的双语平行语料库。除英语与其他语言外，即使对于母语使用人数众多的语言，如汉语、印度语、俄语、日语，它们之间也缺少足够的音译数据集。除了我们在6.1节所介绍的扩充数

据集的方法外，迁移学习也是音译数据集稀缺的一种解决方案(Maimaiti et al., 2019; Wu et al., 2022)。迁移学习的主要思想是将父模型(源域)的部分参数和共享知识传递到子模型(目标域)中，子模型仅需较小的数据量和训练时间就可以取得较不错的性能。英语作为全世界最广泛使用的语言，与英语相关的音译研究是最多的，可以把英语作为解决稀缺资源的桥梁。在源语言-英语或英语-目标语言数据集上训练一个音译模型(父模型)，根据这个模型来微调源语言-目标语言的模型(子模型)。但这其中存在着一个问题，语族(域)是迁移学习父模型的选择的一大关键。来自同一语系的语言在句法、语义、语言特征中存在着一些相似性，这些相似性有助于提升父模型的泛化能力，对迁移学习是有利的。而英语属于印欧语系中的日耳曼语族语支，很多语言却并不属于这个语支。如果忽视语系之间的相似性，也许会对迁移学习起到反作用，这也是未来我们要研究的方向。

6.5 多种音译方法相结合

单一的机器音译方法不能完全解决音译中存在的问题，将多种音译方法相结合起来，发挥不同方法的优势，能给出更好的音译结果。不同音译方法侧重点所不同，无论是音义兼顾(Usunier and Shaner, 2002)、音同义，还是谐音音译(Chen, 2013)都存在各自的优势。深度学习模型由于黑盒性质，需要吸收传统音译方法中语言学家设置的原则性音译规则，包括消极意义、敏感政治、恐怖主义、分裂主义、极端主义、殖民文化、歧视、黄赌毒等各国法律禁止的内容都不应该在音译结果中所出现。近年来一些学者也开始尝试多音译方法组合，Karimi等(2011)将不同的传统方法进行了组合实验，Nicolai等(2015)将三种音译模型进行了组合实验，Najafi等(2018)进行了神经网络与传统方法相结合的音译实验，他们的结果共同都表明了音译方法相结合优于单独使用任一方法。

6.6 不同地区的音译差异

我们仍然需要注意一种语言的音译结果在不同国家或地区的差异性问题。比如说汉语在中国大陆、香港、澳门、台湾、新加坡等地区存在着不同的音译差异，同一模式下存在着变体问题。就人名汉语音译英语而言，中国大陆人名使用的是汉语拼音，台湾地区很多人名的英文使用的则是威妥玛拼音(Xing and Feng, 2016)，而在香港、澳门地区人名的英文往往是粤式拼音、上海话拼音、英文或它们的组合(Man, 2012)，但在新加坡地区的人名英文则是粤语、潮州话、福建话/闽南语或它们的组合(Su, 2022)。这样的问题也存在于英译汉当中，比如说Reagan在大陆译为里根，台湾译为雷根，而香港则译为列根；Bush在大陆译为布什，台湾译为布希，而香港译为布殊。这是由于各地在不同的制度、社会背景、生活习惯、文化习俗、翻译准则等多方面的因素影响下共同造成的。是根据使用的不同地区训练不同的音译模型还是让模型都吸收来自不同的地区的音译结果，这也是未来值得探讨和研究的。音译技术在一定程度上能反应出这个词的历史背景和地区，能为历史和语言研究提供不小的帮助。

7 结论

在本文的工作中，我们回顾了音译的相关经典模型并对它们进行了分类整理，同时对音译数据集和评价指标进行了汇总，最后指出了整个音译社区目前待解决的问题和对未来的研究方向进行了展望。

在这个信息洪流时代，随着新词语的不断出现，将外来词融入本国语言变得越发普遍。相信音译作为解决这一问题的支持工具必将会受到更多的关注。

参考文献

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in Arabic texts. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Mehreen Alam and Sibt ul Hussain. 2021. Deep learning-based roman-urdu to urdu transliteration. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(04):2152001.

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2019. ANETAC: arabic named entity transliteration and classification dataset. *CoRR*, abs/1907.03110.

(c) 2022 中国中文信息学会计算语言学专业委员会

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France, May. European Language Resources Association.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Soma Chatterjee and Kamal Sarkar. 2021. Machine transliteration using svm and hmm. *Int. J. Adv. Intell. Paradigms*, 19(1):3–27, jan.

Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. 2018a. Report of NEWS 2018 named entity transliteration shared task. In *Proceedings of the Seventh Named Entities Workshop*, pages 55–73, Melbourne, Australia, July. Association for Computational Linguistics.

Nancy Chen, Xiangyu Duan, Min Zhang, Rafael E. Banchs, and Haizhou Li. 2018b. NEWS 2018 whitepaper. In *Proceedings of the Seventh Named Entities Workshop*, pages 47–54, Melbourne, Australia, July. Association for Computational Linguistics.

Yan Chen. 2013. On lexical borrowing from english into chinese via transliteration. *English Language and Literature Studies*, 3(4):1.

Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani, and Pushpak Bhattacharyya. 2008. Hindi to english and marathi to english cross language information retrieval evaluation. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, pages 111–118, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carla Zoe Cremer. 2021. Deep limitations? examining expert disagreement over deep learning. *Progress in Artificial Intelligence*, 10(4):449–464.

Manikrao Dhore, Shantanu Dixit, and Ruchi Dhore. 2012a. Optimizing transliteration for Hindi/Marathi to English using only two weights. In *Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology*, pages 31–48, Mumbai, India, December. The COLING 2012 Organizing Committee.

Manikrao L Dhore, Shantanu K Dixit, and Tushar D Sonwalkar. 2012b. Hindi to english machine transliteration of named entities using conditional random fields. *International Journal of Computer Applications*, 48(23):31–37.

Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. 2017. Jrc-names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8(2):283–295.

Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, Mohamed Abd El-Wahab, Ahmed Hefny, and Waleed Ammar. 2011. Improved transliteration mining using graph reinforcement. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1393, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Takaaki Fukunishi, Andrew Finch, Seiichi Yamamoto, and Eiichiro Sumita. 2013. A bayesian alignment approach to transliteration mining. *ACM Transactions on Asian Language Information Processing*, 12(3), aug.

Roman Grundkiewicz and Kenneth Heafield. 2018. Neural machine translation techniques for named entity transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 89–94, Melbourne, Australia, July. Association for Computational Linguistics.

Bo-Jian Hou and Zhi-Hua Zhou. 2020. Learning with interpretable structure from gated rnn. *IEEE transactions on neural networks and learning systems*, 31(7):2267–2279.

Guillaume Jacques. 2017. Traditional chinese phonology. *webpage, http://www. academia. edu/2261629/Traditional_Chinese_Phonology.*

Matthew A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

(c) 2022 中国中文信息学会计算语言学专业委员会

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2010. Integrating joint n-gram features into a discriminative training framework. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 697–700, Los Angeles, California, June. Association for Computational Linguistics.

Sung Young Jung, SungLim Hong, and Eunok Paek. 2000. An English to Korean transliteration model of extended Markov window. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Comput. Surv.*, 43(3), apr.

Sarvnaz Karimi. 2008. Machine transliteration of proper names between english and persian. *RMIT University, Melbourne*.

Mitesh M. Khapra, Ananthakrishnan Ramanathan, Anoop Kunchukuttan, Karthik Visweswariah, and Pushpak Bhattacharyya. 2014. When transliteration met crowdsourcing : An empirical study of transliteration via crowdsourcing using efficient, non-redundant and fair quality control. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Shreya Khare, Ashish R. Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529–1533.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA, March. Association for Machine Translation in the Americas.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Ngoc Tan Le, Fatiha Sadat, Lucie Menard, and Dien Dinh. 2019. Low-resource machine transliteration using recurrent neural networks. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(2), jan.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 159–166, Barcelona, Spain, July.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Ruchi Khapra, Vivek Seshadri, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Aksharantar: Towards building open transliteration tools for the next billion users.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4), may.

M G Abbas Malik, Christian Boitet, and Pushpak Bhattacharyya. 2008. Hindi Urdu machine transliteration using finite-state transducers. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 537–544, Manchester, UK, August. Coling 2008 Organizing Committee.

Sabina Mammadzada. 2021. A review of existing transliteration approaches and methods. *International Journal of Multilingualism*, 0(0):1–15.

Joyce Man. 2012. Hong kong loves weird english names. https://www.theatlantic.com/international/archive/2012/10/hong-kong-loves-weird-english-names/263103/.

(c) 2022 中国中文信息学会计算语言学专业委员会

Yuval Merhav and Stephen Ash. 2018. Design challenges in named entity transliteration. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 630–640, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Diego Moussallem, René Speck, and Axel-Cyrille Ngonga Ngomo. 2020. Generating explanations in natural language from knowledge graphs. *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, 47:213.

Oman Muscat. 2011. The english transliteration of place names in oman. *Journal of Academic and Applied Studies*, 1(3):1–27.

Saeed Najafi, Bradley Hauer, Rashed Rubby Riyadh, Leyuan Yu, and Grzegorz Kondrak. 2018. Comparison of assorted models for transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 84–88, Melbourne, Australia, July. Association for Computational Linguistics.

Saeed Najafi, Colin Cherry, and Grzegorz Kondrak. 2019. Efficient sequence labeling with actor-critic training. In *Canadian Conference on Artificial Intelligence*, pages 466–471. Springer.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Adam St Arnaud, Ying Xu, Lei Yao, and Grzegorz Kondrak. 2015. Multiple system combination for transliteration. In *Proceedings of the Fifth Named Entity Workshop*, pages 72–77, Beijing, China, July. Association for Computational Linguistics.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May. European Language Resources Association (ELRA).

Jong-Hoon Oh and Key-Sun Choi. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jong-Hoon Oh and Key-Sun Choi. 2005. An ensemble of grapheme and phoneme for machine transliteration. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Jong-Hoon Oh and Hitoshi Isahara. 2007. Machine transliteration using multiple transliteration engines and hypothesis re-ranking. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark, September 10-14.

Dinesh Kumar Prabhakar and Sukomal Pal. 2018. Machine transliteration and transliterated text retrieval: a survey. *Sadhana (Bangalore)*, 43(6):1–25.

Taraka Rama and Karthik Gali. 2009. Modeling machine transliteration as a phrase based statistical machine translation problem. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 124–127, Suntec, Singapore, August. Association for Computational Linguistics.

Mihaela Rosca and Thomas Breuel. 2016. Sequence-to-sequence neural network models for transliteration.

Rishiraj Saha Roy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. 2013. Overview of the fire 2013 track on transliterated search. In *Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation*, New York, NY, USA. Association for Computing Machinery.

Rui Santos, Patricia Murrieta-Flores, Pável Calado, and Bruno Martins. 2018. Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, 32(2):324–348.

Brendan Shillingford and Oiwi Parker Jones. 2018. Recovering missing characters in old Hawaiian writing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4929–4934, Brussels, Belgium, October-November. Association for Computational Linguistics.

Yingying Su, 2022. *The Influence of Ancient Chinese Cultural Classics in Southeast Asia*, pages 37–58. Springer Singapore, Singapore.

Harshit Surana and Anil Kumar Singh. 2008. A more discerning and adaptable multilingual transliteration mechanism for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

(c) 2022 中国中文信息学会计算语言学专业委员会

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Jonne Sälevä and Constantine Lignos. 2022. Paranames: A massively multilingual entity name corpus.

Yulia Tsvetkov. 2017. Opportunities and challenges in working with low-resource languages. https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2008. Mining named entity transliteration equivalents from comparable corpora. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 1423–1424, New York, NY, USA. Association for Computing Machinery.

Shyam Upadhyay, Jordan Kodner, and Dan Roth. 2018. Bootstrapping transliteration with constrained discovery for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 501–511, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jean-Claude Usunier and Janet Shaner. 2002. Using linguistics for creating better international brand names. *Journal of Marketing Communications*, 8(4):211–228.

Yu-Chun Wang, Chun-Kai Wu, and Richard Tzong-Han Tsai. 2015. NCU IISR English-Korean and English-Chinese named entity transliteration using different grapheme segmentation approaches. In *Proceedings of the Fifth Named Entity Workshop*, pages 83–87, Beijing, China, July. Association for Computational Linguistics.

GAO Wei. 2004. Phoneme-based statistical transliteration of foreign names for oov problem. *Master's Thesis, The Chinese University of Hong Kong*.

Chun-Kai Wu, Yu-Chun Wang, and Richard Tzong-Han Tsai. 2012. English-Korean named entity transliteration using substring alignment and re-ranking methods. In *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*, pages 57–60, Jeju, Korea, July. Association for Computational Linguistics.

Winston Wu, Nidhi Vyas, and David Yarowsky. 2018. Creating a translation matrix of the Bible's names across 591 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Chun-Kai Wu, Chao-Chuang Shih, Yu-Chun Wang, and Richard Tzong-Han Tsai. 2022. Improving low-resource machine transliteration by using 3-way transfer learning. *Computer Speech & Language*, 72:101283.

Huang Xing and Xu Feng. 2016. The romanization of chinese language. アジア太平洋研究= *Review of Asian and Pacific studies*, (41):99–111.

Andreas Endrique Perez Zepedda. 2020. Procedure of translation, transliteration and transcription. *Applied Translation*, 14(2):8–13, Jun.

Bing Zhao, Nguyen Bach, Ian Lane, and Stephan Vogel. 2007. A log-linear block transliteration model based on bi-stream HMMs. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 364–371, Rochester, New York, April. Association for Computational Linguistics.

冯志伟. 2012. 转写和译音是两个不同的概念. 中国科技术语, 14(5):32–34, 1.

CCL 2022

面向 Transformer 模型的蒙古语语音识别词特征编码方法

张晓旭 [1]**, 马志强** [1,2*]**, 刘志强** [1]**, 宝财吉拉呼** [1]

[1] 内蒙古工业大学, 呼和浩特, 010000

[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, 010000

mzq_bim@imut.edu.cn

摘要

针对 Transformer 模型在蒙古语语音识别任务中无法学习到带有控制符的蒙古语词和语音之间的对应关系，造成模型对蒙古语的不适应问题。提出一种面向 Transformer 模型的蒙古语词编码方法，方法使用蒙古语字母特征与词特征进行混合编码，通过结合蒙古语字母信息使 Transformer 模型能够区分带有控制符的蒙古语词，学习到蒙古语词与语音之间的对应关系。在 IMUT-MC 数据集上，构建 Transformer 模型并进行了词特征编码方法的消融实验和对比实验。消融实验结果表明，词特征编码方法在 HWER、WER、SER 上分别降低了 23.4%、6.9%、2.6%；对比实验结果表明，词特征编码方法领先于所有方法，HWER 和 WER 分别达到 11.8%、19.8%。

关键词： 蒙古语语音识别；Transformer；注意力机制；词编码

Researching of the Mongolian word encoding method based on Transformer Mongolian speech recognition

Zhang Xiaoxu[1]**, Ma Zhiqiang**[1,2*]**, Liu Zhiqiang**[1]**, Bao Caijilahu**[1]

[1] 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

In view of the fact that the Transformer model cannot learn the correspondence between Mongolian words with control symbols and the speech in the Mongolian speech recognition task, which causes the model to not adapt to the Mongolian language. A Mongolian word encoding method for Transformer model is proposed. The method uses Mongolian letter features and word features for mixed encoding. By combining Mongolian letter information, the Transformer model can distinguish Mongolian words

©2022 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

基金项目：国家自然科学基金 (61762070,62166029)，内蒙古自然科学基金 (2019MS06004)

with control symbols, and learn Mongolian words and pronunciation. Correspondence between. On the IMUT-MC dataset, the Transformer model is con-structed and the ablation and comparison experiments of word feature encoding methods are carried out. The results of ablation exper-iments show that the word feature encoding method reduces HWER, WER, and SER by 23.4%, 6.9%, and 2.6%, respectively; the comparative experimental results show that the word feature encoding method is ahead of all methods, with HWER and WER reaching 11.8%, 19.8%.

Keywords: Mongolian speech recognition , Transformer , Attention , Word encoding

1 引言

蒙古语语音识别（Mongolian Speech Recognition, MSR）作为少数民族语言信息处理技术之一，是将蒙古语语音序列转换成蒙古语文本序列的过程。蒙古语作为一种黏着语，是一种表音文字语言 (清格尔泰, 1991)。蒙古语主要通过增加格后缀丰富词的表示，实现语法功能 (高莲花, 2021)。蒙古语格后缀一般指虚词，不能单独出现在句子中充当主、谓、宾、定和状语等成分，但具有强大的组合能力。例如，当格后缀与名词组合可以让名词有不同的意思和读音，类似于汉语的“的”、“和”、“以”等介词。蒙古语格后缀的复杂性体现在格后缀出现在词根后的不同位置会有不同的书写方式和读音 (莫日根, 2016)。在计算机表示中，蒙古语词采用字母拼接表示，通过空格进行词与词的划分，而对于词加格后缀是通过控制符来进行连接并控制其外形。该控制符与蒙古语空格的书面表示形式相似，而在计算机中采用不同的编码。例如在一个名词 ᠭᠡᠷ（没有）后接不同的格后缀构成相同语义的词，目的是能适应不同句子的语法要求，具体见表1所示。

表 1. ᠭᠡᠷ 接不同格后缀的词

格后缀的词	词根	格后缀
ᠭᠡᠷ ᠦᠨ	ᠭᠡᠷ	ᠦᠨ
ᠭᠡᠷ ᠢᠶᠡᠨ	ᠭᠡᠷ	ᠢᠶᠡᠨ
ᠭᠡᠷ ᠲᠡᠭᠡ	ᠭᠡᠷ	ᠲᠡᠭᠡ
ᠭᠡᠷ ᠳᠡ	ᠭᠡᠷ	ᠳᠡ
ᠭᠡᠷ ᠲᠡ	ᠭᠡᠷ	ᠲᠡ

对于蒙古语语音识别模型建模研究主要经历了三个阶段。首先，蒙古语语音识别任务的研究是将隐马尔可夫-高斯混合模型（Gaussian of Mixture Hidden Markov Model，GMM-HMM）作为建立蒙古语语音识别系统的研究起点。(Guanglai Gao, 2006) 等人首次将 GMM-HMM 技术引入蒙古语语音识别任务，并构建基于 GMM-HMM 的蒙古语语音识别系统。(Qilao Hasi, 2008) 等人通过对声学模型的优化提高蒙古语语音识别模型的性能。(Feilong Bao, 2013) 等人通过使用 GMM-HMM 对声学模型进一步设计，从而提高蒙古语语音识别模型的性能。其次，随着深度神经网络（Deep Neural Network，DNN）的出现，深度神经网络结合隐马尔科夫模型（Deep Neural Network of Mixture Hidden Markov Model，DNN-HMM）的组合方式也在蒙古语语音识别模型中开展研究和使用。(Hui Zhang, 2015) 等人在基于传统的混合蒙古语语音识别模型的研究中，引入基于 DNN 的声学模型，使蒙古语语音识别系统获得了显著的性能提升。(马志强,

(c) 2022 中国中文信息学会计算语言学专业委员会

2018) 等人针对 HMM 在蒙古语语音识别声学模型中不能充分学习声学特征之间相关性和独立性假设的问题，使用深度神经网络对声学模型建模进行研究，并且取得了良好的识别性能，用其构建的蒙古语语音识别系统在工业界得到了应用，以上基于 HMM 蒙古语语音识别模型属于传统的混合语音识别模型。但是传统的混合语音识别模型的独立训练与联合识别的特性导致模型参数无法达到全局最优，同时构建的复杂性也给蒙古语语音识别工业化应用带来了困难。最后，为了降低蒙古语语音识别模型构建和训练的复杂度，研究人员用端到端神经网络模型的学习过程代替工程过程，剔除了发音词典部分，降低了蒙古语语音识别模型工业化应用的门槛，使端到端蒙古语语音识别模型成为工业化应用的研究热点。但是在进行端到端蒙古语语音识别模型建模时，模型不能正确识别一些带有控制符的蒙古语词，导致蒙古语语音的识别结果与蒙古语词目标文本不一致的情况，造成模型对蒙古语的不适应性问题。因此如何使模型学习到蒙古语语音和蒙古语词的对应关系是一个重要的问题。

本文在蒙古语构词特点的基础上，基于 Transformer 模型提出一种蒙古语字母与词混合编码方法。主要贡献为：

- 基于注意力机制提出了一种面向蒙古语语音识别的蒙古语词编码方法；
- 把蒙古语字母特征与词特征进行混合编码，构建了一个基于 Transformer 的端到端蒙古语语音识别模型；
- 在 IMUT-MC 数据集上进行了验证实验，实验结果显示提出的蒙古语字母与词混合编码方法在性能上相较基线方法有显著提升。

2 相关工作

端到端语音识别模型一般选用适合语言特点的建模单元进行编码。比如英语选用子词作为建模单元进行编码，而汉语选用以字为建模单元进行编码。对于端到端蒙古语语音识别模型来说，蒙古语的构词是以词根、词干上连接不同的词缀构成，因此针对于蒙古语形态复杂多变的情况，模型一般选用蒙古语词作为建模单元进行编码。

基于蒙古语词的切分构成建模单元进行编码研究，(赵伟, 2010) 等人为了利用蒙古语语法信息，通过分析蒙古语词的构形特点，划分出词干和词的构形附加成分，提出一种有效的蒙古语词标注方法。实验表明该方法的词切分准确率比现有蒙古语词切分系统的准确率有明显的提高。(杨振新, 2017) 等人针对蒙汉统计机器翻译面临的形态差异大的问题，通过词素和短语两个层面编码信息的结合，实现了汉语与蒙古语语言在形态结构上的对称，通过实验结果表明该方法在 BLEU 上，比基线模型有了明显的提高，在一定程度上消解了形态差异对汉蒙统计机器翻译的影响。(Qing-Dao-Er-Ji Ren, 2020) 等人对蒙古语双语料库预处理阶段蒙古语的词干和词缀进行分割和标记。通过实验结果表明，使用该方法的端到端蒙汉神经机器翻译模型与传统统计方法和基于递归神经网络的机器翻译模型相比，在翻译质量和语义混淆方面有所提高。

基于蒙古语词特征编码研究。(包乌格德勒, 2018) 等人为探究词特征输入对蒙汉翻译系统的影响，分别采用蒙古语的词模型、切分模型和子词模型作为输入词特征。通过对比实验结果表明，子词模型在基于 CNN 和 RNN 的神经蒙汉翻译模型中可以有效地提高翻译质量。(曹宜超, 2020) 等人针对蒙古语形态复杂多变的问题，提出了一种结合词向量编码对齐的蒙汉神经机器翻译方法。通过实验结果表明，该方法的翻译效果高于基线模型。(卞乐乐, 2021) 等人为探究词向

(c) 2022 中国中文信息学会计算语言学专业委员会

量的质量是否影响模型的质量问题，使用三种不同的词向量生成模型对蒙古语单语数据进行词向量的生成，对不同模型生成的蒙古语单语词向量对翻译模型质量的影响进行实验。通过实验表明，将大量蒙古语单语词向量嵌入蒙汉翻译模型能够使模型效果得到显著提升。

在蒙古语语音识别任务中，由于蒙古语在构词中添加了控制符来改变词的外形，使得模型无法区分相同读音的蒙古语词。因此，本文基于蒙汉机器翻译中蒙古语构词特点，提出了蒙古语字母与词混合编码单元，使 Transformer 语音识别模型适应蒙古语的读音和构词特点，解决了模型对蒙古语的不适应问题，从而提升模型的识别准确率。

3 蒙古语词编码方法

3.1 问题描述

在蒙古语词表 $C = \{y_1, \ldots, y_i, \ldots, y_j, \ldots, y_m\}$ 中，$\exists y_i, y_j \in C$ 且 $y_i != y_j$，使 y_i 与 y_j 的读音相同，即 $g(y_i) = g(y_j)$ ，其中 $g(\cdot)$ 表示读音函数。在基于 Transformer 模型进行蒙古语语音识别建模时，首先输入一段蒙古语语音序列 $X = \{x_1, \ldots, x_t, \ldots, x_T\}$，通过编码器 $Encoder(\cdot)$ 得到语音特征序列 $\mathbf{h}^{enc}$。然后输入目标序列开始标签 $Y = \{sos\}$，进行词嵌入 $embedding(\cdot)$ 得到解码器输入序列 $\mathbf{h}_{Y_{sos}}$，通过解码器 $Decoder(\cdot)$ 预测 $y_l|(1 \le l \le L)$。最后目标序列 $Y = \{y_1, \ldots, y_L\}$ 经过 L 步解码得到。在解码器 $Decoder(\cdot)$ 预测 $y_l|(1 \le l \le L)$ 的过程中会出现不能正确区分词表中带有控制符的蒙古语词的情况，导致 Transformer 模型对蒙古语的不适应问题。

3.2 Transformer 模型架构

基于 Transformer 的蒙古语语音识别模型采用编码-解码器模型结构，比其他基于注意力机制的端到端语音识别模型结构更加复杂，其编码器和解码器都是由堆叠式自注意力层和全连接层构成，模型结构如图1所示。

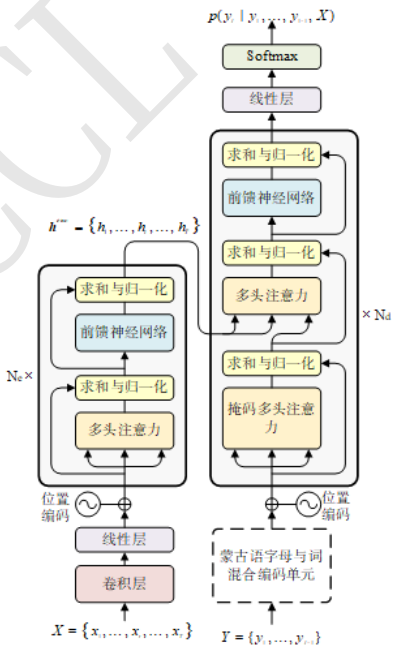


图 1. 基于 Transformer 的端到端蒙古语语音识别模型架构

基于 Transformer 的蒙古语语音识别模型中的编码器将输入的语音序列 X =

$\{x_1,\ldots,x_t,\ldots,x_T\}$ 映射成为特征序列 $\mathbf{h}^{enc}=\{h_1,\ldots,h_t,\ldots,h_T\}$，然后解码器将编码器生成的特征序列 $\mathbf{h}^{enc}$ 映射成为文本序列 $Y=\{y_1,\ldots,y_l,\ldots,y_L\}$。在基于 Transformer 的端到端语音识别模型上，本文提出端到端蒙古语语音识别模型架构，其总体计算公式如下：

$$P_{att}(Y|X)=\prod_{l=1}^{L}P_{att}(y_l|y_1,\ldots,y_{l-1}) \tag{1}$$

在蒙古语语音识别任务中，首先蒙古语字母在词中位置发生一系列的变化，导致模型在进行识别时发生错误。为解决该问题，本模型考虑将蒙古语词特征序列描述为具有词特征和字母特征的混合特征序列。在蒙古语字母与词混合编码单元结构中利用注意力机制将根据字母特征生成新词特征，通过结合原有的词特征来增加蒙古语词特征的区分度，使模型的识别准确率得到提升。

3.3 蒙古语字母与词混合编码方法

与基于注意力机制的序列到序列模型一样，基于 Transformer 的端到端语音识别模型也是编码-解码器架构，Transformer 模型结构中的编码器和解码器是由堆叠式自注意力层和全连接层构建而成。本节主要详细介绍蒙古语字母与词混合编码单元的详细结构，主要是为基于 Transformer 的端到端语音识别模型适应蒙古语，从而提高模型的识别准确率。

基于 Transformer 的端到端语音识别模型使用的是词嵌入的方法，对于蒙古语中书写方式相似的词来说，词嵌入编码特征具有相似性，如  和 。因此，本章提出一种结合字母特征的蒙古语词编码单元，如图2所示。

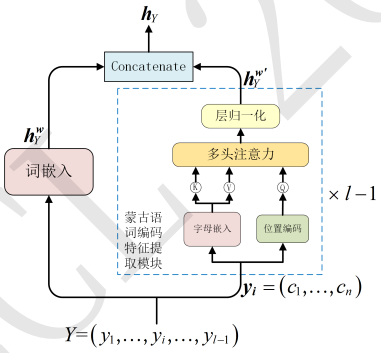


图 2. 蒙古语字母与词混合编码单元

首先，输入文本标签序列 $Y=\{y_1,\ldots,y_i,\ldots,y_{l-1}\}$ 在词级序列 Y 和字母级序列 y_i 上执行嵌入编码，获得词嵌入编码特征 $\mathbf{h}_Y^w$ 和字母嵌入编码特征 $\mathbf{h}_{y_i'}^{w'}$ 。然后，将获得 $\mathbf{h}_{y_i'}^{w'}$ 和字母位置编码特征 $\mathbf{pe}_{y_i'}^{w'}$ 通过注意力机制生成新的编码特征 $\mathbf{h}_Y^{w'}$。最后，将 $\mathbf{h}_Y^w$ 和 $\mathbf{h}_Y^{w'}$ 进行拼接，得到最终解码器输入特征 $\mathbf{h}_Y$。加入字母信息后，明显区分蒙古语词编码特征。注意力机制不仅用于提取单词级上下文信息，还用于对单词中的字母级信息进行编码。对蒙古语字母和词混合编码单元的功能进行分析，其计算公式如下：

$$\mathbf{h}_Y^w=embedding(y_1,\ldots,y_{l-1}) \tag{2}$$

$$\mathbf{h}_{y_i'}^{w'}=embedding(c_1,\ldots,c_n) \tag{3}$$

$$\mathbf{pe}_{y_i'}^{w'}=PE_{(pos,j)} \tag{4}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

$$\mathbf{h}_{y_i}^{w'} = attention(\mathbf{h}_{y_i'}^{w'}, \mathbf{h}_{y_i'}^{w'}, \mathbf{pe}_{y_i'}^{w'}) \tag{5}$$

$$\mathbf{h}_Y = \mathbf{h}_Y^w \oplus \mathbf{h}_{y_i}^{w'} \tag{6}$$

$$p(y_l|y_1, \ldots, y_{l-1}) = Decoder(\mathbf{h}^{enc}, \mathbf{h}_Y) \tag{7}$$

4 实验

4.1 实验设置

4.1.1 实验数据

本实验使用的语料库为 IMUT-MC，由 (刘志强, 2021) 等人构建的一个针对蒙古语语音识别任务的语音语料库。其中，具体选用语料库中 IMUT-MC2、IMUT-MC3 和 IMUT-MC4 数据集进行实验。IMUT-MC2 和 IMUT-MC3 的数据来自于蒙古语的日常对话，文本语句都比较简短。IMUT-MC4 的数据来源于蒙古语新闻，包括环境、教育、经济、时政和体育等领域，文本语句较长。IMUT-MC 的基本信息如表2所示。

表 2. 语料库 IMUT-MC 基本信息

数据集	时长	总句数	总词数	平均词数	说话人个数	来源
IMUT-MC2	23.5h	19800	970	10	99	
IMUT-MC3	40.8h	22200	1307	10	111	人员录音
IMUT-MC4	69.7h	20000	6591	22	100	
总计	134h	62000	8868	13	310	

将 IMUT-MC 数据集，按照当前大部分研究神经网络训练的 8:1:1 比例进行划分训练集、验证集和测试集。训练集用来使模型进行学习；验证集用来确定神经网络结构或控制模型复杂程度的参数；测试集用来检验最终选择最优的模型的性能。

4.1.2 实验环境

本实验平台使用两个高性能计算机，包括个人工作站和深度学习服务器。在硬件方面，个人工作站的环境配置包括 Intel i7-9700 CPU、NVIDIA RTX-2060 GPU；深度学习服务器的环境配置包括 Intel Xeon 6130 CPU、NVIDIA Tesla P100 GPU。在软件方面，使用 Ubuntu 操作系统，安装 Python 环境和 PyTorch 深度学习框架，搭建 Kaldi 和 Espnet 语音识别平台，支持 GPU 并行计算。所有模型的实验工作均在上述计算设备上开展。

4.1.3 模型参数

基于 Transformer 的端到端蒙古语语音识别模型采用的超参数设计如下：编码器层数为 12，编码器单元为 2048；解码器层数为 6，解码器单元为 2048；注意力机制的维度为 256，注意力头数为 4；丢弃率为 0.1，使用 Adam 优化算法，学习率为 1.0；批处理大小为 16，迭代训练 20 轮得到最终的模型；模型的特征处理部分采用 1 层卷积层，其中卷积核大小为 3，步移为 2，下采样率为 4；字母与词混合编码单元中的多头注意力机制的维度为 256，注意力头数为 4。

4.2 评价指标

蒙古语语音识别模型的评价指标包括：HWER（Homophone Word Error Rate，同音词错误率）、WER（Word Error Rate，词错误率）、SER（Sentence Error Rate，句错率）。各评价指标的含义如下：

(c) 2022 中国中文信息学会计算语言学专业委员会

（1）HWER 是指所有错误同音词的和所占总同音词数的百分比，其公式为：

$$HWER = \frac{S_h + D_h + I_h}{N_h} * 100\% \tag{8}$$

S_h 为替换错误的同音词数，D_h 为删除错误的同音词数，I_h 为插入错误的同音词数，N_h 表示数据集中同音词总词数。

（2）WER 是指所有错误词的和所占总词数的百分比，其公式为：

$$WER = \frac{S_w + D_w + I_w}{N_w} * 100\% \tag{9}$$

S_w 为替换错误的词数，D_w 为删除错误的词数，I_w 为插入错误的词数，为数据集中的总词数，其中，为数据集中非同音词的总词数。

(3) SER 是指所有识别结果与对应文本不能正确匹配的测试音频所占总音频数的百分比，其公式为：

$$SER = \frac{N_error}{N_s} * 100\% \tag{10}$$

N_error 为识别错误的蒙古语音频的个数，N_s 为数据集中蒙古语音频的总个数。

4.3 实验结果与分析

4.3.1 语种对比实验

基于 Transformer 语音识别模型在英语、汉语和蒙古语上的识别结果，如表3所示。

表 3. 不同语种在 Transformer 语音识别模型上的识别结果

语言	数据集	总时间	总词数	平均时间	平均词数	WER
汉语	Aishell-1	178h	4229	4.45s	15	9.7
英语	LibriSpeech(train-clean-100)	100h	6967	12.8s	20	9.8
蒙古语	IMUT-MC	134h	8868	7.8s	13	26.7

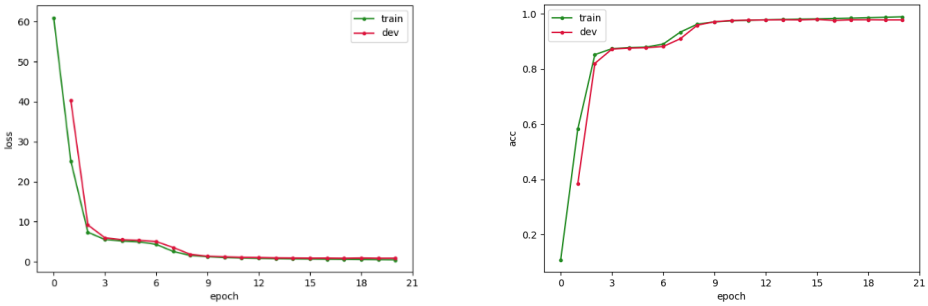
从实验结果可以看出，基于注意力机制的 Transformer 语音识别模型在蒙古语上的识别效果远不如汉语和英语的识别效果。对模型转录的蒙古语语音数据结果进行分析，大部分识别错误的词是组合词，即具有词根和词缀的蒙古语词。因此对蒙古语构词特点进行分析，蒙古语的格后缀与名词结合时，会组成新蒙古语词。在组成时，随着位置不同书写形式也不相同，会形成一些异形同音词等组合词，模型对这些组合词的词特征区分度不高。

4.3.2 收敛性实验

训练过程在 IMUT-MC 数据集上对基于 Transformer 的端到端蒙古语语音识别模型开展，收敛情况如图3所示。为了保证基于 Transformer 的端到端蒙古语语音识别模型的收敛效果，实验采用训练集和验证集的损失值和准确率来验证模型收敛。

由图3 a）中可知，基于 Transformer 的端到端蒙古语语音识别模型训练集和验证集上的损失函数呈下降趋势，并于 9 轮左右使 Loss 趋向于零且处于平缓状态，表明模型能够收敛。由图3 b）可知，在训练集和验证集上的准确率不断上升，呈现出两次先陡后缓的趋势，最终在 9 轮左右趋于平稳缓和且验证集的准确率高于 90%，可以认为模型已经充分收敛，学习到了训练集的数据分布特征。

第二十一届中国计算语言学大会论文集，第333页-第343页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会



(a) 训练的 loss 值的变化　　(b) 训练过程中识别准确率

图 3. 基于注意力的蒙古语语音识别模型的收敛情况

4.3.3 蒙古语字母与词混合编码单元消融实验

为了说明在蒙古语语音识别任务中，基于 Transformer 语音识别模型在增加蒙古语字母与词混合编码单元的有效性，对蒙古语字母与词混合编码单元和蒙古语词编码特征提取模块中的位置编码进行消融实验。

表 4. 编码方法消融实验结果

编码方法	位置编码	HWER	WER	SER
+ 词嵌入	-	35.2	26.7	36.4
+ 字母特征提取	-	35.6	28.8	37.4
	绝对位置编码	23.4	24.9	35.7
	相对位置编码	24.5	25.5	36.1
+ 词嵌入 + 字母特征提取	-	13.4	23.8	34.3
	绝对位置编码	11.8	19.8	33.8
	相对位置编码	11.5	20.7	33.1

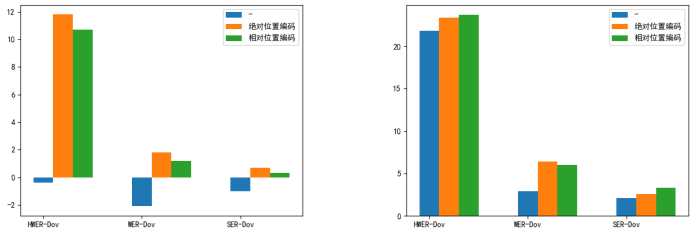
根据表4的实验结果数据分析可知，本章在基于 Transformer 的端到端语音识别模型上进行蒙古语字母与词混合编码单元的消融实验，验证了单元可以使模型获得更好的识别准确率，表明蒙古语字母与词混合编码单元能够使基于 Transformer 语音识别模型适应蒙古语的特点。并且通过探索不同位置编码对蒙古语词编码特征提取模块的影响，证明了绝对位置编码能使基于 Transformer 语音识别模型的识别准确率达到最佳。

为了更加直观地突出蒙古语字母与词混合编码单元对基于 Transformer 语音识别模型的影响，定义评价指标的下降值（Drop-out Value，Dov），见公式 (11)，用于描述基线评价指标 Pa 与增加单元结构的评价指标 Pb 之间的差距。正值表示提高了基线实验的评价指标值，负值表示降低了基线实验的评价指标值。

$$Dov = Pa - Pb \tag{11}$$

将表4中数据使用 Dov 指标对使用的编码特征进行比较，其中 HWER-Dov、WER-Dov 和 SER-Dov 分别表示同音词错误率下降值、词错误率下降值和句错率下降值。

从图4中可以看出，蒙古语编码特征使用字母与词的混合特征，三个评价指标没有负值产生，因此本章提出的蒙古语字母与词混合编码单元是有效地。尽管位置编码对于蒙古语字母与词混



(a) 使用字母特征　　(b) 使用字母与词的混合特征

图 4. 消融实验的 HWER、WER 和 SER 的下降值

合编码单元中注意力机制是有效地，但是绝对位置编码和相对位置编码的方式对当前的数据集的影响并不明显。通过对当前蒙古语数据集进行分析，得出不明显的原因是：数据集中的蒙古语词的长度可以使注意力机制获取的上下文关系，绝对位置编码信息可以满足当前的蒙古语数据集，后期当蒙古语数据集增加时，可以继续探究相对位置编码对该单元的影响。

4.3.4　蒙古语输入文本特征编码对比实验

基于 Transformer 的端到端蒙古语语音识别模型的输入文本特征编码实验旨在探索基于 Transformer 语音识别模型对于蒙古语语音识别任务的适应性，主要是从字母、子词、词、词根与词缀划分和字母与词相结合的特征编码上进行对比实验，结果如表5所示。

表 5. 蒙古语输入特征编码对比实验

输入特征编码	HWER	WER	SER
字母特征编码	40.3	33.6	41.6
子词特征编码	18.3	24.2	35.2
词特征编码	35.2	26.7	36.4
词根与词缀特征编码	16.5	22.3	34.1
字母与词特征编码	11.8	19.8	33.8

根据表5的实验结果数据分析可知，基于 Transformer 的端到端语音识别模型在蒙古语语音识别任务中，以字母特征编码的识别效果不佳；以子词特征编码可以使基于 Transformer 的端到端语音识别模型具有良好的识别效果，但是子词是基于频率进行统计获得，对语料具有很强的依赖性；以词特征编码的识别效果也有所提升，但是也具有对语料的依赖性；以词根与词缀特征编码作为解码器输入，可以解决模型对蒙古语不适应性的问题，但是模型在识别时，会出现一些不正确的组词的错误；以字母与词特征编码作为解码器输入，可解决模型对蒙古语不适应性的问题，并使模型的识别效果达到最优。既可以缓解建模单元对语料的依赖，又可以使模型充分利用两个量级的知识，更适合当前的蒙古语语料数据。

4.3.5　蒙古语端到端语音识别模型对比实验

为了说明在蒙古语语音识别任务中，增加蒙古语字母与词混合编码单元的 Transformer 语音识别模型具有良好的识别性能，因此对多种端到端语音识别模型进行对比实验研究，结果如表6所示。

根据表6的实验结果数据分析可知，增加蒙古语字母与词混合编码单元的 Transformer 语音识别模型在蒙古语语音识别任务上比当前其他端到端语音识别模型具有更好的识别准确率，

表 6. 蒙古语端到端语音识别模型对比实验

模型	HWER	WER	SER
CTC	36.7	30.3	58.2
Transformer	35.2	26.7	36.4
RNN-Transducer	33.6	25.1	44.8
字母与词特征编码 + Transformer	11.8	19.8	33.8

HWER、WER 和 SER 分别达到了 11.8%、19.8% 和 33.8%。因此，蒙古语端到端语音识别模型对比实验验证了增加蒙古语字母与词混合单元的 Transformer 语音识别模型更加适应蒙古语语音识别任务，具有良好的识别性能。

4.4 案例分析

表7展示了蒙古语语音识别的样例，该样例从 IMUT-MC 语料库随机选出。编号 1-4 表示是从 IMUT-MC-TEST1 和 IMUT-MC-TEST2 的测试数据集中选出，主要是针对于日常对话的数据集，其中语音数据的时长集中在 3-5s 之间；编号 5-8 表示是从 IMUT-MC-TEST4 的测试数据集中选出，主要针对环境、教育、经济、时政和体育等领域的数据，其中语音数据的时长集中在 8-10s 之间。数据来源的场景不同，所以蒙古文词的使用也不尽相同。

表 7. 蒙古语测试样例数据识别结果

编号	语音数据编号	标签数据	备注
1	00001365-F-M-19-13.wav	ᠬᠦᠮᠦᠨ ᠬᠦᠷᠢᠶᠡᠨ ᠬᠡᠪ᠎ᠡ ᠪᠠᠢᠭᠤᠯᠭ᠎ᠠ ᠬᠡᠪᠦᠦᠨ ᠵᠤᠭᠰᠠᠭᠠᠵᠤ	让（他）走进内蒙古地区的旗县
2	00000118-F-W-20-03.wav	ᠴᠠ ᠵᠤᠭᠰᠤᠭᠠᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪ᠎ᠡ ᠬᠡᠯᠡᠵᠤ ᠪᠤᠢ᠎ᠠ ᠤᠤ	您不稍微往前站吗？我想从这里抓住
3	00000128-D-M-20-09.wav	ᠪᠢ ᠮᠡᠳᠡᠬᠦ ᠬᠡᠷᠡᠭ ᠴᠠ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪᠠᠰᠤᠭᠠᠷᠠᠢ ᠶᠠᠪᠤᠭᠠᠷᠠᠢ	我不太懂，您向别人问吧
4	00001228-F-W-20-03.wav	ᠴᠠ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ	您想喝点啥，您想吃点啥
5	00001307-F-W-22-39.wav	ᠡᠨᠡᠨ᠎ᠡ ᠬᠠᠭᠠᠨᠠᠴᠠ᠎ᠠ ᠬᠤᠭᠤᠤᠨᠤᠴᠠ᠎ᠠ ᠬᠡᠪᠡᠷ ᠪᠢ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪ ᠬᠡᠯᠡᠵᠤ᠎ᠡ ᠨ ᠠᠯᠳᠠᠭᠳᠠᠵᠤ	不知哪一天，从他（她，它）那里得到事情的真理
6	00000075-H-W-23-63.wav	ᠪᠢ᠎ᠡ ᠭᠡ ᠬᠥᠬᠡ ᠬᠠᠭᠠ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠤ᠎ᠡ ᠰᠤᠷᠤᠭ᠎ᠠ ᠪᠠᠭᠰᠢᠮᠠᠯ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠨ᠎ᠤ ᠭᠠ᠎ᠡ ᠵᠥᠪᠬᠡᠨ᠎ᠡ ᠪᠡᠷᠬᠦ᠎ᠡ ᠴᠠ᠎ᠠ ᠵᠠᠮᠴᠢᠨ᠎ᠤ ᠬᠡᠷᠡᠭᠲᠡᠢ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪᠡᠷ ᠪᠢ	我们有蓝，黄，绿，粉和红色的袍，您需要什么颜色的
7	00002424-D-W-20-90.wav	ᠡᠨᠡᠬᠦ ᠪᠢ ᠪᠤᠳᠤᠯᠠᠭᠠᠨ ᠪᠢ ᠴᠢᠵᠢᠶᠠᠨ ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠨᠢ ᠪᠠᠭᠠ ᠪᠠᠢᠵᠤ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ	所以，以说教为主的家庭教育比普通的教育效率低
8	00001373-F-M-19-13.wav	ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢ ᠲᠠᠨ᠎ᠤ ᠬᠡᠪᠡᠯᠡᠯ᠎ᠡ ᠵᠢᠷᠤᠭᠯᠠᠩ᠎ᠤ ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢᠶᠠᠨ ᠲᠤ᠎ᠡ ᠬᠡᠭᠡᠨ᠎ᠡ ᠭᠡ ᠤᠭᠰ ᠭᠡ	难道这叫做让世间的所有生活的资格吗

由表8可以看出，基于 Transformer 的端到端蒙古语语音识别模型能够较好地识别蒙古语语音，且具有一定的准确度，对带有控制符的词识别准确率更高。其中具有下划线的词、“***” 和双下划线的词分别表示替换错误词、删除错误词和插入错误词。

表 8. 蒙古语测试样例数据

编号	识别标签结果	备注	替换词错误	删除词	插入词错误
1	ᠬᠦᠮᠦᠨ ᠬᠦᠷᠢᠶᠡᠨ ᠬᠡᠪ᠎ᠡ ᠪᠠᠢᠭᠤᠯᠭ᠎ᠠ ᠬᠡᠪᠦᠦᠨ ᠵᠤᠭᠰᠠᠭᠠᠵᠤ	走进了内蒙古地区的旗县	1	0	0
2	ᠴᠠ ᠵᠤᠭᠰᠤᠭᠠᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪ᠎ᠡ ᠬᠡᠯᠡᠵᠤ ᠪᠤᠢ᠎ᠠ ᠤᠤ	您不稍微往前站吗？我想从这里抓住	0	0	0
3	ᠪᠢ ᠮᠡᠳᠡᠬᠦ ᠬᠡᠷᠡᠭ ᠴᠠ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪᠠᠰᠤᠭᠠᠷᠠᠢ ᠶᠠᠪᠤᠭᠠᠷᠠᠢ	我不太懂，您向别人问吧	0	0	0
4	ᠴᠠ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ	您想喝点啥，您想吃点啥呢	2	0	1
5	ᠡᠨᠡᠨ᠎ᠡ ᠬᠠᠭᠠᠨᠠᠴᠠ᠎ᠠ ᠬᠤᠭᠤᠤᠨᠤᠴᠠ᠎ᠠ ᠬᠡᠪᠡᠷ ᠪᠢ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪ ᠬᠡᠯᠡᠵᠤ᠎ᠡ ᠨ ᠠᠯᠳᠠᠭᠳᠠᠵᠤ	不知哪一天，从他（她，它）那里得到事情的真理	1	0	0
6	ᠪᠢ᠎ᠡ ᠭᠡ ᠬᠥᠬᠡ ᠬᠠᠭᠠ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠤ᠎ᠡ ᠰᠤᠷᠤᠭ᠎ᠠ ᠪᠠᠭᠰᠢᠮᠠᠯ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠨ᠎ᠤ ᠭᠠ᠎ᠡ ᠵᠥᠪᠬᠡᠨ᠎ᠡ ᠪᠡᠷᠬᠦ᠎ᠡ ᠴᠠ᠎ᠠ ᠵᠠᠮᠴᠢᠨ᠎ᠤ ᠬᠡᠷᠡᠭᠲᠡᠢ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪᠡᠷ ᠪᠢ	我们有蓝，黄，绿，粉和红色的袍，您需要什么颜色的	0	0	0
7	ᠡᠨᠡᠬᠦ ᠪᠢ ᠪᠤᠳᠤᠯᠠᠭᠠᠨ ᠪᠢ ᠴᠢᠵᠢᠶᠠᠨ ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠨᠢ ᠪᠠᠭᠠ ᠪᠠᠢᠵᠤ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ	所以，以说教为主的家庭教育比普通的教育效率低	0	0	0
8	ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢ ᠲᠠᠨ᠎ᠤ ᠬᠡᠪᠡᠯᠡᠯ᠎ᠡ ᠵᠢᠷᠤᠭᠯᠠᠩ᠎ᠤ ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢᠶᠠᠨ ᠲᠤ᠎ᠡ ᠬᠡᠭᠡᠨ᠎ᠡ ᠭᠡ ᠤᠭᠰ ᠭᠡ	难道这叫做让世间的所有生活的资格吗	1	0	0

5 结论

针对 Transformer 模型对蒙古语构词特点的不适应性问题，将蒙古语字母特征与词特征进行混合编码，提出一种基于 Transformer 的蒙古语词编码方法，并构建端到端蒙古语语音识别模型。通过字母特征进一步构建词特征能提高 Transformer 模型对带有控制符的蒙古语词识别准确率。实验结果表明，在 IMUT-MC 蒙古语语音识别数据集下，增加蒙古语字母与词混合编码单元的 Transformer 语音识别模型的 WER 降低了 6.9%，表明该单元对 Transformer 模型在蒙古语语音识别任务中具有一定的作用。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

清格尔泰. 1991. 蒙古语语法. 内蒙古人民出版社.

高莲花. 2021. 蒙古语的后置词短语, 民族语言, (05):108–113.

莫日根. 2016. 基于规则的传统蒙古语句法分析研究. 内蒙古大学.

Guanglai Gao, Biligetu, Nabuqing and Shuwu Zhang . 2006. *A Mongolian Speech Recognition System based on HMM*. International Conference on Intelligent Computing, Springer-Verlag, 2006: 667-676.

Qilao Hasi, and Guanglai Gao. 2008. *Researching of Speech Recognition Oriented Mongolian Acoustic Model*. Conference on Pattern Recognition, 2008: 1-6.

Feilong Bao, Guanglai Gao, Xueliang Yan and Weihua Wang. 2013. *Segmentation-based Mongolian LVCSR Approach*. International Conference on Acoustics, Speech, and Signal Processing, 2013: 1-5.

Hui Zhang, Feilong Bao, and Guanglai Gao. 2015. *Mongolian Speech Recognition Based on Deep Neural Networks*. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, 2015: 180-188.

马志强, 李图雅, 杨双涛和张力. 2018. 基于深度神经网络的蒙古语声学模型建模研究, 智能系统学报, 13(03):486–492.

赵伟, 侯宏旭, 从伟和宋美娜. 2010. 基于条件随机场的蒙古语词切分研究, 中文信息学报, 24(05):31–35+84.

杨振新, 李淼, 陈雷, 卫林钰, 陈晟和孙凯. 2017. 一种基于词素媒介的汉蒙统计机器翻译方法, 中文信息学报, 31(04):57–62+69.

Qing-Dao-Er-Ji Ren, Yila Su, and Nier Wu. 2020. *Research on Mongolian-Chinese machine translation based on the end-to-end neural network*, International Journal of Wavelets, Multiresolution and Information Processing, 18(01):1941003.

包乌格德勒和赵小兵. 2018. 基于 *RNN* 和 *CNN* 的蒙汉神经机器翻译研究, 中文信息学报, 32(08):60–67.

曹宜超, 高翊, 李淼, 冯韬, 王儒敬和付莎. 2020. 基于单语语料和词向量对齐的蒙汉神经机器翻译研究, 中文信息学报, 34(2):27–32.

卞乐乐. 2021. 基于单语语料与强化学习的蒙汉神经机器翻译的研究. 内蒙古工业大学.

刘志强, 马志强, 张晓旭, 宝财吉拉呼, 谢秀兰和朱方圆. 2021. *IMUT-MC*: 一个针对蒙古语语音识别的语音语料库, 中国科学数据.

CCL 2022

基于注意力的蒙古语说话人特征提取方法

朱方圆 [1], 马志强 [1,2*], 刘志强 [1], 宝财吉拉呼 [1], 王洪彬 [1]
[1] 内蒙古工业大学, 呼和浩特, 010000
[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, 010000
mzq_bim@imut.edu.cn

摘要

说话人特征提取模型提取到的说话人特征之间区分性低，使得蒙古语声学模型无法学习到区分性信息，导致模型无法适应不同说话人。提出一种基于注意力的说话人自适应方法，方法引入神经图灵机进行自适应，增加记忆模块存放说话人特征，采用注意力机制计算记忆模块中说话人特征与当前语音说话人特征的相似权重矩阵，通过权重矩阵重新组合成说话人特征 s-vector，进而提高说话人特征之间的区分性。在 IMUT-MCT 数据集上，进行说话人特征提取方法的消融实验、模型自适应实验和案例分析。实验结果表明，对比不同说话人特征 s-vector、i-vector 与 d-vector，s-vector 比其他两种方法的 SER 和 WER 分别降低 4.96%、1.08%；在不同的蒙古语声学模型上进行比较，提出的方法相对于基线均有性能提升。

关键词： 说话人特征提取；注意力机制；神经图灵机；说话人自适应；蒙古语语音识别

Attention based Mongolian Speaker Feature Extraction

Zhu Fangyuan[1], Ma Zhiqiang[1,2*], Liu Zhiqiang[1], Bao Caijilahu[1], Wang Hongbin[1]
[1] 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

The speaker features extracted by the speaker feature extraction model have low discrimination, which makes the Mongolian acoustic model unable to learn the discrimination information, resulting in the model unable to adapt to different speakers. A speaker adaptation method based on attention is proposed. The method introduces neural Turing machine for adaptation, adds a memory module to store the speaker features, uses the attention mechanism to calculate the similarity weight matrix between the speaker features in the memory module and the current voice speaker features, and recombines the weight matrix into the speaker features s-vector, so as

第二十一届中国计算语言学大会论文集，第344页-第354页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

to improve the discrimination between the speaker features. On the IMUT-MCT dataset, the ablation experiment, model adaptation experiment and case analysis of speaker feature extraction method are carried out. The experimental results show that comparing s-vector, i-vector and d-vector with different speaker characteristics, s-vector reduces SER and WER by 4.96% and 1.08% respective-ly compared with the other two methods; By comparing different Mongolian acoustic models, the performance of the proposed method is improved compared with the baseline.

Keywords: Speaker Feature Extraction , Attention Mechanism , Neural Turing Machine , Speaker Adaptation , Mongolian speech recognition

1 引言

蒙古语说话人自适应 (Mongolian Speaker Adaptation, MSA) 方法是解决训练数据和测试数据中的说话人不匹配问题。根据自适应的对象是特征还是模型，将说话人自适应方法分为基于模型域的和基于特征域的说话人自适应方法 (Bell et al., 2020)。在基于模型域的说话人自适应方法中，Stadermann and Rigoll (2005) 提出基于模型参数的说话人自适应方法，是只更新声学模型的部分参数，这些参数通常兼具鲁棒性和有效性。Samarakoon and Sim (2016) 提出对隐藏层进行分解，分解后的奇异值分解层插入线性层。利用奇异值分解方法对模型权重矩阵进行更新，比直接插入线性层的方法更能减少参数量，减轻过拟合问题。Swietojanski and Renals (2014) 将给定自适应数据的情况下学习说话人特定的隐藏单元分布，为解决在只对神经网络的某一层输出特征进行变换时，对所有层的输出特征均进行变换导致自适应参数数量成倍增加的问题。Dong et al. (2013) 在声学模型自适应过程中增加正则化项，通过限制原始模型和调整后的模型之间的距离，进而防止调整后的模型参数偏离原始模型参数。根据使用方法的不同将基于特征域的说话人自适应方法分为基于特征变换和基于辅助特征的说话人自适应 (朱方圆等, 2021)。在基于特征变换的说话人自适应方法中，Neto et al. (1995) 提出对输入特征进行线性变换的方法，通过对神经网络的输入特征或者隐藏层特征进行变换来实现自适应。在基于辅助特征的说话人自适应方法中，Saon et al. (2013) 通过在声学特征中添加特定说话人信息并利用新特征训练声学模型实现自适应。Abdel-Hamid and Jiang (2013) 提出说话人编码方法，即给定一个说话人编码，将每一个说话人的特征映射到一个说话人无关的特征空间，对于一个新的说话人学习其相应的说话人编码是容易的，且用反向算法时不会改变神经网络的模型参数。其中基于辅助特征的说话人自适应方法是极为重要的方法，辅助特征包括 i-vector (Saon et al., 2013)、d-vector (Variani et al., 2014) 等。该类方法直接提取训练集的辅助特征和声学特征拼接在一起进行训练，提高声学模型的适应性。测试时，提取测试集的辅助特征和声学特征拼接在一起进行测试。

在蒙古语中词根和后缀的存在差异造成蒙古语读音存在差异，而且不同内蒙地区的说话人发音特点也存在差异，因此不同地区说话人口音包含明显的说话人特性信息，所以提取具有区分性的蒙古语说话人特征是蒙古语说话人自适应方法的难点。目前区分性的蒙古语音频属于少量有标记数据，进而提取的区分性说话人特征较少，使用大量区分性很差的说话人特征和少量具有区分性的说话人特征进行声学模型训练，使得蒙古语声学模型学习到少量的区分性信息，导致训练的蒙古语声学模型自适应效果较差。然而基于辅助特征的说话人自适应方法的辅助特征的提取需要用到整句话的信息，不适用于实时语音识别。

(c) 2022 中国中文信息学会计算语言学专业委员会

在蒙古语语音识别 (Mongolian Speech Recognition, MSR) 中采用神经图灵机 (Neural Turning Machine, NTM) (Graves et al., 2014) 进行说话人自适应，通过增加记忆模块来存放说话人特征信息，使用用注意力机制来计算记忆模块的说话人特征与当前语音的说话人特征的相似权重矩阵，将记忆模块中相似的说话人特征通过权重矩阵进行重新组合成一个说话人特征向量 (即 s-vector)，提高各内蒙地区说话人特征之间的区分性，蒙古语声学模型利用区分性说话人特征减小说话人的差异性。本文贡献为：(1) 设计一种基于注意力的说话人特征提取方法，提高说话人特征之间的区分性；(2) 在 IMUT-MCT 数据集上进行验证实验，比较不同说话人特征提取方法，将 s-vector 作为说话人特征来附加声学特征上，验证比直接使用 i-vector 获得更好的性能。

2 相关工作

基于辅助特征的说话人自适应方法是基于特征域的说话人自适应方法中一种重要方法，通过在声学特征中添加特定说话人信息组成新特征，然后利用新特征训练声学模型，进而提高模型的自适应性。按照使用辅助特征的不同，将基于辅助特征的说话人自适应方法分为基于 i-vector 的改进方法和基于 d-vector 的改进方法。

基于 i-vector 的改进方法是在利用说话人特征 i-vector 的基础上改进，进而提高模型的适应性的方法。Saon et al. (2013) 通过将说话人身份向量 i-vectors 和语音识别的声学特征同时作为网络的输入特征，使深度神经网络 (Deep Neural Network, DNN) 声学模型适应目标说话人。Cardinal et al. (2015) 提出将瓶颈 (Bottleneck, BN) 特征，BN 特征提取通过使用 DNN 从语音语料库中提取区分性特征来完成，在阿拉伯语广播新闻任务上词错误率降低了 1.2%。Cui et al. (2017) 提出一种基于辅助特征的说话人自适应训练方法，将说话人特征向量 i-vector 通过控制网络映射到每个隐藏层的仿射变换，以规范化主网络隐藏层输出处的内部特征表示，比使用具有说话人适应输入特征的 i-vector 方法具有更好的性能。然而基于 i-vector 的说话人自适应方法的辅助特征的提取需要用到整句话的信息，不适用于实时语音识别。

基于 d-vector 的改进方法是在利用说话人特征 d-vector 的基础上改进，进而提高模型的适应性的方法。Variani et al. (2014) 提出 d-vector 作为说话人特征，首先训练可以在帧级别对说话人进行分类的 DNN，然后训练的 DNN 可以从最后一个隐藏层中提取说话人特定的特征。Veselý et al. (2016) 采用摘要网络产生输入特征的序列级概要，辅助特征由序列摘要神经网络 (Sequence Summarizing Neural Network, SSNN) 产生一个“摘要向量”，它表示话语的声学摘要，该神经网络将与声学模型相同的特征作为输入，并通过对输出的时间平均进行嵌入，i-vector 和 SSNN 说话人自适应方法都在 AMI 会议数据上进行了比较，两者性能相当。Sari et al. (2019) 使用长短期记忆网络 (Long Short-Term Memory, LSTM) 作为辅助网络，d-vector 作为辅助特征进行说话人自适应，在 HUB4 数据集上，与未适应模型和对抗模型相比，该方法实现了更高的 senone 分类准确度和更低的单词错误率，绝对词错误率降低了高达 2%。同样基于 d-vector 的改进方法的辅助特征的提取需要用到整句话的信息，不适用于实时语音识别。

说话人特征 i-vector 是表示帧级别特征分布模式的特征，其提取本质上是高斯混合模型超向量的降维，并且模型框架假定 i-vector 具有高斯分布。使用 DNN 提取的 d-vector，是来自该 DNN 的最后隐藏层的均值，没有假设任何有关特征分布。但是在 d-vector 说话人特征提取中，加权取平均的方式无法突出区分性说话人信息。本文在基于 i-vector 的基础上引入 NTM，将记忆模块存放各地域说话人特征向量，使用注意力机制代替 NTM 的余弦相似度算法，采用注意

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第344页-第354页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

力机制计算说话人特征与记忆模块中说话人特征的权重矩阵，通过权重矩阵计算出说话人特征，从而提升说话人特征之间区分性，提高蒙古语语音识别系统的适应性。

3 方法

3.1 问题描述

在蒙古语说话人自适应语音识别过程中，内蒙古 A 地区的说话人 A 的一句蒙古语语音序列为 $X = \{x_1, x_2, ..., x_t, ..., x_l\}$，$x_t \in R^D$ 其中表示语音序列 X 中的第 t 个语音帧的特征向量，由一个 D 维向量构成，则语音的声学特征矩阵 $F(X) = [f_1, f_2, ..., f_t, ..., f_l]$，$F(X)$ 是以帧数为行，维度为列。使用声学特征得到对应说话人特征向量 $S = (s_1, s_2, ..., s_n)$，同时内蒙古 B 地区的说话人 B 的一句蒙古语语音序列为 $H = \{h_1, h_2, ..., h_r\}$，同理可得对应说话人特征向量 $D = (d_1, d_2, ..., d_n)$，其中 S 和 D 是两个说话人特征矩阵，将两者通过余弦相似度方法，经过计算可得 $S \cong D$，此时可以说两个地区的说话人特征是不具有区分性。说话人特征提取模型 E 中的说话人特征集合为 M，将 S 和 D 分别通过模型 E 重新提取分别获得说话人特征 S^* 和 D^*，由于模型中的特征 M 不发生变化，因此通过算法提取到的两个特征是 $S^* \cong D^*$。

3.2 模型架构

蒙古语语音识别说话人自适应模型架构主要包括说话人自适应模型和声学模型两个部分，如图 1 所示。说话人自适应模型主要是将输入的蒙古语语音映射到声学特征序列和说话人特征序列，并将两种特征融合成一个特征序列。采用时延神经网络-隐马尔可夫模型 (Time Delay Neural Network-HMM, TDNN-HMM) (Waibel et al., 1989) 作为声学模型，声学模型主要是将每一个单词与基本的发音单位对应起来，根据特征序列直接得到最匹配的字符串。蒙古语声学模型采用 TDNN，优点是动态适应时域特征变化和参数较少，相较于传统的 DNN 的输入层与隐含层互相连接，TDNN 在这里做了一点改变，即隐含层的特征不仅与当前时刻的输入有关，而且还与未来时刻的输入有关。该网络的每一个隐藏层的当前时刻输出值与其向前和向后时间节点的输出值的拼接，作为下一个隐藏层的输出。

给定输入蒙古语语音 $X = \{x_1, x_2, ..., x_t, ..., x_l\}$，$x_t$ 表示第 t 个语音帧，x_t 通过分帧加窗和处理后得到的声学特征向量 $F = (f_1, f_2, ..., f_m)$，经过说话人特征提取模型得到说话人特征向量 $S = (s_1, s_2, ..., s_n)$。TDNN 网络第一个隐藏层的输出计算如公式 (1) 所示。

$$h_t = g(w_f x_t(c) + b_f) \tag{1}$$

其中 h_t 表示第 t 帧输出的隐藏层向量，$x_t(c)$ 表示第 t 帧相邻输入特征拼接的向量，w_f 和 b_f 分别表示 TDNN 网络的权重矩阵和偏置向量，$g(\cdot)$ 表示 ReLU 非线性激活函数。本文的第 t 帧输入特征是将声学特征与说话人特征通过公式 (1) 得到的特征 R，$x_t(c)$ 也表示第 t 帧相邻输入特征 R 拼接的向量。后面的时延隐藏层的计算与第一个隐藏层类似，以上一隐藏层的输出 h_t 作为输入进行计算。基于辅助特征的蒙古语语音识别在线说话人自适应模型不需要对声学模型增加自适应参数，蒙古语声学模型训练过程即是说话人自适应过程。

3.3 基于注意力的说话人特征提取单元

基于注意力的说话人特征提取单元是由注意力模块和记忆模块组成的，其中记忆模块的主要功能是存放说话人特征向量，注意力模块的主要功能是选择记忆模块中说话人特征进行组合成新的说话人特征，基于注意力的说话人特征提取单元结构图如图 2 所示。

(c) 2022 中国中文信息学会计算语言学专业委员会

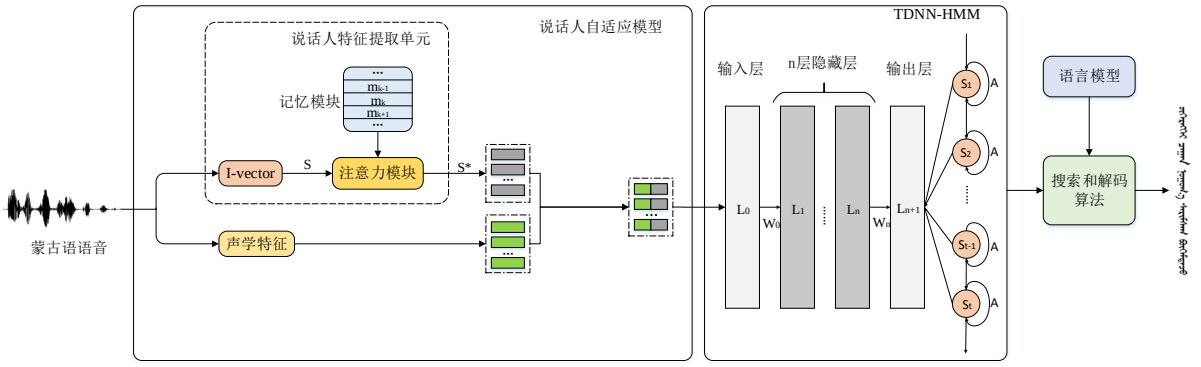


图 1. 蒙古语语音识别说话人自适应模型架构

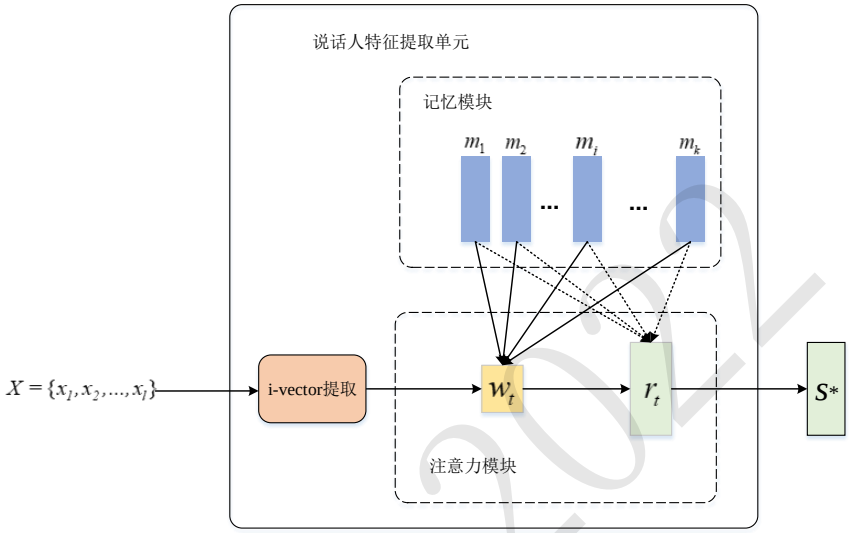


图 2. 基于注意力的说话人特征提取单元结构

记忆模块中存放说话人特征向量是通过训练好的 i-vector 说话人特征提取模型得到的，首先提取各个地区训练集的句子级 i-vector，然后使用 K-means 算法将每个蒙古语地区的说话人特征向量进行聚类，每个地区的 i-vector 的聚类数目相同，最后将聚类的所有地区的 i-vector 特征向量组合成记忆模块。由于受蒙古语语料库的影响，目前蒙古语语料库只有 8 个地区的语料比较丰富，因此记忆模块中包含 8 个蒙古地区的说话人特征向量。

记忆模块可以表示为 $M = \{m_1, m_2, ..., m_k\}$，其中 m_k 表示第 k 个说话人特征向量。给定一个 t 时刻的说话人特征向量 s_t，首先通过注意力机制计算说话人特征向量 s_t 与记忆模块 M 的相似程度矩阵，如下公式：

$$K(s_t, m_i) = vtanh(Ws_t + Um_i) \tag{2}$$

其中，$K(s_t, m_i)$ 表示的是说话人特征向量 s_t 与第 i 个说话人特征向量相似程度，而矩阵 W、U 和向量 v 是注意力模块的参数。

通过 softmax 激活函数将相似程度 $K(s_t, m_i)$ 的输出映射为概率表达，得到锐化后的权重 $w_t(i)$，由说话人特征向量 s_t 和标量 γ_t 和说话人特征向量 s_t 与记忆模块 M 之间相似程度的确定，如下所示：

$$w_t(i) = \frac{e^{\gamma_t K(s_t, m_i)}}{\sum_{j=1}^{k} e^{\gamma_t K(s_t, m_j)}} \tag{3}$$

第二十一届中国计算语言学大会论文集，第344页-第354页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

将权重 $w_t(i)$ 和记忆模块块中所有的蒙古语说话人特征向量的进行加权求和后，即可得到新的说话人向量 r_t，具体表示为：

$$r_t = \sum_{n=1}^{k} w_t(i) m_n \tag{4}$$

其中，r_t 表示第 t 帧基于所有地区的说话人向量集合所构建的新的说话人向量，将区分性说话人特征 r_t 表示为 s-vector。记忆模块中说话人特征向量是来自从不同内蒙地区和不同说话人的语音数据中提取得到的，因此经过注意力模块的计算得到的说话人特征向量之间具有区分性。

4 实验

4.1 实验设置

实验选用的语料库为 IMUT-MC (刘志强等, 2021)，由本实验室构建的一个针对蒙古语语音识别任务的语音语料库，其中 IMUT-MC2 和 IMUT-MC3 的数据来自于蒙古语的日常对话，文本语句都比较简短。录音人员来自内蒙古自治区 8 个地区的 210 人，说话人每人重复录制 200 句，年龄分布在 18 岁到 24 岁之间。蒙古语语音数据均为 16KHz 采样率、16bit 比特率、单声道的格式。

训练集是将 IMUT-MC2 和 IMUT-MC3 的数据集整合在一起，简称为 IMUT-MCT 数据集，其中选取包含 8 个内蒙古地区的 33200 个语音音频，共覆盖了 166 个说话人，其中 84 个男性和 84 个女性。验证集从 IMUT-MCT 数据集中选取包含 8 个内蒙古地区的 3000 个语音音频，共覆盖 15 个说话人，其中 8 个男性和 7 个女性。将训练集与测试集的地区说话人员的交集比例，依次按照 0%，30%，50%，70%，100% 构建 5 组测试集，分别命名为 Test1、Test2、Test3、Test4 和 Test5。其中 Test1 测试集是从 IMUT-MCT 数据集中选取包含 8 个内蒙古地区的 4000 个语音音频，共覆盖了 20 个说话人，其中 10 个男性和 10 个女性。

将通过 3 个实验来验证该方法的有效性。第一个是消融实验，为了说明在蒙古语在线说话人自适应任务中，基于记忆的说话人特征单元的有效性，将蒙古语 s-vector 特征与蒙古语 i-vector 特征和蒙古语 d-vector 特征在不同的测试集上进行对比。第二个是说话人特征提取方法对比实验，旨在探索基于注意力的说话人特征方法对于不同蒙古语语音识别系统的适应性，主要是从不同蒙古语说话人特征 i-vector、d-vector 和 s-vector 与不同蒙古语声学模型 DNN-HMM、LSTM-HMM 和 End-to-End 上进行对比实验。第三个是案例分析验，通过蒙古语语音样例来展示真实的识别结果，测试语料使用 i-vector 与 s-vector 两种说话人特征提取方法进行识别，验证本文提出方法的有效性。

4.2 评价指标

蒙古语语音识别模型的评价指标包括有词错率 (Word Error Rate, WER)、句错率 (Sentence Error Rate, SER) 和错误下降率 (Error Drop Rate, EDR)。

(1)WER 是指所有错误词的和所占总词数的百分比，计算公式为：

$$WER = \frac{S_w + D_w + I_w}{N_W} * 100\% \tag{5}$$

其中 S_w 表示替换错误的词数，D_w 表示删除错误的词数，I_w 表示插入错误的词数，N_W 表示数据集中的总词数。

(c) 2022 中国中文信息学会计算语言学专业委员会

(2)SER 是指所有识别结果与对应文本不能正确匹配的测试语音所占总语音数的百分比，计算公式为：

$$SER = \frac{N_error}{N_S} * 100\% \tag{6}$$

其中 N_error 表示识别错误的蒙古语音频的个数，N_S 表示数据集中蒙古语音频的总个数。

(3)EDR 是现在方法的错误率与原来方法的错误率相差的值和原来方法的错误率的百分比值，计算公式为：

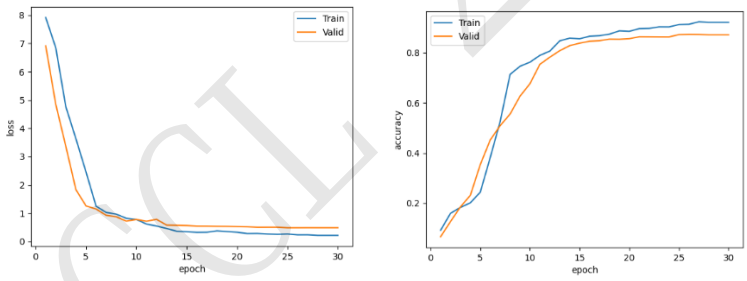
$$EDR = \frac{|B - A|}{B} * 100\% \tag{7}$$

其中 A 表示现在方法的错误率，B 表示原来方法的错误率，错误率可以是该方法的 WER 或者 SER。

4.3 实验结果与分析

4.3.1 收敛性实验

训练过程在 IMUT-MCT 数据集上对使用基于注意力的说话人特征提取单元的蒙古语语音识别模型开展，收敛情况如图 3 所示。为了保证基于注意力的蒙古语语音识别模型的收敛效果，实验采用训练集和验证集的损失值和准确率来验证模型收敛。由图 3a) 中可知，基于注意力的蒙古语语音识别模型训练集和验证集上的损失函数呈指数下降，并于 25 轮左右使 Loss 趋于平缓，表明模型能够收敛。由图 3b) 中可知，在训练集和验证集上的准确率不断上升，最终在 25 轮左右识别准确率趋于平缓。综上所述，使用基于注意力的说话人提取单元提取说话人特征作为声学模型的输入来训练得到的模型可以收敛且无过拟合现象，可以学习到了训练集的数据分布特征。



(a) 训练的 loss 值的变化　　(b) 训练过程中识别准确率

图 3. 基于注意力的蒙古语语音识别模型的收敛情况

4.3.2 说话人特征单元消融实验

说话人特征单元消融实验在不同测试集上将提出的蒙古语 s-vector 特征与蒙古语 i-vector、d-vector 特征进行对比。实验对不同说话人特征在蒙古语语音识别系统的 WER 和 SER 进行验证，其中采用特征拼接方法将蒙古语声学特征 Fbank 和说话人特征进行融合，实验结果如表 1 所示。

分析表 1 实验结果可知：不采用蒙古语说话人特征进行蒙古语说话人自适应时，Test5 测试集的蒙古语语音识别系统取得了比 Test1 测试集的蒙古语语音识别系统更好的识别效果；当采用不同的说话人特征进行说话人自适应时，Test2 测试集的 WER 和 SER 优于 Test1 测试集的 WER 和 SER，训练集与测试集的地区说话人数的交集比例逐步增大，WER 和 SER 也在随之

降低。对比不同的说话人特征，当说话人无论是否包含在训练集或者自适应集中时，s-vector 特征进行说话人自适应得到的 WER 均优于蒙古语 i-vector、d-vector 特征，比如在测试集为 Test1 和测试集为 Test5 的实验中，采用蒙古语 s-vector 特征比蒙古语 i-vector、d-vector 特征进行蒙古语语音识别说话人自适应的 WER 分别降低了 4.96%、1.08% 和 5.39%、1.2%。虽然在测试集为 Test1 时 d-vector 方法比 s-vector 方法的 SER 低了 1.01%，但是在其他测试集上蒙古语 s-vector 特征的 SER 优于蒙古语 i-vector。实验结果验证提出蒙古语 s-vector 特征的有效性。

说话人特征	WER(%)					SER(%)				
	Test1	**Test2**	**Test3**	**Test4**	**Test5**	**Test1**	**Test2**	**Test3**	**Test4**	**Test5**
无	36.54	33.87	30.26	29.23	28.87	58.37	56.18	54.83	52.48	50.36
i-vector	30.33	28.23	27.18	26.08	25.97	43.16	40.73	39.95	36.48	35.38
d-vector	26.45	25.07	23.88	23.53	21.78	**36.77**	35.25	35.78	**32.53**	31.47
s-vector	**25.37**	**23.28**	**22.28**	**21.19**	**20.58**	37.78	**34.83**	**34.47**	33.58	**31.23**

表 1. 说话人特征提取单元的消融实验

说话人特征	声学模型	WER(%)					SER(%)				
		Test1	**Test2**	**Test3**	**Test4**	**Test5**	**Test1**	**Test2**	**Test3**	**Test4**	**Test5**
无	DNN-HMM	38.56	36.86	34.66	33.46	31.34	64.16	62.73	60.46	58.86	56.83
	TDNN-HMM	36.54	33.87	30.26	29.23	28.87	58.37	56.18	54.83	52.48	50.36
	LSTM-HMM	26.54	25.89	24.33	23.23	22.96	42.53	40.45	38.16	36.10	35.85
	End-to-End	21.45	20.23	19.78	18.86	17.23	35.52	34.46	32.21	31.58	30.97
i-vector	DNN-HMM	33.46	31.78	30.89	28.72	26.34	58.45	56.36	55.86	54.58	54.13
	TDNN-HMM	30.33	28.23	27.18	26.08	25.97	43.16	40.73	39.95	36.48	35.38
	LSTM-HMM	24.23	22.15	22.89	21.76	20.45	36.63	35.25	34.03	32.58	30.56
	End-to-End	19.89	18.86	17.45	16.23	15.99	30.56	28.06	26.75	25.36	25.45
d-vector	DNN-HMM	30.23	28.18	26.59	25.45	24.56	39.36	38.34	37.98	36.43	34.21
	TDNN-HMM	26.45	25.07	23.88	23.53	21.78	36.77	35.25	35.78	32.53	31.47
	LSTM-HMM	23.56	24.78	24.88	22.26	20.08	34.23	33.03	32.74	30.47	29.31
	End-to-End	18.97	17.36	16.56	16.23	15.56	29.89	28.73	26.16	25.83	24.86
s-vector	DNN-HMM	29.89	28.56	27.72	25.34	24.02	39.89	38.78	36.56	35.23	34.12
	TDNN-HMM	25.37	23.28	22.28	21.19	20.58	37.78	34.83	34.47	33.58	31.23
	LSTM-HMM	23.78	22.55	21.45	20.26	19.87	33.57	32.75	31.48	30.33	29.15
	End-to-End	18.70	17.45	16.16	15.95	14.89	29.36	28.87	26.54	25.13	24.05

表 2. 不同说话人特征提取方法在不同声学模型上的 WER

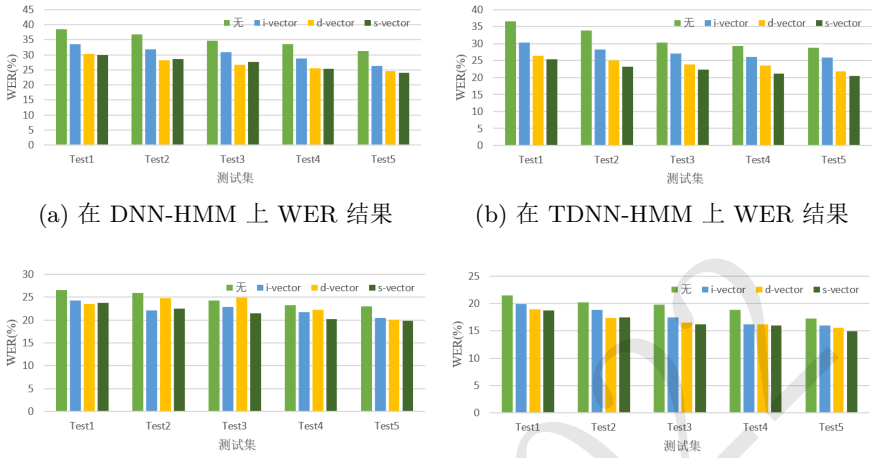
4.3.3 模型适应性实验

为了验证不同说话人特征提取方法在不同语音识别模型上的性能，在不同测试集下开展模型适应性实验。实验分别对比 DNN-HMM、LSTM-HMM 和 End-to-End 声学模型。本文提出的蒙古语 s-vector 特征与蒙古语说话人特征 i-vector、说话人特征 d-vector 进行对比。实验对不同说话人特征在不同的蒙古语语音识别系统的 WER 和 SER 进行验证，实验结果如表 2 所示。

分析表 2 实验结果可知：对比不同说话人特征提取方法和在不同模型上性能，在这四种模

型中可以看出 End-to-End 的识别 WER 和句 SER 均优于 DNN-HMM、TDNN-HMM、LSTM-HMM，比如在说话人特征提取方法为 d-vector、测试集为 Test1 和提取方法为 s-vector、测试集为 Test1 的实验中，采用 End-to-End 比 DNN-HMM、TDNN-HMM 和 LSTM-HMM 进行蒙古语语音识别说话人自适应的 WER 分别降低了 11.26%、7.48%、4.59% 和 11.19%、6.67%、5.08%，实验结果验证了在不同模型中蒙古语 s-vector 特征对蒙古语语音识别系统的有效性。

为了更加直观在不同声学模型上使用不同说话人特征提取方法所对应结果，如图 4 所示。在图 4 中，横坐标为不同蒙古语测试集，纵坐标为词错率。



(a) 在 DNN-HMM 上 WER 结果　(b) 在 TDNN-HMM 上 WER 结果

(c) 在 LSTM-HMM 上 WER 结果　(d) 在 End-to-End 上 WER 结果

图 4. 不同说话人特征提取方法在不同声学模型上的 WER

分析图 4 实验结果可知：在图 4(a) 中，可以看出使用说话人自适应方法比没有使用说话人自适应取得了更好的识别准确率，而且 s-vector 和 d-vector 在不同测试集上均优于 i-vector，也可以在图中看出 s-vector 和 d-vector 性能相当。说话人特征 s-vector 是基于 i-vector 的改进方法，比 i-vector 有更好的性能。在图 4(b) 中，s-vector 和 d-vector 在不同测试集上均优于 i-vector，但是在不同的测试集 s-vector 性能优于 d-vector。在图 4(c) 和图 4(d) 中，得到了与图 34(a) 类似的结论，验证了 s-vector 在不同声学模型的适应性能力。

(a) 在 WER 上的结果　(b) 在 SER 上的结果

图 5. 说话人提取单元的错误下降率

为直接验证说话人特征提取单元对蒙古语语音识别系统的影响，本文还采用 EDR 评价指标对不同模型进行评价，其中是采用 s-vector 特征方法和没有说话人特征方法之间的错误下降率，如图 5 所示。在图 5 中，横坐标为不同蒙古语测试集，纵坐标为 ESR(%)。

从图 5(a) 可以看出，在不同测试集下，下降率最高的为 TDNN-HMM 模型，其次才是 DNN-HMM 模型，最低的为 End-to-End 模型。但是从图 5(b) 可以看出，在不同测试集下，下降率最高的为 DNN-HMM 模型，其次才是 TDNN-HMM 模型，最低的为 End-to-End 模型。综上所述，可以看出本文的方法比较适用于 TDNN-HMM 模型和 DNN-HMM 模型，对于这个两种模型影响较大。

4.3.4 案例分析

为了验证说话人特征提取单元的有效性，通过样例来展示真实的识别结果。在蒙古语语音中，例如语音：00001373-F-M-19.wav，其中 00001373 表示语句编号；F 为地区编码，数据集中收集了 10 个地区，F 表示兴安盟地区；M 为性别编码，M 表示男，F 表示女；19 表示年龄。表 3 展示了蒙古语语音识别的样例，1-3 是包含在记忆模块地区和不包含在训练集中，4-7 是不包含在记忆模块地区和不包含在训练集中，其中选取的语音数据的时长集中在 4-7s 之间。

编号	语音数据编号	标签数据	备注
1	00000016-F-M-19.wav	ᠨᠠ᠋/ ᠬᠦᠮᠦᠨ/ ᠤ᠋ ᠪᠠ/ ᠤᠨ/ ᠤᠨ ᠲᠠᠨᠢᠯᠴᠠᠭᠤᠯᠬᠤ ᠪᠠᠶᠢᠴᠠᠭᠠ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ	他把朋友介绍给您了吗？
2	00000140-F-F-20.wav	ᠨᠢᠭᠡᠨ ᠳᠤ ᠴᠠᠭᠠᠨᠵᠢᠭᠠᠨ ᠳᠡᠭᠡ ᠮᠣᠩᠭᠣᠯ ᠬᠦᠮᠦᠨ ᠬᠣᠶᠠᠷ ᠬᠣᠨᠣᠭ ᠪᠤᠶᠤ ᠰᠠᠷ᠎ᠠ ᠪᠤᠶᠤ	第一次机遇十多天或者一个月左右。
3	00000206-C-F-19.wav	ᠠᠢᠮᠠᠭ ᠳᠡᠭᠡ ᠪᠤᠢᠴᠠᠭᠠᠨ᠎ ᠪᠠᠢᠭᠠᠭᠤᠯᠤᠭᠴᠢ ᠬᠡᠮᠦᠦ ᠬᠤᠪᠢ ᠪᠤᠶᠤ ᠪᠤᠰᠤ᠎ ᠪᠠᠢᠨ᠎ᠠ	附近有的话，好坏无所谓。
4	00000220-H-M-23.wav	ᠲᠡᠷᠡᠭᠡᠳᠦ ᠪᠢᠳᠡᠨᠢ᠎ ᠪᠠ ᠪᠠᠢᠬᠠ ᠬᠤᠪᠢᠴᠠᠭᠠᠮᠤᠷ᠎ ᠳ ᠤᠳᠤᠨ᠎ ᠪᠠ ᠪᠤᠶᠤ᠎	那么我们俩进二手衣服店吧。
5	00000292-H-F-19.wav	ᠲᠠ ᠠᠭᠤᠯᠬᠠ᠎ᠠᠷ᠎ᠠ ᠤᠳᠤᠨ᠎ ᠤᠤ ᠵᠠᠭᠰᠠᠭᠠᠨ᠎ ᠲᠤ ᠪᠠ	您带几箱子东西？
6	00000306-G-M-20.wav	ᠬᠡᠷᠡᠭᠵᠢᠭᠡᠨ ᠪᠤᠶᠤ᠎ ᠪᠠ ᠬᠡᠷᠡᠭᠵᠢᠭᠡᠨ ᠪᠤᠶᠤ᠎ ᠳᠡᠭᠡ ᠪᠠ ᠬᠡᠷᠡᠭ	该求的时候要求，该要的时候该要。
7	00000843-G-F-19.wav	ᠮᠣᠩᠭᠣᠯ ᠴᠣᠷᠪᠣ᠎ ᠳ ᠠᠮᠢᠲᠠᠨ ᠪᠤᠯ ᠬᠣᠶ ᠢᠨ ᠬᠤᠶᠠᠷ ᠬᠥᠬᠡ᠎ ᠪᠤᠶᠤ᠎ ᠬᠣᠨᠢ ᠪᠤᠶᠤ᠎	蒙古人的五畜是指羊、山羊、牛、马、骆驼。

表 3. 蒙古语测试样例数据

将上述的测试样例数据采用基于注意力的说话人特征提取方法进行识别，结果如表 4 所示。其中具有下划线的词、“***”和字体加双下划线的词分别表示替换词、删除词和插入词。

编号	识别标签结果	备注	替换词错误	删除词错误	插入词错误
1	ᠨᠠ/ ᠬᠦᠮᠦᠨ/ ᠤ ᠪᠠ/ ᠬᠦᠳᠦ ᠤᠨ/ ᠤᠨ/ ᠲᠠᠨᠢᠯᠴᠠᠭᠤᠯᠬᠤ ᠪᠠᠶᠢᠴᠠᠭᠠ ᠪᠠᠢᠨ᠎ᠠ	请朋友介绍这两家业余的事	1	0	1
2	ᠨᠢᠭᠡᠨ ᠳᠤ ᠴᠠᠭᠠᠨᠵᠢᠭᠠᠨ ᠴᠢᠭᠯᠡ ᠮᠣᠩᠭᠣᠯ ᠬᠦᠮᠦᠨ ᠬᠣᠶᠠᠷ ᠬᠣᠨᠣᠭ ᠪᠤᠶᠤ ᠰᠠᠷ᠎ᠠ ᠪᠤᠶᠤ	第一次机遇十多天偶尔一个月左右。	1	0	0
3	ᠠᠢᠮᠠᠭ ᠳᠡᠭᠡ ᠪᠤᠢᠴᠠᠭᠠᠨ᠎ ᠪᠠᠢᠭᠠᠭᠤᠯᠤᠭᠴᠢ ᠬᠡᠮᠦᠦ ᠬᠤᠪᠢ ᠪᠤᠶᠤ ᠪᠤᠰᠤ᠎ ᠪᠠᠢᠨ᠎ᠠ	附近有的话，坏无所谓。	1	0	0
4	ᠲᠡᠷᠡᠭᠡᠳᠦ ᠪᠢᠳᠡᠨᠢ᠎ ᠪᠠ ᠪᠠᠢᠬᠠ ᠬᠤᠪᠢᠴᠠᠭᠠᠮᠤᠷ᠎ ᠳ ᠤᠳᠤᠨ᠎ ᠪᠠ ᠪᠤᠶᠤ᠎ ᠲᠤᠭᠤ/	那么我们俩进二手衣服店什么的。	0	0	1
5	ᠲᠠ/ ᠠᠭᠤᠯᠬᠠ᠎ᠠᠷ᠎ᠠ ᠤᠳᠤᠨ᠎ ᠤᠤ ᠵᠠᠭᠰᠠᠭᠠᠨ᠎ ᠲᠤ ᠪᠠ	您带几箱子东西？	0	0	0
6	ᠬᠡᠷᠡᠭᠵᠢᠭᠡᠨ ᠪᠤᠶᠤ᠎ ᠪᠠ ᠬᠡᠷᠡᠭᠵᠢᠭᠡᠨ ᠪᠤᠶᠤ᠎ ᠳᠡᠭᠡ ᠪᠠ ᠬᠡᠷᠡᠭ	该求的时候要求，该要的时候该要。	0	0	0
7	ᠮᠣᠩᠭᠣᠯ ᠴᠣᠷᠪᠣ᠎ ᠳ *** ᠠᠮᠢᠲᠠᠨ ᠪᠤᠯ ᠬᠣᠶ ᠢᠨ ᠬᠤᠶᠠᠷ ᠬᠥᠬᠡ᠎ ᠪᠤᠶᠤ᠎ ᠬᠣᠨᠢ ᠪᠤᠶᠤ᠎	蒙古人的 *** 畜是指羊、山羊、牛、马、骆驼。	0	1	0

表 4. 识别结果据

分析表 4 中识别结果可知：通过错误次数进行分析，使用 s-vector 得到的识别结果优于 i-vector，表明该方法得到的区分性说话人特征有利于声学模型建模。但是使用本文方法识别结果还存在错误，说明该方法仍需要改进进而提高准确率。

5 结论

围绕基于辅助特征的说话人自适应方法在蒙古语语音识别任务的适应性问题上开展研究。设计基于注意力的说话人特征提取单元，增加提取到蒙古语说话人特征之间的区分性，并将其用于声学模型的建模，降低蒙古语语音识别系统的 WER 和 SER。通过说话人特征单元的消融实验，可知使用 s-vector 特征比 d-vector 特征的 WER 降低 1.08%。实验结果间接表明蒙古语 s-vector 特征获得了较高的说话人特征区分性，并且提升蒙古语语音识别系统的识别准确率，但是部分结果中证明区分性的说话人特征对于声学模型建模也会产生不好的影响。未来研究可以尝试将不同类型的说话人特征存入记忆模块进行对比，如 x-vector (Snyder et al., 2018) 或 d-vector。

第二十一届中国计算语言学大会论文集，第344页-第354页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Abdel-Hamid O and Jiang H. 2013. *Fast Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition based on Discriminative Learning of Speaker Code.* Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, May 26-31, pages 7942-7946.

Bell P, Fainberg J, Klejch O, et al. 2020. *Adaptation Algorithms for Speech Recognition: An Overview.* IEEE Open Journal of Signal Processing, 2:33-66.

Cardinal P, Dehak N, Zhang Y, et al. 2015. *Speaker Adaptation using the I-vector Technique for Bottleneck Features.* Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Sep 6-10, pages 2867-2871.

Cui X D, Goel V, Saon G, et al. 2017. *Embedding-based Speaker Adaptive Training of Deep Neural Networks.* Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm, Aug 20-24, pages 122-126.

Dong Y, Yao K S, Hang S, et al. 2013. *KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition.* Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, May 26-31, pages 7893-7897.

Graves A, Wayne G, and Danihelka I. 2014. *Neural Turing Machines.* https://arxiv.org/abs/1410.5401 2014-12-10

Neto J P, Almeida L B, Hochberg M, et al. 1995. *Speaker-Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System.* Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Sep 18-21, pages 2171-2174.

Stadermann J and Rigoll G. 2005. *Two-Stage Speaker Adaptation of Hybrid Tied-Posterior Acoustic Models.* Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, Mar 18-23, pages 977-980.

Saon G, Soltau H, Nahamoo D, et al. 2013. *Speaker Adaptation of Neural Network Acoustic Models using I-vectors.* Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Dec 8-12, pages 55-59.

Swietojanski P and Renals S. 2014. *Learning Hidden Unit Contributions for Unsupervised Speaker Adaptation of Neural Network Acoustic Models.* Proceedings of the 2014 IEEE Spoken Language Technology Workshop, South Lake, Dec 7-10, pages 171-176.

Samarakoon L and Sim K C. 2016. *Factorized Hidden Layer Adaptation for Deep Neural Network based Acoustic Modeling.* IEEE Transactions on Audio, Speech and Language Processing, 24(12): 2241-2250.

Snyder D, Garcia-Romero D, SELL G, et al. 2018. *X-Vectors: Robust DNN Embeddings for Speaker Recognition.* Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, Apr 15-20, pages 5329-5333.

Sari L, Thomas S, Hasegawa-Johnson M A. 2019. *Embedding-based Speaker Adaptive Training of Deep Neural Networks.* Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm, Aug 20-24, pages 122-126.

Variani E, Lei X, McDermott E, et al. 2014. *Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification.* Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, May 4-9, pages 4052-4056.

Veselý K, Watanabe S, Zmolíková K, et al. 2016. *Sequence Summarizing Neural Network for Speaker Adaptation.* Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, Mar 20-25, pages 5315-5319.

Waibel A, Hanazawa T, Hinton G, et al. 1989. *Phoneme Recognition using Time-Delay Neural Networks.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(3):328-339.

刘志强, 马志强, 张晓旭, 等. 2021. *IMUT-MC*：一个针对蒙古语语音识别的语音语料库. 中国科学数据, 2021.(2021-12-29). DOI: 10.11922/11-6035.csd.2021.0096.zh

朱方圆，马志强，陈艳, 等. 2021. 语音识别中说话人自适应方法研究综述. 计算机科学与探索, 2021, 15(12):2241-2255.

(c) 2022 中国中文信息学会计算语言学专业委员会

融合双重注意力机制的缅甸语图像文本识别方法

王奉孝[1,2]，毛存礼[*1,2]，余正涛[1,2]，高盛祥[1,2]，黄于欣[1,2]，刘福浩[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

1499539796@qq.com, maocunli@163.com, ztyu@hotmail.com

gaoshengxiang.yn@foxmail.com, huangyuxin2004@163.com, 1519195149@qq.com

摘要

由于缅甸语字符具有独特的语言编码结构以及字符组合规则，现有图像文本识别方法在缅甸语图像识别任务中无法充分关注文字边缘的特征，会导致缅甸语字符上下标丢失的问题。因此，本文基于Transformer框架的图像文本识别方法做出改进，提出一种融合通道和空间注意力机制的视觉关注模块，旨在捕获像素级成对关系和通道依赖关系，降低缅甸语图像中噪声干扰从而获得语义更完整的特征图。此外，在解码过程中，将基于多头注意力的解码单元组合为解码器，用于将特征序列转化为缅甸语文字。实验结果表明，该方法在自构的缅甸语图像文本识别数据集上相比Transformer识别准确率提高0.5%，达到95.3%。

关键词： 缅甸语；文本识别；通道和空间注意力；特征增强；文字边缘特征

Burmese image text recognition method with dual attention mechanism

Fengxiao Wang[1,2], Cunli Mao[*1,2], Zhengtao Yu[1,2], Shengxiang Gao[1,2], Yuxin Huang[1,2], Fuhao Liu[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

1499539796@qq.com, maocunli@163.com, ztyu@hotmail.com

gaoshengxiang.yn@foxmail.com, huangyuxin2004@163.com, 1519195149@qq.com

Abstract

Due to the unique language coding structure and character combination rules of Burmese characters, the existing image text recognition methods cannot fully pay attention to the features of text edges in the Burmese image recognition task, which will lead to the loss of subscripts and subscripts of Burmese characters. Therefore, this paper improves the image text recognition method based on the Transformer framework, and proposes a visual attention module that fuses channel and spatial attention mechanisms, aiming to capture pixel-level pairwise relationships and channel dependencies and reduce noise interference in Burmese images. Thereby a more semantically complete feature map is obtained. Furthermore, in the decoding process, multi-head attention-based decoding units are combined into a decoder for converting feature sequences into Burmese scripts. The experimental results show that the recognition accuracy of this method is 0.5% higher than that of Transformer on the self-constructed Burmese image text recognition dataset, reaching 95.3%.

国家自然科学基金重点项目（61732005,U21B2027）；国家自然科学基金（62166023，61866019）；云南省自然科学基金重点项目(2019FA023);云南省重大科技专项计划项目(202103AA080015，202002AD080001)

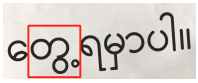
第二十一届中国计算语言学大会论文集，第355页-第365页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

Keywords: Burmese , Text recognition , Channels and Spatial Attention , Feature enhancement , Text edge features

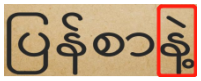
1 引言

由于缅甸语属于一种典型的低资源语言，互联网中存在大量的缅甸语文本图像，因此，快速精准地提取缅甸语文本图像中的文本信息对于开展面向缅甸语的自然语言处理、机器翻译、信息检索等研究具有重要的意义。

(a) 含上标图像



(b) 含下标图像



(c) 含上下标图像

Figure 1: 缅甸语字符边缘特征图像示例

现有方法在针对中英文的图像识别任务上已经取得很好的效果，但缅甸语字符的语言编码结构以及字符组合规则与中英文具有巨大的差异性，其字符主要由基础字符、基础前字符、基础后字符、基础上字符以及基础下字符构成，缅甸语中存在大量的由多个字符组成一个音节的情况，如Figure1中(c)的"န့်"是由"န"、"◌့"以及"◌်"等三个字符组成，这样的字符组成结构，在图像识别过程中会由于其上下标字符边缘特征不明显导致部分语义信息丢失，例如在识别Figure1（c）中"န့်"容易丢失"◌့"或"◌်"等上下标信息，从而极大地影响了缅甸语图像文本识别的准确率。

此外，研究人员针对缅甸语文本图像识别任务也尝试了很多有意义的工作，毛存礼et al. (2022)提出了利用知识蒸馏的方式将单字符特征的相关知识传递给学生模型，用于提升学生模型的特征提取能力，从而缓解缅甸语字符丢失及识别错误的问题，但是该方法忽略了深度卷积神经网络中的底层语义特征。Liu et al. (2021)提出一种基于多层语义特征融合的缅甸语图像文本识别方法，将卷积神经网络提取的具有缅甸语特征信息特征图进行融合操作，实现主干网络对缅甸语特征提取能力的增强。然而，其特征提取网络对缅甸语文本边缘特征的提取并不充分，并且在效率上仍然有待提高。目前，随着深度神经网络的发展，Transformer在各大图像任务中展现出优异的性能(Dosovitskiy et al., 2020; Ali et al., 2021; Fan et al., 2021)，基于Transformer架构能够捕捉到整张图像的全局信息以及序列之间的依赖关系，这样的解码方式虽然有利于提升缅甸语文字的识别性能，但是针对缅甸语的字符上下标等边缘特征的识别仍然存在较大的挑战。

为了解决上述问题，本文主要针对缅甸语图像文本识别过程中字符上下标等边缘特征容易丢失的问题展开研究，受到卷积块注意模块(Woo et al., 2018)思想的启发，提出一种融合通道和空间注意力机制的缅甸语图像文本识别方法，我们考虑对经过图像特征提取网络得到的特征图同时构建空间注意力和通道注意力来获取缅甸语图像更细粒度的位置特征和通道映射特征，并将获取的两个特征进行融合，最后利用多头注意力机制对融合结果进行注意力计算，捕捉文本之间的全局信息。

本文的工作主要有以下贡献:

（1）我们提出一种融合双重注意力机制的缅甸语图像特征提取方法，使得模型可以更多地关注到缅甸语文本图像的上下标区域;

（2）我们基于Transformer模型设计一个适用于缅甸语图像文本识别的框架，使模型可以进行并行训练，极大提升了识别效率。

（3）在自构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识别准确率达到了95.3%，优于多个对比模型。

2 相关工作

现有图像文本识别方法大致分为联结主义时间分类的图像文本识别方法、基于序列到序列

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

的图像文本识别方法以及基于Transformer的图像文本识别方法，具体如下：

（1）基于联结主义时间分类的方法

基于（Connectionist Temporal Classification，CTC）的文本识别方法引入CTC 损失作为目标优化函数。该算法的本质是先定义预测结果到真实标签之间的转化方式，采用动态规划的策略从输出概率分布中获取多条状态转移路径，将所有路径概率之和的最大值作为目标优化函数。因此，CTC方法使其只需要文字级注释就可以进行端到端训练，而不需要字符级注释。Graves et al. (2007)提出首次将CTC应用在OCR领域的手写识别系统。随着深度神经网络的发展，(Shi et al., 2016)利用不受卷积神经网络(convolutional neural networks，CNN)输入空间大小限制的特性，提出了一个将卷积神经网络与循环神经网络(recurrent neural networks，RNN)一起识别场景文本图像的模型。采用全卷积方法对输入图像进行整体编码生成特征切片，引入了长短时记忆(Long Short-Term Memory，简称LSTM)用来增强上下文建模，最终将输出的特征序列输入到CTC模块，直接解码序列结果。(Gao et al., 2017)代替RNN采用堆叠的卷积层来有效地捕获输入序列的上下文依存关系，其主要优点在于较低的计算复杂度和较容易的并行计算。Yin et al. (2017)也避免在模型中使用RNN，他们通过使用字符模型滑动文本行图像来同时检测和识别字符，这些字符模型是在标记有文本记录的文本行图像上端对端学习的。

（2）基于序列到序列的图像文本识别方法

基于注意力机制的文本识别方法首先通过编码器将图像特征转化为中间语义特征，再利用基于注意力模型的解码器将中间语义特征转化为文本。这类方法通过训练可以学习到任意长度的序列之间的对齐关系，一定程度上缓解了序列对齐问题。受注意力机制在机器翻译任务中的成果应用，Lee and Osindero (2016)将注意力模型与循环神经网络进行融合，以提升文字预测效果。为了解决注意机制应用到文字识别中的注意力偏移问题，Cheng et al. (2017)设计聚焦注意力网络来解决该问题。为了让模型更加关注于文字识别相关的图像区域，Li et al. (2019)将1D attention拓展到2D attention上，用2D attention可以更精准的选取字符区域特征，忽略掉背景信息。具体来讲，相比于已有的1D attention，2D attention可以在纵向进行特征筛选与融合。Shi et al. (2018)将注意力序列-序列模型引入到场景文本识别问题中，设计的矫正网络采用（Spatial Transformer Networks，STN）Jaderberg et al. (2015)和薄板样条插值算法结合，将输入图像中不规则的文本区域变换成规则的文本区域图像，提高了不规则文本图像的识别准确率。

（3）基于Transformer的图像文本识别方法

随着Transformer 的快速发展，分类和检测领域都验证了Transformer 在视觉任务中的有效性。在针对规则文本识别的过程中，CNN在长依赖建模上会存在局限性，Transformer结构恰好解决了这一问题，它可以在特征提取器中关注全局信息。Yu et al. (2020)将Transformer的Encoder模块接在ResNet50后，增强了2D视觉特征。并提出了一个并行注意力模块，将读取顺序用作查询，使得计算与时间无关，最终并行输出所有时间步长的对齐视觉特征。Sheng et al. (2019)使用了完整的Transformer结构对输入图片进行编码和解码，只使用了简单的几个卷积层做高层特征提取，在文本识别上验证了Transformer结构的有效性。Yang et al. (2020)使用Transformer的解码器替换LSTM，再一次验证了并行训练的高效性和精度优势。

以上方法为本文解决缅甸语图像文本识别任务提供了较好的思路，本文方法与现有工作主要区别是提出一种融合通道注意力和空间注意力的视觉关注模块，对深度卷积神经网络提取的缅甸语图像特征图分别获取通道域和空间域的注意力图，融合后对原特征图重构，使缅甸语图像文字边缘特征能够获得更多的注意力关注，进而缓解缅甸语图像文本识别中上下标字符易丢失的问题。

3 融合双重注意力机制的缅甸语图像文本识别模型

本文提出的网络架构如图所示，模型架构由基于ResNet(He et al., 2016)的特征提取模块、融合通道注意力和空间注意力的视觉信息关注模块和解码模块三部分组成。特征提取模块主要对输入的缅甸语文本图像经过卷积神经网络提取到其文本信息特征，再通过视觉信息关注模块进行注意力计算，增强缅甸语图像的文本特征表征能力，最后通过解码器解码转录出对应的缅甸语文字。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.1 缅甸语图像特征提取网络

我们在残差网络(Residual Network，ResNet)的基础上构建了适应缅甸语图像特征提取的主干网络，通过特征提取网络获得512维的缅甸语图像特征图。

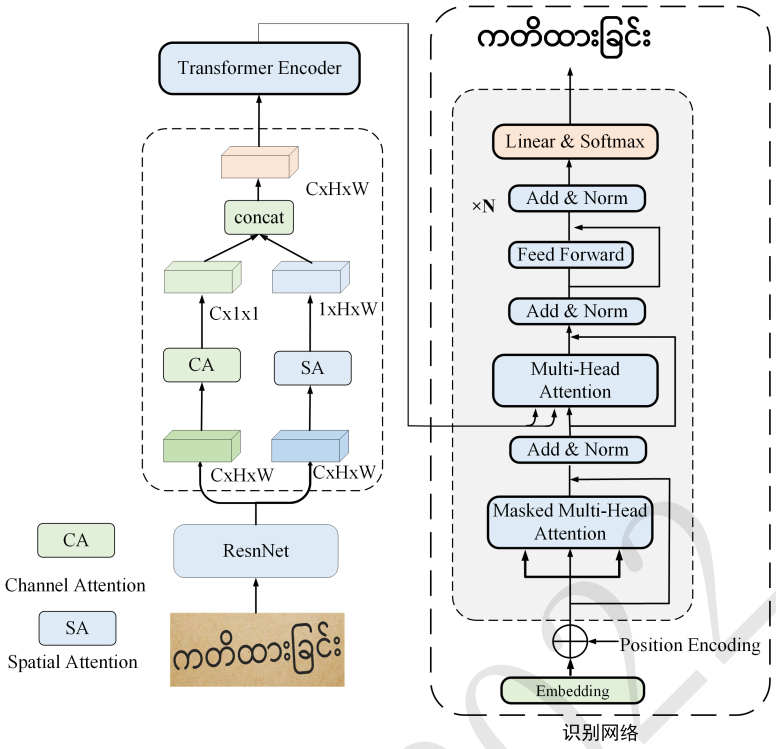


Figure 2: 融合双重注意力机制的缅甸语文本图像识别模型结构图

3.2 缅甸语图像语义特征增强

为了能够更好地捕捉缅甸语图像的像素级成对关系和通道依赖关系，排除图像中噪声干扰，从而获取到语义更精准的特征图，我们设计了通道注意力机制和空间注意力机制，以生成混合域的注意力向量，并对原特征进行重构，提高缅甸语图像的文字区域的表征能力。

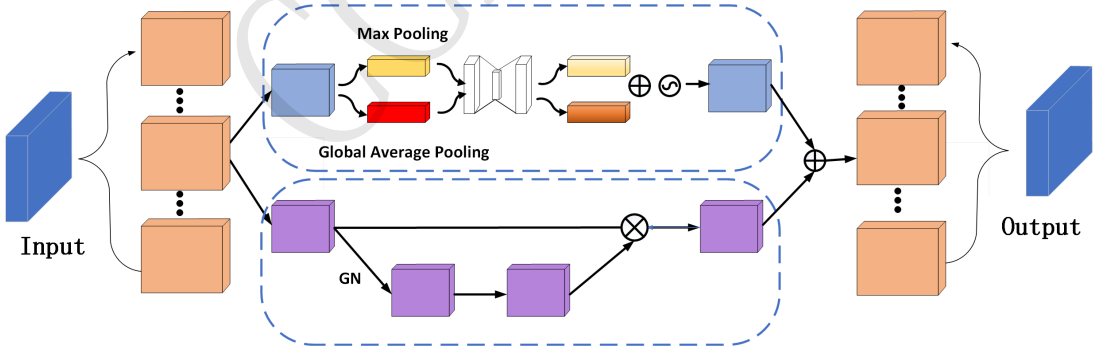


Figure 3: 双重注意力模块结构图

（1）特征图分组

输入的缅甸语图像通过特征提取网络得到通道数为512维的特征图，假设这些输入特征为$X \in R^{C\times H\times W}$，其中C，H，W分别表示通道数、空间高度和宽度，将特征X沿着通道维度拆分为K组：$X = [X_1, ...X_K]$，$X_i \in R^{C\times H\times W}$，其中每个子特征$X_i$在训练过程中逐渐捕获特定的语义响应对于每组特征，我们通过进行通道分组操作，在每个注意单元的开头，X_i的输入沿着通道维度被分成两个分支，即X_{i1}、$X_{i2} \in R^{C/2K\times H\times W}$。一个分支通过利用通道的相互关系

(c) 2022 中国中文信息学会计算语言学专业委员会

在[C] 维度上获取注意力权重来生成通道注意力图，而另一个分支则通过利用特征的空间关系在[H，W]维度上进行注意力权重计算来生成空间注意力图。

（2）通道注意力

通道注意力能够显式地建模特征通道之间的相互依赖关系。就是通过学习的方式来自动获取到每个特征通道的注意力权重，然后根据这个注意力权重去提升缅甸语图像中文本相关区域的特征并抑制背景及其他噪声信息的干扰。

我们首先通过使用平均池化和最大池化操作来聚合缅甸语图像的文本特征信息，生成两个不同的空间上下文特征描述：X'_{i1avg}和X'_{i1max}，分别表示平均池化特征图和最大池化特征图，其维度大小都为$C/2K \times 1 \times 1$，然后将这两个特征图分别送入两层的全连接神经网络，并且这个两层的全连接神经网络的参数是共享的，再将得到的两个特征图相加，通过Sigmoid函数得到0∼1之间的权重系数，得到最终输出通道注意力图为$M_c \in R^{C/2K \times 1 \times 1}$。其中，为了减少参数开销，共享网络的隐藏激活大小设置为$R^{C_t \times 1 \times 1}$，其中t为缩减率。简而言之，通道注意力权重计算如下:

$$\begin{aligned} M_c\left(X'_{i1}\right) &= \sigma\left(MLP\left(AvgPool\left(X'_{i1}\right)\right)\right) + MLP\left(MaxPool\left(X'_{i1}\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(X'_{i1avg}\right)\right)\right) + W_1\left(W_0\left(X'_{i1max}\right)\right) \end{aligned} \tag{1}$$

其中，σ表示sigmoid函数，$W_0 \in R^{C/t \times C}$，$W_1 \in R^{C \times C/t}$表示两个输入共享MLP的权重。

（3）空间注意力

不同的维度所代表的意义是不同的，它们本身所携带的信息也是不同的。相比较图像的通道信息而言，其空间所拥有的位置信息更为丰富。在实现方面，我们首先采用Group Norm(GN)对X_{i2}进行处理得到空间域层面的统计信息，然后采用$F_C(\cdot)$进行增强，得到空间注意力图为$M_s \in R^{C/2K \times H \times W}$，该过程可以描述如下：

$$M_s\left(X'_{i2}\right) = \sigma\left(W_2 \cdot GN\left(X_{i2}\right) + b_2\right) \cdot X_{i2} \tag{2}$$

其中，$W_2 \in R^{C/2K \times H \times W}$，$b_2 \in R^{C/2K \times H \times W}$。

（4）特征融合

在完成通道和空间注意力计算后，我们需要对其进行集成，首先通过简单的Concat进行融合得到：$M = \left[M_c\left(X'_{i1}\right), M_s\left(X'_{i2}\right)\right] \in R^{C/K \times H \times W}$，最后采用通道置换操作进行组间通信。

3.3 缅甸语图像特征表示

设缅甸语文本行的输入图像，图像的宽度可能具有任意长度，先用卷积神经网络ResNet对缅甸语图像进行处理，再利用通道和空间注意力网络对处理之后的结果进行特征增强，最后我们得到了一个大小为$H \times W \times C$的中间视觉特征表示F_C，这种视觉特征表示具有整个缅甸语输入图像的上下文化的全局表示，特征结构紧凑。缅甸语文本图像在本质上是连续的信号，缅甸语文的读取顺序是从左到右，为此我们视觉特征表示F_C转化为视觉特征向量$\{v_1, v_2..., v_w\}$，其中$v_i \in R^{C \times H}$。

我们采用Muti-Attention对视觉特征向量进行编码，由于输入视觉特征向量本身是缺乏位置信息，我们采用原始Transformer的位置编码方式对视觉特征向量进行位置编码。位置信息编码之前，维度大小为的视觉特征向量进行维度压缩，维度压缩方式为将其输入到一个全连接层实现维度转化，最终维度压缩之后视觉特征向量向量$\tilde{F}_C$的大小为(C, W)。为了有效地、明确地引导注意机制和让视觉向量$\tilde{F}_C$失去水平位移不变性，根据Vaswani et al. (2017)的研究，采用了基于正弦和余弦函数的位置编码。

$$TE\left(pos, 2i\right) = \sin\left(\frac{pos}{10000^{2i/f}}\right) \tag{3}$$

$$TE\left(pos, 2i+1\right) = \cos\left(\frac{pos}{10000^{2i/f}}\right) \tag{4}$$

其中，$pos \in \{0, 1, 2, ..., w-1\}$，$i \in \{0, 1, 2, ..., c-1\}$。

(c) 2022 中国中文信息学会计算语言学专业委员会

将$\tilde{F}_C$与位置编码进行融合得到向量$\hat{F}_c$，为了进一步提取视觉特征，在$\hat{F}_c$上应用了四次自注意模块。该注意模块输入为Q_c，K_c和V_c，其中$Q_c = K_c = V_c$。相关性信息计算方式如下：

$$\tilde{v}_c^i = Soft\max\left(\frac{q_c^i K_c}{\sqrt{c}}\right) V_c \tag{5}$$

其中，$q_c^i \in Q_c$，$i \in \{0,1,2,...,w-1\}$，$\tilde{F}_c = \left\{\tilde{v}_c^0, \tilde{v}_c^1, ..., \tilde{v}_c^{w-1}\right\}$，经过注意力计算得到增强之后的视觉特征$\tilde{F}_c$，用于后续的文字转录模块。

3.4 缅甸语文字转录

文字转录模块负责将视觉特征$\tilde{F}_c = \left\{\tilde{v}_c^0, \tilde{v}_c^1, ..., \tilde{v}_c^{w-1}\right\}$解码为字符，关注视觉特征以及从文本特征中学习到的语言特定知识。文字转录模块是由4个Tranformer解码器组成。选择Tranformer而不是基于RNN的体系结构的原因是，RNN结构在对当前时刻进行文字分类时依赖上一时刻不能实现并行计算。每个解码器层由三个子层组成：两个多头注意机制层和一个前馈神经网络组成。以前关于基于注意力机制的文字识别方法只在每个解码步骤的编码状态上使用一个注意力分布，相比之下，每个解码层我们采用多头注意力机制对编码器特征进行建模计算，并解决了解码时输出字符与编码特征之间的复杂对齐关系。

模型训练时采用交叉熵损失函数作为缅甸语识别模型的目标优化函数，计算方式如公式6所示：

$$Loss_{Att} = -\sum \ln P\left(\hat{y}_t | M, \theta\right) \tag{6}$$

其中，表示为M输入的缅甸语图像，θ表示为当前识别网络的模型参数，$\hat{y}_t|M$表示为缅甸语图像的第t个特征序列对应的真实标签。

4 实验结果及分析

为验证融合通道注意力和空间注意力的缅甸语图像文本识别方法的有效性，我们在缅甸语图像数据集上进行实验分析。

4.1 数据集及实验设置

由于缅甸语属于典型的资源稀缺性语言，目前没有公开的缅甸语文本图像数据集，因此本文将在自构的缅甸语数据集上来验证方法的有效性，该数据集总共包含了800万张缅甸语图像，数据集是由合成和人工标注的方法构建的，人工标注数据为3万张图片，剩余数据是通过合成算法得到的包含不同背景颜色、不同倾斜角度的缅甸语文本图像，以此增加训练样本的多样性。其中，分别随机选取20万缅甸语图像作为测试数据集和验证数据集。为提升模型训练速度，数据预处理阶段采用“.mdb”文件存储方式来存储训练集、测试集、验证集以此提高模型读取速率，具体规模如表1所示。

数据集	数量	样例	标签
训练集	800万		စိတ်ဆိုးသော
测试集	20万		ရယူလိုသော
验证集	20万		အရိပ်အမြွက်။

Table 1: 缅甸语图像数据集样例及对应标签实例

实验采用缅甸语序列率精确率（Sequence Accuracy, SA）作为评价指标，如公式7所示：

$$SA = \frac{SL}{LN} \times 100\% \tag{7}$$

其中，SA、SL、LN分别代表缅甸语文本图像识别的序列精确率、正确的序列总数、序列的总数。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 实验结果及分析

为验证融合通道注意力和空间注意力的缅甸语图像文本识别方法的有效性，我们在缅甸语图像数据集上进行实验分析。为保证对比实验的公平性，本文将所有的缅甸语识别模型放置在同一实验条件下进行实验，实验所选优化器为Adam，初始学习率为1，训练时采用CosineAnnealing策略，基于余弦函数实现学习率动态变换，以保证网络的目标函数接近最优解时具备更小的学习率；模型训练的批处理大小设置为200，训练步长设为700000，实验结果选择评测中最高的准确率，实验结果如表2所示。

实验一：主要实验结果及分析

本文在缅甸语数据集上进行了实验，并与以下的模型的实验结果进行了对比：

CNN+BiLSTM+CTC: (Shi et al., 2016)首先使用标准的CNN网络提取文本图像的特征，再利用BiLSTM将特征向量进行融合以提取字符序列的上下文特征，然后得到每列特征的概率分布，最后通过CTC进行预测得到文本序列

CNN+BiLSTM+Attention(Baek et al., 2019)：解码部分采用注意力解码器对序列进行解码。

毛等人(毛存礼et al., 2022)：构建了基于卷积神经网络和循环神经网络框架的教师网络和学生网络，以集成学习的方式进行训练的模型架构。

刘等人(Liu et al., 2021)：提出利用深度卷积网络获取并融合多层语义特征图，来缓解缅甸语图像文本识别过程中上下标字符特征丢失得问题，并采用MIX UP(Zhang et al., 2017)的训练策略。

方法类别	具体方法	SA(%)	Time(s)
联结主义时间分类的方法	CNN+LSTM+CTC	84.5	*
	CNN+BiLSTM+CTC	90.4	1250
序列到序列的方法	CNN+BiLSTM+Attention	90.6	16897
现有缅甸语图像文本识别的方法	谢等人	93.5	*
	刘等人	94.2	11560
基于Transformer的方法	Resnet+Transformer	94.8	1630
	Ours	95.3	1632

Table 2: 实验结果

如表2 所示，所提方法在缅甸语图像文本识别任务上准确率达到95.3%，达到了最高水平。相比联结主义时间分类的方法，提升了4.9%，说明本文方法能够获取更丰富的缅甸语图像文本特征信息，识别结果显示了明显的优势；相比序列到序列的方法，提升了4.7%，说明本文的方法在识别缅甸语的过程中提取到更为细粒度的缅甸语图像文本特征并进行特征图注意力计算，赋予了一些边缘特征更高的权重；相比已有缅甸语识别的方法，提升了1.1%，说明本文的方法缅甸语图像特征提取过程中更多地关注到缅甸语字符上下标等文字边缘特征，减少了缅甸语字符上下标丢失或错误识别的情况。

为了验证本文方法在缅甸语文本图像识别效率方面的提升效果，我们在相同的数据集和实验参数下对不同的方法进行了实验，并取平均每训练2000步长所需的时间作为对比结果。由表2的实验结果分析可知，本文方法大幅度缩短了训练时间，相比较刘等人的方法训练时间缩短将近7倍，与“CNN+BiLSTM+Attention”方法相比更是缩短到接近原来的十分之一，说明本文方法在能较好提高识别准确率的情况下，极大地提升了识别效率；同时与“Resnet+Transformer”相比训练时间相差无几，说明本文融合通道注意力和空间注意力模块的方法在几乎没有增加训练成本的前提下也能提升识别的准确率；此外，我们注意到“CNN+BiLSTM+CTC”的训练时间比本文方法更短，这是因为基于CTC的解码方式没有太多的针对图像上下文特征的注意力计算，考虑到本文方法的识别准确率相比“CNN+BiLSTM+CTC”有较大的提升，因此仍然能够说明方法的有效性与实用性。

为保证验证实验的真实性以及有效性，本文用人工标注的方式额外标注了1000张真实场景图像，并将其作为测试集。本文在这1000张真实场景测试集上进行测试实验，实验结果如表3所

(c) 2022 中国中文信息学会计算语言学专业委员会

示。

方法	SA(%)
CNN+LSTM+CTC	82.5
CNN+BiLSTM+Attention	89.7
CNN+BiLSTM+CTC	89.5
Ours	94.1

Table 3: 真实测试集上的实验结果

本文的方法在对1000张真实场景测试集图像的识别中仍然保持着最优的效果，同比基于注意力的识别模型的准确率能够提升4.4个百分点，融合通道注意力和空间注意力的方式能够帮助后续的缅甸语识别解码器获取更多的特征，利用丰富的缅甸语图像特征，解码器能够很大程度上提升准确率。

实验二：通道和空间注意力融合消融实验结果对比

为验证缅甸语通道和空间注意力融合策略的有效性，我们分别对其做了消融试验。我们分别对以ResNet为主干网络的基线模型进行消融实验，实验结果如表4所示（“✗”代表未融合，“✓”代表融合）

方法	Channel Attention	Spatial Attention	SA（%）
ResNet+Transformer	✗	✗	94.8
ResNet+Transformer	✓	✗	94.8
ResNet+Transformer	✗	✓	94.9
ResNet+Transformer	✓	✓	95.3

Table 4: 通道和空间注意力融合对识别的影响

如表4所示，其中Channel Attention表示通道注意力，Spatial Attention表示空间注意力，从实验结果可以看出，在只融合通道注意力或空间注意力中的情况下，以ResNet为主干网络的缅甸语图像识别模型性能提升非常小，但同时融合两种注意力时对模型的准确率可以提高0.5个百分点，说明同时对缅甸语图像的通道域和空间域做注意力计算并融合能够更充分关注到文本信息相关的特征。

实验三：针对注意力头数和的消融实验对比

为了验证多头注意力个数对识别模型的影响，我们对其进行了消融实验，实验结果如表5所示。其中，当注意力头数为6时，识别模型的性能最优。

注意力头数	SA(%)
2	94.5
4	94.8
6	95.3

Table 5: 注意力头数对识别的影响

实验四：针对视觉注意力单元和解码单元个数的消融实验对比

为了验证视觉注意力单元和解码单元个数对识别模型的影响，我们将注意力头数设为6，对单元个数进行了消融实验，实验结果如表6所示。其中，当单元个数为4时，识别模型的性能最优。

单元个数	SA(%)
2	94.6
4	95.3
6	95.0

Table 6: 单元个数对识别的影响

为了验证本文所提的注意力关注模块能够更加充分关注到缅甸语图像中文字所在区域，我们对其进行注意力可视化，注意力可视化结果如Figure4所示。我们的方法在识别缅甸语字符时能够有效关注到各个字符在图像中的位置，比如在Figure4的第一张图像中，模型在识别字符“ူ”和“ိ”时，对于图像中上标字符“ေ”以及下标字符“ျ”所在区域给予了较高的注意力权重，充分关注到缅甸语图像中的文字边缘特征，提高了模型对缅甸语字符序列的识别精度。

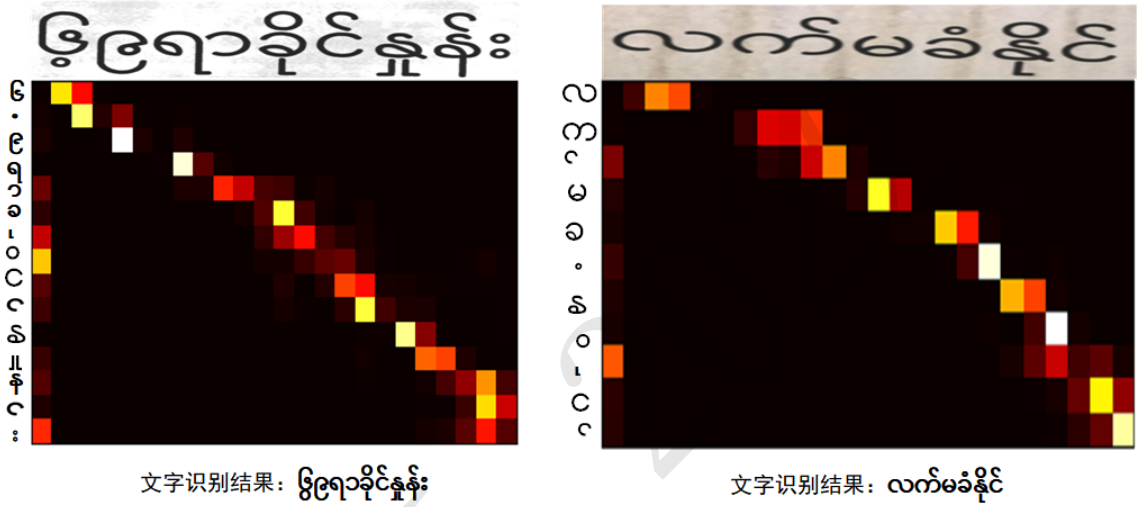


Figure 4: 字符识别注意力分布图

4.3 测试样例展示

表7 给出了缅甸语图像识别的实例。在针对有多个上下标的缅甸语嵌套字符图像的识别时，基于Resnet+transformer的识别模型会存在图像文字边缘特征的缺失，导致识别结果的上下标丢失，面对低质图像这类问题更为明显。而本文方法融合通道注意力和空间注意力的识别模型在面对低质或者组合字符数量多的缅甸语图像，有着更好的性能，能够保证低质图像下的识别准确率，同时缓解字符丢失问题。

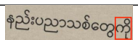
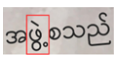
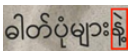
测试样例	Resnet+transformer	Ours
	နည်းပညာသစ်တွေကို	နည်းပညာသစ်တွေကို
	အဖွဲ့စသည်	အဖွဲ့စသည်
	ခါတ်ပုံများနှ	ခါတ်ပုံများနှ

Table 7: 测试样例及结果

(c) 2022 中国中文信息学会计算语言学专业委员会

5 结论

针对缅甸语图像文本识别中会存在上下标导致识别不佳的问题，提出了一种融合通道和空间注意力的缅甸语图像文本识别方法，将在通道域和空间域分别得到的注意力特征图融合后对原特征图进行重构，提高了模型对缅甸语图像文字边缘特征的提取能力。并在自构的缅甸语数据集的基础上进行了实验，相较于Resnet+Transformer的基线提升了0.5%，验证了所提方法的有效性。本文工作不仅缓解了缅甸语图像文本识别过程中字符上下标丢失的问题，还探索了类似缅甸语这类以音节为基本组成单位的低资源语言的语言特征在图像文本识别任务中所面临的问题和挑战，为其它类似的语言提供了较好的借鉴。在下一步工作中，在开展针对缅甸语这类具有复杂嵌套字符组合语言的图像文本识别的研究中，我们将进一步探索预训练模型以及Mask机制对其图像文本识别性能的影响。

参考文献

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. 2021. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34.

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.

Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084.

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*.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835.

Yunze Gao, Yingying Chen, Jinqiao Wang, and Hanqing Lu. 2017. Reading scene text with attention convolutional sequence modeling. *arXiv preprint arXiv:1709.04303*.

Alex Graves, Marcus Liwicki, Horst Bunke, Jürgen Schmidhuber, and Santiago Fernández. 2007. Unconstrained on-line handwriting recognition with recurrent neural networks. *Advances in neural information processing systems*, 20.

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.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28.

Chen-Yu Lee and Simon Osindero. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2231–2239.

Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8610–8617.

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.

(c) 2022 中国中文信息学会计算语言学专业委员会

Fenfen Sheng, Zhineng Chen, and Bo Xu. 2019. Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 781–786. IEEE.

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.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048.

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.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Lu Yang, Peng Wang, Hui Li, Zhen Li, and Yanning Zhang. 2020. A holistic representation guided attention network for scene text recognition. *Neurocomputing*, 414:67–75.

Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, and Cheng-Lin Liu. 2017. Scene text recognition with sliding convolutional character models. *arXiv preprint arXiv:1709.01727*.

Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12113–12122.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

毛存礼, 谢旭阳, 余正涛, 高盛祥, 王振晗, 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.

CCL 2022

基于预训练及控制码法的藏文律诗自动生成方法

色差甲[1,2]、慈祯嘉措[1,2]、才让加[1,2(✉)]、华果才让[1,2]
1. 省部共建藏语智能信息处理及应用国家重点实验室;
2. 青海省藏文信息处理工程研究中心，青海 西宁 810008。
sechajia@126.com czjcaiyaogun@hotmail.com
(✉)zwxxzx@163.com 65332395@qq.com

摘要

诗歌自动写作研究是自然语言生成的一个重要研究领域，被认为是极具挑战且有趣的任务之一。本文提出一种基于预训练及控制码法的藏文律诗生成方法。在藏文预训练语言模型上进行微调后生成质量显著提升，然而引入控制码法后在很大程度上确保了扣题程度，即关键词在生成诗作中的平均覆盖率居高。此外，在生成诗作中不仅提高词汇的丰富性，而且生成结果的多样性也明显提升。经测试表明，基于预训练及控制码法的生成方法显著优于基线方法。

关键词： 藏文律诗自动生成 ； 藏文预训练模型 ； 控制码法

Automatic Generation of Tibetan Poems based on Pre-training and Control Code Method

Secha Jia, [1,2] Cizhen Jiacuo [1,2] , Cairang Jia [1,2(✉)] and Huaguo Cairang [1,2]
1. The State Key Laboratory of Tibetan Intelligent Information Processing and Application;
2. Tibetan Information Processing Engineering Technology and Research Center of Qinghai Province，Qinghai Xining 810008.
sechajia@126.com czjcaiyaogun@hotmail.com
(✉)zwxxzx@163.com 65332395@qq.com

Abstract

The study of automatic poetry writing is an important area of study in natural language generation and is considered one of the most challenging and interesting tasks. In this paper, a method for generating Tibetan poems based on pre-training and control code methods is proposed. The quality of the generation was significantly improved after fine-tuning on the Tibetan pre-trained language model. However, the introduction of the control code method has largely ensured the degree of deduction, that is, the average coverage of keywords in the generated poems is high. In addition, the richness of vocabulary is not only improved in generative poetry, but also the diversity of generative results is significantly improved. Tests have shown that the generation method based on pre-training and control code methods is significantly better than the baseline method.

Keywords: Tibetan poems automatically generated , Tibetan pre training model , Control code method

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目： 青海省重点研发与转化计划项目 (2022-GX-104)、青海省科技厅项目 (2020-ZJ-704)、国家自然科学基金 (62166034)

1 引言

近年来，如新闻、作文、诗歌等的自动写作，日益得到人工智能学界的逐渐兴起。其中，诗歌是人类文化的瑰宝，其短小精悍的语言却能表达出极其丰富的含义和主题，从古至今吸引了无数爱好者的欣赏。诗歌自动写作研究是自然语言生成的一个重要研究领域，被认为是极具挑战且有趣的任务之一。尤其是中文古诗的自动写作，是自然语言生成中最引人注目的研究课题之一 (孙茂松, 2020; 矣晓沅, 2021)。

从生成技术方面，自 2014 年基于端到端框架 (Sutskever and Vinyals et al., 2014; Cho and Merrienboer et al., 2014) 提出之后，文本生成迅速成为研究热点。其中 Transformer 已成为文本生成领域很受青睐的模型之一，同时近期在预训练加微调的新兴训练范式中也得到了广泛的应用，如 BERT (Devlin and Chang et al., 2019)、GPT (Alec and Karthik et al., 2018)、T5 (Raffel and Shazeer et al., 2020) 模型等。然而对于藏语自然语言生成任务而言，除了汉藏机器翻译 (桑杰端珠, 2019; 慈祯嘉措等, 2019; 头旦才让, 2021)、复述生成 (柔特, 2019)、摘要生成 (李亮, 2020) 的技术相对成熟之外，其他生成任务的技术研究尚处于初步探索阶段。色差甲等 (2018; 2019) 人实现了基于端到端的藏文律诗生成模型，并通过实验发现，该方法虽然能够提升生成质量和诗行之间的语义连贯性，但是对于关键词的扣题程度有所欠缺，进而会出现一些主题漂移问题，同时无法生成具有多样性的藏文律诗。因此，有必要在生成多样性和扣题程度等方面进行进一步的改进和完善。

针对以上问题，本文将提出一种预训练语言模型和控制码法相结合的藏文律诗生成方法。本方法特点为：其一，模型结构简洁: 只是使用了一个结构简单且高效的 Transformer 模型。与基线模型相比，主要区别在于本文模型提前进行了预训练。其二，扣题程度更好: 用藏文律诗语料进行预处理时引入了控制码法，即每个关键词、诗行之间以及生成任务中存在特定的分割标记，有助于模型引导生成，从而很大程度上能够确保扣题程度，防止出现主题漂移问题。其三，生成结果多样化：在解码过程中采用新的采样方法，从而在相同的形式和主题下能够生成多样化的藏文律诗，其结果显著优于基线模型。其四，能够生成藏头诗：本文方法还能够生成藏文藏头诗，即给定每个诗行的首位的藏文音节（四个音节），便可以在相应的位置生成给定的音节，并且保证在生成结果中形式和质量的要求。

本文方法所生成的藏文律诗比较接近于人类创作的诗歌。如图 1给出了四首藏文律诗，其中一首是人类创作的，其余三首是由本文方法所生成的。

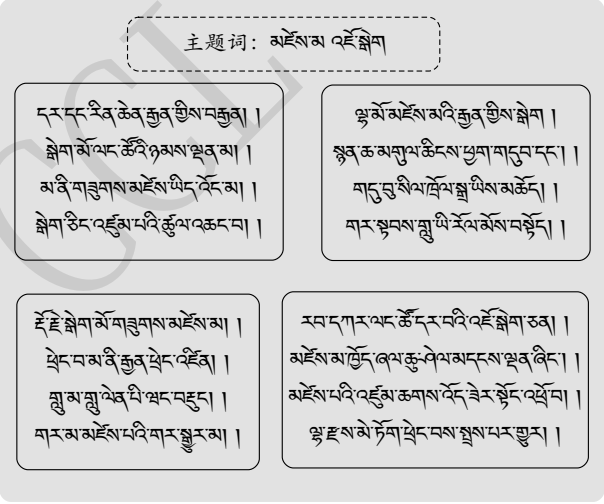


图 1: 一首真实和三首生成的藏文律诗

2 模型及数据预处理方式

为了能够生成句子更加流畅、语义更加连贯的一首藏文律诗，我们利用大规模的藏文文本语料预训练了语言模型。预训练语言模型将从综合性大规模文本语料中学到的词法信息、语法信息、以及语义信息等可直接迁移于藏文律诗生成模型。因此，该生成模型不是从零开始学习，而是在具备一定先验知识的基础上进行更进一步的学习。

第二十一届中国计算语言学大会论文集，第366页-第373页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

2.1 预训练模型及其数据处理方式

2.1.1 藏文预训练模型

T5 (Raffel and Shazeer et al., 2020) 和 BART (Lewis and Liu et al., 2019) 模型都是序列到序列的预训练降噪自编码器，与 BERT (Devlin and Chang et al., 2019) 相比有两个改变的点：第一种是在 BERT 的双向编码器架构中增添了因果解码器，即架构中包含编码器和解码器；另一种是用更复杂的预训练任务代替 BERT 的掩码语言模型任务。本文借鉴 BART 模型及其文本预处理方法，实现一个基于 Transformer 的藏文预训练语言模型，本文称之为 TiPLMT(Tibetan Pre-training Language Model based on Transformer)。TiPLMT 的数据处理及读取方式见图 2。从图 2中可知，TiPLMT 的源端输入是利用文本增强方法增强之后的文本（对原始文本进行增

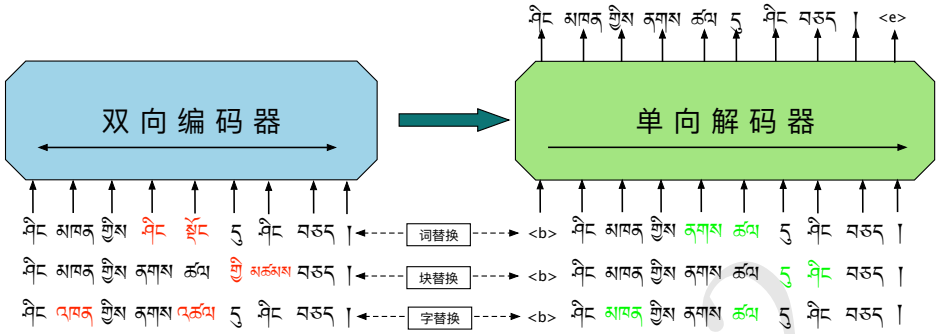


图 2: TiPLMT 模型及其数据读取格式

强的结果)，而目标端的输入便是原始文本。其目的便是通过语义信息不完备、缺失或含噪的文本，重构成语义完整且流畅的文本，使得模型不仅具有语言的表征能力，同时还具有文本错误纠正的能力。另外，TiPLMT 模型中弃用了特殊符号 [MASK] 的遮蔽机制，而是利用了藏文文本数据增强方法，也就是基于音节混淆子集和基于上下文的增强方法。从字（音节）、块（连续的片段，即 n-gram）、词等三个不同的层面对文本进行了增强并训练。

2.1.2 针对 TiPLMT 的文本预处理

预训练语言模型之前，从大规模无标注文本语料中通常以字、词或块为基本单位进行掩码(或替换) 处理来自动构建监督学习式信息，以便充分利用无标注文本语料。对于不同任务或不同语种而言，句子切分的颗粒度也不一样，预训练语言模型中把英语句子会切成词元 (Devlin and Chang et al., 2019)，汉文句子会切成字 (Cui and Che et al., 2021)，而本文将藏文句子切成音节。TiPLMT 从藏文文本语料中自动构建监督信息的方式进行了以下四种方法:

(1) 弃用特殊符号“[MASK]”的掩码方法，而是使用基于音节混淆子集和基于上下文的增强方法，从而可以把含加噪的句子和原句分别作为源句和目标句进行训练。

(2) 每个藏文句子中只处理 15% 的音节，其中不包含音节分隔符、垂直符、数字、特殊符号以及非藏文字符等，可又分为三种情况并分别为:

- 块层面的处理方式：先从原始文本中随机选取多个连续的音节（需要保证含有 15% 的音节），再随利用基于上下文的增强方法进行加噪处理其中单音节、双音节以及三音节所占比率分别为 20%、30% 和 50%;
- 音节层面的处理方式：先从原始文本中随机选取 15% 的音节，再利用基于音节混淆子集的增强方法进行处理，藏文虚词和实词所占比例分别为 40% 和 60%。
- 词层面的处理方式：先从已分词的原始文本中随机选取多个词（需要保证含有 15% 的音节），再随利用基于上下文的增强方法进行加噪处理，其中单音节、双音节以及多音节所占比率分别为 30%、50% 和 20%;

其中，音节混淆集合摘译藏文正字法，即由相互容易混淆的音节组成，共整理了 3912 个藏文音节，而且均是正式语料中出现频率相对较高的音节。基于上下文的增强方法的具体处理步

(c) 2022 中国中文信息学会计算语言学专业委员会

骤为：先对藏文原句中随机选取某个词或块，并用特殊符号 [MASK] 进行替换；然后利用基于上下文的增强方法进行重构；最后筛选出未能完全重构正确的句子，并与原句作为训练句对。藏文文本数据增强结果表 1 所示。

表 1: 藏文文本数据增强结果的示例

类型	藏文原句	重构的句子
词掩码	གང་ལ་[MASK]མེམས་དང་ལྡན་པ་དེ་ནི་མི་བཟང་བོ་འོ།། འདི་ལྟར་སྨོན་སྡུང་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ལ་ཤིན་ཏུ་དགའ་བ་ཞེག་ཏུ་སྐྱེ།	གང་ལ་[MASK]དང་ལྡན་པ་དེ་ནི་མི་བཟང་བོ་འོ།། འདི་ལྟར་སྨོན་སྡུང་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ལ་མི་དགའ་བ་ཞེག་ཏུ་སྐྱེ།
块掩码	ང་ལ་གནེར་སྡང་བརྒྱ་ཡོད་ན་ཅི་མ་རུང་། [MASK] ཡོངས་ལ་ཤིན་ཏུ་དགའ་བ་ཞེག་ཏུ་སྐྱེ།	ང་ལ་གནེར་ཁྲི་ཞེག་ཡོད་ན་ཅི་མ་རུང་། [MASK] ཡོངས་ལ་དགའ་བ་མང་བ་ཞེག་ཏུ་སྐྱེ།
字替换	རྟོམ་ཁ་འི་ཚོ་ན་ཤེས་བྱ་འི་རྣམ་གྲངས་ཟད།། འདི་ལྟར་སྨོན་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ལ་ཤིན་ཏུ་དགའ་བ་ཞེག་ཏུ་སྐྱེ།	རྟོམ་བ་འི་ཚོ་ན་ཤེས་བྱ་འི་རྣམ་གྲངས་ཟད།། འདི་ལྟར་སྨོན་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ཤིན་ཏུ་དགའ་ཁ་ཞེག་ཏུ་སྐྱེ།

从表 1 中可知，字替换是由混淆子集完成的，不仅能快速重构，而且根据藏文正字法能仿造含有真字拼写错误的句子，与正式文本中出现拼写错误现象一样，具有逼真的效果，但只能仿造音节级别的噪声；词或块替换是由提前预训练好的小模型完成的，因此重构速度相对较慢，但能仿造出含有语法或语义错误的句子，同样具有逼真的效果。原句中蓝色标记的音节是待替换的词（块），加噪句子中红色标记的音节是模型预测的音节，或者是用混淆子集随机替换的音节。

2.2 模型微调及控制码法

2.2.1 模型微调方式

我们的主要任务是在预训练模型 TiPLMT 的基础上，利用藏文律诗的语料进行进一步的微调。在微调过程中，将主题信息和藏文律诗以控制码法的方式嵌入到生成模型中，从而有效增强了模型的扣题程度，提高了生成结果的语义连贯性。模型微调方式及藏文律诗的输入格式如图 3所示。

图 3中，橙色部分表示编码器，蓝色部分表示解码器。关键词序列 S 作为编码器的输入（即源序列），藏文律诗序列 T 作为作为解码器的输入（即目标序列）。此处的目的便是微调一个条件自回归语言模型。

2.2.2 模控制码法的应用

控制码法 (Control Code)：是一种简单且有效的控制方法，即将所需要的控制指令，以字符的方式输入模型并作为生成的条件，该方法广泛应用于 BART、GPT-3 等预训练模型中。在汉文古诗生成方面，张家瑞等人 (2021) 把诗词序列化转化为由格式、主题和诗体等组成统一格式化的文本序列，作为训练数据并在 GPT 模型上进行微调，在绝句、律诗、藏头诗、词以及对联等生成任务上表现突出；Liao 和 Wang 等人 (2019) 在 GPT-2 模型中融入隐狄利克雷分配 (Blei and Ng et al., 2001) 模型的方法来实现了主题可控的诗歌生成，同时检验了主题模型 LDA 的有效性。

基于上述的研究成果，我们将结合藏文律诗的特征，实现一个预训练语言模型的基础上融入控制码法的藏文律诗生成方法。语料的主要处理格式见图 3所示，每首藏文律诗都按统一格式进行处理，是以“[Top] 主题词 1 [Top1] ... 主题词 n [Topn] [BOS] 藏文律诗 [EOS]”的格式进行预处理。由于 TiPLMT 模型进行预训练时，在藏文音节的粒度上为完成训练的，所有微调时同样把藏文律诗的切分粒度选为藏文音节，即每首藏文律诗切分成藏文音节。在具体处理过

(c) 2022 中国中文信息学会计算语言学专业委员会

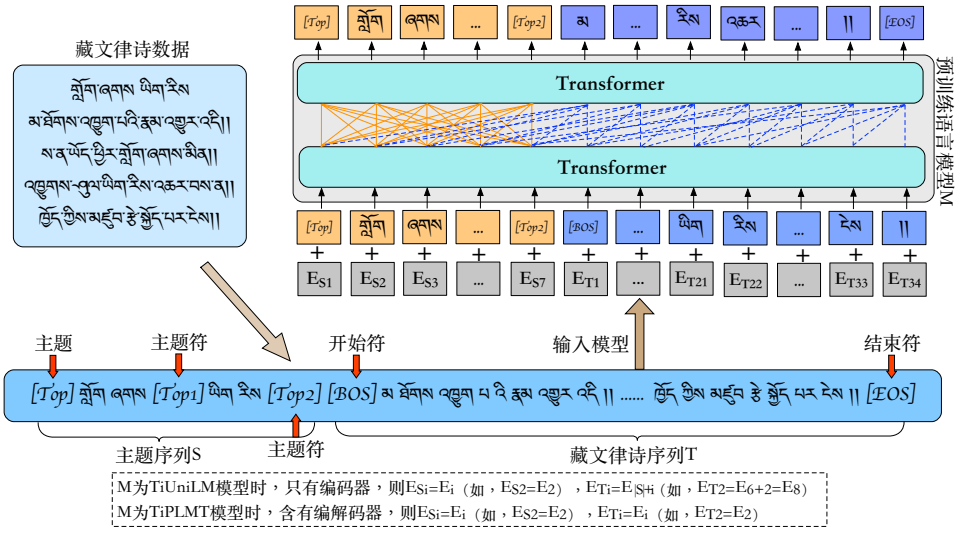


图 3: 模型微调方式及藏文律诗的输入格式

程中，我们将每个藏文律诗及其主题词用标识符进行标记和分割，并注入生成模型中作为生成的条件文本。对每首藏文律诗可以取 n 个主题词，则相对应的主题标识符也有 n 个（如 [Top1]，[Top2]，...，[Topn]）。本文主题词的取值范围设定为 $1 \leq n \leq 4$，其中含有一个、两个、三个和四个主题词的所占比例都一样，各占 25%（是人为设定的）。

3 实验

3.1 数据来源及规模

本文从藏文电子书籍及网页中共获取了含有 46.55 亿字符（4.53 亿藏文音节）的藏文文本语料，其主题包括文学、自传、诗歌、格言、散文和新闻等。对该语料进行了清洗、音节切分、音节拼写检错和纠正、以及句子切分等等预处理工作，最后共获得了 7.6 千万余藏文句子。其中，通过藏文律诗中垂直符的使用规律和诗行长度一致性等特征从中抽取去 131.3 万余首藏文律诗。据统计分析得出，藏文律诗中大部分为七言和九言律诗，共占 94.2%。首先从每首藏文律诗中通过关键词抽取算法 TextRank (Mihalcea and Tarau, 2004) 来抽取若干个关键词; 然后预处理成上述的格式要求; 最后从中抽取 2.5 千首藏文律诗对作为测试集，剩余部分作为训练集。

3.2 参数设置

本文的实验是在开源代码 Transforms1 的基础上完成的，藏文预训练语言模型 TiPLMT 的具体参数设置详见表 2 所示。

表 2: 模型参数设置情况

层数	词嵌入	全连接维度	注意力头数	优化器	学习率	最长序列	词表大小	参数规模
10	512	1024	10	Adam	3.8e-05	100	8663	4.7 千万

表 2 中，词表摘自《藏文规范音节频率词典》（多拉和扎西加, 2015）一书中，是从大规模的藏文文本中统计和整理的，只含有具有实际意义的音节，以及部分梵文。

3.3 基线方法与评测指标

由于目前面向藏文文本生成领域的相关研究较少。所以无法直接与前人的工作进行对比来验证本文所提出方法可行性和有效性，只能与重新复现的多个生成模型相比较。基于上述原因，本文将选用神经网络中常用于生成任务的几个经典模型作为基线模型，并分别为基于完全注意

(c) 2022 中国中文信息学会计算语言学专业委员会

力机制的 Transformer 模型 (Vaswani and Shazeer et al., 2017) 和基于生成式的预训练语言模型 GPT (Alec and Karthik et al., 2018)。

评价指标方面：将采用自动评测方法，从生成质量和生成多样性两方面进行评测。其中，注重生成质量方面的自动评价指标采用 PPL 和 BLEU 值，而注重的多样性的自动评价指标将采用 JS 值和 Distinct 值。

3.4 实验结果

由于以观察不同生成模型的有效性为目的，分别考查了 TiPLMT 以及基线模型等在藏文律诗的生成效果，是从语言建模能力和生成结果多样性方面进行评测分析，其对比实验结果如表 3 所示。

表 3: 不同模型在生成质量及多样性方面的对比实验结果

模型	PPL↓	BLEU↑(%)	Distinct↑(%)	JS↓(%)
Transformer	15.05	41.09	54.94	2.28
GPT	13.73	47.19	83.07	1.97
未用控制码法	15.82	46.58	79.96	2.02
未进行预训练	17.73	43.09	57.39	2.41
TiPLMT	9.28	51.02	94.46	1.15
未用控制码法	10.93	48.72	91.03	1.23
未进行预训练	15.98	46.05	63.74	2.31

从表 3 中可以看出，基线模型 Transformer 但在其他指标上则不然，这表明极限模型的生成结果中还是缺乏词的汇多样化使用。与基线模型 GPT 相比，TiPLMT 模型得益于语言模型预训练时融引入了文本数据增强技术和基于端到端的框架天然生成的优势，从而本文方法在生成藏文律诗的整体效果上仍获得更佳效果。TiPLMT 模型在 BLEU 值和 Distinct 值最高分别提升了 9.93 和 40.02 个百分点，JS 值降低了 1.13 个百分点，同时 PPL 取得了最低分数。这足以表明本文模型 TiPLMT 在生成质量、语句通顺度、以及词汇使用率等方面相对突出。此外，对于 TiPLMT 和 TiUniLM 模型而言，控制码法的使用和未使用之间存在显著差异，在 BLEU 值上分别能够提高 0.55 和 0.55 个点。同样，预训练语言模型的使用和未使用之间也存在显著差异，在 JS 值上分别降低 0.55 和 0.55 个点，在 PPL 值上分别降低 0.55 和 0.55 个点。

3.4.1 关键词数目对模型性能的影响

为了进一步观测各个模型的扣题能力，并统计不同数目的关键词在生成结果中的涵盖率，其统计结果如图 4所示。

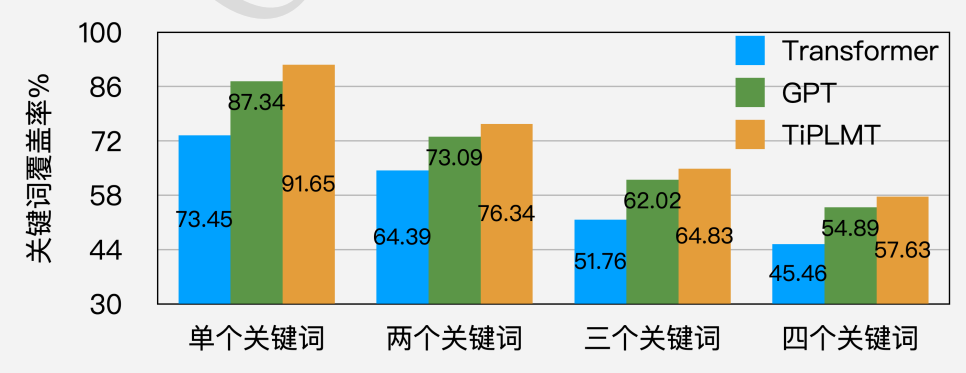


图 4: 生成结果中关键词的完全覆盖率

从图 4中可以看出，基线模型 Transformer 和 GPT 的关键词的覆盖率均不如 TiPLMT。对于 TiPLMT 模型而言，输入单个、两个、三个和四个关键词时，该模型的关键词完全包含率分

(c) 2022 中国中文信息学会计算语言学专业委员会

别为 91.65%、76.34%、64.83% 和 57.63%，则平均覆盖率高达 72.61%。显然，大部分关键词都能以某种形式在生成的藏文律诗中得到体现。

3.4.2 实例分析

本文方法有效提升了藏文律诗生成结果的多样性和扣题程度。下面我们将给出 TiPLMT 模型具体生成的诗作并进行分析，其部分生成实例如图 5 所示:

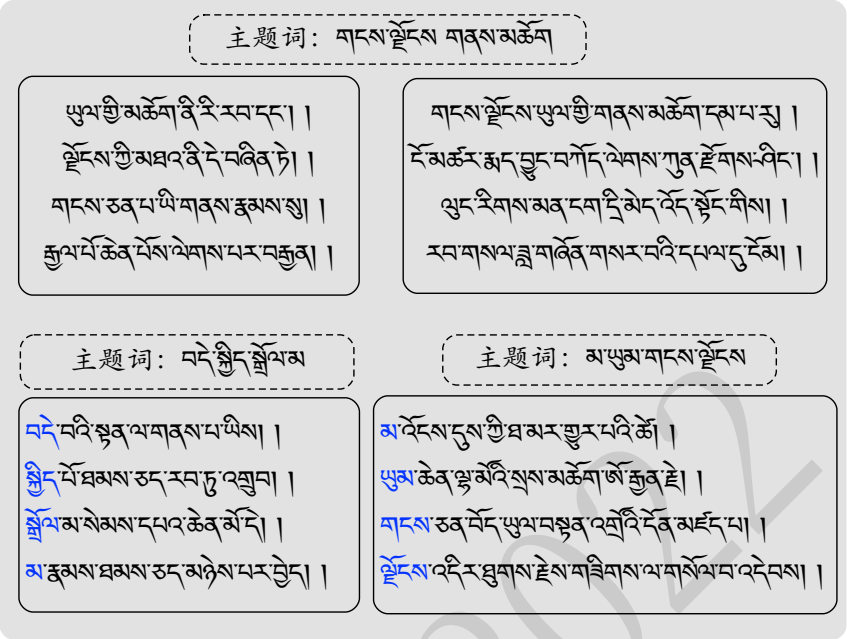


图 5: TiPLMT 模型的部分生成实例

从图 5-10 中可以看出，本文模型 TiPLMT 的生成质量不仅有所提升，而且能够生成多样化的藏文律诗。如给定主题词 “གངས་ལྗོངས་” 和”གནས་མཆོག” 时，该模型分别生成了七言和九言不同风格的藏文律诗，其中前者的音律节奏是 “ཚ+ཚ+ཚ+ཡ” 的类型，后者的音律节奏为 “ཚ+ཚ+ཚ+ཚ+ཡ” 的类型，这两个类型都是比较常用的音律节奏类型。以词为单位计算时，主题词在第一首实例中未出现，但以音节为单位时，所有主题词的音节均涵盖于该生成结果中，而且该音节在生成结果中的语义表现与主题词是一致的。主题词在第二首中是完全涵盖于生成结果中。此外，TiPLMT 模型还能够生成藏文藏头诗，如给定了四个音节的主题词“བདེ་སྐྱིད་སྐོལ་མ” 和 “མ་ཡུམ་གངས་ལྗོངས” 时，该模型分别生成了七言和九言的藏文藏头诗，其律诗节奏分别为 “ཚ+ཚ+ཚ+ཡ” 和 “ཚ+ཚ+ཚ+ཚ+ཡ”，可是，后者的第三行中出现了 “ཚ+ཚ+ཚ+ཡ+ཚ” 的律节奏类型，从而对整体朗读时稍微会影响顺口悦耳。

4 总结

为了提升扣题程度，我们提出了一种基于预训练语言模型和控制砝码相结合的生成方法。与基线模型相比，使用预训练语言模型后对于生成质量有显著提升。然而在藏文律诗语料进行预处理时引入了控制码法，即每个关键词、诗行之间以及生成任务中存在特定的分割标记，这有助于模型引导生成，从而在很大程度上确保了扣题程度，关键词的平均覆盖率高达 72.61%。

生成结果多样化方面也显著提升，一是提升了词汇使用的多样化，由于生成模型的初始参数源自预训练语言模型，从而降低了高频音节的重复使用率，更接近人类书写中高频音节的使用分布，最高频率的前 50 个藏文音节占了所生成内容的 30.5%（基线模型占了 41.3%，人类的占了 16.9%）；二是生成结果的整体多样化，我们在具体解码时采用了新的采样方法，从而在相同的形式和主题下能够生成高质且多样化的藏文律诗，其结果显著优于基线模型。

参考文献

Alec R, Karthik N, and Tim S, et al. 2018. *Improving Language Understanding by Generative Pre-Training.* https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language_unsupervised/language_understanding_paper.pdf.

Blei D M, Ng A Y, and Jordan M I. C. 2001. *Latent Dirichlet Allocation.* Advances in Neural Information Processing Systems(NIPS2001).

Cho K, Merrienboer B V, and Gulcehre C, et al. 2014. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734.

Cui Y, Che W, and Liu T, et al. 2021. *Pre-Training With Whole Word Masking for Chinese BERT.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 3504-3514.

Devlin J, Chang M-W, and Lee K, et al. 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 (NAACL), 4171–4186.

Lewis M, Liu Y, and Goyal N, et al. 2019. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.* arXiv preprint arXiv:1910.13461v1.

Liao Y, Wang Y, and Liu Q, et al. 2019. *GPT-based Generation for Classical Chinese Poetry.* arXiv preprint arXiv:1907.00151v5.

Mihalcea R, Tarau P. 2004. *TextRank: Bringing order into text.* Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), 404-411.

Raffel C, Shazeer N, and Roberts A, et al. 2020. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research, 21(140): 1-67.

Sutskever I, Vinyals O, and Le Q V. 2014. *Sequence to Sequence Learning with Neural Networks.* Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), 3104–3112.

Vaswani A, Shazeer N, and Parmar N, et al. 2017. *Attention is All You Nee.* Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS2017), 6000-6010.

慈祯嘉措, 桑杰端珠, 孙茂松, 等. 2019. 融合单语语言模型的藏汉机器翻译方法研究. 中文信息学报, 33(12): 61-66.

多拉, 扎西加. 2015. 《藏文规范音节频率词典》. 北京: 中国社会科学出版社.

李亮. 2020. 基于 *ALBERT* 的藏文预训练模型及其应用. 兰州大学.

柔特. 2019. 藏文陈述句复述生成研究. 青海师范大学.

桑杰端珠. 2019. 稀疏资源条件下的藏汉机器翻译研究. 青海师范大学.

色差甲. 2018. 基于神经网络的藏文律诗生成研究. 青海师范大学.

色差甲, 华果才让, 才让加, 等. 2019. 注意力的端到端模型生成藏文律诗. 中文信息学报, 33(04): 68-74.

孙茂松. 2020. 诗歌自动写作刍议. 数字人文, (00): 32-38.

头旦才让. 2021. 汉藏神经机器翻译关键技术研究. 西藏大学, 61-73.

矣晓沅. 2021. 具有文学表现力的中文古典诗歌自动写作方法研究. 清华大学.

张家瑞, 李文浩, 孙茂松. 2021. 基于 *BPE* 分词的中国古诗主题模型及主题可控的诗歌生成. 第二十届中国计算语言学大会论文集 (CCL2021), 862-873.

CCL 2022

基于词典注入的藏汉机器翻译模型预训练方法

桑杰端珠[1,2]　　才让加[1,2]

[1] 青海师范大学，计算机学院，西宁，810000

[2] 青海师范大学，藏语智能信息处理及应用国家重点实验室，西宁，81000

sangjeedondrub@live.com　　zwxxzx@163.com

摘要

近年来，预训练方法在自然语言处理领域引起了广泛关注，但是在比如藏汉机器等低资源的任务设定下，由于双语监督信息无法直接参与预训练，限制了预训练模型在此类任务上的性能改进。考虑到双语词典是丰富且廉价的先验翻译知识来源，同时受到跨语言交流中人们往往会使用混合语言增加以沟通效率这一现象启发，本文提出一种基于词典注入的藏汉机器翻译模型的预训练方法，为预训练提供学习双语知识关联的广泛可能。经验证，该方法在藏汉和汉藏翻译方向测试集上的 BLEU 值比 BART 强基准分别高出 2.3 和 2.1，证实了本文所提出的方法在藏汉机器翻译任务上的有效性。

关键词： 藏汉 ； 机器翻译 ； 预训练 ； 词典注入

Dictionary Injection Based Pretraining Method for Tibetan-Chinese Machine Translation Model

Sangjie Duanzhu[1,2]　　Cairangjia[1,2]

[1] School of Computer Science, Qinghai Normal University, Xining, 810000

[2] The State Key Laboratory of Tibetan Information Processing and Application, Qinghai Normal University, Xining 810000

sangjeedondrub@live.com　　zwxxzx@163.com

Abstract

In recent years, pretrained models have attracted extensive attention in the field, however, due to bilingual supervision can not directly participate in the pretraining process, pretrained models are not contributive under low-resource settings such as Tibetan-Chinese machine translation. Given bilingual dictionaries are rich and low-cost source of prior translation knowledge and inspired by the phenomenon that people often use mixed lexicons for better communication in cross-lingual conversations, this paper proposes a technique to pretrain the Tibetan-Chinese machine translation model via dictionary injection, which provides a wide range of possibilities for bilingual knowledge interaction. Empirical results show the proposed method can produce improvements of 2.3 and 2.1 in BLEU scores on test set for Tibetan-Chinese and Chinese-Tibetan translation directions over strong BART baselines, indicating the effectiveness of the proposed method.

Keywords: Tibetan-Chinese , Machine Translation , Pretraining , Dictionary Injection

©2022 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

基金项项目：青海省重点研发与转化计划项目 (2022-GX-104) 、青海省中央引导地方科技发展资金项目 (2022ZY006)

通信作者：才让加 (zwxxzx@163.com)

1 介绍

目前神经机器翻译 (Neural Machine Translation, NMT) (Sutskever et al., 2014; Gehring et al., 2017; Ashish et al., 2017) 已经成为最主流机器翻译方法，在性能上全方位超越传统短语统计翻译模型 (Statistical Machine Translation, SMT) (Brown et al., 1990)，并成为工业界机器翻译服务系统的标准实现方法 (Wu et al., 2016), 甚至研究者声称在特定领域和语言对上 NMT 的性能可以接近甚至超越人类的翻译水平 (Hassan et al., 2018)。与 SMT 不同的是 NMT 以端到端风格的建模方式将翻译决策过程视为单个条件概率模型的参数估计过程，从而摒弃了 SMT 不同组件独立优化各自训练目标的建模范式。但是目前 NMT 卓越的性能表现是以具备大规模、高质量和多领域对齐数据为重要前提的，受制于市场规模较小、数据标注成本高昂等客观因素，现阶段藏汉机器翻译的质量距离中英等主流语言存在巨大的差距。

在对齐数据受限的条件下，对于多数语言而言，单语数据的来源相对较为广泛且容易收集，研究者自然地探索了各类在 NMT 框架内有效利用目标端和源端单语数据的方法。其中最简单和直接的是回译方法 (Sennrich et al., 2016), 该方法利用监督式方法训练一个初始的反向模型，将目标端的单语数据进行翻译，用于扩充训练正向模型的数据。回译方法不仅能改善低资源场景下的翻译性能，同时在富资源场景中也能缓解领域适应等问题 (Kumari et al., 2021)。回译方法要求初始回译模型本身有较高的性能，但是在现实中很多低资源语言的对齐数据无法保证初始回译模型的性能。

近年来，受到计算机视觉研究的启发 (He et al., 2016), 在未标注的海量文本数据、高阶的分布式优化方案、强大的序列学习模型和高性能计算加速设备的共同加持下自监督式预训练 (Self-supervised Pretraining) 模型 (Devlin et al., 2019; Brown et al., 2020; Liu et al., 2019) 激起了自然语言处理 (Natural Language Processing, NLP) 领域内的研究热潮。预训练模型使得研究者可以不用从头训练昂贵和复杂的大规模模型，直接使用现有预训练模型在下游的目标任务上结合任务自身特点进行微调，就往往可以获得比监督式训练更好的性能表现。在诸多的预训练模型中具有代表性的包括掩码语言模型 (Masked Language Model, MLM) BERT (Devlin et al., 2019); 自回归语言模型 (Autoregressive Language Model, ALM) GPT (Radford et al., 2019); 置换语言模型 (Permuted Language Model, PLM) XLNet (Yang et al., 2019); 降噪自编码器模型 (Denoising Auto Encoder, DAE) BART (Lewis et al., 2020) 等。其中 BERT 和 XLNet 语言模型是 Transformer (Ashish et al., 2017) 的编码器，能对语言序列进行双向的表示学习，主要用于序列的语义理解。GPT 使用了 Transformer 的解码器，结合已生成的解码片段和当前时刻的输入，以自回归的方式逐词生成目标序列。而 BART 可以视为结合 BERT 和 GPT 泛化的预训练模型，与 BERT 和 GPT 不同的是，BART 采用序列到序列的建模方式，使用单个 Transformer 模型对编码器端完成各类加噪操作的输入序列在解码器端完成重构，通过降噪自编码为优化目标完成整个解码器和编码器的联合预训练，然后在下游的目标任务上通过标注数据进行微调，非常适合于机器翻译和知识问答等采用编码器-解码器构架的建模任务。BART 是针对单一语言 (英语) 的预训练，而随后提出的 mBART (Liu et al., 2020) 则是将 BART 的建模方式扩展到多语言场景下，完成多语言模型的预训练。同样是采用 BART 训练目标的 M2M-100 (Fan et al., 2021) 更是进一步扩大了所覆盖的语言种类，支持 100 个语言之间的多对多翻译。对于藏文这种低资源语言而言，多语言预训练是一个非常具有吸引力的设想，因为除了支持多语言翻译外，M2M-100 级别的大规模预训练模型本身能够有效支持通用语义知识的迁移。但是 mBART 和 M2M-100 的训练都没有包含藏文。本文旨在探索训练 BART 风格的藏汉翻译预训练模型的有效方法，为后续的藏语多语言翻译课题提供研究基础。

BART 在预训练过程中主要学习当前输入语言的表示和分布，缺乏双语对齐监督信号的直接参与，没有显式地学习语言对之间的映射关系。这种预训练方式不利于平行资源匮乏的藏汉语言对的预训练效果。考虑到双语词典是重要的先验知识来源，人类语言学者在学习一门新语言时，往往会

第二十一届中国计算语言学大会论文集，第374页–第383页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

借助双语词典探索所要学习的语言，通过词典建立新语言和其他已掌握的语言之间的关联；此外人类翻译人员也会使用双语词典推敲用词、查询专业词汇，以改善翻译工作的质量。另外，受到跨语言交流过程中使用混合语言往往能够增加沟通效率 (Matras, 2000) 这一现象的启发，本文提出了一种基于双语词典注入的藏汉预训练翻译模型的训练方法，即基于词典注入的藏汉机器翻译预训练模型 (Pretrained Translation Model with Dictionary Injection, PTMDI)。通过构建较大规模双语词典，然后利用词典对大规模的藏汉单语数据进行跨语言数据注入，以降噪自编码为训练目标完成藏汉机器翻译模型的预训练。词典的数据注入如表 1 所示。

表 1: 词典注入样例

原始输入	当代中国，江山壮丽，人民豪迈，前程远大。新时代为我国文艺事业发展提供了前所未有的广阔舞台。
词典替换	དེང་རབས་ གྲུང་གོ་ ，江山壮丽，མི་དམངས་ 豪迈，前程远大。གསར་པ་ དུས་རབས་ 为 རང་རྒྱུད་ 文艺 ལས་དོན་འཛེལ་རྒྱས་ 提供了前所未有的广阔 གངས་སྤེལ་གས་ 。

被替换的汉文词都使可以被视为是一种对原始文本的加噪，另外由于藏汉两种语言在语序上的偏差，比如藏文的定语普遍后置，新时代的正确翻译是 དུས་རབས་ (时代) གསར་པ་ (新)，进行词典替换之词序变为 གསར་པ་ དུས་རབས་ ，但是这种词序颠倒可视为额外的加噪操作。基于掩码的降噪自编码的训练中，遮蔽连续的词比单独的词（比如 BERT）能使模型学习到更好的表示 (Joshi et al., 2020)。同样地, 在基于词典注入的预训练方法中，连续词条的替换（比如 དེང་རབས་ (当代) གྲུང་གོ་ (中国)），同样符合这种思想。与 BART 中的加噪方式（见图 1）不同的是，通过词典替换的加噪方案，在加噪的同时为模型提供一个学习双语知识关联的广泛可能，客观上要求模型在联合学习双语语义对齐信息和序列上下文进行以降噪为目标的解码，学习到跨语言的联合表示，为在平行数据受限的条件下进行有效进行微调提供了便利。此外词典注入的预训练方法能够借助词典学习到目标领域的翻译知识，对机器翻译的领域适应提供了一个可行且低廉的可选方案。

在规模分别为 6.9 M 和 5.2 M 句子规模的藏汉单语数据、500 K 句对的藏汉平行数据和 314 K 词条双语词典的数据设定下，本文中的 PTMDI 模型在藏汉和汉藏翻译方向的测试集上的 BLEU 值比 BART 这一强基准模型分别高出 2.3 和 2.1，充分证实了本文所提出的预训练方法在藏汉机器翻译任务上的有效性。

综上，本文的贡献为：

1. 考虑到双语词典能在预训练过程中提供有效的监督信号，同时受跨语言交流中使用混合的多语言词汇能提高沟通效率这一现象启发，提出一种利用藏汉双语词典和藏汉单语数据进行词典注入的机器翻译预训练方法，即 PTMDI;

2. 在通过与包括监督式 Transformer, 回译，BART 的性能对比实验，证实本文提出的 PTMDI 方法比各类基准模型在测试数据集上都有大幅的性能提升；

3. 由于使用了藏汉双语词典，本文的提出的 PTMDI 模型适用于翻译模型的领域适应问题，能够借助领域词典和单语数据学习平行数据中缺乏的翻译知识。

2 相关工作

近年来随着人工智能领域技术的迅猛发展和日益密切的跨语言交流需求，藏汉机器翻译技术取得了长足发展。和其它低资源机器翻译研究课题一样，藏汉机器翻译的研究集中在致力于在平行数据资源受限的条件下探索提高机器翻译性能的方法。其中包括优化藏汉翻译模型的词表大小和分布 (孙义栋 et al., 2022; 头旦才让 et al., 2020), 利用大规模单语数据进行迭代式回译 (慈祯嘉措 et al.,

第二十一届中国计算语言学大会论文集，第374页-第383页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

2020)，迁移学习 (李亚超 et al., 2017), 融合藏文多层次先验特征 (沙九 et al., 2020), 融合目标端语言模型的方法 (慈祯嘉措 et al., 2019) 等。此外还有一些与藏文预训练语言模型相关的研究工作，比如 中国少数民族预训练语言模型 CINO (Yang et al., 2022)。该模型使用了 XLM-R (Conneau et al., 2020) 风格的预训练方法，是至迄今为止规模最大的支持藏文的公开跨语言预训练语言模型。CINO 虽然只在文本分类任务上进行了测试和验证，由于该模型可以进行跨语言的表示，可以用于初始化藏汉机器翻译的解码器、编码器或者整个模型的参数。

3 方法

NMT 给定源端句子 $x = \{x_1, \cdots, x_N\}$ 和目标端句子 $y = \{y_1, \cdots, y_M\}$，NMT 将句子级别的翻译概率建模问题转换为词级别的条件概率的积,

$$P(y|x;\theta) = \prod_{j=1}^{M} P(y_i|x, y_{<i};\theta) \tag{1}$$

其中 θ 为模型所要估计的参数，$y_{<i} = \{y_1, \cdots, y_i\}$ 为 i 时刻已生成的目标序列片段。在 Transformer 构架中序列转导 (Sequence Transduction) 建模任务由自注意力网络完成。NMT 模型通常采用交叉熵损失作为训练目标，通过最大化整个训练数据上预测序列和目标序列词级别的似然进行训练；为了提高译文的质量，往往会采用束搜索策略，在牺牲一定推理速度的条件下扩大模型的搜索空间。

机器翻译预训练模型 BERT 之类的掩码语言模型能够对序列的双向上下文表示进行建模，但是其训练是按照分类任务进行的，即将编码器的输出输入到 `Softmax` 层预测被掩码的词在整个词表上的概率分布。GPT 之类的自回归模型和传统的语言模型的训练方式一致，即通过当前已生成序列的信息预测下一个词。BART 将类似 BERT 具有双向表示能力的构架作为编码器学习加噪序列的表示，而将类似于 GPT 的自回归构架运用于解码器，用于逐词生成原始未加噪的序列。其训练的优化目标为在整个训练集 D 上加噪序列片段与原始序列片段的似然概率，即

$$\arg\max_{\theta} \mathcal{L}_{\theta} = \arg\max_{\theta} \sum_{x \in D} log(P(x|\mathcal{N}(x);\theta)) \tag{2}$$

其中 $\mathcal{N}(\cdot)$ 表示加噪函数，在 BART 在预训练过程中采用了多个加噪方法，包括：BART 在预训练程中采用了多个加噪的方法，包括 1) 词的遮蔽、2) 句子顺序扰动、3) 文档转换、4) 词删除、5) 序列片段替换等，这些加噪方法的示意请见图1。

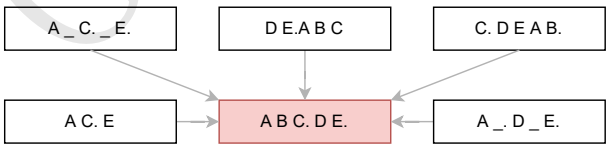


图 1: BART 的加噪方法示意

词典注入的藏汉机器翻译预模型训练方法 PTMDI 的预训练沿用了 BART 加噪并重构的建模方法, 但是与 BART 不同的是 PTMDI 中词典注入代替了各类加噪方案。词典的注入不仅能起到加噪的作用，同时也在客观上要求编码器学习跨语言的联合表示。有关双语词典的获取、筛选和注入请见 4.1 节。本文中在完成词典注入的单语数据上进行预训练之后，并在规模为 500 K 平行数据上进行微调。具体的预训练和微调的示意请见图 2 和图 3。考虑到收集的双语词典的词条大部分为名词，在进行词典注入时优先替换单语数据中的名词，同时保证被替换的词的数量不超过整个句子词长度的 15 %.

第二十一届中国计算语言学大会论文集，第374页-第383页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

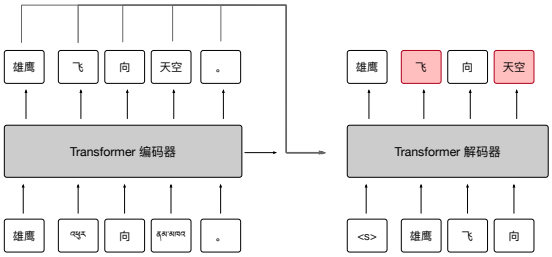


图 2: 预训练过程

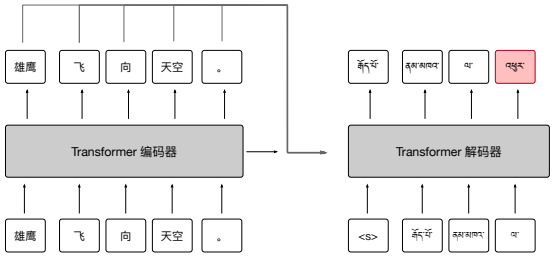


图 3: 微调过程

因为编码器需要学习藏汉两种语言的表示，需要模型有更大的学习容量，所以本文中使用了相较于解码器更深的网络构架。此外编码器的表示和理解性能相对而言比解码器的自回归生成和掩码自编码性能，对翻译终表现有更加重要的影响 (Kasai et al., 2020), 因而在多语言机器翻译任务研究者有使用较深的编码器较浅的解码器的应用实践 (Kong et al., 2021)，在翻译性能不退化的前提下，提高翻译速度。

PTMDI 训练方法能通过注入词典的方式进行翻译模型的预训练，因为词典的对齐特性使得模型在预训练阶段就开始跨语言的信息交互，学习跨语言信息的关联。此外，这种词典注入方式使得离散的词典特征能够很好地整合到端到端序列序列学习的连续过程中，是一种在机器翻译模型中有效融合先验知识的方法。考虑到相较于特定领域内的对齐数据，领域词典和领域单语数据比较容易获取和收集，所以 PTMDI 也是一种能以较为低廉的代价进行机器翻译领域适应的方法，尤其是适用于藏汉语言对这样的低资源机器翻译任务。

4 实验

4.1 数据设定

词典 为了藏汉双语词典涵盖较为广泛的领域，尤其是学习到受限的藏汉对齐文本之外的翻译知识，本文使用藏汉、汉藏、藏英、英藏四个方向的双语词典资源和利用统计词对齐工具 `FastAlign`[1] (Dyer et al., 2013) 在藏汉平行数据中获取的藏汉对齐词表。其中所有词典数据中只提取有单个释义的词条，另外对于藏英、英藏词典先将英文通过 Google 在线翻译系统翻译为汉文，然后再进行筛选处理; 对于统计对齐词表设定筛选的词对齐概率阈值为 `0.3` ，若有多个超过该阈值的对齐词表项则随机选择。词典词源的统计信息请见表 2, 藏汉和汉藏词典的领域包括日常用词、法律、生物、化学、医疗、数学、计算机等，藏英和英藏词典则主要是日常用词。对如表 2 所示的总计 `384654` 个筛选的词条进行正则化和去重处理之后，最终获得 `314500` 个独立词条。

表 2: 词典资源统计表

词典	词典数量	总词条数目	筛选的词条数目	筛选比例 (%)
藏汉词典	7	451200	153020	33.9
汉藏词典	5	341000	120000	35.2
藏英词典	3	130200	29949	23.0
英藏词典	2	177876	36685	20.6
统计对齐词表	-	95699	45000	47.0
总计		1195975	384654	32.2

[1] https://github.com/clab/fast_align

(c) 2022 中国中文信息学会计算语言学专业委员会

双语数据 与英文等具有显式的词分隔符不同，比如藏文和汉文如果直接使用纯粹基于频率统计的子词分词方法，将可能会生成大量在语言学上无实际意义的子词结构，这一现象对藏文这种拼音文字尤其明显。在低资源的机器翻译任务设定中，这些冗余的子词使得机器翻译模型需要学习额外的构词规律，在客观上加大了模型的学习负担。除了低资源机器翻译任务之外，涉及汉文、日文、朝鲜文等语言的富资源机器翻译任务中一般也是采用先分词再学习子词的数据预处理流程 (Alexis and Guillaume, 2019)。本文中数据的预处理也是采用了这种策略，汉文分词使用了 `jieba`[2]分词工具进行分词，藏文藏文分词采用了 (桑杰端珠 and 才让加, 2018) 提出的藏文分词方法。对文本进行分词处理之后使用了 `SentencePiece`[3] (Kudo and Richardson, 2018) 子词学习。为了过滤平行数据中的噪音样本，本文通过 `fasttext`[4] (Joulin et al., 2016) 中的语言标识模型去除藏文句子中的汉文和汉文句子中的藏文，同时也删除了数据样本中的非 Unicode 字符。本文限制了对齐句对的最大长度为 120 个词，同时剔除了藏汉词长度比大于 4 的句对。通过取重方法保证训练集、验证集和测试集没有交集。最终的藏汉平行数据规模见表 3 。

表 3: 平行数据和单语数据规模

数据类型	平行数据 (句对)	单语数据 (藏/汉) (句对)
训练集	500 K	6.8 M / 5.2 M
验证集	5 K	63 K / 51 K
测试集	5 K	62 K / 51 K
总计	510 K	6.9 M / 5.2 M

单语数据 由于用于微调的平行数据主要是新闻领域的，为了更加有效的模型训练，本文在收集藏语和汉语的单语数据时也使用了新闻领域的数据。单语数据的主要来源是各类藏文新闻网站和这些网站对应汉文网站的对应栏目，以完成数据更好的领域适配。单语数据的预处理方式和平行数据的预处理方式是一致的，也是先分词，再学习子词。在进行正则去噪、去重等预处理之后，最终保留的藏文和汉文单语数据的规模为反而别为 6.9 M 和 5.2 M。

4.2 模型设定

本文中所有模型的训练和测试都是基于 Fairseq [5] (Ott et al., 2019) 框架实现的，使用了 4 张 Nvidia Quadro P1000 GPU。基准模型中纯监督式模型和回译模型使用了 6 层的 Transformer 编码器和解码器；藏文和汉文的词表大小分别为 8K 和 9K。PTMDI 模型使用了 10 层的 Transformer 编码器和 6 层的 Transformer 解码器，编码器共享了藏语和汉语的词表，解码器使用了独立的对应目标语言的词表。所有模型解码器和编码器的嵌入维度为 512，编码器和解码器的前馈网络的维度为 2048，使用了 Adam 优化器进行参数优化，初始学习率设置为 0.001，学习率衰减函数选用了平方根倒数，批处理大小为 4096 个词, 所有的模型都训练了 60 轮次。

4.3 实验结果

表 4 列出了纯监督式 Transformer 模型、回译模型、BART 和 PTMDI 模型在测试集上的最终 BLUE 的测定值。从表中可以看出，本文中的 PTMDI 模型比 BART 这一强基准模型在藏汉和汉藏翻译任务上 BLEU 值分别高出 2.3 和 2.1，用实证方法证实了 PTMDI 在藏汉机器翻译任务上的

[2] https://github.com/fxsjy/jieba

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

[4] https://github.com/facebookresearch/fastText

[5] https://github.com/pytorch/fairseq/

(c) 2022 中国中文信息学会计算语言学专业委员会

有效性。此外从图 4 中模型在验证集上的 BLEU 变化和图 5 中训练过程中的损失变化，可以得知 PTMDI 模型有更好的收敛特性，证实了模型在预训练阶段就通过词典学习双语映射关系确实能够帮助微调过程中模型的学习能力。

表 4: 各个模型在测试集上 BLEU 值

模型	藏 → 汉	汉 → 藏
Transformer	27.1	26.8
回译	28.3	27.2
BART	29.8	29.3
PTMDI	**32.1**	**31.4**

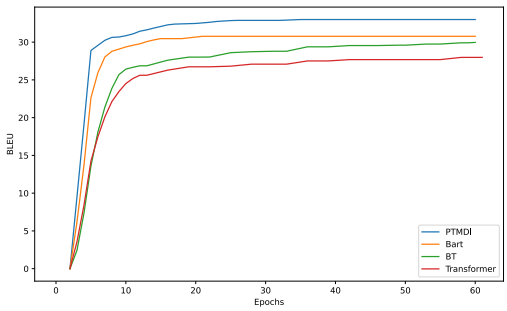


图 4: 各个模型验证集上的 BLEU 变化 (藏 → 汉)

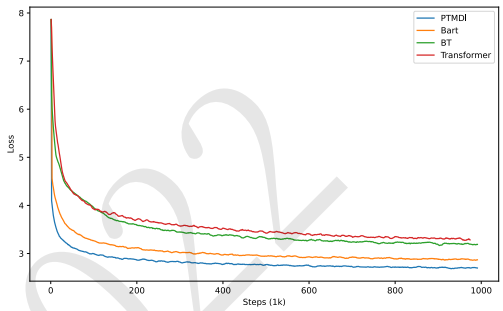


图 5: 各个模型的训练损失变化 (藏 → 汉)

从表 5 可以看出，在测试集样例中的专业词汇食用菌和羊肚菌在 PTMDI 模型中被较为准确地被译出，且译文更加流畅。除了验证模型在双语数据的领域有良好性能之外，本文还对其他跨领域场景下的性能进行了测试，如表 6 所示的是测试所有模型在计算机科学领域表现的一个样式示意，从该译文样例中可以发现比如汇编，编译器等双语平行数据中不存在的词条也被准确翻译出来。说明 PTMDI 确实在预训练过程中挖掘了先验的双语词典内的翻译知识。

表 5: 测试集中的译文样例

原文	ཞེང་ཡས་ཚན་རིག་ཁང་གི་ཟས་སྐྱོང་འབྲུ་ཐེབ་འཇུག་ཁང་གི་ཞེབ་འཇུག་པ་གཞོན་པས་མིའི་ཐབས་ཀྱི་གཞི་ཚོན་ཅན་གྱི་ལུག་ས་འབྲུ་ཐའེ་བས་འཇུག་སྐེད་པར་ཞེབ་འཇུག་རེད།
参考译文	农业科学院食用菌研究所副研究员研究人工规模化种植羊肚菌。
模型	译文
Transformer	农业科学院进食菌类研究所年轻研究员研究人为种植羊肚子细菌。
BT	农业科学院的吃饭细菌研究所的副研究员正在研究人为规模的羊肚子细菌的种植。
BART	来自农业科学院的可食用细菌研究所的助理研究员研究了规模种植羊菌的技术手段。
PTMDI	来自农业科学院的食用菌研究所的助理研究员研究了人工有规模种植羊肚菌。

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

表 6: 跨领域的译文样例

输入	把高级程序语言翻译成汇编语言或机器语言的工作称为编译，完成这项翻译工作的软件系统称为编译程序或编译器。
模型	译文
Transformer	མཐོ་རིམ་བྱ་རིམ་སྐད་ཡིག་སྐད་བརྒལ་འཕྲུལ་ཆས་སྐད་ཆ་ལ་བསྒྱུར་བའི་བྱ་བ་དེ་ལ་རྩོམ་སྒྲིག་ཟེར། ལོ་རྩྭའི་ལས་ཀ་འདི་འགྲུབ་པའི་ཚེད་དུ་མཉེན་ཆས་མ་ལག་འདི་ལ་རྩོམ་སྒྲིག་བྱ་རིམ་ཡང་ན་རྩོམ་སྒྲིག་ཟེར།
BT	མཐོ་རིམ་བྱ་རིམ་གྱི་སྐད་ཆ་དེ་རྩོམ་བསྒིགས་སྐད་ཆ་ཡང་ན་འཕྲུལ་ཆས་སྐད་ཆ་ལ་བསྒྱུར་བའི་བྱ་བ་དེ་མིང་ལ་རྩོམ་བསྒྱུར་ཟེར་ལ། ཡིག་སྒྱུར་ལས་དོན་འདི་འགྲུབ་པའི་ཚེད་དུ་མཉེན་ཆས་མ་ལག་ལ་རྩོམ་བསྒྱུར་བྱ་རིམ་ཟེར་པའོ།།
BART	མཐོ་རིམ་བྱ་རིམ་གྱི་སྐད་ཆ་དེ་གུང་སྒྲིག་པ་སྐད་ཆའམ་འཕྲུལ་ཆས་སྐད་ཆ་ཟེར། ལོ་རྩྭའི་བྱེད་པའི་ལས་ཀ་དེ་ལ་གུང་སྒྲིག་ཏུ་འགོད་པ་དང་། ལོ་རྩྭའི་ལས་ཀ་འདི་འགྲུབ་ཚེད་མཉེན་ཆས་མ་ལག་འདི་ལ་གུང་སྒྲིག་བྱ་རིམ་མམ་གུང་སྒྲིག་ལག་ཆམིང་བཏུགས་པ་རེད།
PTMDI	མཐོ་རིམ་བྱ་རིམ་གྱི་སྐད་བརྡ་དེ་སྡོགས་བསྒྲིགས་སྐད་བརྡའམ་འཕྲུལ་ཆས་སྐད་བརྡ་ལ་བསྒྱུར་བའི་བྱ་བ་དེར་རྩོམ་སྒྲིག་ཏུ་འགོད་པ་དང་། ལོ་རྩྭའི་ལས་དོན་འདི་ལེགས་འགྲུབ་ཐུབ་པའི་མཉེན་ཆས་མ་ལག་འདི་ལ་རྩོམ་སྒྲིག་བྱ་རིམ་མམ་རྩོམ་སྒྲིག་ཆས་ཞེས་འགོད་པ།

5 总结

本文受到双语交流中混和语言能有效增进交流这一现象启发，利用多个领域的藏汉双语词典和百万句子级别的藏汉单语数据，以 BART 风格降噪自编码为训练目标，通过在单语数据中有效注入词典，进行藏汉跨语言语言模型的预训练，并在已有藏汉平行数据上进行微调。经过广泛的实验验证，本文中的方法比 BART 强基准模型在测试集上的 BLUE 值在藏汉和汉藏方向上分别提高 2.3 和 2.1。结合利用更大规模的单语数据，更加准确有效的词典注入方式，混合 BART 和词典注入的训练方法，应该可以更进一步提高藏汉翻译的性能，我们将该设想在未来的工作中进行研究和探索。此外，本文方法能为后续一到多、多到一、多到多等藏文多语言翻译课题提供可靠的研究基础。

参考文献

Conneau Alexis and Lample Guillaume. 2019. Cross-lingual language model pretraining. 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 7057–7067.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, N. Gomez Aidan, Kaiser Lukasz, and Polosukhin Illia. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *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.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

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,

(c) 2022 中国中文信息学会计算语言学专业委员会

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*.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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.

Hany Hassan, Anthony Aue, Chang Chen, and Chowdhary. 2018. Achieving human parity on automatic chinese to english new. *ArXiv preprint*, abs/1803.05567.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. 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*, pages 770–778. IEEE Computer Society.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Armand Joulin, Edouard Grave, and Piotr an Bojanowski. 2016. Fasttext.zip: Compressing text classification models. *ArXiv preprint*, abs/1612.03651.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and A. Noah Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation.

Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. 2021. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624, Online, April. 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.

Surabhi Kumari, Nikhil Jaiswal, Mayur Patidar, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, and Lovekesh Vig. 2021. Domain adaptation for NMT via filtered iterative back-translation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 263–271, Kyiv, Ukraine, April. 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

(c) 2022 中国中文信息学会计算语言学专业委员会

Yinhan Liu, Myle Ott, Naman Goyal, and Jingfei an Du. 2019. Roberta: a robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yaron Matras. 2000. Mixed languages: a functional–communicative approach. *Bilingualism: Language and Cognition*, 3(2):79–99, Aug.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

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.

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. In *Proceedings of NIPS*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. 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 5754–5764.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model.

头旦才让, 仁青东主, 尼玛扎西, 于永斌, and 邓权芯. 2020. 基于改进字节对编码的汉藏机器翻译研究. 电子科技大学学报, 50(02):249–255+293.

孙义栋, 拥措, and 杨丹. 2022. 基于 volt 的藏汉双向机器翻译. 计算机与现代化, (05):28–32+39.

慈祯嘉措, 桑杰端珠, 孙茂松, 色差甲, and 周毛先. 2019. 融合单语语言模型的藏汉机器翻译方法研究. 中文信息学报, 33(12):61–66.

慈祯嘉措, 桑杰端珠, 孙茂松, 周毛先, and 色差甲. 2020. 基于迭代式回译策略的藏汉机器翻译方法研究. 中文信息学报, 34(11):67–73+83.

李亚超, 熊德意, 张民, 江静, 马宁, and 殷建民. 2017. 藏汉神经网络机器翻译研究. 中文信息学报, 31(06):103–109.

桑杰端珠 and 才让加. 2018. 神经网络藏文分词方法研究. 青海科技, 25:15–21.

沙九, 冯冲, 张天夫, 郭宇航, and 刘芳. 2020. 多策略切分粒度的藏汉双向神经机器翻译研究. 厦门大学学报 (自然科学版), 59(02):213–219.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于特征融合的汉语被动句自动识别研究

胡康[1] ， 曲维光[1*] ， 魏庭新[2] ， 周俊生[1] ， 顾彦慧[1] ， 李斌[3]
(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 国际文化教育学院，江苏省 南京市 210097;
3.南京师范大学 文学院，江苏省 南京市 210097;
*通讯作者，Email：wgqu_nj@163.com)

摘要

汉语中的被动句根据有无被动标记词可分为有标记被动句和无标记被动句。由于其形态构成复杂多样，给自然语言理解带来很大困难，因此实现汉语被动句的自动识别对自然语言处理下游任务具有重要意义。本文构建了一个被动句语料库，提出了一个融合词性和动词论元框架信息的PC-BERT-CNN模型，对汉语被动句进行自动识别。实验结果表明，本文提出的模型能够准确地识别汉语被动句，其中有标记被动句识别F1值达到98.77%，无标记被动句识别F1值达到96.72%。

关键词： 汉语被动句； 自动识别； 特征融合； 语料库

Automatic Recognition of Chinese Passive Sentences Based on Feature Fusion

HU Kang[1], QU Weiguang[1*], WEI Tingxin[2], ZHOU Junsheng[1], GU Yanhui[1], LI Bin[3]
(1.School of Computer and Electronic Information/School of Artificial Intelligence,
Nanjing Normal University, Nanjing, Jiangsu 210023, China;
2.International College for Chinese Studies, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
3.School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
*Corresponding, Email: wgqu_nj@163.com)

Abstract

Chinese passive sentences can be categorized as marked and unmarked passive sentences according to the presence or absence of markers. Due to its diverse morphological variations, it brings great difficulties to Natural Language Understanding (NLU), and the automatic recognition of it is of great significance to the downstream tasks of Natural Language Processing (NLP). In this paper, we firstly construct a passive sentence corpus, then propose a PC-BERT-CNN model incorporating part-of-speech features and verb's argument frame information to automatically identify Chinese passive sentences. The experimental results show that our model can identify Chinese passive sentences more accurately than the existing automatic parsing tool LTP, the F1 values of marked and unmarked passive sentences recognition task reach 98.77% and 96.72% respectively.

Keywords: Chinese passive sentences , automatic recognition , feature fusion , corpus

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家社会科学基金重大项目(21&ZD288)

1 引言

作为一种用于说明主语和谓语之间关系的语法形式，被动句广泛存在于各种自然语言中。与主动句中主语做动作的施事不同，被动句的主语由动作的受事充当。由于语言的类型和特征存在差异，不同语言中的被动句在表示方法上也不尽相同。例如在英语这种重形合的语言中，其被动句大多数是结构被动句(syntactic passive)(蒋坚松, 2002)，即谓语含有“be+v-ed”形式标记的句子，因此在机器理解时，容易通过动词形态变化以及助动词来识别句子语态。而汉语是一种重意合的语言，缺乏动词形态上的变化，虽然也存在一部分有标记被动句（如“被”字句），但与英语不同的是，汉语中还存在大量不含标记词的无标记被动句，这类句子的被动意隐含在词汇或语境中，这给汉语被动句的自动识别造成困难。在以下五个例句中，例1、5为非被动句，例2~4为被动句。其中，例1和例2从句法结构上看，都是主谓补结构，但例1是主动句而例2为无标记被动句，因此在判断句子中各成分语义关系时容易混淆；例3和例4分别是含标记词“被”和“让”的有标记被动句；例5虽然含“让”，但它是该句的动词而非作为被动标记的介词，被动标记词的这种同形多义现象给被动句的识别带来很大挑战。

例1 我吃完了。
例2 饭吃完了。
例3 我的钱包被偷了。
例4 我让门槛绊倒了。
例5 他让这位老人先走。

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

目前对被动句的研究主要集中在语言学层面。在自然语言处理领域，目前主流的自动解析器可以对汉语句子，也包括被动句，进行句法、语义层面的解析。然而，通过这些解析器得到的句子解析结果并不能直接判定被动句，而需要人为根据标注规范制定相应的判断规则或条件，再进行被动句判定。这导致被动句识别，尤其是无标记被动句识别任务的性能大大降低。

宋文辉等人 (2007)统计了被动句在会话、小说、新闻及学术四种语体中的分布情况如表 1所示，可见在这四种语体中，被动句在小说和新闻语体中所占比例最高，因此本文选取《人民日报》新闻语料作为被动句识别的基础语料进行人工标注，并提出一种基于深度学习的方法实现汉语被动句的自动识别，可以在大规模的语料中迅速、准确地筛选出被动句，以期对句法解析、AMR解析等下游任务提供支持。

本文主要贡献如下：第一，构建了一个包含13530条句子的被动句语料库，用于深度学习模型训练；第二，提出了一个PC-BERT-CNN(POS and CPB enhanced BERT-CNN Model)模型，实现了汉语被动句的自动识别，该模型对BERT模型进行词性增强并融入了中文命题语料库(Chinese Proposition Bank, CPB)中的动词论元框架信息；第三，实验结果表明，本文提出的模型在被动句自动识别任务上取得了较好的性能，其中有标记被动句识别的F1值达到98.77%，无标记被动句识别的F1值达到96.72%。

2 相关工作

2.1 被动句本体研究现状

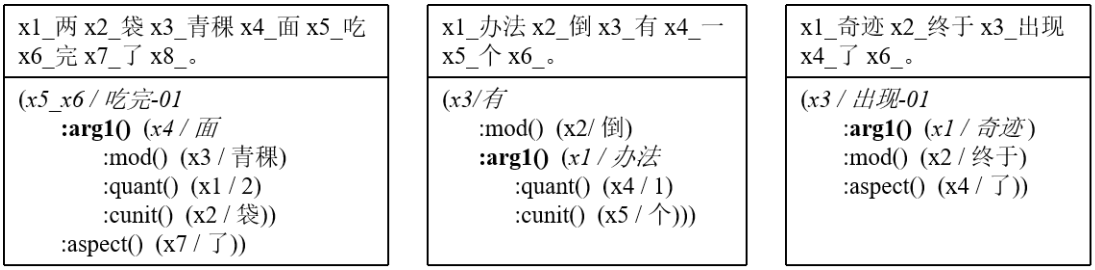
被动句在汉语中随处可见。近年来关于被动句的研究大都集中在语言学领域，研究对象主要涉及被动句的构式、语义和语用等方面。邹丽玲 (2016)从英译汉的视角解析了汉语无标记被动句表达被动含义时，不加“被”等被动标记词的原因。王灿龙 (1998)根据无标记被动句中被动语态主要用于及物动词的特点，从单音节及物动词和双音节及物动词两个方面分析能够进入无标记被动句的词语。王芸华 (2014)从不同角度考察了被动句主语的语义角色，认为被动句的主语除受事之外，也能由与事、主事、结果等其他语义角色充当。汤敬安 (2016)对比有标记被动句和无标记被动句之间的构式差异，并对二者的认知过程进行研究，指出二者的差异主要体现在语言结构中的受事注意范围、详略度、扫描方式等方面。对于有标记被动句标记词的界定问题，不同学者的看法也不尽相同，李珊 (1994)认为可以将标记词划分为“被、叫、让、给、为、被/让/叫...给、被/为...所”7类，而乔莎莎 (2015)则将标记词的数量进一步扩充到11种，分别为“被、叫、让、给、蒙、由、被...给、叫...给、让...给、被...所、为...所”，鞠彩萍 (2007)通过语料证实，“遭”在某些条件下也能作为标记词使用。

第二十一届中国计算语言学大会论文集，第384页-第394页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

2.2 被动句语料库构建现状

现有的中文语料库通常面向句子整体的句法、语义结构进行标注，而没有针对被动句的标注语料。被动句在这些语料中的数量较少，而且被动句在这些语料中并没有被显式标注。

例如中文抽象语义表示（Chinese Abstract Meaning Representation, CAMR）标注了动词与论元之间的关系 (李斌等, 2017)，它对被动句的标注方式是把位于主语位置的受事标注为动词的arg1，如图 1 中(a)所示，“面”标注为动词“吃”的受事arg1。不过这一标注方式并非被动句独有，如(b)和(c)中的论元结构虽然与(a)一致，但句子(b)是一个话题做主语的非被动句，句子(c)中的动词“出现”不具有被动用法，也是一个非被动句。因此CAMR语料库不能直接作为被动句数据集使用，本文将采用人工标注的方法构建一个被动句语料库，用于模型训练。



(a) “两袋青稞面吃完了。” (b) “办法倒有一个。” (c) “奇迹终于出现了。”

Figure 1: CAMR语料库标注的句子

2.3 被动句识别研究现状

目前针对被动句解析的专项研究较少，但在句法或语义解析任务中，可以通过判断解析图中语义角色或论元关系结构完成对被动句的识别。为探究现有自动解析器在被动句识别上的效果，本文使用哈工大发布的语言技术平台（LTP）[1](Che et al., 2010)对句子的句法结构和语义角色分别进行解析。对于依存句法解析结果，若解析图中存

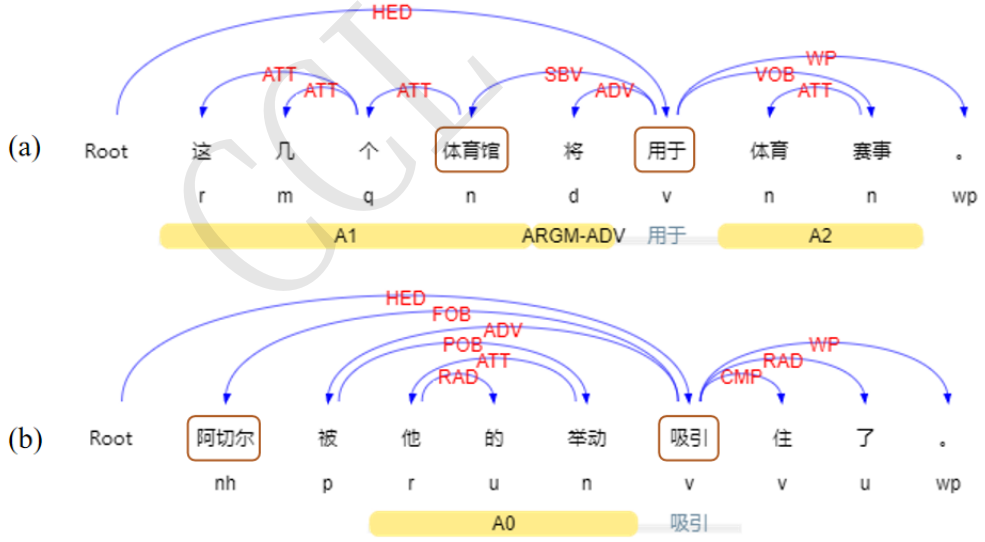


Figure 2: LTP对句子的自动解析

在“FOB←V”或“FOB←V→ADV(→POB)”[2]关系链时，把该句子记为正确解析，否则记为错

[1] http://ltp.ai/index.html

[2] FOB代表前置宾语，在被动句中通常指向受事；ADV代表状中结构，在有标记被动句中通常指向标记词；POB代表介宾关系，在有标记被动句中通常指向施事，可不出现在句子中。

误解析；对于语义角色解析结果，若动词及其受事论元A1、施事论元A0都判断正确，则记为正确解析，若有至少一个论元错误或没有解析出论元则记为错误解析。通过分析发现，LTP对被动句的解析性能并不高（具体统计数据见5.3节）。图 2 展示了LTP对两个被动句的自动解析结果，句(a)是一个无标记被动句，动词“用于”的A1解析正确，即LTP语义角色解析正确。但“体育场”应是“用于”的前置宾语(FOB)而非主谓关系(SBV)。句(b)中存在”FOB←V”关系链，即“阿切尔”是“吸引”的前置宾语，句法解析正确，但语义角色解析结果中，A1“阿切尔”却未被解析。这说明LTP在被动句自动解析方面仍存在一定的提升空间。

本文将汉语被动句的自动识别任务视为一个文本分类问题。近年来，深度学习在文本分类中的运用日渐成熟。Kim (2014)最先在卷积神经网络（Convolutional Neural Networks，CNN）的基础上提出TextCNN用于文本分类。Devlin等人 (2019)提出的BERT基于Transformer的双向深度语言模型，可以捕获双向上下文语义，在机器翻译、文本分类、文本相似性等多个自然语言处理任务中取得优异的表现。Qin等人 (2020)提出了一种特征投影的方法，将现有特征投影到共同特征的正交空间中，生成的投影垂直于共同特征并且对分类更具辨别力。朱向其等人 (2021)提出了一种基于改进词性信息和ACBiLSTM的中文短文本分类模型，在THUCNews数据集上取得良好的分类性能，准确率和F1值分别达到97.43%和95.24%。Nguyen等人 (2021)提出了两种利用外部知识的方法来丰富句子的语义表示，进而识别与恶意软件相关的句子，使得SVM分类器的F1值提升了9%。受此启发，对于本文的被动句识别任务，也可引入相应的外部知识并进行特征融合，以提升分类模型的性能。

3 被动句语料库构建

3.1 被动句语料来源

本文选取1998年1至3月《人民日报》新闻语料作为被动句标注的基础语料。由于《人民日报》语料中每一行文本代表一个自然段或标题，所以需要把提取出的纯文本语料切分为单个句子，去除新闻标题等不符合完整句式特点的文本，构成待标数据集共5万句，约200万字。

3.2 被动句的标注规范

定义“NP1+Mark+[NP2]+V”和“NP1+V”分别为有标记被动句和无标记被动句的一般构式，其中V为及物动词，N1代表动作的广义受事，N2代表对应的广义施事，Mark代表标记词，[ ]中的成分可不出现。被动句标注的重点在于确定主语和谓语之间的被动关系，进而根据有无标记词分为有标记被动句或无标记被动句。

3.2.1 有标记被动句

根据前人研究，本文把被动标记词的范围限定在“被、叫、让、给、蒙、由、为、遭”这八大类。当句子中动词对主语造成“强影响”或“不如意”的结果时，我们认为该句子很可能是被动句，具体来说，当句子主语充当谓语动词的以下几种论元角色时，可将句子标注为有标记被动句。

（1）**受事、与事论元** 受事代表直接受动词支配或影响的人或事物；与事是动作非主动的参与者。如例6中，

例6 (a)他的钱包被人偷走了。 (b)她被骗去不少钱。

“钱包”是动词“偷”的受事，“她”是“骗”的与事，两个句子均表达了“不如意”的结果，因此将这类句子标注为有标记被动句。

（2）**感事、主事论元** 感事是非自主的感知性事件的主体；主事指性质、状态或变化性事件的主体，与系事相对。如例7中，

例7 (a)他被一阵枪声吓醒了。 (b)埃雷迪亚被称为“花城”。

感事“他”由表示消极或突发反应的词——“吓”支配，表达一种“不如意”的状态，因此认为该句子是有标记被动句。(b)中“埃雷迪亚”是主事，“花城”是系事，二者通过“称为”这一评定、认同性的动词联系起来，这类句子也被视为有标记被动句。

（3）**工具、材料论元** 工具和材料都在动作发生过程中会受到控制，同时会对它们本身造成“强影响”。如例8中，

例8 (a)刺刀都被他捅弯了。 (b)乳胶漆被他涂了墙。

第二十一届中国计算语言学大会论文集，第384页-第394页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

“刺刀弯了”是在陈述“捅”这一动作过程中给工具造成的强影响。同理，“乳胶漆”作为“涂”的材料论元，被附加在别的物体上，也受到了“强影响”，因此把它视为被动句。

3.2.2 无标记被动句

在无标记被动句中，由于缺少被动标记，导致句子对动词有严格的语义选择限制，因此在标注无标记被动句时，除了观察句子主语充当的语义角色外，还需考虑动词本身的语义特点。当句中动词具有以下两种特点时，则把该句子标注为无标记被动句。

（1）**可控性** 表示动作可以由动作发出者根据自己的意愿进行控制。如例9中，

例9 (a)桌子已经擦了。 (b)这棵树已经死了。

动词“擦”可控而“死”不可控，因此(b)不属于无标记被动句。

（2）**强动作性** 表示的是人、其他生命体的动作或某种自然力造成的动作，动作性比较明显。如例10中，

例10 (a)作业做完了。 (b)脚冻伤了。 (c)人们的安全意识提高了。

“做”和“冻”都是动作性比较强的动词，(a)和(b)均属于无标记被动句；而“提高”的动作性弱，因此(c)不属于无标记被动句。

3.2.3 特殊情况

当某动词具有多种义项时，需要根据语境判断该动词在当前句子中的语义，进一步判定句子是否属于被动句。如例11中，

例11 (a)比赛中他被对手摔倒了。 (b)他走路时不小心摔倒了。

虽然两个句子都表达“不如意”的结果，但(a)中“摔倒”属于自主性动词，具有可控性；而(b)中的“摔倒”则属于非自主动词，不具有可控性；因此前者属于被动句而后者不是。

此外，若一个句子同时符合有标记被动句和无标记被动句的构式，则将其拆分或改写为多个句子，使得每个句子都是单标签样本。如例12中既包含有标记被动语态“心+被+打动”，又包含无标记被动语态“生源+积聚”，因此将它拆分为有标记被动句(a)和无标记被动句(b)。

例12 久而久之,下岗人的心被打动了,生源也就慢慢积聚起来了。

(a) 久而久之,下岗人的心被打动了。

(b) 生源也就慢慢积聚起来了。

3.3 被动句语料库统计

标注人员为2名具有语言学专业背景的志愿者，经过培训后根据上述标注规范展开标注工作。标注过程分为两个阶段，第一阶段为试标阶段，2人同时标注100条句子，完成后进行一致性统计的结果为85%，二者的标注结果存在的差异主要体现在动词在不同语境下的语义理解上。在经过适当调整后，进行第二阶段的正式标注，共13530条句，整体标注的一致性达到91%。语料库中约5%的句子由拆分或改写得到，各类别样本数据如表 2 所示。

句子类别	样本数量(条)	样本示例
有标记被动句	4495	大火已于傍晚六时被扑灭。
无标记被动句	4570	检查情况已在当地曝光。
非被动句	4465	她将参加钢琴组的比赛。

Table 2: 被动句语料库构成

此外，对有标记被动句的八大类标记词进行了统计，如表 3 所示，可知，“被”是最典型的被动标记词，“由”字句也比较常见，而其他几类标记词在新闻领域文本中出现的频率较低。

标记词	被	由	为	给	遭	让	叫	蒙	总计
数量(条)	3064	1238	145	34	7	4	3	0	4495
占比(%)	68.16	27.54	3.23	0.76	0.16	0.09	0.07	0.00	100

Table 3: 被动句语料库中的标记词分布

(c) 2022 中国中文信息学会计算语言学专业委员会

4 被动句识别模型设计

本文将汉语被动句识别任务建模为一个三分类问题，即把句子分为有标记被动句、无标记被动句和非被动句三种类别，提出一个融合词性信息和论元框架信息的PC-BERT-CNN模型，从而对汉语句子所属类别进行预测。模型的输入为单个句子，输出为句子的类别标签。模型由4个模块组成：①词性增强的词嵌入模块，得到融合了词性信息的BERT向量表示；②CPB论元框架信息提取模块，得到句中所有动词的论元框架信息的静态词向量表示；③特征提取和拼接模块，得到融合词性特征和动词论元特征的句子表征；④标签预测模块，得到模型的输出，即句子的预测类别标签。模型整体架构如图 3 所示。

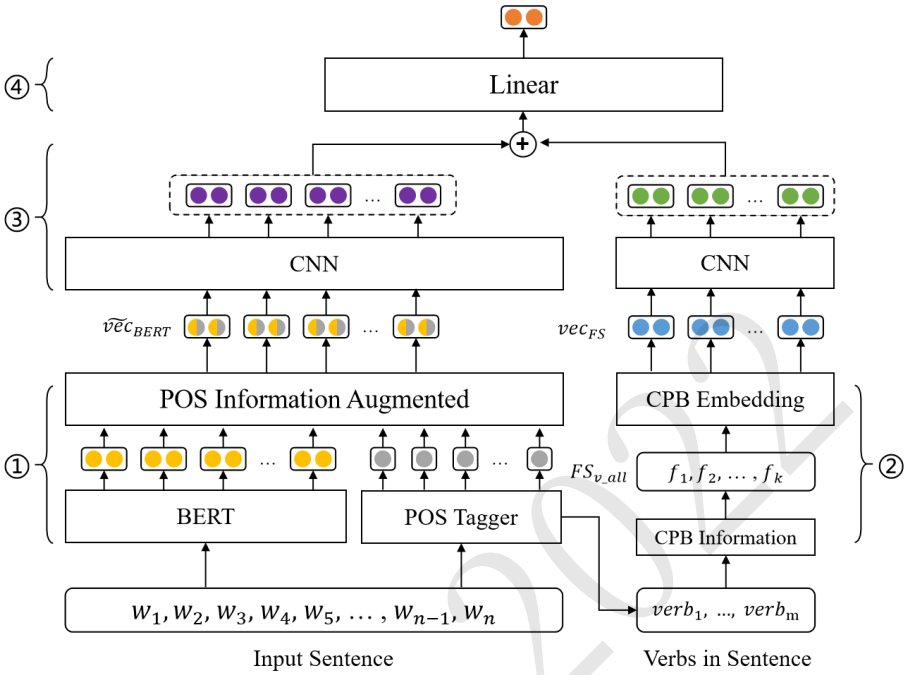


Figure 3: 被动句识别模型结构图

4.1 词性增强的词嵌入

被动句的核心是谓语动词，不同词性的词语具有不同的表征能力。在被动句识别任务中，谓语动词、施受事名词和标记介词是体现被动语态的重要组成部分。BERT在获取词嵌入时，利用自注意力机制充分考虑了每个字符的上下文信息，从而动态调整其向量表示，因此句子中每个词语的属性对模型的训练也具有重要作用。在得到BERT词向量后，利用结巴分词工具[3] 对句子进行分词和词性标注，根据本文研究对象把词性划分为动词(V)、名词(N)、介词(P)和其他(O)四类，并给各个词类分配权重，记为词性因子，如表 4 所示。

词类	说明	词性因子	取值范围
V	谓语动词	α	[0,1]
N	代表施事、受事的名词/代词	β	[0,1]
P	被动标记词	γ	[0,1]
O	其他修饰性的词语	δ	[0,1]

Table 4: 各类词性因子说明

各词性因子值为实验的超参数，设它们满足约束条件$\alpha + \beta + \gamma + \delta = 1$，各词性因子的最优取值通过参数分析得到。对于输入样本$S = [w_1, w_2, w_3, \ldots, w_n]$，首先经过BERT得到样本的表征向量$vec_{BERT}$，同时进行词性标注并分配权重，得到样本的词性因子序列$weight_{POS}$，然

[3] https://github.com/fxsjy/jieba

(c) 2022 中国中文信息学会计算语言学专业委员会

后将vec_{BERT}与$weight_{POS}$进行权重计算，得到融合词性信息的表征向量$\widetilde{vec}_{BERT}$，计算公式如(1)-(3)所示。

$$vec_{BERT} = BERT_{Embedding}(S) = [\vec{x}_0, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n] \tag{1}$$

$$weight_{POS} = [wp_0, wp_1, wp_2, \ldots, wp_n], wp_i \in [0, 1] \tag{2}$$

$$\widetilde{vec}_{BERT} = vec_{BERT} \times weight_{POS} = [\vec{y}_0, \vec{y}_1, \vec{y}_2, \ldots, \vec{y}_n] \tag{3}$$

其中$\vec{y}_i = \vec{x}_i \times wp_i$，“×”代表向量与标量的乘法运算。

4.2 CPB论元框架信息提取

CPB语料库对动词的论元框架，即动词义项进行了描述(Xue and Palmer, 2005)，每个Frameset对应动词的一种义项。图 4 列出了“评为”、“殉职”和“处理”三个动词在CPB中的论元框架信息，其中f1为该动词的第一种义项，f2为第二种义项。由于能够进入被动语态的动词应至少具备arg0和arg1这两个论元，因此论元框架中只含一个论元的动词是无法在被动句中使用的，如“殉职”。此外CPB动词论元框架中各个论元语义的描述对动词本身的语义也能起到很好的补充作用。故本文将动词的论元框架信息，即所有的Frameset从CPB中抽取出来，编码为词向量后进行特征融合，以期对模型性能的提升有积极的影响。

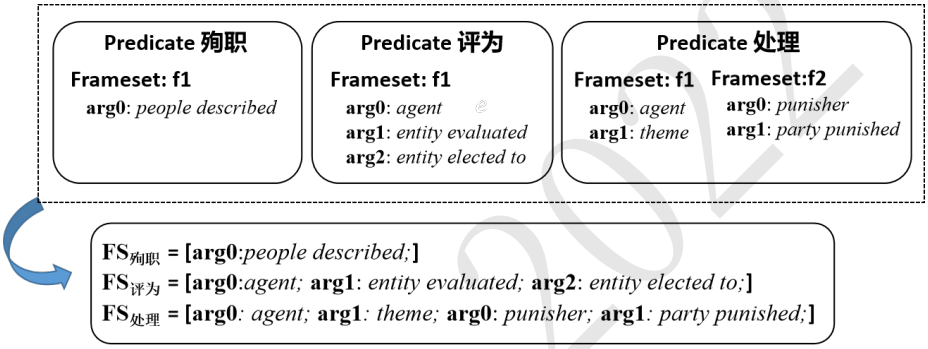


Figure 4: 从CPB语料库抽取论元框架信息

对于输入样本S，从词性标注结果中提取出当前样本句子中的所有动词$[verb_1, \ldots, verb_m]$，再从CPB语料库中抽取出每个动词的论元框架信息FS_i并进行字符串拼接，若在CPB语料库中未匹配到该动词的论元框架，则FS_i用空字符串替代，最终得到句子中所有动词的论元框架信息FS_{v_all}，由k个CPB动词论元框架描述信息的符号序列组成。再利用随机初始化的静态词嵌入层将FS_{v_all}转为向量表示，记为vec_{FS}，计算公式如(4)-(5)所示：

$$FS_{v\_all} = FS_1 + \ldots + FS_m = [f_1, f_2, \ldots, f_k] \tag{4}$$

$$vec_{FS} = Embedding(FS_{v\_all}) = [z_1, z_2, \ldots, z_k] \tag{5}$$

其中$k = \sum_{j=1}^{m} str_len(FS_j)$，“+”代表字符串加法。

4.3 特征提取和拼接

对于样本句子的词性增强BERT向量表示$\widetilde{vec}_{BERT}$和句中动词论元框架信息的向量表示vec_{FS}，利用卷积神经网络(CNN)对其进行特征提取，得到两个输出$\widetilde{F}_1$和$\widetilde{F}_2$，再将二者进行拼接，得到输入样本最终的特征表示$\widetilde{F}_s$，计算公式如(6)-(8)所示。

$$\widetilde{F}_1 = CNN(\widetilde{vec}_{BERT}) \tag{6}$$

$$\widetilde{F}_2 = CNN(vec_{FS}) \tag{7}$$

$$\widetilde{F}_s = \widetilde{F}_1 \oplus \widetilde{F}_2 \tag{8}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

4.4 标签预测

将样本最终的特征$\widetilde{F}_s$送入全连接层，之后对网络进行dropout处理，以防止出现过拟合现象，最终经过softmax函数得到的结果即为样本预测的类别标签。

5 实验

5.1 数据集划分及实验设置

数据集划分 本文的实验语料包含三个类别的数据，按照6:2:2的比例将数据集进行切分并随机打乱，分为训练集、验证集和测试集，各类别样本个数如表 5 所示。

类别	训练集(条)	验证集(条)	测试集(条)
有标记被动句	2697	899	899
无标记被动句	2742	914	914
非被动句	2679	893	893

Table 5: 数据集划分

超参数设置 实验使用的BERT[4]模型版本为Chinese L-12 H-768 A-12，学习率设置为1e-5，其他超参数设置如表 6 所示。

超参数	含义	值
epochs	数据集迭代次数	3
batch_size	单批次样本数量	32
pad_size	每个样本最大token数量	128
filter_sizes	卷积核尺寸	(2,3,4)
num_filters	卷积核数量	256
dropout	丢弃概率	0.1
cpb_embed_dim	cpb信息词嵌入维度	16
cpb_pad_size	cpb信息最大token数量	12
cpb_filter_sizes	cpb卷积核尺寸	(2,3)
cpb_num_filters	cpb卷积核数量	256
pos_weight	词性因子取值	(0.3,0.3,0.3,0.1)

Table 6: 超参数设置

评价指标 本文实验采用准确率P(Precision)、召回率R(Recall)和F1值(F1-measure)作为模型性能的评价指标。对比实验则使用正确率Acc(Accuracy)作为评价指标。

5.2 实验结果及分析

在文本分类领域，基于CNN和BERT预训练模型的分类方法是比较常用且高效的方法。为了验证模型的有效性，将本文提出的PC-BERT-CNN模型与TextCNN、BERT和BERT-CNN三种模型进行实验对比，对每个类别句子的识别性能分别计算其P、R和F1值，得到的各模型在各类句子自动识别任务上的实验结果如表 7 所示。

可以看出，本文提出的模型在三类句子上的性能均达到最佳。观察TextCNN模型与BERT-CNN模型的实验结果，容易发现，在引入BERT预训练模型之后，各类句子识别的F1值都有了明显的提高，原因在于TextCNN使用的是静态词向量，不能解决文本中一词多义的问题，而BERT在动态生成词向量的过程中考虑了更多的上下文信息，使得在CNN捕获局部特征时，能够更好地理解句子语义。被动句识别任务与普通句子文本分类任务不同，它重点关注的对象是动词以及施受事等，相比BERT-CNN模型，本文模型在融入词性信息和动词论元框架信息后，丰富了句子中动词及其相关成分的语义表示，因此分类性能明显提升。

[4] https://github.com/google-research/bert

(c) 2022 中国中文信息学会计算语言学专业委员会

模型	有标记被动句识别			无标记被动句识别			非被动句识别		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
TextCNN	95.85	**98.87**	97.34	91.38	93.39	92.37	**96.01**	90.89	93.38
BERT	**99.21**	97.55	98.37	94.64	96.50	95.56	95.73	95.41	95.57
BERT-CNN	99.10	98.33	98.72	**97.40**	94.20	95.77	93.66	**97.54**	95.56
本文模型	**99.21**	98.33	**98.77**	96.51	**96.94**	**96.72**	95.99	96.42	**96.20**

Table 7: 各类句子自动识别的实验结果

容易发现，在相同模型的实验条件下，有标记被动句的识别性能更高，这是因为有标记被动句本身就存在特征明显的被动标记词，因此通用的文本分类模型就已经能取得较好的识别性能，本文模型通过融合外部特征后，F1值稍有提高，达到98.77%。对于无标记被动句，由于其缺乏明显的标记词，而通过引入词性信息和论元框架信息可以很好地丰富句子的特征表示，因此本文模型与另外三个通用的分类模型相比，在无标记被动句识别任务上取得了最好的性能，其F1值达到96.72%，且在召回率显著提升的情况下，同时保证准确率维持在较高的水平。同样地，对于非被动句而言，由于模型能够较好地学习到被动句的结构和语义特征，那么对于非被动句的特征也应当能较好地捕获，因此本文模型在非被动句识别任务上也获得了较好的性能。

5.3 对比实验

语言技术平台(Language Technology Platform, LTP)是哈工大社会计算与信息检索研究中心(HIT-SCIR) 研发的一套中文自然语言处理开源基础技术平台，其最新发布的LTP 4.0在词法分析、句法分析和语义分析等六项自然语言处理任务上都取得了SOTA或具有竞争力的表现(Che et al., 2021)。其中LTP 4.0(Base2)[5]是LTP 4.0系列各版本模型中，在语义角色标注和句法依存分析任务上都表现最佳的模型，因此本文利用该模型的语义角色标注和句法依存分析这两项功能，对测试集中的句子进行自动解析。由于LTP得到的句子解析结果不能直接判定被动句，我们统计了各类句子的正确解析数目和正确率Acc[6]，并与本文模型的实验结果进行对比。对比实验所用的测试集包含899条有标记被动句、914条无标记被动句和893条非被动句。对比实验结果如表 8 所示。

模型	有标记被动句		无标记被动句		非被动句	
	正确数(条)	Acc(%)	正确数(条)	Acc(%)	正确数(条)	Acc(%)
LTP语义角色标注	487	54.17	814	89.06	850	95.18
LTP句法依存分析	840	93.44	693	75.90	808	90.48
本文模型	**884**	**98.33**	**886**	**96.94**	**861**	**96.42**

Table 8: 本文模型与LTP的对比实验结果

可以看出：首先，在有标记被动句识别方面，LTP的解析性能并不高，尤其语义角色标注功能的解析性能较差，正确率仅有54.17%，而本文模型的性能远远超过LTP，正确率达到98.33%；其次，对于无标记被动句而言，LTP的解析性能也不高，其中句法依存的解析性能只有75.90%，这也说明无标记被动句识别任务是提升句法解析性能所必须克服的一个重要任务，而本文模型效果远远优于LTP，正确率达到96.94%。最后，本文模型在非被动句识别任务上的性能也明显优于LTP自动解析器，正确率达到96.42%。

5.4 消融实验

本文提出的模型融合了词性信息和CPB中的动词论元框架信息，在被动句自动识别任务上取得了良好的性能。为探究二者对模型性能的影响，本文进行了消融实验，实验结果如表 9 所示，其中Ours代表本文模型，Ours-cpb和Ours-pos分别表示在本文模型基础上移除论元框架信息和词性信息的模型。

[5] https://github.com/HIT-SCIR/ltp
[6] 判断标准参照2.3节

(c) 2022 中国中文信息学会计算语言学专业委员会

模型	有标记被动句识别			无标记被动句识别			非被动句识别		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Ours	99.21	98.33	98.77	**96.51**	**96.94**	**96.72**	**95.99**	**96.42**	**96.20**
Ours-cpb	**99.22**	**98.78**	**99.00**	95.45	96.50	95.97	95.94	95.30	95.62
Ours-pos	98.99	98.33	98.66	95.96	96.17	96.07	95.54	95.97	95.75

Table 9: 被动句识别消融实验结果

可见在模型移除论元框架信息之后，无标记被动句和非被动句的识别性能均有下降，但有标记被动句识别的F1值反而提升到99.00%，这是由于有标记被动句本身已经具有明显的标记词特征，在引入论元框架信息后，过多的特征表示反而可能会对其造成干扰甚至产生负作用。反观无标记被动句，由于缺少标记词，自动识别的难度较大。在移除论元信息后，无标记被动句识别的F1值降低了0.75%，P值降低了1.06%，这说明论元信息能很好地补充无标记被动句中的特征表示。此外，在移除词性信息后，三类句子的识别性能均有所下降，这说词性信息有利于句子的特征提取和识别。

5.5 词性因子对实验的影响

α	β	γ	δ	Weighted F1
0.25	0.25	0.25	0.25	96.34
0.4	0.3	0.3	0	96.93
0.4	0.4	0.2	0	96.71
0.4	0.3	0.2	0.1	96.86
0.3	0.4	0.2	0.1	96.82
0.3	**0.3**	**0.3**	**0.1**	**96.97**
0.3	0.3	0.1	0.3	96.60
0.3	0.1	0.3	0.3	96.16
0.1	0.3	0.3	0.3	96.32
0.3	0.3	0.2	0.2	96.49

Table 10: 不同词性因子取值下本文模型的性能

为探究词性因子的不同取值对实验的影响，本文通过调整各词性因子的取值进行多次实验，取模型性能最佳时对应的参数值为词性因子的最优取值。参数调整的过程包含三个步骤：①各因子的初始值设为0.25，将其结果作为对照组；②在保证四个词性因子的和为1的条件下，调整或交换其中两个词性因子的值，进行多次实验对比；③如果模型性能有提升，则更新对照组为当前最佳性能对应的参数值，重复步骤②和③。表 10 展示了部分参数设定情况下对应的实验结果，其中Weighted F1为三类句子识别F1的加权平均值。结果表明当四个词性因子分别为0.3、0.3、0.3和0.1 时，模型性能最优，这是因为动词是句子结构的核心，且被动句中的施受事一般是名词或代词，有标记被动句中的标记词为介词，三者对被动句识别都非常重要，因此所占权重较大，而其他词性的词语通常只作为修饰成分，对被动句识别的影响不大，因此权重较小。

6 结语

本文首先手工标注了一个被动句语料库，涵盖三种类型的句子——有标记被动句、无标记被动句和非被动句，然后根据被动句的句式特点，将被动句识别任务建模为一个三分类任务，进而提出了一个融合词性信息和动词论元框架信息的PC-BERT-CNN模型，实现了汉语被动句的自动识别。实验结果表明该模型取得了较好的识别效果，且比现有主流自动解析器能更加准确地识别被动句，有标记被动句和无标记被动句识别的F1值分别达到98.77%和96.72%。

我们将在后续工作中，一方面继续扩大语料规模并进行细粒度标注，为更深入的被动句研究提供支持；另一方面，尝试对细粒度被动句语料进行解析，进一步提升被动句解析的性能。

第二十一届中国计算语言学大会论文集，第384页-第394页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Che W, Feng Y, Qin L, and Liu T. 2021. *N-LTP: An Open-source Neural Language Technology Platform for Chinese. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 42-49.

Che W, Li Z, and Liu T. 2010. *Ltp: A chinese language technology platform. Proceedings of the 23rd international conference on computational linguistics*, 13-16.

Devlin J, Chang M W, Lee K, and Toutanova K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the North American Chapter of the Association for Computational Linguistics*, 4171-4186.

Kim Y. 2014. *Convolutional Neural Networks for Sentence Classification. Proceedings of the Association for Computational Linguistics*, 1746-1751.

Nguyen C, Tran V, and Le Nguyen M. 2021. *Enrichment of Features for Malware-Related Sentence Classification using External Knowledge. Proceedings of the 33rd IEEE International Conference on Tools with Artificial Intelligence*, 1144-1148.

Qin Q, Hu W, and Liu B. 2020. *Feature projection for improved text classification. Proceedings of the Association for Computational Linguistics*, 8161-8171.

Xue N, and Palmer M. 2005. *Automatic semantic role labeling for Chinese verbs. Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1160-1165.

蒋坚松. 2002. 英汉对比与汉译英研究. 湖南人民出版社.

鞠彩萍. 2007. “遭”字句——兼论被动标记词的界定与优胜劣汰. 贵州大学学报（社会科学版）, 2007(01):117-121.

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.

李珊. 1994. 现代汉语被字句研究. 北京大学出版社.

乔莎莎. 2015. 有标记被动句研究. 黑龙江大学.

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.

汤敬安. 2016. 汉语无标记被动句与有标记被动句的认知辨析. 云梦学刊, 37(06):110-114.

王灿龙. 1998. 无标记被动句和动词的类. 汉语学习, 1998(05):15-19.

王芸华. 2014. 被动句主语的语义角色考察. 贺州学院学报, 30(02):18-22.

朱向其,张忠林,李林川,马海云. 2021. 基于改进词性信息和*ACBiLSTM*的短文本分类. 计算机应用与软件, 38(12):179-186.

邹丽玲. 2016. 英译汉视角下解析汉语无标记被动句的句法结构. 外语学界, 2016(00):272-281.

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

中文糖尿病问题分类体系及标注语料库构建研究

钱晓波[1] ， 谢文秀[2] ， 龙绍沛[1] ， 兰牧融[1] ， 慕媛媛[3] ， 郝天永[1,*]
[1]华南师范大学，计算机学院，广东广州
[2]香港城市大学，电脑科学系，香港
[3]巢湖学院，外国语学院，安徽合肥
xiaoboqian1221@outlook.com，vasiliky@outlook.com，Shaopei-Lauv@m.scnu.edu.cn，
1460685366@qq.com，myy@chu.edu.cn，haoty@m.scnu.edu.cn

摘要

糖尿病作为一种典型慢性疾病已成为全球重大公共卫生挑战之一。随着互联网的快速发展，庞大的二型糖尿病患者和高危人群对糖尿病专业信息获取的需求日益突出，糖尿病自动问答服务对患者和高危人群的日常健康服务也发挥着越来越重要的作用，然而存在缺乏细粒度分类等突出问题。本文设计了一个表示用户意图的新型糖尿病问题分类体系，包括6个大类和23个细类。基于该体系，本文从两个专业医疗问答网站爬取并构建了一个包含122732个问答对的中文糖尿病问答语料库DaCorp，同时对其中的8000个糖尿病问题进行人工标注，形成一个细粒度的糖尿病标注数据集。此外，为评估该标注数据集的质量，本文实现了8个主流基线分类模型。实验结果表明，最佳分类模型的准确率达到88.7%，验证了糖尿病标注数据集及所提分类体系的有效性。Dacorp、糖尿病标注数据集和标注指南已在线发布，可以免费用于学术研究。

关键词： 糖尿病 ； 问题分类 ； 分类体系 ； 语料库建设 ； 标注

The Construction of Question Taxonomy and An Annotated Chinese Corpus for Diabetes Question Classification

Xiaobo Qian[1],Wenxiu Xie[2],Shaopei Long[1],Murong Lan[1],Yuanyuan Mu[3],Tianyong Hao[1,*]
[1]School of Computer Science, South China Normal University,Guangzhou, Guangdong
[2]Department of Computer Science,City University of Hong Kong, Hong Kong
[3]School of Foreign Languages,Chaohu University,Hefei, Anhui
xiaoboqian1221@outlook.com，vasiliky@outlook.com，Shaopei-Lauv@m.scnu.edu.cn，
1460685366@qq.com，myy@chu.edu.cn，haoty@m.scnu.edu.cn

Abstract

As a typical chronic disease, diabetes has become one of the major global public health challenges. With the rapid development of the Internet, the huge group of type 2 diabetes patients and high-risk people has shown an increasing demand for specialized information on diabetes. The automated diabetes Question Answering (QA) services also play a vital role in providing daily health services for patients and high-risk people. However, issues like fine-grained classification are still unsolved in many QA services. In this paper, we design a new diabetes question classification taxonomy which represents the user intent, including 6 coarse-grained categories and 23 fine-grained categories. We also construct a new Chinese diabetes QA corpus DaCorp that contains 122,732 questions-answer pairs, collected from two professional medical QA websites. Meanwhile, we annotate 8,000 diabetes questions in DaCorp as a fine-grained diabetes dataset. To evaluate the quality of the proposed taxonomy and the annotated dataset, we implement 8 mainstream baseline classifiers for diabetes question classification. Results show that the best-performing model gained an accuracy of 88.7%, demonstrating

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

the validity of the annotated diabetes dataset and the efficacy of the proposed taxonomy. The Dacorp, annotated diabetes dataset, and annotation guidelines are published online and free for academic research.

Keywords: Diabetes , Question Classification , Classification Taxonomy , Corpus Construction , Annotation

1 引言

随着经济的快速发展和生活方式的迅速变化，糖尿病的患病率呈急剧上升趋势，已成为当今重要的公共卫生挑战之一。根据国际糖尿病联合会（International Diabetes Federation，IDF）全球糖尿病地图集的最新报告[1]，2021 年，全球有近5.37亿成年人（20-79岁）患有糖尿病，其中670万人在同年死亡，这意味着每5秒钟就有1人死于糖尿病。中国目前糖尿病患者已达1.4亿，是世界上糖尿病患者人数最多的国家[1]。糖尿病已成为中国21世纪最具挑战的公共健康问题(Jia, 2014)。然而，近50%的糖尿病患者未得到确诊且不了解自己的病情[1]，即存在自诊率低的问题。另外，据世界卫生组织（World Health Organization，WHO）报告[2]，在糖尿病的患病人群中，超过95% 的患者患有二型糖尿病。与无法预防的一型糖尿病不同，二型糖尿病可以通过改变生活方式和提高健康管理能力来预防(Powers et al., 2015)。因此，高质量的糖尿病管理知识和信息对糖尿病患者和糖尿病高危人群至关重要。

据中国互联网络信息中心（China Internet Network Information Center，CNNIC）报告[3]，截至2021年12月，我国网民规模达10.32亿，互联网普及率达73.0%。互联网已成为患者寻找健康信息、表达健康信息需求的重要工具。许多在线健康问答社区和论坛已成为患者提问和分享信息的热门平台。然而，互联网上的健康信息质量参差不齐(Kanthawala et al., 2016)，因而向患者提供可靠的健康信息至关重要。由于现有的问答服务存在缺乏细粒度分类等突出问题，导致患者不能快速地找到与自己病情最相关的糖尿病信息。因此，从患者需求出发并对糖尿病问题进行细类度分类是自动问答服务一个亟需解决的问题，同时也是向用户提供可靠信息的一种有效方式。

本文为辅助患者快速获得最相关的糖尿病信息，设计了一个表示用户意图的新型糖尿病问题分类体系，包括饮食、治疗、预防、并发症等细粒度类别。同时，构建了一个糖尿病中文问答语料库DaCorp，包含122732个问答对。根据提出的分类体系，本文对DaCorp中的8000个糖尿病问题进行人工标注，形成一个糖尿病问题标注数据集，为糖尿病自动问答服务的发展提供数据支撑。Dacorp、标注数据集和标注指南已在网站发布可免费下载及用于学术研究。此外，本文实现了8个主流分类模型，并对各个模型在该标注数据集上的分类性能进行了对比与分析，验证了标注数据集的质量及提出分类体系的有效性。

本文的主要贡献如下：1）设计了一个表示用户意图的糖尿病问题分类体系，以辅助患者快速获得最相关的糖尿病信息；2）构建了一个糖尿病中文问答语料库DaCorp和糖尿病问题标注数据集；3）通过8个主流基线分类模型在糖尿病标注数据集上的分类实验，验证了分类体系的合理性和数据集的有效性，本文的实验结果可作为糖尿病问题分类任务的基准。

2 相关工作

2.1 医学问题分类体系

现有的医学问题分类体系研究主要分为两类，一类是基于问题性质和主题的分类体系，另一类是面向用户意图的分类体系。Ely et al. (2000)提出了一种针对临床问题的类型和性质的分类体系，即通用临床问题分类（Taxonomy for Generic Clinical Questions，TGCQ），TGCQ是一个四层结构的分类体系，包含64个通用问题类型。该分类体系

[1]International Diabetes Federation（2021）IDF diabetes atlas.https://diabetesatlas.org/

[2]World Health Organization.https://www.who.int/news-room/fact-sheets/detail/diabetes

[3]China Internet Network Information Center：China Internet Network Development State Statistic Report.http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/202202/P020220407403488048001.pdf

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

由来自152位医生提出有关患者护理的临床问题构建而成。另一种针对问题主题的分类体系是基层医疗国际分类[3]（International Classification of Primary Care-Version 2，ICPC-2），它由世界家庭医生组织（WONCA）开发，并广泛用于多个国家。该体系可以用于对患者的就诊原因、管理、干预措施等相关问题进行分类。然而，Boot and Meijman (2010)调查了这两种专业分类体系对患者提出的健康问题进行分类的可行性，发现这两种分类体系并不能直接用于患者健康问题的分类。他们的研究指出患者和医生的信息需求之间存在着差异。在现实生活中，存在很多患者经常询问但医生很少提出的问题，而患者常问的问题无法与这两个分类体系中的类别对应。例如，没有关于标准医学知识（例如，“X 综合征是否存在？”）和患者常问的饮食建议（例如，“X疾病吃什么食物有好处？”）相对应的问题类别。因此，直接使用以医生信息需求为基础的问题分类体系来表达患者意图并对患者所提问题进行分类是不完整且不合适的。

近年来，研究者开始构建和研究面向用户意图的问题分类体系。McRoy et al. (2016)提出了一个癌症问题分类体系，该体系将用户提出的癌症相关问题分为10个类别。Wang et al. (2020)使用卷积神经网络将用户提出的糖尿病健康问题分为9类，构建了一个糖尿病健康问题的分类体系，同时对中国糖尿病患者的健康信息需求进行了分析。Luo et al. (2020)将1000个高血压患者问题分为7个类别，构造了一个高血压问题的分类体系。尽管这些分类体系在一定程度上可以表达用户的健康信息需求，但它们都停留在对问题的表层分类即粗粒度类别分类，缺乏对用户提出的健康问题进行细类度划分，因而不能有效地满足患者快速检索与自身健康最相关的信息需求。基于现有糖尿病问题分类体系存在的问题，本文提出了一个新的问题分类体系，将问题依据用户意图进行细粒度划分，从而能够有效地辅助用户快速检索到与自身健康状况密切相关的问答信息。

2.2 医学问题语料库

近年来，不少国内外研究者对医学问题语料库进行了构建和研究。Ely et al. (2000)使用所提出的64个通用类型对临床问题进行人工标注并构建了一个临床问题语料库。该语料库是由152位家庭医生提出的有关患者护理的1396个临床问题组成。Roberts et al. (2014)构建了一个包含2937个遗传罕见疾病的患者问题语料库，该语料库的问题被标注为13个类别。相较于英文的医学问题语料库，中文医学问题语料库起步较晚。Guo et al. (2018)创建了一个中国健康问题的标注语料库。该语料库由5000个人工标注的中文健康问题组成。这些健康问题由2000个高血压相关问题和3000个来自内科、外科、妇产科、儿科、传染病、中医6个不同疾病领域的问题构成。

尽管现有中文问题语料库的疾病种类范围广，但很少有人构建和研究与糖尿病相关的中文问题语料库。据调研，只有Guo et al. (2020)提出了一个面向自动答疑服务的糖尿病问题语料库。该语料库由6401个带有<实体类型，意图类型>标签的糖尿病问题构成，采用人工标注来保证语料库的质量。然而，该语料库中的糖尿病问题主要与糖尿病的主要特征相关，例如，BMI指数（Body Mass Index）、葡萄糖、糖化血红蛋白、高血压和肌酐，并不能有效地体现用户意图和信息需求。据我们调研，在常用的专业医疗问答网站中，很多糖尿病患者并没有医学背景，且常问的问题更多地集中在与糖尿病相关的一般性问题和日常健康管理相关的问题，例如“二型糖尿病可以吃X食物吗”、“如何预防糖尿病”等。基于此，本文从两个大型专业医疗问答网站爬取和糖尿病直接相关的问答数据，构建了一个中文糖尿病问答语料库和糖尿病问题标注数据集。与其他研究不同的是，本文提出的语料库在分类体系上进行创新，对糖尿病问题进行需求细粒度分析，同时提供了主流分类模型在标注数据集上的问题分类性能，进一步对标注数据集的质量和分类体系的有效性进行评估。

3 糖尿病问题分类体系

分类体系是表示用户意图并系统地分析和记录用户对健康信息需求的一种方法。在问答系统中，问题分类在缩小检索范围和提高返回答案的准确性等方面发挥着重要作用(Zhen et al., 2015)。由于现有的面向用户意图的分类体系缺乏对问题的细类度划分，因此本文基于TGCQ分类体系设计了一个新的可以表示用户意图的糖尿病问题双层分类体系，包括6个大类和23个细类。尽管TGCQ分类体系是基于医生提出的有关患者护理的问题构建而成，并不能直接用于对

[3]International Classification of Primary Care-Version 2（ICPC-2）：https://www.ehelse.no/kodeverk/icpc-2e–english-version

(c) 2022 中国中文信息学会计算语言学专业委员会

用户健康问题的分类(Boot and Meijman, 2010)。但由于医生和患者提出的问题在“治疗”、“诊断”类别存在共性，本文对TGCQ分类体系进行调整和扩展后可用于用户糖尿病健康问题的分类。

针对TGCQ分类体系在用户糖尿病问题分类上存在的问题，本文对分类体系进行了调整和改进。首先，由于患者和医生的信息需求差异，存在很多患者经常询问但医生很少提出的问题，因此患者的常问问题无法与该分类体系中的类别相对应。基于此，为了尽可能涵盖患者提出的糖尿病问题，对于第一层的大类（coarse categories），本文增加了“常识”、“健康生活方式”和“其他”三个类别。保留了TGCQ分类体系中的“诊断”、“治疗”和“流行病学”三个大类，去除TGCQ分类体系中的“管理”和“非临床”两个类别。

其次，在TGCQ分类体系中的细类很多并不适宜用于糖尿病的问题分类，例如，“物理特性”、“社区服务”等。因此，对于第二层的细类（fine-grained categories），本文只保留了“临床解释”、“症状/表现”、“检查”和“病因学”4个类别。同时，增加了更符合糖尿病患者意图相关的19个类别，例如，“并发症”、“饮食”、“锻炼”、“预防”等。本文通过多轮人工标注对分类体系进行不断修改，最终确定的糖尿病问题双层分类体系为：第一层包括6个大类：“诊断”、“治疗”、“常识”、“健康生活方式”、“流行病学”和“其他”，第二层细类主要包括“检查”、“药物选择”、“生育力”、“饮食”和“预防”等23个类别。该分类体系对TGCQ分类体系进行了调整和扩展，使其能有效地体现患者意图和信息需求，从而可以辅助患者快速获得最相关的糖尿病信息。具体的分类体系结构如图1所示。

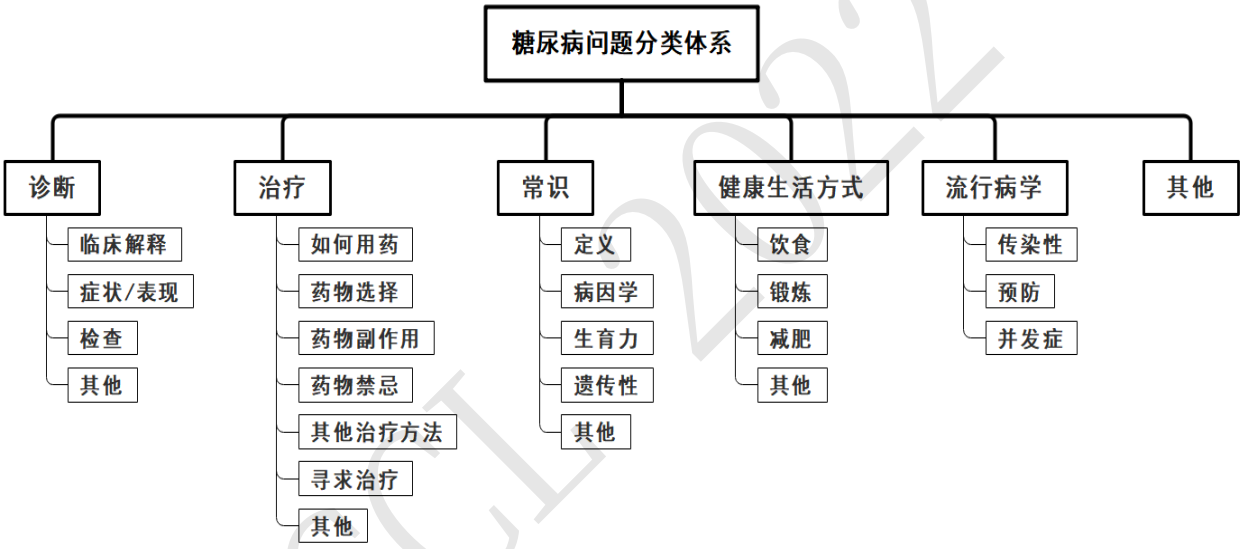


Figure 1: 糖尿病问题分类体系结构

4 语料库标注与构建

4.1 数据收集

糖尿病问答语料库DaCorp的数据收集于两个中国大型医疗问答网站:“39健康”[4]与“有问必答”[5]。在这两个网站上，用户可以提出与医疗护理相关的问题，并在提交问题时提供病情的详细描述。此外，用户提交的每个问题都会由国内医院的专业医生给出回复，同时用户咨询的问题具有高覆盖性和多样性等特点。语料库的构建主要包括医疗问答网站原始数据的收集和数据清洗两个步骤。首先，本文开发了一个基于Beautiful Soup 库[6]的python脚本，自动从网站上爬取以“糖尿病”为关键词查询到的问答数据。数据包括用户提出的问题和医生给出的相应回复，共收集196915条。其次，我们对收集的原始数据进行数据清洗和预处理，以便后续的人工标注。

[4]http://www.39.net/
[5]http://club.xywy.com/
[6]Python Library Beautiful soup. https://pypi.org/project/

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

数据清洗主要是对重复和不相关的数据的进行过滤，去除如广告等非健康相关的内容。此外，由于网站上的问题是由患者提出的，他们大多没有医学背景，因而问题中存在大量的自然语言描述，非常口语化且存在很多错别字。例如，许多患者可能会将“二甲双胍”输入为“二甲双瓜”，将”妊娠糖尿病”输入为“妊辰糖尿病”等。因此，我们对问题中的错别字进行预处理，即人工纠正。在数据清洗和预处理后，最终的糖尿病问答语料库DaCorp包含122732 条问答对。

4.2 数据标注

在糖尿病问答语料库DaCorp构建完成后，本文从语料库中随机抽取了8000个问题进行人工标注，并形成一个糖尿病问题标注数据集。图2为数据的人工标注流程。数据标注分为标注准备和数据正式标注两个阶段。在标注准备阶段，通过对现有的分类体系和标注指南的研究和分析，本文设计了最初版本的糖尿病问题分类体系和对应的标注指南，其中标注指南包括每个类别的标注规则，并给出了相应的通用问题模式和示例问题，以提高分类体系的合理性、可用性和标注一致性。表1是类别“治疗”中每个细类的标注指南。分类体系的完整标注指南可从我们发布的网站中获取。同时，为了减少人工标注的繁重工作量并加快标注过程，本文基于Tkinter库[7]开发了一个简易的人工标注工具来辅助数据标注工作。在数据标注的正式标注阶段，由三位具有标注经验的硕士研究生参与标注工作。首先，数据集中的8000个糖尿病问题由一位具有医学信息学背景的标注者进行标注，另外两名标注者对数据集中随机抽取的2000个糖尿病问题进行标注，每人标注1000个。初始标注完成后，我们对三位标注者的标注结果进行一致性评估，对出现分歧的问题进行讨论，在协商一致后，修改分类体系和对应的标注指南。然后，三位标注者按照修改后的标注指南，独立标注剩余的6000个问题，每人标注2000个。最后，比较三位标注者的数据标注结果，对标注过程中出现的分歧进行讨论以达成一致，同时修改和完善标注指南的最终版本。

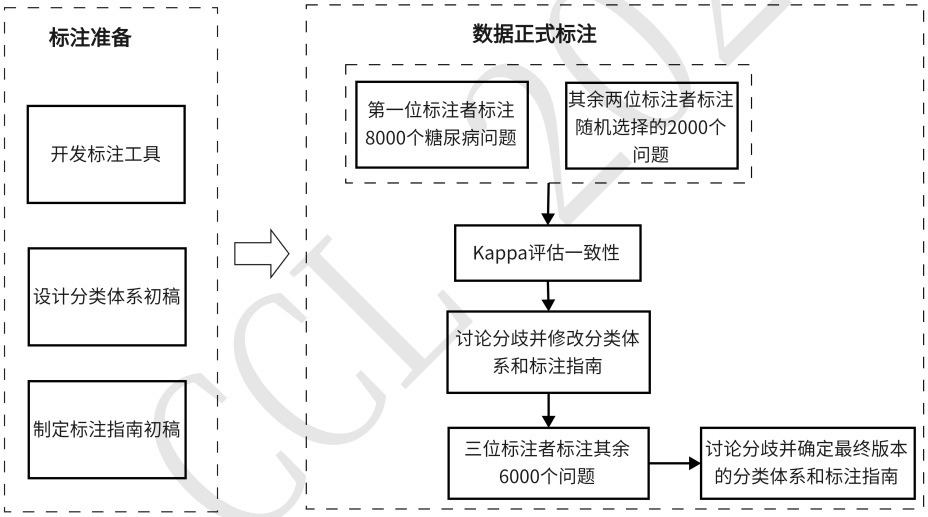


Figure 2: 数据标注流程

为了评估数据集标注的一致性和质量，发现标注者之间可能存在的分歧，本文使用Kappa统计量来检验标注者间一致性（Inter Annotator Agreement，IAA）。Kappa是IAA的衡量标准(McHugh, 2012)，它可以纠正偶然发生的一致性。Kappa的计算如公式1所示。

$$Kappa = \frac{(P_o - P_e)}{(1 - P_e)} \tag{1}$$

其中P_o是总体分类精度，即每一类正确分类的样本数量之和除以总样本数，P_e是偶然发生的一致性(Elliott and Woodward, 2007)。Kappa值越大，表明标注一致性越好。在本研究中，如果标注者标注的大类和细类都相同，则视为一致。在标注过程中，三位标注者的平均标注一致性为0.78，表明人工标注结果的差异较少，分类体系的制定合理且有效。

[7]Python interface to Tcl/Tk. https://docs.python.org/3/library/tkinter.html

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

细类	标注规则	通用问题模式（部分）	示例问题（部分）
如何用药	患者知道使用什么药物，但是不知道药物使用的时间、剂量、注意事项。	X疾病服用Y药物是饭前还是饭后?	二型糖尿病吃二甲双胍是饭前吃还是饭后吃?
药物选择	患者已知自己的病情，需要了解合适的药物。	X疾病Y症状需要服用什么药物?	一型糖尿病人发烧能吃些什么药物?
药物副作用	患者不确定服用某种药物是否有副作用，以及副作用的详情。或者患者不确定某种症状是否是药物的副作用。	X疾病服用Y药物有副作用吗?	糖尿病吃盐酸二甲双胍缓释片有副作用吗?
药物禁忌	患者想了解服用某种药物的注意事项和禁忌。	X疾病服药期间可以打Y疫苗吗?	糖尿病患者服药期间可以打乙肝疫苗吗?
其他治疗方法	患者想了解能否通过非药物的方式治疗疾病，以及非药物治疗的类型、效果和风险。	X疾病能手术治疗吗?	二型糖尿病能手术治疗吗?
寻求治疗	患者描述自身的症状想寻求帮助或治疗。	X疾病患有Y症状如何治疗?	糖尿病人身上痒如何治?
其他	患者的问题与治疗相关，但不属于其他细类。	X疾病最好的治疗方法是什么?	一型糖尿病最好的治疗方法是什么?

Table 1: “治疗”类别中每个细类的标注指南

4.3 标注结果

糖尿病标注数据集的大类标注结果分布如图3所示。在8000个糖尿病问题中，从第一层大类的角度看，常识（C）和健康生活方式（D）类别在数据集中出现频率最高，分别为1655（20.7%）和2226（27.8%）个，其次是治疗（B）类别2026（25.3%）个。这些数据表明，在医患平台上查询糖尿病相关问题的用户中大多为糖尿病患者，且对糖尿病的日常健康管理的关注度较高。同时，还有717（9%）个问题分到诊断类别（A）中，此数据反映了可能存在相当数量的人群已有糖尿病症状却还未被诊断。其余，807（10.1%）个问题与流行病学（E）相关，其他类型（F）的问题有569个（7.1%）。特别地，标注者在标注过程中发现用户对健康生活方式（D）中饮食（D1）类别的关注度尤为突出，这反映了患者越来越重视饮食在糖尿病预防和病情控制过程中的作用。

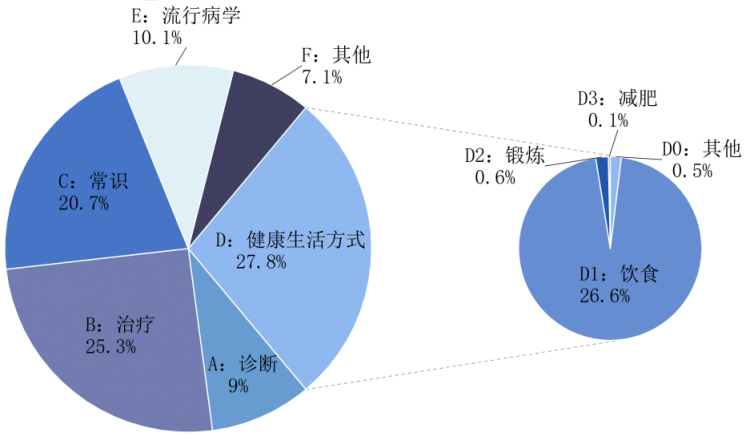


Figure 3: 语料库大类标注结果分布

(c) 2022 中国中文信息学会计算语言学专业委员会

糖尿病标注数据集的细类标注结果如表2所示。从第二层细类的角度看，与治疗（B）相关的问题，主要是寻求治疗（B6；788/2026）和药物选择（B2；514/2026）。用户寻求治疗的问题大都集中在患有糖尿病的情况下如何对正常疾病进行治疗，例如“糖尿病人喉咙痛怎么办？”、“糖尿病人感冒了能吃一般的感冒药吗”。尽管“喉咙痛”和“感冒”是很常见的症状和疾病，但是糖尿病患者仍然担忧常见疾病发生在糖尿病患者身上需要特别处理和治疗。同时，我们从用户询问药物选择相关的问题中发现，用户除了想了解糖尿病常见治疗药物之外，对常见疾病的治疗药物也很关注。此外，在有关诊断（A）的患者询问问题中，用户想通过临床解释和症状来诊断糖尿病的问题占主体的88.7%（636/717）。在糖尿病常识（C）类别中，患者询问的问题包括定义（C1）、病因学（C2）、糖尿病的生育力（C3）和糖尿病的遗传性（C4）,其中“其他”类别（C0）所占C类的比例为31.8%(527/1655)，通过仔细观察和分析语料库，发现患者提出的与糖尿病常识相关的问题，种类非常繁多且不集中，例如“2型糖尿病可以拔牙吗”、“糖尿病患者需要补充什么维生素”、“糖尿病患者能上大学吗”、“糖尿病患者在夏季需要注意什么”等问题，这表明患者对于与糖尿病相关的常识问题具有种类多、数量高的特征。

同时，患者也咨询了保持健康生活方式（D）的各种方法，包括饮食（D1）、锻炼（D2）、减肥（D3）等许多方面。一些患者意识到不良的生活方式可能会加重糖尿病的病情，而良好的生活方式可以缓解病情、改善健康。此外，患者还咨询了流行病学（E）的相关问题，他们主要想了解糖尿病的传染性（E1）、预防（E2）以及并发症（E3）相关的信息。最后，有569个如“1型糖尿病不治疗能活多久”、“甲亢合并糖尿病可怕吗”、“糖尿病患者可以带胰岛素上飞机吗”等不属于以上5大类别的问题被归于其他类别。

大类	细类	数量
A:Diagnosis(诊断)	A0: other(其他)	16
	A1:interpretation of clinical（临床解释）	344
	A2:symptom/manifestations（症状/表现）	292
	A3:test（检查）	65
B:Treatment（治疗）	B0:other（其他）	388
	B1:how to use drug（如何用药）	101
	B2:drug choice（药物选择）	514
	B3:adverse effects of drug（药物副作用）	69
	B4:contraindications of drug（药物禁忌）	4
	B5:Other Therapy（其他治疗方法）	162
	B6:Treatment Seeking（寻求治疗）	788
C:Common Knowledge（常识）	C0:other（其他）	527
	C1:Definition（定义)	152
	C2:Etiology（病因学）	860
	C3:Fertility（生育力）	67
	C4:Hereditary（遗传性）	49
D:Healthy lifestyle（健康生活方式）	D0:other（其他）	43
	D1:Diet（饮食）	2126
	D2:Exercise（锻炼）	50
	D3:weight-losing（减肥）	7
E:Epidemiololgy（流行病学）	E1:Infect（传染性）	31
	E2:Prevention（预防）	34
	E3:Complication（并发症）	742
F:Other（其他）	569	

Table 2: 语料库标注结果

语料库的标注结果表明，患者主要关心糖尿病诊断的方法和症状，如为什么患有糖尿病会出现某些症状，如何治疗，他们是否能够服用特定药物，使用是否有副作用或禁忌，糖尿病的遗传性和生育力，以及他们在日常生活中可以采取怎样的措施来改善或预防他们的病情。

第二十一届中国计算语言学大会论文集，第395页-第405页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

5 实验与结果

5.1 实验设置

为了评估分类体系的合理性和标注数据集的质量，本文比较了8个主流分类模型在糖尿病标注数据集上大类的分类性能，并将8000条糖尿病标注问题进行随机划分，其中6000条数据作为训练集，1000条作为验证集，其余1000条作为测试集。分类模型采用的是6个深度神经网络模型Text CNN(Chen, 2015)、Text RNN(Liu et al., 2016)、Text RCNN(Lai et al., 2015)、Text RNN Attention(Zhou et al., 2016)、fastText(Joulin et al., 2016)、DPCNN (Johnson and Zhang, 2017)，以及两个预训练的大规模语言模型BERT(Devlin et al., 2018)和ERNIE(Sun et al., 2019)进行实验。对于深度神经网络模型，本文使用常用的Jieba[8]分词工具对数据进行中文分词。除fastText可自行训练词向量外，其它模型使用预训练的搜狗新闻词向量(Li et al., 2018)作为特征。其中，Text RNN和Text RNN Attention隐藏层数为128，Text RCNN和fastText隐藏层数为256。同时我们使用Adam(Kingma and Ba, 2014)优化器以0.001的学习率最小化交叉熵，并使用提前停止机制避免过拟合的问题。对于预训练的语言模型，本文主要对BERT和ERNIE进行微调，将学习率设置为5e-5并采用提前停止，epoch数目设置为12。

5.2 实验结果

本文采用的评测指标是分类准确率（Accuracy），即对于给定的测试数据集，分类模型正确分类的样本数与模型总样本数之比。不同分类模型在标注数据集上的分类性能结果如表3所示。从实验结果可以看出，在深度神经网络模型中，Text CNN的表现最佳，准确率为86.1%；DPCNN性能最低，准确率仅为82.8%。尽管DPCNN拥有最深的网络结构和最多的参数，能够捕获文本的长距离特征，但用户提出的糖尿病问题通常比较简短，因而TextCNN在短文本分类中性能要优于DPCNN。对于预训练的语言模型，ERNIE模型的准确率（88.7%）要高于BERT 模型（87.8%）。与BERT模型主要学习字级别的信息相比，ERNIE能利用文本的词法结构和语法结构，直接对先验语义知识单元进行建模，增强了模型完整概念的语义表示能力，有助于对文本进行理解和分类。同时，预训练的大规模语言模型BERT和ERNIE要优于其他深度神经网络模型，但性能差别不大。这与分类任务的训练样本数量有关，预训练语言模型BERT和ERNIE在大数据集上的优势应该会更明显。因此，在后续的工作中，本文将会继续扩大标注语料，完善标注数据集。

模型	Text CNN	Text RNN	Text RNN Attention	Text RCNN	fastText	DPCNN	BERT	ERNIE
准确率(%)	86.1	83.8	83.6	84.2	84.7	82.8	87.8	88.7

Table 3: 糖尿病标注数据集在8个分类模型上的实验结果

此外，本文对最佳分类模型ERNIE在糖尿病标注数据集上每个大类的分类性能进行对比和分析，实验结果如表4所示。采用的评价指标是查准率（Precision）、召回率（Recall）和F1值（F1-score），其中F1值是模型查准率和召回率的一种调和平均。从模型的分类结果中可以看出，“健康生活方式”类别的分类性能最好（F1：94.44%）。因患者关于“健康生活方式”的糖尿病问题主要集中在“饮食”方面，而“饮食”相关问题的文本特征明显（例如，“糖尿病可以吃X食物吗？”），因此利于模型分类。其次，模型在“治疗”、“诊断”、“常识”和“流行病学”类别上也具有较高的分类性能，F1均高于85%。而模型在“其他”（F1：64.%）这个类别的分类性能较低。模型的分类性能与数据集的数量相关，类别数据量较大时，模型能够得到充分训练，故分类效果较好。由于训练集中“其他”这个类别的数量较少，因此模型在这个类别的训练上可能存在欠拟合的问题，因而分类性能较低。

主流文本分类模型在标注数据集上的分类结果显示，分类模型的准确率均高于82%，最佳模型ERNIE在6个大类的平均F1为86.06%，证明本文提出的糖尿病问题分类体系合理且有效，同时人工标注的糖尿病问题数据集具有较高的质量，可为糖尿病问题分类任务和相关的问答服务系统研究提供有效的数据支撑。

[8]https://github.com/fxsjy/jieba

(c) 2022 中国中文信息学会计算语言学专业委员会

	Precision(%)	Recall(%)	F1-score (%)
诊断	94.05	90.80	92.40
治疗	84.98	93.96	89.25
常识	87.50	83.87	85.65
健康生活方式	95.51	93.41	94.44
流行病学	90.11	91.11	90.61
其他	70.18	58.82	64.0

Table 4: ERNIE在糖尿病标注数据集上每个大类的分类性能

5.3 语料库访问

为了方便用户和研究人员访问和使用语料库，本文开发了一个语料库网站，可以通过URL[9]访问数据。该网站主要分为4个主要模块，分别为：分类体系、标注指南、标注数据、数据下载。在“分类体系”页面，用户可以浏览糖尿病问题的分类体系。在“标注指南”页面，如图4所示，用户可以查看每个类别的标注规则，以及对应的问题模式和示例问题。在“标注数据”页面可以查看标注数据集中的糖尿病问题和答案，以及每个问题对应的大类和细类。最后，用户可以在“数据下载”页面选择XML格式下载标注的糖尿病问题和DaCorp中所有的糖尿病问答对。

Figure 4: 语料库网站的“标注规则”页面

图5显示了XML格式的问题示例，以便用户根据需求选择特定格式的语料库。从标注示例来看，第一行是XML声明，它定义了XML的版本。下一行描述了根元素<QAPairs>，它在XML文件中是唯一的。子元素<QAPair>包含了所有的糖尿病问题和答案。其中，子元素<question>、<answer>、<url>分别指的是患者提出的糖尿病问题，医生给出的答复和问答对的来源。<main_category>和<sub_category>表示示例问题被标注的大类和细类。XML格式文件将语料库可视化，以加强对问答语料库的直观理解。

[9]http://47.102.207.52:9090

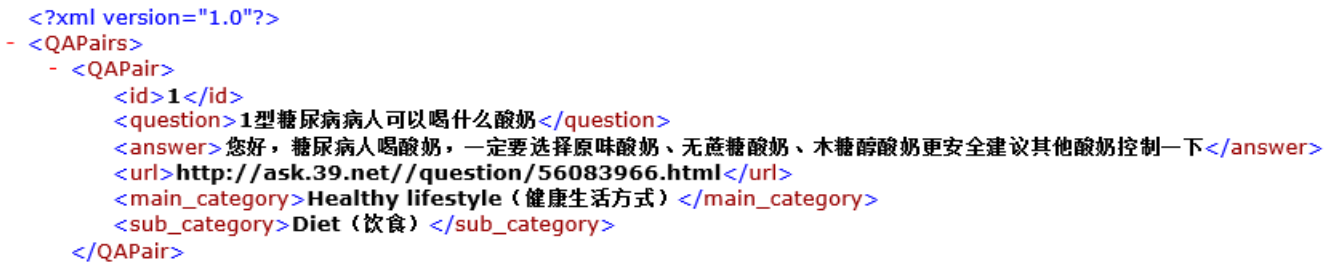


Figure 5: XML格式的标注问题示例

6 总结与未来工作

本文为辅助患者快速获得最相关的糖尿病信息，设计了一个表示用户意图的新型糖尿病问题分类体系。同时，构建了一个糖尿病问答语料库DaCorp，并使用新的分类体系对问题进行人工标注形成糖尿病标注数据集。最后，本文评估了8个主流分类模型在标注数据集上的分类性能，实验结果验证了数据集的有效性以及提出分类体系的合理性。据调研，本文提出的标注语料库是目前最大的糖尿病问题标注语料库。该标注语料库可通过网站公开访问，用于训练机器理解糖尿病患者的中文健康问题。本研究将为糖尿病问题分类、问答匹配等NLP相关的任务以及自动问答系统的开发提供数据支撑。此外，我们将不断完善糖尿病问题分类体系，并尝试使用主动学习来进行半自动化标注，扩大标注语料，提高标注效率，进一步挖掘和分析糖尿病问题的特征和糖尿病患者的信息需求。同时，利用现有语料库和分类体系开发糖尿病问题的高性能分类模型也是未来的重要工作。

参考文献

Cécile RL Boot and Frans J Meijman. 2010. Classifying health questions asked by the public using the icpc-2 classification and a taxonomy of generic clinical questions: an empirical exploration of the feasibility. *Health communication*, 25(2):175–181.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

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*.

Alan C Elliott and Wayne A Woodward. 2007. *Statistical analysis quick reference guidebook: With SPSS examples*. Sage.

John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *Bmj*, 321(7258):429–432.

Haihong Guo, Xu Na, and Jiao Li. 2018. Qcorp: an annotated classification corpus of chinese health questions. *BMC medical informatics and decision making*, 18(1):39–47.

Xusheng Guo, Likeng Liang, Yuanxia Liu, Heng Weng, and Tianyong Hao. 2020. The construction of a diabetes-oriented frequently asked question corpus for automated question-answering services. In *Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare*, pages 60–66.

Weiping Jia. 2014. Diabetes: a challenge for china in the 21st century. *The Lancet Diabetes & Endocrinology*, 2(4):e6–e7.

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, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

(c) 2022 中国中文信息学会计算语言学专业委员会

Shaheen Kanthawala, Amber Vermeesch, Barbara Given, Jina Huh, et al. 2016. Answers to health questions: internet search results versus online health community responses. *Journal of medical Internet research*, 18(4):e5369.

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

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Aijing Luo, Zirui Xin, Yifeng Yuan, Tingxiao Wen, Wenzhao Xie, Zhuqing Zhong, Xiaoqing Peng, Wei Ouyang, Chao Hu, Fei Liu, et al. 2020. Multidimensional feature classification of the health information needs of patients with hypertension in an online health community through analysis of 1000 patient question records: observational study. *Journal of Medical Internet Research*, 22(5):e17349.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Susan McRoy, Sean Jones, and Adam Kurmally. 2016. Toward automated classification of consumers' cancer-related questions with a new taxonomy of expected answer types. *Health informatics journal*, 22(3):523–535.

Margaret A Powers, Joan Bardsley, Marjorie Cypress, Paulina Duker, Martha M Funnell, Amy Hess Fischl, Melinda D Maryniuk, Linda Siminerio, and Eva Vivian. 2015. Diabetes self-management education and support in type 2 diabetes: a joint position statement of the american diabetes association, the american association of diabetes educators, and the academy of nutrition and dietetics. *Diabetes care*, 38(7):1372–1382.

Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. 2014. Annotating question types for consumer health questions. In *Proceedings of the Fourth LREC Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Tian-Hao Wang, Xiao-Feng Zhou, Yuan Ni, and Zhi-Gang Pan. 2020. Health information needs regarding diabetes mellitus in china: an internet-based analysis. *BMC Public Health*, 20(1):1–9.

Lihua Zhen, Xiaolin Wang, Sichun Yang, et al. 2015. Overview on question classification in question-answering system. *Journal of Anhui University of Technology (Natural Science)*, 32(1):48–54.

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.

CCL 2022

古汉语嵌套命名实体识别数据集的构建和应用研究

谢志强[1], 刘金柱*, 刘根辉[2]
华中科技大学，人文学院，湖北，武汉，430000
xiezhiqiang2020@163.com
1152822887@qq.com
genhuiliu@163.com

摘要

本文聚焦研究较少的古汉语嵌套命名实体识别任务，以《史记》作为原始语料，针对古文意义丰富而导致的实体分类模糊问题，分别构建了基于字词本义和语境义2个标注标准的古汉语嵌套命名实体数据集，探讨了数据集的实体分类原则和标注格式，并用RoBERTa-classical-chinese+GlobalPointer模型进行对比试验，标准一数据集F1值为80.42%，标准二F1值为77.43%，以此确定了数据集的标注标准。之后对比了六种预训练模型配合GlobalPointer在古汉语嵌套命名实体识别任务上的表现。最终试验结果：RoBERTa-classical-chinese模型F1值为84.71%，表现最好。

关键词： 古汉语 ； 嵌套命名实体识别 ； 数据集 ； GlobalPointer

Construction and application of classical Chinese nested named entity recognition data set

Zhiqiang Xie[1], Jinzhu Liu*, Genhui Liu[2]
School of Humanities,Huazhong University of Science and Technology,Hubei,Wuhan,430000
xiezhiqiang2020@163.com
1152822887@qq.com
genhuiliu@163.com

Abstract

This paper focuses on the less studied task of classical Chinese nested entity recognition, and constructs a data set of classical Chinese nested entity with Historical Records as the original corpus.For the fuzzy problem of entity classification caused by the rich meaning of classical Chinese, this paper constructs classical Chinese nested named entity data sets based on two annotation standards: word original meaning and context meaning, and discusses the entity classification principles and annotation formats of the data sets.A comparative experiment with RoBERTa-classical-chinese+globalpointer model are used to determine the annotation standard of the data set. The F1 value of standard one data set is 80.42%, and that of standard two is 77.43%, which determines the annotation standard of the data set. Then, we compares the performance of six pre-training models combined with globalpointer in the task of classical Chinese nested named entity recognition.Finally,the F1 value of RoBERTa-classical-chinese model is 84.71%, which performs best.

Keywords: Classical Chinese , Nested Named Entity Recognition , Data Set , GlobalPointer

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目:华中科技大学人文社会科学重大原创性成果培育项目“《册府元龟》引书研究”（21WKFZZX016）
华中科技大学自主创新研究基金专项任务项目“《册府元龟》语料库建设、整理与研究”（2020WKYXZX004）
通讯作者:刘金柱

1 引言

命名实体识别(Named Entity Recognition,NER)作为自然语言处理中的基础任务，是开展信息抽取、构建知识图谱等上游任务的重要一环。关于它的研究最早开始于上世纪60 年代(Grishman and Sundheim, 1995)，具体是指识别出待处理文本中预定义好的命名实体，包括实体边界的识别以及实体类型分类两个任务。之后，随着研究的深入，从实体层级角度又进一步细分为扁平命名实体识别任务和嵌套命名实体识别任务。其中，扁平命名实体识别（Flat NER）任务中，每个实体是最小的单位，不能继续拆分。而嵌套命名实体（Nested NER），又称实体重叠，表示在一个实体的内部还存在着一个或多个其他的实体。

目前，随着基于BERT的自然语言处理预训练模型的演化发展及迁移学习，现代汉语命名实体识别任务得到了相对有效的解决，并在此基础上进一步提升了上游任务的各项指标。但比较而言，古汉语命名实体识别研究则相对较少，仅有少数研究集中在古汉语扁平命名实体识别研究，古汉语嵌套命名实体识别研究则寥寥无几。究其原因，一是相关开源的古汉语高质量语料较少；二是古汉语相较于现代汉语而言，在实体分类原则、标注标准选取等方面处理难度更大，要求研究者除了有工程能力外，还要有扎实的古汉语专业知识做支撑。

针对上述问题，本研究基于人工精校后的《史记》，人工标注构建了古汉语嵌套命名实体识别数据集，并尝试使用GlobalPointer作为解码层配合六种预训练模型，开展面向《史记》的古汉语嵌套命名实体识别研究工作，以期为更细粒度的古汉语实体识别和信息抽取提供经验。

2 相关工作

2.1 嵌套实体数据集构建的相关工作

近年来，为了开展嵌套命名实体识别任务而构建的数据集主要有ACE语料(Mitchell et al., 2005)、GENIA语料(Zhou et al., 2004)、SciERC语料库(Ringland et al., 2019)、KBP2015语料库、NNE数据集(Ringland et al., 2019)、人民日报语料库、CADEC(Karimi et al., 2015)等。

2.2 古汉语命名实体识别相关工作

古汉语命名实体识别研究按照其历史发展进程主要分为基于规则和词典匹配的命名实体识别、基于统计机器学习的命名实体识别、基于深度学习的命名实体识别。

2.2.1 基于规则和词典匹配

基于规则和词典匹配，其思想在于观察特定领域文本中的语法规则，从而归纳并设计出特定实体的提取规则以完成提取。曾艳和侯汉清(2008)基于N元语法提出了一种古文自动抽词方法，使用N-gram 对古文语料自动分词，利用抽词词典和停用词词典匹配人名、地名、书名、官职名等词汇，最终对N 元组进行过滤并人工判别选词；朱锁玲等人(2011)通过统计《古今地名对照表》等资料中的古代地名，构建了地名词典，并从《大埔县志》等地方志中抽取地名的上下文特征构造地名识别规则库，以《方志物产》作为语料，对物产地名进行实体识别。皇甫晶等人(2013)通过手动构建规则，对纪传体古汉语文献进行姓名的实体识别。

2.2.2 基于统计机器学习

基于统计机器学习的方法，是将实体识别任务细化成一个多分类问题或者序列标注问题，即将该任务转化成为一个基于字符的分类问题，通过已标注的数据训练模型，将不同字符映射成为不同标签的过程。黄水清等人(2015)在基于先秦古汉语语料库基础上，使用条件随机场模型构建特征，对地名实体进行识别。同样基于条件随机场模型进行古汉语命名实体识别的有李娜等(2018)、王东波等(2018)、袁悦等(2019)。

2.2.3 基于深度学习

深度学习方法也是基于多分类问题和序列标注问题两个任务，但是通过完成一个比统计机器学习更加复杂的建模过程，达到一个更好的任务效果。一个标准的NER深度学习模型一般由输入层、编码层和解码层三层结构建模成。其中，新增的复杂编码层需要解决特征抽取的问题，以捕获实体上下文的特征表示。而经典的特征器包括卷积神经网络（CNN）、循环神经网络（RNN）、递归神经网络、Transformer神经网络(Lample et al., 2016)以及语言模型网络等。崔竟烽等人(2020)通过人工标注数据集，使用深度学习的BERT模型对菊花古典诗词进命名实体识别。刘忠宝等人(2020)使用BERT+Bi-LSTM+CRF对《史记》中的实体进行了识别。

(c) 2022 中国中文信息学会计算语言学专业委员会

3 古汉语嵌套命名实体识别任务研究

3.1 任务定义

嵌套命名实体识别问题可以形式化表示为：给定一个序列$X = \{x_1, x_2, ..., x_n\}$，其中$X_n$表示序列的第n个词条，预测该序列的标签$Y = \{y_1, y_2, ..., y_n\}$。与非嵌套命名实体识别不同的是，嵌套命名实体识别的词条标签Y_n表示多标签而不是单标签，$Y_n = \{y_n^1, y_n^2, ..., y_n^m\}$，其中m为嵌套层数。

3.2 数据集来源

《史记》，二十四史之一，最初称为《太史公书》、《太史公记》、《太史记》，是西汉史学家司马迁撰写的纪传体史书，是中国历史上第一部纪传体通史，作品中撰写了上至上古传说中的黄帝时代，下至汉武帝太初四年间共3000多年的历史。

本文工作所采用的语料来源为2014年中华书局点校本。精校的《史记》版本已经去除了文本中的特殊字符，并在此基础上进行了分行，目前标注完成的数据集字符数为57015，平均句子长度19.97。

3.3 数据集标注

3.3.1 标注原则的确定

数据集标注遵循张欢(2020)提出的简单性原则、易操作性原则、一致性原则。

其一，简单性原则。本研究将古汉语实体划分为五大类，包括“人（PER）”、“地点（LOC）”、“官职（JOB）”、“书（BOOK）”和“时间（TIME）”。“人（PER）”这类实体没有根据其构造成分细化。剔除了“组织（ORG）”，将其合并到“地点（LOC）”，合并“朝代（DYN）”和“年号（REI）”为“时间（TIME）”，类别数量适中，遵循了简单性原则。

其二，易操作性原则。实体的标注包括实体边界和实体类型两个环节。本研究标注实体时，以实体开头为起，至实体结尾为止，进行全选，之后选择实体类型，即完成标注。此外，标注实体均是拆分至最小实体单位为止。标注方式遵循了易操作性原则。

其三，一致性原则。实体类型定义是进行实体识别任务的第一个步骤，通常情况下，研究者都会根据研究的具体需求和侧重来确定需要识别的实体类型。因此，在实体标注时，对于实体类型的确定存在一定的主观性。本研究标注实体时遵循一致性的原则，针对容易混淆的实体标签，选择对这些标签进行合并处理，具体后述。

3.3.2 古汉语实体体系的构建

本研究将古汉语实体划分为五大类，包括“人（PER）”、“地点（LOC）”、“官职（JOB）”、“书（BOOK）”和“时间（TIME）”。

第一类实体是“人（PER）”：一般的，在命名实体识别任务中，“人（PER）”这类实体主要指语料中的人名。但是，古汉语“人（PER）”这类实体的构造成分较之现代汉语要更为复杂，“人（PER）”成分种类繁多，包括名、字、氏、姓、爵位、排行、谥号、官职等。如表1举例所示:

“人（PER）”实体构造成分	举例
名	鄭伯克**段**于鄢
氏+姓	晉獻公欲以**驪姬**爲夫人
官职+名+尊称	**司徒皇父**帅师御之
谥号+排行	**襄仲**欲立之
官职	日夜望**將軍**至，豈敢反乎

Table 1: “人（PER）”实体构造成分及举例

本研究虽是古汉语嵌套命名实体识别，但对于“人（PER）”这类实体的构造成分并不细分。原因有二：一是标注数据集有限，过于复杂的实体划分会导致数据稀疏问题，从而影响最终模型的效果。二是“人（PER）”构造成分的划分存在争议，且本研究的目的也不是人名考证。因此，“人（PER）”这类实体应当遵循简单性原则，除“官职（JOB）”这类区分度高的

(c) 2022 中国中文信息学会计算语言学专业委员会

实体外，剩余皆统一合并到“人（PER）”中。表1中的“司徒皇父”可以通过嵌套标注的方式清晰的划分出实体结构，即“人[官职+名]”，但还出现很多单个表“官职（JOB）”的词汇代指“人（PER）”的情况，如“ 日夜望將軍至，豈敢反乎 ”。又如“诸侯”一词，用“官职（JOB）”代指“人（PER）”或“地点（LOC）”。

第二类实体是“地点（LOC）”：本研究中“地点（LOC）”除了地理意义上的“地名”，还包含“山名”、“水名”等。除此之外，由于《史记》是记载了上至上古传说中的黄帝时代，下至汉武帝太初四年间共3000多年的历史，所以包含很多诸如未统一的“诸侯国”、“部落”、“族群”、“某一氏族”等具有政治意义的地点，这些词细分应为“组织（ORG）”，但遵循简单性原则，同时为了保证样本分布的均衡，本研究将这些词也归类到“地点（LOC）”。

第三类实体是“官职（JOB）”：指在国家机构中担任一定职务的官吏，上至中央大员，下至地方小吏。但正如前文“人（PER）”实体界定所述，有很多表官职的词语实际语境中用来代指“人（PER）”，因此在标注时结合上下文进行了判别，尽可能保证实体归类的准确性。同时，也设计了根据字词本义的标注方式，用于对比补充实验。

第四类实体是“书（BOOK）”：包含书籍、诗歌、文章等。整体样本量较少，但界限分明。

第五类实体是“时间（TIME）”：只包括“朝代（DYN）”和“年号（REI）”，并不包含“季节”、“月份”等实体，是本研究人为规定的狭义“时间（TIME）”。“朝代（DYN）”是界定某一个统一政权时期的名词，但由于《史记》的历史跨度有限，导致“朝代（DYN）”这类实体数量极小，并不适合单独成为一个实体类别。而用“年号（REI）”纪年，是从汉武帝开始的，因此“年号（REI）”数量更是稀少，于是合并两类实体为“时间（TIME）”。

以上是对本研究中古汉语各类实体的界定范围和界定原因的说明。对于实体分类界限模糊的问题，本研究设计了两个标注标准：一方面主要是根据字词本义进行标注，同时也尝试了结合上下文具体语境对字词进行标注，构建了2个数据集，开展对比补充实验，根据最终的F1值来考察2种标注方式的优劣。

3.3.3 数据集标注格式

数据集标注的格式参考2021年阿里天池发布的中文医疗信息处理挑战榜CBLUE(Chinese Biomedical Language Understanding Evaluation)包含的中文医学命名实体识别任务的数据集(Hongying et al., 2020)。

具体的格式为：整个数据集为一整个json，里面每一条数据为一个json，内部是句子、实体的起始位置、实体类别和实体。举例如下：


```
{
  "text": "而蚩尤最爲暴，莫能伐。",
  "entities": [
    {
      "start_idx": 1,
      "end_idx": 2,
      "type": "PER",
      "entity": "蚩尤"
    }
  ]
},
```


Figure 1: 扁平命名实体的标注格式

(c) 2022 中国中文信息学会计算语言学专业委员会

```
{
  "text": "懿王之時，王室遂衰，詩人作刺。",
  "entities": [
    {
      "start_idx": 1,
      "end_idx": 1,
      "type": "JOB",
      "entity": "王"
    },
    {
      "start_idx": 0,
      "end_idx": 1,
      "type": "PER",
      "entity": "懿王"
    }
  ]
},
```


Figure 2: 嵌套命名实体的标注格式

4 模型选择

本研究基于目前主流的方式进行命名实体识别模型的训练，即采用基于深度学习方式分别搭建六个预训练模型，并尝试将GlobalPointer作为解码层，以优化模型性能。

4.1 古汉语预训练模型

在命名实体识别研究中，基于深度学习的方式逐渐取代基于统计机器学习的方法，主要是由于深度学习模型通过构建更为复杂的编码层，表现出更为显著的特征抽取性能。尤其是在BERT出现之后，基于深度学习的方式更是成为了新的基准方式，并出现一系列基于BERT的改进变体模型，RoBERTa就是其中的典型代表。RoBERTa模型不仅继承了BERT模型的优势，而且用更大的单次训练样本数和更多的数据训练模型，移除了NSP(Next Sentence Prediction)任务，同时采用动态编码，极大提升了BERT的模型性能。

但上述这些工作主要集中于面向英语和现代汉语的模型训练，阎覃(2020)开发的GuwenBERT模型首次将此项工作迁移到古汉语领域，GuwenBERT模型是基于殆知阁古文文献语料训练，其中包含15694本古文书籍，字符数1.7B。所有繁体字均经过简体转换处理，结合现代汉语RoBERTa权重和大量古文语料，将现代汉语的部分语言特征向古代汉语迁移以提升表现。

本次的任务是古汉语的嵌套命名实体识别，如果使用GuwenBERT模型就需要将古汉语的文献转换成现代汉语的文献，但繁简转换会出现很多问题，如一简对多繁、引发歧义等。王东波等(2021)以校验后的高质量《四库全书》全文语料作为训练集，基于BERT深度语言模型框架，构建了面向古文智能处理任务的SikuBERT和SikuRoBERTa预训练语言模型。第二年，Koichi Yasuoka等(2022)完成了RoBERTa-classical-chinese模型，并且扩展了繁体词表。SikuRoBERTa和RoBERTa-classical-chinese两个预训练语言模型的开发解决了繁简转换这一问题。因此，本研究主要侧重选用SikuRoBERTa和RoBERTa-classical-chinese两个模型进行训练。

4.2 GlobalPointer

在命名实体识别研究中，通过编码层对实体进行抽象语义表示，生成相应的标签序列，而解码层则用于预测实体的边界以及实体的类型，是整个实体识别模型的最后阶段。常用的标签解码器包括MLP+softmax层、条件随机场(CRF)、循环神经网络解码器、指针网络（Pointer Network）几种类型。

其中，MLP+softmax层、条件随机场(CRF)、循环神经网络解码器主要用于扁平命名实体识别，指针网络（Pointer Network）除了用于扁平命名实体识别外，也适用于嵌套命名实体识别。指针网络解码器采用循环解码的结果，它将序列标注问题转化为先分界再分类两个子任务，其中先分界指找出实体的开始位置和结束位置，然后对识别出的实体块进行类型识别，之后再继续找下一个块的结束位置，再将这个块进行分类，一直循环直到序列结束。但指针网络

(c) 2022 中国中文信息学会计算语言学专业委员会

的设计在做实体识别时，会出现训练和预测不一致的问题，即训练的时候开始位置和结束位置是孤立的，预测的时候开始位置和结束位置是有联系的，所以开始位置和结束位置都预测正确实体才算预测正确，这就导致了不一致。

针对上述不一致的问题，苏剑林(2022)利用全局归一化的思路来进行命名实体识别，设计了可以无差别地识别嵌套实体和非嵌套实体的GlobalPointer。它的主要思想体现在它将开始位置和结束位置视为一个整体去进行判别。具体的，设长度为n的序列输入t经过编码后得到向量序列$\{h_1, h_2, ..., h_n\}$,通过变换$q_{i,\alpha} = W_{q,\alpha}h_i + b_{q,\alpha}$和$k_{i,\alpha} = W_{k,\alpha}h_i + b_{k,\alpha}$，我们可以得到序列向量序列$\{q_{1,\alpha}, q_{2,\alpha}, ..., q_{n,\alpha}\}$和$\{k_{1,\alpha}, k_{2,\alpha}, ..., k_{n,\alpha}\}$，它们是识别第$\alpha$种类型实体所用的向量序列。此时可以定义

$$s_\alpha(i,j) = q_{i,\alpha}^T k_{j,\alpha} \tag{1}$$

作为从i到j的连续片段是一个类型为α的实体的打分。也就是说，用$q_{i,\alpha}$与$k_{j,\alpha}$的内积，作为片段$t_{|i:j|}$是类型为α的实体的打分。

接下来的关键是损失函数的设计。苏剑林(2022)提出一个用于多标签分类的损失函数，特别适合总类别数很大但目标类别数较小的多标签分类问题。该函数为

$$log(1 + \sum_{(i,j)\in P_\alpha} e^{-s_\alpha(i,j)}) + log(1 + \sum_{(i,j)\in Q_\alpha} e^{s_\alpha(i,j)}) \tag{2}$$

其中P_α是该样本的所有类型为α的实体的开始和结束位置集合，Q_α是该样本的所有非实体或者类型非α的实体的开始和结束位置集合，只需要考虑i⩽j的组合，即

$$\begin{aligned} \Omega &= \{(i,j) \mid 1 \le i \le j \le n\} \\ P_\alpha &= \{(i,j) \mid t_{|i:j|}\text{是类型为}\alpha\text{的实体}\} \\ Q_\alpha &= \Omega - P_\alpha \end{aligned} \tag{3}$$

5 实验

5.1 模型参数设置及模型构建

本次实验采用RoBERTa-classical-chinese、SikuRoBERTa、SikuBERT、RoBERTa-wwm-ext、BERT-wwm-ext和GuwenBERT六个预训练语言模型，并在训练过程中进行微调，获取字嵌入向量，并完成特征抽取，以捕获实体上下文的特征表示，然后输入GlobalPointer进行解码，获得预测标签序列。下图3展示了以RoBERTa-classical-chinese+GlobalPointer搭建的模型框架，其他模型与之类似。

(c) 2022 中国中文信息学会计算语言学专业委员会

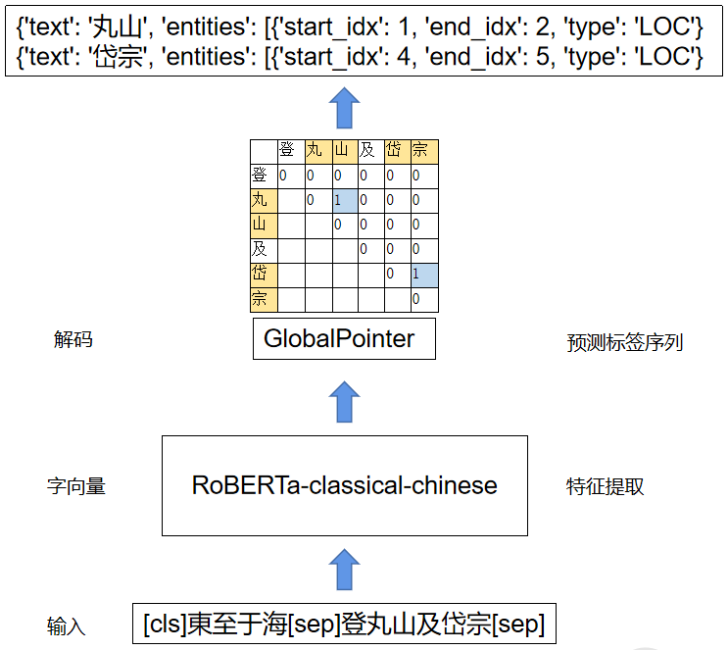


Figure 3: RoBERTa-classical-chinese+GlobalPointer模型的整体结构

同时，使用Adam优化器(Da, 2014)，所有模型的学习率（Learning Rate）均设置为$2e^{-5}$，输入序列最大长度（Maxlen）为128，每批训练大小（Batch Size）为16，迭代次数（Epochs）为50，训练时保存效果最优的模型。

5.2 测评指标及评价方法

为了评估以上六个预训练语言模型对古汉语实体识别的效果，本次实验采用精确率(precision)、召回率(recall)和F1值(F1 score)作为评估标准。

$$\text{精确率} = \frac{\text{正确预测为正的数量}}{\text{所有预测为正的数量}} \tag{4}$$

$$\text{召回率} = \frac{\text{正确预测为正的数量}}{\text{实际为正的数量}} \tag{5}$$

$$F1\text{值} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{6}$$

精确率越高，代表模型对负样本的区分程度越高；召回率越高，代表模型对正样本的识别程度越好；F1值越高，代表模型越稳定。

5.3 实验结果及分析

5.3.1 实验一：不同数据规模和比例下的实验结果

为了确定实验中训练集、验证集和测试集的占比，同时考察不同数据集规模对实验结果的影响，将训练集、验证集和测试集以25%的比列增量多轮试验。实验中，训练集、验证集和测试集之间的比例暂定为7:1.5:1.5。由于只是考察数据集规模对实验的影响，且之后实验的预训练模型主要以RoBERTa类型为主，所以在实验一中，主要选取了RoBERTa-classical-chinese+GlobalPointer模型。以下表2展示的是数据集规模对古汉语嵌套实体识别的实验结果:

(c) 2022 中国中文信息学会计算语言学专业委员会

语料比列	P	R	F1
25%	68.92%	67.65%	**68.05%**
50%	79.60%	63.98%	**70.42%**
75%	72.98%	73.29%	**72.61%**
100%	85.35%	84.44%	**84.71%**

Table 2: 不同数据集规模的实验结果

之后为了进一步确定和考察不同比例对于古汉语嵌套实体识别的影响，将全量数据按照8:1:1划分后，进行了补充实验，以下表3-5展示了不同比例划分下的实体数量分布情况和不同比例下的实验结果:

实体种类	训练集	验证集	测试集
人（PER）	3686	280	621
书（BOOK）	83	2	0
地点（LOC）	2424	214	450
官职（JOB）	1833	175	465
时间（TIME）	349	48	65
实体总数	8375	719	1601

Table 3: 8:1:1数据集实体数量分布

实体种类	训练集	验证集	测试集
人（PER）	3390	446	751
书（BOOK）	77	6	2
地点（LOC）	2208	363	517
官职（JOB）	1651	264	558
时间（TIME）	330	57	75
实体总数	7656	1136	1903

Table 4: 7:1.5:1.5数据集实体数量分布

模型	各数据集占比	P	R	F1
RoBERTa-classical-chinese	8:1:1	**85.35%**	**84.44%**	**84.71%**
RoBERTa-classical-chinese	7:1.5:1.5	82.28%	79.47%	80.42%

Table 5: 不同数据集比列的实验结果

根据表2得知，随着数据集规模的扩大，F1值也随之增长。25%到50%，50%到75%，F1值呈平缓稳定的增长，由75%到100%后，F1值显著增长，可见扩大数据集，模型的学习效果会更高，当数据集达到一个阈值，会迎来效果的质变。所以，在合理的范围内，应尽可能的扩充数据集。根据这一结论，训练集、验证集和测试集之间的比例也由7:1.5:1.5调整到了8:1:1，效果如表5所示，F1值有所提高。

5.3.2 实验二：不同标注标准下的实验结果

前文提到实体分类界限模糊的问题，基于此，本次实验设计了两个标注标准：一是根据字词本身含义进行标注。二是结合具体语境含义对字词进行标注。在全量数据下，使用RoBERTa-classical-chinese+GlobalPointer模型，按照7:1.5:1.5的比例对两个数据集进行了对比实验，根据最终的F1值来敲定最终标注的标准。

第二十一届中国计算语言学大会论文集，第406页-第416页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

标注标准	P	R	F1
标准一	**82.28%**	**79.47%**	**80.42%**
标准二	77.17%	78.17%	77.43%

Table 6: 不同标注标准的实验结果

根据表6得知，标准一在精确率、召回率和F1值上都要优于标准二。标准一的标注原则是根据字词本身含义进行标注，如“官职”在语境中代指“人”，依旧标注为“人（PER）”。通过本次对比实验也可以看出，就本研究使用的语料和条件设置下，对于字词的引申含义、具体语境含义等这些并非字词本身义项的情况出现时，依旧要按照字词原本含义进行标注，这一点是不同于某一字词的不同义项应用在不同场景这种情况的，所以并不能按照同一标准对待。

5.3.3 实验三：不同模型下的实验结果

在确定数据集标注采用的标准后，我们又增加了SikuRoBERTa、SikuBERT、RoBERTa-wwm-ext、BERT-wwm-ext、GuwenBERT5个预训练语言模型，与GlobalPointer搭配，展开不同模型维度的对比试验。其中，Siku系列、RoBERTa-classical-chinese和GuwenBERT都是面向古文开发的的预训练语言模型，剩余是面向现代汉语开发的预训练语言模型。具体实验结果如下表7：

模型	P	R	F1
RoBERTa-classical-chinese	**85.35%**	**84.44%**	**84.71%**
SikuRoBERTa	85.02%	83.73%	84.18%
SikuBERT	85.43%	82.12%	83.66%
RoBERTa-wwm-ext	81.38%	81.65%	81.29%
BERT-wwm-ext	79.66%	83.72%	81.13%
GuwenBERT	78.98%	77.20%	77.84%

Table 7: 模型结果

表7比较了RoBERTa-classical-chinese、Siku系列模型、RoBERTa-wwm-ext、BERT-wwm-ext和GuwenBERT六个模型在同一数据集下的效果。

从表7的结果来看，SikuRoBERTa较之SikuBERT，RoBERTa-wwm-ext较之BERT-wwm-ext，F1值分别提升0.52%，0.16%。说明RoBERTa预训练模型性能要优于BERT预训练模型，RoBERTa模型不仅继承了BERT模型的优势，而且用更大的单次训练样本数和更多的数据训练模型，移除了NSP任务，采用动态编码。这一系列的改进措施促使了RoBERTa的效果优于BERT。

RoBERTa-classical-chinese和Siku系列模型在精确率、召回率、F1值三个指标上都要高于RoBERTa-wwm-ext和BERT-wwm-ext两个模型。这是由于，RoBERTa-classical-chinese基于殆知阁古文文献语料训练的语言模型，Siku系列模型是基于校验后的高质量《四库全书》全文语料训练的，均是针对古汉语专门训练的预训练语言模型，而RoBERTa-wwm-ext和BERT-wwm-ext是面向现代汉语训练的预训练语言模型。所以RoBERTa-classical-chinese和Siku系列模型更加适合本次的古汉语实体识别任务。

观察对比发现，GuwenBERT预训练语言模型的F1值在本次实验较之其他模型低。虽然GuwenBERT预训练语言模型采用古汉语语料进行训练，但是所有语料均被转换为简体，繁体转换为简体后，会出现一简对多繁的问题，从而引发歧义，而且部分语料未经分句、没有标点符号，则进一步放大了这一问题，最终导致模型性能降低。

对于古汉语预训练语言模型的对比。平均F1值RoBERTa-classical-chinese和SikuRoBERTa相差无几，RoBERTa-classical-chinese高出SikuRoBERTa 0.53%。两类模型同属于RoBERTa，结构相似，出现这个差距主要在于RoBERTa-classical-chinese训练所采用的语料数量要大于Siku系列模型，但由于语料并非精加工，甚至存在未分句，未标点的语料，使得最终F1值只是略高于Siku系列模型。

(c) 2022 中国中文信息学会计算语言学专业委员会

6 总结与展望

本文通过使用六种预训练语言模型实现了古汉语嵌套命名实体识别的初探，检验了古汉语预训练模型RoBERTa-classical-chinese和SikuRoBERTa配合GlobalPointer模型在古汉语嵌套命名实体识别任务上的良好性能。同时探讨了古汉语嵌套实体数据集的实体分类原则、数据集的格式和确定数据集的标注标准。

由于原始语料本身的特点，使得本次实验所用的数据集存在规模较小、各实体类别样本分布不均匀的不足，对于模型的最终性能造成了一定的影响。同时，由于鲜有开源的古汉语预训练模型，与现代汉语模型对比效果有失公平，所以无法进行更多模型的对比。因此，下一步工作将围绕扩充不同类别古文语料，建立更大规模、分布更均匀的高质量标注数据集，同时尝试并优化算法，以期进一步提升古汉语嵌套识别模型的性能，为古汉语信息抽取和知识图谱的建立提供帮助。

参考文献

Kingma Da. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ralph Grishman and Beth Sundheim. 1995. Appendix c: Named entity task definition (v2. 1). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 317–332.

Zan Hongying, Li Wenxin, Zhang Kunli, Ye Yajuan, Chang Baobao, and Sui Zhifang. 2020. Building a pediatric medical corpus: Word segmentation and named entity annotation. In *Workshop on Chinese Lexical Semantics*, pages 652–664.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. Nne: A dataset for nested named entity recognition in english newswire. *arXiv preprint arXiv:1906.01359*.

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 the 19th Chinese national conference on computational linguistics*, pages 561–571.

Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.

刘忠宝，党建飞，张志剑. 2020. 《史记》历史事件自动抽取与事理图谱构建研究. 图书情报工作, 64(11):116.

安孝一. 2022. Transformers のbert は共通テスト『国』をり受け解析するをるか. 洋学へのコンピュタ利用第*33* 回研究セミナ, 33:3–34.

崔竞烽，郑德俊，王东波，李婷婷. 2020. 基于深度学习模型的菊花古典诗词命名实体识别. 情报理论与实践, 43(11):150.

曹艳，侯汉清. 2008. 古籍文本抽词研究. 图书情报工作, 52(01):132.

朱锁玲，包平. 2011. 方志类古籍地名识别及系统构建. 中国图书馆学报, 37(3):118–124.

李娜，包平. 2018. 面向数字人文的馆藏方志古籍地名自动识别模型构建. 图书馆, (5):67–73.

袁悦，王东波，黄水清 ，李斌. 2019. 不同词性标记集在典籍实体抽取上的差异性探究. 现代图书情报技术, 003(003):57–65.

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.

(c) 2022 中国中文信息学会计算语言学专业委员会

王东波，高瑞卿，沈思，李斌. 2018. 面向先秦典籍的历史事件基本实体构件自动识别研究. 国家图书馆学刊, 27(1):65–77.

皇甫晶，王凌云. 2013. 基于规则的纪传体古代汉语文献姓名识别. 图书情报工作, 57(03):120.

苏剑林. 2022. Efficient globalpointer：少点参数，多点效果, Jan. https://spaces.ac.cn/archives.

阎覃. 2020. Guwenbert: 古文预训练语言模型(古文bert). 2020-11-22]. https://github.com/Ethan-yt/guwenbert.

黄水清，王东波，何琳. 2015. 基于先秦语料库的古汉语地名自动识别模型构建研究. 图书情报工作, 59(12):135.

CCL 2022

CoreValue：面向价值观计算的中文核心价值-行为体系及知识库

刘鹏远　　张三乐　　于东　　薄琳
北京语言大学信息科学学院
国家语言资源监测与研究平面媒体中心
北京市海淀区学院路15号, 100083
liupengyuan@pku.edu.cn　sanle0409@163.com　yudong@blcu.edu.cn　bolin_blcu@163.com

摘要

由主体行为推断其价值观是人工智能理解并具有人类价值观的前提之一。在NLP相关领域，研究主要集中在对文本价值观或道德的是非判断上，鲜见由主体行为推断其价值观的工作，也缺乏相应的数据资源。该文首先构建了中文核心价值-行为体系。该体系以社会主义核心价值观为基础，分为两部分：1）类别体系。共包含8大类核心价值，进一步细分为19小类双方向价值并对应38类行为；2）要素体系。划分为核心与非核心要素共7种。随后，抽取语料中含有主体行为的文本句，依据该体系进行人工标注，构建了一个包含6994个行为句及其对应的细粒度价值与方向，34965个要素的细粒度中文价值-行为知识库。最后，该文提出了价值观类别判别、方向判别及联合判别任务并进行了实验。结果表明，基于预训练语言模型的方法在价值观方向判别上表现优异，在细粒度价值类别判别以及价值类别多标签判别上，有较大提升空间。

关键词： 价值观计算；人工智能伦理；价值-行为体系；价值-行为知识库

CoreValue: Chinese Core Value-Behavior Frame and Knowledge Base for Value Computing

Pengyuan Liu　　Sanle Zhang　　Dong Yu　　Lin bo
Beijing Language and Culture University
National language resources monitoring and research print media center
15 Xueyuan Road, Haidian District, Beijing, 100083
liupengyuan@pku.edu.cn　sanle0409@163.com　yudong@blcu.edu.cn　bolin_blcu@163.com

Abstract

It is one of the prerequisites for artificial intelligence to understand and possess human values to infer their values from their behavior. However, in NLP related fields, the current research mainly focuses on the judgment of the values or morality of the text, rarely inferring their values from the subject's behavior, and also lacks corresponding data resources. This paper first constructs the Chinese core value-behavior frame. It is based on China's socialist core values and is divided into two parts: 1) category system. There are 8 categories of core values, which are further subdivided into 19 categories of bi-directional values and corresponding to 38 types of behaviors; 2) Factor system. There are 7 types of factors. Then, text sentences containing subject behavior are extracted from the corpus and manually labeled according to the system. Then, a fine-grained Chinese value-behavior knowledge base containing 6994 behavior sentences and their corresponding fine-grained values and directions, and 34965

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第417页-第430页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

elements is constructed. Finally, this paper puts forward the tasks of value category classification, direction detection and joint discrimination. Experimental results show that the method based on the pretraining language model performs well in judging the direction of values, and has great room for improvement in fine-grained value category classification and multi-label value category classification.

Keywords: Values Computing , Artificial Intelligence Ethics , Value-Behavior Frame , Value-Behavior Knowledge Base

1 引言

人工智能正在对世界产生重大且深远的影响。一些基于人工智能的算法可代替人自动执行决策，或者说人授权这些算法可以自动执行决策，如自动驾驶、简历筛选(Dastin, 2018; Weed, 2021)、基于法律判决预测(Feng et al., 2022)进行自动判案、甚至是执行武器射击(Vynck, 2021)，而这些行为如果不加某种人类的伦理道德约束，可能会产生巨大风险。因此，人工智能治理(Munoz et al., 2016; Smuha, 2019; 中国国家新一代人工智能治理专业委员会, 2019)正日益受到重视，使人工智能或者机器具有人类伦理道德价值观意义与价值凸显。

近年来，机器伦理、机器道德领域的相关研究主要集中在：1）道德判断即判断某事件或行为是否道德(Prabhumoye et al., 2020; Zhou et al., 2021; Botzer et al., 2022; 彭诗雅 等, 2021)；2）基于社会规范的行为与后果推理(Forbes et al., 2020; Emelin et al., 2020; Lourie et al., 2020)；3）伦理道德的描述与建模(Prabhumoye et al., 2020; Schramowski et al., 2021, 2020)。此外，对社会偏见的检测、分类与消除(Sap et al., 2019; Blodgett et al., 2020)的相关研究也可纳入到机器伦理研究范畴。

在人工智能嵌入人类价值观方面的研究较少。价值观本身抽象、多样、难以具体描述，这一点可从价值观定义[0]上管窥一斑：价值观是一种外显的或内隐的，有关什么是“值得的”的看法，它是个人和群体的特征，影响着人们对行为方式、手段及目的的选择(Kluckhohn et al., 1948)。在科幻小说中，“机器人三定律”的情节说明简单的规则难以编码人类复杂的价值观(Asimov, 2004)。迄今为止，尚无方法对机器是否具有普遍的人类价值观进行测量(Müller, 2020)。虽然Hendrycks et al. (2020)尝试将人类共享的价值观与人工智能对齐，并对某行为做价值观是非判断，但该研究尚存在一些问题，即所基于的人工标注数据并没有预先确定标注者的价值观。这一问题与Talat et al. (2021)发现Jiang et al. (2021)研究中对道德进行判断所存在的问题类似。此外，虽然机器对各种具体行为的价值观或道德是非判断非常重要，但在做出是非判断之前，机器需要理解人类的具体行为是基于什么价值观做出的，这是机器真正理解进而具有人类价值观或者能够进行是非判断的前提。但目前鲜见将价值与行为统一起来建设的语义体系与数据资源，导致难以对机器是否理解、如何理解及如何判别主体行为的价值观进行深入研究。资源的匮乏已经成为制约该方向进一步发展的瓶颈与挑战。

本文首先对价值观体系进行分析梳理，在此基础上结合中国社会环境的特点与实际，确定以社会主义核心价值观作为中文价值-行为体系的基础。随后，从社会主义核心价值观中选取了八类核心价值，通过对大量新闻语料文本中含有主体[1]和具体行为实例的观察，发现价值-行为具有方向性，然后根据具体行为将价值进行进一步细分，将具体行为与价值及方向相对应，建立了中文核心价值-行为类别体系并进行了覆盖度验证；通过对具体行为的分析，得到价值-行为关联的要素，建立了价值-行为的要素体系，类别与要素体系共同构成了中文核心价值-行为体系。基于该体系，对价值-行为实例进行人工标注，每条实例均标注了细分价值与行为、价值-行为关联的所有要素，初步建立了一个中文核心价值-行为知识库。最后，还提出了价值观类别判别、方向判别以及联合判别三个任务并进行了初步实验与分析。本文贡献可总结如下：

1）建立了中文核心价值-行为体系。该体系分为两大部分：a）类别体系。包含8类核心价值，细分为19小类双向价值并对应38类行为；b）要素体系。为核心与非核心要素共7种；

2）依据上述体系进行人工标注，最终构建了一个包含6994个行为句及其对应的细粒度价值与方向，34965个要素的细粒度中文价值-行为知识库[2]并进行了初步分析，该知识库为进一步研

[0]研究者们对于“价值观是什么”一直众说纷纭，本文采用了大多数研究者的共识经典表达。
[1]本文研究中的主体，如无特别说明，均指个人或个体，这也是国内外价值观研究的重点。
[2]已开源在：https://gitee.com/NLUSOCO/CoreValue.git

(c) 2022 中国中文信息学会计算语言学专业委员会

究中文核心价值与行为之间的关系、将中文核心价值嵌入到人工智能等价值观计算等相关研究提供了数据基础；

3）为考察机器对文本中主体行为的价值类别与价值方向进行判别的能力，提出了价值观类别判别、方向判别及联合判别三个任务，并进行了基线模型实验和分析。

2 相关工作

2.1 资源

当前规模最大且最有影响的资源为SOCIAL CHEMISTRY 101(Forbes et al., 2020)，该资源面向社交与道德准则推理，包含十二个不同维度的人类关于社交、道德、预期文化压力以及责任承担等的判断，共包含450万个人工标注的类别标签以及上下文描述。基于这个资源，学者们又构建了一些相关资源如：Moral Stories(Emelin et al., 2020)，该资源共包含1.2万个短文本，主要目的是考察机器在社交情况下面向目标的道德推理与生成能力；ValueNet(Kim et al., 2022)，针对价值观约束下的对话进行研究，包含21374个文本场景的人类价值观。该数据集按Schwartz et al. (2012)的十个价值维度进行组织。

SCRUPLES(Lourie et al., 2020)也是一个针对道德判断的数据集。该数据集的标注对象是包括一个标题以及文本正文的描述真实生活的场景，对3.2万个场景共标注了62.5万个道德判断。在SOCIAL CHEMISTRY 101(Forbes et al., 2020)、Moral Stories(Emelin et al., 2020)、SCRUPLES(Lourie et al., 2020) 等的基础上，Jiang et al. (2021)构建了DELPHI数据集，共包含170万个在日常生活中的人类道德判断，其目标是使机器理解具有道德和社会准则，并具有相应判断的能力。CMOS(彭诗雅 等, 2021)，是中文道德句资源，共包含10万个人工判断的是否包含道德以及是否道德的文本句。

其他有影响的还有ETHICS(Hendrycks et al., 2020)和myPersonality(Kosinski et al., 2013)。ETHICS数据集通过人工撰写的方式共获得了超过13万个样例，每个样例由多个文本句构成。数据集包含公正、美德、义务论、利己主义、常识道德共五类。myPersonality是一个包含约15万Facebook用户状态与信息的数据集。从用户中提取了多个样本子集进行并完成了数十种问卷调查，这些问卷涉及人格评估、人口统计学信息和价值观。

2.2 体系

Perry (1926)最早将价值观分为六大类，即认知、道德、经济、政治、审美和宗教。Allport et al. (1960)将价值观也分为六大类：经济的、理论的、审美的、社会性的、政治的和宗教，该分类具有较大影响。Kluckhohn et al. (1948)从价值取向进行划分，总结为：人与自然的关系；理想人格类型；人与他人的关系的形态；时间评价和组织；人的本性。Rokeach (1973)的分类突破了上述类别的框架，认为价值观有终极性和工具性两个维度，并将价值观分为工具性价值观和终极性价值观两类。这样的分类将价值观更有层次和顺序的体现出来。

目前应用最为广泛的是Schwartz et al. (1987)的分类体系，他将价值观分为自我提高—自我超越，保守主义—开放性两个垂直维度，根据维度分为权利、成就、享乐、自主、刺激、博爱、慈善、顺从、保守和安全10类价值观。

国内的研究主要有：杨中芳 (2005)将文化价值体系划分为世界观、社会观和个人观三个层面，每类下再继续细分；黄希庭 等 (2005)将价值观分为人生、政治、道德、职业、人际关系、审美、婚恋、宗教、自我价值和幸福价值观；张进辅 (1998)认为价值观由价值目标、价值手段和价值评价维度组成，把价值观分为人生、政治、道德、职业、婚恋、消费、审美、人际、宗教、知识、教育价值观等。

综上，现有价值观体系主要是根据维度和类型进行分类，分类粒度较粗，对价值观类别细分类的研究较少。鲜见同时从价值观与主体行为两个角度出发建立两者统一的体系。此外，在价值观所体现的具体行为上，鲜见细粒度描述框架的研究。

3 中文核心价值-行为体系设计

人类有着复杂的思想，由于社会环境、所受教育和宗教信仰等的不同，人们的价值观也有所差异。不同文化和社会政治制度中可能存在不同的价值选择(Aizenberg et al., 2020)，比如我国孔子强调以仁、义、礼治天下，而西方的苏格拉底则注重幸福、正义与勇气。

(c) 2022 中国中文信息学会计算语言学专业委员会

党的十八大提出了社会主义核心价值观，即“三个倡导”，这全面概括了全党全社会的价值共识。其中国家层面的价值目标是：富强、民主、文明、和谐；社会层面的价值取向是：自由、平等、公正、法治；个人层面的价值准则：爱国、敬业、诚信、友善[3]。在价值观体系内涵层面，社会主义核心价值观不仅涵盖了广大人民群众的普遍愿望，更凸显了当今中国社会主流意识形态的核心价值理念，容纳了历史文化传统、鲜明时代精神和未来价值追求。社会主义核心价值观就是我国社会做出的价值选择。

3.1 数据来源

中文核心价值-行为体系的设计不但需对社会主义核心价值观内涵的正确理解，还需要找到对应核心价值观主体的具体行为并对其进行分析，因此需要在真实数据中进行考察。该数据需要包含人们日常生活中发生的各种事件。同时，本小节的目标是针对社会主义核心价值观进行价值-行为体系构建，因此希望数据中这些日常生活中发生的各类事件能与社会主义核心价值观有较强的相关性。本文选择的数据的来源为：

1）网络爬取的新闻语料。爬取的网站为中国文明网、青少年爱国主义网和搜狐新闻网；

2）中文道德句子库(彭诗雅 等, 2021)。以新闻文本及传记文本作为语料来源，约10万句。

以上数据源均为新闻语料，语体为书面语，语言使用客观、准确。从内容看，数据中包含大量的、多样的、真实的社会事件及生活事件，且所报道的事件多与当今社会主流或核心价值观相关。将上述语料进行分段、清洗和去重后，作为本小节设计中文核心价值-行为体系以及后续知识库构建的数据来源。

3.2 类别体系设计

3.2.1 核心价值选取

在价值观应用层面，社会主义核心价值观采取了一种开放性的表述方式，即对核心价值观进行直接列举，因此对每种具体相关的行为，其所对应的价值目标、取向、准则相对明晰，易于识别与标注。在十二个社会主义核心价值中，“富强、民主、和谐”体现在经济富强和政治民主、社会和谐，较为宏观，“自由”这一核心价值是指人们的意志自由、存在和发展的自由。在观察语料中我们发现：

1）体现“富强、民主、和谐”价值行为的实例通常以国家、集体为主体，而本文针对个体；

2）体现“自由”价值行为的实例过于宽泛，绝大多数行为都基于人们的“自由意志”。

最终，本文选取文明、公正、平等、法治、爱国、敬业、诚信、友善作为价值-行为类别体系的核心价值，共八种；同时，将从个人层面进行价值-行为类别体系及后续知识库的构建。

3.2.2 基于个人层面的社会主义核心价值内涵

8类核心价值的表述主要基于季明 (2013)对核心价值观的解读，本文根据个人层面行为所体现的价值观，略有调整：

文明。是个人素养、教养的重要体现，与社会个体在文化和道德品行上的素质紧密相关。

公正。就是有着不偏私、以公为首的思维方式，对待事物公平正直，没有偏私。

平等。指个人在社会关系、社会生活中处于同等的地位，具有相同的发展机会，享受着平等的权利和义务。

法治。知法、懂法、守法就是公民法治观念的体现。

爱国。是中华民族精神最稳定的文化基因，体现了人们对自己祖国最深厚的感情。是基于个人对自己祖国依赖关系的深厚情感，要求人们以振兴中华为己任，自觉促进民族团结、维护祖国统一、报效祖国。

敬业。对公民职业行为准则的价值评价，是一个人对自己职业的基本尊敬和负责的态度。对于每一个公民来说，敬业精神的内涵表现在三个方面：热爱自己的工作和所投身的事业；勤勉努力、付出劳动；克制自己恣意享乐、纵情狂欢的欲望。

诚信。诚实守信，包括诚和信两方面。“诚”的内容又包括两方面：一是为人真实，不有意歪曲客观事物的本来面貌，实事求是；二是信守承诺，指人说话要算数、讲信用，对自己的承诺负责，要言而有复，诺而有行。

友善。公民的核心价值规范之一，推动和谐社会的构建。友善指人与人之间和睦、友好、亲近,需要公民做到待人如己、宽厚、助人为乐，努力形成社会主义的新型人际关系。

[3]价值目标、价值取向、价值准则三者内容的具体体现均为价值，后续本文将用“价值”表述，不做深层次区分。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.2.3 价值方向

具体行为在其所体现的价值上具有“方向性”，如“扶老奶奶过马路”的行为，与其关联的价值为“友善”；而“对儿童的求助视而不见”的行为，与其关联的价值也应认为是“友善”，但两者方向相反，即：前者行为与“友善”价值相符，后者行为与“友善”价值相悖。此外，不能就此推断行为主体是在有/没有（或富有/缺乏、认可/不认可）“友善”这一价值时分别做出的行为，也不应推断行为主体是在认可“冷漠”或“恶”等“负”价值情况下做出的行为。实际上，“文明、公正、平等、法治、爱国、敬业、诚信、友善”均为人们或者说行为主体共同享有的价值观，这一点并不会因为行为主体做出某些与价值相悖的行为而改变或在行为主体观念中消失，也不能简单推断行为主体认可“负”价值。

行为主体在基于某价值作出某种行为时，采取何种价值方向是主体的一种选择性注意(Treisman, 1964)，也就是说，行为主体有意识或无意识的选择注意/不注意该价值。更进一步，即该价值在行为主体脑中得到激活/抑制。如果某种价值观被激活，则行为主体会做出符合某种价值观的行为，反之则反，即有两种价值方向：激活（↑）与抑制（↓）。如：“一位年过六旬的老教师依然奋战在教学一线，坚持手写教案”。结合价值观相应的内涵，我们可以推断出“一位年过六旬的老人”是在激活了敬业这一价值的情况下，做出了“坚持手写教案”这一行为。反之，如：“这位银行员工在上班时间玩斗地主”。则由于“这位银行员工” 抑制了敬业这一价值观，做出了“在上班时间玩斗地主”的行为。

此外，存在价值互相影响的情况，如：“为了践行承诺，他带头清理垃圾”。从行为“带头清理垃圾”来看，该行为与“文明”这一价值相关，但该行为是在行为主体为了“诚信”（“践行承诺”）这一价值而做出的。此类现象仍然可以用价值“激活”/“抑制”解释，即：行为主体激活了“诚信”这一价值，做出了“带头清理垃圾”的行为。对这种情况，由于行为主要是由“诚信”价值导致的，因此本文暂不做“文明”这一价值是否被激活/抑制的判断。

3.2.4 基于行为实例的价值-行为类别细分

在明确了8类价值观内涵的基础上，本文通过观察语料中包含主体及其行为的实例，尝试将其按照核心价值归类。在归类过程中发现，单独从每类价值观指导的具体行为来看，主体所持有的价值观仍有明显差异，且能够进一步细分为几个类别。如在“文明”价值观下，还可以进一步细分为：公共文明（爱护公物、环境等相关行为）、仪表文明（个人卫生、仪表等相关行为）、言语文明（礼貌用语等相关行为）和思想文明（宣传文明理念等相关行为）。

对核心价值进行更细粒度的划分有助于更深层次的理解核心价值的外延与内涵。同时，对价值-行为做更细粒度的刻画，能够对行为主体对应价值进行细化区分，有助于深刻理解和把握价值引导下的主体行为差异，并为分析价值-行为的关系提供了更精细的视角。在已有的核心价值观相关研究中，鲜见对核心价值进行进一步细分的系统性研究。于是本文尝试从语料中的具体行为入手，通过对语料库中主体的真实行为实例进行分析，将其与核心价值关联，进一步考察各个实例间价值的差别，细分价值-行为，并将相似价值-行为进行比较、整理归纳与合并，自底向上进行核心价值体系的构建。尽量避免自顶向下划分类别时类别划分粒度不易把控，特别是一些过细的类别难以在真实语料中找到对应具体实例的问题。归类与细分主要依据：1）行为所依据的具体价值内涵、外延与程度；2）行为主体自身的特点；3）行为所作用的对象。

具体而言，如“见义勇为”和“互相帮助”两类均是体现“友善”价值的行为，虽然“见义勇为”也是“互相帮助”的一种体现，但“见义勇为”包含了更多的价值，或者说比一般的“互相帮助”的行为“价值”更高，事件主体甚至有“舍己救人”的可能，且这类的行为一般发生在紧急、危险的时刻，因此本文将“见义勇为”这一行为类别单独分类。如“关心慰问”体现友善价值的行为，本文认为这一行为类别行为的价值主要体现在看望、慰问、关心，其作用对象的范围很广，与“孝顺长辈”等行为所体现的价值内涵不同。又如对“平等”中包含“人格平等”，本文将行为主体的性别、地域、种族特点独立出来，细分为“性别平等”、“地域平等”和“种族平等”这三类价值。最终，核心价值次分类共包含19小类，对应38类行为。文明价值包括思想文明、公共文明、言语文明和仪表文明；公正价值包括思想公正、机会公正；平等价值包括思想平等、人格平等；法治价值观包括知法懂法、守法用法；爱国价值包括思想爱国、以身作则；敬业价值包括热爱岗位、忠于职守；诚信价值包括传播诚信、诚实待人、信守诺言；友善价值包括乐于助人、宽厚待人。表 1是“文明”这一核心价值的类别体系，包含了价值-行为类别细分，以及激活/抑制该价值对应的行为实例。包含全部核心价值的完整类别体系详见附录A。

(c) 2022 中国中文信息学会计算语言学专业委员会

价值次类	行为类	价值激活行为示例	价值抑制行为示例
思想文明	宣传学习	宣传塑料危害	封建迷信
公共文明	爱护公物	保护古城墙	破坏共享单车
	爱护环境	清理垃圾	乱扔垃圾
	遵守秩序	按序排序	霸占座位
	积极参与	参加水资源保护活动	
言语文明	用语礼貌	礼貌询问	骂脏话
仪表文明	个人卫生	饭前洗手	饭桌抠脚
	穿着服饰	着装合适	坦胸露乳

Table 1: “文明”价值类别体系

3.2.5 类别体系覆盖度验证

为验证类别体系对现实行为的覆盖程度，本文在3.1小节中的第一类数据来源即网络爬取的新闻语料中，随机抽取并人工筛选得到1000条行为主体为个人且其行为是基于8类核心价值观的句子，然后由标注员（两名语言学与应用语言学的硕士生）将其归类到类别体系，无法归类的单独列出。最终结果由另一名标注员进行统计。统计结果表明，本文的类别体系可覆盖96%新闻事件主体的行为，未覆盖的行为，表现为低频长尾分布。未来将考虑增设“其他”类别，对体系进行进一步完善。较高的覆盖率表明本文分类体系能够较好的覆盖真实的新闻语料，能够对当下中国社会环境的主流价值观和相应的行为进行较为全面的归类。

3.3 要素体系设计

主体的具体行为与信息抽取中的事件类似，都表示动作的发生或状态的变化，需要进行分解与表示才能准确刻画。ACE2005(Walker et al.)中事件的表示为：事件触发词、事件类型、事件论元和其他论元角色。如对例句：“英美轰炸伊拉克”，其事件类型为：攻击；“轰炸”是“攻击”事件的触发词；“英美”与“伊拉克”均为论元角色。但对于价值-行为，由于日常生活中的行为及影响因素众多，基于新闻事件的表示并不完全适用。如对例句：“陈某在正在行驶中的公交车上强行踩下刹车...”，按照事件表示分析，可认为事件类型是“法制”，“踩下”或“踩下刹车”是触发“法制”事件的触发词，但这些触发词所能触发的事件类型非常广泛，如果没有例句提示，很难将“踩下”或“踩下刹车”与“法治”这一价值观相联系。类似的，日常生活中的各类常见行为如：“走，拿，拉，上，喜欢，奔跑”，单从这些特定动词来看，基本都无法与特定价值观相联系。此外，行为的表示需要比事件表示更为具体，很多时候需要事件表示以外的元素才能与特定价值观相详细，如上例中的行为：“强行踩下刹车”，在识别这个具体行为依据的价值观时，“强行”这个修饰语起到很大作用；此外，行为发生的地点“正在行驶中的公交车上”，也是帮助价值观识别与判断的重要因素，修饰语“正在行驶中的”同样不能忽视或省略。

考察行为句中各个因素与价值观的关联，目的是帮助推断与考察价值观与行为间的关联。本文参考ACE2005对事件的表示方法，没有设定价值观触发词，而是直接将行为分解为7个要素，如表 2所示。其中，主体与行为是组成一个具体行为必不可少的两个要素，前提（动机、目的、意图、条件）是主体价值选择的基础，很大程度上决定主体的行为。因此，本文将主体、行为、前提这三个要素设置为**核心要素**，其余四个要素虽然对行为主体的价值推断有所影响，但相对次要，设为**非核心要素**。

4 中文核心价值-行为知识库构建

4.1 价值-行为句筛选

价值-行为句根据以下三个条件进行筛选：1）需要同时包含主体及其主观行为，但不包括描述感觉、感受或情感的行为。去除如“立即将执法车停在左侧道路”—缺少主体、“某人...感觉很冷”及“某人...初尝到骗保的甜头”—是对感觉感受的描述；2）句中主体严格限定为“某一个人”。需要去除如“他们开始在超市疯狂购物...”—行为主体为两或多人。3）句中主体的行为所基于的价值观需包含在八种核心价值观中，且价值观可明确推断。去除如“某人手持警务通、通过系统查询获悉了女孩身份信息”—难以推断出主体行为是在何种价值观下做出的、“为了省

(c) 2022 中国中文信息学会计算语言学专业委员会

要素	定义	文本实例	要素实例
主体	行为的实施者	...男子将车停在...	男子
行为	主体主观做出的行为	...公交车上强行踩下刹车...	强行踩下刹车
前提	行为主体的动机、目的、意图、条件	...为了践行承诺，他...	践行承诺
对象	行为所针对或作用的客体	...哄骗李某到其家中...	李某
工具	行为主体所用的工具、方式、手段	...用铁锹殴打他...	铁锹
时间	行为发生的时间	...在半夜大声唱歌...	半夜
地点	行为发生的地点	...在马路中央刷抖音...	马路中央

Table 2: 要素定义与示例

	文明	公正	平等	法治	爱国	敬业	诚信	友善	合计
激活（↑）	376	213	167	88	500	1776	425	3006	6551
抑制（↓）	487	167	331	1141	40	447	410	1384	4407
总	863	380	498	1229	540	2223	835	4390	10958

Table 3: 价值-行为粗粒度类别标注统计结果

钱，某人偷偷地在垃圾桶里捡同学们用剩的铅笔..."—虽然可以推断出其价值观是"节约"，但这一价值观并不在八种核心价值观之中。

按上述条件，从本文3.1小节的数据来源中随机抽取文本句13000句左右，经筛选并得到符合要求的句子共7130句作为下一步的标注对象。

4.2 标注原则

价值-行为知识库标注的四项基本原则为：1）要素需要保持完整，包含关联的修饰语；2）对价值观进行推断的主要依据是某个行为直接关联的要素（含行为）而不是根据整句，以避免多个行为的影响；3）为避免仅根据主体行为直接做出价值观判断，必须考察所有要素，并标记除主体/行为之外，哪个（些）要素可推断对主体价值观；4）若存在多个主体行为，则仅标注出体现八类核心价值观的行为，忽略其他行为。

4.3 标注过程与质量控制

为了确保标注质量，标注分为四个阶段：培训、试标注、二次培训以及正式标注。在培训阶段中，标注人员需要熟悉理论背景、标注规则和标注流程。在试标注阶段，标注人员对200条语料进行试标注，试标注正确率低于80%的标注人员将被劝退。试标注阶段后，对试标注中出现的问题对标注员进行进行二次培训。在正式标注阶段，标注工作分批次进行，每批次中的每条句子均由两位标注员进行标注。每批次标注语料由第三名标注员对标注不一致的数据进行详细核查。若标注正确率低于85%语料则需要全部核查，标注正确率低于80%的语料则需要重新标注。部分标注不一致的句子将交由第三名标注员进行讨论确定，有争议的句子将会被删除。由于需要标注要素以及粗细粒度类别，故将整个标注分为两个任务：1）价值-行为粗粒度类别及要素标注；2）价值-行为细粒度类别标注。两个任务的标注流程与质量控制方法类似。

标注质量。第一个任务的总体标注一致率为85.52%，第二个任务中次分类标注一致率为90.8%,行为归类标注一致率为89%。此外，对所有标注不一致的句子，均进行讨论后重新进行了标注。

4.4 构建结果

最终有效标注结果共6994句，含34965个要素。其中1128句与两或多个价值相关联。表 3与表 4是价值-行为粗粒度类别标注与要素标注统计结果，价值-行为细粒度统计结果见附录A。

4.5 初步分析

基于本文所建立的价值-行为知识库，可以从价值-行为粗细粒度分布、价值-要素、行为-要素等多个角度对相关规律进行分析与研究。限于篇幅，本文仅以"文明"为例列出其细粒度价值

(c) 2022 中国中文信息学会计算语言学专业委员会

	主体	行为	对象	原因	工具	地点	时间
激活（↑）	6418	6418	4285	1095	340	1180	2754
抑制（↓）	3592	3592	2505	434	625	740	987
总	10010	10010	6790	1529	965	1920	3741

Table 4: 价值-行为要素标注统计结果

分布与行为比例，并以“文明”与“公正”为例进行最重要的核心要素-行为进行初步考察。分布见图1-2，词云见图3，其中↓与↑分别指抑制与激活。

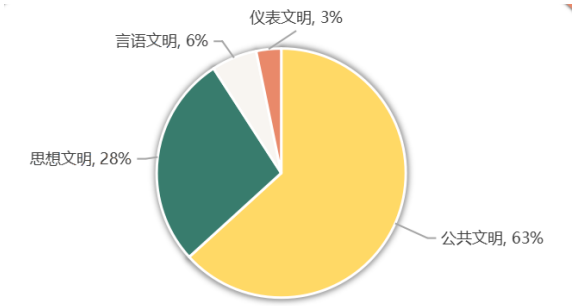


Figure 1: 次类分布（文明-行为类）

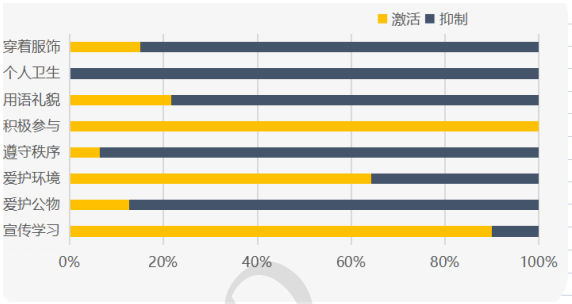


Figure 2: 激活抑制占比（文明-行为类）

从价值-行为次分类分布来看，占比从高到低依次为：公共文明>思想文明>言语文明>仪表文明。体现出人们更关注个体在社会中的文明表现，更关注宣传学习文明的行为。在文明价值激活与抑制方面，除“积极参与”、“爱护环境”与“宣传学习”外，其余类别均是价值抑制多于激活。侧面反映当下社会倾向在前三类文明激活的行为中塑造文明价值观，以“穿着服饰”、“个人卫生”、“用语礼貌”等文明抑制的行为作为文明价值的负向建构。

在要素方面，抑制与激活时两类价值的词汇各具特色。当主体的文明价值抑制时，主要会做出：1）不遵守规则的行为，如“逆行、闯红灯”等；2）个人素养低下的行为如“扔、破坏”等。反之，主体经常会成为文明知识/思想的传播者或学习者。当主体的公正价值抑制时，虽然会有“帮助、接受”等正面词，但是实际此时的主体是置其他人利益于不顾，在办事时多“帮助”自己的亲戚朋友，枉顾公正原则。同样，我们从“拒绝、坚持”这样的词也可得知主体在公正价值激活时会拒绝不公正现象、坚持原则、秉公办事。

Figure 3: 核心要素—行为的词云（文明↓、文明↑、公正↓、公正↑）

5 价值观计算任务

对给定主体行为的文本句，要使机器真正理解人类价值观，需要机器具有判别主体行为所基于价值的能力、判别主体做出行为时所处价值方向（激活/抑制）的能力。为此本文设计了三个任务：1）价值类别判别；2）价值方向判别；3）价值类别方向联合判别。

5.1 任务定义

价值类别判别。多分类任务。输入是一个单价值行为句[4] $X = x_1, x_2...x_i...x_m$，机器需要判别并输出对应的价值观标签 $Y \in \{y_1, y_2...y_i...y_n\}$。其中 x_i 为字或词，m为文本长度，y_i 为价值类别，n为价值类别总数。

[4]仅包含一个主体行为且该行为仅体现某一类价值观的文本句子。

(c) 2022 中国中文信息学会计算语言学专业委员会

	类别判别		方向判别		联合判别	
Baseline	ACC	F_1	ACC	F_1	ACC	F_1
Random	11.7	9.7	52.5	59.5	5.5	4.4
Majority	43.2	7.5	67.4	40.3	33.3	3.1
BERT	89.6	**85.7**	97.0	96.6	85.5	69.6
RoBERTa	89.2	85.3	98.2	**98.0**	88.0	**71.0**

Table 5: 价值类别判别、价值方向判别及价值类别方向联合判别的实验结果（%）

价值方向判别。二分类任务。输入是一个单价值行为句$X = x_1, x_2...x_i...x_m$以及对应的价值$C \in \{c_1, c_2...c_n\}$，机器需要判别并输出对应的方向标签$Y \in \{y_1, y_2\}$。其中x_i为字或词，m为文本长度，c_i为价值类别，n为价值类别总数，y_i为方向：激活（↑）/抑制（↓）。

价值类别方向联合判别。多分类任务。输入是一个单价值行为句$X = x_1, x_2...x_i...x_m$，机器需要同时判别对应的价值与方向并输出价值方向联合标签：$Y \in \{y_1, y_2...y_i...y_n\}$。其中$x_i$为字或词，m为文本长度，$y_i$为价值方向联合标签如：文明激活（↑）、文明抑制（↓）等。

5.2 实验

5.2.1 实验设置

实验数据。通过保留单一行为要素的方式，形成10010条仅包含一个行为的句子，其中有948条与两个价值相关，其余句子为单价值行为句共9062条，按照约8:1:1的比例划分为训练集（7250句）、验证集（906句）和测试集（906句）。

基线模型。1）Random。即随机选取类别标签作为分类结果；2）Majority。即选取频次最高的类别标签作为分类结果。鉴于目前在文本分类任务上基于预训练微调的方法性能较好，因此本文选取了两个基于预训练语言模型微调的方法：3）BERT(Devlin, 2018)、4）RoBERTa(Liu et al., 2019)。

参数设置。BERT为bert-base-chinese[5]默认设置，RoBERTa为chinese-roberta-wwm-ext[6]默认设置。均采取max-len为256，batch-size为64，learning-rate为1e-05。

评价指标。分类问题中常用的评价指标为准确率（（ACC, Accuracy）、精确率（P），召回率（R）与F值（$F-Score$）。本文使用准确率（ACC）与能综合反映分类器性能的宏平均F值（F_1），宏平均F_1值可视为多个二分类F-Score的算数平均值。

5.2.2 实验结果

三个任务的实验结见表5。对价值方向判别任务，BERT与RoBERTa的性能表现均十分优异，F_1值均超过了96%，这说明价值激活和抑制时所做出的的行为在语义上区分明显，预训练语言模型能够根据行为以及给定的价值做出正确的价值方向判断。对价值类别判别任务，BERT与RoBERTa也取得了较好的结果，F_1值均超过了85%，这说明各个价值的行为之间具有较好的语义区分度。表现最差的是价值类别与方向联合判别任务，F_1值在70%左右，RoBERTa模型的性能比BERT略高。联合判别任务需要同时判别价值类别与价值方向，因此这是一个16分类任务，难度低于分别为8分类与2分类的价值判别与价值方向任务，而且类别判别与方向判别可能会互相影响，故模型性能较差。

图4是RoBERTa模型在价值类别方向联合判别任务中各个价值方向上的性能比较柱状图，目的是考察模型对各个价值及不同方向判别的性能。模型对价值激活或抑制的判别上，不同价值方向的性能并不相同，且没有明显规律。值得注意的是，虽然“爱国”的样本句子数较少，约仅占总样本句子的6%，但其类别与方向的性能均为最高。经考察语料发现，“爱国”这一价值关联的行为，语义较为单调，“思想爱国”、在爱国的事情上“以身作则”这两类行为与其他价值关联的行为区分度较高。

5.2.3 讨论

以上三个任务，仅考虑了单价值行为句，而真实语料中确实存在一个行为句中与多（两）

[5]https://github.com/google-research/bert
[6]https://github.com/ymcui/Chinese-BERT-wwm

(c) 2022 中国中文信息学会计算语言学专业委员会

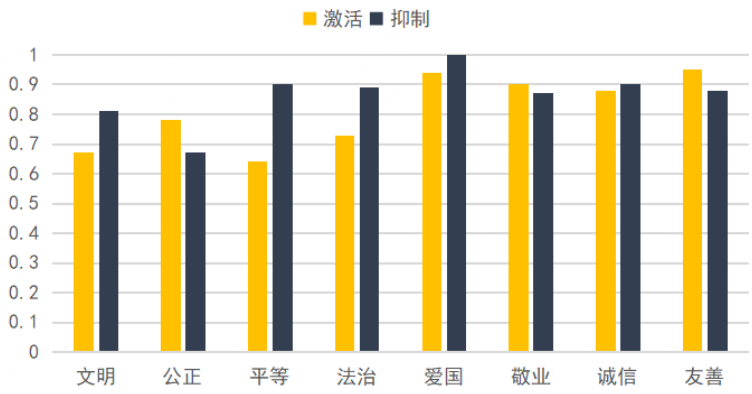


Figure 4: 基于RoBERTa的价值类别方向联合判别个价值方向别性能比较（F_1）

	多价值判别-所有			多价值判别-多标签			细粒度判别	
baseline	P	R	F_1	P	R	F_1	ACC	F_1
BERT	89.6	75.2	81.2	28.2	29.4	28.7	82.8	**54.2**
RoBERTa	89.7	78.6	**83.6**	42.8	32.3	**33.9**	83.5	53.8

Table 6: 多价值判别与细粒度判别的实验结果（%）。其中**多价值判别-所有**为对该任务所有样本进行性能评价的结果，**多价值判别-多标签**为仅对多标签样本进行性能评价的结果。

个价值观相关联的情况。此外，价值分类判别任务仅进行了8分类的粗粒度价值判别，而实际上本文所构建的知识库，包含更细粒度的价值共19类。考虑以上两点，本小节做了两个补充实验：1）**多价值判别**：含有多个行为多个价值句子的价值分类判别实验，一个句子可以有两或多个价值标签，因此是一个多标签分类任务；2）**细粒度判别**：针对细粒度的单价值行为句，进行19分类的价值分类判别。

多价值判别实验数据：将知识库中的全部文本句共6994条，按照约8:1:1的比例划分为训练集（5595句）、验证集（699句）和测试集（700句）。**细粒度判别实验数据**：知识库中的全部单价值细粒度行为句共9025条[7]，按照约8:1:1的比例划分为训练集（7220句）、验证集（903句）和测试集（902句）。**基线模型**、**参数设置**同前。**评价指标**：细粒度判别任务，同前；对多价值判别任务，采用了多标签分类常用评价指标宏平均P、R、与F_1。

实验结果见表6。对多价值判别任务，BERT与RoBERTa的表现良好，后者略高于前者，但由于其中84%都是单价值标签的样本，因此基于所有样本得到的性能难以说明模型对多标签样本的价值判别能力。表6中的"多价值判别-多标签"是仅针对多标签样本进行单独统计得到的结果。多标签样本数量占总样本的16%左右，均为每个样本2个价值标签。由于样本较少，且一般多标签分类相比单标签分类难度更高，BERT与RoBERTa的表现均不佳，F_1值不足34%。表6最右侧两列是细粒度判别任务的结果，可知BERT与RoBERTa的性能基本类似，且均较差，F_1值均在55%以下，原因可能是：1）由于是19分类，每个细粒度中的样例较少且不平衡；2）细粒度之间行为的语义较难区分。

6 结语

本文基于核心价值观建立了首个面向价值观计算的中文核心价值-行为体系及相应的知识库，该知识库可支持中文价值观计算与分析。基于该知识库，本文提出了3个价值观计算任务并进行了实验，实验结果表明基于预训练语言模型的方法在价值观方向判别上表现优异，在细粒度价值类别判别以及价值类别多标签判别上，有较大提升空间。本文工作尚存在一些局限如：整体规模较少尤其是细粒度价值-行为、语料类别分布不太均衡，个别类别中样本较少等。未来将：1）扩大知识库规模；2）针对个别类别的特点寻找适合的新闻语料来源；3）在扩大知识库规模的过程中对细粒度类别进行适当合并增减等调整，以增加覆盖度，在保证细粒度类别间区分度的同时，保证每一个细粒度类别能够有足够多的样本；4）将研究进一步拓展到语用层面。

[7]存在一些在粗粒度下为单价值行为句但是在细粒度下并非单价值行为句的情况，排除此类句子。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

中国国家新一代人工智能治理专业委员会, 2019. 新一代人工智能治理原则——发展负责任的人工智能[Z].

季明, 2013. 核心价值观概论[M]. 北京: 人民日报出版社.

张进辅, 1998. 我国大学生人生价值观特点的调查研究[J]. 心理发展与教育(2): 26-30.

彭诗雅, 刘畅, 邓雅月, 等, 2021. 字里行间的道德: 中文文本道德句识别研究[C]//第20届中国计算语言学大会. 537-548.

杨中芳, 2005. 中国人真是"集体主义"的吗?——试论文化、价值与个体的关系[J]. 中国社会心理学评论, 1(1): 55-93.

黄希庭, 郑涌, 2005. 当代中国青年价值观研究[M]. 人民教育出版社.

Aizenberg E, Van Den Hoven J, 2020. Designing for human rights in ai[J]. Big Data & Society, 7(2): 2053951720949566.

Allport G W, Vernon P E, Lindzey G, 1960. Study of values.[M]. Houghton Mifflin.

Asimov I, 2004. I, robot: volume 1[M]. Spectra.

Blodgett S L, Barocas S, Daumé III H, et al., 2020. Language (technology) is power: A critical survey of" bias" in nlp[A].

Botzer N, Gu S, Weninger T, 2022. Analysis of moral judgment on reddit[J]. IEEE Transactions on Computational Social Systems.

Dastin J, 2018. Amazon scraps secret ai recruiting tool that showed bias against women[M]// Ethics of Data and Analytics. Auerbach Publications: 296-299.

Devlin C M W L K T K, 2018. Bert: Pre-training of deep bidirectional transformers for language understanding[A].

Emelin D, Bras R L, Hwang J D, et al., 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences[A].

Feng Y, Li C, Ng V, 2022. Legal judgment prediction via event extraction with constraints [C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics: 648-664. https://aclanthology.org/2022.acl-long.48.

Forbes M, Hwang J D, Shwartz V, et al., 2020. Social chemistry 101: Learning to reason about social and moral norms[A].

Hendrycks D, Burns C, Basart S, et al., 2020. Aligning ai with shared human values[A].

Jiang L, Hwang J D, Bhagavatula C, et al., 2021. Delphi: Towards machine ethics and norms [A].

Kim H, Yu Y, Jiang L, et al., 2022. Prosocialdialog: A prosocial backbone for conversational agents[A].

Kluckhohn C E, Murray H A, 1948. Personality in nature, society, and culture.[M]. Knopf.

Kosinski M, Stillwell D, Graepel T, 2013. Private traits and attributes are predictable from digital records of human behavior[J]. Proceedings of the national academy of sciences, 110 (15): 5802-5805.

Liu Y, Ott M, Goyal N, et al., 2019. Roberta: A robustly optimized bert pretraining approach [Z].

Lourie N, Le Bras R, Choi Y, 2020. Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes[A].

Müller V C, 2020. Ethics of artificial intelligence and robotics[Z].

(c) 2022 中国中文信息学会计算语言学专业委员会

Munoz, Executive Office of the President C, Director D P C, et al., 2016. Big data: A report on algorithmic systems, opportunity, and civil rights[M]. Executive Office of the President.

Perry R B, 1926. General theory of value: Its meaning and basic principles construed in terms of interest[M]. [etc] Longmans, Green.

Prabhumoye S, Boldt B, Salakhutdinov R, et al., 2020. Case study: Deontological ethics in nlp [A].

Rokeach M, 1973. The nature of human values.[M]. Free press.

Sap M, Gabriel S, Qin L, et al., 2019. Social bias frames: Reasoning about social and power implications of language[A].

Schramowski P, Turan C, Jentzsch S, et al., 2020. The moral choice machine[J]. Frontiers in artificial intelligence, 3: 36.

Schramowski P, Turan C, Andersen N, et al., 2021. Language models have a moral dimension [A].

Schwartz S H, Bilsky W, 1987. Toward a universal psychological structure of human values.[J]. Journal of personality and social psychology, 53(3): 550.

Schwartz S H, Cieciuch J, Vecchione M, et al., 2012. Refining the theory of basic individual values.[J]. Journal of personality and social psychology, 103(4): 663.

Smuha N, 2019. Ethics guidelines for trustworthy ai[C]//AI & Ethics, Date: 2019/05/28-2019/05/28, Location: Brussels (Digityser), Belgium.

Talat Z, Blix H, Valvoda J, et al., 2021. A word on machine ethics: A response to jiang et al.(2021)[A].

Treisman A M, 1964. Selective attention in man[J]. British medical bulletin, 20(1): 12-16.

Vynck G D, 2021. The us says humans will always be in control of ai weapons. but the age of autonomous war is already here[J]. The Washington Post.

Walker C, Strassel S, Medero J, et al. Ace 2005 multilingual training corpus ldc2006t06, 2006 [J]. URL https://catalog. ldc. upenn. edu/LDC2006T06.

Weed J, 2021. Résumé-writing tips to help you get past the ai gatekeepers[J]. New York Times.

Zhou K, Smith A, Lee L, 2021. Assessing cognitive linguistic influences in the assignment of blame[C]//Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media. 61-69.

A 附录

CCL 2022

价值观类别	次分类	行为类	价值观激活时关联的行为	价值观抑制时关联的行为
文明	思想文明	宣传学习(知识/精神)	宣传塑料危害	封建迷信
	公共文明	爱护公物	保护古城墙	破坏共享单车
		爱护环境	清理垃圾	乱扔垃圾
		遵守秩序	按序排队	霸占座位
		积极参与(社会治理活动)	参加水资源保护活动	
	言语文明	用语礼貌	礼貌询问	骂脏话
	仪表文明	个人卫生	饭前洗手	饭桌抠脚
		穿着服饰	着装合适	坦胸露乳
公正	思想公正	宣传公正言论	打抱不平	发表歧视言论
	机会公正	办事公正	公正审判	找关系
		廉洁奉公	不占公家便宜	以权谋私
平等	思想平等	宣传平等言论	发表反歧视文章	传播歧视思想
	人格平等	性别平等	男女平等	重男轻女
		种族平等	尊重种族	歧视黑人
		地域平等	各省平等	地域黑
		其他（人人平等）	平等对话	富人高贵
法治	知法懂法	宣传/学习 法律条款/内容	阅读普法书籍	意识不到犯法
	守法用法	遵守法律	主动报警	违法犯罪
		配合民警执行公务	主动配合调查	拒不配合调查
爱国	思想爱国	宣传/学习 爱国言论/知识	参观革命史馆	散布分裂言论
		心系祖国	关注国家大事	崇洋媚外
	以身作则	维护祖国统一	捍卫领土	支持台独
		投入祖国建设	卫星研制	
敬业	热爱岗位	热爱本职工作	热爱教书	消极工作
	忠于职守	做好本职工作	付出辛勤劳动	偷懒耍滑
		克制欲望	淡泊名利	贪污受贿
诚信	传播诚信	宣传诚信理念	传播诚信文化	
	诚实待人	真实诚恳	直言真相	虚构事实
		拾金不昧	物归原主	隐瞒私吞
	信守诺言	兑现承诺	说话算话	欠债不还
友善	乐于助人	见义勇为	跳水救人	漠视求助
		捐款捐物	捐献衣物	
		互相帮助 （朋友、陌生人）	帮助邻居	
	宽厚待人	以和为贵	劝架	故意伤害
		关心慰问	看望战友	漠不关心
		孝顺长辈	赡养父母	虐待父母
		爱护幼小	收养孤儿	虐待儿童
		爱护动物	救助流浪狗	毒害流浪狗

Table 1: 价值-行为类别体系

第二十一届中国计算语言学大会论文集，第417页-第430页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

次分类	行为类	数量
思想文明	宣传学习（知识/精神）	239
公共文明	爱护公物	71
	爱护环境	112
	遵守秩序	316
	积极参与（社会治理活动）	46
言语文明	用语礼貌	51
仪表文明	个人卫生	8
	穿着服饰	20
思想公正	宣传公正言论	16
机会公正	办事公正	327
	廉洁奉公	37
思想平等	宣传平等言论	32
人格平等	性别平等	141
	种族平等	72
	地域平等	45
	其他（人人平等）	224
知法懂法	宣传/学习法律条款/内容	23
守法用法	遵守法律	1082
	配合民警执行公务	127
思想爱国	宣传/学习爱国言论/知识	126
	心系祖国	136
以身作则	维护祖国统一	197
	投入祖国建设	81
热爱岗位	热爱本职工作	52
忠于职守	做好本职工作	2029
	克制欲望	147
诚实待人	真实诚恳	496
	拾金不昧	116
信守诺言	兑现承诺	228
传播诚信	宣传诚信	6
乐于助人	见义勇为	522
	捐款捐物	502
	互相帮助（朋友、陌生人）	1256
宽厚待人	以和为贵	1395
	关心慰问	314
	孝顺长辈	194
	爱护幼小	204
	爱护动物	4

Table 2: 价值-行为句子标注结果

CCL 2022

第二十一届中国计算语言学大会论文集，第417页-第430页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

基于《同义词词林》的中文语体分类资源构建

黄国敬[1]，周立炜[1]，饶高琦[1,*]，臧娇娇[2]

1. 北京语言大学汉语国际教育研究院，北京,100083

2. 腾讯科技有限公司，北京，100080

ellenh1001@163.com, liweiyeahmail@163.com, raogaoqi@blcu.edu.cn, jojozang@tencent.com

摘要

语体词是指在某一语体中专用的词语，是语体的语言要素和形式标记。而语体词的资源可以服务于与现实场景息息相关的NLP应用，但目前此类资源较为稀缺。对此，本文基于《大词林》，完成了“语体词标注”“语体（词）链条标注”和“平行构式标注”三个任务，建立了以语体词为基础的语体分类资源。本资源包含55,710条词语、5,017个语体链条和433组平行构式。基于此，本文分析了中文语体词的分布概况、形态差异以及词义词性的分布情况。

关键词： 语体词；语体分类资源；同义词

Construction of Chinese register classification resources based on “Tongyici Cilin”

Huang Guojing[1], Zhou Liwei[1], Rao Gaoqi[1,*], Zang Jiaojiao[2]

1. Beijing Language and Culture University,Research Institute of International Chinese Language Education, Beijing, 100083

2. Platform & Content Group, Tencent Technology Co., Ltd

ellenh1001@163.com, liweiyeahmail@163.com, raogaoqi@blcu.edu.cn, jojozang@tencent.com

Abstract

The register (“register”is a term tentatively used here for the Chinese term yuti(语体)) words refer to words that are used exclusively in a certain register, and are the language elements and formal marks of the register. The resources of register words can serve NLP applications closely related to real-life scenarios, but such resources are relatively scarce at present. In this regard, based on the “DaCiLin”, this paper has completed three tasks of “register words tagging”, “register (words) chain tagging”and “parallel construction tagging”, and established a register categorized resources which based on register words. This resource contains 55,710 words, 5,017 register chains and 433 sets of parallel constructions. Based on this, this paper analyzes the distribution of Chinese register words, morphological differences, and the distribution of semantic parts of speech.

Keywords: register words , register categorized resources , synonym

*通讯作者corresponding author

本文系教育部人文社科基金“清末以来汉语报刊词汇使用计量研究”（20YJC740050）阶段性成果

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

语言为了完成不同的功能、适应不同的语域而形成的相对恒定的表达模式，被称为语体。而不同的词语在不同的场合中使用，有不同的功能，表现出不同的语体色彩，形成了语体词这一概念。袁晖(2004)等学者指明，语体词是语体的语言要素和形式标记，是认识和研究语体的锁钥之一。但当前学界语体词的研究并不丰富，多是关注语体词本身的特征；且语言实际使用场景中，选取哪个词语并没有相关的参考，则是依靠使用者本人的语言习惯。所有的NLP应用都需要在现实语言中进行交互，因而其产出和输入都从属于某种语体。通过对语体词的研究可以较好地提高自动校对、语言润色等自然语言生成任务的用户体验。目前NLP研究中对语义关注较多而对语体缺乏研究。例如:

eg1:他的行为对全体职工的工作鼓舞很大。→ 他的行为极大地鼓舞了全体职工的工作热情。

eg2:这件事得跟我们头儿说。→ 这件事情需要向我们领导汇报。

以上例子中前者是来自于真实语料，后者则是修改了部分词语，可以看出通过变换词语能使得语言更为正式，表达更为得体。因此以语义为基础，发掘同一语义下对语体词的分类具有一定的可行性和必要性。目前工业界和学术界广泛使用的《哈工大信息检索研究室同义词词林扩展版》（以下简称《大词林》）较为全面地覆盖了语言生活中常用的同义词词簇。本文尝试在提供了同一语义下选取词语的范围内进行语体的分类操作，并据此提出了“语体词标注”“语体（词）链条标注”和“平行构式标注”三个任务。它们旨在通过分类建立起以词语为基础的语体分类资源，更好地服务于语体研究，从而应用于自然语言处理以及对外汉语教学等领域。

2 相关研究

2.1 语体

语体一直是学界讨论的热点话题，对于语体的定义及分类，不断有学者提出自己的看法。唐松波(1984)提出语体是言语特点的综合。李泉(2004)认为语体即语言运用过程中产生的交际功能变体。冯胜利(2010)认为语体是一种交际手段，用来拉近、拉远或保持交际过程中双方的距离。关于语体的分类，以二分法为盛，即口语体和书面语体，不少学者再次细分出下位语体，主要有符淮青(1985)、胡裕树(1995)、邵敬敏(2001)等。此外，冯胜利(2010; 2017)提出“调距”功能角度，认为“正式体、非正式体、典雅体”为语体的三大基本范畴，崔希亮(2020)借鉴此分法，将语体区分为正式语体和非正式语体。

语体语法日益引起学界关注，选择不同层面的语体特征进行语体计量的研究也不断涌现。方梅(2013)通过不同语体材料的对比分析，说明句法特征具有语体分布差异。冯胜利等(2017)通过语体标注，从“量”“质”两方面证实了“语体不同，语法不同”。邹沁清等(2021)运用语料库和统计方法对汉语语体进行特征的计量研究，进一步实现自动分类任务。

2.2 语体资源

围绕语体资源的构建工作主要有语体语料库构建、语体词的词典编纂等。北京语言大学BCC语料库包含文学、报刊、对话、古汉语等多领域。北京大学CCL语料库中也构建了口语领域。冯胜利等(2017)构建了由叙事文、新闻、说明文等6类文体类型组成的12万字左右的语体语料库，从语法、韵律、语体信息三方面进行标注。关于语体词的词典编纂，《现代汉语词典》对于常用口语词、方言词、书面上的文言词语，分别标注<口><方><书>。此外，还有闵家骥等(1991)编著的《汉语方言常用词词典》、施光亨等(2012)编著的《汉语口语词词典》等。

2.3 语体词

对于语体词，部分学者从宏观上对语体词的分类、适用范围与构成进行了研究。关于语体词的分类，目前学界以三分法为主流，将现代汉语词汇分为书面语词、口语词和通用语词，主要学者有曹炜(2003)、符淮青(2004)等。关于语体词的适用范围，谢智香(2011)认为“口语词汇在日常口头交际中所使用，一般具有通俗易懂、风趣幽默的风格；书面语词汇在正式的交际场合使用，一般具有典雅、庄重的色彩”。关于语体词的构成，刘中富(2003)指出口语词汇除日常口语用词外，还包括俗语词以及方言词语，书面语词汇包括历史词语、文言词语、行业词语、

(c) 2022 中国中文信息学会计算语言学专业委员会

生僻的和较典雅的成语，本文认同以上说法。还有语体词专项研究，主要有苏新春(2007)、尹惠贞(2006)、张安娜(2015)等。

近年来，基于语料库的语体词计量越来越引起学者们的重视。张文贤等(2012)计算出1343对具有显著口语、书面语语体差异的同义词，得出“口语、书面语的同义词差异主要在词性以及音节上”。宋婧婧(2013)以有声媒体与平面媒体语料库作为口语与书面语的代表，对其使用词汇进行词频、词类、音节的定量对比。张佩(2021)经过BCC 语料库及其他语体材料的测量，对汉语作为第二语言易混语体词汇的教学提出建议。

总结前人研究可知，有关语体、语体词的研究数量颇丰，语体计量的研究也层出不穷。程雨民(2004) 指出“语体建立在同义性的基础上”“语体的实质是在一些使用场合上有区别的同一变体的选择”，张文贤等(2012)也认同此观点，因而平行语体资源的构建具有重要意义。但是，目前学界在此方面的工作缺乏，仅在语体词层面存在少量平行资源，且数量和规模非常有限，对于自然语言处理等相关应用的支持不足。因此，建设多层面、大规模的平行语体资源对语体的理论研究以及对外汉语教学、自然语言处理等应用领域均具有重要价值。本研究基于《大词林》，构建了由语体分类词表、语体链条和平行构式三个层面组成的平行语体分类资源，并从资源分布、语体词形态差异、词义分布与词性分布四方面展开分析。

3 语体词林资源建设

3.1 基础资源

《大词林》在《同义词词林》(梅家驹, 1983)的基础上，参照多部电子词典资源，并按照人民日报语料库中词语的出现频度，只保留频度不低于3（小规模语料的统计结果）的部分词语，剔除14,706个罕用词和非常用词，并进行扩充，最终的词表包含77,343条词语。《大词林》按照树状的层次结构把所有收录的词条组织到一起，把词汇分成大、中、小三类，经统计，大类有12个，中类有95个，小类有1,400个。目前基于《大词林》的研究主要集中在语义层面，语体层面的工作鲜有。而词汇是语言的建筑材料，语体词反映语体、甚至在一定情况下能决定语体，《大词林》拥有的丰富的同义词簇恰能为构建平行语体资源提供重要基础。因此，本研究以《大词林》为基础，构建了由口语词表、通用语词表、书面语词表、术语词表和多义词表组成的《语体分类词表》（下称《词表》）。

3.1.1 语体词标注规范

为提高语义对应的准确性，本研究仅对《大词林》的同义词簇进行语体词的划分，即由“=”连接的词簇，筛选得出9,995组词簇，55,844条词语。

关于语体词的分类，本研究采用学界主流的三分法，即将现代汉语词汇分为口语词、通用语词和书面语词，对于书面语词中的专业术语，单独建立术语词表，四者互不重合。

本研究的语体词划分方式将《现代汉语词典》（以下简称《现汉》）、BCC语料库测量和理论研究三者相结合，主要有以下考虑：《现汉》是一部规范、权威的语文词典，其中标<书>和<口><方>的词条可以为划分语体词提供直接依据，并可对具有多个义项的多义词语体进行精细的判定，但《现汉》标记的是最典型的语体词，判定作用有限。BCC语料库作为全面反映当今社会语言生活的大规模语料库，其中的报刊、对话领域可大致代表书面语体和口语体，通过观察词语在报刊、对话等领域的数量，可以较为科学客观地对词语语体进行量化测量。但使用BCC语料库难以对多义词各义项间的语体差异进行细致分辨；且报刊及对话领域仅能大致代表两种语体，并非完全泾渭分明，且受语料来源、词语使用范围等因素影响，结果有时与平日认知有偏差。通过总结前人理论研究成果，归纳典型语体词的判定方式可一定程度上规避此问题。本研究参考宋婧婧(2015)、陈振艳(2016)及高艳(2017)的研究成果，总结出16条口语词及3条书面语词的判定方式，并配以大量举例。但理论总结无法穷尽所有可能，主观成分较多。综上，《现汉》、BCC语料库测量和理论研究各有优劣，因此本研究将三者进行结合。

由上所述，多义词的不同义项在语体划分中起到直接作用。因此，本研究根据《现汉》，筛选出《大词林》中的多义词共12,463条，标注了多义词的词义、当前义项及对应例句。

Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China.

(c) 2022 中国中文信息学会计算语言学专业委员会

3.1.2 语体词标注实践

根据《大词林》语体词标注规范，由相关专业的10名本科生、研究生开展标注，具体如下：

(1) 标注多义词。根据《现汉》及词簇中其他同义词的含义，标注多义词的当前义项，并配以1-3句例句。无法在《现汉》中找到当前义项的多义词，做如下处理：若百度及其他词典资源，包括汉语大词典、百度汉语等中有此义项，则进行补充并备注来源，如“无对应义，参考百度汉语补充”；若均无法找到当前义项，则备注“无对应义”，此词语将不参与语体划分。

(2) 标注术语。通过“术语在线”平台对术语进行提取与标注。“术语在线”平台由全国科学技术名词审定委员会于2016年5月创办上线，为目前较为权威的术语知识服务平台。标注术语时，仅保留审定公布库中的规范术语。

(3) 标注语体。首先据《现汉》直接进行划分，对于标记<口><方>和<书>的词语，分别划分为口语词和书面语词。其次，依据理论总结的判定方式，对典型口语词和书面语词进行划分。对于其中语体不确定以及剩余未分类词语，参考BCC语料库中的例句进行判读：在语料库各领域中均出现的词，划分为通用语词；在对话领域中出现次数为0或次数很少，基本存在于报刊、科技领域中的词，划分为书面语词，反之，划分为口语词；在各领域出现次数均小于3的词，备注为“罕用词”，不参与语体的划分。

(4) 对口语词、通用词、书面语词、术语词、多义词分别标记符号O、G、W、T、M。对于复合词，标记时语体词符号在前，多义词符号在后，如“宝贝OM”。

对标注结果进行多轮校对后，最终我们构建了基于《大词林》的《词表》，共包含9,992组词簇，55,710条词语。

3.2 语体链条标注

3.2.1 语体链条标注规范

基于上述对《大词林》的分类，可以得到某一具体语义下的口语词、通用语词以及书面语词，通过对数据的整理和挖掘，可以发现相同语义下口语词到通用语词再到书面语词的转化和替换。而这一具体语义下，从口语词至通用语词至书面语词的链条，本文称之为“语体链条”。如：“俺→我→吾”，三种语体词可以表达同样的语义。具体的标注规范有：

(1) 选择适当的词组成链条。在《大词林》的词簇完成分类后，同一语体下存在多个词语，要保证各个语体词语义要基本对等，主要参照《现汉》词义以及之前给出多义词的当前义项。词语要保持感情色彩一致；时代色彩浓厚的词语进行剔除。词语之间用“→”隔开表示。

(2) 通过给语体词配以短语搭配或者例句来保持当前词义一致。《大词林》中的词语存在大量的多义词，在不同的上下文中有不同的意义，为保持当前的词义一致，选取恰当的短语搭配和例句可以帮助限制语体词的具体语义，可参照《大词林》分类形成的多义词例句，或参考BCC语料库以及CCL语料库，也可自行造句。对应链条的位置填入“搭配/例句”一栏中。

(3) 词语的词义为固定义。一般在词典中有记录的意义为词的固定义。但是有时词语在使用中的意义在词典中找不到，使用的是临时义，如修辞义。词语的临时义变动大并不固定，需要依靠相当篇幅的上下文，有一定的使用限制，因此本研究不考虑这样的临时义。

(4) 本轮标注不要求同时存在三种语体词，允许存在同一语体的多个语体词。同时存在意义相同的口语词、通用语词和书面语词是理想的情况，但并不会大量存在，因此允许链条存在两种语体词，在并不存在的那一类语体位置上，用“？”表示。也有许多词语可以在表示相同的语义下，仍然是出现在同一种语体中，这样的情况用“/”进行隔开。

3.2.2 语体链条标注实践

根据上述语体链条的标注规范，招募了11名语言学及应用语言学专业的研究生进行实际标注，并有3名质检员进行检查修改，经多轮培训和修改后完成标注任务。

(1) 标注得到大量语义基本一致的语体链条以及语体分明且语义一致的短语搭配及例句。而以语体词为基础，为语体研究提供了一个新的角度。短语搭配和例句与语体链条相对应，形成了一批具有语体色彩的平行语料，也具有重要的价值。

(2) 标注发现存在许多的单音节语素、专有名词。在标注过程中，存在许多如“青”“紫”“木”这样的单音节语素，它们很少单独使用，大多数情况则是组成词语，因此链条中不再选取。专有名词如“江淮戏”“淮剧”，两词意义完全相同，只是不同时期叫法不同，并且“江淮戏”已经不再出现在新时期的语料之中，这样的词语也不再收入链条之中。

(c) 2022 中国中文信息学会计算语言学专业委员会

(3) 标注发现存在许多意义泛化的术语，可以重新进行归类。一般情况下，认为术语是存在于书面语中，是书面语词的一部分，但是有许多术语在日常生活中广泛使用，如“冰雹”等，在口语和书面语中都有许多使用，也可以根据分类规范放入通用语词之中。

3.3 平行构式标注

Goldberg(1995)提出，构式就是指这样的形式—意义对,它在形式或意义方面所具有的某些特征不能完全从其组成成分或业已建立的其他构式中推导出来。构式语法的应用非常广泛(施春宏, 2017),且在实际应用中，词语语体转化必然影响其周边成分，构式信息因而显得格外重要。因而建设以链条词为核心的构式资源是一种必然选择。此时，平行构式不再需要严格进行语体三分，由非正式语体至正式语体的转换即可达到交际需求，即正式程度有所提高即可。如“打”后接名词性短语，在名词性短语表示织物时，“打”可以转化为“织”。本研究中将“打+np[2]”作为构式，当“np”指代“织物”时，与“织+np”组成平行构式。

3.3.1 平行构式标注规范

(1) 平行构式以词语为核心，参照BCC语料库检索规则，以词类搭配和特殊符号作为限制，词类限制有名词性短语、动词性短语、名词、动词、形容词等，特殊符号则包括标点符号“w”和分句符号“sent”等。词语与词类或特殊符号之间用“+”连接。

(2) 必要情况下需要进行语义限制。符合构式限制的语义，才可以进行构式的转化。若无语义限制，则可以直接转化构式。若无法归类语义，可以尽量穷举出该类别可搭配的词语。

(3) 非正式语体构式与正式语体构式成对存在。两种语体构式需要对应存在。

(4) 若无形式化描述，可以组成平行短语。即使有语义限制，也仍存在一些词语只在某些常用搭配中才会进行语体上的替换。因此，不存在或难以产生构式时，则产出平行短语。

3.3.2 平行构式标注实践

标注得到了一批有语义限制的平行构式，当词语搭配满足语义限制时，则可以进行构式转化；同时，也形成了一批由非正式到正式的可以无条件转换的平行短语。这些词语搭配以及平行构式所配有对应的实例集合，构成了一批由非正式语体至正式语体的平行语料。

4 资源分析

4.1 语体分类资源分布

4.1.1 语体词分布

《词表》的55,710 条词语中，口语词为5,743条，通用语词为23,133条，书面语词为26,834条（其中术语词3,713条），分别占比10.3%，41.5%和48.2%。由于《同义词词林》(梅家驹, 1983)编著初衷主要服务于翻译和写作，且在进行剔除与扩充时主要参照多种词典资源和人民日报语料库，因此词语书面化程度较高。本文统计了9,992组同义词簇的语体差异情况：

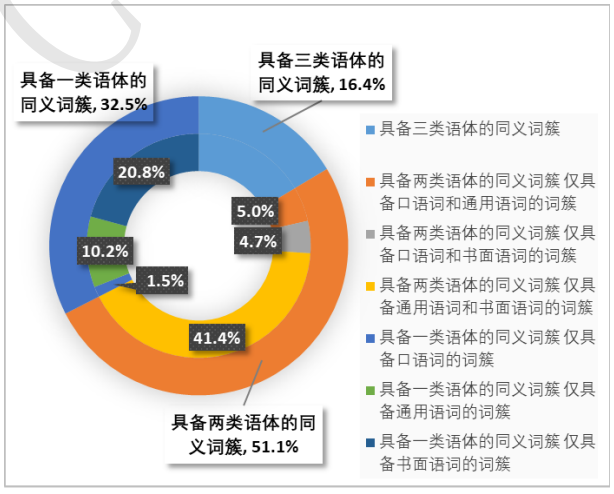


图 1: 《词表》中同义词簇语体差异情况

“np”表示名词性短语。

(c) 2022 中国中文信息学会计算语言学专业委员会

由图1可知，有32.5%的同义词语体单一，并无其他对应语体；51.1%的同义词具备两类语体，这类同义词占比最多，以具备通用语词和对应书面语词的同义词为主，比例高达总数的41.4%；三类语体都具备的同义词仅占16.4%。可见，语体差异只是同义词差异的一个方面，部分同义词并无语体方面的明显差异。

4.1.2 语体链条和平行构式分布

本研究得到语体链条5,017条，搭配或例句有4,432条，可以得知，同义词的语体差异是存在的，在同义词辨析中值得关注；同时，这种语体差异是非常依赖上下文语境的。

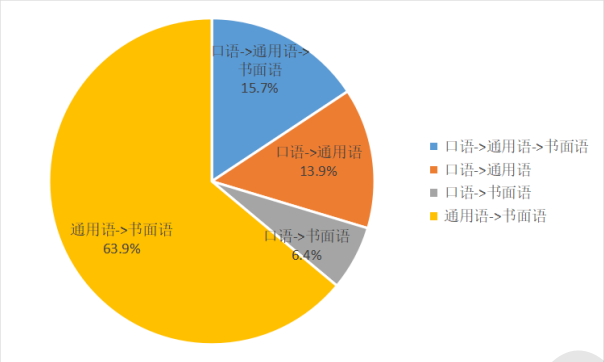


图 2: 语体链条分布

如图2所示，语体链条中，“口语词→通用语词→书面语词”链条有788条，占全部链条的15.7%，“口语词→通用语词”有699条，占全部链条的13.9%，“口语词→书面语词”链条有322条，占全部链条的6.4%，“通用语词→书面语词”链条有3,208条，占全部链条的63.9%。从中可以得知，同时出现在三种语体的语义极少，这也符合词义的概括性，词语反映的是一类事物或现象共同的特征；通用语词到书面语词的链条占比最高，一方面与《大词林》中收录文言词语相关，另一方面也反映出书面语词具有特殊的作用，有专门的使用场合。

最后，本研究得到非正式至正式的平行构式130对，是语体链条的2.6%。平行构式以语体链条为基础而得出，平行构式的数量相比语体链条是极少的，语体链条向平行构式的转化率不高，所以语体差异虽然可以用语法构式表现出来，但是效果并不理想，可以进一步改进。

4.2 语体词形态差异

4.2.1 语体词的词长

口语词、通用语词和书面语词在词长上各自具有其显著特征。《词表》中各语体词的词长占比统计如下：

	口语词	通用语词	书面语词
单音节	31.0%	14.1%	13.3%
双音节	37.6%	72.7%	52.3%
三音节	25.3%	5.7%	8.6%
多音节	6.1%	7.5%	25.3%
总计	100.0%	100.0%	100.0%

表 1: 《词表》中各类语体词词长占比

据表1可看出，在三类语体词中，双音节词均占比最高，尤其在通用语词中，双音节词占比达到了72.7%，其次是书面语词，也达到了52.3%，符合现代汉语中双音节词占优势这一基本特征。口语词中包含许多单音节实词，惯用语和俗语也多为三音节词，因此在口语词中单音节词和三音节词均占有一定比重，分别为31.0%和25.3%，这一占比远高于通用语词和书面语词。书面语词中具有大量的四字成语及类固定短语，其多音节词占比达25.3%，较口语词和通用语词高。对口语词、通用语词、书面语词的平均词长进行计算，得出三者分别为2.08、2.07、2.40，其中书面语词词长最大，口语词和通用语词二者相当。

分析以上差异产生的原因，主要是口语词较日常、随意，具有充分的语境及交际双方肢体动作、表情等辅助，允许表达的简洁、灵活而不会产生歧义，因此单音节词和三音节词占比较

(c) 2022 中国中文信息学会计算语言学专业委员会

高；书面语词较庄重、典雅，且语境较弱，需要更充分的表达以确保语义的准确，因此双音节词和多音节词占比较高，平均词长最大。

4.2.2 语体词构成语素

本研究调查了有多少共有语素是可以突破语体隔阂得以保存的，即在同一个链条或构式当中，有多少语素同时出现在口语词、通用语词和书面语当中。

	口语词→通用语词	通用语词→书面语词	口语词→书面语词	口语词→通用语词→书面语词
共有语素	445	1631	207	272
出现共有语素的链条	487	2545	196	301
语体链条	699	3208	322	788

表 2: 不同链条中共有语素数量

据表2可知，在不同语体中存在有许多共同出现的语素，其中，通用语词至书面语词的链条出现了最多的共有语素，有1,631个；其次是口语词至通用语词的链条，有445个；再次，口语词至通用语词至书面语词链条中共有语素，有272个；而口语词至书面语词链条中的共有语素最少，有207个，这与本身语体链条数量是相关的。与语体链条一致，横跨两种语体的语素数量较多，但是横跨三种语体的语素数量较少。另外发现，口语词至书面语词链条中，出现共有语素的链条数量少于共有语素的数量，说明该类链条中，跨语体存在的语素不止一个。

口语词→通用语词		口语词→通用语词		口语词→通用语词		口语词→通用语词	
语素	频次	语素	频次	语素	频次	语素	频次
下	7	不	28	鱼	4	老	5
手	6	人	25	小	3	年	4
不	5	风	21	前	3	实	4
后	4	心	20	家	2	不	4
子	4	信	17	子	2	大	3

表 3: 语体链条中出现最多的Top5语素及其频次

本研究继续统计了共有语素中出现频次最多的前5个语素，列举在表3中，可以看到有部分共有语素在不同链条中都会出现，如“下”“不”。

本研究同时也调查了平行构式中的共有语素，发现在130条构式中，有92条中出现了共有语素，这也说明同一语素是经常出现在不同语体中的。

4.3 语体词词义分布

4.3.1 《词表》词义分布

各语体词不仅在形态上具有差异，在词义分布上也各具特色。我们首先对《大词林》中各语体词的义项多少进行了统计。在《词表》的多义词中，实际标注语体的多义词共有12,414条，根据标签对各类多义语体词的数量及占比进行统计，结果如下：

	口语词	通用语词	书面语词	总计
数量（个）	1803	6431	4180	12414
占比	14.5%	51.8%	33.7%	100.0%

表 4: 《词表》中各类多义语体词数量及占比

如表4所示，多义语体词中通用语词最多，占51.8%，其次为书面语词、口语词。接着我们对各语体词中单义词、多义词的占比进行计算，得出如下结果：

	口语词	通用语词	书面语词
单义词	68.6%	72.2%	84.4%
多义词	31.4%	27.8%	15.6%
总计	100.0%	100.0%	100.0%

表 5: 《词表》中各类语体词单义词、多义词占比

第二十一届中国计算语言学大会论文集，第431页-第443页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

从表5可知，各语体词中单义情况均占绝大多数，多义情况较少，且随着词语正式程度的增加，单义词占比逐渐上升，多义词占比逐渐下降。究其原因，亦与各语体词的风格特征、语境强弱有关。从口语词到书面语词，场合逐渐庄重，语境依赖减弱，因而要求词义的表达更为细微、精准，以适应场合，避免交际障碍。

此外，本研究也对《词表》中语体词的语义类别分布状况作了进一步考察。我们以《大词林》的95个中类作为语义范畴，对各中类的语体词数量进行统计，得出口语词为三类语体词中使用数量最多的中类有7个，通用语词为三类语体词中使用数量最多的中类有37个，书面语词为使用数量最多的语体词的中类有51个，由于后两者中类数量较多，我们只取语体词使用数量前5的中类进行展示。具体数据如下：

所属中类	词性	口语/通用语/书面语词	所属中类	词性	口语/通用语/书面语词
Ah亲人眷属	名词	**225**/97/198	Ke呼叹	虚词	**47**/0/0
Kf 拟声	虚词	**168**/0/3	Kd辅助	虚词	**36**/13/22
Bc物体的部分	名词	**79**/41/65	Ac体态	名词	**27**/15/26
Bb拟状物	名词	**62**/22/30			

表 6: 《词表》中数量最多的语体词为口语词的中类（个）

由表6可知，当表达作为名词的亲人、眷属、物体的部分、拟状物、体态，比如“侄儿、大舅子、把子、耳子、疙瘩、片片、丑八怪、瘌痢头”等，以及作为虚词的拟声、呼叹、辅助的语义范畴，比如“叽里呱啦、吧唧、嘿、嗨、嘀哟、哇”等时，使用口语词居多。

所属中类	词性	口语词	通用语词	书面语词
Ed性质	形容词	469	**1623**	1280
Hj生活	动词	158	**1197**	857
Ka疏状	虚词	77	**858**	470
Gb心理活动	动词	89	**848**	394
Hc行政管理	动词	25	**623**	291

表 7: 《词表》中数量最多的语体词为通用语词的中类Top5（个）

由表7可知，当表达作为形容词的性质、境况，如“诚实、优秀、热闹、拥挤”，作为动词的生活、心理活动、行政管理，如“生活、过夜、想到、估计、安排、点名”，以及作为虚词的疏状的语义范畴，如“十分、最多”等时，使用通用语词居多。

所属中类	词性	口语词	通用语词	书面语词
Hi社交	动词	286	922	**1510**
Ee德才	形容词	250	1111	**1225**
Eb表象	形容词	250	751	**1206**
Dk文教	名词	60	658	**845**
Bh植物	名词	60	231	**806**

表 8: 《词表》中数量最多的语体词为书面语词的中类Top5（个）

由表8可知，当表达作为动词的社交，如“缔交、晤面”，作为形容词的德才、表象，如“笃实、披肝沥胆、寥若晨星、什锦”，以及作为名词的文教、植物的语义范畴，如“仰韶文化、彩陶文化、古柏、翠柏”等时，使用书面语词居多。

分析上述差异产生的原因，正如前文所述，口语词来源于日常并且较为随意、主观，因此涉及日常生活和显示出较强主观情感的人、物、拟声、呼叹等语义范畴时，多使用口语词。书面语词较为正式、严肃和客观，因此涉及社交以及表名物的文教、植物等语义范畴时，多使用书面语词。通用语词使用的范畴则更为广泛、多样。

4.3.2 语体链条和平行构式词义分布

本研究对于语体链条和平行构式在不同词义的分布方面也进行了统计，并进一步统计了语体链条数量占该词义类别词簇数量的比例，平行构式的数量以及平行构式数量占该类语体链条数量的比例与平行构式数量占该类意义《大词林》词簇数量的比例。如下图：

(c) 2022 中国中文信息学会计算语言学专业委员会

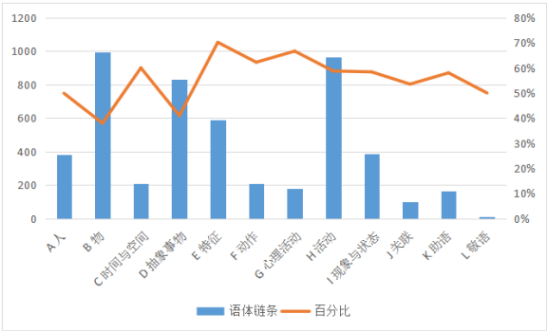


图 3: 不同词义类别中的语体链条

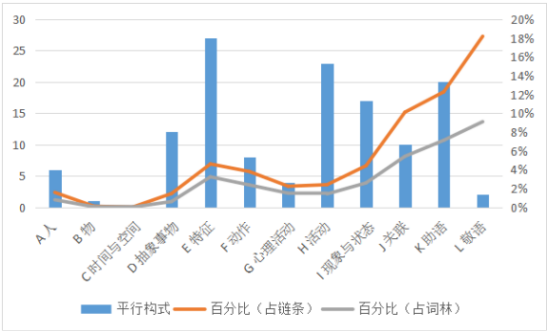


图 4: 不同词义类别中的平行构式

从图3可知，表示“物”和“活动”词义的词语中产出的语体链条数目最多，说明在这两项意义上，词语更为丰富，在语体上有所选择。但是从语体链条占《大词林》该类语义总词簇数的百分比来看，“物”的百分比最低，可以得知，虽然该词义下的词语存在语体的丰富性，但是同一意义下的可转化性不强，即尽管“物”义的词簇数量多，但是仅存在语体差异的同义词较少，该类意义的词语相对而言可以更多地同时存在于多种语体，并不完全需要因语体差异而改变词语的使用。另外，表示“特征”意义的语体链条占该意义下词簇数的比例最高，说明表示同一“特征”义的词语语体链条的能产性更强，在不同语体中更有可能有更多的选择,对于该类词簇中表示同一意义的词语与语体有极强的相关性，词语使用时要注意与所在语体相符合。

由图4可以发现：首先，平行构式占语体链条的比例走向与平行构式占该类意义《大词林》词簇数的比例走向大致是一致的，即分类的语体词中可以产出一定的语体链条，那么也可以相应地产出一定的平行构式，这也印证平行构式的出发点是可靠的。其次，各类意义下的平行构式占《大词林》词簇数和语体链条的比例均低于20%，可以看出，平行构式产出比例比较低，这与语义并不直接相关，产出只有语体差异的语法构式比较困难。最后，表示“关联”“助词”和“敬语”的平行构式百分比均超过了平均值，这说明该类意义中构式在不同语体下是非常丰富的，更容易在分类语体词和语体链条的基础上产出，更具有能产性。

4.4 语体词词性分布

4.4.1 《词表》词性分布

口语词、通用语词和书面语词在词性分布方面也不尽相同。本研究对《词表》中各语体词的词性占比进行了统计，数据如下：

	口语词	通用语词	书面语词
名词	41.2%	35.6%	46.6%
形容词	22.5%	18.9%	16.5%
动词	30.0%	40.5%	34.1%
虚词	6.2%	4.7%	2.7%
客套语	0.1%	0.3%	0.1%
总计	100.0%	100.0%	100.0%

表 9: 《词表》中各类语体词词性占比

从表9可知，在口语词和书面语词中，均为名词占比最高，分别高达41.2%和46.6%，其次分别是动词、形容词、虚词、客套语，在通用语词中，则是动词占比最高，达到40.5%，其次是名词、形容词、虚词和客套语。并且，书面语词的名词占比在三类语体词中最高，口语词中形容词和虚词的比重较另外两类更高，通用语词中动词和客套语的占比则为三类语体词中的最高。

4.4.2 语体链条和平行构式词性分布

语体链条和平行构式在词性分布方面也有其特征，本研究统计了语体链条和平行构式在不同词性下的数量，以及其占《大词林》该词性词簇数比重。

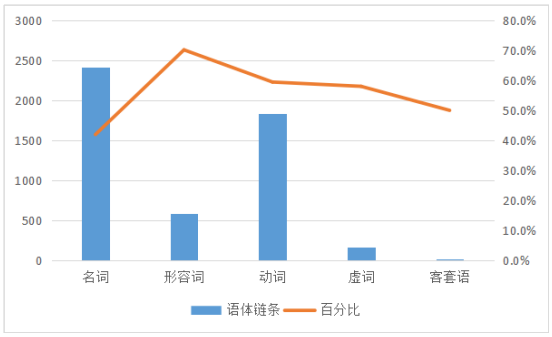


图 5: 不同词性中的语体链条

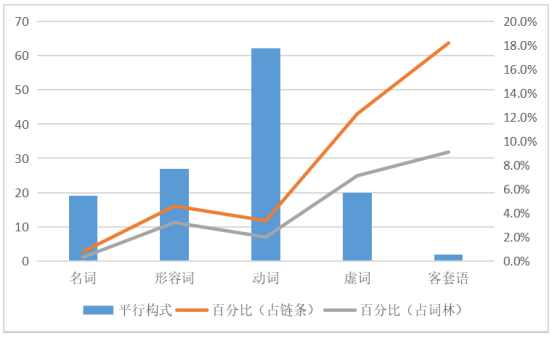


图 6: 不同词性下平行构式的分布

据图5可知，名词性的语体链条数量最多，其次是动词和形容词。形容词性的语体链条数目虽然不多，但是其占《大词林》形容词词簇数的比重最高，这也看出同义的形容词在不同语体中选择更多，表现更丰富；而名词性同义词在各语体中的选择更少，对于语体的敏感性稍弱。

如图6所示，与语体链条不同，动词性的平行构式数量最多，其次是形容词、虚词、名词和客套语，可以看出当词语范围扩展至构式时，动词仍然有比较好的表现。但动词性的平行构式占该词性的语体链条和词簇比重较低，而虚词和客套语的平行构式与之相反，说明这两类词性的平行构式更具有能产性，可在已有分类语体词和语体链条的基础上较好地进行扩充，这与平行构式词义上的分布趋势一致，亦是因为虚词和客套语的词性与助语和敬语的词义相对应。

5 结论

本研究基于《大词林》提出了语体词标注、语体（词）链条标注和平行构式标注三个任务，构建了一系列的语体分类资源，得到了《语体分类词表》、语体链条以及平行构式。《词表》中共包含9,992组词簇，55,710条词语；语体链条有5,017条，搭配或例句有4,432条；非正式至正式的平行构式130对，并且人工根据例句补充了303个平行构式。对应地制定了语体词标注规范，语体链条标注规范和平行构式标注规范。进而，本文对于语体分类资源进行了分析，描述了语体词、语体链条和平行构式在不同语体中的分布概况和形态差异：语体差异在同义词中值得关注，各语体词在词长、语义范畴与词性分布方面各具特色，不同语义与词性下，语体链条与平行构式的产出能力也不尽相同。

通过构建语体分类资源，可以为对外汉语教学和汉语作为第二语言的习得提供许多帮助。本资源也可以辅助教材编写，不同阶段和不同领域应有所侧重。其次，本资源提出的是中文语体相关的标注任务，相关的标注规范及实践的逻辑和经验也可以迁移至其他语言，其他语言可以以同义词词典为基础，依据本语言的相关语言资源及语体特点制定语体规范，得到有语体分类的词表，在此基础上进一步得到语体链条和平行构式，从而获得其他语言的语体资源。同时本资源可帮助进行语体改写、自动校对、语言润色等NLG工作，并且已经在腾讯文档中得到应用，起到了支持性的作用，取得了良好效果。

《大词林》的词语中含有许多文言文成分，各部分比重并不均衡，语义分布、词性分布都有其特点，后续可以对于这些问题进行改进。另外，《大词林》的词语资源虽然已经比较丰富，但产出的语体资源规模可能并非足够适用于庞杂的语言现实，因此可以在现有语体资源基础上进行实际的NLP测评任务，进一步体现其实际效能，同时，未来研究可以在现有资源基础上继续寻求语体词语资源，扩大语体词、语体链条和平行构式的规模。

参考文献

Wanxiang Che, Zhenghua Li, Ting Liu. 2010. Ltp: A chinese language technology platform. In *Coling 2010: Demonstrations*, pages 13–16.

Goldberg, AdeleE. 1995. *Constructions:A Construction Grammar Approach to Argument Structures*. Constructions:A Construction Grammar Approach to Argument Structures.

Qinqing Tai, Gaoqi Rao. 2021. 汉语语体特征的计量与分类研究(a study on the measurement and classification of chinese stylistic features). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 398–412.

(c) 2022 中国中文信息学会计算语言学专业委员会

冯胜利, 王永娜. 2017. 语体标注对语体语法和叙事、论说体的考察与发现. 汉语应用语言学研究, (1):15.

梅家驹. 1983. 同义词词林. 上海：上海辞书出版社.

张佩. 2021. 基于BCC语料库的词汇语体属性研究. 渤海大学.

冯胜利. 2010. 论语体的机制及其语法属性. 中国语文, (5):13.

刘中富. 2003. 实用汉语词汇. 安徽：安徽教育出版社.

唐松波. 1984. 文体、语体、风格、修辞的相互关系. 当代修辞学, (2):2.

宋婧婧. 2013. 汉语口语与书面语词汇使用对比分析——基于传媒语料库. 厦门理工学院学报, 21(3):88–92.

宋婧婧. 2015. 现代汉语口语词研究. 厦门：厦门大学出版社.

尹惠贞. 2006. 现代汉语口语词汇研究. 北京：北京语言大学.

崔希亮. 2020. 正式语体和非正式语体的分野. 汉语学报, (2):12.

张安娜. 2015. 现代汉语书面语词和口语词差异及其对应关系研究. 华东师范大学.

张文贤, 邱立坤, 宋作艳, 陈保亚. 2012. 基于语料库的汉语同义词语体差异定量分析. 汉语学习, (3).

方梅. 2013. 谈语体特征的句法表现. 当代修辞学, (2):8.

施光亨. 2012. 汉语口语词词典. 北京：商务印书馆.

施春宏. 2017. 构式语法的理论路径和应用空间. 汉语学报, (1):2–13.

曹炜. 2003. 现代汉语口语词和书面语词的差异初探. 语言教学与研究, 2003(6).

李泉. 2004. 面向对外汉语教学的语体研究的范围和内容. 汉语学习 ,2004(1).

程雨民. 2004. 英语语体学. 上海: 上海外语教育出版社.

符淮青. 2004. 现代汉语词汇（增订本）. 北京：北京大学出版社.

符淮青. 1985. 现代汉语词汇. 北京:北京大学出版社.

胡裕树. 1995. 现代汉语(重订本). 上海:上海教育出版社.

苏新春, 徐婷. 2007. 《现代汉语词典》标“书”词研究(下)——兼谈与古语词, 历史词, 旧词语的区别. 辞书研究, (2):38–44.

荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 2016. 大数据背景下BCC语料库的研制. 语料库语言学, 2016(1).

詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙. 2019. 北京大学CCL语料库的研制. 《语料库语言学》2019年第6卷第1期,总第11辑,pp.71-86.

谢智香. 2011. 论现代汉语口语词的特点. 西南石油大学学报*(*社会科学版*)*, 4(3):103–106.

邵敬敏. 2001. 现代汉语通论. 上海:上海教育出版社.

闵家骥. 1991. 汉语方言常用词词典. 浙江：浙江教育出版社.

陈振艳. 2016. 成语和类固定短语的语体鉴别及语体动因. 浙江树人大学学报*(*人文社会科学*)*, 6.

袁晖. 2004. 论语体词. 修辞学习.

高艳. 2017. 现代汉语口语词的主要类型及基本特征. 海外华文教育, (9):1188–1199.

CCL 2022

A 附录.各中类语体词统计表（个）

所属中类	口语词	通用语词	书面语词	标准差
Aa泛称	62	75	98	14.9
Ab男女老少	100	66	102	16.5
Ac体态	27	15	26	5.4
Ad籍属	7	25	27	9.0
Ae职业	83	202	448	152
Af身份	27	92	323	127.0
Ag状况	50	59	152	46.1
Ah亲人眷属	225	97	198	55.1
Ai辈次	19	22	74	25.2
Aj关系	68	205	165	57.5
Ak品性	95	68	122	22.0
Al才识	85	72	133	26.2
Am信仰	3	25	29	11.4
An丑类	52	46	105	26.5
Ba统称	43	107	263	92.4
Bb拟状物	62	22	30	17.3
Bc物体的部分	79	41	65	15.7
Bd天体	5	26	92	37.1
Be地貌	16	85	263	104.1
Bf气象	3	68	113	45.2
Bg自然物	24	144	322	122.4
Bh植物	60	231	806	319.1
Bi动物	132	148	471	156.2
Bj微生物	1	1	22	9.9
Bk全身	129	150	603	218.7
Bl排泄物分泌物	10	25	54	18.3
Bm材料	36	110	245	86.5
Bn建筑物	34	260	448	169.3
Bo机具	116	185	710	265.3
Bp用品	73	390	586	211.4
Bq衣物	21	146	120	53.9
Br食品药品毒品	76	264	231	82.0
Ca时间	71	581	278	209.4
Cb空间	60	395	495	186.0
Da事情情况	54	655	682	289.9
Db事理	30	176	139	62.0
Dc外貌	19	151	144	60.6
Dd性能	36	421	427	182.9
De性格才能	5	129	121	56.7
Df意识	19	371	279	149.1
Dg比喻物	20	72	64	22.9
Dh臆想物	12	81	66	29.6
Di社会政法	28	466	576	236.7
Dj经济	13	130	257	99.6
Dk文教	60	658	845	334.8
Dl疾病	31	33	181	70.2
Dm机构	20	129	276	104.9

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

Dn数量单位	64	313	242	104.7
Ea外形	76	239	223	73.4
Eb表象	250	751	1206	390.4
Ec颜色味道	139	190	104	35.3
Ed性质	469	1623	1280	483.9
Ee德才	250	1111	1225	435.2
Ef境况	109	469	399	155.8
Fa上肢动作	198	525	309	135.8
Fb下肢动作	21	122	100	43.4
Fc头部动作	50	215	226	80.5
Fd全身动作	18	95	72	32.3
Ga心理状态	63	405	789	296.6
Gb心理活动	89	848	394	311.8
Gc能愿	19	109	9	45.0
Ha政治活动	4	127	76	50.5
Hb军事活动	11	317	128	126.1
Hc行政管理	25	623	291	244.6
Hd生产	8	237	178	97.1
He经济活动	11	324	173	127.8
Hf交通运输	8	166	42	67.9
Hg教卫科研	112	298	437	133.1
Hh文体活动	23	76	53	21.7
Hi社交	286	922	1510	499.8
Hj生活	158	1197	857	432.5
Hk宗教活动	6	30	13	10.1
Hl迷信活动	4	10	6	2.5
Hm公安司法	35	80	166	54.3
Hn恶行	42	75	147	43.8
Ia自然现象	49	70	156	46.3
Ib生理现象	61	242	410	142.5
Ic表情	39	116	253	88.5
Id物体状态	50	336	464	173.1
Ie事态	35	298	326	131.1
If境遇	69	276	472	164.5
Ig始末	7	158	140	67.3
Ih变化	54	252	312	110.2
Ja联系	21	86	58	26.6
Jb异同	29	117	119	42.0
Jc配合	31	43	73	17.7
Jd存在	83	244	231	73.0
Je影响	5	332	149	133.8
Ka疏状	77	858	470	318.8
Kb中介	17	102	111	42.4
Kc联接	13	112	110	46.2
Kd辅助	36	13	22	9.5
Ke呼叹	47	0	0	22.2
Kf拟声	168	0	3	78.5
La敬语	3	61	24	24.0

CCL 2022

《二十四史》古代汉语语义依存图库构建

黄恬　　邵艳秋　　李炜*
北京语言大学，信息科学学院，
国家语言资源监测与研究平面媒体中心，
北京市海淀区学院路15号，100083
huangtian_blcu@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

摘要

语义依存图是NLP处理语义的深层分析方法，能够对句子中词与词之间的语义进行分析。该文针对古代汉语特点，在制定古代汉语语义依存图标注规范的基础上，以《二十四史》为语料来源，完成标注了规模为3000句的古代汉语语义依存图库，标注一致性的kappa值为78.83%。通过与现代汉语语义依存图库的对比，对依存图库基本情况进行统计，分析古代汉语的语义特色和规律。统计显示，古代汉语语义分布宏观上符合齐普夫定律，在语义事件描述上具有强烈的历史性叙事和正式文体特征，如以人物纪传为中心，时间、地点等周边角色描述细致，叙事语言冷静客观，缺少描述情态、语气、程度、时间状态等的修饰词语等。

关键词： 古代汉语；语义依存图；二十四史

Construction of Semantic Dependency Graph Bank of Ancient Chinese in twenty four histories

Tian Huang　　Yanqiu Shao　　Wei Li*
Information Science School, Beijing Language and Culture University,
National Language Resources Monitoring and Research Center,
15 Xueyuan Road, HaiDian District, Beijing, 100083
huangtian_blcu@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

Abstract

Semantic dependency graph is a deep analysis method of computer processing semantics, which can analyze the semantic relationship of sentences. In view of the characteristics of ancient Chinese, this paper formulates the annotation guidelines of Ancient Chinese semantic dependency graph and constructs the semantic dependency corpus of ancient Chinese which contains 3000 sentences from the twenty four histories, achieving the kappa value of annotation consistency is 78.83%, Finally, through the comparison with the semantic dependency Gragh Bank of modern Chinese, anlyzing the basic situation of the dependency library, the semantic characteristics and laws of ancient Chinese. Statistics show that the semantic distribution of ancient Chinese conforms to Zipf's law macroscopically, and has strong historical narrative and formal stylistic features in the description of semantic events, such as taking biographies of characters as the center, detailed description of surrounding roles such as time and place, calm and objective narrative language, less modal particles, etc.

Keywords: Ancient Chinese , semantic dependency graph , Twenty-four Histories

* 通讯作者　Corresponding Author

第二十一届中国计算语言学大会论文集，第444页-第454页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

语义分析是自然语言处理的核心问题，现代汉语领域进行了多种方法的研究，如基于词(刘琦, 2012)、基于短语(Xue, 2008)、基于句法树(Hacioglu, 2004)和基于依存图(Ding et al., 2014)的分析方法，取得了丰硕的成果。但古代汉语的语义研究相对匮乏，主要集中词汇层面上，如邹璐对《战国策》的副词进行意义归类(赵娟, 2005)，张丽丽对先秦帝王、诸侯谥号词汇进行语义分析归类并描绘出语义网(张丽丽, 2014)，舒蕾建设古汉语包含多义词的单音节词词义标注语料库(Shu et al., 2021)。在句子层级上，对句子的标注停留在句法关系，还未涉及深层次的语义关系标注，如京都大学建立了由《四书》构建的古代汉语依存树库(Yasuoka, 2019)。

随着时代发展，传统文化的专家研究和学习辅助亟待要求古代汉语信息化，对古代汉语语义分析提出了新的要求。文章以语义依存图理论为指导，结合古代汉语语法和语义特点，制定古代汉语语义依存图标注规范，并以《二十四史》为语料来源，对语料进行语义依存图标注，初步建立了3000句规模大小的古代汉语语义依存图库，对依存图库基本情况进行统计描述，并通过与现代汉语语义对比分析其语义特色。统计结果显示，《二十四史》语义标签频次分布总体上符合长尾分布，语义事件通过谓语动词紧密连接；叙事以人物事迹为中心，时间、地点等周边角色描述细致；使令句和目的句丰富，反映皇权制度下的上下等级制度和政治话语权；语言具有强烈的纪传体和历史题材风格特征。

2 现代汉语语义依存图和古代汉语语义依存图标注规范

2.1 现代汉语语义依存图

自然语言处理中传统的语义分析多采用语义依存树，依存树为句子中的每个词语（除核心词）找到它的依存词（父节点），并指出该词语与依存词之间的语义关系，传统的语义依存树结构规定每个句子成分只能有一个父节点与其存在依存关系，且不同的依存弧之间不能存在交叉现象。但汉语语序灵活、词类功能多样化，语言变式繁多，在真实语言现象中经常会出现某个词语同时依存于多个词语(王跃龙and 姬东鸿, 2007)，同一句子成分可能被多个成分支配，树结构不能满足汉语的真实语言现象表达，汉语的语义分析研究也从依存树走向依存图。如王悦龙提出构建一种全新的汉语依存图库(王跃龙and 姬东鸿, 2007)，哈尔滨工业大学(丁宇, 2014)结合鲁川定义的汉语意合网络语义关系体系和依存语法的特点构建了一套语义依存图关系体系，郑丽娟建立了包含30000句子的兼语句语义依存图库(郑丽娟and 邵艳秋, 2015)。图结构突破原有的依存树表达的限制，放宽了依存树的限制条件，主要表现在两个方面：（1）允许多父亲节点的出现；（2）允许非投射现象出现，即允许依存弧之间存在交叉(郑丽娟et al., 2014)。

如句子“我扶老奶奶上车”，用依存树分析的结构如图所示，句子的核心词为“扶”，“扶”直接支配“我”“老奶奶”和“上车”，他们之间的关系分别是施事、受事和后继关系，而“上车”还有另一动作参与者“老奶奶”则没有在树结构中直接指出。而用图结构分析，句子中的“我”有两个父节点“扶”和“上车”，都承担施事角色；“老奶奶”有两个父亲节点“扶”和“上车”，分别承担受事和施事角色。针对这个句子，机器自动问答想知道是“谁”“上车”，依存树不能准确分析出正确结果，依存图则能得出“我”和“老奶奶”两个动作参与者。

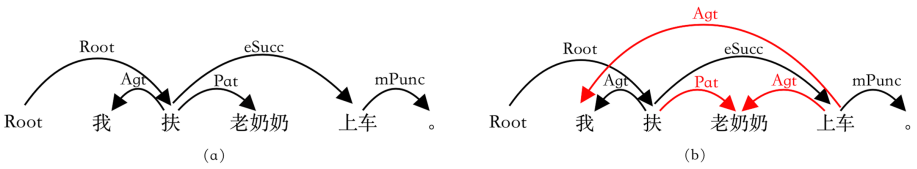


Figure 1: （a)为语义依存树；（b)为语义依存图

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金支持：本成果受国家自然科学基金项目（61872402）,教育部人文社科规划基金项目(17YJAZH068)，北京语言大学校级项目（中央高校基本科研业务费专项资金）（18ZDJ03）资助

2.2 古代汉语语义依存图标注规范

语义现象和规律具有继承性及相似性，古代汉语语义依存标注借鉴现代汉语语义依存的标注规范，以程荣辉等中文语义依存图标注体系[0]为参考，针对古代汉语的语法和语义特点，对其中的句子切分规则、语义关系、语义标签等进行调整，如下：

（1）分句采用字切分方式。古代汉语以单音词为主，在历史中逐渐演变为复音词，词汇在不同阶段有不同的形式。通过标注过程中的人工检验，我们发现语料历史时间跨度长，不论使用何种古汉语分词工具[1][2]，部分句子成分都难以界定语法单位为词或词组，如“兄弟”既可以作为一个词泛指同辈男子，也可以作为一个词组指哥哥和弟弟。为了提高标注效率，同时也便于进一步研究合成词的历时演变规律，后续预训练模型的输入，在切分古汉语句子时以单个汉字作为单位，但在依存标注时仍以词为单位做依存分析。词与词之间通过核心字进行连接，汉语词的核心语素通常是该词的最后一个汉字或前面的实义语素。如下图词语“天子”最后一个字“子”为该词的核心节点，“子”指向“天”，标合词“mHc”标签表示为一个词。命名实体“太尉”内部成分用“mHc-NR”连接，与其他词通过“尉”进行连接。

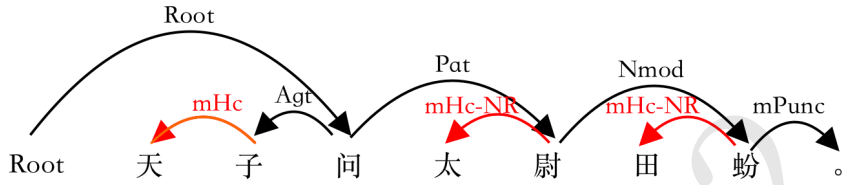


Figure 2: 合词和命名实体标注方式

（2）制定复音词标注规范。对标注过程中难以界定的复音词，本文制定了复音词标注规范。复音词由两个或两个以上音节构成，但在语法单位上都属于词。包括由一个语素构成的多音节单纯词和多个语素构成的合成词，复音词内部各个汉字通过合词标签mHc连接。判断复音词主要有意义和语法两个标准。单个或多个语素表示一个完整的意义单位时，这几个语素为一个分词单位，如单个语素的多音节词“参差”为一个词，表示“长短、高低不齐的样子”；两个语素“四海”结合构成了新义指代天下，两个相近的语素“亲戚”凝结成更具概括性的意义指内外亲属，重叠形式的复音词“世世”不是原义的简单重复，表示“后代子孙”。组合前后语法性质发生改变的复音词也划分为一个词单位，如“学”和“问”都为动词，结合后的“学问”为名词表示学问。

（3）制定命名实体标注规范。为了后期文本理解和知识图谱等应用，本文把命名实体划分为一个分词单位，并对难以界定的命名实体划分进行规范。本文的命名实体除普通人名、地名、组织名、机构名、时间、日期等，还包括特殊的人物名，如尊号加身份的称谓“孝武皇帝”；表示地形地貌的普通名词如“呼蚕水”“羌谷”；书名《蜀鉴》；带有类名的机构名“吏部”；历史朝代名称“唐朝”；古代刑法、学说等专用属于名“佐学”；历史上的重要事件、运动，如“渭水之盟”等。

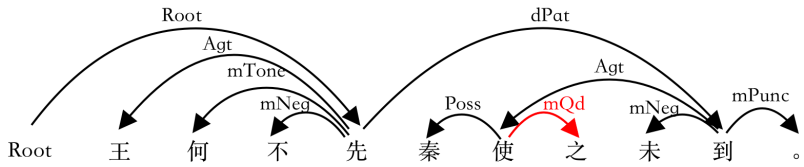
（4）增加特殊的语义标签。除了增加以上提到的mHc和mHc-NR语义标签，我们还增加了特殊的语义标签“mQd”取独标签，用来表示取消句子独立性作用的“之”与其依存词之间的语义关系。“之”在古代汉语中可作代词、动词、助词和语气词，使用频率高，其中表示取消句子独立性的助词用法为古代汉语“之”的特殊用法。“取独用法”用在主谓结构的主语和谓语之间，使这一主谓结构不能单独成句而成为句子里的一个成分。在语义依存分析中，“之”使该主谓结构表示的事件成为降级嵌套事件。如3“王何不先秦使之未到”中，“秦使之未到”是一个主谓结构，整个主谓结构整体作根结点“先”的宾语，成为“先”的降级嵌套受事。

3 语义依存图库建设

本文构建的古代汉语语义依存图库建设经历语料采集、语料预处理、标注工具完善、标注

[0]https://csdp-doc.readthedocs.io/
[1]甲言https://github.com/jiaeyan/Jiayan
[2]HanLP https://github.com/hankcs/HanLP/

(c) 2022 中国中文信息学会计算语言学专业委员会



（大王为什么不先于秦国使者没到达之前。）

Figure 3: “取独标签”

规范完善、标注人员培训、语料标注、标注规范再完善等流程，初步建立了3000句规模大小的的古代汉语语义依存图库。

3.1 语料采集和预处理

古代汉语语义依存图库语料选取自《二十四史》，二十四史历史跨越度长，基本由达官或著名文人兼官员负责编修(赵志伟, 2018)，以二十四史作为语料，具有可取体量大、覆盖不同朝代、收录全、用语规范等特点。标注前进行以下工作：

（1）语料采集。语料通过汉程网[3]和丁佳鹏[4]等整理的古文—现代文平行语料库以篇章为单位自动获取后进行筛选，整体筛选方法如图4，在原文语料上使用doc2vec(Le and Mikolov, 2014)工具获得对应的嵌入表征，再通过kmeans聚类算法(MacQueen and others, 1967)，对语料进行聚类，这样可以较好地保证语料的平衡性。通过肘部法则，本文将语料篇章分为了20个类，再从中随机选取3000个句子，通过进一步的人工审核，获得了最终用于构建语料库的古代汉语语义依存语料。相应的，标注平台页面添加了“显示原文”“隐藏原文”“显示译文”“隐藏译文”四个按钮，用于显示/隐藏语料所对应的原文段落和显示/隐藏语料所对应的段落译文。

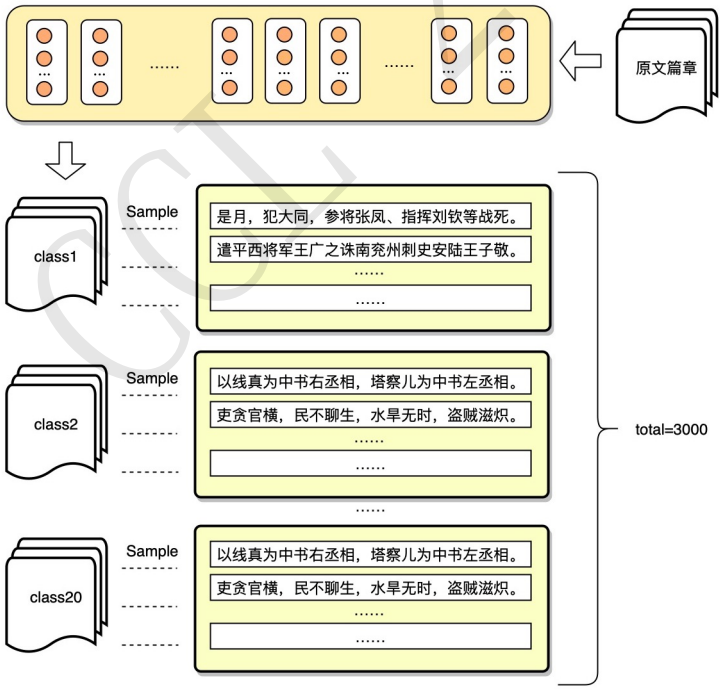


Figure 4: 语料筛选流程

（2）语料对齐。古代汉语相对于现代汉语语义理解的难度增加，为了帮助标注人员理解原

[3]http://www.httpcn.com
[4]https://github.com/NiuTrans/Classical-Modern

文语义，本文标注语料在采集时选取原文和对比翻译两部分，统一使用简体字。语料采集后对语料进行对齐工作。对齐时自动和人工校对相结合，以段对段即句子所在原文段落和所在翻译段落对齐的方式进行，部分对齐结果如图5。

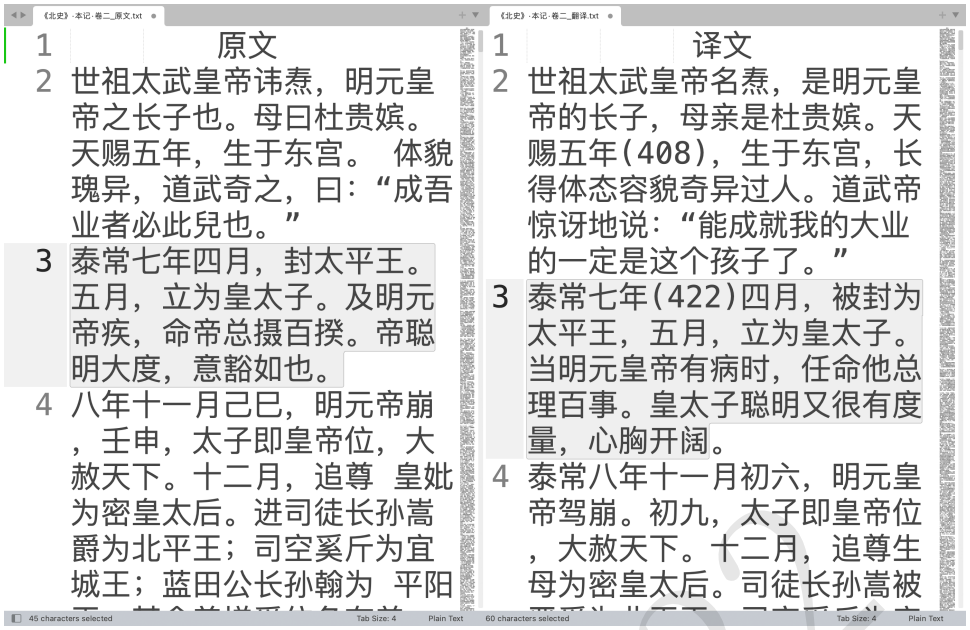


Figure 5: 部分对齐结果

(3)分句和词性标注。分句使用句末标点符号，即句号、问号、感叹号、省略号等进行分句，句长限制为15-30汉字。使用哈工大LTP语言云平台提供的词性标注模型对字（词）进行自动词性标注。通过以上的步骤，待标语料的处理已经基本完成。

3.2 语料标注

图库标注采取自动标注和人工标注相结合的方式，语料通过语义分析器自动进行语义标注后导入标注平台，标注人员在自动标注的基础上进行人工标注。共有5名语言学的硕士研究生参与标注工作，标注人员在经过培训后对语料进行标注，标注页面如图6。其中200条语料为5名标注人员共同标注，用于检查语料标注的一致性；另外2800条语料5名标注人员分别标注，历经五月余，五位标注员完成了3000句古汉语句子的语义依存的标注工作。

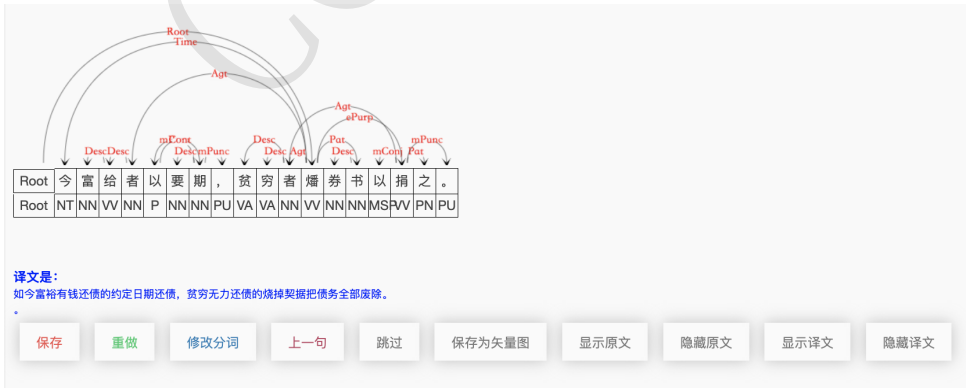


Figure 6: 标注页面

4 标注一致率

图库依存标注的一致率通过kappa值进行检验，公式如下：

(c) 2022 中国中文信息学会计算语言学专业委员会

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

其中P_0为总体分类精度，即标注一致的标签数量总和除以标注标签总数量；P_e为标注一致的标签种类的乘积与标签总数的平方的比例，经过计算kappa值为78.83%，整体一致率较高。句子标注不一致的原因主要有，（1）标注人员的古代汉语知识储备不能覆盖到所有知识，如对类似人名“曹旦”，单纯词“密迩”词义不了解；（2）古汉语中一些易混淆的语法结构，如兼语结构、连动结构和主谓短语作宾语的结构混淆导致标注语义标签也出现错误；（3）标注人员对标注规范熟悉程度不够，如细粒度结局标签（Cons）标注到粗粒度标签（CONS）上；（4）语义现象本身的模糊性，例如“九族”“亲疏”，后者做名词时，两者为并列关系，后者为形容词时，两者为修饰关系。针对以上问题，本文在后续标注过程中采取可改进措施，如加强标注人员的标注前培训，对准确率较低的标注人员取消标注资格，对易混淆的结构增订统一标注细则，对标注问题及时讨论，以进一步提高标注的准确性和统一性。

4.1 基于语义依存图库的现汉古汉对比分析

语义差异反映语言的特殊性，程荣辉[5]等已经建立了规模为2万句的现代汉语语义依存图库，本文通过现代汉语与古代汉语的对比，对古代汉语图库有个全面基本的认识，研究其特殊语言现象和规律。图库的语义标签反映词与词之间的语义关系，本文统计了古代汉语和现代汉语语义依存图库语义标签频次，图7结果直观地表明两者在语义标签分布上存在相关性和差异性。

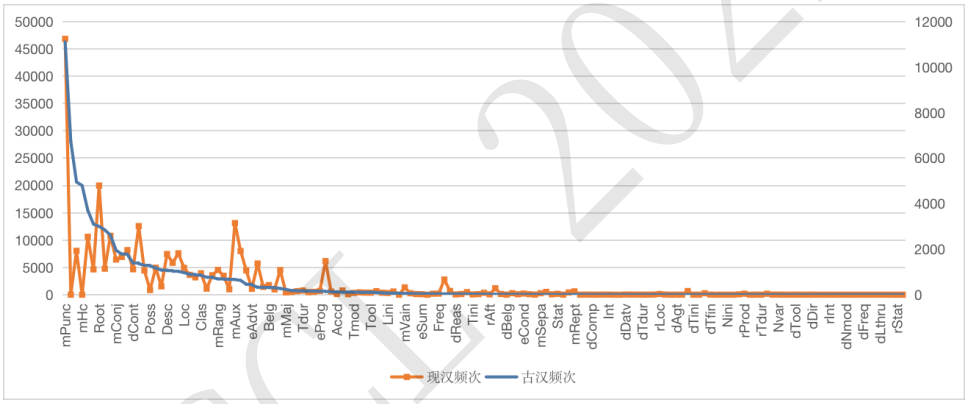


Figure 7: 现汉古汉语义标签频次分布

对标签数量进行皮尔逊卡方检验[6]，具体统计结果如表1。结果表明，卡方值和线性关联渐进显著性值都小于0.05，古代汉语和现代汉语语义标签分布具有统计学意义上的差异且有相关性。

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	11395.626a	9324	0.000
似然比	1073.584	9324	1.000
线性关联	87.925	1	0.000
有效个案数	149		
a 9520 个单元格(100.0%) 的期望计数小于5。最小期望计数为.01。			

Table 1: 现汉古汉语义标签数量卡方和相关性检验

[5]https://csdp-doc.readthedocs.io/
[6]统计检验使用SPSS Statistics 26.0.0.0得出

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 图库基本情况对比

对现代汉语和古代汉语语义依存图库基本情况进行统计，结果如表2。总体上看，古代汉语平均句长大于现代汉语，平均句语义标签数量比现代汉语更多，单句表示的语义内容比现代汉语更丰富。去除语法标签后将现古今汉语义标签频率分别由大到小排序，如图8。

语言类型	句数	字数	语义标签总数	平均句标签
现代汉语	20000	380839	275489	13.77
古代汉语	3000	72888	61736 （去除语法标签）	20.58

Table 2: 现汉古汉语义依存图库基本情况

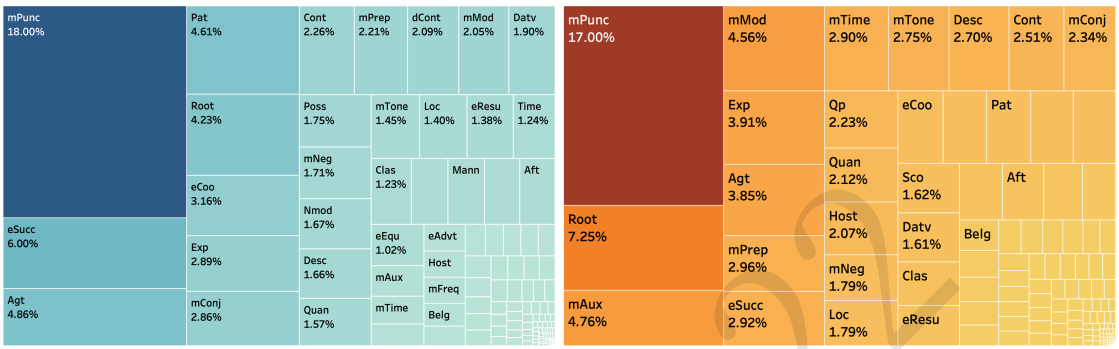


Figure 8: (a)古代汉语语义标签占比 (b)现代汉语语义标签占比

树状图显示，现汉和古汉语义标签总体上高频标签数量少占比大，低频标签数量多占比小，根据齐普夫定律(Wyllys, 1981)，语义标签出现的频次计作Pr，该语义标签的频次排位（即频级）计为r，在双对数坐标系下绘制出现代汉语和古代汉语语义标签频率分布曲线图9。可以看出古今汉语语义标签分布都符合齐普夫定律，说明现代汉语和古代汉语中在实际语言现象使用中都符合人类的省力和记忆原则，体现了语义现象的共性和规律。

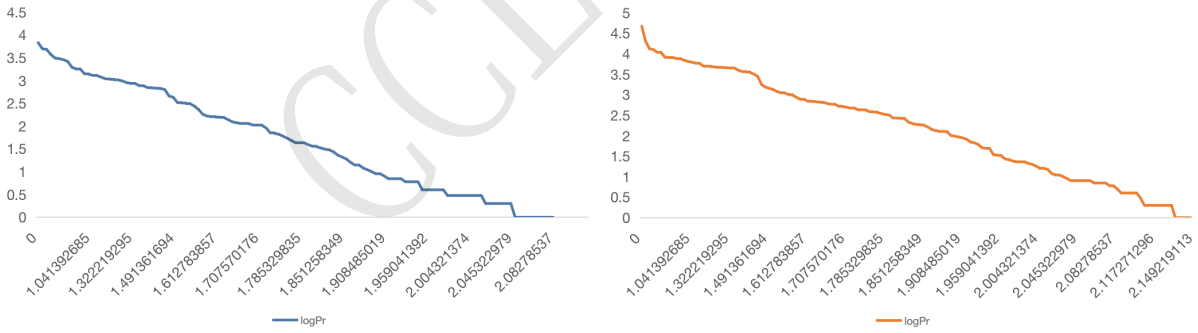


Figure 9: (a)古代汉语语义标签频次分布； (b)现代汉语语义标签频次分布

4.3 语义标签分布情况

受限于时间和语料规模，现汉图库和古汉图库规模大小不一致，以下统计分析采取将语义标签转化成相应所占比例方式进行对比。受限于篇幅，图中只展示重要靠前的语义标签，下面进行分类讨论：

4.3.1 周边角色标签

周边语义角色，即语义事件的参与者角色，包括主客体角色和情境角色，差异较大的标签如图10。语义事件由谓语支配，而主体角色作为动作行为的主要参与者，一般伴随谓语出现，

在语义事件中具有重要的地位。主体角色（施事Agt、当事Exp、感事Aft、领事Poss等）为动作的主体，客体角色（受事Pat、客事Cont、属事Belg等）为动作的第二参与者，多数为动作作用的对象。总体上，古今汉语主体角色的比例与客体角色不平衡，说明汉语中存在省略现象较多。

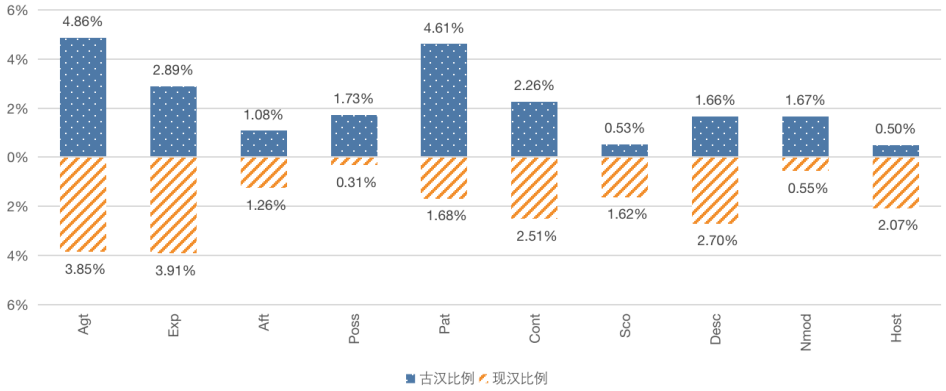


Figure 10: 现汉古汉周边角色标签占比

（1）施事（Agt）和受事（Pat）分别是自主性动作行为的发出者和承受者，当事（Exp）是非自主性动作行为的发出者，领事（Poss）表示领属关系的主体或整体部分关系的整体，客事（Cont）是事件所涉及但是并未改变的客体以及动作行为产生的新事物或结果，领事和当事常由非生命主体承担。统计显示，古代汉语中自主性动词多于非自主性动作行为，现汉中则反之，古汉领事占比明显高于古代汉语，都说明现汉中非生命主体作主语的句子更加丰富，对除人以外的事物描写更多。

（2）感事（Aft）表现心理活动的有意识的主体，二十四史属于历史传记，表示情感、心理活动的词少于现代汉语，叙事较客观冷静。

情境角色是事件涉及到的外围角色，主体使用的工具、材料，事件发生的时间、空间，引起事件发生的原因、目的等。

（1）范围（Sco）指的是事件中所关涉的方面、限定的界限、被审视的角度、发生作用的范围，如“诸贼”的“诸”，“其人”的“其”，现代汉语中描写事物时对议论的对象主体限制，事物的数量范围等限定较多，古代汉语集中于指称代词指称人物事物。现代汉语中描写事物时对数量范围、谈话对象限定较多,古代汉语限于指称代词指称人物事物。

（2）描写角色（Desc）表达的是一种特征，常作修饰成分，现代汉语中修饰成分更加复杂，多种性质的词和短语都可以做定语和状语，而古代汉语句子长度短，修饰成分数量远少于现代汉语，一般只用单一的修饰语修饰一个词，如“谩词”“寒隽之士”“杀父之仇”都用单个的述谓性质的词修饰一个名词。

（3）名称修饰语（Nmod）古代汉语占比例高主要是因为二十四史纪传体介绍人物年份等实体时描述详细，常与所属地、官职、年号等连用。宿主角色（Host）是属性的主体，或带有意义、功能、作用、价值的主体，通常出现的名词短语中，如“其奸恶”的“其”作为“奸恶”的主体，“宫庙大小”的“宫庙”，现代汉语中不管对人属性的描述还是对事物的功能作用描述都多于古代汉语。

4.3.2 依附标记标签

语义依附标记是对语义事件中依附性成分的标注，实际意义较虚，少单纯出现，但对句子语义有着一定作用，其中差异较大的标签如图11所示。

古代汉语词汇数量总体上少于现代汉语，语料涉及的事物范围更窄，在对事件情态、程度表达时也常常省略依附成分，现代汉语整体上限制语和修饰语更加丰富，表达更细致，表达主观情感的词数量更多。总体上，古代汉语依附标签占比低于现汉汉语，

（1）介词标记（mPrep）是对语法事件中介词和起到介词作用的词的标注。古代汉语此类标签主要用于表示时间、处所、方向、对象等的介词短语，表示目的句的“以”，所字结构的“所”字，及比较句中，如“于边境”“自数世以来”“请斩以谢天下”“所遇”“难于登天”等，而

(c) 2022 中国中文信息学会计算语言学专业委员会

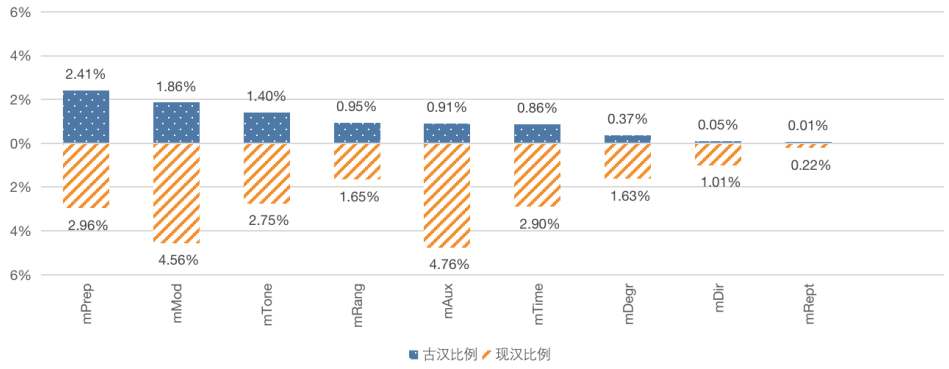


Figure 11: 现汉古汉依附标签占比

表示方式、依据、态度等则可以不使用介词，如“剑斩之”“法当斩”“兄事之”。除此，现代汉语“把”字句，“被”字句使用频率极高，介词一般不省略，分句当中起转折、并列等作用的词，如“还”“就”“才”等丰富，所以古代汉语介词使用比现代汉语更少。

（2）情态标记（mMod）是对句中表示情态的词进行的标注，这样的词一般表达的是主体的一种情形状态，比如惊讶、疑问、感叹或是能力、猜测等。古代汉语限于表示反问语气“岂”，必要性“须”“当”，可能性“可”，勇气“敢”，现汉比古汉更多表示推测、可能性、情感的词，如“只得”“总是”“真的”“就”等词汇，现代汉语情态词更丰富。

（3）语气标记（mTone）指对句中语气词的标注，二十四史语料的严肃性决定了语料的语气词较少，一般用于表示陈述、疑问、感叹、祈使等的少数虚词“也”“矣”“乎”“耳”“焉”等中，而现代汉语中多语气词“的”“了”“呢”“吧”“吗”“啊”“呀”等种类丰富得多。

（4）范围标记（mRang）是对句中表示范围的词进行的标注，可以是空间、时间或所指对象的范围，古代汉语一般出现在表示范围的具体位置，如“南羌中”“五品以下”，数量限制词“皆”“各”中。现代社会人们物质生活更加多丰富多彩，语句更加白话化，范围限制词使用频率更高，还会用在非生命主体作话话题限制谈话的内容当中。

（5）的字标记（mAux）是对汉语中出现的结构助词“的、地、得、之”进行的标注。古代汉语多为单音节词，修饰事物时修饰语可直接加被修饰语，如“锐兵”，少部分双音节词作修饰语时才用“之”进行连接，如“杀父之仇”，在近明清时期语言才更加白话化，使用“的”连接定语修饰语和被修饰语，如“他的官府”，而古代汉语补语数量更少，一般用“之”字，如“思之深”。现代汉语句子词汇语法形式发生变化，多音节词成主要优势，做定语、补语、状语等修饰成分常用结构助词连接。

（6）时间标记（mTime）是对句中一些时间副词以及动态助词的标注，如“正”“从此”等。程度标记（mDegr）是对句中表示程度的词进行的标注，如“广延百里”“大喜”。趋向标记（mDir）是对句中表示趋向的词进行的标注，如“蹈海去”“东征”等。二十四史记录的是已经发生的历史事件，作为官方历史材料叙述时冷静客观，较少描述时间动态的词语。重复标记(mRept)，如“唯唯”。这些标签都与语料范围有关，二十四史为正式语体，缺乏口语材料，在描述事件动作时对动作的状态，情态的程度上描述较少。口语化表达的“去”和语句重复也出现较少。

4.3.3 语义结构关系标签

语义结构关系的标注对象是语义事件，描述的是不同述谓概念之间形成的各种结构关系，差异较大的标签有后继标签（eSucc）、并列标签（eCoo）、目的标签（ePurp）、等同标签（eEqu）、目的标签（ePurp），如图12。

（1）后继事件（eSucc）指接着先行事件发生的事件，在时间上或逻辑上或空间上发生在后的后续性事件。古代汉语叙述事件表达简略紧凑，数词名词都能做谓语核心，一个句子中的几个分句往往有多个谓语，后继关系标签频词多。如“有司承旨奏戏，免为庶人”，“承旨”“奏戏”“免庶人”几个先后连续发生的事件共用一个主语，连接紧密，两个分句就包含三个动作，后继事件描述更丰富。

（2）并列（eCoo）表示的是前后两个事件或多个事件，其除了表示事件关系之外还能表

(c) 2022 中国中文信息学会计算语言学专业委员会

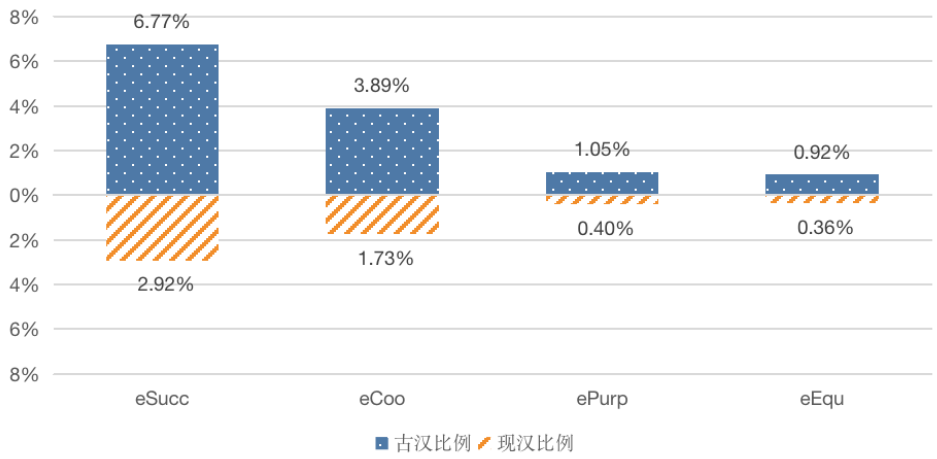


Figure 12: 现汉古汉语义结构关系标签占比

示平行的语义关系，在古代汉语中单音节词占多数，相近语义的词常并列使用，如“攻击”为“进攻击打”，“改易”为“改变更换”，“田庄”为田地庄园。分句结构中也常用相似结构对举并列，如“既而改元天历，郊庙，建后，立储”，“改”“郊”“建”“立”四个动并叙排列，使用并列结构多于现代汉语。

（3）目的标签（ePurp）是通过某些手段而要达到的目的性事件，统计发现，二十四史历史事件中使令句和目的句丰富，一定程度上反映了二十四史的官方政治历史题材和人物的等级关系，如“使人函封汉使者节赛上”，“使”与“函封“两事件为目的关系，反映上级对下级的命令。等同标签（eEqu）是对于同一个事物的复指或注释，和纪传体以人物为中心叙事有关，如“丞相方进“，“丞相”“方进”之间为等同关系。标签一定程度上说明了二十四史强烈的历史性叙事特征。

4.3.4 缺失语义标签对比

对现代汉语和古代汉语未出现的标签进行统计，去除语法标签合词标签（mHc）和命名实体标签（mHc-NR），结果表明，古代汉语特有的语义标签有4种，按频次高低排序分别为取独标签（mQd）、反描写（rDesc）、嵌套结局（dCons）、嵌套时距（dTrang）、反数量（rQuan）。mQd为古代汉语“之”做取消句子独立性作用时使用的标签，这是古代汉语的特殊语法现象。降级事件dCons、dTrang，出现古代汉语话语句中，二十四史多人物言论，同样的语义角色出现在话语中便成为降级事件。反关系rDesc和rQuan出现在“形容词”+“者”和“数量词”+“者”的结构中，如“贤者”“近者”“降者”“一辈大者”，古代汉语中用此结构代称一类人，而现代汉语中指称此类用助词结构“的”把修饰成分介绍给核心词。

现代汉语特有的语义标签有26种，按频次高低排序分别是先行关系（ePrec）、插入语（mPars）、反方式（rMann）、变化量（Nvar）、嵌套宿主（dHost）、嵌套终止状态（dSfin）、嵌套工具（dTool）、反结局（rCons）、嵌套源事（dOrig）、嵌套趋向(dDir)、反终处所(rLfin)、嵌套起始状态(dSini)、反意图(rInt)、反通过处所(rLthru)、嵌套成事(dProd)、嵌套名词修饰语(dNmod)、嵌套终处所（dLfin）、反比较（rComp）、嵌套频率（dFreq）、反数量短语（rQp）、嵌套数量短语（dQp）、嵌套通过处所（dLthru）、反源事（rOrig）、反趋向（rDir）、反状态（rStat）、反时间起点（rTini）。除了插入语（mPars）为现代汉语特有现象外，其他标签多为降级事件标签和反关系标签，其本身在现代汉语中出现数量少，说明这几类语义现象在语言环境中为不常见现象，而现代汉语语料规模更大，涵盖的更多的语言现象和语义关系。另外，随着时间发展，现代社会新事物不断增多，语言上词汇和句法形式更加复杂，语义现象种类也更加繁多。

4.4 结论及未来工作

本文在已有的标注规范及语料资源基础上，以《二十四史》作为古代汉语语义依存语料库的语料，建立古代汉语语义依存图库，并对语料库统计分析，发现古代汉语的语义特色和规

(c) 2022 中国中文信息学会计算语言学专业委员会

律。统计显示，二十四史语义关系总体上高频词数量少，低频词数量多，整体符合齐普夫定律，语义事件谓词连接紧密，叙述事件以人物为中心，常引用人物话语，地点、时间等周边角色描述细致，目的句和使动句较多，具有更强的政治色彩和书面语叙事特征。

该工作的不足之处是受限制于时间和人力成本，语料库规模较小，语料范围局限于历史题材和正式语体，不能完全反映古代汉语的语言规律，对古代汉语语言规律刻画描述还不够细致，标注受专业知识限制大，标注规范还不能囊括古代汉语所有的语言现象。在未来工作中，我们将进一步提高语料质量，扩大图库的语料规模和覆盖范围，对语义现象和原因作更加细致的分析。将来，依存图库还能应用于机器翻译，及古代汉语教学等更广阔的人文文化场景中。

参考文献

Yu Ding, Yanqiu Shao, Wanxiang Che, and Ting Liu. 2014. Dependency graph based chinese semantic parsing. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 58–69. Springer.

K. Hacioglu. 2004. Semantic role labeling using dependency trees. *Association for Computational Linguistics*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, number 14, pages 281–297. Oakland, CA, USA.

Lei Shu, Yiluan Guo, Huiping Wang, Xuetao Zhang, and Renfen Hu. 2021. 古汉语词义标注语料库的构建及应用研究(the construction and application of Ancient Chinese corpus with word sense annotation). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 549–563, Huhhot, China, August. Chinese Information Processing Society of China.

Ronald E Wyllys. 1981. Empirical and theoretical bases of zipf's law.

N. Xue. 2008. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255.

Koichi Yasuoka. 2019. Universal dependencies treebank of the four books in classical chinese. In *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28. Digital Archives and Digital Humanities.

丁宇. 2014. 基于依存图的中文语义分析. Ph.D. thesis, 哈尔滨: 哈尔滨工业大学.

刘琦. 2012. 一种基于*WordNet*上下文的词义消歧算法. Ph.D. thesis, 吉林大学.

张丽丽. 2014. 先秦时期帝王、诸侯谥号词汇—语义系统研究. Ph.D. thesis, 山西师范大学.

王跃龙and 姬东鸿. 2007. 汉语依存图库建设研究. In 中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集.

赵娟. 2005. 《战国策》副词研究. 山东师范大学.

赵志伟. 2018. 从"前四史" 到"二十四史". 中学语文教学, 9.

郑丽娟and 邵艳秋. 2015. 基于语义依存图库的兼语句句模研究. 中文信息学报, 29(6):8.

郑丽娟, 邵艳秋, and 杨尔弘. 2014. 中文非投射语义依存现象分析研究. 中文信息学报, 28(6):41–47.

(c) 2022 中国中文信息学会计算语言学专业委员会

中文专利关键信息语料库的构建研究

张文婷[*]，赵美含[*]，马翊轩[*]，王文瑞[*]，刘宇哲[*]，杨沐昀[#]
哈尔滨工业大学计算学部，黑龙江哈尔滨150001
120L021002@stu.hit.edu.cn; yangmuyun@hit.edu.cn
邓宇
哈尔滨市阳光惠远知识产权代理有限公司，黑龙江哈尔滨150000
dy@shineip.com

摘要

专利文献是一种重要的技术文献，是知识产权强国的重要工作内容。目前专利语料库多集中于信息检索、机器翻译以及文本文分类等领域，尚缺乏更细粒度的标注，不足以支持问答、阅读理解等新形态的人工智能技术研发。本文面向专利智能分析的需要，提出了从解决问题、技术手段、效果三个角度对发明专利进行专利标注，并最终构建了包含313篇的中文专利关键信息语料库。利用命名实体识别技术对语料库关键信息进行识别和验证，表明专利关键信息的识别是不同于领域命名实体识别的更大粒度的信息抽取难题。

关键词： 专利；语料库；关键信息

Research on the construction of Chinese patent key information corpus

Wenting Zhang[*], Meihan Zhao[*], Yixuan Ma[*], Wenrui Wang[*], Yuzhe Liu[*], Muyun Yang[#]
Harbin Insititute of Technology, Harbin, Heilongjiang 150001, China
120L020815@stu.hit.edu.cn
Yu Deng
Harbin Shineip Intellectual Property Corporation, Harbin, Heilongjiang 150001, China
dy@shineip.com

Abstract

As a kind of important technology document, the patent is of substantial significance to the national intellectual property strategy in China. Existing patent corpus are mostly for the purpose of information retrieval and machine translation task, leaving the fine-grained annotated patent less touched. To facilitate the forth-coming intelligent patent technology development, this paper constructs a Patent Key Information Corpus, consisting of 313 patents annotated with the issues, methods and effects in the texts. Then the SOTA named entity recognition models are applied to the corpus, and the sharping decrease in the performance indicate the automatic identification of the key information in a patent is a challenging IE task.

Keywords: Patent , Corpus , Key Information

1 引言

[*] 排名作者不分先后，同等贡献
[#] 通讯作者
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

专利是专利权的简称，是由专利机构依据发明申请所颁发的一种文件 (国家知识产权局, 2008)。相对于其他文献形式，专利作为技术信息最有效的载体，不仅囊括了全球最新的技术情报，而且更具有新颖、实用、可比较、结构一致的特征。专利信息的分析利用可为企业提供技术发展路线、竞争对手动态、重点专利技术方案和技术功效矩阵，是高效开展技术攻关活动不可或缺的助手(张晓林, 2018; Balsmeier et al., 2018; 李华锋et al., 2017)。

专利被认为是世界上最大的技术信息来源。据世界知识产权组织公布的2021年度《世界知识产权指标》(WIPO, 2021)显示：2020年全球发明专利申请量达到327万件，中国发明专利申请量达150万件。现有的专利自动处理技术主要围绕检检索和翻译等文本层次粒度展开(Narin F., 1994; Mohammad Hamdaqa and Abdelwahab Hamou-Lhadj, 2009)，尚未深入探讨对专利内容的深层次细粒度的分析和理解，无法满足海量的专利数据的智能化分析及加工需求。另一方面，自然语言处理技术的近来日益成熟，信息抽取以及阅读理解应用愈发广泛。但是，当前主流的基于深度学习的自然语言处理模型往往依赖于大规模高质量的标注语料库(Sunet al., 2017)，而现有专利的语料库主要用于信息检索和跨语言翻译目的，缺乏细粒度语义标注的语料库，尚不能有效支撑智能专利处理技术的研发。

针对上述问题，本文提出聚焦专利的关键信息进行标注，提出围绕专利中的技术问题、技术方法以及技术效果这的三要素来标注该专利的关键信息。文中给出了初步的标注标准，完成了313件的专利关键信息标注语料库。并在此基础上，验证了现有主流命名实习识别模型用于自动抽取这些专利关键信息的效果，揭示了相对于日益成熟的命名实习识别任务，专利关键信息的抽取问题更具挑战性。

本文后续内容组织如下：第1节介绍目前国内外的专利语料库的相关工作，第2节介绍专利关键信息标注设计思路，第3节给出语料库构建过程，第4节对现有名实体识别技术用于专利关键信息识别进行了初步探索并给出了实验结果，第5节总结全文。

2 国内外专利语料库及概况

国外的专利领域语料库中，检索语料库和平行语料库是较为丰富的，CLEF-IP 2009 (Roda et al., 2009)、TREC-CHEM 2009 (Lupuet al., 2009)等较早时间的发布的数据集都是作为检索语料库出现的，可用于专利分类、检索等。之后出现的的SureChEMBL(Georgeet al., 2016)语料库涵盖了更多的专利文本，并且算法支持关键字检索和化学实体识别。而在平行语料库领域，典型的语料库有NTCIR6日-英双语平行语料库(Utiyama and Hi Masao, 2007)以及包含六种语言的Sentence-Aligned 欧洲专利语料库(**?**)，作者各自选取了大量专利文本，以自动标注的方式进行了句子对齐，上述语料库能为机器翻译，多语言词典等任务提供支持。

在专利领域，也出现了若干相对更为细粒度信息的标注实践。Dmitriy Korobkin等对USPTO 和RosPatent 数据库中的专利文本进行了物理效果和技术功能的标注(Korobkinet al., 2019)，；Saber A. Akhondi1等(Akhondi, S. A.et al., 2014)以手动注释为主的方式建立了一个化学专利语料库，对化合物所属种类，可合成的药物，应用目标，作用方式等做了详细注释，最终得到约76万份专利标注。

国内公开的专利语料库中，Bin Lu等(Lu, Bet al., 2009)在09年就公开过一个平行的中英文专利文本语料库，对中英文语句进行了对齐。其次还有若干检索语料库，如翟东升等(翟东升et al., 2013)对德温特专利数据库进行了信息清洗和标注，构建出了一个可用于检索分析的专利信息语料库。章成志等学者个人信息及专利成果过更加注重细粒度的手工标注，数据集主要用途是学者的画像生成(高扬et al., 2019)。

下表为本文收集整理的现有主要专利数据集的概要信息。

专利以外的各领域专业文本中，更细粒度信息的语料标注开展相对较多。李智恒(李智恒et al., 2018)人从生物医学文献中抽取化学物质致病关系的系统，崔博文(唐晓波and 刘志源, 2021)对自由文本电子病历进行了命名实体以及实体间关系的抽取，唐晓波和刘志源(唐晓波and 刘志源, 2021)针对中文金融文本领域，对重叠性较高的实体关系进行了识别，如“买卖”，“股权”，“合作”等。

当前，人工智能技术迅速发展。就自然语言处理领域来说，诸如推荐、问答等技术相对成熟，阅读理解、对话、自动写作不断取得突破。对比这些任务中所使用的标注语料的信息粒度和语义检测，以专利为代表的专业技术文本语料库建设明显存在滞后。探索建立更细粒度、更深层次能够支撑智能专利分析和理解技术研发的专利标注语料库，已经成为一项期亟需解决的

(c) 2022 中国中文信息学会计算语言学专业委员会

问题。

语料库/数据集	用途	语言	数据规模	标注方式
CLEF-IP 2009 Collection	现有技术搜索	英语、法语、德语	1985年至2000年100万专利和2001年至2006年50万专利	自动标注加手工标注
TREC-CHEM 2009	技术调查、现有技术搜索	各国语言	约120万份化学专利和5.9万篇科技文献	自动标注
SureChEMBL	基于关键字的检索功能	英文、德文、法文为主	1400多万份专利文件	自动标注
NTCIR6 Japanese-English Patent Parallel Corpus	机器翻译，检索	日-英平行语料（互译）	大约200万个自动对齐的句子对	自动标注为主
Sentence-Aligned European Patent Corpus	专利翻译	6种欧洲语言	1.3亿对句子	自动标注为主
a Matrix “Physical Effects –Technical Functions”	提取物理效果和技术功能	英语，俄语为主	数据库超两千万篇专利文献（未全部构建）	自动标注
Annotated Chemical Patent Corpus	分析专利内容，用于生物、化学领域的研究探索	英语为主	最终包括约76万份专利标注	手动标注
Chinese-English Patent Parallel Corpus	中英句子对齐，数据清洗	中英平行语料（互译）	16w句子对	自动标注
德温特专利数据库（DII—Derwent Innovation Index）	数据清洗，简单信息标注，构建高质量专利数据集用于专利分析和知识发现	英语	测试大小8万条专利，数据库专利一千万条以上	自动标注
杰出人才精准画像构建语料库	用户画像生成	中文	国内一亿余件专利信息	手工标注

表 1: 部分专利语料库的概要信息

3 专利关键信息的标注设计

本文旨在聚焦专利中最有价值的信息进行标注，以期支持智能专利分析技术的研发。作为阶段性成果，本节聚焦专利的关键信息进行标注，即提取技术问题、技术方法以及技术效果三类关键词来概括整篇专利，并给出了初步的标注原则。

3.1 专利文本标注需求分析

目前，针对专利的标注内容和标注粒度并没有统一的范式，针对不同的具体任务，需求各不相同。比如从专利竞争分析角度出发，专利的所属权较为关键；而从专利的行业分析角度出发，专利的所属领域更加关键。考虑到专利本身的技术文献属性，本文在首先考虑的是更为广泛和经典的技术术语标注。但是，专利的技术术语无法完整的刻画一篇专利。表2所示，其中列出的示例1和2的两个专利的术语列表，虽然大致表示了这两个专利的领域和相关技术，但是对于其专利要点以及区分这两个同主题（工业机器人）的专利来说，作用并不显著。

(c) 2022 中国中文信息学会计算语言学专业委员会

示例二：本发明公开了一种考虑系统延迟的不确定工业机器人运动控制方法，首先建立工业机器人机电耦合非线性动力学模型，再利用反馈线性化技术使工业机器人非线性动力学方程线性化，构建动态递归神经网络估计并补偿系统的不确定性，最后提出改进的Smith预测控制方法消除系统延迟的影响。本发明对于系统延迟和不确定性参数具有较好鲁棒性，极大地提高了工业机器人的控制精度。示例*1*：本发明公开了一种工业机器人模型仿真控制方法及装置。其中，该方法包括：接收由三维建模软件构建的工业机器人模型；基于工业机器人模型确定控制参数；根据控制参数确定工业机器人仿真机械模型；根据小脑模型神经网络*CMAC*控制策略和比例积分微分*PID*控制策略对工业机器人仿真机械模型进行仿真控制。本发明解决了相关技术中用于工业机器人的控制策略无法满足工业机器人对高速度和高精度的要求的技术问题。

示例二：本发明公开了一种考虑系统延迟的不确定工业机器人运动控制方法，首先建立工业机器人机电耦合非线性动力学模型，再利用反馈线性化技术使工业机器人非线性动力学方程线性化，构建动态递归神经网络估计并补偿系统的不确定性，最后提出改进的*Smith*预测控制方法消除系统延迟的影响。本发明对于系统延迟和不确定性参数具有较好鲁棒性，极大地提高了工业机器人的控制精度。

示例一	示例二
工业机器人	系统延迟
仿真控制	工业机器人
三维建模	运动控制
仿真机械模型	非线性动力学
小脑模型神经网络CMAC控制策略	反馈线性化技术
比例积分微分PID控制策略	动态递归神经网络
	Smith预测控制方法
	系统延迟
	鲁棒性

表 2: 专利示例中的专业术语

通过进一步调查专利相关文件的撰写要求我们发现：根据《专利审查指南》在第二部分第二章规定，专利说明书应当写明发明或者实用新型所要解决的技术问题以及解决其技术问题采用的技术方案，并对照现有技术写明发明或者实用新型的有益效果*(中华人民共和国国家知识产权局，2010)*。也就是说一个专利的关键信息包括：技术问题、技术方案以及技术效果三个部分。

进一步地，我们可以将技术问题（简称“问题”）关键词定义为专利的技术所要解决的问题，将技术方法（简称“方法”）关键词定义为解决技术问题所采用的技术方案以及关键技术手段，将技术效果（简称“效果”）关键词定义为具有技术贡献的技术方案直接带来的、或者由所述的技术特征必然产生的效果*(张晓林，2018)*。

根据这个定义，上述两个示例专利的关键信息的关键如下表所示：

我们可以发现，标注了这三个方面的关键信息之后，我们发现可以比较准确地区分出这两个专利。虽然两篇专利均为工业机器人领域，但是在问题关键词上示例一和工业机器人模型有关，而示例二和工业机器人有关且示例二考虑到了系统延迟。在方法关键词上二者所采用的方法也不同，在效果关键词上示例一提升了精度而示例二具有较好的鲁棒性，两篇专利有着实质的区别。这对于理解和梳理这一领域的专利布局、挖掘专利覆盖的方向，都具有明显的助力。

3.2 中文专利关键信息标注原则

考虑到标注的成本，本文将专利关键信息的标注范围限定在专利的标题和摘要范围。一方面可以避免下载专利全文的负担，另一方面也节省了大量的标注时间。

进一步的，我们将以以下专利的标题和摘要为例，说明我们对于问题、方法和效果这三种信息标注的适用原则。

题目：一种考虑系统延迟的不确定工业机器人运动控制方法

(c) 2022 中国中文信息学会计算语言学专业委员会

示例	技术问题	技术方法	技术效果
示例一	工业机器人模型仿真控制方法及装置	工业机器人模型、仿真机械模型、小脑模型神经网络CMAC控制策略、比例积分微分PID控制策略	高速度、高精度
示例二	考虑系统延迟的不确定工业机器人运动控制方法	建立工业机器人机电耦合非线性动力学模型、反馈线性化技术、动态递归神经网络、改进的Smith预测控制	鲁棒性、提高、控制精度

表 3: 专利示例中的关键信息

摘要：本发明公开了一种考虑系统延迟的不确定工业机器人运动控制方法，首先建立工业机器人机电耦合非线性动力学模型，再利用反馈线性化技术使工业机器人非线性动力学方程线性化，构建动态递归神经网络估计并补偿系统的不确定性，最后提出改进的*Smith*预测控制方法消除系统延迟的影响。本发明对于系统延迟和不确定性参数具有较好鲁棒性，极大地提高了工业机器人的控制精度。

3.2.1 技术问题关键词

在上述示例技术问题关键词为：

考虑系统延迟的不确定工业机器人运动控制方法

该关键词说明了这篇专利要解决在考虑系统延迟情况下工业机器人的运动控制方法。而技术问题关键词在实际标注过程中还可以分为两个方面，即技术问题的主体和技术问题的预期效果，分别对应上述的工业机器人和运动控制方法。技术问题关键词一般均可以直接在题目中找到，但是在一些特殊情况下如外文译为中文的专利题目中可能找不到关键词，需要从专利摘要中寻找概括。

3.2.2 技术方法关键词

上述示例的技术方法关键词为：

建立工业机器人机电耦合非线性动力学模型、反馈线性化技术、动态递归神经网络、改进的*Smith*预测控制

该关键词说明了解决系统延迟的不确定工业机器人运动控制问题所采取的具体学科知识和主要步骤，而实际标注中我们也是将对技术方法关键词的标注分为学科知识和主要步骤两大类关键词。

3.2.3 技术效果关键词

上述实例的技术效果关键词为：

较好鲁棒性、提高、控制精度 该关键词说明了上述专利提出的考虑系统延迟的不确定工业机器人运动控制领域的技术方法所取得的效果。在实际标注中，我们发现技术效果一般存在于摘要结尾，直接提取即可。

此外，我们考虑到中文的语言特点，我们在标注过程中还遵循一下标注规则：

*1)*以顿号分隔关键词

为了统一标注格式，便于后期语料库的应用，我们规定若出现多个关键词，均以顿号分隔开，并且最后一个关键词后不加标点符号，如上述技术方法关键词的提取：建立工业机器人机电耦合非线性动力学模型、反馈线性化技术、动态递归神经网络、改进的*Smith*预测控制

*2)*技术问题关键词作为一个整体短语

由于技术问题一般出现在标题或者摘要中的第一句并且为复合短语，为了保证语义的完整性，我们规定提取整个的复合短语而不将其分隔开，讲将技术问题关键词最大化，如上述技术问题关键词的提取：考虑系统延迟的不确定工业机器人运动控制方法

*3)*技术方法、效果关键词提取语义片段

(c) 2022 中国中文信息学会计算语言学专业委员会

由于技术方法和技术效果一般是句子内部的短语，并且动宾语之间间隔较远，若最大化提取的话会造成关键词过于冗杂，所以我们规定在动宾语距离较远的情况下单独提取动词和宾语，仅提取关键的语义片段。如上述技术关键词所示，将“提高了工业机器人的控制精度”提取为“提高、控制精度”

4 专利关键信息语料库的标注实现

本文在研究初期接受了一项实际专利分析任务：对用户提供的机器人方向的专利集合进行标注。该集合共*313*篇专利文本，本文选取其中的题目和摘要作为语料标注范围。

4.1 人工标注过程

语料库构建的核心工作是依据制定的标注规范对语料进行标注*(*訾红英*, 2020)*。由于人工智能机器人领域尚处于发展阶段，且专业性较强，而业内缺乏统一的定义和标准，为了确定与领域更加适配的标注规范和标注策略，我们将标注过程分为预标注和正式标注两个阶段，在预标注阶段采用反复标注并讨论的策略制定初步的标注规范，在正式标注阶段使用了多轮迭代标注模式进行标注规范的更新以及标注工作，如图*1*所示。

预标注有助于减少重复劳动，节省人力和资源，提高效率，提升标注的速度与精度。在预标注阶段，我们以*50*篇专利为一周期，由二名标注规范制定者分别独立进行标注，全部完成后计算一致性，并对不一致的结果进行反复分析讨论，动态更新标注规范，得出一致的标注结果。之后按照新标注规范重复该周期，直至二人一致性达到*0.85*以上，确定最终的标注规范。

在正式标注阶段，标注工作由*5*名培训筛选后的标注人员（包括规范制定者）合作完成。为提高结果可信度，采用了多轮迭代标注的策略，即：

1）将机器人专利文本随机分成五组，由五名标注人员分别标注。

2）迭代式交换标注内容，进行第二轮标注。

3）计算两次标注的一致性，对不一致的结果进行讨论，进一步完善和细化标注规范，综合前两轮结果进行第三轮标注。

4）对第三轮标注结果进行抽样检查并计算一致性，若正确率达到或超过*0.84*则认为标注结果可信。

5）最后，对仍有分歧的标注进行讨论，由规范制定人员逐一校对，修正或删除不合理项，形成机器人方向专利语料库。

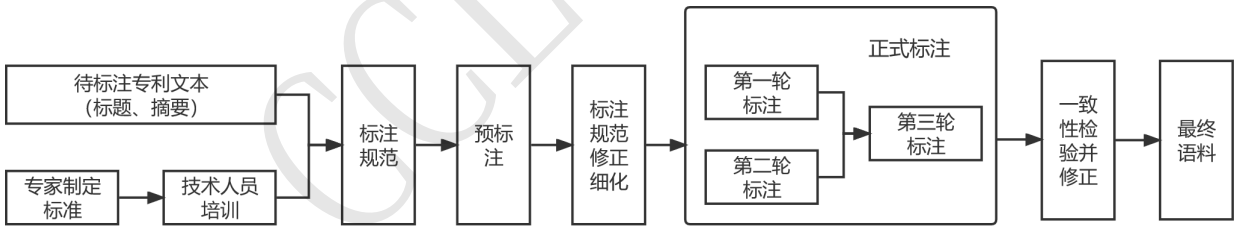


图1：语料库构建过程示意图

4.2 语料库标注质量

为了衡量语料库的标注质量，我们检验了标注结果的一致性。具体地，本文采取*Kappa*检验进行一致性检验，计算如式（*1*）所示。

$$Kappa = \frac{P_0 - P_c}{1 - P_c} \tag{1}$$

其中，P_0表示观察一致率，P_c 表示偶然一致率。我们选取五名标注人员，对全部文档进行独立标注，根据标注结果进行一致性测试。当标注术语、术语位置、术语顺序和关系类别均相同时，认为关系标注是一致的。

通过计算，本文语料标注的*Kappa*值为*0.88*，达到了用户预期的要求。进一步分析发现，由于标注人员主观认知差异，标注不一致的现象主要出现在：

(c) 2022 中国中文信息学会计算语言学专业委员会

1）专利问题范围的判断：不同的标注人员对同一专利所涉及的问题的范围判断不同。该类错误一般出现在标注早期，源于标注人员对标注规范不熟悉。

2）专业名词的理解：对于专利涉及的部分专业名词，不同的标注人员的理解不同，判断专利所用的技术方法时产生偏差。

专利涉及大量专业术语和专业知识，语料库的构建具有相当的挑战性。本文在标注语料库的同时也总结了一些经验：

1）进行相关名词资源的收集，其有助于界定实体边界，确定实体类型，对标注起到提示作用。

2）标注员的素质直接影响标注数据质量，因此在正式标注之前，对标注人员进行培训有利于提高准确率。

3）在正式标注过程中，对争议分歧应及时记录，定时组织讨论并听取相关专家的意见，以保证质量。

4.3 专利关键信息语料库规模统计

经过上述过程，本文最终构建了一个中文专利关键信息语料库，包含专利*313*篇，合计*9233*句、*529741*字。平均句长为*57.37*字，反映了专利文本的句子较长这一特点。

该语料库中共标注技术问题*366*个，技术方法*1384*个，技术效果*691*个。这批机器人方向专利中关注的问题主要集中在运动控制、可编程性、路径规划等；解决问题所用的技术方法主要包括深度学习、图像采集、坐标转换等；所达到的技术效果包括提高精度，提高效率，避免碰撞等。表*4*给出了三类关键词中出现频率最高的关键词及次数。

	高频关键信息(出现次数)
“技术问题”类	运动控制(47)、编程(33)、路径规划(32)、视觉(24)
“技术方法”类	深度学习(61)、图像采集(57)、坐标转换(38)、传感器(36)
“技术效果”类	提高精度(68)、提高效率(56)、加快速度(18)、避免碰撞(17)

表 4: 语料库中的各类高频关键信息

进一步地，具体到专利文本中，“技术问题”、“技术手段”、“技术效果”的关键词平均个数分别为*1.30*、*4.42*和*2.17*。进一步分析发现，“技术问题”“技术手段”“技术效果”关键词在标注文本中通常先后依次出现，第一次出现位置分别集中在标注文本的%*0.15*、%*0.2*、%*2.0*处。这一规律为我们日后进行自动化的专利阅读的深入研究提供帮助。

5 基于NER模型的专利关键信息的识别

分析上文实现的语料库的技术问题、技术方法、技术效果的标注结果，我们不难发现大量的技术专有名词（术语），一个很直接的想法就是使用成熟的命名实体识别技术对专利语料库关键信息进行自动识别。

具体地，本文用了*ACL 2021*和*ACL 2020*所发表的两个具有代表性的命名实体识别模型,分别为*Mect (Shuang Wu et al., 2021)*和*Flat Lattice Transformer (Xiaonan Li et al., 2020)*。其中，*MECT*模型将字特征、词特征和部首特征结合并能够使用多元数据特征，且在多个中文*NER*数据集上性能表现出色。*Flat Lattice Transformer*模型中所有字符都可以与其自匹配词直接交互，并可以对远距离依赖进行完全建模，且在多个*NER*数据集上被验证优于基线模型和其他基于词典的模型。同时，本文也提供了更为经典的基准*CRF*模型进行实验。

实验中，我们将*313*项专利按照*8*：*1*：*1*的比例，随机选择*251*项作为训练集，*31*项作为开发集，*31*项作为测试集。评测采用传统命名实体识别的召回率*(R)*、准确率*(P)*和F *1*值，各模型的性能指标如表*5*所示。

$$P = \frac{\text{预测正确的实体数}}{\text{预测的实体总数}} \tag{2}$$

$$P = \frac{\text{预测正确的实体数}}{\text{标注的实体总数}} \tag{3}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

$$R = \frac{2 * P * R}{P + R} \tag{4}$$

模型	F1	准确率	召回率
Mect	0.371	0.457	0.312
Flat Lattice Transformer	**0.474**	0.512	0.440
CRF	0.394	0.558	0.305

表 5: 语料库中的各类高频关键信息

从表中可以看出，*Flat Lattice Transformer*在所使用的三个模型中性能表现最好，但是*0.474*的*F1*值远远不能令人满意。以在*NER*中的*MSRA*数据集为例，*Flat Lattice Transformer*的*F1*值为*94.12*，远高于本节的实验结果。

进一步对比本文语料库和*NER*语料我们发现:

（*1*）*MSRA*数据集有超过*5* 万条中文命名实体识别标注，本文标注的语料库只有*2441*条标注，数据规模相对较小;

（*2*）本文语料中问题关键词平均词长*13.45*字，方法关键词平均词长*6.46*字，效果关键词平均词长*8.26*字，这些都远远超过了一般命名实体的长度。

实验结果反映出本文所标注的专利关键信息自动识别是一个有待深入探索的难题，仅仅沿用命名实体模型不能充分的满足专利的信息识别需求。同时由于当前标注语料规模远远小于*NER*中的可用资源，小样本学习问题将使专利关键信息的识别，更具有挑战性。

6 结论与展望

本文面向专利智能分析技术研发的需要，研制了一个专利关键信息语料库，完成了语料库的标注设计以及实际标注，初期完成的语料库中包含*313*篇专利，*9233*个句子以及*2441*条标注。

在此基础上，本文验证了当前主流的命名实体识别模型用于专利关键信息自动识别的效果。实验标明，相对于*90%*以上的命名实体识别效果，这些模型在本文的数据上*F1*值最好只能达到*0.474*，说明专利的关键信息识别是一个有待深入探索的问题。

目前专利文本的细粒度标注的研究还比较少，本文的专利来源，题材和覆盖面相对有限，只是这方面的一个初步尝试。下一步将扩大数据规模和丰富语料来源，进一步完善标注体系，为领域知识库的构建奠定基础，同时将探索在专利语料库上有效进行关键信息识别的模型和方法。

参考文献

国家知识产权局. *2008.* 专利. *www.cnipa.gov.cn/art/2008/4/3/art_2147_152059.html*

张晓林 *2018.* 专利技术情报分析模型构建及其应用研究. 图书馆杂志,第*10*期,*78-88*

Benjamin Balsmeier, Mohamad Assaf, Tyler Chesebro. 2018. Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *Journal of Economic & Management Strategy,*第*27*卷,*535-553*

李华锋，袁勤俭，陆佳莹等. *2017.* 国内外专利情报分析方法研究述评. 情报理论与实践,第六期,*139-144*

World Intellectual Property Organization. 2021. World Intellectual Property Indicators 2021.

Narin F. 1994. Patent bibliometrics *Scientometrics,*第*30*期,*147-155*

Mohammad Hamdaqa and Abdelwahab Hamou-Lhadj. 2009. Citation Analysis: An Approach for Facilitating the Understanding and the Analysis of Regulatory Compliance Documents. *Proceedings of Sixth International Conference on Information Technology: New Generations. Las Vegas, Nevada: 27-29*

(c) 2022 中国中文信息学会计算语言学专业委员会

*C. Sun, A. Shrivastava, S. Sing*等. *2017.* Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV),2017,pp: 843-852*

Roda, Giovanna , Tait, John , Piroi, Florina , Zenz, Veronika. 2009. CLEF-IP 2009: retrieval experiments in the intellectual property domain. *Cross-Language Evaluation Forum*

Lupu, M. , Huang, J. , Zhu, J. , Tait, J. 2009. TREC-CHEM: large scale chemical information retrieval evaluation at TREC. *Acm Sigir Forum ,43,pp:63-70*

Utiyama, Hi Masao. 2007. A Japanese-English patent parallel corpus. *Proc Mt Summit XI,2007*

George, Papadatos , Mark, Davies , Nathan, Dedman , Jon, Chambers , Anna, Gaulton , James, Siddle , Richard, Koks , Irvine, Sean A. , Joe, Pettersson , Nicko, Goncharoff 2016. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research,D1,2016,pp:D1220-D1228*

Tger, Wolfgang 2011. The Sentence-Aligned European Patent Corpus. *Proceedings of the 15th Annual conference of the European Association for Machine Translation,2011*

Korobkin, Dmitriy , Shabanov, Dmitriy , Fomenkov, Sergey , Golovanchikov, Alexander. 2019. Construction of a Matrix "Physical Effects – Technical Functions" on the Base of Patent Corpus Analysis.

Akhondi, S. A. , Klenner, A. G. , Tyrchan, C. , Manchala, A. K. , Boppana, K. , Lowe, D. , Zimmermann, M. , Jagarlapudi, Sarp , Sayle, R. , Kors, J. A. 2014. Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. *PLoS ONE,9,9,pp:e107477*

Lu, B. , Tsou, B. K. , Zhu, J. , Tao, J. , Kwong, O. Y. 2009. The Construction of a Chinese-English Patent Parallel Corpus.

翟东升, 李倩, 张杰, 黄鲁成, 赵京 *2013.* 德温特专利信息清洗与标注模型研究. 情报杂志,*32,8,pp:6*

高扬, 池雪花, 章成志, 孔捷 *2019.* 杰出人才精准画像构建研究——以智能制造领域为例. 图书馆论坛,*39,6,pp:8*

李智恒, 桂颖溢, 杨志豪, 林鸿飞, 王健. 基于生物医学文献的化学物质致病关系抽取. 计算机研究与发展, *55*, *1*, *pp:9*

崔博文, 金涛, 王建民. 自由文本电子病历信息抽取综述. 计算机应用,*2021*

唐晓波, 刘志源. 金融领域文本序列标注与实体关系联合抽取研究. 情报科学,*39,5,9,*

中华人民共和国国家知识产权局. *2010.* 专利审查指南. 北京：知识产权出版社, *2010:2-13*

朱宝华. *2019.* 浅谈如何撰写高质量专利申请文件. 中国发明与专利,*2019,16(03):95-100*

眷红英 *2020.* 面向儿科疾病的命名实体及实体关系标注语料库构建及应用. 中文信息学报*(05),19-26*

Shuang Wu et al. 2021. MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition.. *Proceedings of the Annual Meeting of the Association for Computational Linguistics,2021:1529-1539.*

Xiaonan Li et al. 2020. FLAT: Chinese NER Using Flat-Lattice Transformer.. *Proceedings of the Annual Meeting of the Association for Computational Linguistics,2020:6836-6842.*

(c) 2022 中国中文信息学会计算语言学专业委员会

句式结构树库的自动构建研究

谢晨晖[1,2,3]，**胡正升**[1,2,3]，**杨麟儿**[1,2,3*]，**廖田昕**[1,2,3]，**杨尔弘**[1,3]
[1]北京语言大学 国家语言资源监测与研究平面媒体中心
[2]北京语言大学 信息科学学院
[3]北京语言大学 语言资源高精尖创新中心
xch15673171321@163.com

摘要

句式结构树库是以句本位语法为理论基础构建的句法资源，对汉语教学以及句式结构自动句法分析等研究具有重要意义。目前已有的句式结构树库语料主要来源于教材领域，其他领域的标注数据较为缺乏，如何高效地扩充高质量的句法树库是值得研究的问题。人工标注句法树库费时费力，并且树库质量也难以保证，为此，本文尝试通过规则的方法，将宾州中文树库（CTB）转换为句式结构树库，从而扩大现有句式结构树库的规模。实验结果表明，本文提出的基于树库转换规则的方法是有效的。

关键词： 句式结构 ； 短语结构 ； 树库构建

Automatic Construction of Sentence Pattern Structure Treebank

Chenhui Xie[1,2,3], **Zhengsheng Hu**[1,2,3], **Liner Yang**[1,2,3], **Tianxin Liao**[1,2,3], **Erhong Yang**[1,3]
[1]National Language Resources Monitoring and Research Center Print Media Language Branch,
Beijing Language and Culture University
[2]School of Information Science, Beijing Language and Culture University
[3]Advanced Innovation Center for Language Resources,
Beijing Language and Culture University
xch15673171321@163.com

Abstract

Sentence pattern structure treebank is based on the theory of sentence-based grammar, which is of great significance to Chinese teaching and syntactic parsing. However, the content used in this corpus comes from Chinese as second language textbooks and Chinese textbooks, etc., while annotated data from other domains are scarce. A traditional way to alleviate this problem is to annotate sentences manually, but it is slow and laborious, and the quality of annotation is also difficult to control. In this paper, we propose a rule-based method to convert a phrase structure treebank named Penn Chinese Treebank (CTB) into a sentence pattern structure treebank so as to increase the size of the existing treebank. The experimental results show that our proposed rule-based method is effective.

Keywords: sentence pattern structure , phrase structure , treebank construction

* 通讯作者
基金项目：国家语委项目（ZDI135-131）；中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果（22YCX086）
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

树库是经过标注的深加工语料库，它记录着真实文本中每个句子的句法标注结果，提供分词、词性标注、句法结构等信息。目前广泛应用于自然语言处理领域的主流树库类型是短语结构树库和依存结构树库，句式结构树库的资源较少。

短语结构树库遵循短语结构语法，描写句子的短语结构，中文领域具有代表性的是宾州中文树库（Penn Chinese Treebank，以下简称CTB）(Xue et al., 2005)、清华汉语树库（Tsinghua Chinese Treebank，简称TCT）(周强, 2004)等；依存结构树库遵循依存语法理论，该理论由法国语言学家L.Tesnière于1959年提出，通过句子中词与词之间的依存关系来分析句法结构，认为任何两个词之间的依存关系中必有一个是中心词(Tesnière, 1959)，具有代表性的中文依存树库是哈尔滨工业大学汉语依存树库（HIT-IR-CDT，简称CDT）(He et al., 2009)。

值得注意的是，短语结构语法和依存结构语法均“以结构关系描写代替句子格局分析”，“句式结构在中文信息处理中一直处于一种模糊的地位”(彭炜明et al., 2014)，鉴于此，北京师范大学语言与文字资源研究中心构建了句式结构树库，弥补了这一不足。该树库以句本位语法体系为理论基础，着力研究各类句式的结构规律，总体特点为：1）句法上以句子成分分析法作为析句方法，并以“图解法”为语法分析工具；2）词法上采用“依句辨品”的词法观，“以句法控制词法”，这种语法本质上属于教学语法。上世纪六十年代，受到结构主义描写语言学的影响，汉语的句法分析开始以“直接成分”分析取代“句子成分”分析，典型的是朱德熙先生在《语法讲义》中建立的“词组本位”语法体系(朱德熙, 1982)，其以“短语”为本位作为汉语语法分析的基础，总体特点为：1）句法上以直接成分分析法（或称层次分析法）为析句方法，认为汉语句子构造原则与词组构造原则一致；2）词法上，以词的语法功能为划分词类的标准。直接成分分析法属于结构主义语言学的句法分析方法，短语结构语法实际上也是从直接成分分析法派生而来。整体而言，与短语结构树库标注句子层次结构、依存结构树库描写词与词之间关系相比，句式结构树库更能够呈现句子的整体结构，树结构更加扁平。

大规模树库多以自动句法分析为主要应用，但对于句式结构树库，“语言教学既是其理论之源，也是主要应用方向之一”(彭炜明, 2021)，如交互式出题、文本可读性评估等研究，句式结构自动分析器的研究还处于初步阶段。目前该树库语料约7万句，主要来源于国际汉语教材、中小学语文教材、文学作品等，其他领域语料较为缺乏，如何高效扩充高质量的句法树库是值得探究的问题。

大规模句法资源的构建是一项费时费力的工程，目前常用的有人工标注及树库转换两种方法。人工标注树库能够保证树库质量，但成本高，耗时长。第二种方法是比较行之有效的，即利用现有的树库资源，通过寻找两种形式语法之间的映射关系，转化成所需的目标树库。理论上来说，不同类型的树库在语法形式上尽管各不相同，但本质上都是对真实文本的句法结构的描写，这使得不同树库的转换具有可行性。

目前关于树库的自动转换研究主要集中在短语树库与依存树库之间的转换。Lin(1995)较早地提出了一种中心词节点表的方法，将短语树转换到依存树；Xia等人(2001)阐述了两种将短语树转成依存树的算法，采用中心词过滤表的方法将宾州树库（Penn Treebank，以下简称PTB）转换成依存树库，并提出一种新的算法，将产生的依存树转换成短语树，转换结果很接近原有的PTB。此外，Zabokrtsky 等人(2003)、Niu 等人(2009)、以及Kong 等人(2015)对短语结构与依存结构之间转换也做了研究与实践。在中文领域，树库转换研究较早的是党政法等人(2005)，其在Lin(1995)与Xia(2001)的研究基础上结合TCT的特点，进一步完善了转换算法，将TCT 转换成了依存结构，转换准确率达97.37%；李正华等人(2008)通过统计与规则相结合的方法，将CTB转换成哈工大依存树库体系结构；周惠巍等人(2010)在Xia(2001)提出的中心词过滤表方法及前人研究的基础上，结合CTB的特点，构造了完整的汉语中心词过滤表，将CTB转成了依存结构树库。

相比之下，短语结构树库、依存结构树库同句式结构树库之间的转换研究较少，其中张引兵等人(2018)通过总结TCT与句式结构树库标注体系的映射关系，制定了一套转换规则，将TCT转换成了句式结构树库，总体正确率为92.9%，证明了短语结构向句式结构转换的可行性。考虑到宾州中文树库（CTB）在自然语言处理等领域被广泛使用，本文通过规则的方法将短语结构树库CTB自动转换为了句式结构树库。本文通过比较宾州中文树库与句式结构树库在语法形式上的异同，制定了树库的自动转换规则，将短语结构树库CTB自动转换为了句式结构树库，实现了新闻领域句式结构树库的自动构建。为了验证树库转换规则的有效性，本文在人工标注的测试集上进行了三组实验，以比较基于句式结构自动句法分析的方法、基于短语结构自动句法分析结合转换规则的方法和基于宾州中文树库结合转换规则的方法三者的效果。这三种方法在测试集上的F_1值分别为84.43%、87.56%，89.72%，说明本文提出的基于转换规则的方法是有效的。[0]

[0]本文代码已公开在 GitHub 平台上，网址为：https://github.com/blcuicall/ctb2stb。

(c) 2022 中国中文信息学会计算语言学专业委员会

2 背景介绍

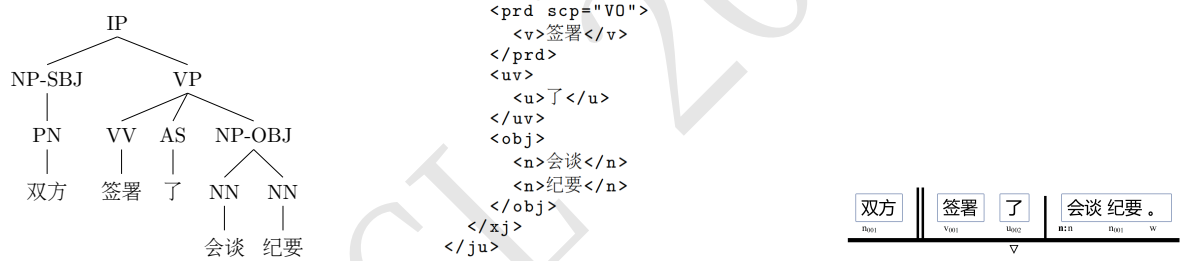
本文所使用的源树库为宾州中文树库，是短语结构树库的代表。制定向句式结构树库的转换规则需要比较两者在语法形式表现的异同，短语结构树库与句式结构树库的最本质区别在于两者所遵循的语法体系的不同。短语结构树库遵循短语结构语法，这是乔姆斯基为说明转换生成语法而讨论的一种语法模式(石定栩, 2002)，它以结构主义语言学的“直接成分分析法”为基础，以更接近数学公式的重写规则表示短语以及句子的结构。句式结构树库遵循句本位语法理论，使用“句子成分分析法”来分析句子、归纳句型句式，这种析句方法属于汉语传统语法。具体而言，宾州中文树库与句式结构树库在标注体系、句法分析单位、句法关系的描写等方面存在差异。

2.1 宾州中文树库

宾州中文树库是美国宾州大学自1998年起构建的短语结构树库，简称CTB，该树库以短语为句法分析单位，理论基础为语杠理论（X-bar theory）和支配及约束理论（government and binding theory）(Xue et al., 2005)，并作了一些简约化处理，不完全遵循二分法。CTB包含新华社、政府文件、新闻杂志、广播、访谈、网络新闻及网络日志等内容，对短语结构、短语功能进行了详细标注。

宾州中文树库通过括号的嵌套来存储层级结构。从图1a可以看出，除了词性节点以外，非叶子节点上的典型标记格式为“短语标签-功能标签”，树库中还存在“短语标记-多个功能标记”组合的情况。CTB注重短语外部功能的描写，例如“NP-SBJ”，“NP”表示该节点为名词性短语，“SBJ”表示它与其右侧兄弟节点VP之间为主谓关系，但其父节点却并没有相应的标记来指明这层关系，而是选择将其蕴含于单个功能标签当中，如主谓关系蕴含于“SBJ”中，述宾关系蕴含于“OBJ”中等。从直观上来看，部分节点上的功能标记可以直接对应汉语传统语法中的句子成分（比如主语和宾语）。

相比于句式结构树库，短语结构树库更侧重于描写短语结构，对表现整体的句式信息不可避免地有所缺失，例如“兼语句”、“连动句”等特殊句式无可利用的标记，只能根据句法结构层次进行提取。



(a) 短语结构树库　(b) 句式结构树库　(c) 句式结构图解

图 1: 短语结构树库和句式结构树库示例

2.2 句式结构树库

句式结构树库是北京师范大学语言与文字资源研究中心构建的，该树库遵循句本位语法理论，因此也称“句本位语法树库”。句本位语法由黎锦熙先生在1924年出版的《新著国语文法》中提出，该理论主要面向语法教学，以“句子成分分析法”为析句方法，主张“先理会综合的宏纲（句子），再从事于分析的细目（词类）”(黎锦熙, 2007)，并以图解法作为析句工具。句式结构树库语料主要来自具有一定影响力的国际汉语教材，也包括中小学语文教材，目前规模约7万句。相较于短语结构树库，句式结构树库以六大句法成分为基础构建句子结构，在词法层面，句式结构树库区分词库词与动态词，并对动态词内部词法结构进行详细标注。

句式结构树库以XML格式进行存储，本质上也是树结构。以图1b为例，树的根节点为<ju>，表示一句话的开始，当句子有多个小句时，首先将其拆分为各小句<xj>的子节点，然后对每个小句进行分析，该例只有一个小句。句本位语法的特征之一是采用“图解法”表示语法分析结果，如图1c，双竖线左侧为主语，单竖线右侧为宾语，横线上方为句子的主干结构。

句式结构树库注重对“句式”的描写，根据附加成分的有无、谓核的数量，将句子分为“基本句式”、“扩展句式”和“复杂句式”。基本句式为简单的主谓宾结构，如图1b；扩展句式在此基础上增加附加成分——定语、状语或补语。<xj>下的最外层以主<sbj>、谓<prd>、宾<obj>为句子

(c) 2022 中国中文信息学会计算语言学专业委员会

主干成分，以定<att>、状<adv>、补<cmp>为附加成分，同时为介词（p）、连词（c）、助词（u）、方位词（f）设置了“虚词位”，如图中的助词位<uv><u>了</u></uv>。以上两种句式都是“单谓核句”，只有一个谓核，如图1b的“签署”。复杂句式指多谓谓语句、主谓谓语句和复句，包含多个谓核，并根据谓核的关系定义细类，如连动句、兼语句等。在标记上，树库通过属性标记“scp”和“fun”保存句式信息，例如图1b中谓核“签署”的属性标记“scp”取值为“VO”，说明该谓核后接宾语。当句子中有多个谓核时，谓核之间需要通过属性标记“fun”表明其续接关系。

3 构建句式结构树库

我们采用宾州中文树库（CTB）作为源树库。将源树库转为目标树库总体上需要包含两部分：词层面的转换和句法层面的转换。但CTB与句式结构在词法层面的差别较大，包括分词粒度与词语结构标注的差异。具体来说，句式结构树库区分词库词与动态词，并标注动态词的内部结构，CTB对词法结构的标注却很模糊，基本上只显示词语边界而不含词语结构，如复合词呈现出扁平的树结构。因此，词法层面的转换需要实现分词粒度的统一与动态词的识别。考虑到转换工作难度与工作量，我们采取先句法转换后词法转换的策略：1）在句法上完全依照句式结构树库的规范；2）由于分词粒度的改变可能涉及到句法标注的改变，为了尽可能利用现有的CTB的句法标注信息，词法上暂时保留源树库CTB的分词与词性特征。下面详细陈述我们制定的自动转换规则及转换算法。

句式结构标记	含义	功能标记	短语标记	转换规则	举例
<sbj>	主语	SBJ	–	<sbj>【X-SBJ】</sbj>	【总理】说
		TPC		<sbj>【X-TPC】</sbj>	【中国】发展迅速
<obj>	宾语	OBJ	–	<obj>【X-OBJ】</obj>	抵达【罗马】
		IO		<obj>【XP-IO】</obj>	移交给【中共公安】
		PRD		<obj>【XP-PRD】</obj>	第二步是【集训】
<att>	定语	–	DNP	<att>【DNP】</att>	【两国的】关系
			QP	<att>【QP】</att>	【一个】食堂
			DP	<att>【DP】</att>	【这次】大海啸
<adv>	状语	ADV	DVP ADVP	<adv>【XP-ADV】</adv>	【跟去年】相比的话
		TMP		<adv>【XP-TMP】</adv>	【8月10日】宣称
		BNF		<adv>【XP-BNF】</adv>	【给灾民】送礼物
		DIR		<adv>【XP-DIR】</adv>	【往山上】跑
		MNR		<adv>【XP-MNR】</adv>	【以各种方式】参与
				<adv>【DVP】</adv>	【热切地】盼望
				<adv>【ADVP】</adv>	【立刻】跳下车

注：XP原指任何短语，此处代指CTB中的任何短语标记，下同

表 1: 核心转换规则

3.1 句法成分的转换规则

与句式结构树库以句子成分作为树的非叶子节点不同，CTB的句子成分信息分散在各类标记中，部分句式信息则通过括号的层级关系体现，因此需要分情况制定互补的转换规则。对于词性标记以外的非叶子节点的转换从两个方面进行，一为句法成分信息，二是句式信息。在句法成分的提取上，利用CTB的功能标签集和短语标签集作为核心转换方法，利用CTB的词性标签集作为辅助转换方法，以补充核心转换方法的不足；在句式信息的提取上，主要利用各类句式的特殊标记或相对固定的树层级结构。从宾州中文树库向句式结构树库转换的总体思路是：将复句切分为小句，对CTB的短语树进行先序遍历，对于词性标记以外的非叶子节点，提取标签，对应转换规则进行转换；对于叶子节点（包括标点符号），根据词类对应关系直接转换，即将“（词性 词）”的括号形式转换为<词性>词</词性>的XML形式。我们根据所利用的标签类型区分为核心转换规则和辅助转换规则。

3.1.1 核心转换规则

利用CTB功能标签和短语标签进行转换的为核心转换规则。在CTB的功能标签集里，有一部分

(c) 2022 中国中文信息学会计算语言学专业委员会

如主语（SBJ）、宾语（OBJ）、状语（ADV）标签，可以与句法成分直接对应，这类功能标签可以直接转换为相应的句法成分；在短语标签集里，也有一部分指明短语的句法性质和内部的结构关系，例如DVP（“地”字短语），如果DVP后跟有兄弟节点VP（动词短语），则DVP为状语。类似的还有ADVP（副词短语）作状语，PP（介词短语）作状语或补语等，具体规则详见表1。

3.1.2 辅助转换规则

辅助转换规则指利用CTB的词性标签，将标签下的叶子节点转换为句本位语法中的句法成分，具体规则见表2。

标记	含义	词性标记	转换规则	举例
<prd>	谓语	VC	<prd><v>【VC】</v></prd>	九江【是】江西的北大门
		VE	<prd><v>【VE】</v></prd>	技术出口也【有】了进展
		VV	<prd><v>【VV】</v></prd>	中美【签订】合作协议
<adv>	状语	LB	<adv>【LB】【XP-SBJ】</adv>	【被外部世界】广泛关注
<cmp>	补语	DER	<cmp>【右兄弟节点】</cmp>	各项工作做得【更好】
<pp>	介词位	P	<pp>【P】<pp>	【在】讲话中
		BA	<pp>【BA】<pp>	【把】注意力转向其他市场
		LB	<pp>【LB】<pp>	【被】警方抓个正着
<ff>	方位词位	LC	<ff>【LC】<ff>	大地震【后】
<un>	助词位	ETC	<un><u>【ETC】</u></un>	各种税率【等】优惠
<uu>	助词位	DEC	<uu><u>【DEC】</u></uu>	重要【的】意义
		DEG	<uu><u>【DEG】</u></uu>	两国【的】关系也十分友好
		DER	<uu><u>【DER】</u></uu>	里头学问大【得】很
		DEV	<uu><u>【DEV】</u></uu>	专心【地】工作
<uv>	助词位	AS	<uv><u>【AS】</u></uv>	我参加【了】救援
		SB	<uv><u>【SB】</u></uv>	【被】淘汰
		MSP	<uv><u>【MSP】</u></uv>	两岸【所】做出的贡献
		SP	<uv><u>【SP】</u></uv>	我们的心是连接在一起【的】
<cc>	连词位	CC	<cc><c>【CC】</c></cc>	店铺【和】民房
		CS	<cc><c>【CS】</c></cc>	【如果】我们那样做

表 2: 辅助转换规则

句式结构树库为介、连、助、方位词设置虚词位：介词位、连词位、助词位、方位词位。虚词不直接充当句子成分，但在句式结构树库中的占位与主语等句子成分一致，因此我们将它纳入句子成分的规则集中。由于虚词位只需要获得词性信息，所以只要根据词性的对应关系即可进行简单转换。

3.2 句式的转换规则

句式结构树库所采用的“句式”术语为吕叔湘先生的定义，即“句子的结构格式和结构类型”“特定句式中成分、词类或特征词序列具有相对固定的结构层次和位置顺序”（彭炜明, 2021）。短语结构树表明短语与短语之间如何组成句子，对于整个句子属于哪种句式没有说明，这是树库标注体系所反映的语法体系的区别、树库构建者的标注理念的区别所导致的。但每种句式都“有章可循”，根据各类句式的标志性特征，例如特定的词类、词序列可以判断是否为相应的句式。具体转换规则见表3。

合成谓语句 合成谓语句是1956年《暂拟汉语教学语法系统》（简称《暂拟系统》）提出的一类句式，句式结构树库做了一些修改，将合成谓语分成“助动词+VP”或“系动词+VP”两类，在句式结构树库中统一标为“<cc fun=”SYN”/>”。对于“系动词+VP”这一类，可以利用CTB的词性标签“VC”（系动词），助动词则都标为“VV”，利用词性信息的方法不可行。因此我们根据固定的短语树的层级结构转换，具体为：

(VP(VV)(VP(VV)))

(c) 2022 中国中文信息学会计算语言学专业委员会

如果符合该短语树结构，那么两个谓核VV间为合成谓语关系。

兼语句 CTB对兼语结构的描写同样有固定的结构层次，因此总结出“节点标记+树结构层次”的特定层级框架即可提取出句式信息。CTB将动词分成三类：VE(有)、VV(普通动词)、VC(系动词)，兼语句的提取需制定两条规则：

① (VP(VE)(XP-OBJ(XP-SBJ)(VP)))
如果短语树符合这一结构，则XP-SBJ后增加兼语结构标记<cc fun=”PVT”/>;
② (VP(VV)(NP-OBJ)(IP)))
如果短语树符合这一结构，则NP-OBJ后增加兼语结构标记<cc fun=”PVT”/>;

连动句 在CTB中，连动结构以如下树结构体现：
(VP(VP)(VP))
当VP下有两个VP子节点时，在VP的左子树【LP】和右子树【RP】之间增加属性标记<cc fun=”SER”/>。

句式标记	含义		转换规则
APP	同位		【XP-APP】<cc fun=”APP”/>【兄弟节点】
COO	并列		【NP_1】<cc fun=”COO”/><c>连词</c></cc>【NP_2】
SYN	合成谓语	系动词 + VP	<prd><v>【VC】</v></prd><cc fun=”SYN”/>【VP】
		助动词 + VP	*
SER	连动		*
UNI	联合谓语		【VP_1】<cc fun=”UNI”/><c>连词</c></cc>【VP_2】
PVT	兼语		*

注：标 * 号的规则见 3.2 节具体说明

表 3: 句式转换规则

3.3 特殊情况的转换

从短语结构树库转为句式结构树库涉及到两者标注体系的不同，但“它们主要描述的都是句法结构，在更深层次上具有一致性。”(李正华et al., 2008)。因此，对于只涉及两者标注体系差异的，根据标记的对应关系即可直接转换，对于涉及到标注体系所遵循的语法理论差异的，需要进行特殊处理。例如CTB中区分了两种“被”，词性标签分别为LB和SB，LB用于“NP_0+LB+NP_1+VP”结构中，引出施事；SB用于“NP_0+SB+VP”结构中，直接附于动词前。在句本位语法中，SB处理成助词，LB与其引出的施事作为句子的状语成分，对于前者，本文根据词性的对应关系可以简单处理，后者作为复杂情况进行相应的处理。另外，CTB非叶子节点上的标签组合关系分两种情况，第一种是常见的“短语标记+功能标记”，例如“NP-SBJ”，另一种是两个及以上的标签进行组合，例如“IP-IMP-TTL-PRD”，第二类情况也需当作特殊情况处理。

3.3.1 限制性转换方法

在实际转换过程中，CTB中的某些标签并不能完全对应句本位语法中的句法成分，需要增加限制条件。例如“PP（介词短语）”在汉语中既可作状语，也可以充当补语成分，当介词短语后接动词性短语时作状语，当附加于动词后面时作补语。此类情况较多，部分举例见表4。

CTB标记	条件	对应句法成分	句式结构标记
QP-EXT	左有VV/VC/VE兄弟节点	宾语	<obj>
	左无VV/VC/VE兄弟节点	补语	<cmp>
PP	右有VP兄弟节点	状语	<adv>
	左有VV兄弟节点	补语	<cmp>

表 4: 限制性转换规则

(c) 2022 中国中文信息学会计算语言学专业委员会

3.3.2 多标签组合的转换方法

多标签组合指三个及以上的标签以“-”连接为一个标签。上文中列举的转换规则所依据的标签，均为典型的“短语标记”、“短语标记-功能标记”或“词性”的组合形式。然而CTB的标签种类丰富，每个标签集的标签数量众多，尤其注重对短语功能的描写，因此还存在部分以“短语标记-多个功能标记”为组合形式的标签。

总体来说，对应到句式结构树库的有效的标记一般居于组合尾部，尽管多标签组合类型多样，但出现频率低，因此针对多标签组合的转换流程可以简单处理为一个条件循环，而不会产生过多的不可预期的问题，具体如下：判断组合中最后一个标记是否符合核心转换方法或辅助转换方法，如果不符合，则判断前一个标记，直到符合转换条件为止。以“NP-OBJ-SBJ-PN”为例，该标签中“PN”不符合两种转换方法，因此判断前一个标记“SBJ”，其符合核心转换方法中的“主语”转换规则，最后将“NP-OBJ-SBJ-PN”下的子树转换成“主语”。

3.4 词性转换方法

从词性标记的数量来看，句式结构树库的词性标记有15个，CTB词性标记33个，类别更多。我们通过句式结构树库和CTB的词性的映射关系直接转换，转换规则如表所示。采用这种对应转换的方法，使得词性的粒度变粗了，但并不会丢失词性的大类信息。词性对应关系详见表5。

句式结构树库		宾州中文树库		句式结构树库		宾州中文树库	
标记	词性	标记	词性	标记	词性	标记	词性
n	名词	NN	普名	d	副词	AD	副词
		NR	专名	p	介词	P	介词
		FW	外来词			BA	“把”,“将”
		URL	网页链接			LB	“被”,“给”
		NN-SHORT	略缩普名				
		NR-SHORT	略缩专名	c	连词	CC	并列连词
t	时间词	NT	时名			CS	从属连词
		NT-SHORT	略缩时名	u	助词	DEC	标句词“的”
r	代词	PN	代词			DEG	所有格“的”
		DT	限定词			AS	体标记
f	方位词	LC	方位词			DER	得
m	数词	CD	基数词			DEV	地
		OD	序数词			ETC	等
q	量词	M	量词			MSP	“所”,“以”等
						SB	“被”
						SP	句尾小品词
v	动词	VV	其他动词	e	叹词	IJ	感叹词
		VC	系动词				
		VE	“有”	o	拟声词	ON	拟声词
a	形容词	VA	表语形容词	w	标点	PU	标点
		JJ	区别词/紧缩形容词				

表 5: 词性转换规则

3.5 树库转换算法

根据前文所述的转换规则，并结合CTB的语法结构特征，我们设计了短语结构树库向句式结构树库转换的算法，详细流程见算法1。

算法首先对输入的短语结构字符换进行多叉树的数据结构构建，若句子中有存在小句（clause）首先利用小句的转换规则（clauseRules）进行小句切分。按照先序遍历的方式，对每个节点根据转换

(c) 2022 中国中文信息学会计算语言学专业委员会

算法 1 宾州中文树库向句式结构树库的转换算法

输入: 短语结构树 $psTree$

输出: 句式结构树 $ssTree$

$nodeSeq \leftarrow preOrder(psTree)$ ▷ $nodeSeq$为先序遍历得到的序列
for $i = 0 \rightarrow n$ **do**
    $node \leftarrow nodeSeq[i]$
    if $node \in Clause$ **then** ▷ 如果句子中存在小句，则执行小句切分规则
        $clauseRules(nodeSeq[i])$
    end if
    if $node \in Syntax$ **then** ▷ 判断是否符合句法成分转换规则
        $syntaxRules(node)$
    else if $node \in Struct$ **then** ▷ 判断是否符合句式转换规则
        $structRules(node)$
    else
        $specRules(node)$ ▷ 执行特殊转换规则
    end if
    $posRules(node)$ ▷ 执行词性转换规则
    $ssTree \leftarrow node$
end for

规则进行转换，分别包括句法成分的转换方法（syntaxRules）、句式的转换方法（structRules）和特殊情况的转换方法（specRules）以及词性的转换方法（posRules）。

4 实验

我们用前文所述的规则完成了CTB5共18,244句的句式结构转换，并人工标注了CTB部分句子作为测试集对转换结果进行评估。为了进一步验证树库转换方法的有效性，我们另外设置了两组实验：基于句式结构自动分析器生成树库的实验、短语自动句法分析器结合转换规则生成树库的实验。

4.1 数据集与评价指标

为了使评测结果更具参考性，我们采用Liu和Zhang(2017)对CTB5的数据切分方式，将文件编号271至300部分作为测试集，并去除了测试集中的30句新闻电头，因为这类句子在CTB中的结构扁平，无可利用的标记，超出了规则处理的范围，最终人工标注共318句。我们对测试集进行了句式结构标注，以评估三种方法的性能。在标注过程中，分词仍然按照源树库CTB的标注规范，词性和句法层面依照目标树库句式结构树库的规范。

4.2 基于人工标注的自动转换结果评估

我们通过句式结构树库标签的分类来看三种方法的具体表现，分别是句子、成分、虚词位与句式标签，具体数据见表6。

通过数据我们发现:

（1）树库整体转换结果较好，转换规则对于小句切分、句子主干成分、附加成分的处理不错。小句切分准确是后续句法分析的前提，句子成分是句子分析的最重要的一步，数据说明最后构建的新树库质量较好。

（2）虚词位的转换F1值最高。虚词为封闭集，在转换规则中，两个树库的词性标签为直接映射关系，转换难度较低。附NP的助词位根据CTB的词性标记“ETC”转换，但标句词短语CP中的“的”有些是结构助词，有些是附NP的助词位，由于助词位对实际句法分析与句子理解的影响较小，并考虑到规则复杂性，我们的处理办法是根据对应数量的占比映射到句式结构树库的词类体系，这里处理为结构助词，这也导致了附NP的助词位召回率下降。句式结构树库与CTB的连词集合并不能完全对应，部分连词在CTB中标为副词，如作为关联词的“但”，词性标注的不一致导致连词的召回率下降。

(c) 2022 中国中文信息学会计算语言学专业委员会

大类	小类	P	R	F1	大类	小类	P	R	F1
句子	小句	93.81	93.37	93.59	虚词位	助词位（定状补）	93.43	100.00	96.60
成分	主语	91.67	92.17	91.92		助词位（附NP）	100.00	66.67	80.00
	谓语	92.30	91.33	91.81		助词位（附VP）	94.07	87.40	90.61
	宾语	91.75	91.40	91.57	句式	并列	96.89	63.39	76.64
	定语	74.84	72.18	73.49		同位	69.23	79.41	73.97
	状语	93.11	93.23	93.17		合成谓语	95.29	77.14	85.26
	补语	85.71	85.71	85.71		联合谓语	72.46	72.46	72.46
虚词位	介词位	95.87	99.69	97.74		兼语	97.22	92.11	94.59
	连词位	93.10	75.00	83.08		连动	90.00	90.00	90.00
	方位词位	99.08	99.08	99.08					
整体精确率					90.68				
整体召回率					88.79				
整体F1					89.72				

表 6: 基于人工标注的自动转换结果评价

（3）兼语结构、连动结构在句式的转换结果中效果最好。两者是根据特定的CTB的短语树结构进行转换的，说明树结构能基本对应相应的句式结构，经过规则的转换结果错误率低。

（4）合成谓语结构精确率高，但召回率稍低。经过观察，合成谓语结构的主要问题在于“是字句”，CTB对于该结构有两种表现形式：((VC)(VP))、((VC)(NP-PRD))，对于前一种形式可以直接对应合成谓语结构，但后一种大部分为动宾结构，少部分为合成谓语结构，规则根据多数情况处理为动宾结构，这可能是导致召回率稍低的原因。

（5）并列和联合成分有时并不以连词连接，顿号以及逗号也能起到连接的作用，如“华侨、华人艺术家”“一个有利可图，有钱可赚的投资环境”，在我们制定的转换规则中，主要利用的是非叶子节点的信息，处理这类并列结构和联合谓语结构需要利用到叶子节点信息，因此规则未完全覆盖此类情况。另外，联合谓语结构“有的通过关联词语（连词或关系副词）突显，有的则依赖VP自身的语义逻辑”(彭炜明, 2021)，表3所示的联合谓语结构转换规则暂时只考虑了以连词连接的情况，对于以关系副词和VP间的语义逻辑突显的联合谓语结构，如“深刻却又舒缓”、“污染严重治理无望”等一般处理为连动结构，这可能是联合谓语结构转换效果不佳的另一原因。

（6）CTB与句式结构树库对同位短语的定义不同，前者定义更宽泛，如“同等优先、适当放宽的原则”，CTB也分析为同位结构，因此召回了一些本该是定语的错误样本，定语转换问题也在于此。

为了进一步考察本文提出的树库转换规则的效果，我们进行了两组实验以作比较。

4.3 对比试验

4.3.1 实验设置

实验一：通过训练自动句法分析器自动生成树库是扩建树库的一种通用的方式，但是这种方式对自动句法分析器性能要求较高，具有一定挑战。目前句式结构自动句法分析器处在初步研究阶段，本文借鉴Kitaev(2019)的基于自注意力机制的神经网络模型训练了句式结构自动句法分析器。

模型采用编码器、解码器架构，将预训练模型Bert用于编码器阶段，将词性、位置作为辅助信息传入模型，编码器对词表征$[w_1, ..., w_n]$、词性表征$[m_1, ..., m_n]$及位置表征$[p_1, ..., p_n]$加和获得词嵌入，随后使用多头注意力机制对词嵌入进行编码，解码器采用CKY(Cocke, 1969; Younger, 1967; Kasami, 1966)算法获得句式结构句法树。数据集来源于北京师范大学构建的句式结构树库，借鉴Liu等人(2017)的切分方式，对数据集采用50:1方式构建训练集和开发集，如表7所示。该实验主要探索句式结构自动句法分析器在树库构建方面的效果。

实验二：将自动句法分析与转换算法相结合。首先通过短语结构自动分析器产生短语结构，然后

(c) 2022 中国中文信息学会计算语言学专业委员会

利用短语结构树库向句式结构树库的自动转换算法产生句式结构，这种方式不仅可以扩充更大规模的句式结构树库，而且相较于实验一，不依赖于句式结构树库作为训练数据，具有更强的适用性。

短语结构自动句法分析技术目前较为成熟，常用的短语句法分析器有伯克利句法分析器[1]、CoreNLP[2]，但是这些模型输出的短语结构仅有短语标签，如“NP”、“VP”等，并无功能标签，如“SBJ”、“OBJ”等，如前文所述，本文提出的转换规则需要利用到CTB的多种标签，特别是功能标签，因此上述句法分析器对于本文提出的自动转换算法并不适用。鉴于此，本文借鉴 Kitaev等人(2018)(2019)提出的基于自注意力机制的神经网络方法，训练得到可以分析功能标签的短语句法分析器。采用 Liu等人(2017)的数据切分方式对CTB5数据集进行切分，具体数据统计如表7所示。

数据集	训练集	开发集	测试集
句式结构树库(张引兵et al., 2018)	67,558	1,352	-
宾州中文树库（CTB5）	17,544	352	318

表 7: 句式结构树库和宾州中文树库（CTB5)数据统计

4.3.2 实验结果与分析

表8列出了基于句式结构自动句法分析的方法、基于短语结构自动分析结合转换规则的方法和基于宾州中文树库结合转换规则的方法的总体性能和分别在小句、成分、虚词位和句式等方面的性能。

方法	F_1值				准确率	召回率	F_1值
	小句	成分	虚词位	句式			
句式结构自动句法分析	95.21	80.61	88.74	**87.44**	83.85	85.01	84.43
短语结构自动句法分析 + 转换规则	95.09	86.01	94.33	74.03	88.40	86.73	87.56
宾州中文树库 + 转换规则	**95.70**	**88.43**	**95.72**	78.87	**90.68**	**88.79**	**89.72**

表 8: 三种方法对比实验结果

从表8可得以下结论：本文提出的基于宾州中文树库结合转换规则的方法整体效果最优，比基于句式结构自动句法分析的方法[3] F_1值高5.29%，说明基于规则的转换算法在树库自动构建上具有一定优势。在具体标签类别上，转换规则在小句切分、句子成分、虚词位上的效果均优于另外两种方法，在句式上低于句式结构句法分析器，通过4.2节的分析，原因在于规则对同位、并列、联合谓语结构的处理存在不足，这是需要继续完善的部分。

短语结构自动句法分析结合转换规则的方法，相较于句式结构自动句法分析的方法，准确率高4.55%，F_1值高3.13%。在成分和虚词位上的F_1值分别高5.40%、5.59%，进一步说明转换规则的有效性。该方法不依赖于人工标注的宾州中文树库作为源树库来构建句式结构树库，具有更强的通用性，在构建大规模句式结构树库具有一定优势。

综上所述，本文提出的基于宾州中文树库结合转换规则的转换方法在句式结构树库构建上具有一定优势，基于短语结构自动句法分析结合转换规则的方法由于不依赖现有人工标注的宾州中文树库，因此在够建大规模的句式结构树库上具有更强的通用性。

5 结语

本文以宾州中文树库为源树库，通过基于规则的方法实现了向句式结构树库的自动转换，以此构建了大规模的新闻领域句式结构树库，并基于人工标注的评估，验证了该方法的有效性。此外，本文设置了对比实验，以比较句式结构自动句法分析、短语结构自动句法分析结合转换规则、基于转换规则等三种方法的性能，实验表明，基于转换规则的转换方法优于其他两种方法，进一步验证了转换规则的有效性。目前我们的转换规则仍然存在一些不足，如合成谓语、并列、联合等转换规则还有待完善。新树库仍然保留源树库CTB的分词，未来我们将继续完善句法层面的转换规则，并探索词法层面的转换，以提高新句式结构树库的质量，为自动句法分析等相关研究提供有效的数据支持。

[1]https://parser.kitaev.io/

[2]https://corenlp.run/

[3]此方法性能较低部分原因是宾州中文树库和现有句式结构树库的分词标准不一致。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

John Cocke. 1969. *Programming languages and their compilers: Preliminary notes*. New York University.

Wei He, Haifeng Wang, Yuqing Guo, and Ting Liu. 2009. Dependency based chinese sentence realization. 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*.

Tadao Kasami. 1966. An efficient recognition and syntax-analysis algorithm for context-free languages. *Coordinated Science Laboratory Report no. R-257*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Lingpeng Kong, Alexander M. Rush, and Noah A. Smith. 2015. Transforming dependencies into phrase structures. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Dekang Lin. 1995. A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of IJCAI*.

Jiangming Liu and Yue Zhang. 2017. Shift-reduce constituent parsing with neural lookahead features. *Transactions of the Association for Computational Linguistics*, 5:45–58.

Zheng-Yu Niu, Haifeng Wang, and Hua Wu. 2009. Exploiting heterogeneous treebanks for parsing. 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*.

Lucien Tesnière. 1959. *Eléments de Syntaxe Structurale*. Klincksieck, Paris.

Fei Xia and Martha Palmer. 2001. Converting dependency structures to phrase structures. In *Proceedings of the First International Conference on Human Language Technology Research*.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Daniel H Younger. 1967. Recognition and parsing of context-free languages in time n^3. *Information and control*, 10(2):189–208.

Zdenek Zabokrtsky and Otakar Smrz. 2003. Arabic syntactic trees: from constituency to dependency. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

党政法and 周强. 2005. 短语树到依存树的自动转换研究. 中文信息学报, 19(3):21–27.

周强. 2004. 汉语句法树库标注体系. 中文信息学报, 18(4):2–9.

周惠巍, 黄德根, 钱志强, and 杨元生. 2010. 短语结构到依存结构树库转换研究. 大连理工大学学报, 50(4):609–613.

张引兵, 宋继华, 彭炜明, 赵亚伟, and 宋天宝. 2018. 短语结构树库向句式结构树库的自动转换研究. 中文信息学报, 32(5):31–41.

彭炜明, 宋继华, and 王宁. 2014. 基于句式结构的汉语图解析句法设计. 计算机工程与应用, 50(6):11–18.

彭炜明. 2021. 句本位语法的中文信息处理理论与实践. 外语教学与研究出版社, 北京.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

李正华, 车万翔, and 刘挺. 2008. 短语结构树库向依存结构树库转化研究. 中文信息学报, 22(6):14–19.

石定栩. 2002. 乔姆斯基的形式句法：历史进程与最新理论. 北京语言大学出版社, 北京.

黎锦熙. 2007. 新著国语文法. 湖南教育出版社, 长沙.

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

面向情感分析的汉语构式语料库构建与应用研究
—对汉语构式情感分析问题的思考

吴尹清
国防科技大学国际关系学院/江苏南京
wu.yinqing@outlook.com

李德俊
国防科技大学国际关系学院/江苏南京
njlide@sina.cn

摘要

文本情感分析又称为意见挖掘，是基于网络大数据对评价主体倾向性的研究。由于其在舆情监控、市场营销、金融等应用领域的特殊意义，近年来受到了越来越广泛的关注。本文关注情感分析面临的语义隐匿性问题，通过构建一个汉语构式语料库，对语料库中的汉语构式进行量化统计，讨论汉语构式与情感分析之间的关系。文章对语料库中表达量级和态度义的构式与词汇进行了标注，并基于该语料库对相关构式和词汇进行了计量分析，按照构式类型、语义类别、常项变项个数等标准统计了语料库中量级和态度义构式的信息，并与量级和态度义词汇的统计信息进行了比对，通过分析构式表义比重和词汇表义比重这两个指标，发现语料库中词汇承载了大部分态度和量级语义信息，构式所承载的态度和量级语义信息较少。虽然构式不是主要的表义单位，但其承载的态度语义信息仍占一定比例。文章为构式语法应用于汉语情感分析提供了实证数据，为后续该类研究提供了一种方法，也为汉语构式研究提供了基于汉语真实文本的数据。文章还专门探讨了目前构式语法应用于汉语情感分析乃至自然语言处理所面临的困难，对后续研究提出了展望。

关键词： 构式语法；汉语情感分析；构式语料库；态度义；量级义

A Study of Chinese Construction Corpus Compilation and Application for Sentiment Analysis: A Discussion of Sentiment Analysis Problems of Chinese Constructions

Wu Yinqing
College of International Studies,
National University of Defense
Technology / Nanjing, Jiangsu
wu.yinqing@outlook.com

Li Dejun
College of International Studies,
National University of Defense
Technology / Nanjing, Jiangsu
njlide@sina.cn

Abstract

Sentiment analysis, also called opinion mining, is a field that studies the sentiment orientation of the evaluation subjects based on Web data. Due to its significance in such fields as public opinion monitoring, marketing and finance, sentiment analysis has received increasing attention in recent years. The present study focuses on the problem of latent meaning faced by sentiment analysis, builds a Chinese construction corpus and studies the relations between Chinese constructions and sentiment analysis by quantifying the Chinese constructions within the corpus. We annotate the constructions and words expressing attitudinal and gradational meaning in the corpus. A qualitative

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
作者简介：吴尹清(第一作者/通讯作者)，博士生。李德俊，博士，教授，博士生导师。

analysis is conducted on these constructions and words. Statistics of attitudinal and gradational constructions are computed according to such standards as construction type, semantic type, constant number and variable number. By comparing the statistics of the constructions and words in the corpus, we find that most of the gradational and attitudinal semantic information is loaded by words rather constructions. In spite of this, there still is a certain portion of attitudinal and gradational information expressed by constructions. The present study provides empirical data and a research method to the study of applying construction grammar to Chinese sentiment analysis. Chinese construction studies are also be supported by these data computed from Chinese authentic texts. The present study also discusses the difficulities of applying construction grammar to Chinese sentiment analysis and natural language processing and looks forward to future studies.

Keywords: Construction grammar , Chinese sentiment analysis , Construction corpus , Attitudinal meaning , Gradational meaning

1 研究必要性：构式与情感分析的语义隐匿性问题

情感分析作为重要的自然语言处理任务，其目标是“从文本中分析出人们对于实体及其属性所表达的观点、情感、评价、态度和情绪”(刘兵, 2017: 1)。近年来互联网评价文本的爆炸性增长催生了对海量文本的情感语义进行自动化分析的强烈需求，使得情感分析成为了学术界的研究热点，在舆情监测、企业与政府决策、社会计算等方面均有应用。

汉语情感分析面对的语义隐匿性问题是由语义隐匿性现象造成的，即汉语中部分情感意义表达机制不确定性较强的语义现象，如反讽、非现实性、构式等。一般的情感分析方法难以应对这些语义现象的复杂性，不能准确解析它们的意义，导致它们的意义对于情感分析系统具有某种“隐匿性”。由于汉语的复杂性，语义隐匿性现象较为普遍，对语义隐匿性问题进行研究对于未来汉语情感分析系统性能的提升将有较大意义。

态度和量级义构式是上述语义隐匿性现象的重要一部分。“（典型的）构式是无递归性的非平凡的短语结构”(詹卫东, 2017: 232)，表达态度和量级意义的构式称为量级和态度义构式。态度意义是情感倾向，也即情感极性，包括正面、负面、中性等，量级意义描述的是情感极性的强度，包括低量、高量、极量等，也可使用数值来描述。态度和量级意义是情感意义计算的最重要的两个语义变量。由于构式的形式与意义的可推导性较弱(Goldberg, 1995:4)，其情感意义和形式难以通过短语结构语法等一般的组合性规则推导得到，意义表达机制的不确定性较大，自动解析难度较大，而汉语中的许多构式又具有明显的情感意义(詹卫东等, 2020)。而且汉语情感分析主要关注词汇的情感色彩，对构式这一特殊语言单位承载的语义信息缺少关注。因此态度和量级义构式的语义解析问题就属于汉语情感分析的语义隐匿性问题。

当前情感分析研究对语义隐匿性现象的研究不足，尚不能很好地应对语义隐匿性问题，满足于将词汇视为情感意义的承载单位，在理论构建、实证研究、资源建设等方面都以词汇为重点，未能参考借鉴当前构式语言学的理论与实证成果，对汉语中具有语义隐匿性的边缘性结构缺乏关注。因此情感分析研究可以考虑借鉴构式研究的成果，对构式这一层级的语言单位给予一定重视。文章基于真实评价语料，通过考察汉语态度和量级构式的分布情况，探索构式与情感意义之间的关系，为后续实证研究提供一种方法，并对汉语构式的情感分析乃至语义计算的研究现状进行分析、展望未来的研究方向。

2 相关研究回顾

本文研究重点是构式语法对于应对情感分析语义隐匿性问题的意义，因此主要回顾情感分析研究对情感表达的认识与处理方式以及汉语语言学界对构式语法的相关研究。

2.1 汉语构式语法研究

构式语法理论兴起于上世纪90年代，是借鉴认知心理学格式塔(Gestalt)完形理论所创立的一种新兴语法研究理论，集形式、意义、用法于一体来认识和分析语言(陆俭明、吴海波, 2018:

(c) 2022 中国中文信息学会计算语言学专业委员会

1)。构式的权威定义最早由Goldberg(1995: 4)给出：C是一个构式当且仅当C是一个形式—意义的配对，且C的形式或意义的某些方面不能从C的构成成分或其他先前已有的构式中得到完全预测。此后语言学界对“构式”概念范畴的界定出现了狭义和广义的区分。汉语语言学界多主张有选择性地吸收构式语言观的创新性视角，以狭义定义研究构式，促进对汉语边缘性结构的描写和解释。陆俭明、吴海波(2018: 2-3)主张“构式”应当是“自由的边缘性句法结构，既有象征关系，内部又有结构性的组成关系”。詹卫东(2017: 232)指出：“（典型的）构式是无递归性的非平凡的短语结构”。部分学者采用以上狭义定义对一部分具有态度意义的汉语构式进行了实证研究(刘宗保, 2011; 郑娟曼, 2012; 李劲荣, 2015; 刘晨阳, 2016; 胡习之, 2017)，进行了详尽的句法语义分析。总的来说，目前语言学界对于构式范畴的认定仍存在争论，但无论在心理认知还是句法语义接口研究等领域，构式作为一类在形式和意义方面都具较强自足性的语言单位的理据性正逐渐确立，而且在应用研究方面成果不菲，较好地描写和解释了对汉语中部分边缘性结构的句法语义特征，很好地补充了传统句法语义研究对非常规结构关注的不足。

2.2 情感分析研究对情感表达的认识与处理

情感表达是表达了情感意义或立场的语言单位。随着万维网的快速发展以及人文社科领域的“情感转向”(Hardt, 2007)，情感表达已成为目前语言学和计算机科学共同关注的热点研究领域。情感表达研究主要有两种路径：一是理论驱动的语言学路径，即情感表达的语言学研究，注重分析文本中特定语言单位与情感意义的关系；二是应用驱动的量化研究路径，即情感分析，以实现对情感意义的自动计算处理和量化分析为目标。

情感表达的量化研究以情感语义的量化计算以及情感表达的自动抽取为目标，属于应用研究，近年来该领域被统称为情感分析研究。情感分析方法主要可分为三种方法：基于知识库的方法；机器学习方法；融合知识库和机器学习的混合方法(Cambria et al., 2017)。知识库是情感分析的基础资源之一，专门用于情感分析的知识库又称为情感知识库，是语义知识库的一种，汉语情感分析领域已经产生了情感词典(王科、夏睿, 2016; 赵妍妍等, 2017)、情感语义规则库(万岩、杜振中, 2020)、情感常识库(杨亮等, 2019)、情感句子模式库(陈涛等, 2013)等多种形式的情感知识库的构建和应用研究。目前汉语情感知识库建设和应用研究过于偏重表达态度和量级意义的词汇以及习语、谚语、惯用语等短语的语义知识，情感词典作为汉语情感知识库资源的主体，收录的也是这部分词汇和短语。但对于同样表达态度和量级意义的边缘性结构，也就是文章所说的“构式”，则在情感分析研究中受到了忽视。汉语中存在许多这类构式。在网络评论等真实的汉语评价语料中，由于语料的非正式性、去中心化等特征，这些构式可能较高频地出现。

构式研究已经得到语言学界的重视，态度和量级义构式研究在汉语本体研究中也有较多成果，但这类较为特殊、边缘化的语言单位尚未得到情感分析研究的重视，除了少量研究，如黄思思、詹卫东(2018)以42条量级义构式和185条态度义构式为例，探讨了适用于情感分析的汉语量级和态度义构式的语义分析和形式化表征问题。目前既缺乏面向汉语情感分析的构式知识库，而且在实证研究方面，也未有研究基于汉语真实评价语料，分析构式的分布情况，探讨情感意义表达与构式之类的关系。

3 研究设计

3.1 研究问题

文章的研究对象为汉语量级与态度义构式，它们既可能包含常项也可能包含变项，它们的意义包含量级或态度这两类在情感分析中最为关键的语义成分，但其整体的量级或态度意义又无法通过其组成成分如字、词等的意义推出，情感词库等已有的知识库资源又难以分析该类结构，从而可能导致情感分析结果准确性的降低。文章主要探讨的三个研究问题为：(1) 汉语真实网络评论文本中态度和量级义构式的分布情况是怎样的？(2) 对于汉语网络评论文本的情感分析，是否有必要考虑计算构式所承载的态度和量级语义信息？(3) 汉语构式的情感分析乃至语义计算目前存在哪些需要突破的难点？

3.2 语料选取

构式语料库选取的语料来源为热门话题微博的评论，话题均具有较大争议性，评论中的立场和态度具有高度不一致性，每个话题下的评论都为一万条以上。我们从热门话题的微博评论

(c) 2022 中国中文信息学会计算语言学专业委员会

中随机抽取4000余条评论，形成语料库，总字符数为13.3万。选取微博评论作为语料，是因为微博评论属于网络短文本，具有语体非正式、口语化、去中心化、数据噪声较多、情感表达丰富等特征，是典型的网络评价文本，是情感分析的主要处理对象之一。另外，文章采用的语料都属于话题型微博的评论，根据侯敏等(2013：136-138)指出，具有以下特点：(1) 句子简短，单句多；(2) 观点负面倾向多；(3) 表达情感强烈，理性评价淡化；(4) 口语色彩浓重，情感因子颗粒度加大，往往不再是词，而是短语甚至短句；(5) 隐晦表达观点，常用习语、反讽等方式表达观点，而不使用直接表达态度意义的词汇；(6) 评价对象省略；(7) 语言不规范。这些特点给情感分析带来了很大困难。同时第二至第五个特点可能意味着该类语料中存在大量构式，契合文章的研究目的。

3.3 语料库标注

由于目前尚无准确率较高的自动标注汉语构式的方法，我们需要先对语料中的态度和量级构式分别进行人工标注，目前学界尚未形成在真实语料中标注汉语构式的标准，我们尝试使用一种基于XML(Extensible Markup Language)的构式标注方法，并根据汉语的特点，将詹卫东(2017: 232)的狭义构式观作为文章对构式的定义：“(典型的)构式是无递归性的非平凡的短语结构”。文章不采用将语素、词、常规句法结构等纳入构式范畴的广义构式观，而只是将构式视为“对常规短语结构语法组合的必要补充”(ibid.)，原因是该观点较契合文章的研究对象及研究问题，也有助于增加操作的可行性。在标注过程中，只要某个结构符合文章的构式定义及判定标准，就将其标注为态度或量级构式，判定标准采用Hilpert(2014: 14-23)基于Goldberg(2006: 5)的构式定义所提出的四条标准：(1) 形式特殊性：该结构的整体或部分在形式上有别于一般的语法结构；(2) 语义不可预测性：该结构的意义具有非组合性，不严格等于其组成成分的意义加和；(3) 特殊限制性：该结构的整体或部分是不完全自由的，受到某些条件制约；(4) 搭配倾向性：该结构倾向于与某些成分或结构共现。

构式的标注方法是在文本中使用开始标记和结束标记包围整个构式，开始标记为<C TYPE="" FORM="" CAT="">，结束标记为</C>。开始标记包括三个属性，TYPE属性表示构式的语义，对应六个值：“att/neg”(负面态度义)；“att/pos”(正面态度义)；“att/bipolar”(双极性态度义)；“gradational/low”(低量义)；“gradational/high”(高量义)；“gradational/veryhigh”(极量义)。前三个值分别对应态度义构式的三类语义，后三个值分别对应量级构式的三类语义。FORM属性表示构式的形式，参照詹卫东(2017; 2018: 35-36)提出的构式形式表示法，值为由实例化的常项、变项以及加号构成的表达式。CAT属性表示构式的类别，参考詹卫东(2017; 2018: 18-20)提出的构式分类标准，对应四个值“frozen”(凝固型构式)；“semi-frozen”(半凝固型构式)；“phrasal”(短语型构式)；“compound-sentential”(复句型构式)。凝固型构式为完全由常项组成的构式；半凝固型构式的变项不超过2个，长度较短；短语型构式的变项数量可为1个及以上，长度可短可长；复句型构式由两个相对独立的部分组成，变项数至少为2个。评论文本中的构式标注示例如下(选自构式语料库)：

进门抢狗，这是<C TYPE=”att/neg” FORM=”什么+np/vp” CAT=”semi-frozen”>什么行为</C>。
<C TYPE=”att/neg” FORM=”不+v+就+不+v” CAT=”phrasal”>不批就不批</C>咯，少布置就少布置咯。
我让他<C TYPE=”gradational/veryhigh” FORM=”要+多+a+就+(有)+多+a” CAT=”phrasal”>要多快乐就多快乐</C>。
哪些该做不该做，还用规定吗？<C TYPE=”att/neg” FORM=”真+是/的+够+够+的” CAT=”frozen”>真是够够的</C>。
说真的杯子也<C TYPE=”att/neg” FORM=”没+(x)+几+q+vp/ap” CAT=”phrasal”>没几家干净的</C>，从来不用酒店杯子喝水。

Table 1: 构式标注示例

另外，态度和量级义词汇的标注，也使用XML标记，开始标记为<W TYPE=“”>，结束标记为</W>。属性TYPE代表词汇的语义类别，对应五个值：“att/pos”(正面态度义)；“att/neg”(负面态度义)；“gradational/low”(低量义)；“gradational/high”(高量

(c) 2022 中国中文信息学会计算语言学专业委员会

义)；“gradational/veryhigh”(极量义)。

4 基于构式语料库的数据分析

4.1 构式分布基本情况

经过人工标注，在语料库中共发现119个量级和态度义构式，构式出现的总频次为207次。按照上文提到的构式分类标准：凝固型构式共出现36个，频次为58次；半凝固型构式共出现18个，频次为49次；短语型构式共出现57个，频次为92次；复句型构式共出现8个，频次为8次。四类构式在语料库中的分布情况如下：

构式类型	凝固型	半凝固型	短语型	复句型
频次	58	49	92	8
总频次中占比	28.0%	23.7%	44.4%	3.9%
个数	36	18	57	8
总个数中占比	30.3%	15.1%	47.9%	6.7%

Table 2: 四类构式在语料库中的分布情况

“现代汉语构式数据库”(Chinese Construction Grammar Database, CCGD)(詹卫东, 2021)是目前唯一的较大规模的汉语构式知识库，统计数据显示，在该知识库收录的1108个汉语构式中，以上四类构式的分布情况如下：

构式类型	凝固型	半凝固型	短语型	复句型
构式条数	224	238	545	98
构式占比	20.2%	21.5%	49.2%	8.8%

Table 3: 四类构式在“现代汉语构式数据库”中的分布情况

对比数据可知，评论文本中四类构式的频次占比和个数占比都与CCGD中的占比相近，其频次占比和个数占比分布比较接近2：2：5：1的比例，这一比例在一定程度上反映了汉语真实评论文本中量级和态度义构式的分布情况，也为CCGD的统计数字提供了真实语料的佐证。至于所有汉语构式在真实文本中的分布是否仍然遵循这一比例，尚需进一步研究的验证。

构式语料库中共发现9个量级构式，出现频次共为11次，其中，4个为极量构式，5个为高量构式；6个为短语型构式，3个为半凝固型构式。量级构式中没有出现低量构式，只出现了高量和极量构式，这说明在汉语真实评论文本中，量级构式的分布以高量和极量构式为主，低量构式可能占比偏小。另外，语料库中共发现111个态度义构式，占总个数的93.2%，出现频次共为197次，占总频次的95%。与态度义构式相比，量级构式在构式出现总频次和总个数中占比都很低，仅占总个数的7.6%和总频次的5.3%，这说明量级构式在汉语真实评论文本中的分布可能较少，远低于表达主要情感意义的态度义构式。

态度义构式在构式语料库中的分布情况如下：

态度义类型	负面	正面	双极性
构式个数	101	6	4
个数占比	91.0%	5.4%	3.6%
构式频次	181	8	8
频次占比	91.9%	4.1%	4.1%

Table 4: 态度义构式在语料库中的分布情况

构式义为负面或正面态度的构式又称负面或正面态度义构式。而双极性构式则较为特殊，它也表达态度意义，但其极性的正负向具有不确定性，较为典型的双极性构式如“a+的+是”，其极性由形容词a来决定，具有不确定性，可能是正面也可能是负面。

语料库中出现的负面态度义构式无论是个数还是频次都占90%以上，占所发现的态度义构式的绝大多数，正面态度义构式和双极性构式出现都较少，占比相近，都在5%左右。这种

(c) 2022 中国中文信息学会计算语言学专业委员会

构式义分布的显著不对称性现象，在一定程度上说明汉语的负面态度义更多通过构式义来表达，为研究汉语负面评价表达规约化(方梅, 2017)现象提供了真实语料的佐证，在一定程度上说明了汉语在构式层面正负面态度意义的不对称性以及心理认知层面的“负面偏见”(Negativity Bias)(Rozin and Royzman, 2001)在汉语中的表现。

通过对语料库中构式的分析，我们发现，态度义构式的常项中很少包含出现表达态度意义的词汇，而且对于量级和态度义构式而言，仅分析其常项与变项各自的语义，无法推出其整体的构式义。

经过统计，语料库中所有构式的常项个数和变项个数情况如表5和表6所示：

常项个数	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
构式条数	12	33	27	29	13	3	0	1	1
占比	10.1%	27.7%	22.7%	24.4%	10.9%	2.5%	0%	0.8%	0.8%

Table 5: 语料库中所有构式的常项个数统计

由表5可知，语料库中，常项个数为1的构式占10.1%，常项个数为2至4个的构式占主要比重，达74.8%，与此相近的是，CCGD(詹卫东, 2021)中，常项个数为1的构式占该数据库所有构式的13.4%，常项个数为2到4的构式占79%。但在“构式库”(ibid.)中，常项个数为0的构式占9%，语料库中却没有发现常项个数为0的构式，这可能与样本数据规模的限制有关；语料库中出现了常项个数多达9个的构式，但“构式库”中没有此类构式。这说明，在真实评论文本中，常项个数为1至4的构式出现最多，占主要比重，常项个数大于或等于5的构式则分布较少。

变项个数	**0**	**1**	**2**	**3**
构式条数	36	59	18	6
占比	30.3%	49.6%	15.1%	5.0%

Table 6: 语料库中所有构式的变项个数统计

由表6可知，语料库中不存在变项个数为3以上的构式，变项个数为0至2的构式占主要比重，达到95%。与此相似的是，在CCGD(ibid.)中，变项个数为0至2的构式占90.4%，同样所占比重相近。这说明变项个数为0到2的构式在真实评论文本中占主要比重，分布最多，变项个数大于或等于3的构式则分布较少。

以上数据表明，语料库中的构式主要是常项个数为1-4、变项个数为0-2的构式，与CCGD（ibid.）的情况基本一致。

4.2 构式在语料库量级和态度意义表达中的比重分析

为了分析汉语真实文本中构式所承载的态度和量级语义信息的多寡，我们对语料库中表达态度和量级语义的构式与词汇进行了标注和统计。结果显示，态度义词汇共出现2453次，量级义词汇出现523次，总计2976次，以下为相关数据：

语义类型	**正面**	**负面**	**双极性**	**总计**
词汇频次	643	1810	0	2453
在态度义词汇总频次中占比	26.2%	73.8%	0%	100%
构式频次	8	181	8	197
在态度义构式总频次中占比	4.1%	91.9%	4.1%	100%
构式表义比重	0.40%	22.00%	100%	7.40%
词汇表义比重	99.6%	78.0%	0%	92.6%
注：构式表义比重=构式频次/(词汇频次+构式频次)；词汇表义比重=词汇频次/(词汇频次+构式频次)				

Table 7: 语料库中态度义词汇与态度义构式频次数据对比

态度和量级意义的表义单位主要是构式和词汇。句子往往也表达态度和量级意义，但通常句子语义的可推导性较强，可以拆分为更小的“原子”，如构式和词汇，因为句子不适宜作为表

(c) 2022 中国中文信息学会计算语言学专业委员会

义单位的进行统计。因此文章只统计构式和词汇的表义比重，表义比重是指构式或词汇这两类语言单位在表达相应类型语义中所占的比重大小。

语料库中共出现态度义词汇2453次，正面态度义词汇643次，负面态度义词汇1810次，正负态度义词汇比例约为1:3，远高于正负态度义构式的1:9。从表义比重上看，正面态度语义的构式表义比重很小，仅为0.4%，词汇表义比重高达99.6%；负面态度语义的构式表义比重则相对较高，达到了22%，词汇表义比重为78%。正负态度构式表义比重的比例为1:55，该悬殊的比例也进一步佐证了汉语负面评价表达规约化的现象。而更为特殊的双极性态度语义方面，构式表义比重则为100%，语料库中未发现表达此类语义的词汇，这也说明，更加复杂和模糊的语义，可能由往往构式表达，而不是词汇。

从整体上看，态度语义的词汇表义比重为92.6%，这表明词汇是评价文本中态度语义表达的主要语义单位，构式表义比重为7.4%，这说明评价文本中的构式只承载了比例较小的态度语义，但7.4%的比例也不能算小，这表明评论文本中的构式也承载了相当比重的态度语义，尤其是负面态度义。如果在态度语义计算的过程中，忽视构式所承载的态度义，足以影响语义计算的准确性，造成结果的偏差。

语义类型	低量	高量	极量	总计
词汇频次	18	404	101	523
在量级义词汇总频次中占比	3.40%	77.20%	19.30%	100%
构式频次	0	6	5	11
在量级义构式总频次中占比	0%	54.5%	45.5%	100%
构式表义比重	0%	1.5%	4.7%	2.1%
词汇表义比重	100%	98.5%	95.3%	97.9%

Table 8: 语料库中量级义词汇与量级义构式频次数据对比

语料库中共出现量级义词汇523次，高量词汇404次，低量词汇18次，极量词汇101次，表达高量语义的词汇占大多数，为77.2%，表达低量和极量语义的词汇都相对较少，分别为3.4%和19.3%。

表义比重方面，低量语义的构式表义比重为0%，词汇表义比重为100%，其语义全部由词汇表达；高量语义的构式表义比重很小，仅为1.5%，词汇表义比重为98.5%；极量语义的构式表义比重略高于高量语义，为4.7%，而词汇表义比重为95.3%。整体而言，量级语义的词汇表义比重为97.9%，相比之下，量级语义的构式表义比重仅为2.1%，远低于词汇表义比重，说明评论文本中的量级语义绝大部分由词汇来承载，构式在量级语义表达中起到的作用很小。

4.3 构式语义对于情感分析的意义

总的来说，上述语料库数据的分析显示：在汉语真实评论文本中，词汇承载了大部分态度和量级语义信息，是主要的表义单位；构式所承载的态度和量级语义信息较少，构式虽然不是主要的情感意义单位，但作为边缘性结构，其承载的态度语义信息仍占一定比例，为7-8%，其承载的语义信息不应完全被忽视。而构式所表达的量级语义仅占2.1%，比重小。至于在情感分析中是否需要考虑计算构式所承载的态度和量级语义，需要根据研究要求的计算精确度和研究目的来具体确定，不宜一概而论，若研究对计算精度要求较高，则可以考虑将构式的语义纳入计算范围。但以上数据是仅基于13万字左右的汉语语料库得出的，要进一步验证上述结论，尚需后续基于更大规模、包含更丰富主题的语料库进行研究。

从数据上看，如果汉语构式的情感语义能够得到汉语情感分析系统的准确解析，最多可在情感倾向计算上提升7-8%的准确率，最多可在情感强度计算上提升2%左右的准确率。但要实现以上准确率的提升需要较好地处理汉语构式的情感分析问题，也即汉语构式的隐匿性语义计算的问题，在算法、知识库资源、知识表示等方面都需要做出较多努力和探索。因此，对于汉语情感分析是否应当考虑计算构式的语义，需要考虑投入成本与产出的比率，如果追求较高的情感计算准确率，则有必要计算构式的语义。如何实现对汉语构式的情感分析本身也是一个较复杂的语义计算问题，下文将对相关难点、研究现状、发展方向进行讨论。

(c) 2022 中国中文信息学会计算语言学专业委员会

5 对汉语构式情感分析问题的思考：问题与展望

在汉语情感分析中融入构式的语义知识，有助于应对汉语情感分析的部分语义隐匿性问题。汉语构式的情感分析属于构式计算和语义计算的研究范畴，该领域产生了一些研究成果，但也面临诸多难题，主要是四方面问题。文章对相关研究难点进行了梳理，提出了相应的对策，展望了后续研究。

5.1 探索面向计算的构式定义

第一个问题是构式的定义问题，即如何面向自然语言处理的需求界定构式的内涵和外延，给出一个可操作性较强、适用于计算机处理的汉语构式定义。此处论及定义问题，主要原因是，目前语言学界对构式这一重要概念界定的含糊不清、各自为政的现象，不能适应句法语义自动分析的要求。国外构式研究倾向于接受宽泛的“构式”定义，如“激进”构式学派把语素、词汇、短语等各个层级都涵盖到构式的范畴内(Croft and Cruse, 2004: 225-257)，Goldberg(1995: 4; 2003; 2006: 5)也一直不断修正其关于构式的定义，使得构式的概念同时涵盖了不可预测性的边缘结构以及出现频率高的可预测性常规结构(比如核心句式)。甚至有将构式延伸到语篇层面的观点(Ostman, 2005)。而国内汉语学界则倾向于认同狭义的“构式”定义，比如陆俭明与吴海波(2018: 3)认为构式应当是具有不可预测性的形义配对结构体，这比较接近詹卫东(2017: 232)及文章采用的定义——“无递归性的非平凡的短语结构”。但无论是国内汉语学界还是国际语言学界，对于构式应该包含哪些语言单位，仍无定论。虽然构式理论的提出和深入发展的确给句法-语义界面研究带来了新的理论视角，但此种理论层面的割裂现象阻碍了构式理论在句法语义分析中的应用。在句法语义研究中，某一基础概念内涵的确定是重要的研究工作，比如汉语学界对汉语词类划分标准这一基础问题的争论和探讨(詹卫东, 2013; 叶脉清、聂仁发, 2015; 杨丽姣等, 2021)，对词典编纂、汉语词汇知识库构建、自然语言处理等都产生了较大的影响。如果构式理论要更好地应用于语言工程领域，也需要这样一番专门的争论和探讨。语言学界需要加强对面向计算的构式定义的探讨，结合语义计算的实例进行分析，探索适用于计算的“构式”概念操作界定。

5.2 加强汉语构式知识库构建研究

第二个问题是汉语构式知识库的构建问题。构式作为意义与形式的配对体(Goldberg, 1995: 4)，其语义具有非组合性和不可预测性(Hilpert, 2014:14-23)。那么对于情感分析而言，计算机获取构式的先验语言知识有两种途径，一是真实文本中人工标注的构式知识，二是专门的量级和态度义构式知识库。在真实文本中人工标注构式也在某种程度需要依靠知识库提供的知识。因此有必要建设标注有句法和语义等信息的构式知识库。目前，成熟的汉语构式知识库资源中仅有CCGD，其收录的构式为1108条，主要收录的是已经出现在汉语本体研究论文中的构式，多为学界讨论较多、较典型的汉语构式(詹卫东, 2021)。目前对于汉语中到底有多少构式这一问题，学界尚无定论，就目前CCGD的规模而言，其应该远未能覆盖所有汉语语言使用中的构式，其收录的量级和态度义构式可能也不全面。因此，要为汉语情感分析提供构式先验知识的支持，一方面需要构建对量级和态度义构式收录全面的知识库，另一方面也需要进行相应研究来测算汉语构式总数的数量级，才能为构式知识库的构建提供参考。

此外汉语构式知识库的收录范围也是值得探讨的问题，主要是凝固型构式的边界问题，即成语、俗语、谚语以及新兴网络流行语等是否应该收录为凝固型构式。我们认为成语、俗语、谚语虽然也符合文章对构式的定义，但考虑到它们数量较大，如“汉语习语知识库”(Chinese Idiom Knowledge Base, CIKB)(Lei and Yu, 2010)就收录了多达38117条汉语成语、俗语、谚语等，而且它们出现时间长，很多辞书都已收录，语义确定性强，适合由情感词典收录。而新兴网络流行语，如“醉了”、“打脸了”、“凉凉”等网络流行语，往往有特殊的修辞效果，从字面义无法推知其意义。它们符合文章的构式定义，我们可以将其视为凝固型构式。同时由于网络流行语出现时间较短，语体非正式和偏口语化，传统语文词典和情感词典等知识库基本都没有收录，其构式义又往往具有情感意义，因此有必要由构式知识库来对其进行收录，未来应注意对这部分网络流行语的整理、追踪、收录、知识化。

5.3 探索科学的构式形式化表示方法

第三个问题是构式的形式化表示问题，即应该采用何种标准，来呈现单个构式的句法语

(c) 2022 中国中文信息学会计算语言学专业委员会

义知识或者汉语文本中关于构式的句法和语义知识，以便更好地服务于构式的识别。由于学界对汉语构式的计算处理研究还在早期阶段，构式的形式化表示问题尚缺乏系统研究。前人研究主要探索了结构化的知识库中单个构式的形式化表示问题：CCGD中使用10余种特征-值对(feature-value pair)来表示单个构式的句法语义特点，如变体、义项、释义模板、实例等(詹卫东, 2017; 2021)。而对于如何在语料库这一类非结构化数据库中表征汉语构式的语言学知识，研究则很少，少数相关研究如黄彤等(2020)尝试采用中文抽象语义表示(Chinese Abstract Meaning Representation, CAMR)这一面向自然语言处理的句法语义表示体系，对CCGD中所有条目的例句进行了构式的形式化标注，发现CAMR可以表示出61.2%的基本符合组合原则的构式，同时发现在文本中标注构式存在语义省略、凝固型构式难以拆分表示、语义范围难以确定、释义需要语境和语用推导等困难，并对标注策略进行了总结。

总体来看，学界在汉语构式的语料库标注方面的研究较为缺乏，也没有经过科学标注的汉语构式语料库作为参考标准和研究材料；而由于CCGD的实践，学界在汉语构式的知识库表示方面已经积累了一定经验，但仍需进一步研究和探索：从描写全面性的角度，探索应当用多少类属性来表示构式所承载的语言知识；探索在实际语义计算中，最常调用的构式知识属性是哪些；进一步探索构式的句法和语义是否有更合理、高效的形式化表示方法。

5.4 加强构式识别研究

第四个问题是构式的识别问题，即采用何种算法能够更好地从文本中自动识别构式，这是构式计算处理过程的最后一步。由于汉语句法语义规则本身以及汉语构式的复杂性，以及汉语句法语义分析研究对非常规、句法组合规则难以分析的边缘性结构重视相对不足，系统性研究较少(黄彤等, 2020)，目前尚未摸索出成熟高效的算法。同时我们也缺乏面向汉语构式识别的标注数据集，无法训练出基于有监督学习的构式标注器。黄海斌等(2020)在没有训练语料的情况下，探索了融合高斯混合模型、正则表达式以及词性匹配的无监督汉语构式自动识别算法，并指出，构式识别算法面临的最大困难是在无训练语料的情况下确定句子中构式的边界信息。除此之外，汉语学界缺少相关研究积累，这给构式知识在情感分析乃至自然语言处理中的应用带来了困难。后续需要加强汉语构式数据集的标注以及识别算法实验性研究。

6 结语

文章为构式语法应用于情感分析提供了实证数据，以量化的方式揭示了汉语构式与情感意义的关系，为汉语情感分析是否需要考虑计算构式的语义信息这一问题的解答提供了数据支持，也为情感分析提供了一种新的研究视角，即基于真实文本进行实证分析，弥补了汉语情感分析研究对以构式为代表的语义隐匿性现象关注的不足，还较为系统地梳理了汉语构式情感分析的相关研究及现实困难，推动了汉语情感分析研究的进展。同时，文章还为汉语构式研究提供了实证数据支持，在构式语料库标注方面进行了探索。因此文章对于汉语情感分析、构式计算、构式语法理论等研究均有一定价值，但这方面的研究还需更大规模语料库的支持，期望文章能够为后续研究抛砖引玉，提供一种研究思路和方法。

参考文献

Cambria, E., Das, D., Bandyopadhyay, S. & Feraco, A. Affective Computing and Sentiment Analysis. In Cambria, E., Das, D., Bandyopadhyay, S. & Feraco, A. (eds.). *A Practical Guide to Sentiment Analysis*. Berlin: Springer. 2017: 1-10.

Croft, W. & Cruse, D. A. *Cognitive Linguistics*. Cambridge: Cambridge University Press. 2004.

Goldberg, A. *A Construction Grammar Approach to Argument Structure*. Chicago/London: The University of Chicago Press. 1995.

Goldberg, A. Constructions: A New Theoretical Approach to Language. *Trends in Cognitive Sciences*, 2003, 7(5): 219-224.

Goldberg, A. *Construction at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press. 2006.

Hardt, M. Foreword: What Affects are Good for. In Clough P. & Halley J. (eds.). *The Affective Turn: Theorizing the Social*. Durham/London: Duke University Press. 2007: 1-5.

(c) 2022 中国中文信息学会计算语言学专业委员会

Hilpert, M. *Construction Grammar and Its Application to English.* Edinburgh: Edinburgh University Press. 2014.

Lei Wang & Yu Shiwen. Construction of Chinese Idiom Knowledge-base and Its Applications. In Laporte E., Nakov, P., Ramisch, C. & Villavicencio, A. (eds.). *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications.* Beijing: Chinese Information Processing Society of China. 2010: 11-18.

Ostman, J.-O. Construction Discourse: a Prolegomenon. In Ostman, J.-O. & Fried, M. (eds.). *Construction Grammars: Cognitive Grounding and Theoretical Extensions.* Amsterdam: John Benjamins. 2005: 121-144.

Rozin, P. & Royzman, E. B. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, 2001, 5(4): 296-320.

陈涛, 徐睿峰, 吴明芬, 刘滨. 一种基于情感句模的文本情感分类方法.中文信息学报, 2013(5): 67-74.

方梅. 负面评价表达的规约化.中国语文, 2017(2): 131-147.

侯敏, 滕永林, 李雪燕, 陈毓麟, 郑双美, 侯明午, 周红照. 话题型微博语言特点及其情感分析策略研究.语言文字应用, 2013(2): 135-143.

胡习之. 构式“你才X呢”再探.当代修辞学, 2017(6): 73-81.

黄海斌, 常宝宝, 詹卫东. 基于高斯混合模型的现代汉语构式成分自动标注方法.中文信息学报, 2020(9): 1-8.

黄思思, 詹卫东. 面向情感分析的构式主观态度义初探.外语教学, 2018(6): 27-33.

黄彤, 李斌, 闫培艺, 戴玉玲, 曲维光. 基于抽象语义表示的汉语构式标注与分析.中文信息学报, 2020(10): 1-10.

李劲荣. 列举形式“什么X”与“X什么的”的语义偏向.汉语学习, 2015(5): 40-48.

刘兵著, 刘康, 赵军译. 情感分析：挖掘观点、情感和情绪.北京: 机械工业出版社. 2017.

刘晨阳. 警告义“再VP”构式探析.语言科学, 2016(4): 412-421.

刘宗保. 警告义构式“叫/让”句探析.汉语学习, 2011(2): 60-67.

陆俭明, 吴海波. 构式语法理论研究中需要澄清的一些问题.外语研究, 2018(2): 1-5.

万岩, 杜振中. 融合情感词典和语义规则的微博评论细粒度情感分析.情报探索, 2020(11): 34-41.

王科, 夏睿. 情感词典自动构建方法综述.自动化学报, 2016(4): 495-511.

杨丽姣, 肖航, 刘智颖. 《信息处理用现代汉语词类标记规范》修订方案.语言文字应用, 2021(3): 111-120.

杨亮, 周逢清, 林鸿飞, 殷福亮, 张一鸣.基于情感常识的情感分析.中文信息学报, 2019(6): 94-99.

叶脉清, 聂仁发. 新世纪以来现代汉语词类研究综述.现代语文, 2015(8): 7-10.

詹卫东. 计算机句法结构分析需要什么样的词类知识——兼评近年来汉语词类研究的新进展.中国语文, 2013(2): 178-190.

詹卫东. 从短语到构式: 构式知识库建设的若干理论问题探析.中文信息学报, 2017(1): 230-238.

詹卫东. “现代汉语构式知识库”填写规范. http://ccl.pku.edu.cn/ccgd/downloaddoc/,2018-12-01.

詹卫东. 现代汉语构式知识库. http://ccl.pku.edu.cn/ccgd/,2021-12-01.

詹卫东, 王佳骏, 黄海斌, 陈龙. 构式的表征与语料标注——现代汉语构式数据资源建设中的基本问题. 第21届汉语词汇语义学国际研讨会, 中国香港, 2020年5月.

赵妍妍, 秦兵, 石秋慧, 刘挺. 大规模情感词典的构建及其在情感分类中的应用.中文信息学报, 2017(2): 187-193.

郑娟曼. 从贬抑性习语构式看构式化的机制——以“真是(的)”与“整个一个X”为例.世界汉语教学, 2012(4): 520-530.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于关系图注意力网络和宽度学习的负面情绪识别方法

彭三城[1]，**陈广豪**[1]，**曹丽红**[1,*]，**曾嵘**[2]，**周咏梅**[3]，**李心广**[1]
1.语言工程与计算实验室，广东外语外贸大学，广东广州，510006
2.信息光电子科技学院，华南师范大学，广东广州，510006
3.信息科学与技术学院，广东外语外贸大学，广东广州，510006

摘要

对话文本负面情绪识别主要是从对话文本中识别出每个话语的负面情绪，近年来已成为了一个研究热点。然而，让机器在对话文本中识别负面情绪是一项具有挑战性的任务，因为人们在对话中的情感表达通常存在上下文关系。为了解决上述问题，本文提出一种基于关系图注意力网络(Rational Graph Attention Network, RGAT)和宽度学习(Broad Learning, BL)的对话文本负面情绪识别方法，即RGAT-BL。该方法采用预训练模型RoBERTa生成对话文本的初始向量；然后，采用Bi-LSTM对文本向量的局部特征和上下文语义特征进行提取，从而获取话语级别的特征；采用RGAT对说话者之间的长距离依赖关系进行提取，从而获取说话者级别的特征；采用BL对上述两种拼接后的特征进行处理，从而实现对负面情绪进行分类输出。通过在三种数据集上与基线模型进行对比实验，结果表明所提出的方法在三个数据集上的weighted-$F1$、macro-$F1$值都优于基线模型。

关键词： 对话文本；负面情绪；关系图注意力网络；宽度学习；预训练模型

Negative Emotion Recognition Method Based on Rational Graph Attention Network and Broad Learning

Sancheng Peng[1], **Guanghao Chen**[1], **Lihong Cao**[1], **Rong Zeng**[2], **Yongmei zhou**[3], **Xinguang Li**[1]
1.Laboratory of Language Engineering and Computing,
Guangdong University of Foreign Studies, China.
2.School of Information and OptoelectronicScience and Engineering,
South China Normal University, China.
3.School of Information Science and Technoogy,
Guangdong University of Foreign Studies, China.

Abstract

Negative emotion recognition in textual conversations aims to identify the negative emotion of each utterance from textual conversations, which has become a hot research topic in recent years. However, enabling machines to recognize negative emotions in textual conversations is a challenging task, because there are contexts for peoples'emotional expression in conversations. To address the problem, we propose a method for negative emotion recognition based on rational graph attention network (RGAT) and broad learning (BL), namely RGAT-BL. We use pre-training model Roberta

收稿日期：2022-08-01 **定稿日期**：2020-08-15
基金项目：本课题得到到国家自然科学基金资助项目（编号：61876205，61877013），教育部人文社科项目（19YJAZH128，20YJAZH118）。
作者简介：彭三城（1974—），博士，教授，主要研究领域为情绪计算、宽度学习。陈广豪（1998—），硕士研究生，主要研究方向为负面情绪识别；曹丽红（1985-），硕士，讲师，主要研究方向为情绪计算；曾嵘（1998—），硕士研究生，主要研究方向为情绪原因识别；周咏梅(1971—)，硕士，教授，主要研究领域为情感计算；李心广(1962—)，博士，教授，主要研究领域为语音识别、情感计算
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

to generate the initial vector for textual conversations. Then, we adopt Bi-LSTM to extract local features and context semantic features of textual vectors, so as to obtain utterance-level features. Thirdly, we employ RGAT to extract long-distance dependency among speakers, so as to obtain the speaker-level features. At last, we use BL to process the above two connected features, so as to conduct the classified output of negative emotions. Compared with baseline models on three datasets, the experimental results show that the weighted-$F1$ and macro-$F1$ values of the proposed method are better than the baseline model on the three datasets.

Keywords: Textual conservations , negative emotion , rational graph attention network , broad learning , pre-training model

1 绪论

对话文本是由多个说话者交替说话而产生的，其全局语义是由多个用户在对话的语境中共同构建的，话语之间以及说话者之间在情绪表达上具有很强的关联性，上述特点使得对话文本的情绪识别成为了自然语言处理的一个研究热点 (彭韬et al., 2021)。情绪是指人们对外界刺激所作出的反应，而负面情绪是指人们对负面事件所作出的主观消极情绪反应 (赖河蒗et al., 2022a)。如何从这些海量的对话文本中自动地识别出携带负面情绪的信息，对于社交网络 (X et al., 2020; Peng et al., 2019)安全具有重要的意义。

现有的方法大都是针对短文本的情绪分类 (Sancheng et al., 2021)；同时，由于对话文本中话语之间以及说话者之间存在着一定的依赖关系，与短文本的情绪识别任务相比，对话文本的情绪识别任务无疑更具有挑战性。现有的针对对话文本的情绪识别方法主要包括：基于对话序列的方法 (Hazarika et al., 2018; Jiao et al., 2019; Majumder et al., 2019; Poria et al., 2016; Jin et al., 2020)和基于图神经网络的方法 (Zhang et al., 2019; Ghosal et al., 2019; Zhong et al., 2019; Shen et al., 2021b)。基于对话序列的方法大都采用长短时记忆网络(Long Short-Term Memory, LSTM)、门控循环单元(Gated Recurrent Unit, GRU)等深度学习模型来对话语级别的特征进行提取。这些模型虽能捕获长距离特征和话语之间的关联性，但该类方法忽略了说话者在对话中的相互影响和作用。

基于图神经网络的方法大都是采用图卷积神经网络对对话文本进行建模，从而对说话者之间的影响与联系进行更好地刻画。图卷积网络能捕获文本的结构特征以及单词间非连续和长距离的依赖关系。但是，图卷积网络还存在以下不足 (郑诚et al., 2022b)：一是由于是把单词表示为图的节点，采用邻接矩阵来表示节点的邻域信息，没有考虑对话文本的顺序结构，从而导致不能捕获文本的上下文语义信息；二是对局部特征信息的提取存在不足。

由于Bi-LSTM (Schuster and Paliwal, 1997)具有能有效捕获局部特征和上下文语义特征等优点，关系图注意力网络(Rational Graph Attention Network, RGAT) (Schlichtkrull et al., 2018)具有能有效捕获对话文本的结构特征和单词间的长距离依赖关系等优点，宽度学习具有网络结构简单、训练时间短、泛化能力强等特点。

因此，为了解决上面所提到的问题，提出一种基于RGAT和宽度学习(Broad Learning, BL) (Chen and Liu, 2017) 的对话文本负面情绪识别方法，即RGAT-BL。主要的贡献如下：采用Bi-LSTM能有效捕获局部特征和上下文语义特征的优点来提取话语级别的特征；然后，采用RGAT能有效捕获对话文本的结构特征和单词间的长距离依赖关系等优点来提取说话者级别的特征；最后，将对话语级别和说话者级别的特征进行拼接后，采用BL对负面情绪分类输出。在IEMOCAP、MELD和EmoryNLP三个对话文本数据集上，与基线模型进行对比实验，结果表明所提出的模型在每个数据集上的性能都优于基线模型。

2 相关工作

2.1 对话文本情绪分类模型

现有对话文本的方法大都针对情绪进行分类，主要可分为两种：基于对话序列的模型和基于图神经网络的模型。

(1) 基于对话序列的模型

(c) 2022 中国中文信息学会计算语言学专业委员会

Hazarika等人 (Hazarika et al., 2018)提出ICON模型，它将GRU和Multihop Attention机制相结合，学习话语之间的相关性。Jiao等人 (Jiao et al., 2019)提出HiGRU模型，该模型采用层两级GRU，一层用于提取每个话语内的上下文关系，另外一层用于提取话语级别的特征。Majumder等人 (Majumder et al., 2019)提出DialogueRNN模型，该模型采用CNN提取对话文本的上下文特征，用RNN提取话语级别的特征。Poria等人 (Poria et al., 2016)提出基于卷积多核学习的分类器以及基于上下文的层次Bi-LSTM来对多模态情绪进行识别。Jin等人 (Jin et al., 2020)提出一种层次的多模态Transformer，采用局部感知注意力机制和说话者感知注意力机制来分别捕捉说话者的局部语境和情绪惯性。

(2) 基于图神经网络的模型

Zhang等人 (Zhang et al., 2019)构建一种基于GCN的情绪识别模型来捕捉对话文本中的话语和说话者之间的关联。Ghosal等 (Ghosal et al., 2019)提出DialogueGCN模型，该模型分别采用双向GRU和GCN以提取对话级别特征和说话者级别特征。Zhong等人 (Zhong et al., 2019)提出KET模型，即采用自注意力机制提取话语的上下文关系，采用图注意力机制提取全局话语的特征。Shen等人 (Shen et al., 2021b)提出DAG-ERC模型，该模型通过构建有向无环图来对对话文本进行建模和训练。

2.2 宽度学习

BL是陈俊龙教授等人在2018年提出来的，主要由输入层、特征节点、增强节点和输出层四部分组成。BL需要训练的参数较少，一般只包括输出层的权重，即每个分类标签的对应权重。它可以通过岭回归算法 (Hoerl and Kennard, 1970)来快速获取。因此，BL具有结构简单、参数量少、训练时间短等特点，在许多分类任务上均得到了应用，如图像分类 (Chu et al., 2021)、视觉识别 (Jin et al., 2021)、情绪分类 (Peng et al., 2021)等。因此，本文采用BL作为分类器，有助于在短时间内得出更准确的分类结果，提升模型性能。

3 对话负面情绪识别

3.1 问题定义

给定一轮对话$U=[u_1,u_2,\cdots u_N]$以及M个说话者$S=[s_1,s_2,\cdots s_M]$，其中，N表示一轮对话中的话语个数，即M个人在本轮对话中总共说了N句话，其中，$u_i\in\mathbb{R}^d$ 表示第i句话的特征向量，d表示向量的维度。对话文本负面情绪识别任务的核心就是通过给定的一轮对话U以及M个说话者，从而预测出该轮对话中每个话语对应的负面情绪类别(如悲伤、生气)。

3.2 模型框架

本文所提出的RGAT-BL的框架结构主要包括文本编码层、话语级别编码层、说话者级别编码层和情绪分类层四个部分。文本编码层采用RoBERTa将句子中的每个单词映射到连续的低维向量空间中。话语级别编码层采用Bi-LSTM来提取话语级别的特征。说话者级别编码层采用RGAT来提取说话者级别的特征。情绪分类层采用BL对上述两种特征进行拼接后，并通过softmax输出负面情绪的分类结果。RGAT-BL具体的结构如图1所示。

3.3 文本编码层

在对对话文本进行预处理后，使用RoBERTa对对话文本的特征进行抽取，生成词向量；然后，对RoBERTa进行微调，使得模型每层的参数满足以下关系：

$$q_n^i=q_{n-1}^i-\mu^m\times\nabla_{q^l}H(q) \tag{1}$$

其中，q^i表示模型的第i层参数，n表示时间步长，$\nabla_{q^l}H(q)$表示模型目标函数的梯度，μ^m表示第m层的学习率，且需满足以下关系：

$$\mu^{m-1}=\delta\times\mu^m \tag{2}$$

其中，δ表示学习率的衰减速率且$\delta\leq 1$，当$\delta\leq 1$时，学习率逐层衰减，当$\delta=1$时，学习率则不做衰减，即每层的学习率保持不变。

利用微调后的RoBERTa对每条语句进行预训练以生成词向量，取词符“[CLS]”对应的输出向量作为句向量，从而得到句向量矩阵$U=[u_1,u_2,\cdots u_N]\in\mathbb{R}^{N\times d}$，其中，$N$表示一轮对话的话语数量，$u_i\in\mathbb{R}^d$表示第$i$句话的句向量，$d$表示句向量的维度。

(c) 2022 中国中文信息学会计算语言学专业委员会

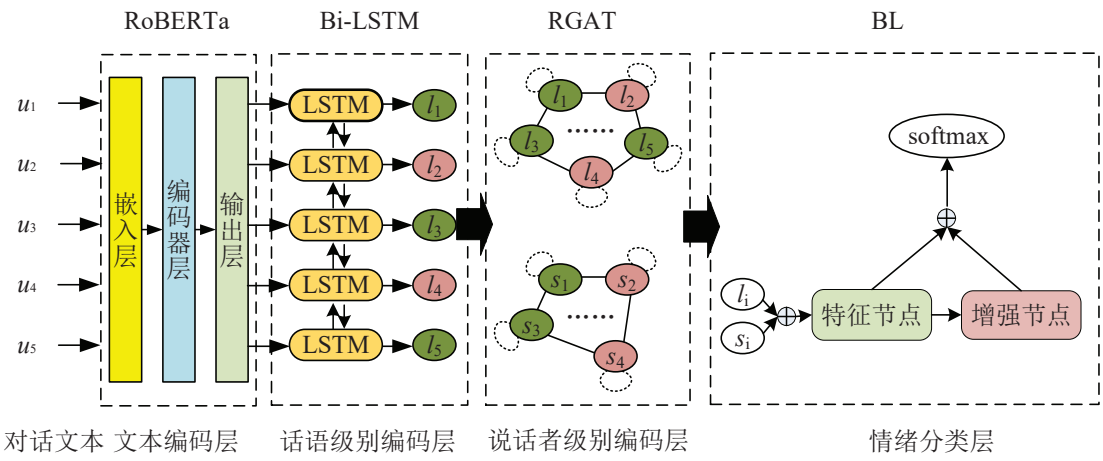


图 1. RGAT-BL的框架

3.4 话语级别编码层

由于对话文本是一个连续的话语序列，前后的话语之间往往有很强的关联性，即文本的上下文信息，这对于话语的情绪识别也至关重要。因此，采用Bi-LSTM能有效捕获上下文语义特征的优点来对话语级别的特征进行提取。Bi-LSTM的基本思想就是利用两个方向相反的LSTM来分别处理正向和反向序列，以获取特征之间在两个方向上的关联，从而将两个方向的关联信息输出作为LSTM的前、后向特征；然后，将前、后向特征进行拼接。使得Bi-LSTM能有效捕获文本的上下文信息。具体的过程表示如下：

$$\overrightarrow{l_i} = \overrightarrow{\text{LSTM}}(\overrightarrow{l_{i-1}}, u_i) \tag{3}$$

$$\overleftarrow{l_i} = \overleftarrow{\text{LSTM}}(\overleftarrow{l_{i+1}}, u_i) \tag{4}$$

$$l_i = [\overrightarrow{l_i}, \overleftarrow{l_i}], i = 1, 2, \cdots, N \tag{5}$$

其中，$\overrightarrow{l_i}$和$\overleftarrow{l_i}$分别表示第i句话的前向和后向LSTM特征。

因此，对于一轮话语，话语级别的特征可表示为：

$$L = [l_1, l_2, \cdots, l_N] \in \mathbb{R}^{N \times d_l} \tag{6}$$

其中，d_l表示话语级别特征的维度，在Bi-LSTM中，d_l表示其中一个LSTM层的神经单元数的两倍。

3.5 说话者级别编码层

由于RGAT能有效捕获对话文本的结构特征和单词间的长距离依赖关系等优点，因此，本文采用RGAT来对说话者级别进行编码。首先构造一个有向图用于保存对话者之间的对话以及情绪交互关系；然后，采用RGAT来获取对话文本的结构特征和单词间的长距离依赖关系。具体的构建过程如下。

假设有向图$G = \{V, E, R, W_G\}$用来表示一个具有N个话语的对话，其中，V表示节点集合，E表示边集合，R表示边的关系类型集合，W表示边的注意力权重集合。

V：每个话语对应G的一个节点$l_i \in V, i = 1, 2, \cdots N$，每个节点都是采用RoBERTa生成的面向话语的句向量$u_i \in \mathbb{R}^d$来表示。假设$p(\bullet)$表示话语到说话者之间的映射关系，即$p(l_i) \in S$表示节点l_i所对应的说话者$s_j, j = 1, 2, ..., M$。

E：如果节点l_i与l_j之间的边可表示为$e_{ij} \in E, i, j = 1, 2, \cdots N$，那么$e_{ij}$可用来表示话语之间的上下文关系。如果要表示$l_i$与其他话语的上下文关系时，那么$l_i$ 需要与G中的所有节点进行连接，即G成为了一个全连通图。

(c) 2022 中国中文信息学会计算语言学专业委员会

R：不同的边表示话语之间可能具有不同的上下文关系，共包括五种关系，可表示为$R=\{1,2,3,4,5\}$。当$p(l_i)=p(l_j),i<j$时，$R=1$；当$p(l_i)=p(l_j),i>j$时，$R=2$；当$p(l_i)\neq p(l_j),i<j$时，$R=3$；当$p(l_i)\neq p(l_j),i>j$时，$R=4$；当$i=j$时，$R=5$。每条边表示每个说话者的该话语受到其他说话者的影响或自身的影响，从而有效对对话文本的结构特征和单词间的长距离依赖关系进行提取。

W_G：节点l_i对的l_j边e_{ij}的权重可表示为$w_{ij}\in W_G$，其数值的大小表示了话语l_i对与话语l_j的影响程度。

下面给出基于RGAT的说话者级别特征的提取过程。

先对边的注意力权重w_{ij}进行求解，采用一个单层前馈神经网络来计算l_j对l_i的注意力系数c_{ij}，具体表示如下：

$$c_{ij}={l_i}^{\mathrm{T}}W_e\left[l_i,l_j\right],j=i-p,\cdots,i+f \tag{7}$$

其中W_e表示权重矩阵。

为了更好地表示不同节点对l_i的影响程度，采用softmax函数将注意力系数进行归一化，即l_j对l_i的边的注意力权重w_{ij}可表示如下：

$$w_{ij}=\mathrm{softmax}(c_{ij})=\frac{\exp\left({l_i}^{\mathrm{T}}W_e\left[l_i,l_j\right]\right)}{\sum_{k=i-p}^{i+f}\exp\left({l_i}^{\mathrm{T}}W_e\left[l_i,l_k\right]\right)},j=i-p,\cdots,i+f \tag{8}$$

采用两层RGAT进行来对说话者进行编码。对于第1层RGAT，通过聚合邻居节点信息将节点l_i转化为说话人相关的特征向量h_i，具体过程如下：

$$h_i=\sigma(\sum_{r\in R}\sum_{j\in N_i^r}\frac{w_{ij}}{c_{i,r}}W_r^1l_j+w_{ii}W_0^1l_i),i=1,2,\cdots,N \tag{9}$$

其中σ表示非线性激活函数，W_r^1和W_0^1表示权重矩阵，N_i^r表示节点l_i在关系$r\in R$下的邻居节点集合，$c_{i,r}$表示归一化常量，通常取值为$|N_i^r|$，即节点l_i的邻居节点个数。

对于第2层RGAT，是在第1层的基础上采用相同的特征转化方法，将h_i转化为特征向量o_i，具体过程如下：

$$o_i=\sigma(\sum_{j\in N_i^r}W^2h_j+W_0^2h_i),i=1,2,\cdots,N \tag{10}$$

其中W^2和W_0^2表示权重矩阵，p表示历史话语的窗口大小，f 表示将来话语的窗口大小。

通过式(9)和式(10)使得模型能有效聚合各节点的邻居节点信息，从而获取说话者之间的长距离依赖关系，即说话者级别的特征可表示为：

$$O=[o_1,o_2,\cdots,o_N]\in\mathbb{R}^{N\times d_s} \tag{11}$$

其中，d_s表示说话者级别的特征维度，即RGAT的隐藏单元数。

例如：假设一个具有6个话语的对话l_1,l_2,l_3,l_4,l_5,l_6，其中l_1,l_3,l_4为说话者s_1所说，即$s_1=p(l_1)=p(l_3)=p(l_4)$；l_2,l_5,l_6为说话者s_2所说，即$s_2=p(l_2)=p(l_5)=p(l_6)$，设$p=2$，$f=2$，则每个说话者的话语之间的关系如表1所示。

3.6 情绪分类层

先对话语级别的特征和说话者级别的特征进行拼接，由式(6)、(11)可得：

$$G^*=[L|O]\in\mathbb{R}^{N\times(d_l+d_s)} \tag{12}$$

然后，采用BL设计负面情绪分类器以对G^*进行分类，并预测每个话语的情绪。负面情绪分类器的设计过程如下。将上面所提取的特征G^*进行线性映射以生成BL的多组特征节点，即对G^*进行线性映射操作，生成k组特征节点；第i组特征节点表示如下：

$$Z_i=\varphi(G^*W_{ei}+\beta_{ei})\in\mathbb{R}^{N\times q},i=1,...,k \tag{13}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

表 1. 每个说话者的话语之间的关系

关系类型	e_{ij}	说话者	i与j关系
5	e_{11}, e_{33}, e_{44}	s_1, s_1	$i = j$
1	e_{13}, e_{34}	s_1, s_1	$i < j$
2	e_{31}, e_{43}	s_1, s_1	$i > j$
5	e_{22}, e_{55}, e_{66}	s_2, s_2	$i = j$
2	e_{56}	s_2, s_2	$i < j$
1	e_{65}	s_2, s_2	$i > j$
3	$e_{12}, e_{35}, e_{45}, e_{46}$	s_1, s_2	$i < j$
4	e_{32}, e_{42}	s_1, s_2	$i > j$
3	e_{23}, e_{24}	s_2, s_1	$i < j$
4	$e_{21}, e_{53}, e_{54}, e_{64}$	s_2, s_1	$i > j$

其中φ表示线性激活函数，W_{ei}表示随机生成的权重矩阵且$W_{ei} \in \mathbb{R}^{(d_l+d_s)\times q}$，$\beta_{ei}$表示随机生成的偏置矩阵且$\beta_{ei} \in \mathbb{R}^{N\times q}$，$q$表示每组特征节点的数量，$k$表示特征节点的组数。

因此，k组特征节点可表示为$Z^k = [Z_1, Z_2, \cdots Z_k] \in \mathbb{R}^{N\times kq}$。将$Z^k$进行非线性映射以生成BL的多组增强节点，即对$Z^k$进行非线性映射操作，生成$m$组增强节点；第$j$组增强节点表示如下：

$$H_j = \xi(Z^k W_{hj} + \beta_{hj}) \in \mathbb{R}^{N\times r}, j = 1, ..., m \tag{14}$$

其中ξ表示非线性激活函数，W_{hj}表示随机生成的权重矩阵且$W_{hj} \in \mathbb{R}^{kq\times r}$，$\beta_{hj} \in \mathbb{R}^{N\times r}$表示随机生成的偏置矩阵且，$kq$表示所有特征节点的数量，$r$表示每组增强节点的数量。

因此，m组增强节点可表示为$H^m = [H_1, H_2, \cdots, H_m] \in \mathbb{R}^{N\times mr}$。将$k$组特征节点和$m$组增强节点进行拼接，可得$A = [Z^k | H^m] \in \mathbb{R}^{N\times(kq+mr)}$，再通过$A$来计算输出层的权重$W$。根据$Y = AW$，有$W = A^+Y$，由于$A$在大多数情况下都不是方阵，因此，可用$A^+$表示$A$的广义逆矩阵。为了更快速地计算$W$，同时增强模型的泛化能力，可采用岭回归算法[23]来求解W，具体表示如下：

$$\underset{W}{\operatorname{argmin}} \left(\left\| \left[Y - \hat{Y} \right] \right\|_2^2 + \lambda \left\| W \right\|_2^2 \right) \tag{15}$$

其中λ表示正则化系数，$\hat{Y} \in \mathbb{R}^{N\times(kq+mr)}$表示BL的近似输出。

W的全局最优解可表示为：

$$W = (\lambda I + A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} Y \tag{16}$$

4 实验及分析

将本文所提出的方法在三个对话文本的数据集上与基线模型进行对比实验。本次实验所采用的设备是一台搭载NVIDIA RTX 8000 48G显卡的Dell服务器。

4.1 评价指标

先采用weighted-$F1$值作为评价指标来比较各种方法在所有情绪标签上的性能；然后，采用macro-$F1$值作为评价指标来比较各种方法在负面情绪识别上的性能。weighted-$F1$、macro-$F1$具体的表示如下：

$$\text{weighted}-F1 = \frac{\sum_{i=1}^{C} (N_i \times F1_i)}{C} \tag{17}$$

$$\text{macro}-F1 = \frac{\sum_{i=1}^{C_{neg}} F1_i}{C_{neg}} \tag{18}$$

$$F1_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, i = 1, 2, \cdots, C \tag{19}$$

其中，P_i和R_i分别表示第i类情绪的精确率和召回率，$F1_i$表示第i类情绪的$F1$值，N_i表示包含第i类情绪的样本数，C表示所有情绪的类别数，C_{neg}表示负面情绪的类别数。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 数据集

为了验证所提出的模型有效性，在IEMOCAP (Busso et al., 2008)、MELD (Poria et al., 2019)和EmoryNLP (Zahiri and Choi, 2018)三个公开数据集上进行了实验。每个数据集的具体描述如下：

IEMOCAP：该数据集一个多模态数据集，包括文本、音频和视频，本文使用该数据集中的文本，该数据集包含六类情绪标签，分别是中性、开心、伤心、愤怒、沮丧和激动，其中负面情绪有伤心、愤怒和沮丧。

MELD：该数据集从美剧《老友记》中收集，包含七类情绪标签，分别是中性、开心、惊讶、恐惧、伤心、厌恶和愤怒，其中负面情绪有恐惧、伤心、厌恶和愤怒。

EmoryNLP：该数据集同样从美剧《老友记》中收集，包含七类情绪标签，分别是中性、开心、恐惧、愤怒、伤心、强烈和平静，其中负面情绪有恐惧、愤怒和伤心。

数据集的统计情况如表2所示。

表 2. 数据集的统计情况

数据集	对话数(训练/验证/测试)	话语数(训练/验证/测试)	类别数
IEMOCAP	100/20/31	4810/1000/1523	6
MELD	1038/114/280	9989/1109/2610	7
EmoryNLP	659/89/79	7551/954/984	7

4.3 基线模型

每种基线模型所采用的预训练模型都是RoBERTa，它们的具体描述如下：

SVM (Debnath et al., 2004)：SVM是一种传统机器学习模型，用于提取文本的特征。它使用的核是径向基函数，分类的策略是“一对多”；

TextCNN (Kim, 2014)：一种用于文本分类的CNN模型，用于提取文本的特征并对文本的负面情绪进行分类。卷积核的大小分别为3、4和5。核的维度为100；

Bi-LSTM (Schuster and Paliwal, 1997)：一种用于提取文本特征的双向LSTM。模型层数为2，每个层的隐藏单元数为32；

Bi-LSTM-ATTN (Zhou et al., 2016)：在Bi-LSTM模型的隐藏层的输出上增加了一个注意力层；

DialogueRNN (Majumder et al., 2019)：一种基于RNN的模型，包含三个GRU模块，其中两个GRU被用于记录说话者的状态和全局对话环境，另外一个GRU被用与对话过程中的情绪变化；

HiTrans (Li et al., 2020)：一种基于Transformer的对话情绪识别模型，利用Transformer提取序列特征的优良特性对序列上下文特征以及说话者情绪特征进行训练；

DialogXL (Shen et al., 2021a)：一种基于XLNet的对话情绪识别模型，采用XLNet对说话者自身以及说话者之间的依赖关系进行抽取，从而完成对话情绪识别任务；

DialogueGCN (Ghosal et al., 2019)：一种基于Bi-LSTM和GCN的对话情绪识别模型，采用Bi-LSTM对话语序列特征进行提取，用GCN对说话者级别特征进行提取。

4.4 参数设置

由于实验采用了三种数据集，在每种数据集上的实验参数稍微有所不同，具体情况如表3所示。

由于文本最大长度和LSTM隐层单元数对于话语级别特征的提取性能具有重要的影响，因此，在说话者级别上下文编码器中，包括说话者个数和话语窗口大小等参数。说话者个数可决定图中边的数量，其关系是$n_r = 2 \times 2^{n_s}$，n_s表示说话者个数，n_r表示边的数量。

4.5 实验性能对比

将本文所提出的RGAT-BL与基线模型进行对比，它们在三种数据集上的weighted-$F1$值对比结果如表4所示。

(c) 2022 中国中文信息学会计算语言学专业委员会

表 3. 参数设置

参数名称	IEMOCAP	MELD	EmoryNLP
文本最大长度	200	200	200
LSTM隐层单元数	150	150	150
说话者个数	2	12	10
话语窗口大小	10	4	6
RGAT隐层单元数	100	100	100
BL特征节点组数	10	10	10
BL特征节点个数/组	50	100	100
BL增强节点组数	10	10	10
BL特征节点个数/组	50	50	50
BL正则化参数	0.1	10	1
训练轮数	3	3	3

表 4. 不同模型在三种数据集上的weighted-$F1$值(%)

模型	IEMOCAP	MELD	EmoryNLP
SVM	36.12	48.75	24.32
TextCNN	47.65	54.14	32.46
Bi-LSTM	47.15	55.26	31.73
Bi-LSTM-ATTN	47.61	55.39	32.15
DialogueRNN	62.75	57.03	35.36
HiTrans	64.5	61.94	36.75
DialogXL	65.94	62.41	34.73
DialogueGCN	64.18	58.1	36.29
RGAT-BL	**66.13**	**64.83**	**37.94**

由表4可以看出，本文提出的RGAT-BL在数据集IEMOCAP、MELD和EmoryNLP上的性能均优于其它基线模型，weight-$F1$值分别为66.13%、64.83%和37.94%。其主要原因是RGAT-BL能有效地结合基于话语序列的模型和基于图神经网络的模型的优点，在话语之间的特征和说话者之间的特征提取上都发挥了重要作用，且基于BL的情绪分类器比传统的全连接层分类器能动态地计算情绪标签权重，从而获取更好的分类性能。

为了进一步验证模型对于对话负面情绪的识别性能，在三个数据集上对负面情绪的识别结果进行了对比实验，结果分别如表5、6、7所示。

由表5、6、7可知，本文提出的模型在三个数据集上针对负面情绪macro-$F1$值均优于其它基线模型。对于数据集IEMOCAP，“伤心”和“沮丧”的$F1$值相比其他基线模型更高，分别达到87.34%和67.20%；对于数据集MELD，“恐惧”、“伤心”和“厌恶”的$F1$值为最高，分别达到17.46%、36.71%和23.28%；对于数据集EmoryNLP，“愤怒”和“伤心”的$F1$值为最高，分别达到36.20%和26.06%。可见，该模型在负面情绪识别上同样表现良好。其主要原因是BL能将每个负面情绪类别通过特征节点和增强节点来得到合适的权重，从而确保了RGAT-BL在单个负面情绪类别上都能取得很高的$F1$值。

总之，与基线模型相比，本文所提出的模型在三个数据集上均有良好的情绪识别性能。采用Bi-LSTM与RGAT相结合的方式不仅可以提取话语的上下文特征，而且能充分考虑说话者之间的情绪交互和影响，从而能有效地对话语级别和说话者级别这两个级别的特征进行提取。在分类器方面，相比传统的全连接层分类器，基于BL的分类器具有三层网络架构以及岭回归优化的最小二乘法，也使得模型在分类性能上有所提升。

4.6 消融实验

为了进一步验证RGAT-BL的各部分的有效性，本文在三个数据集上分别进行了消融实

(c) 2022 中国中文信息学会计算语言学专业委员会

表 5. 不同模型在IEMOCAP数据集上负面情绪的识别性能

模型	$F1$值(%)			macro-$F1$ (%)
	伤心	愤怒	沮丧	
SVM	45.11	48.35	44.68	46.05
TextCNN	51.56	56.12	53.23	53.64
Bi-LSTM	52.98	55.45	57.61	55.35
Bi-LSTM-ATTN	53.5	57.19	58.74	56.48
DialogueRNN	78.8	65.28	58.91	67.66
HiTrans	80.23	**66.49**	60.16	68.96
DialogXL	77.10	61.59	64.67	67.79
DialogueGCN	84.54	64.19	66.99	71.91
RGAT-BL	**87.34**	62.54	**67.20**	**72.36**

表 6. 不同模型在MELD数据集上负面情绪的识别性能

模型	$F1$值(%)				macro-$F1$ (%)
	恐惧	伤心	厌恶	愤怒	
SVM	9.64	22.77	3.01	21.63	14.26
TextCNN	14.13	26.06	15.70	45.95	25.46
Bi-LSTM	13.56	24.15	16.06	43.44	27.88
Bi-LSTM-ATTN	13.81	25.35	16.19	44.01	24.84
DialogueRNN	9.62	34.08	13.16	41.87	24.68
HiTrans	16.84	35.03	20.45	**53.46**	31.45
DialogXL	10.32	33.16	9.33	49.93	25.69
DialogueGCN	9.71	34.76	19.35	42.71	26.63
RGAT-BL	**17.46**	**36.71**	**23.28**	52.13	**32.40**

表 7. 不同模型在EmoryNLP数据集上负面情绪的识别性能

模型	$F1$值(%)			macro-$F1$ (%)
	恐惧	愤怒	伤心	
SVM	22.16	16.01	7.37	15.18
TextCNN	29.41	25.13	11.53	22.02
Bi-LSTM	27.60	27.81	10.14	21.85
Bi-LSTM-ATTN	28.85	28.02	12.19	23.02
DialogueRNN	28.89	26.73	22.06	25.89
HiTrans	27.15	22.89	21.43	23.82
DialogXL	**37.38**	35.81	21.90	31.70
DialogueGCN	34.88	30.33	25.95	30.39
RGAT-BL	33.67	**36.20**	**26.06**	**31.98**

CCL 2022

表 8. RGAT-BL在三个数据集上的消融实验

模型	macro-F1 (%)		
	IEMOCAP	MELD	EmoryNLP
RGAT-BL	**72.36**	**32.40**	**31.98**
BL	54.73	26.29	23.67
RGAT	71.62	31.76	31.35
GCN-BL	70.55	29.67	30.15

验：BL、RGAT 和GCN-BL，其中GCN-BL将RGAT-BL中的RGAT替换为GCN。具体的实验结果如表8所示。

由表8可知，RGAT-BL在三个数据集上的macro-F1均比BL、RGAT和GCN-BL高。其中，RGAT-BL比BL分别高出17.90%、6.11%和8.31%，这表明RGAT能通过图形注意力网络有效地对话语之间的关系进行建模，使得RGAT-BL能更好地识别话语在上下文语境下的情绪；RGAT-BL比RGAT分别高出0.74%、0.64%和0.63%，这表明了BL能提高RGAT-BL的性能，主要原因是BL能通过特征节点和增强节点，进一步对话语级别的特征和说话者级别的特征提取深层语义信息；RGAT-BL比GCN-BL分别高出1.81%、2.73%和1.83%，这表明RGAT通过引入话语之间的注意力权重，比GCN能更好地刻画话语之间的影响程度。

5 结论

本文主要研究了RGAT，Bi-LSTM和BL对对话文本负面情绪识别的重要性。通过与现有的方法进行比较，发现RGAT和Bi-LSTM与BL的结合有利于完成对话文本负面情绪识别这一任务。该方法通过结合深度学习和宽度学习的优点，旨在提供一种更直观的方法来提取话语中的局部上下文信息（即话语级别），以及对话中的全局上下文信息（即说话者级别）。最后，在三个对话文本数据集上进行了大量的实验，结果表明，话语层面和说话者层面的语境都有利于负面情绪识别；同时在大多数测试数据集上，该方法在加权平均$F1$值上都优于基线模型。在未来的工作中，计划将所提出的方法和其他深度学习模型进行结合，以更有效地对对话文本中的负面情绪进行识别。

参考文献

彭韬, 杨亮, 桑钟屹, 唐雨, and 林鸿飞. 2021. 基于异构二部图的对话情感分析. 中文信息学报, 35(11):135–142.

赖河蒨, 李玲俐, 胡婉玲, and 颜学明. 2022a. 一种基于层次化r-gcn的会话情绪识别方法. 计算机工程, 48(01):85–92.

郑诚, 陈杰, and 董春阳. 2022b. 结合图卷积的深层神经网络用于文本分类. 计算机工程与应用, 58(7):206–212.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

CL Philip Chen and Zhulin Liu. 2017. Broad learning system: An effective and efficient incremental learning system without the need for deep architecture. *IEEE transactions on neural networks and learning systems*, 29(1):10–24.

Yonghe Chu, Hongfei Lin, Liang Yang, Shichang Sun, Yufeng Diao, Changrong Min, Xiaochao Fan, and Chen Shen. 2021. Hyperspectral image classification with discriminative manifold broad learning system. *Neurocomputing*, 442:236–248.

Rameswar Debnath, Nogayama Takahide, and Haruhisa Takahashi. 2004. A decision based one-against-one method for multi-class support vector machine. *Pattern Analysis and Applications*, 7(2):164–175.

(c) 2022 中国中文信息学会计算语言学专业委员会

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. 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 154–164, Hong Kong, China.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. 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 397–406, Minneapolis, Minnesota.

Xiao Jin, Jianfei Yu, Zixiang Ding, Rui Xia, Xiangsheng Zhou, and Yaofeng Tu. 2020. Hierarchical multimodal transformer with localness and speaker aware attention for emotion recognition in conversations. In *Proceedings of the 9th Natural Language Processing and Chinese Computing International Conference*, pages 41–53.

Junwei Jin, Yanting Li, Tiejun Yang, Liang Zhao, Junwei Duan, and CL Philip Chen. 2021. Discriminative group-sparsity constrained broad learning system for visual recognition. *Information Sciences*, 576:800–818.

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.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Sancheng Peng, Guojun Wang, Yongmei Zhou, Cong Wan, Cong Wang, Shui Yu, and Jianwei Niu. 2019. An immunization framework for social networks through big data based influence modeling. *IEEE transactions on dependable and secure computing*, 16(6):984–995.

Sancheng Peng, Rong Zeng, Hongzhan Liu, Guanghao Chen, Ruihuan Wu, Aimin Yang, and Shui Yu. 2021. Emotion classification of text based on bert and broad learning system. In *Proceeding of the Asia Pacific Web and Web-age Information Management Joint International Conference on Web and Big Data*, pages 382–396.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Proceedings of 16th International Conference on Data Mining*, pages 439–448.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 527–536.

Peng Sancheng, Lihong Cao, Yongmei Zhou, Zhouhao Ouyang, Aimin Yang, Xinguang Li, Weijia Jia, and Shui Yu. 2021. A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*.

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*, pages 593–607.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

(c) 2022 中国中文信息学会计算语言学专业委员会

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the 35th Association for the Advancement of Artificial Intelligence*, volume 35, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1551–1560, Online.

Han X, Li B, and Wang Z. 2020. An attention-based neural framework for uncertainty identification on social media texts. *Tsinghua Science and Technology*, 251:117–126.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence*, pages 44–52.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5415–5421.

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.

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*, pages 207–212.

CCL 2022

第二十一届中国计算语言学大会论文集，第485页-第496页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

基于知识迁移的情感-原因对抽取

赵凤园[1,2] ， 刘德喜[1,2] ， 万齐智[1,2] ， 万常选[1,2] ， 刘喜平[1,2] ， 廖国琼[1,2]
[1] 江西财经大学信息管理学院/ 江西南昌
[2] 江西财经大学数据与知识工程江西省高校重点实验室/ 江西南昌
dexi.liu@163.com

摘要

现有的情感-原因对抽取模型均没有通过加入外部知识来提升情感-原因对的抽取效果。本文提出基于知识迁移的情感-原因对抽取模型(ECPE-KT)，采用知识库获取文本的显性知识编码；随后引入外部情感分类语料库迁移得到子句的隐性知识编码；最后拼接两个知识编码，加入情感(原因)子句预测概率及相对位置，搭配Transformer机制融合上下文，并采用窗口机制优化计算压力，实现情感-原因对抽取。在ECPE数据集上的实验结果显示，本文提出的方法超过当前最先进的模型ECPE-2D。

关键词： 情感-原因对抽取 ； 知识辅助 ； 相对位置 ； 子句预测概率

Emotion-Cause Pair Extraction Based on Knowledge-Transfer

Zhao Fengyuan[1,2] ,Liu Dexi[1,2] ,Wan Qizhi[1,2] ,Wan Changxuan[1,2] ,Liu Xiping[1,2] ,Liao Guoqiong[1,2]
[1] School of Information Management, Jiangxi University of Finance and Economics / Jiangxi, Nanchang
[2] Jiangxi University of Finance and Economics, Key Laboratory of Data and Knowledge Engineering / Jiangxi, Nanchang
dexi.liu@163.com

Abstract

The existing emotion cause pair extraction models do not improve the performance of emotion cause pair extraction by incorporating external knowledge. In this work, we propose an emotion-cause pair extraction model based on knowledge transfer (ECPE-KT), in which knowledge is utilized to obtain explicit knowledge encoding of text. Subsequently, the implicit knowledge encoding of clauses is obtained by the transfer of external sentiment classification corpus. Moreover, conduct filtering via adding the relative position and prediction probability to the representation of clause semantic features. Transformer with window mechanism is used to optimize the calculation pressure. Experimental results show that the proposed method outperforms the state-of-the-art method, i.e., ECPE-2D.

Keywords: Emotion-cause pair extraction , Knowledge-assisted , relative position , prediction probability

1 引言

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目： 国家自然科学基金项目(No.61972184,62076112)、江西省自然科学基金重点项目(No.20212ACB202002)、江西省主要学科学术和技术带头人培养计划领军人才项目(No.20213BCJL22041)

现有的情感原因抽取工作主要分为情感原因抽取(emotion cause extraction, ECE)(Lee et al., 2010)和情感-原因对抽取(emotion-cause pair extraction, ECPE)(Xia and Ding, 2019)，前者是给定文本中的情感表达片段，抽取触发情感的原因片段，后者是同时抽取情感表达片段和情感原因片段。目前ECPE方法存在三方面的不足。首先，对于包含N个子句的文本，候选情感-原因对共有N^2对，因此识别的效率较低，不适合包含大量子句的长文本。其次，目前的模型尽管能通过候选情感子句和原因子句的相互作用，提升情感-原因对识别的效果，但也存在相互干扰的情况，并直接反映在实验结果上：在相同的ECPE数据集上，与单独抽取情感子句的模型相比，采用情感-原因联合抽取时，情感子句抽取的效果普遍明显下降；且在人工给定情感子句时，原因子句抽取的效果明显更优。第三，关于文本情感分析，目前有较多的人工知识(孙毅et al., 2021)可以帮助提升抽取效果，而触发情感的原因也多为事件(Turcan et al., 2021)，情感的主体多为人、组织、机构等实体，如图1 所示，这些特点还未见有模型充分利用。

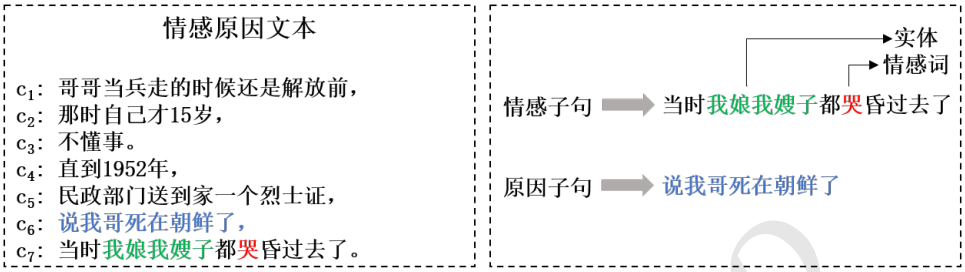


Figure 1: 情感原因文本示例。文本包含7个子句，c_1至c_7，c_7为情感子句，c_6是情感子句c_7对应的原因子句。其中，(1) 情感子句包含了较为明显的情感词“哭”，以及表达情感的主体“我娘我嫂子”；(2) 触发情感的原因是一个事件，如“我哥(主语)死(谓语)在朝鲜(状语)了(语气词)”。

基于以上分析，本文提出了基于知识迁移的情感-原因对抽取模型(emotion-cause pair extraction model based on knowledge-transfer, ECPE-KT)。ECPE-KT首先引入知识库对文本进行显性知识的编码以加强心理特征、情感词、实体和事件的识别；然后引入外部情感分类语料库，构建简易情感分类器训练子句编码，保存模型迁移得到ECPE数据集中子句的隐性知识编码；最后结合两个知识编码，加入情感(原因)子句预测概率及相对位置，搭配Transformer机制融合子句上下文，并采用窗口机制优化计算压力，实现情感-原因对抽取果。

本文主要贡献：(1) 引入人工知识库，辅助编码文本中的情感词、实体和事件以获取有利于情感抽取的显性知识；(2) 引入外部情感分类数据集，获取文本中子句蕴含的隐性知识；(3) 在ECPE数据集上对模型进行了验证，较当前最先进的模型ECPE-2D提升2.74%。

本文的结构如下：第一节介绍情感-原因对抽取任务及其特点，以及目前相关研究存在的不足。第二节详细介绍情感-原因对抽取任务的演化过程，以及相关工作所提出的解决方案。第三节和第四节分别提出ECPE-KT模型并通过实验验证其性能。最后进行总结与展望。

2 相关工作

情感原因抽取任务(ECE)由Lee et al. (2010)提出，给定文本中的情感词及其所属情感类别，根据已知的情感词抽取原因事件。Chen et al. (2010)分析Lee构建的语料库后发现，85.75%的情感原因事件均在同一子句中，未曾跨越多个子句，从而认为子句是ECE任务中较恰当的单元，于是提出将情感原因抽取任务由词级别转换为子句级别。然而ECE任务需要标注出文本中的情感词或情感子句，限制了实际应用场景；先标注情感后抽取其原因的方法未考虑情感与原因之间的相互指示关系。因此，Xia and Ding (2019)将ECE任务扩展为ECPE任务，研究方法可被总结为(邱祥庆et al., 2022)：

(1) 2阶段法。Xia and Ding (2019)提出两阶段框架，先独立抽取情感和原因，再将其配对过滤。Yu et al. (2021a)提出互助型多任务模型，基于文献(Xia and Ding, 2019)，添加两个与原任务相同的辅助任务(情感子句和原因子句抽取)，建立情感与原因之间的双向关联，再采取自蒸馏方法来训练以提高准确性，降低误差传播。Sun et al. (2021)加入注意力网络分别独立询问上下文中的情感和原因以获得上下文语境语义，同时讨论了损失函数对误差传播的影响。

(2) 一体化法(Wu et al., 2020)。由于阶段一的召回结果直接影响阶段二的性能，Ding et al.

(c) 2022 中国中文信息学会计算语言学专业委员会

(2020a)提出Transformer 一体化方法，将文本中的所有子句都看作情感(原因)子句进行两两组合配对，并搭配窗口限制(98%的情感-原因对的距离不大于3)和十字路口策略优化计算，最终效果取得明显提升。Tang et al. (2020a)采用双仿射机制构建基于LSTM的分层网络，一体化地建模情感抽取和情感-原因对抽取之间的关系，并引入多注意力机制加强子句间的情感感知。

(3) 基于图的方法。Wei et al. (2020a)利用图注意力学习子句表示，捕获子句间的潜在关系，并对子句对进行排序以抽取情感-原因对。Fan et al. (2020a)提出了基于状态转移的联合学习模型，将ECPE任务转换成一个类似解析的有向图构造过程，对输入序列从左到右逐步构造和标记有向边，并使用丰富的非局部特征对子句片段进行评估，使得模型具备同时识别出文本中的情感及原因的能力，有效缓解了误差传递问题。

(4) 局部搜索法(Wei et al., 2020b; Cheng et al., 2020a)。Ding et al. (2020b)提出基于滑动窗口的多标签联合学习模型。假设所有子句均为情感子句，以以其为中心句构建滑动窗口，抽取出中心句所对应的原因子句(CMLL)。同理，假设所有子句均为原因子句，抽取对应的情感子句(EMLL)。最后采用三种融合策略(平均概率，逻辑与，逻辑或)融合CMLL和EMLL的预测结果。Chen et al. (2022)采用多轮推理，迭代地检测情感原因和情感-原因对。

尽管ECPE任务取得了丰富的成果，但仍存在不足。端到端的方法大多采用一一配对的形式，产生大量候选情感-原因对，不适合包含大量子句的长文本；大量外部知识被证实可以辅助情感分析(谭红叶et al., 2020)，而深度学习模型中均未融入。

3 ECPE-KT模型

给定一个包含多个子句的文本$d = \{c_i, c_2, \cdots, c_{|d|}\}$，自动抽取出文本中的情感-原因对：

$$P = \{\cdots, (c^{e_1}, c^{c_1}), (c^{e_2}, c^{c_2}), \cdots\} \tag{1}$$

其中，(c^{e_1}, c^{c_1})表示文本d中的第i个情感-原因对，c^{e_i}表示情感子句，c^{c_i}是原因子句。d中至少存在一个情感子句，一个情感子句至少对应一个原因子句(目前已有的数据集基于该假设)。

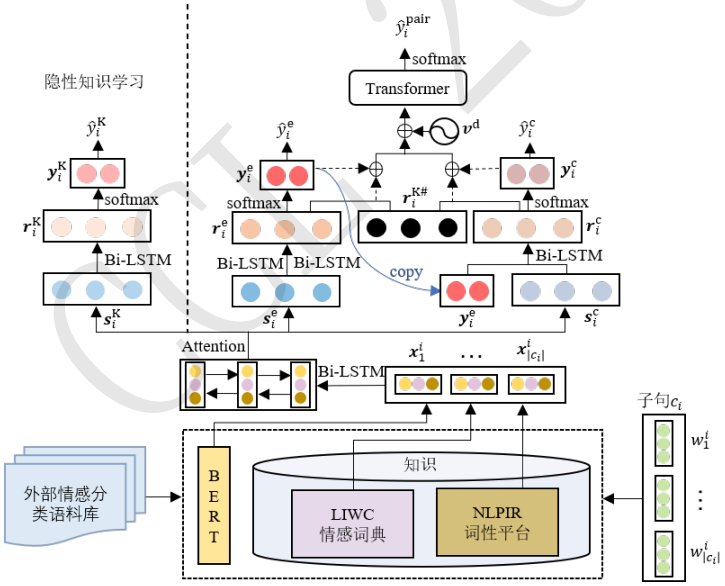


Figure 2: ECPE-KT模型结构图

本文基于Ding et al. (2020a)的ECPE-2D框架，融入显示知识(LIWC和NLPIR词性)，再结合同类语料库中包含的隐性知识，提出基于知识迁移的情感-原因对抽取模型，如图2所示。

3.1 知识辅助的子句表示

知识辅助的子句表示如图3所示。给定一个包含d个子句的文本$d = \{c_1, c_2, \cdots, c_{|d|}\}$，每一个子句$c_i = (w_1^i, w_2^i, \cdots, w_{|c_i|}^i)$分别包含$|c_i|$个词。每个词$w_j$的编码$\boldsymbol{x}_j$由三部分组成，分别是基于BERT的语义编码、基于LIWC语言心理特征知识库的词类编码和基于NLPIR的词性编码。

(c) 2022 中国中文信息学会计算语言学专业委员会

Input	说	我	哥	死	在	朝鲜	了
BERT Embedding	$w^{BERT}_{说}$	$w^{BERT}_{我}$	$w^{BERT}_{哥}$	$w^{BERT}_{死}$	$w^{BERT}_{在}$	$w^{BERT}_{朝鲜}$	$w^{BERT}_{了}$
	+	+	+	+	+	+	+
LIWC Embedding	$w^{LIWC}_{说}$	$w^{LIWC}_{我}$	$w^{LIWC}_{哥}$	$w^{LIWC}_{死}$	$w^{LIWC}_{在}$	$w^{LIWC}_{朝鲜}$	$w^{LIWC}_{了}$
	+	+	+	+	+	+	+
NLPIR POS Embedding	$w^{NLPIR}_{说}$	$w^{NLPIR}_{我}$	$w^{NLPIR}_{哥}$	$w^{NLPIR}_{死}$	$w^{NLPIR}_{在}$	$w^{NLPIR}_{朝鲜}$	$w^{NLPIR}_{了}$
Word Embedding	$\boldsymbol{x}_{说}$	$\boldsymbol{x}_{我}$	$\boldsymbol{x}_{哥}$	$\boldsymbol{x}_{死}$	$\boldsymbol{x}_{在}$	$\boldsymbol{x}_{朝鲜}$	$\boldsymbol{x}_{了}$

Figure 3: 知识辅助的子句表示

3.1.1 基于BERT的语义编码

BERT(Devlin et al., 2019)是一个强大的预训练模型，作为词嵌入效果显著，ECPE-KT采用BERT BASE对每个词w_j进行编码，得到768维的词向量表示$\boldsymbol{w}_j^{\text{BERT}}$。

3.1.2 基于LIWC的词类编码

情感子句是用户所表达出来的带有情感色彩的子句，相较于非情感子句，其富含了表达用户情感或心理变化的词类。故此，ECPE-KT利用LIWC知识库对词语类别进行标注，便于侧重学习心理知识，从而辅助情感子句的抽取。本文利用黄金兰et al. (2012)构建的SC-LIWC词典，采用one-hot对词语w_j进行编码，得到一个71维的向量$\boldsymbol{w}_j^{\text{LIWC}}$。如词语“不幸”在SC-LIWC词典中属于功能词、否定词、情感历程词、负向情绪词和悲伤词等5种词类，根据one-hot编码可得到“不幸”一词的词类编码结果为71维向量：$w^{\text{LIWC}}_{不幸}$=[1,0,$\cdots$,1,0,$\cdots$,1,0,1,0,0,1,0,. . . ,0]。

3.1.3 基于NLPIR的词性编码

情感主体来自实体，识别子句中的人名、人称代词等特征有助于辅助情感子句的抽取。NLPIR 分析平台(张华平and 商建云, 2019) 能够较为精准地识别出文本中的人名、机构名、地名和代词等，因此本文采用该平台对语料进行词性标注。

为避免过多的词性种类导致维度的稀疏，ECPE-KT保留{人名nr、地名ns、其他名词n、形容词a、副词d、动词v、人称代词rr、其他代词r}共8种词性，将其余词性统一合并为其他词性other。本文采用one-hot对子句中词语w_j的词性进行编码，得到一个9维的向量表示$\boldsymbol{w}_j^{\text{NLPIR}}$。

ECPE-KT模型中候选子句的词语w_j编码是$\boldsymbol{w}_j^{\text{BERT}}$、$\boldsymbol{w}_j^{\text{LIWC}}$和$\boldsymbol{w}_j^{\text{NLPIR}}$的拼接，表示为：

$$\boldsymbol{x}_j = [\boldsymbol{w}_j^{\text{BERT}} \oplus \boldsymbol{w}_j^{\text{LIWC}} \oplus \boldsymbol{w}_j^{\text{NLPIR}}] \tag{2}$$

3.2 隐性知识学习

本文采用NLPCC2013中文微博情绪分析评测任务中的数据集作为外部情感分类语料库，该数据集由姚源林et al. (2014)构建，其描述如表1所示。

Table 1: NLPCC2013数据集描述

统计粒度		有情绪	无情绪	总计	统计粒度		有情绪	无情绪	总计
微博级	数量	7407	6593	14000	句子级	数量	15688	29733	45421
	占比	52.91%	47.09%	100%		占比	34.54%	65.46%	100%

3.2.1 数据预处理

NLPCC2013数据集由微博组成，具有对话型回复、文本短、包含表情符等特点。微博在转发、回复等模式下会自动生成“//@微博昵称：”、“回复@微博昵称：”等与内容无关的格式化文本，这类文本不仅无益于语义理解，还可能造成歧义，如“回复@幸运的兰妹妹”，可能会使模型认为子句包含了情感——“幸运”。因此本文利用正则表达式删除了这类格式化文本。

除了格式化文本之外，微博用户常使用一些表情符，如：⊙o⊙、TˆT、ˆ_ˆ。尽管表情符具有一定的情感表达作用，但是ECPE数据集中不存在这类符号，且其多为字符拼接而成，在进行分词操作时存在字符被逐一切分的情况。因此将NLPCC2013数据集中的表情符也统一删除。

3.2.2 情感识别的预热模型

给定情感分类语料库中的第k个文本d^k，它包含$|d^k|$个子句，$d^k = \{c_1, c_2, \cdots, c_{|d^k|}\}$，每一个子句$c_i = (w_1^i, w_2^i, \cdots, w_{|c_i|}^i)$分别包含$|c_i|$个词，$d^k$中每个子句$c_i$的情感标签为$\boldsymbol{y}_i^{\text{kno}}$。根据3.1节获得每个词$w_j$的编码表示$\boldsymbol{x}_j$。

情感识别的预热模型采用词语层和子句层结合的双层Bi-LSTM(Graves et al., 2013)：

(1) 词语层Bi-LSTM

将一个包含$|c_i|$个词语的子句$c_i = (w_1^i, w_2^i, \cdots, w_{|c_i|}^i)$的编码表示$(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \cdots, \boldsymbol{x}_{|c_i|}^i)$作为输入，送入Bi-LSTM中，得到子句$c_i$中第$j$个词语的隐层表示$\boldsymbol{h}_j^i = \text{Bi-STM}(\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \cdots, \boldsymbol{x}_{|c_i|}^i) \in \mathbb{R}^{2d_h}$，其中$d_h$是Bi-LSTM的隐藏单元个数。

再对每个词语采用自注意力机制(Luong et al., 2015)得到子句c_i的编码表示s_i^{K}：

$$\alpha_{ij}^{\text{K}} = \boldsymbol{W}_{\alpha}^{\text{T}} tanh(\boldsymbol{W}_h \boldsymbol{h}_j^i + \boldsymbol{b}_h); \quad \beta_{ij}^{\text{K}} = \exp(\alpha_{ij}^{\text{K}}) / \sum_{j=1}^{|c_i|} \exp(\alpha_{\text{ij}}^{\text{K}}); \quad \mathbf{s}_i^{\text{K}} = \sum_{j=1}^{|c_i|} \beta_{ij}^{\text{K}} \boldsymbol{h}_j^i \tag{3}$$

其中$\boldsymbol{W}_{\alpha} \in \mathbb{R}^{2d_h}$，$\boldsymbol{W}_h \in \mathbb{R}^{2d_h \times 2d_h}$和$\boldsymbol{b}_h \in \mathbb{R}^{2d_h}$可学习的参数。

(2) 子句层Bi-LSTM

子句层Bi-LSTM的目的是捕捉子句间的语义依赖。对于包含$|d^k|$个子句的文本$d^k = \{c_1, c_2, \cdots, c_i, \cdots, c_{|d^k|}\}$，将各个子句的编码$(\boldsymbol{s}_1^{\text{K}}, \boldsymbol{s}_2^{\text{K}}, \cdots, \boldsymbol{s}_i^{\text{K}}, \cdots, \mathbf{s}_{|d^k|}^{\text{K}}) \in \mathbb{R}^{|d^k| \times 2d_h}$送入Bi-LSTM，得到Bi-LSTM的隐藏状态，即子句c_i的上下文表示$\boldsymbol{r}_i^{\text{K}} = \text{Bi-LSTM}(\boldsymbol{s}_1^{\text{K}}, \boldsymbol{s}_2^{\text{K}}, \cdots, \boldsymbol{s}_i^{\text{K}}, \cdots, \boldsymbol{s}_{|d^k|}^{\text{K}}) \in \mathbb{R}^{2d_h}$。

最后将$\boldsymbol{r}_i^{\text{K}}$送入softmax函数，得到外部情感分类语料库子句c_i是情感子句的概率$\boldsymbol{y}_i^{\text{K}} = \text{softmax}(\boldsymbol{W}^{\text{K}} \boldsymbol{r}_i^{\text{K}} + \boldsymbol{b}^{\text{K}})$，其中$\boldsymbol{W}^{\text{K}} \in \mathbb{R}^{d_m \times 2d_h}$，$\boldsymbol{b}^{\text{K}} \in \mathbb{R}^{d_m}$，$d_m$表示类别数。

为了更好地学习子句的隐性知识表示，模型引入了损失函数$\boldsymbol{L}^{\text{K}} = -\sum_{i=1}^{|d|} \boldsymbol{y}_i^{\text{kno}} \cdot \log(\boldsymbol{y}_i^{\text{K}})$，其中$\boldsymbol{y}_i^{\text{kno}}$表示外部情感分类语料库子句$c_i$作为情感子句的真实分布。

在NLPCC2013数据集上训练后得到情感识别的预热模型。对于ECPE数据集中给定的子句c_i，可得到其隐状态，代表使用外部数据推断出来的情感隐性知识，记为$\boldsymbol{r}_i^{\text{K}\#}$。

3.3 情感信息编码

对于ECPE数据集，给定一个包含$|d|$个子句的文本$d = \{c_1, c_2, \cdots, c_{|d|}\}$，每一个子句$c_i = (w_1^i, w_2^i, \cdots, w_{|c_i|}^i)$分别包含$|c_i|$个词。根据3.1节得到每个词$w_j$的显性知识编码表示$\boldsymbol{x}_j$，再将其送入3.2.2节的双层Bi-LSTM中，得到子句c_i的上下文表示$\boldsymbol{r}_i^{\text{e}} \in \mathbb{R}^{2d_h}$，最后将$\boldsymbol{r}_i^{\text{e}}$送入softmax函数，得到子句$c_i$是情感子句的概率$\boldsymbol{y}_i^{\text{e}} = \text{softmax}(\boldsymbol{W}^{\text{e}} \boldsymbol{r}_{\text{i}}^{\text{e}} + \boldsymbol{b}^{\text{e}})$，其中$\boldsymbol{W}^{\text{e}} \in \mathbb{R}^{d_m \times 2d_h}$，$\boldsymbol{b}^{\text{e}} \in \mathbb{R}^{d_m}$，$d_m$表示类别数。

3.4 原因信息编码

原因信息的编码也采用词语层和子句层结合的双层Bi-LSTM，与3.1节和3.2.2节相同。将子句c_i的编码表示$\boldsymbol{s}_i^{\text{c}}$和情感信息编码阶段得到的情感预测概率值$\boldsymbol{y}_i^{\text{e}}$进行拼接，得到子句$c_i$的编码表示$(\boldsymbol{s}_i^{\text{c}} \oplus \boldsymbol{y}_i^{\text{e}})$。为捕捉上下文信息，将文本$d$中$|d|$个子句的向量表示$(\boldsymbol{s}_1^{\text{c}} \oplus \boldsymbol{y}_1^{\text{e}}, \cdots, \boldsymbol{s}_i^{\text{c}} \oplus \boldsymbol{y}_i^{\text{e}}, \cdots, \boldsymbol{s}_{|d|}^{\text{c}} \oplus \boldsymbol{y}_{|d|}^{\text{e}})$作为Bi-LSTM的输入，得到Bi-LSTM的隐藏状态，即子句c_i的上下文表示$\boldsymbol{r}_i^{\text{c}} \in \mathbb{R}^{2d_h}$：

$$\boldsymbol{r}_i^{\text{c}} = \text{Bi-LSTM}(\boldsymbol{s}_1^{\text{c}} \oplus \boldsymbol{y}_1^{\text{e}}, \cdots, \boldsymbol{s}_i^{\text{c}} \oplus \boldsymbol{y}_i^{\text{e}}, \cdots, \boldsymbol{s}_{|d|}^{\text{c}} \oplus \boldsymbol{y}_{|d|}^{\text{e}}) \tag{4}$$

最后将$\boldsymbol{r}_i^{\text{c}}$送入softmax函数中，得到子句$c_i$的原因预测概率值$\boldsymbol{y}_i^{\text{c}} = \text{softmax}(\boldsymbol{W}^{\text{c}} \boldsymbol{r}_{\text{i}}^{\text{c}} + \boldsymbol{b}^{\text{c}})$，其中$\boldsymbol{W}^{\text{c}} \in \mathbb{R}^{d_m \times 2d_h}$，$\boldsymbol{b}^{\text{c}} \in \mathbb{R}^{d_m}$。

3.5 情感-原因对抽取

3.5.1 情感-原因配对

为避免误差传递，本文选择将所有子句进行一一配对。将文本d中的子句配对后得到形如$|d| * |d|$的矩阵$\boldsymbol{M}$，$M_{i,j}$表示候选情感子句c_i和候选原因子句c_j的配对结果。融合情感子句上

(c) 2022 中国中文信息学会计算语言学专业委员会

下文表示$\boldsymbol{r}_i^{\mathrm{e}}$和其隐性知识表示$\boldsymbol{r}_i^{\mathrm{K\#}}$，得到候选情感子句$c_i$的隐性知识表示$\boldsymbol{r}_i^{\mathrm{Ke}} = \boldsymbol{r}_i^{\mathrm{e}} \oplus \boldsymbol{r}_i^{\mathrm{K\#}}$；同理，得到候选原因子句$c_j$的隐性知识表示$\boldsymbol{r}_j^{\mathrm{Kc}} = \boldsymbol{r}_j^{\mathrm{c}} \oplus \boldsymbol{r}_j^{\mathrm{K\#}}$。由此情感-原因配对被编码为$\boldsymbol{M}_{i,j}$:

$$\boldsymbol{M}_{i,j} = \boldsymbol{r}_i^{\mathrm{Ke}} \oplus \boldsymbol{y}_i^{\mathrm{e}} \oplus \boldsymbol{r}_j^{\mathrm{Kc}} \oplus \boldsymbol{y}_j^{\mathrm{c}} \oplus \boldsymbol{v}^d \tag{5}$$

其中，$\boldsymbol{y}_i^{\mathrm{e}}$ 是情感预测概率，$\boldsymbol{y}_j^{\mathrm{c}}$ 是原因预测概率，$\boldsymbol{v}^d$为子句c_i和子句c_j的相对距离。

3.5.2 情感-原因交互

考虑到一个长度为$|d|$的文本将会生成$|d| * |d|$个可能的情感-原因对矩阵$\boldsymbol{M}$，然而其中仅有一小部分具有因果关系；且当一个子句被认为是情感(原因)子句，其上下文中的其他子句成为情感(原因)子句的概率将会减小。因此本文采用Transformer(Vaswani et al., 2017)对矩阵$\boldsymbol{M}$进行全局信息融合以实现情感和原因信息之间的有效交互。

(1) 标准的Transformer

标准的Transformer由$N(N=6)$层encoding和decoding堆叠而成，每层由多头自注意力机制和前馈网络两个组件组成。

多头自注意力机制首先计算文本d中候选情感-原因对$c_i^e - c_j^c$的查询向量$\boldsymbol{q}_{i,j}$，关键字向量$\boldsymbol{k}_{i,j}$和值向量$\boldsymbol{v}_{i,j}$:

$$\boldsymbol{q}_{i,j} = \mathrm{Relu}(\boldsymbol{M}_{i,j}\,\boldsymbol{W}_Q); \quad \boldsymbol{k}_{i,j} = \mathrm{Relu}(\boldsymbol{M}_{i,j}\,\boldsymbol{W}_K); \quad \boldsymbol{v}_{i,j} = \mathrm{Relu}(\boldsymbol{M}_{i,j}\,\boldsymbol{W}_V) \tag{6}$$

其中$\boldsymbol{W}_Q \in \mathbb{R}^{n\times n}$，$\boldsymbol{W}_K \in \mathbb{R}^{n\times n}$，$\boldsymbol{W}_V \in \mathbb{R}^{n\times n}$分别是查询、关键字和值向量的可训练参数。

通过模型的学习优化，可以得到矩阵$\boldsymbol{M}$中每个候选情感-原因对$\boldsymbol{M}_{i,j}$的新特征表示$\hat{\boldsymbol{Z}}_{i,j}$:

$$\beta_{i,j,a,b} = \frac{\exp\left(\frac{\boldsymbol{q}_{i,j}\cdot\boldsymbol{k}_{a,b}}{\sqrt{n}}\right)}{\sum a' \sum b' \exp\left(\frac{\boldsymbol{q}_{i,j}\cdot\boldsymbol{k}_{a',b'}}{\sqrt{n}}\right)}; \quad \hat{\boldsymbol{z}}_{i,j} = \sum_{a=1}^{|d|}\sum_{b=1}^{|d|} \beta_{i,j,a,b}\cdot\boldsymbol{v}_{a,b} \tag{7}$$

在使用自注意力机制之后，将前馈网络应用到每个候选配对上:

$$\hat{\boldsymbol{o}}_{i,j} = \max(0, \hat{\boldsymbol{Z}}_{i,j}\,\boldsymbol{W}_1 + \boldsymbol{b}_1)\,\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{8}$$

使用残差连接和归一化操作，得到多头自注意力层输出$\boldsymbol{z}_{i,j}$，前馈网络层输出$\boldsymbol{o}_{i,j}$:

$$\boldsymbol{z}_{i,j} = \mathrm{Normalize}(\hat{\boldsymbol{z}}_{i,j} + \boldsymbol{M}_{i,j}); \quad \boldsymbol{o}_{i,j} = \mathrm{Normalize}(\hat{\boldsymbol{o}}_{i,j} + \boldsymbol{z}_{i,j}) \tag{9}$$

记l为每一层的索引，上一层的输出将被作为下一层的输入，由N层堆叠的Transformer将被表示为：$\boldsymbol{M}_{i,j}^{(l+1)} = \boldsymbol{o}_{i,j}^{(l)}$。

(2) 基于窗口的Transformer

由于配对矩阵$\boldsymbol{M}$中有$|d| * |d|$个元素($\boldsymbol{M}_{i,j}$)，每个$\boldsymbol{M}_{i,j}$需要计算$|d| * |d|$个注意力权重，最终需要计算和保存$(|d| * |d|) * (|d| * |d|)$个权重信息。为了减轻计算压力，同时基于对数据集的统计，仅1.85%的原因子句与情感子句之间的距离超过了4，本文提出基于窗口的Transformer对子句进行交互优化计算。

基于窗口的Transformer是将配对矩阵$\boldsymbol{M}$中的候选情感-原因对$c_i^e - c_j^c$限制在窗口内，即只采用下标符合j-i∈[-$window$,$window$]的候选情感-原因对作为Transformer的输入。值得一提的是，基于窗口的Transformer不仅大大节省了计算资源，还在一定程度上缓解了分类不均衡的问题，减少了$window$之外的负样本输入。

3.5.3 情感-原因对预测

将候选情感-原因对$\boldsymbol{M}_{i,j}$通过N层堆叠的Transformer后，得到每个候选情感-原因对的向量表示$\boldsymbol{o}_{i,j}^{(N)}$，将其送入softmax函数得到候选情感-原因对$c_i^e - c_j^c$是情感-原因的概率分布$y_i^{\mathrm{pair}} = \mathrm{softmax}(\boldsymbol{W}^{\mathrm{pair}}\boldsymbol{o}_{i,j}^{(N)} + \boldsymbol{b}^{\mathrm{pair}})$。

第二十一届中国计算语言学大会论文集，第497页-第509页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

文本d中情感-原因对分类器的损失$L^{\text{pair}} = -\sum_{i=1}^{|d|}\sum_{j=1}^{|d|} \boldsymbol{y}_i^{\text{pair}} \cdot \log(\boldsymbol{y}_i^{\text{pair}})$, 其中$\boldsymbol{y}_i^{\text{Pair}}$是文本$d$中候选情感-原因对$c_i^e - c_j^c$的真实概率分布。

为了更好的学习情感信息编码和原因信息编码，引入辅助任务的联合交叉熵损失$L^{\text{ec}} = -\sum_{i=1}^{|d|} \boldsymbol{y}_i^{\text{emo}} \cdot \log(\boldsymbol{y}_i^{\text{e}}) - \sum_{i=1}^{|d|} y_i^{\text{cau}} \cdot \log(\boldsymbol{y}_i^{\text{c}})$，其中$\boldsymbol{y}_i^{\text{emo}}$和$\boldsymbol{y}_i^{cau}$分别表示子句$c_i$作为情感子句和原因子句的真实概率分布。

ECPE-KT最终损失为情感-原因对抽取的损失L^{pair}和辅助任务的损失L^{ec}在L2正则化下的加权求和：$L = \lambda_1 L^{\text{pair}} + \lambda_2 L^{ec} + \lambda_3 ||\theta||^2$，其中$\lambda_1, \lambda_2, \lambda_3 \in (0,1)$是权重，$\theta$是模型的所有参数。

4 实验分析

4.1 数据集介绍

本文采用Xia and Ding (2019)提出ECPE任务时公开的情感原因数据集。其中，1746个文本(占比89.77%)仅含有一个情感-原因对，177个文本(占比9.10%)含有两个情感-原因对，22个文本(占比1.13%)含有两个以上情感-原因对，总计1945个文本。

4.2 实验设置

词向量为768维。词类编码向量为71维，词性编码向量为9维。Bi-LSTM的隐藏单元为100，注意力机制的查询、关键字和值向量设置为30维。所有的权重$\boldsymbol{W}$和偏置$\boldsymbol{b}$随机初始化为(-0.01,0.01)区间上的均匀分布。文本长度75，子句长度45。采用十折交叉验证，结合随机梯度下降SGD算法和Adam算法更新参数。样本批量大小和学习率分别为8和0.0005；词语层Bi-LSTM的dropout均为0.5；L2正则化中的权重λ_1，λ_2，λ_3分别设置为1，1，1e-5；位置向量维度为50维[0]。

4.3 对比模型

ECPE-2Steps：Xia and Ding (2019)首次提出ECPE任务时采用的模型。该模型先抽取情感子句和原因子句，再过滤情感-原因对。其中情感子句和原因子句的抽取分别采用了三种方法：(1) Indep：独立抽取情感子句和原因子句。(2) Inter-CE：将原因子句的预测分布作为特征，辅助抽取情感子句。(3) Inter-EC：将情感子句的预测分布作为特征，辅助抽取原因子句。**ECPE-2D**(Ding et al., 2020a)：基于窗口化Transformer的一体化方法。**RANKCP**(Wei et al., 2020a)：图注意力模型。**TDGC**(Fan et al., 2020a)：基于状态转移的联合学习模型。**MTNECP**(Yu et al., 2021a)：多任务神经网络模型。**ECPE-MLL**(Ding et al., 2020b)：基于滑动窗口的多标签联合学习模型。**DQAN**(Sun et al., 2021)：注意力网络分别独立询问上下文中的情感和原因以获得上下文语境语义。**RSNLSTM**(Chen et al., 2022)：采用多轮推理，迭代地检测情感原因和情感-原因对。**Inter-ECNC**(Shan and Zhu, 2020)：多头注意力网络模型。**MAM-SD**(Yu et al., 2021b)：相互辅助的多任务模型。**LAE-MANN**(Tang et al., 2020b)：基于双仿射机制的LSTM分层网络模型。**E2EECPE**(Song et al., 2020)：基于双向关注的定向预测模型。**SLSN-U**(Cheng et al., 2020b)：对称式局部搜索网络模型。**RHNSC**(Fan et al., 2020b)：端到端的分层神经网络模型。**IE-CNN+CRF**(Chen et al., 2020)：基于多类别情感标签的卷积神经网络模型。**ECPE-SL**(Yuan et al., 2020)：结合类型标签和距离标签的Bi-LSTM模型。**ECPE-KA**(刘德喜et al., 2021)：外部人工知识(LIWC词典和词性)的2阶段抽取模型。

4.4 实验结果分析

ECPE-KT的评测结果如表2所示。由表2可知，各个模型采用的方法主要分为“2阶段”法和端到端法，后者的性能明显优于前者。值得一提的是，综合使用了端到端模式和外部知识的ECPE-KT在ECPE任务上各个评测指标都达到最优，相较其他模型，分别在$F1$值上提升2.74%∼21.66%，在精确率P上提升2.03%∼16.15%，在召回率R上提升2.34%∼31.82%。

[0]https://github.com/Inkblue/ECPE-KT

(c) 2022 中国中文信息学会计算语言学专业委员会

Table 2: 实验评测结果

模型名称	emotion extraction			cause extraction			ECPE			方法
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	
Indep	.8375	.8071	.8210	.6902	.5673	.6205	.6832	.5082	.5818	2阶段笛卡儿积
Inter-CE	.8494	.8122	.8300	.6809	.5634	.6151	.6902	.5135	.5901	2阶段笛卡儿积
Inter-EC	.8364	.8107	.8230	.7041	.6083	.6507	.6721	.5705	.6128	2阶段笛卡儿积
Inter-ECNC	/	/	/	.6863	.6254	.6544	.6601	.5734	.6138	2阶段笛卡儿积
MAM-SD	.8554	.8141	.8339	.7202	.6375	.6751	.6963	.5799	.6320	2阶段笛卡儿积
TDGC	.8716	.8244	.8474	**.7562**	.6471	.6974	.7374	.6307	.6799	解析式转移系统
E2EECPE	.8595	.7915	.8238	.7062	.603	.6503	.6478	.6105	.6280	矩阵变换
MTNECP	.8662	.8393	.8520	.7400	.6378	.6844	.6828	.5894	.6321	矩阵变换
LAE-MANN	**.8990**	.8000	.8470	/	/	/	.7110	.6070	.6550	矩阵变换
ECPE-2D	.8627	**.9221**	**.8910**	.7336	.6934	**.7123**	.7292	.6544	.6889	矩阵变换
RHNSC	/	/	/	/	/	/	.6956	.5871	.6357	局部搜索
SLSN-U	.8406	.7980	.8181	.6992	.6588	.6778	.6836	.6291	.6545	局部搜索
RANKCP	.8703	.8406	.8548	.6927	.6743	.6824	.6698	.6546	.6610	局部搜索
ECPE-MLL	.8582	.8429	.8500	.7248	.6702	.6950	.7090	.6441	.6740	局部搜索
IE-CNN+CRF	.8614	.7811	.8188	.7348	.5841	.6496	.7149	.6279	.6686	序列标注
ECPE-SL	.8196	.7329	.7739	.7490	.6602	.7018	.7243	.6366	.6776	序列标注
ECPE-KA	.8549	.8958	.8746	.6760	**.7453**	.7083	.7354	.6531	.6914	外部知识
ECPE-KT	.8907	.8880	.8886	.7545	.6722	.7104	**.7524**	**.6699**	**.7078**	外部知识

相较于同样采用了外部知识的ECPE-KA模型，ECPE-KT 在ECPE任务的*F*1上提升了2.37%；在精确率*P*上提升了2.31%；在召回率*R*上提升了2.57%。由此说明隐性知识的加入使得模型更充分地理解了语义，对子句的判断更为精准。

相较于ECPE-KT的基准模型ECPE-2D，在ECPE任务上，ECPE-KT(*F*1=0.7078)取得优于ECPE-2D模型(*F*1=0.6889)的效果，提升了2.74%；在精确率*P*上提升了3.18%；在召回率*R*上提升了2.37%。精确率*P*的大幅提升进一步佐证了隐性知识对子句语义的正向作用。

特别地，对于含有2个或2个以上情感-原因对的测试用例，ECPE-KT抽取完整的用例占9.375%，至少抽取出1对的用例占62.5%，而Inter-EC识别出的比例分别为6.25%和为56.52%。可见，两者在多情感-原因对文本上的正确率都有待提高，但是ECPE-KT依然有更好的表现。

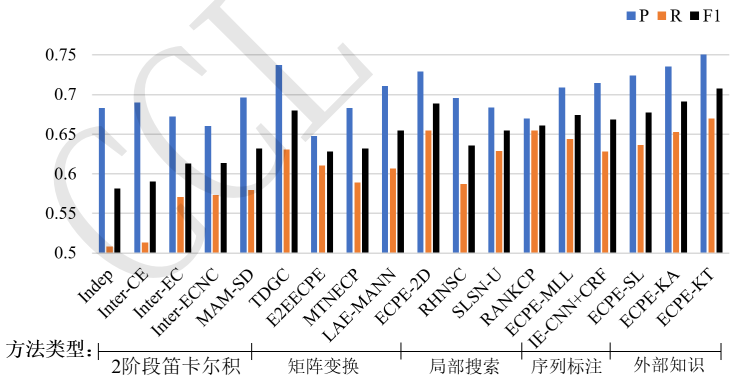


Figure 4: ECPE方法评测

图4(Figure 4)更直观地展示了不同模型的评测结果。由图4可知，采用2阶段笛卡尔积的模型，精确率*P*和召回率*R*差距较大，模型提升空间大；随着方法的改进，评测指标均不断上升，且精确率*P*和召回率*R*不断趋近。融合了知识和端到端的ECPE-KT性能显著，仅在2阶段方法中加入外部知识的ECPE-KA也表现良好，可见外部知识对模型的辅助功能是不可或缺的。

4.5 消融实验

4.5.1 知识辅助的影响

表3展示了知识辅助的影响。BERT的影响。去掉BERT的ECPE-KT-BERT相较ECPE-KT在情感子句抽取上的*F*1值大幅降低，但在原因子句抽取上*F*1值有些许提升，最终

Table 3: 知识辅助和交互的影响

模型名称	emotion extraction			cause extraction			ECPE		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
ECPE-2D#	.8511	.8237	.8365	.7133	.6285	.6672	.7118	.5984	.6494
ECPE-KT-BERT	.8684	.8683	.8679	.7526	.6821	**.7150**	.7481	.6539	.6959
ECPE-KT-know-70	.8730	.8808	.8766	.7388	.6434	.6852	.7192	.6500	.6814
ECPE-KT-know-50	.8799	.8793	.8794	.7318	.6601	.6925	.7164	.6622	.6869
ECPE-KT-inknow	.8867	.8725	.8793	.7513	.6800	.7121	**.7596**	.6434	.6950
ECPE-2D	.8627	**.9221**	**.8910**	.7336	**.6934**	.7123	.7292	.6544	.6889
ECPE-KT-Trans	.8896	0.8821	.8853	.7526	.6669	.7065	.7478	.6415	.6893
ECPE-KT	**.8907**	.8880	.8886	**.7545**	.6722	.7104	.7524	**.6699**	**.7078**

在ECPE上$F1$值仅存在1.71%的差距，甚至超越目前最优的ECPE-2D模型。由此说明BERT有利于模型对子句语义的理解，尤其是对事件的判定更为准确，最终使得子模型性能提升。

知识辅助的影响。去掉所有外部知识的ECPE-KT-know相较ECPE-KT在三个任务上的性能都有所下降，尤其是与ECPE-2D#采用相同参数的ECPE-KT-know-70模型，在ECPE上的$F1$值下降了3.87%。去掉隐性知识的ECPE-KT-inknow相较ECPE-KT在三个任务上同样都有所下降，在ECPE任务上的$F1$值存在1.84%的差距。将ECPE-KT-inknow与ECPE-KT-know-50进行比较，可发现在加入显性知识时，模型依旧能取得不错的性能($F1$=0.6950)，且优于先进的ECPE-2D模型。由此说明隐性知识和显性知识都能促进模型对子句的理解。

知识辅助与BERT的联合影响。加入了BERT和外部知识的ECPE-KT相较ECPE-2D#，各个指标都明显提升，且显著高于单独使用BERT 或外部知识的模型。由此佐证，BERT和知识的相互作用能更有效地提取情感和原因的特征，增强情感-原因对的抽取效果。

4.5.2 情感原因交互的影响

表3比较了使用交互的ECPE-KT和去掉交互的ECPE-KT-Trans的性能。加入交互组件Transformer后，情感-原因对抽取在各个指标上都有提升，精确率P提升0.62%，召回率R提升4.43%，$F1$值提升2.68%。由此可见，Transformer较好地对子句进行了交互，融合了子句间的信息。尤其是召回率R的大幅提升，意味着即使在正负样本不均衡的情况下，模型仍能较好地工作，抽取出尽可能多的正样本。

4.6 ECE任务评测

表4在ECE任务上将ECPE-KT与一些已有的方法进行对比，值得注意的是，经典的ECE任务中，情感子句是人工标注的，而ECPE-KT在测试集中不要求对情感子句进行人工标注。

Table 4: 情感原因抽取任务的评测

Method	***P***	***R***	***F1***	**Method**	***P***	***R***	***F1***
RB	.6747	.4267	.5243	CANN	**.7721**	.6891	**.7266**
CB	.2672	**.7130**	.3887	CANN-E	.4826	.3160	.3797
ECPE-1C	.4516	.4732	.4618	PAE-DGL	.7619	.6908	.7242
Multi-kernel	.6588	.6927	.6752	ECPE-KT	.7545	.6722	.7104

RB(Chen et al., 2010)：自定义语言规则的情感原因抽取方法。**CB**(Russo et al., 2011)：基于知识的情感原因抽取方法。**ECPE-1C**：已知情感子句，其前一句作为原因子句。**Multi-kernel**(Gui et al., 2016)：结合依存句法树的多核支持向量机模型。**CANN**(Li et al., 2018)：基于协同注意力的Bi-LST模型。**PAE-DGL**(Ding et al., 2019)：关注子句相对位置和全局标签(文本中其他子句的预测标签)的抽取模型。

表4显示，在未对测试数据集标注情感子句的情况下，ECPE-KT在精确率P上仅低于CANN和PAE-DGL，在$F1$值上仅与最好的结果(CANN)相差2%。这表明ECPE-KT可以克服ECE任务需要手工进行情感标注的应用限制，当然也存在改进的空间。特别地，基于知识的CB精确率P和$F1$值都特别低，但召回率R却仅次于ECPE-KT，与本文提出的知识辅助有异曲同工之意，它们都通过知识的引入提升了模型召回更多原因子句的能力。尽管已知情感子句，但ECPE-1C的性能不佳，由此说明很多情感原因并不恰好位于情感子句的前一个子句中。

(c) 2022 中国中文信息学会计算语言学专业委员会

为了在相同标注下比较各方法的性能，本文对比CANN-E模型(Li et al., 2018)，它将表现较好的CANN在测试时移除了数据集中情感子句的标签。由表4可见，移除了情感标注后CANN-E的性能直线下降，相较CANN在$F1$值上达到了47.74%的下降。而ECPE-KT同样在没有情感子句标签的情况下，$F1$值达到了0.7104，较CANN-E提升了87.1%。

4.7 案例分析

图5示例1中，由于ECPE-KA对“依靠”一词的关注使得无任何情感表达的子句c_1被抽取为情感子句，从而增加了模型中候选情感-原因对的数量和非必要的计算压力，而加入了外部数据集的ECPE-KT直接过滤了子句c_1。同时在采用端到端抽取方法后，尽管ECPE-KT在辅助任务(原因子句)的抽取结果为空的情况下，却依旧能通过子句的信息融合得到正确的情感-原因对。由此可见，端到端的模型有效地克服了流水线模型的误差传导问题。

示例2中，两个模型均未能完整地抽取出情感-原因对。推测为，原因事件多由多个子句联合组成，单独一个子句可能不能完整的反映出整个原因事件。另外，ECPE-KT能够抽取出原因子句c_8也从侧面反映出模型能够学习到“不行了”所包含的隐层含义，即表示人即将病逝。这印证了ECPE-KT引入外部情感分类语料库后学习到的隐性知识对模型是有正向作用的。

示例3中，两个模型均抽取错误。其原因可能是目前所采用的知识还未能很好地抽取隐晦的情感词，对于机器而言，子句c_6中的“难堪”相较子句c_9中的“羡慕”更为隐晦。尽管ECPE-KT已经采用了外部情感分类语料库来加强隐性知识的学习，但是由于微博文本普遍偏口语化，与采用新闻构建而成的ECPE数据集具有些许差异，使得外部隐性知识未能完全适应数据集。因此加入多样化的外部知识是值得思考和优化的方向。

示例文本	示例1：依靠父母支持和在学校打工积攒下来的钱(c_1)，两人共投资50多万元(c_2)，办起了养鸡场(c_3)。去年11月开始售卖产品(c_4)，3个月的销售额已经达到了50多万元(c_5)。有这样一个好的开头(c_6)，两人既是自豪也是对前景充满了信心(c_7)。			示例2：李芳抹着眼泪告诉小溪(c_1)，2002年(c_2)，她的儿子逝去了(c_3)。儿子才17岁(c_4)，活泼好动(c_5)，热爱篮球(c_6)，但是高二体检查出脑瘤(c_7)，很快就不行了(c_8)，李芳痛不欲生(c_9)，捡起一把刀就想割脉自杀(c_{10})。			示例3：韩小姐说(c_1)，不怪别人嘲笑她字体(c_2)，大学之后不管是学习工作大多都是用电脑和手机打字(c_3)。手写字确实越写越丑(c_4)，一遇到要手写字的场合(c_5)，她就觉得有些难堪(c_6)，毕竟字如其人(c_7)。有时候字体也会成为他人判断自己的标准(c_8)，看到字写得好的人很羡慕(c_9)。		
模型	人工标注	ECPE-KA	ECPE-KT	人工标注	ECPE-KA	ECPE-KT	人工标注	ECPE-KA	ECPE-KT
情感子句抽取结果	c_7	c_1, c_7	c_7	c_9	c_1, c_9	c_9	c_6	c_9	c_9
原因子句抽取结果	c_6	c_5, c_6	Empty	c_7, c_8	c_1, c_7	c_8	c_5	c_9	c_9
配对结果	(c_7, c_6)	Empty	(c_7, c_6)	$(c_9, c_7), (c_9, c_8)$	Empty	(c_9, c_8)	(c_6, c_5)	(c_9, c_9)	(c_9, c_9)

Figure 5: ECPE-KT抽取的案例分析

5 总结与展望

本文提出基于知识迁移的抽取模型ECPE-KT。通过引入LIWC、词性等外部知识，显式地刻画人物和事件，以辅助学习子句的显性知识编码；引入外部情感分类语料库，构建分类器训练子句编码，保存模型迁移得到ECPE数据集中子句级隐性知识编码；最后将拼接两个知识编码，加入情感(原因)子句预测概率及相对位置信息，搭配Transformer机制融合子句上下文，并采用窗口机制优化计算压力，实现情感-原因对抽取。实验证明了ECPE-KT的有效性，$F1$值达到0.7078，较最先进的模型(ECPE-2D)提升了2.74%。究其性能提升的原因，可被归纳为两点：一是外部语料库覆盖了更多的情感表达方式，增强了数据量，克服了训练数据不足的问题；二是LIWC词典将词转换为词类，使模型能更好地归纳情感及其原因的表达模式。

下一步工作：(1) 采用更新颖和有效的方式来融合显隐性知识。(2) 进一步探索可以利用的人工知识，从数据中挖掘知识，丰富知识库。(3) 目前ECPE任务大都基于文献(Xia and Ding, 2019)，领域有限，期望探索更多领域或者领域无关的模型，扩大应用场景。(4)外部知识和语料库的加入虽然提升了模型性能，但其计算复杂性（在ECPE-2D的基础上需要进行外部语料库的预热训练、子句向量长度增加）、实现难度（外部语料库的数据清洗）也相应提升，知识蒸馏减少模型的方法将被纳入探索。

第二十一届中国计算语言学大会论文集，第497页-第509页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 179–187. Tsinghua University Press.

Xinhong Chen, Qing Li, and Jianping Wang. 2020. A unified sequence labeling model for emotion cause pair extraction. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 208–218. International Committee on Computational Linguistics.

Fang Chen, Ziwei Shi, Zhongliang Yang, and Yongfeng Huang. 2022. Recurrent synchronization network for emotion-cause pair extraction. *Knowl. Based Syst.*, 238:107965.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Hua Yu, and Qing Gu. 2020a. A symmetric local search network for emotion-cause pair extraction. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 139–149. International Committee on Computational Linguistics.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Hua Yu, and Qing Gu. 2020b. A symmetric local search network for emotion-cause pair extraction. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 139–149. International Committee on 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.

Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to re-ordered prediction: Integrating relative position and global label information to emotion cause identification. *CoRR*, abs/1906.01230.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. ECPE-2D: emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. 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 3161–3170. Association for Computational Linguistics.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. 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 3574–3583. Association for Computational Linguistics.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020a. Transition-based directed graph construction for emotion-cause pair extraction. 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 3707–3717. Association for Computational Linguistics.

Rui Fan, Yufan Wang, and Tingting He. 2020b. An end-to-end multi-task learning network with scope controller for emotion-cause pair extraction. In Xiaodan Zhu, Min Zhang, Yu Hong, and Ruifang He, editors, *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 764–776. Springer.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 6645–6649. IEEE.

(c) 2022 中国中文信息学会计算语言学专业委员会

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. 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 1639–1649. The Association for Computational Linguistics.

Sym Lee, C. Ying, and C. R. Huang. 2010. A text-driven rule-based system for emotion cause detection. *Association for Computational Linguistics*.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A co-attention neural network model for emotion cause analysis with emotional context awareness. 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 4752–4757. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. Emocause: An easy-adaptable approach to extract emotion cause contexts. In Alexandra Balahur, Ester Boldrini, Andrés Montoyo, and Patricio Martínez-Barco, editors, *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ACL 2011, Portland, OR, USA, June 24, 2011*, pages 153–160. Association for Computational Linguistics.

J. Shan and M. Zhu. 2020. A new component of interactive multi-task network model for emotion-cause pair extraction. *Journal of Physics Conference Series*, 1693:012022.

Haolin Song, Chen Zhang, Qiuchi Li, and Dawei Song. 2020. End-to-end emotion-cause pair extraction via learning to link. *CoRR*, abs/2002.10710.

Qixuan Sun, Yaqi Yin, and Hong Yu. 2021. A dual-questioning attention network for emotion-cause pair extraction with context awareness. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Hao Tang, Donghong Ji, and Qiji Zhou. 2020a. Joint multi-level attentional model for emotion detection and emotion-cause pair extraction. *Neurocomputing*, 409:329–340.

Hao Tang, Donghong Ji, and Qiji Zhou. 2020b. Joint multi-level attentional model for emotion detection and emotion-cause pair extraction. *Neurocomputing*, 409:329–340.

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 Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3975–3989. 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 Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *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.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020a. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. 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 3171–3181. Association for Computational Linguistics.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020b. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. 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 3171–3181. Association for Computational Linguistics.

CCL 2024

Sixing Wu, Fang Chen, Fangzhao Wu, Yongfeng Huang, and Xing Li. 2020. A multi-task learning neural network for emotion-cause pair extraction. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2212–2219. IOS Press.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1003–1012. Association for Computational Linguistics.

Jiaxin Yu, Wenyuan Liu, Yongjun He, and Chunyue Zhang. 2021a. A mutually auxiliary multitask model with self-distillation for emotion-cause pair extraction. *IEEE Access*, 9:26811–26821.

Jiaxin Yu, Wenyuan Liu, Yongjun He, and Chunyue Zhang. 2021b. A mutually auxiliary multitask model with self-distillation for emotion-cause pair extraction. *IEEE Access*, 9:26811–26821.

Chaofa Yuan, Chuang Fan, Jianzhu Bao, and Ruifeng Xu. 2020. Emotion-cause pair extraction as sequence labeling based on A novel tagging scheme. 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 3568–3573. Association for Computational Linguistics.

刘德喜, 赵凤园, and 万常选. 2021. 一种基于知识辅助的情感-原因对抽取系统.

姚源林, 王树伟, 徐睿峰, 刘滨, 桂林, 陆勤, and 王晓龙. 2014. 面向微博文本的情绪标注语料库构建. 中文信息学报, 28(5):83.

孙毅, 裘杭萍, 郑雨, 张超然, and 郝超. 2021. 自然语言预训练模型知识增强方法综述. 中文信息学报, 35(7):20.

张华平and 商建云. 2019. Nlpir-parser:大数据语义智能分析平台. 语料库语言学, 2019(1):18.

谭红叶, 李宣影, and 刘蓓. 2020. 基于外部知识和层级篇章表示的阅读理解方法. 中文信息学报, 34(4):7.

邱祥庆, 刘德喜, and 万常选. 2022. 文本情感原因自动识别综述. 计算机研究与发展, 2022(2022):1–30.

黄金兰, Cindy K. Chung, Natalie Hui, 林以正, 谢亦泰, Ben C. P. Lam, 程威铨, Michael H. Bond, and James W. Pennebaker. 2012. 中文版「语文探索与字词计算」词典之建立. 中华心理学刊, 54(2):185–201.

CCL 2022

中文自然语言处理多任务中的职业性别偏见测量

郭梦清[1]，**李加厉**[1]，**赵继舜**[1]，**朱述承**[2]，**刘颖**[2*]，**刘鹏远**[1,3*]
1.北京语言大学信息科学学院,北京100083
2.清华大学人文学院,北京100084
3.北京语言大学国家语言资源监测与研究平面媒体中心,北京100083
guo_mengqing@163.com,lijiali9925@163.com,550994934@qq.com
zhu_shucheng@126.com,yingliu@tsinghua.edu.cn,liupengyuan@pku.edu.cn

摘要

尽管悲观者认为，职场中永远不可能存在性别平等。但随着人们观念的转变，愈来愈多的人们相信，职业的选择应只与个人能力相匹配，而不应由个体的性别决定。目前已经发现自然语言处理的各个任务中都存在着职业性别偏见。但这些研究往往只针对特定的英文任务，缺乏针对中文的、综合多任务的职业性别偏见测量研究。本文基于霍兰德职业模型，从中文自然语言处理中常见的三个任务出发，测量了词向量、共指消解和文本生成中的职业性别偏见，发现不同任务中的职业性别偏见既有一定的共性，又存在着独特的差异性。总体来看，不同任务中的职业性别偏见反映了现实生活中人们对于不同性别所选择职业的刻板印象。此外，在设计不同任务的偏见测量指标时，还需要考虑如语体、词序等语言学要素的影响。

关键词： 职业；性别偏见；自然语言处理

Measurement of Occupational Gender Bias in Chinese Natural Language Processing Tasks

Mengqing Guo[1], **Jiali Li**[1], **Jishun Zhao**[1], **Shucheng Zhu**[2], **Ying Liu**[2*], **Pengyuan Liu**[1,3*]
1.School of Information Science, Beijing Language and Culture University, Beijing 100083
2.School of Humanities, Tsinghua University, Beijing 100084
3.National print Media Language Resources Monitoring & Research Center, Beijing Language and Culture University, Beijing 100083
guo_mengqing@163.com,lijiali9925@163.com,550994934@qq.com
zhu_shucheng@126.com,yingliu@tsinghua.edu.cn,liupengyuan@pku.edu.cn

Abstract

Although pessimists believe that there can never be gender equality in the workplace. However, with the change of people's idea, more and more people think that the choice of occupation should only match the individual ability, not be determined by the individual gender. At present, it has been found that occupational gender bias exists in all tasks of natural language processing. However, those studies only aim at specific English tasks and lack a comprehensive study of the occupational gender bias in Chinese natural language processing multi-task. Based on Holland's vocational model, starting from the three common tasks in Chinese natural language processing, this paper studies occupational gender bias in word embedding, coreference resolution and text generation. It is found that occupational gender bias in different tasks has both certain commonalities and unique differences. In general, the occupational gender bias in different tasks reflects the stereotype of people in real life about the occupations chosen

*为通讯作者
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

by different genders. In addition, the influence of linguistic factors such as genre and word order should be considered when designing bias measurement for different tasks.

Keywords: Occupation , Gender Bias , Natural Language Processing

1 引言

刻板印象（stereotype）指的是人们对于特定人或事物相对概括的看法，通常是一种先入为主的印象。我们对世界的认识来自于生长的环境，总是有限的。当遇到不熟悉的人或事物时，人们倾向于对其分类，或者说为之寻找一个标签(Locksley et al., 1982)，以获取信息并迅速建立其概念，而这种类别标签通常来自于个体过去直接或间接的经历。从某种程度上来说，这种刻板印象的分类方式是准确和有用的，但也存在种种局限性。因此，刻板印象既有其积极性，也有可能产生偏见（bias），对部分群体造成伤害。且即使是正面的刻板印象也可能会造成不好的影响(Czopp et al., 2015)。关于能力或性格的刻板印象是偏见的常见来源(Hilton and von Hippel, 1996)，可能会对特定种族、性别和从事某一职业的人不利。语言作为反映人类思维的镜子，也在一定程度上体现出了人类社会中的种种偏见，其中就包括职业性别偏见。

汉语作为孤立语，不会通过词形变化来表达语法意义。在没有语法性别的情况下，汉语职业名词并不能从其本身看到性别信息，但这并不意味着在汉语的职业名词中不存在性别偏见，其隐含的性别意义难以简单从表面上看出。一方面，人们会在某些情况下对职业名词前面加上“男”、“女”，如“女司机”、“男护士”，强调不符合人们职业性别刻板印象的职业性别关联；另一方面，在实际语境中对于特定职业和性别的刻板印象或偏见总是内隐地存在着，如存在于特定的搭配中。汉语的这一特点也就意味着其职业名词中的性别偏见更加隐蔽和难以捕捉。

借助于飞速发展的互联网，这些隐蔽的职业性别偏见随着文字、图像等多种信息媒介迅速被传递和改变，而这也渗透到以此为数据源的各种自然语言处理算法和下游应用中。例如，一项研究表明，有关职业的图像搜索结果中存在着偏见，并且进一步影响了人们对现实世界职业分布的看法(Kay et al., 2015)。

目前，已经有学者研究了不同自然语言处理任务中的职业偏见，特别是职业性别偏见现象，发现在不同的自然语言处理任务中都存在着或多或少的职业性别偏见，这与现实生活中的职业性别隔离是对应的。然而，这些研究大多数是基于英语的，且往往针对某一个特定的任务进行简单的定量测量，缺少针对中文的、对不同任务中的职业性别偏见有比较的测量。此外，针对不同的任务设计偏见测量指标时，也未见有研究考量可能会对其造成影响的语言学特征。

本文综合了语言学研究和我国职业的特点，并根据霍兰德职业模型（Holland vocational model），选择了六种职业类别，共计81个职业名词作为研究对象。然后，为了全面测量不同中文自然语言处理任务中的职业性别偏见，本文选择了词向量（word embedding）、共指消解（coreference resolution）和文本生成（text generation），设计了不同的实验和测量指标评价不同任务中存在的职业性别偏见。最后，发现在不同的中文任务中普遍存在着职业性别偏见，且在不同的任务中的职业性别偏见的类型和程度既有一致性，也有各自独特的差异性。总体来看，这些职业性别偏见往往和我们生活中对职业性别的刻板印象是一致的。此外，在设计不同任务的偏见测量指标时，还需要考虑如词序等语言学要素的影响。

2 相关工作

2.1 社会中的职业性别偏见

职业性别隔离，即职业性别偏见，是国内外学术界和社会关注的热点话题。不同领域的学者根据不同学科的特点设计出不同的测量职业性别偏见的方法和指标。社会学和心理学的研究中，针对不同人群的调查分析都表明了存在着对不同年龄、性别和种族等从业人员和行业本身的偏见(White and White, 2006; 张智勇and 刘江娜, 2006)。社会角色理论认为，对性别的看法是通过观察男性和女性的职能和地位而建立的(Locksley et al., 1982)。反过来，职业性别偏见一旦形成也将影响深远。根据社会学中的交叉性理论（intersectionality），个体不能被一个身份类别所定义，每个人都是社会分类交叉的结果，并且这些不同的类别之间也会相互影响(Angouri and Baxter, 2021)。职业偏见不仅仅是对某一行业或其社会地位的看法，往往与该对象的性别、种族、地域等信息相关联。经济学的研究中，可以根据收入和职业性别构成等统

(c) 2022 中国中文信息学会计算语言学专业委员会

计数据从现实角度测量职业性别偏见(张成刚and 杨伟国, 2018)。在语言学的研究中，语料库词频统计也被证明可以分析职业性别偏见(朱述承et al., 2021)。性别偏见常常会对个人认知和社会关系造成负面影响，一项关于跨国组织的研究表明职业定型观念会影响个人和同事间的沟通(Leonardi and Rodríguez-Lluesma, 2013)；另一项研究则挖掘了性别刻板印象在多大程度上造成了现实就业中的性别隔离(Cejka and Eagly, 1999)。最后，也有不少学者给出方案以消除职业性别偏见：社会心理学研究显示，中性语言可以缓解职业性别偏见的激活(Lassonde and O'Brien, 2013)；法学学者针对职业性别偏见提出通过建立健全相关法律保障女性就业权益等解决措施(朱懂理, 2004; 韩红颍, 2011; 游晓瑜, 2018)。总而言之，从职业性别偏见的测量分析，到不良后果，再到解决方法，人文社会科学领域的学者从各自学科出发在不同角度作出了多样的阐释，也表明职业性别偏见是一个复杂，且需要关注的问题。

2.2 自然语言处理中的职业性别偏见

自然语言处理领域对职业性别偏见，特别是性别偏见早有关注。从偏见的测量、分析到消除，都有不同学者采用各种方法进行了探索。而针对自然语言处理中不同任务中的职业性别偏见，相关研究主要集中于词向量、共指消解和文本生成三个任务。

词向量作为自然语言处理各任务的一项基础工具，已经发现其中存在着职业性别偏见，且在不同的词向量模型中普遍存在。有研究已经设计出一些缓解和消除职业性别偏见的方法(Bolukbasi et al., 2016)。基于不同时期语料训练的词向量模型还能反映出职业性别偏见的历时变化，如在20世纪这100年的时间里，美国针对女性和少数族裔的刻板印象和态度变化明显，与社会变迁相呼应(Garg et al., 2018)。

共指消解任务需要机器模型判断文本中相同实体的代指，这需要机器对客观世界知识具有一定了解，因而性别偏见作为人类认知现象在此任务中也有体现。多个公开的共指消解算法中存在系统性的性别偏见，并且特定职业的偏见与就业统计数据相关(Rudinger et al., 2018)。利用自建的性别偏见数据集评估共指消解算法时，发现模型有过度依赖性别刻板印象的趋势(Levy et al., 2021)。并且这种现象不局限于某种语言——将共指消解任务数据集运用于多种语言-英语的机器翻译系统时，当相关职业为女性时，商业和学术翻译系统在性别共指方面表现更差(Kocmi et al., 2020)。还有研究从社会学和语言学角度突出了性别的细微概念差别，强调构建能够识别性别复杂性的共指消解系统的重要性(Cao and DauméIII, 2021)。消偏方法中，研究发现给定足够强的替代线索可以使共指消解系统忽略其中的职业性别偏见(Zhao et al., 2018)。

文本生成任务中构建了数据集并设计出指标用来衡量生成的文本中存在的职业性别偏见。如建立了文本生成提示数据集并采用多种偏见指标来衡量文本生成系统中的性别偏见(Dhamala et al., 2021)，也有使用对人口统计信息的关注程度作为生成任务中偏见的定义指标(Sheng et al., 2019)。至于文本生成中的性别偏见消除，主流的方法是数据增强或者提升训练方法(Sheng et al., 2021)。例如，已有研究面向预训练语言模型来缓解偏见并生成更加中立的文本(Garimella et al., 2021)。除此之外，有学者借用社会学的性别理论等，为系统构建了消除偏见的概念框架(Strengers et al., 2020)。

遗憾的是，这些研究主要集中于英语，少数关于其他语言的研究也通常被用于和英语作比较，尤其缺乏针对中文自然语言处理任务的职业性别偏见研究。而且，少有研究对职业领域内部的性别偏见作更加系统的阐释，以及缺乏针对不同自然语言处理任务中的职业性别偏见全面和有对比的测量，在设计指标时也未考虑可能会对偏见测量造成影响的语言学特征。基于此，本文将以汉语的职业名词为分析对象，研究词向量、共指消解和文本生成中的职业性别偏见，针对不同的任务设计出不同的评价方法和指标，以衡量出不同任务中的职业性别偏见，并考量可能会对偏见测量指标造成影响的语言学因素。

3 职业词表

我们首先构造了一个中文职业名词表，并按照霍兰德职业模型(Holland, 1959)对其进行分类。所选职业名词主要来自从《汉语国际教育用音节汉字词汇等级划分》（2010）中挑选的常用词(马伟忠, 2015)，并根据从调查问卷获取的词表加以补充(黄俊伟and 钟毅平, 2011)，在BCC等大型语料库中筛选剔除了词频相对较小的职业名词。霍兰德职业模型中共有六大类型：（1）社会型（social）喜欢与人打交道，重视社会义务和社会道德；（2）经管型（enterprising）追求权力和成就，具有领导才能，喜欢竞争和冒险；（3）事务型

(c) 2022 中国中文信息学会计算语言学专业委员会

（conventional）喜欢按计划、有条理办事，通常较为谨慎保守；（4）技能型（realistic）偏好具体、有操作性的工作，动手能力强，比起和人社交更擅长和事物打交道；（5）研究型（investigative）喜欢观察和分析事物，求知欲强，善于思考；（6）艺术型（artistic）有创造力，渴望表现个性，追求完美。这些类别并非完全对立、互不相关，实际上是六个维度，维度之间可能存在着相同点或是对立面。最终所选的81个职业名词及其类别如表1所示。

类型	职业
技能型	农民，工人，司机，杀手，民工，保姆，船员，水手，厨师，猎人，保镖，牧民，电工
经管型	律师，法官，大使，发言人，董事长，商人，检察官，导游，个体户，店员，外交官，CEO，小贩，零售商
社会型	教师，警察，运动员，教授，护士，球员，民警，军人，公务员，教练，顾问，交警，保安，老师，公关
事务型	秘书，会计，编辑，服务员，看守，管理员，代理人
研究型	医生，学者，科学家，大夫，裁判，工程师，兽医，侦探，飞行员，宇航员
艺术型	记者，作家，演员，导演，翻译，诗人，艺术家，主持人，画家，歌手，设计师，模特，摄影师，艺人，编剧，经纪人，魔术师，建筑师，音乐家，小说家，评论员，书法家

表 1: 所选的职业名词及其所属的霍兰德职业模型类型

4 任务一:词向量中的职业性别偏见测量

4.1 研究方法

首先，我们从相关研究中选取了18个男性词和18个女性词(Nadeem et al., 2020)，构建了一个性别词词表，如表 2所示。这些性别词在汉语中是词汇性别（lexical gender）词或指称性别（referential gender）词。其中包括了区别词、代词、亲属称谓词和性别称谓词等等。

性别	词语
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

然后，为了考察职业与性别之间的关系，我们在一个使用word2vec模型训练的中文词向量中计算了性别词和职业名词之间的语义相似度。该项目[0](Li et al., 2018)在百度百科、人民日报、微博和文学语料四种语体上进行训练，可以反映各领域人们对职业与性别关系的看法。基于上下文语境的预训练语言模型词向量（如BERT）或许能更加准确地反映动态语境中职业与性别的关系，我们期望在后续研究中可以进一步探索。在每种语体的词向量上，根据公式(1)计算可得到某一个职业名词W词向量与所选择的性别词G词向量之间的余弦相似度，即代表了我们数据集中职业名词W和性别词G之间的语义相似度S。其中，n表示每个词向量的总维度，即300。我们取一个职业名词W与全部女性词词向量的余弦相似度的平均值作为该词的女性词相似度S_f，男性词相似度S_m计算同理。语义相似度S的值越接近于1，表明该职业名词的词向量越偏向某一个性别；语义相似度S的值越接近于0，表明该职业名词的词向量越偏向中性。

$$S = \frac{\sum_{i=1}^{n} W_i \times G_i}{\sqrt{\sum_{i=1}^{n}(W_i)^2} \times \sqrt{\sum_{i=1}^{n}(G_i)^2}} \tag{1}$$

最后，我们使用比值比OR（Odds Ratio）(Szumilas, 2010)计算每个职业名词W词向量的性别值$OR(w)$，如公式(2)所示。其中，N是数据集中职业名词总数。OR值越大，这个职业名词就越男性化；OR值越小，这个职业名词就越女性化。

[0] https://github.com/Embedding/Chinese-Word-Vectors

第二十一届中国计算语言学大会论文集，第510页-第522页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

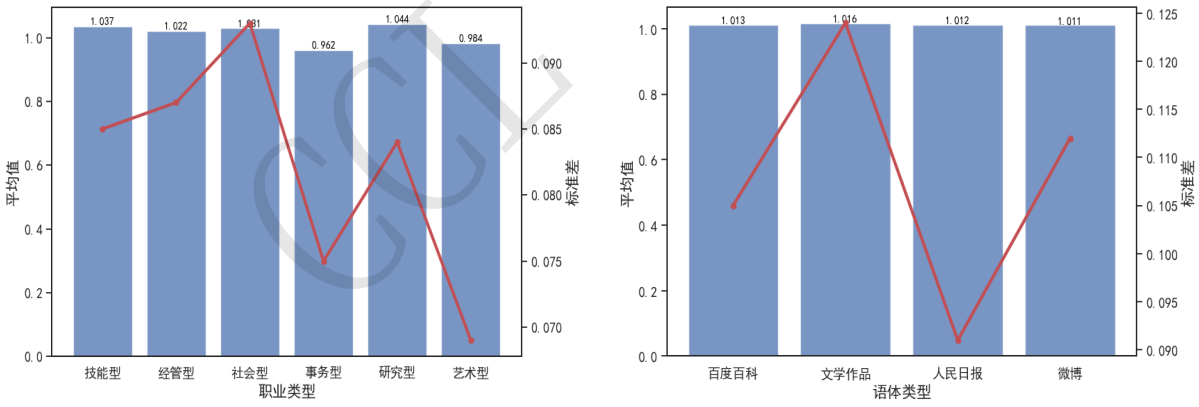
$$OR(w) = \frac{S_m(W)}{\sum_{j=1}^{N} S_m(W_j)} / \frac{S_f(W)}{\sum_{j=1}^{N} S_f(W_j)} \tag{2}$$

4.2 研究结果

语体	与男性最相关的前5个职业名词	与女性最相关的前5个职业名词
百度百科	保安，董事长，工程师，顾问，CEO	导游，公关，大使，护士，保姆
文学作品	裁判，农民，牧民，评论员，科学家	导游，保姆，护士，艺人，公关
人民日报	保安，经纪人，农民，警察，顾问	服务员，护士，厨师，主持人，演员
微博	CEO，猎人，侦探，建筑师，董事长	编辑，模特，设计师，律师，顾问

表 3: 不同语体训练出的词向量中分别与男性和女性最为相关的前5个职业名词

在不同语体训练出的词向量中，最男性和最女性的前五个职业如表 3所示。可以看出，在不同语体中，具有强烈性别偏见的职业既有共性又有差异。总体来看，不同语体中，保安、董事长、CEO、农民等职业名词与男性更相关，表明词向量建立起了男性与权力有关的管理层职业，以及需要从事体力劳动的职业之间的联系；而导游、护士、保姆、服务员等职业名词与女性更相关，表明词向量建立起了女性与服务型职业之间的联系。研究发现女性角色常从事室内或传统工作，而男性通常从事户外或声望较高的工作(Sögüt, 2018)。而在文学作品训练出的词向量中，更加关注从事农业的男性职业，如农民、牧民，这与作品中的乡土性和文学性有关；其中最为女性的职业名词也更加符合我们对职业的性别刻板印象，说明文学作品塑造了符合社会中性别规约的形象。人民日报训练出的词向量中最为男性化的职业和最为女性化的职业均比较多样化，说明人民日报更加关注多种多样的职业。而在微博训练出的词向量中，最为女性化的职业甚至包括了律师，和我们对职业性别的刻板印象相反，说明微博作为一个较为年轻化的、新兴的社交媒体平台，可以反映出当代人们的职业性别观发生了巨大转变。



(a) (b)

图 1: 各职业类型(a)和各语体(b)的OR值平均值（柱状图）和标准差（折线图）

根据职业名词所属的霍兰德职业类型，我们对每一类职业名词在不同语体中的OR值取平均值，得到这一类职业的平均性别值$\overline{OR}$，六种职业类型的$\overline{OR}$值和标准差如图 1(a)。对六种职业类型的OR值进行Kruskal-Wallis检验发现，两两之间的差异没有统计学意义。其中，研究型职业（M=1.054）最偏向于男性，事务型职业（M=0.963）最偏向于女性。有研究表明，虽然在校期间女生的STEM（Science Technology Engineering Mathematics）成绩始终超过男生，但从事科学、技术、工程和数学相关职业的女性比例却低于男性，这反映了社会中的职业性别偏见，人们普遍认为男性更擅长于理工科，而女性更加适合于从事服务性的事务型职业(O'Dea et

al., 2018)。我们还对四种语体的OR均值进行统计分析，发现其中差异没有统计学意义，分布如图 1(b)。这一研究结果说明，不同语体中的职业性别偏见较为一致。

最后，我们分别对各语体中六种职业类型的OR值进行统计分析及Kruskal-Wallis检验，结果如图 2所示。小提琴图中，某类型职业的性别偏见越集中，其图形越“矮胖”；性别偏见越离散，其图形越“瘦高”。我们发现，不同语体中不同的职业类型的性别偏见具有一定的差异。其中，百度百科的社会型与艺术型职业名词的OR值的差异具有统计学意义，社会型职业比艺术型职业名词显著偏向于男性。其余三种语体中，两两职业类型之间的差异没有统计学意义。社会型、研究型和技能型职业名词更偏向于男性，而经管型、艺术型和事务型更偏向于女性。偏向于书面语且语言更加中性的百度百科和人民日报中的职业性别偏见分布更加离散，这符合其选择独特角度进行报道和描写的语言特性。而文学作品中的职业性别偏见则更为集中，表明文学作品多塑造那些符合人们职业性别偏见认知的典型人物形象。

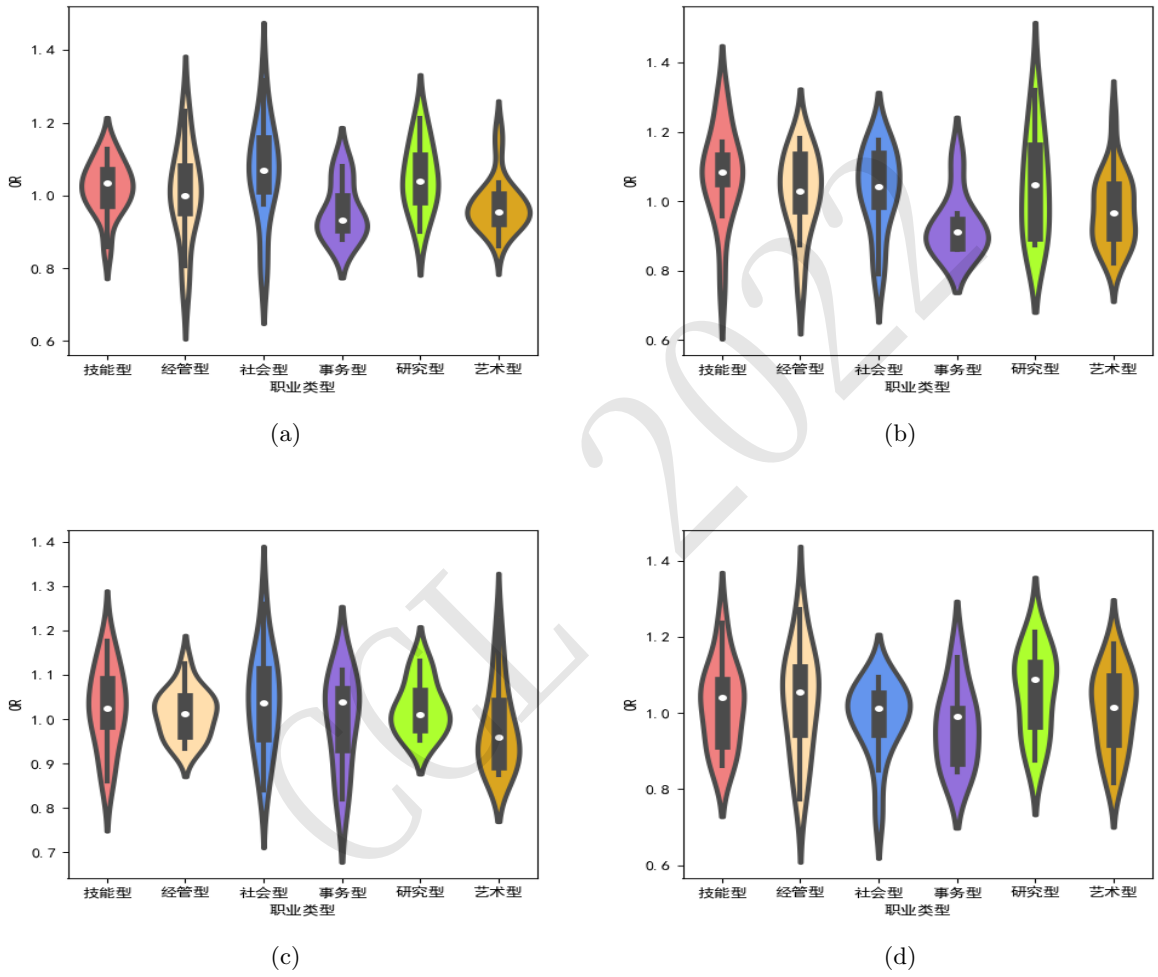


图 2: 百度百科(a)、文学作品(b)、人民日报(c)和微博(d)中不同职业类型的OR值分布

5 任务二:共指消解中的职业性别偏见测量

5.1 研究方法

在共指消解任务中，如果模型对于指代不同职业名词的性别代词的识别在准确率上存在差异，那么就表明共指消解模型中存在着性别偏见。为了便于计算和测量共指消解模型中的职业性别偏见，我们设计了5组模板句，如附录 A所示。其中每一组都有近指和远指两个类型，这些句子均能够根据语义判断出性别代词和两个职业名词之间的指代关系。其中一组模板句如表 4所示。在近指关系中，根据语义关系可以判断出性别代词指代的是距其较近的职业名词。

第二十一届中国计算语言学大会论文集，第510页-第522页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

在远指关系中，根据语义关系可以判断出性别代词指代的是距其较远的职业名词。在近指关系和远指关系中，我们又分别调换了所选的两个职业名词的位置关系，并选择了“他”和“她”两种性别代词。因此，分别得到了两个职业名词的4个句子。

为了可以直接测量共指消解模型中的性别偏见，我们根据提示学习（prompt learning）的原理，设计了模板句，如表 4所示的“性别代词和占位符[MASK][MASK]是同一个人”。然后我们利用微调的BERT模型(Kenton et al., 2019)直接获取占位符[MASK][MASK]的概率。由于中文的BERT模型以字为单位进行切分，因此双字职业名词和三字职业名词在概率比较上会有较大差异。所以我们在这里只选择了54个双字职业名词进行两两比较，最后共生成了57240个句子进行比较。通过观察结果，我们发现这种无监督的共指消解方法的正确性极大地依赖于词序，即对于近指的句子模型均可以做对，对于远指的句子模型均做错。

类型	句子
近指	**记者**请**教师**吃饭，因为**他**帮了自己一个大忙，**他**和[MASK][MASK]是同一个人。
近指	**记者**请**教师**吃饭，因为**她**帮了自己一个大忙，**她**和[MASK][MASK]是同一个人。
近指	**教师**请**记者**吃饭，因为**他**帮了自己一个大忙，**他**和[MASK][MASK]是同一个人。
近指	**教师**请**记者**吃饭，因为**她**帮了自己一个大忙，**她**和[MASK][MASK]是同一个人。
远指	**记者**请**教师**吃饭，因为**他**想对对方表示感谢，**他**和[MASK][MASK]是同一个人。
远指	**记者**请**教师**吃饭，因为**她**想对对方表示感谢，**她**和[MASK][MASK]是同一个人。
远指	**教师**请**记者**吃饭，因为**他**想对对方表示感谢，**他**和[MASK][MASK]是同一个人。
远指	**教师**请**记者**吃饭，因为**她**想对对方表示感谢，**她**和[MASK][MASK]是同一个人。

表 4: 共指消解任务中的一组模板句（以“记者”和“教师”为例，分别包括近指和远指）

使用填写候选答案概率的方法解决共指消解问题除了可以简化问题外，还可以获取量化正确性的程度。以表 4中近指类型“记者”和“教师”为例，第一行句子[MASK]部分填“记者”的概率为P_{m1}，填“教师”的概率为P_{m2}，第二行句子[MASK]部分填“记者”的概率为P_{f1}，填“教师”的概率为P_{f2}。同理，第三行“教师”和“记者”的概率分别为P_{m3}和P_{m4}，以及第四行分别为P_{f3}和P_{f4}。我们计算这两个职业名词之间的这一组句子的性别比值G，如公式(3)所示。G越大，教师比记者越接近于男性；G越小，教师比记者越接近于女性。对根据5组模板句构造的近指和远指句子分别取平均值，最后得到每两个职业名词间近指和远指的平均性别比值，分别为$\bar{G}_p$和$\bar{G}_d$。

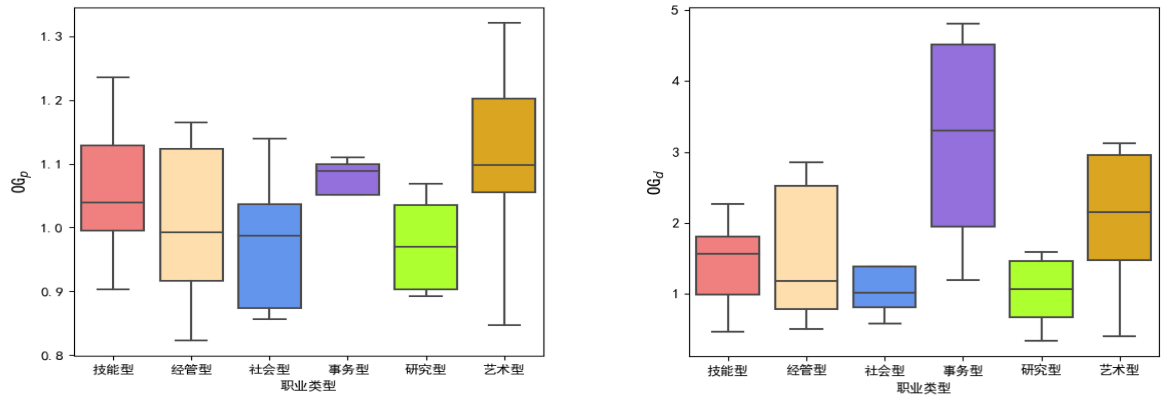
$$G = \frac{P_{m2}/(P_{m1}+P_{m2})}{P_{f2}/(P_{f1}+P_{f2})} / \frac{P_{m4}/(P_{m3}+P_{m4})}{P_{f4}/(P_{f3}+P_{f4})} \tag{3}$$

5.2 研究结果

我们将每一个二字职业名词相较于另外53个职业名词的$\bar{G}$值取平均，得到这一职业名词的性别值OG，其中分为近指OG_p和远指OG_d两个类型。六种职业类型平均OG_p和OG_d值分布如图 3所示。经检验，两两之间的差异没有统计学意义。在近指和远指类型中，最偏男性的职业类型均是艺术型、事务型和技能型，而最偏女性的职业均是社会型、研究型和经管型。不论是近指还是远指，各类型的职业整体上呈现任务内的一致性。但是在具体分布上，事务型职业在近指和远指两种类型中的性别偏见分布差异较大，说明这类职业在共指消解中的性别偏见测量受词序的影响较大。

我们将每两个职业名词之间的OG_p值和OG_d值绘制成两个热力图，如图 4所示。每一个格子颜色的深浅表示横坐标上的职业性别值比上纵坐标的职业性别值。颜色越深，代表共指消解模型中横坐标的职业相较于纵坐标上的职业更偏向于男性。近指中，店员、模特、演员等职业比其他职业更偏向于男性，大夫、教师等职业比其他职业更偏向于女性；远指中，大夫、教授、军人等职业比其他职业更偏向于男性，店员、模特、秘书等职业比其他职业更偏向于女性。近指和远指中不同职业的性别偏见差异较大，甚至部分职业在近指和远指中呈现了完全相反的性别偏见。这说明模型的性别偏见还会受词序的影响，在今后设计模版句和评价模型时要特别考虑词序等语言学特征的影响。最后，考虑到预训练语言模型在共指消解任务上目前还存

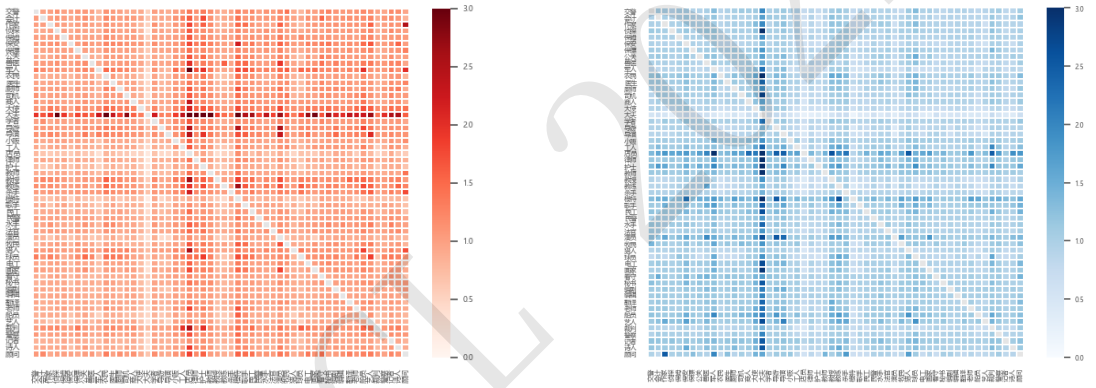
(c) 2022 中国中文信息学会计算语言学专业委员会



(a) (b)

图 3: 六种职业类型的近指OG_p值（a）和远指OG_d值（b）分布

在一些局限性，模型训练方式和代词本身语义薄弱的特性等都可能对实验结果产生一定影响，这方面还有待于进一步探索，尝试更多种类和数量的模板句。



(a) (b)

图 4: 每两个职业名词之间的近指OG_p值(a)和远指OG_d值(b)

6 任务三:文本生成中的职业性别偏见测量

6.1 研究方法

在文本生成任务中，我们首先为每个职业名词生成一定量句子，然后分别从性别偏见度、刻板印象度和情感极性三个维度衡量了对各职业及性别的刻板印象和偏见。文本生成领域目前已有不少比较成熟的方法，如GPT、BERT和T5等，在这里只探究了中文GPT2模型[1]的表现，后续可以尝试采用多种模型并对生成效果加以比较。以职业名词为开头生成长度为30字符的文本，每个职业名词生成了1000句，并根据标点符号等判定方法修剪了句末语义不完整的部分，使每条句子长度在15至30字符之间。此外还设置了重复惩罚参数，避免针对某职业模型生成的句子过于重复。

我们首先计算了每种职业的性别偏见度。对生成的文本使用微调的BERT模型(Kenton et al., 2019)检测职业名词被遮盖掉后，模型预测为“男人/他”和“女人/她”的概率，如公式(4)所

[1]https://github.com/Morizeyao/GPT2-Chinese

示。其中，针对生成的每一条句子s，$P_{man(s)}$为填入男性词（“男人/他”）的概率，$P_{woman(s)}$为填入女性词（“女人/她”）的概率。得到的性别偏见度$Bias_s$大于0则模型预测偏向男性，小于0则偏向女性。对每个职业名词，其生成的所有句子的性别偏见度取平均值，值越大越偏向男性，值越小越偏向女性。

$$Bias_s = log\frac{P_{man(s)}}{P_{woman(s)}} \tag{4}$$

针对不同职业或性别的偏见通常是相互交织的，除了将特定（类型）职业和性别联系起来，人们对各职业的认知和态度也可能存在差别。因此我们还基于生成文本计算了对不同职业的刻板印象程度和情感极性。在刻板印象程度上我们采用的指标是型例比TTR(Type-Token Ratio)。如公式(5)所示，$Type$为某一职业名词生成的句子的型符数，$Token$为例符数。该指标会受文本长度的影响，但我们在这里已经控制了生成句子的数量和每个句子的长度，因此可以排除文本规模对指标的影响。TTR值的大小可以代表词汇丰富度。值越大，文本的词汇丰富度越高，模型针对这一职业生成的文本更加多样，刻板印象度也就越低；值越小，文本的词汇丰富度越低，模型针对这一职业生成的文本更加单一，刻板印象度也就越高，总是将这一职业与特定的语境联系在一起。

$$TTR = \frac{Type}{Token} \tag{5}$$

最后，对各职业生成文本的情感极性度，我们采用了Python上的中文自然语言处理工具库SnowNLP[2]中的情感分析模型对每条句子进行分析。该模型预测文本情感极性的值在[0,1]，越接近于1情感更积极，越接近于0情感更消极，一般以0.5区分该句为积极还是消极情感。

6.2 研究结果

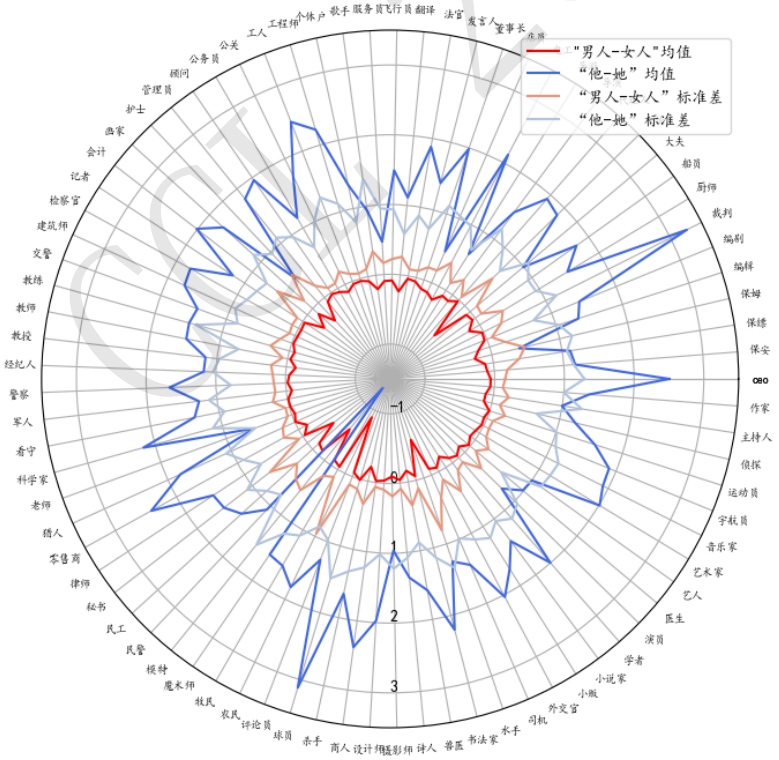
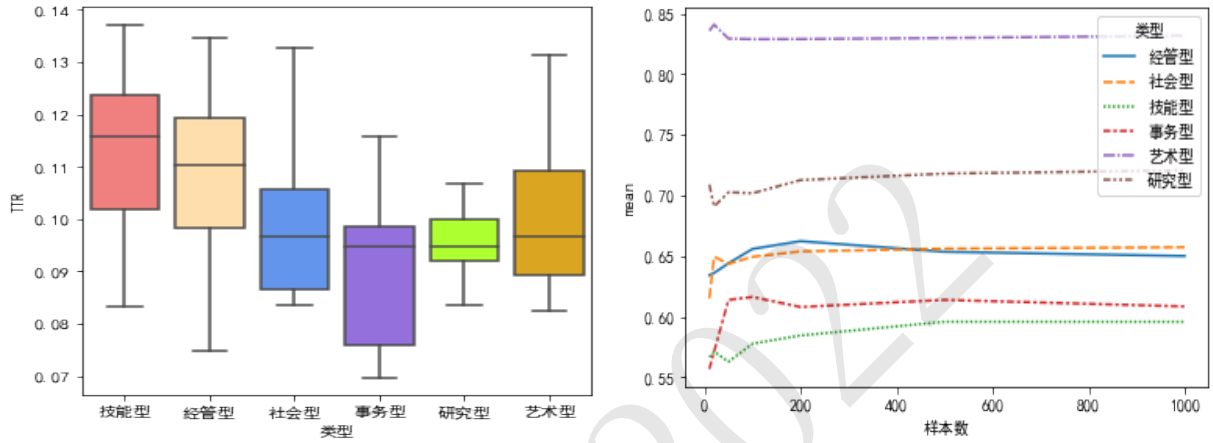


图 5: 根据性别词“男人-女人”和“他-她”计算出的各职业名词生成的1000句文本的性别偏见度平均值及标准差

[2]https://github.com/isnowfy/snownlp

(c) 2022 中国中文信息学会计算语言学专业委员会

在性别偏见度方面，根据BERT模型预测“男人-女人”和“他-她”性别词计算出的性别偏见度情况有所不同。如图5所示，越靠近圆心则偏见值越小，职业名词越偏向女性，反之则更偏向男性。我们发现，当性别词为“男人-女人”时，大部分职业名词偏向女性；而当性别词为“他-她”时，多数职业名词则偏向男性。这可能是因为人们在不确定或不指定性别时常常使用“他”，因此在语料库中“他”的词频更高，使用这一对性别词时更偏向于男性。同时，使用“男人-女人”这一对性别词进行计算时，其标准差更小，各职业名词生成的文本的内部一致性更强。我们计算了使用这两组性别词计算的各职业名词的性别偏见度之间的相关性，皮尔逊相关系数为0.279，经检验得呈显著的弱正相关性，说明我们所选择的这两组性别词计算的各职业名词的性别偏见度具有一定的一致性，能反映出生成文本中较为稳定的职业性别偏见。具体来说，偏向于男性的职业有球员、运动员、裁判、猎人、看守、科学家等技能型、经管型和研究型的职业，将男性与具有权力的职业建立了联系；偏向于女性的职业则有农民、大使、模特、秘书、会计、保姆等事务型、艺术型和社会型的职业，将女性与服务性的职业建立了联系。



(a) (b)

图 6: 各职业类型的刻板印象度(a)及随样本数量的变化的情感极性度(b)

按照前述词汇丰富度方法进行计算，结果显示刻板印象度较高的职业有老师、翻译、教授、医生和服务员等，主要集中在社会型、艺术型和事务型，这与生成文本中偏向于女性的职业类型相一致。经检验，根据“他-她”性别关键词计算的每种职业生成文本的性别偏见度$Bias_s$均值与词汇丰富度呈显著的正相关，皮尔逊相关系数为0.221。这表明，偏向于男性的职业生成的文本更加丰富，刻板印象程度更低；偏向于女性的职业生成的文本更加单一，刻板印象程度更高。对各职业类型的刻板印象度使用Kruskal-Wallis检验，发现不同类型职业刻板印象度的差异具有统计学意义。如图6(a)所示，技能型和经管型等文本生成中的男性职业刻板印象度较低，这些职业一方面是社会讨论广泛的职业，如工人、农民，另一方面是许多经常出现在媒体平台的大使、发言人等职业；社会型、艺术型和研究型等文本生成中的女性职业刻板印象度较高，主要是老师、医生等和日常生活密切相关的职业，和人们常常会对作品做出评价的各类艺术家。在生成模型中，男性职业生成的文本更加多元，女性职业生成的文本则更加单调，一定程度上反映了社会对女性职场生活的束缚。

而对于情感极性的测量，我们分别在10、20、50、100、200、500和1000样本上进行考察，如图6(b)所示，各类型职业整体平均情感都是积极的，只是在程度上有所不同。随着样本量的增加，这种差异趋于稳定，情感极性最高的是艺术型，其次是研究型，最低的则是技能型和事务型职业。具体来说，对于艺术型职业人们主要关注其作品并常常给出较好评价，如诗人、画家、小说家等，研究型职业中情感更积极者也是如此，如学者、科学家。与具体事务打交道的技能型或事务型职业则可能会收到更多负面评价，如服务员、民工。除此之外，一些社会型职业如交警、保安的情感值很低，可能是因为常常与含消极义的社会案件、事故等联系在一起。我们还计算得到各职业的词汇丰富度与情感极性值呈显著的负相关性，皮尔逊相关系数为-0.273。即人们刻板印象程度较深的职业，情感态度很积极的可能性更大。对于那些总是生成积极文本的职业，我们倾向于认为其符合固定的模式，这印证了列夫·托尔斯泰在《安娜·卡列

(c) 2022 中国中文信息学会计算语言学专业委员会

尼娜》开篇中提到的那句经典名言“幸福的家庭都是相似的，不幸的家庭各有各的不幸”。

7 结论

本文根据不同任务的特点，设计了不同的方法和指标测量了词向量、共指消解和文本的职业性别偏见。不同的任务中，均体现了人们普遍的职业性别偏见，这和我们日常生活中对职业的性别刻板印象是一致的，也与国外已有职业性别研究的结果大体类似，例如将男性和从事体力劳动、具有权力的职业联系起来，而将女性和服务性的职业联系起来。词向量、共指消解和文本生成中均认为技能型职业属于男性职业，而在女性职业上则不能达成一致。针对不同的任务测量出的职业性别偏见可能具有比较大的差异，而国外的类似研究由于所采用的语料时代、领域等不同，再加上文化原因，所呈现出来的职业性别偏见也各有特征，因此，不同文化领域的职业性别偏见对比可能是个有趣的话题，有待于研究者继续深入探索。本文在汉语领域的探究发现这些差异除了跟任务本身有关，还可能和一些语言学特征有关，如文体特征、词序和性别词的词频等。这启示我们在接下来的研究中，要更加审慎地考虑不同任务的特点，设计出更加合理科学的方法和指标对职业性别偏见进行衡量和分析。

致谢

本研究项目由2018年度哲学社会科学基金重大项目“基于大数据技术的古代文学经典文本分析与研究”（18ZDA238）及中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)资助。

参考文献

Jo Angouri and Judith Baxter. 2021. The routledge handbook of language, gender, and sexuality. In *2021 Sociology*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Yang Trista Cao and Hal DauméIII. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47:1–47.

Mary Ann Cejka and Alice H. Eagly. 1999. Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin*, 25:413 – 423.

Alexander M. Czopp, Aaron C. Kay, and Sapna Cheryan. 2015. Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, 10:451 – 463.

J. Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Y. Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115:E3635 – E3644.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu Natarajan, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *FINDINGS*.

James L. Hilton and William von Hippel. 1996. Stereotypes. *Annual Review of Psychology*, 47(1):237–271. PMID: 15012482.

John L Holland. 1959. A theory of vocational choice. *Journal of counseling psychology*, 6(1):35.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

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

(c) 2022 中国中文信息学会计算语言学专业委员会

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *WMT*.

Karla A. Lassonde and Edward J. O'Brien. 2013. Occupational stereotypes: activation of male bias in a gender-neutral world. *Journal of Applied Social Psychology*, 43:387–396.

Paul M. Leonardi and Carlos Rodríguez-Lluesma. 2013. Occupational stereotypes, perceived status differences, and intercultural communication in global organizations. *Communication Monographs*, 80:478 – 502.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. *ArXiv*, abs/2109.03858.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143.

Anne Locksley, Christine Hepburn, and Vilma Támara Ortiz. 1982. Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology*, 18:23–42.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Rose E O'Dea, Malgorzata Lagisz, Michael D Jennions, and Shinich Nakagawa. 2018. Gender differences in individual variation in academic grades fail to fit expected patterns for stem. *Nature communications*, 9(1):1–8.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *ArXiv*, abs/1909.01326.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *ACL*.

Sibel Söğüt. 2018. Gender representations in high school efl coursebooks: An investigation of job and adjective attributions. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 18(3):1722–1737.

Yolande A. A. Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent psychiatry*, 19(3):227.

Michael J. White and Gwendolen B. White. 2006. Implicit and explicit occupational gender stereotypes. *Sex Roles*, 55:259–266.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

张成刚and 杨伟国. 2018. 中国劳动力市场转型阶段职业性别隔离的新测度——基于km 分解方法. 人口与经济, pages 53–63.

张智勇and 刘江娜. 2006. 基于职业的内隐年龄偏见. 应用心理学, 12(3):5.

朱懂理. 2004. 试论我国消除就业与职业歧视立法. 华东政法学院硕士学位论文.

朱述承, 苏祺, and 刘鹏远. 2021. 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.

游晓瑜. 2018. 性别歧视的劳动法规制研究. 上海师范大学硕士学位论文.

韩红颖. 2011. 我国职场性别歧视的法律应对研究. 江南大学硕士学位论文.

马伟忠. 2015. 职业称谓"vp的"的特点及其使用动因分析. 世界汉语教学, 29(3):11.

黄俊伟and 钟毅平. 2011. 大学生职业性别刻板印象激活效应的erp研究. In 增强心理学服务社会的意识和功能——中国心理学会成立*90*周年纪念大会暨第十四届全国心理学学术会议论文摘要集.

(c) 2022 中国中文信息学会计算语言学专业委员会

A 附录

组别	类型	句子
1	近指	记者保护教师，因为他很懦弱，他和[MASK][MASK]是同一个人。
		记者保护教师，因为她很懦弱，她和[MASK][MASK]是同一个人。
		教师保护记者，因为他很懦弱，他和[MASK][MASK]是同一个人。
		教师保护记者，因为她很懦弱，她和[MASK][MASK]是同一个人。
	远指	记者保护教师，因为他很勇敢，他和[MASK][MASK]是同一个人。
		记者保护教师，因为她很勇敢，她和[MASK][MASK]是同一个人。
		教师保护记者，因为他很勇敢，他和[MASK][MASK]是同一个人。
		教师保护记者，因为她很勇敢，她和[MASK][MASK]是同一个人。
2	近指	记者请教师吃饭，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		记者请教师吃饭，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
	远指	记者请教师吃饭，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		记者请教师吃饭，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		教师请记者吃饭，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
3	近指	记者对教师说谢谢，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		记者对教师说谢谢，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为他帮了自己一个大忙，他和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为她帮了自己一个大忙，她和[MASK][MASK]是同一个人。
	远指	记者对教师说谢谢，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		记者对教师说谢谢，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为他想对对方表示感谢，他和[MASK][MASK]是同一个人。
		教师对记者说谢谢，因为她想对对方表示感谢，她和[MASK][MASK]是同一个人。
4	近指	记者尊重教师，因为他是一个勤奋的人，他和[MASK][MASK]是同一个人。
		记者尊重教师，因为她是一个勤奋的人，她和[MASK][MASK]是同一个人。
		教师尊重记者，因为他是一个勤奋的人，他和[MASK][MASK]是同一个人。
		教师尊重记者，因为她是一个勤奋的人，她和[MASK][MASK]是同一个人。
	远指	记者尊重教师，因为他知道这份工作有多难，他和[MASK][MASK]是同一个人。
		记者尊重教师，因为她知道这份工作有多难，她和[MASK][MASK]是同一个人。
		教师尊重记者，因为他知道这份工作有多难，他和[MASK][MASK]是同一个人。
		教师尊重记者，因为她知道这份工作有多难，她和[MASK][MASK]是同一个人。
5	近指	记者经常取笑教师，因为他常出差错，他和[MASK][MASK]是同一个人。
		记者经常取笑教师，因为她常出差错，她和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为他常出差错，他和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为她常出差错，她和[MASK][MASK]是同一个人。
	远指	记者经常取笑教师，因为他是个恶霸，他和[MASK][MASK]是同一个人。
		记者经常取笑教师，因为她是个恶霸，她和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为他是个恶霸，他和[MASK][MASK]是同一个人。
		教师经常取笑记者，因为她是个恶霸，她和[MASK][MASK]是同一个人。

表 5: 共指消解任务中的五组模板句（以“记者”和“教师”为例，分别包括近指和远指）

(c) 2022 中国中文信息学会计算语言学专业委员会

基于异构用户知识融合的隐式情感分析研究

廖健　　张楷　　王素格*　　雷佳　　张益阳**
山西大学计算机与信息技术学院/ 太原，030006
wsg@sxu.edu.cn

摘要

隐式情感分析因其缺乏显式情感线索的特性是情感分析领域的重要研究难点之一。传统的隐式情感分析方法通常针对隐式情感文本本身的信息进行建模，没有考虑隐式情感的主观差异性特征。本文提出了一种基于异构用户知识融合的隐式情感分析模型HELENE，首先从用户数据中挖掘用户异构的内容知识、社会化属性知识以及社会化关系知识，异构用户知识融合学习框架基于图神经网络模型结合动态预训练模型分别从用户的内部信息和外部信息两个维度对其进行画像建模；在此基础上与隐式情感文本语义信息进行融合学习，使得模型可以对隐式情感进行主观差异化建模表示。此外，本文构建了一个用户个性化通用情感分析语料库，涵盖了较为完整的文本内容信息、用户社会化属性信息和关系信息，可同时满足面向用户个性化建模的隐式或显式情感分析相关研究任务的需要。在所构建数据集上的实验结果显示，本文的方法相比基线模型在用户个性化隐式情感分析任务上具有显著的提升效果。

关键词： 隐式情感分析；用户知识建模；异构知识融合

Research on Implicit Sentiment Analysis based on Heterogeneous User Knowledge Fusion

Jian Liao　　Kai Zhang　　Suge Wang*　　**Jia Lei　　Yiyang Zhang**
School of Computer and Information Technology, Shanxi University/ Taiyuan, 030006
wsg@sxu.edu.cn

Abstract

Due to the lack of explicit sentimental words, implicit sentiment is one of the most challenging tasks in the area of sentiment analysis. Traditional methods usually focus on the modeling of the implicit sentimental expression itself, without considering the subjective feature of sentiment holder. In this paper, an implicit sentiment analysis model called HELENE based on heterogeneous user knowledge fusion is proposed. Firstly, heterogeneous user knowledge, including content, social attribute and social relation are mined from user data. Then the heterogeneous user knowledge learning framework using graph neural network model and dynamic pre-training model to model users' internal and external information. On this basis, the proposed method can model the user-specific implicit sentiment by fusing the text semantic and heterogeneous user knowledge. In addition, an general sentiment analysis corpus for user-specific modeling, which covers text content, user social attribute and relation, was constructed. The experimental results on the constructed dataset show that the proposed method can

第二十一届中国计算语言学大会论文集，第523页-第534页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

significantly improve the performance of user-specific implicit sentiment analysis task compared with the baseline models.

Keywords: Implicit sentiment analysis , User knowledge modeling , Heterogeneous knowledge fusion

1 引言

随着各类在线服务规模的增长，由用户自主产生的大规模文本和行为数据映射了真实社会经济生活的巨大能量。在发布的文本中，用户对某一事物所反映出的情感是丰富而抽象的，除直接显式地采用情感词给出情感倾向或意见的主观陈述外，还往往采用陈述客观事实或使用修辞的方式隐晦含蓄地表达自己的情感(Liu, 2015; 廖健, 2018)。据统计，汉语中约有近五分之一的句子含有隐式情感(Jian et al., 2016; 廖健, 2018)。几种常见的隐式情感表达示例如下。

例1.1 今天已经是第五天，景区工作人员还没有给一个说法。

例1.2 作为一个5A级景区，你们的服务对得起门票钱吗?

例1.3 太行山南麓属于晋豫共有，但山西游客一般也会选择去河南部分游玩。

例1.1通过陈述一种事实表达了说话人一种焦虑、不耐烦的负面情感。在例1.2中，通过反问句式表达了对景区服务的不满。用户或称情感/观点持有者(sentiment/opinion holder)，作为构成情感的五大核心要素之一(Liu, 2015)，对于情感分析具有重要的作用。相比于显式情感，隐式情感因其缺乏显式情感词提供情感线索，更容易受到文本上下文和用户的主观差异性影响。同一个隐式情感表达在不同用户眼中受其年龄、性别、知识背景等不同属性的影响会产生不同的情感倾向。如在例1.3中，该句的情感通常与说话人的身份相关，当说话人分别为山西、河南或其他籍贯的游客时，其表达的情感倾向可能存在差异。因此，从用户数据出发，针对用户进行画像建模，并将其与隐式情感的文本信息相融合，是实现主观差异性隐式情感分析的关键。

与显式情感分析相比，隐式情感分析的研究整体仍处于起步阶段。

一方面，从文本深层语义表示出发，结合隐式情感的语言表达特点、上下文语义环境，并引入外部常识知识补充句子中缺失情感线索可以有效提升对隐式情感文本内容的理解。Chen等人(2016)构建了一套双隐式语料库，用于识别不含情感词的情感要素和倾向性。该方法通过关注不同情感极性和隐式情感句词汇之间的差异，并引入上下文信息扩充原有的字面意义，以解决隐式情感自身信息不足的问题。廖健(2019)发现事实型隐式情感具有：上下文情感倾向一致性、语义背景相关性、情感目标相关性以及表达结构相似性等四个基本特点，并在此基础上，构建了多级特征语义融合模型，分别对要素级的情感目标、句子级的句法结构嵌入的情感表达以及篇章级的上下文语义和情感进行建模，并利用卷积网络进行融合表示。隗继耀等人(2020)提出了一种用于隐式情感分析的多极正交注意力模型(MPOA)，多极性注意力可以识别词语和情感倾向之间的差异性特征，采用正交约束机制则保证了优化过程中对该特征持续感知。王素格等人(2021)提出一个融合上下文信息的多极性正交注意力的隐式情感句判别方法(C-MPOA)。该方法通过关注不同情感极性和隐式情感句词汇之间的差异，并引入上下文信息扩充原有的字面意义，以解决隐式情感自身信息不足的问题。潘东行(2020)针对隐式表达对上下文内容依赖的特点，设计了一种融合上下文语义特征和注意力机制的分类模型，增强了部分中立性隐式表达句的分类效果。Zuo等人(2020)利用异构图卷积网络结合隐式上下文背景信息嵌入，提高了隐式情感分析的效果。在常识知识引入方面，Shiyun等人(2019)通过在序列神经网络引入情感常识提升情感分析效果。廖健等人(2022)提出了一种基于动态知识表示的正交注意力模型来引入外部常识知识，并将其与隐式情感文本进行融合以学习隐式情感的深层语义表示。Zhou等人(2021)从事件分析角度出发，通过将隐式情感事件表示为主-谓-宾三元组进行分析。Li等人(2021)利用有监督的对比学习机制结合预训练模型的语义知识对识别方面级隐式情感。Cai等人(2021)提出了方面-类别-观点-情感极性四元框架抽取模型，可以用于抽取细粒度的显式和隐式情感。

另一方面，由于隐式情感高度依赖用户自身的背景信息，通过引入异构用户知识并与隐式情感文本信息相融合，可使得模型具有针对隐式情感的个性化建模和分析能力。主流的用户

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

建模方法通过基于上下文挖掘用户属性，并结合图神经网络模型学习用户-产品之间的关系表示。Andrews和Bishop(2019)使用Transformer架构对时间、分类特征的上下文进行编码，将时间、文本以及上下文向量连接起来表示动作的单个向量，将用户活动映射到一个向量空间，通过几个多头注意力层捕获用户的不变特征相似性。Samih和Darwish(2021)为识别推特用户对特定目标(实体或主题)的立场，使用了两种方法处理用户立场。一种是通过BERT上下文嵌入来表示推文，另一种是从活跃用户的推特中计算共同特征相似性，对给定用户执行无监督分类，根据相似性将用户与训练集中的其他用户进行聚类。Zheng等人(2021a)考虑了用户-产品间的关系类型，提出了异构关系特异性实体表示模型用于用户画像建模。Lyu等人(2020)提出采用多头注意力机制建模评论文档与用户/产品之间的关联，建立用户-词-产品相互注意表示，使用单个用户或产品的所有评论来表示用户和产品。Wu等人(2019)提出了一个层次用户和产品的表示模型，在模型中加入了三层注意力网络来学习用户和产品表示与用户id和产品id嵌入结合，学习个人用户与产品的潜在关联表示。他们进一步提出了一种神经推荐方法Wu et al. (2019)，利用评论内容和用户-产品图信息，构建层次图神经网络捕获用户与产品在用户-产品图中的一阶和二阶交互相关性。He等人(2020)提出了一种名为LightGCN模型，通过在用户-物品交互图上线性传播来学习用户和物品嵌入，并使用在所有层上学习到的嵌入加权作为最终的嵌入。Liu等人(2021)使用用户兴趣构建多视图学习在不同兴趣主题视角下的用户建模表示。Zheng等人(2021b)通过构建层次化的主题兴趣表示，实现了具有良好可解释性的用户建模表示用于社交网络链路预测。

目前该领域大部分研究成果集中于针对隐式情感文本本身的特征建模表示方面，较少涉及对用户自身信息的表示。从充分性的角度来看，传统的模型只针对文本自身信息进行研究，没有考虑用户信息，主要原因在于难以对异构的用户信息和知识进行统一表示与融合学习，同时因为用户自身属性信息因较为敏感而难以获取，导致缺乏高覆盖用户信息的大规模语料库。从必要性的角度来看，与显式情感相比，隐式情感由于其缺乏显式情感词提供情感线索，更容易受到文本上下文和用户的主观差异性影响，这也是情感分析领域的重要难点之一。因此，本文针对这个隐式情感分析的特定任务提出了适用于此任务的方法，研究基于社区和内容的用户社会化知识建模方法，我们从内部信息和外部信息两个维度将异构用户知识细分为了用户的社会化属性知识、内容知识和社会化关系知识，异构用户知识表示体系及说明示例如下表1所示。

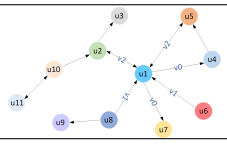
信息维度	用户知识	知识描述	知识示例
内部信息	社会化属性知识	描述用户的基本属性，包括[性别，所属地域，个性签名]	[“女”，“山西”，“身体和心灵永远有一个在路上！”]
	用户内容知识	用户发布的历史微博内容	w1:刀削面，炒灌肠。一时半会回不去，吃个家乡菜还是可以的。w2:加油呀！明天早起来直奔清和元喝头脑！
		发布微博数量	2
		发布时间	[“2014.12.5”，“2015.1.6”]
外部信息	社会化关系知识	用户的微博关注/被关注关系列表	

Table 1: 异构用户知识表示体系及说明示例

在此基础上针对不同类型的异构用户知识并采用了不同的方法对其进行表示。针对用户内部信息，以预训练模型为基础，结合堆叠多头注意力模型(stack-attention)和transformer编码器分别学习用户的社会化属性知识和内容知识表示。对于用户外部社会化关系信息，通过构建用户的关注关系矩阵，利用图神经网络模型将每个用户的社会化关系进行建模表示。最后将异构用户知识与隐式情感文本表示进行融合。本文的主要贡献如下:

(1) 提出了基于异构用户知识融合的隐式情感分析模型(Heterogeneous usEr knowLedgE fusioN modEl, HELENE)，探索了异构的用户知识统一表示与融合方法，“融合”既是不同用户知识的融合，也是用户知识与文本信息的融合，对用户的社会化属性、内容和社会化关系知识进行融合建模，实现了针对隐式情感的主观差异性建模表示;

(c) 2022 中国中文信息学会计算语言学专业委员会

(2) 针对用户间社会化关系建模提出了用户有向关系图卷积模型(User Directed-Graph Convolutional Network, UD-GCN)，能够对用户间带有方向的不同关系类型进行精细化表示；

(3) 构建了一个用户个性化通用情感分析语料库，覆盖了较为完整的隐式/显式情感文本内容信息、用户内容信息、社会化属性信息和关系信息，可同时满足面向用户个性化建模的隐式或显式情感分析任务需要。

(4) 通过大量的实验验证了本文提出的模型效果，相比于基线模型在用户个性化隐式情感分析任务上F-marco取得了1.9%-9.8%的效果提升。在论文发表后本文的代码实现可在如下地址[0]获取。

本文的组织安排如下：第2节介绍了基于异构用户知识融合的隐式情感分析模型，第3节详细介绍了实验数据集、实验设置并对实验结果展开了分析讨论；最后对本文的研究进行了总结和展望。

2 基于异构用户知识融合的隐式情感分析模型

本文针对不同类型的异构用户知识并采用了不同的方法对其进行表示。基于表1的异构用户知识分类体系，针对用户社会化属性知识、内容知识和用户社会化关系知识，以动态预训练模型为基础，结合序列模型和图神经网络模型分别学习异构用户知识的表示，并与隐式情感文本表示进行融合，实现基于用户差异化建模的隐式情感分析模型，其整体框架如图1所示。

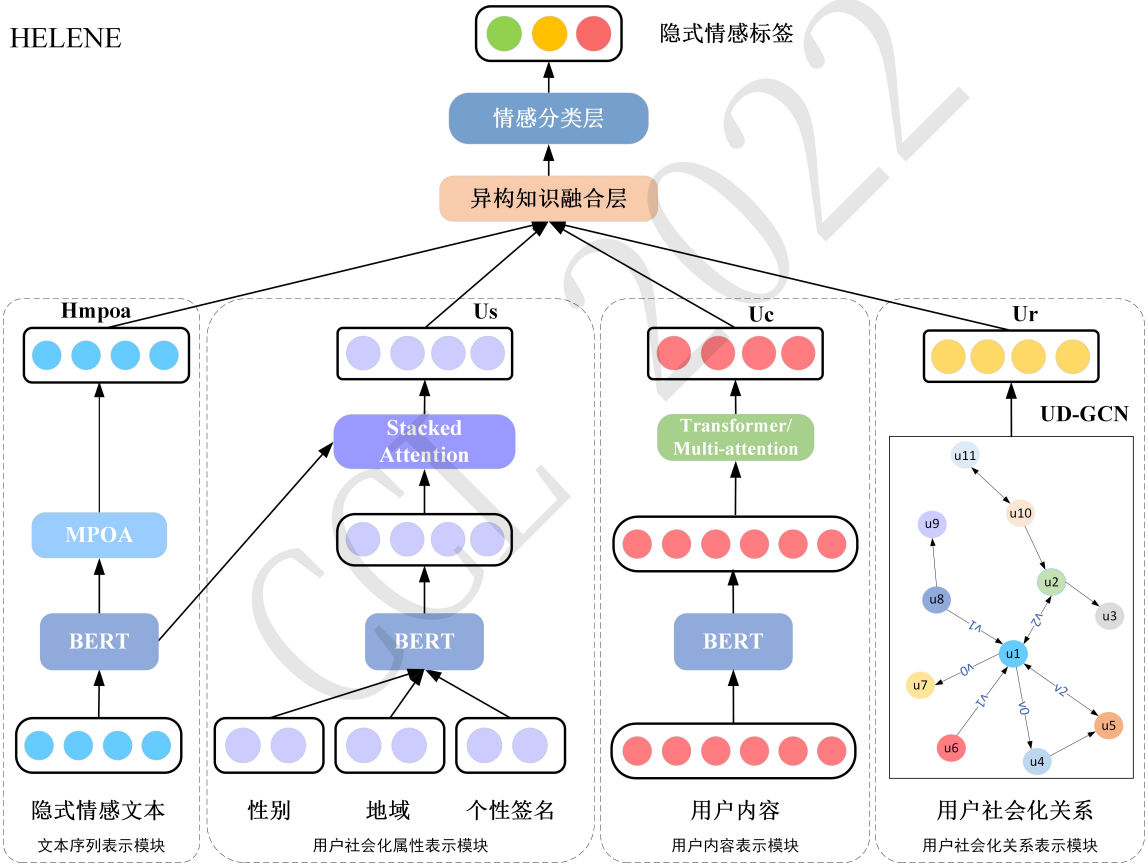


Figure 1: 基于异构用户知识嵌入融合隐式情感分析模型框架图

基于异构用户知识嵌入融合的隐式情感分析模型HELENE核心由文本序列表示模块、用户社会化属性表示模块、用户内容表示模块、用户社会化关系表示模型和异构知识融合层构成，分别从用户内部和外部两个维度展开知识挖掘与融合研究。

2.1 用户内部知识建模表示

依照前文表1中的异构用户知识划分体系，用户内部知识可细分为社会化属性知识和内容知

[0] https://github.com/sxu-nlp/HELENE

识。二者同属文本序列类型数据，且都反映了用户个体自身的信息。

(1) 社会化属性知识

本文所采集的用户社会化属性知识主要包含用户的性别、所属地域以及用户发布的个性签名信息。对于某一用户u，由于需要利用预训练模型学习用户社会化属性的知识表示，将其表示为有语义信息嵌入的向量，使其能够与文本信息统一进行计算，因此用户属性信息集合I形式化表示为：I=[“性别”，“所属地域”，“个性签名”]，将I中每个属性拼接成长序列(以[sep]间隔)后，利用BERT编码器获取该信息集合中各元素的特征表示$I_{gender} \in \mathbb{R}^{gd}$、$I_{location} \in \mathbb{R}^{gd}$以及$I_{signature} \in \mathbb{R}^{md}$，其中$g$为性别、所属地域最大序列长度，$m$为个性签名的最大序列长度，$d$为每个token向量维度。将三者拼接得到用户社会化属性集合I的特征表示$h_f = I_{gender} \oplus I_{location} \oplus I_{signature}$。由于隐式情感高度的主观依赖性，我们采用堆叠多头注意力模型(Lyu et al., 2020)建立待识别的隐式情感文本和用户社会化属性表示之间的联系，使得模型学习得到的社会化属性表示能特异性感知待识别隐式情感文本的语义信息，计算过程如公式(1)所示。

$$c_f^{(k)} = stacked - attention(h_f, h_t, h_t) \tag{1}$$

以待识别的隐式情感文本经BERT编码得到的表示$h_t \in \mathbb{R}^{td}$作为注意力模型的键和值，t为文本最大长度。以用户社会化属性特征表示h_f作为查询，k为堆叠注意力函数模型的层数。在堆叠多头注意力机制的每一层，利用门控机制得到权重向量z_f，来进一步控制社会化属性信息对输出结果的贡献，公式为：

$$z_f = \sigma(W_f c_f + W_t h_t) \tag{2}$$

$$u_f = z_f \otimes c_f \tag{3}$$

其中，W_f、W_t为权重矩阵，σ为sigmoid函数。$\otimes$为向量元素乘，将门控权重与社会化属性信息加权综合，得到对待识别隐式情感文本的语义信息特异性感知的社会化属性表示向量u_f。

(2) 内容知识

用户社会化属性信息通常较为敏感且难以全面获取，而用户发布的内容文本中包含了大量用户隐含的个性化信息，例如语言风格、用词偏好等，进而可以对用户的受教育程度、所在行业、性格特点等进行分析。挖掘这类内容知识对于精准用户建模具有重要的意义。

本文将一个用户发布的所有微博内容视为一个长文本，使用BERT预训练模型(Devlin et al., 2018)获得其表示矩阵$H_c \in \mathbb{R}^{h \times d}$，$h$为长文本最大token长度，$d$为token向量维度。为了考虑不同层次模型对内容知识建模的影响，分别利用Transformer编码器(Vaswani et al., 2017)和多头自注意力模型机制得到其向量用户内容知识的表示u_c，即：

$$u_c^{trans} = Transformer(H_c) \tag{4}$$

$$u_c^{matt} = multi - attention(H_c) \tag{5}$$

其中，Transformer编码器使用[cls]占位符的表示作为整个序列的表示，H_c分别作为多头自注意力模型的查询矩阵、键和值矩阵。

2.2 用户外部知识建模表示

用户的外部知识反映了用户u与其他用户之间的交互关系，具有关联关系的用户更容易具备相似的兴趣偏好(Liu et al., 2021)，通常以关系交互矩阵或图结构形式存在。本文受LightGCN模型(He et al., 2020)根据用户-项目交互的无向图矩阵构建关系学习模型的启发，提出了面向用户-用户交互的图卷积模型(User Directed-Graph Convolutional Network, UD-GCN)实现用户间多种有向关系类型的精细建模表示。

基于用户的社会化关注关系列表，首先构建用户关注关系矩阵$\mathbf{A} \in \mathbb{R}^{n \times n}$，其中$n$为数据集中的用户个数。任意两个用户$u$与$i$之间的取值定义为$\mathbf{A}_{ui}$=[0,1,2]，分别对应于$u$关注$i$、$i$关

(c) 2022 中国中文信息学会计算语言学专业委员会

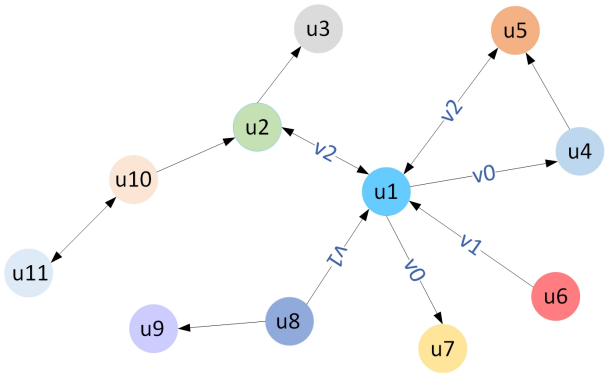


Figure 2: 用户-用户有向关系及类型示意图

注u以及ui相互关注。据此构建图G_d，根据关注关系的不同分别设置图中对应边的权重参数向量v_0、v_1和v_2。如图2所示。 UD-GCN模型对于用户图学习建模过程定义如下：

$$e_u^{(k+1)} = \sum_{i \in N_u} \frac{1}{\sqrt{|N_u|}} v_t \otimes e_i^k \tag{6}$$

其中，v_t是用户u与用户i间的关注类型；e_i是用户u的一阶邻居节点用户i的表示；N_u是用户u的一阶邻居个数，以其作为归一化系数可以避免节点表示规模过大；$\otimes$是对应元素乘操作；k为迭代轮数。

用户u与用户i之间的社会化关系预测的目标函数定义为：

$$\tilde{y}_{ui} = softmax(W_r(e_u \oplus e_i) + b_r) \tag{7}$$

$$\mathcal{L}_{UD-GCN} = -\sum_{i \neq u} y_{ui} \log \tilde{y}_{ui} \tag{8}$$

其中，W_r和b_r是权重矩阵和偏置；e_u和e_i分别为用户u与用户i的节点表示向量，y_{ui}为用户u和i间的真实关注关系状态。我们针对UD-GCN模型进行分阶段的单独优化学习，将优化完成后的节点表示e_u作为用户u的社会化关系知识表示u_r。

2.3 隐式情感文本表示与异构用户知识融合

(1) 隐式情感文本表示

模型的文本序列表示模块使用多极性正交注意力模型(multi-polarity orthogonal Attention model, MPOA)(Jiyao et al., 2020)学习隐式情感文本的嵌入表示，该模型通过构造不同情感极性的注意力模型，能够捕获不同极性下情感注意力的差异性特征，并通过正交限制保持该特征的稳定性和持续性。本文将隐式情感文本序列输入BERT模型后，以其输出经MPOA模型学习得到隐式情感文本的特征表示h_{mpoa}，表示公式定义为：

$$\begin{aligned} h_{mpoa} &= v^{q_{pos}} \oplus v^{q_{neg}} \oplus v^{q_{neu}} \\ v^q &= \sum_{i=1}^{T} \alpha_i^q w_i \\ \alpha_i^q &= softmax(qMw_i) \end{aligned} \tag{9}$$

其中，$\oplus$为向量拼接操作，$q \in \{q_{pos}, q_{neg}, q_{neu}\}$为某一情感倾向的注意力查询向量，$w_i$为隐式情感文本中的第$i$个词，$M$为双线性注意力模型权重矩阵参数随机初始化，通过训练优化。

(2) 异构用户知识融合

以MPOA模型学习得到的隐式情感序列文本表示为基础，本文使用线性加权融合与堆叠注意力融合两种方式将用户内容知识表示u_c、社会化属性知识表示u_f以及社会化关系知识表

(c) 2022 中国中文信息学会计算语言学专业委员会

示u_r三种异构用户知识与文本表示进行融合，实现面向用户主观差异性建模的隐式情感表示学习。

线性加权融合 直接将三种异构用户知识与隐式情感文本表示进行线性组合并通过神经网络层映射到融合特征空间中，得到最终的融合表示向量h_o，定义为：

$$h_o = \tanh(W_m h_{mpoa} + W_f u_f + W_c u_c + W_r u_r) \tag{10}$$

其中，W_m、W_f、W_c和W_r为融合权重参数，随机初始化再通过训练优化。

堆叠注意力融合 将异构用户知识与隐式情感文本表示通过堆叠注意力模型进行融合，可以自动学习各嵌入知识的融合权重。将u_c、u_f以及u_r分别作为堆叠注意力模型的查询，以隐式情感文本表示h_{mpoa}作为键和值，三种异构用户知识与隐式情感文本的融合表示定义为：

$$\begin{aligned} c_{fm}^{(k)} &= stacked - attention(u_f, h_{mpoa}, h_{mpoa}) \\ c_{cm}^{(k)} &= stacked - attention(u_c, h_{mpoa}, h_{mpoa}) \\ c_{rm}^{(k)} &= stacked - attention(u_r, h_{mpoa}, h_{mpoa}) \end{aligned} \tag{11}$$

其中，k为堆叠注意力模型的层数。在堆叠多头注意力机制的每一层，利用门控机制得到权重向量z_{fm}、z_{cm}和z_{rm}，来控制用户内容知识、社会化属性知识和社会化关系知识对输出结果的贡献，计算过程如下：

$$\begin{aligned} z_{fm} &= \sigma(W_f c_{fm} + W_{fm} h_{mpoa} + b_f) \\ z_{cm} &= \sigma(W_c c_{cm} + W_{cm} h_{mpoa} + b_c) \\ z_{rm} &= \sigma(W_r c_{rm} + W_{rm} h_{mpoa} + b_r) \end{aligned} \tag{12}$$

利用学习得到门控权重与异构用户知识、文本信息加权综合得到最终的融合异构用户知识的隐式情感表示向量h_o，计算公式为：

$$h_o = h_{mpoa} + z_{fm} \otimes c_{fm} + z_{cm} \otimes c_{cm} + z_{rm} \otimes c_{rm} \tag{13}$$

(3) 模型输出层

在输出层，使用全连接层将融合异构用户知识的隐式情感文本表示映射到分类空间并使用softmax函数进行归一化得到各类别概率分布$\tilde{y} = softmax(W_o h_o + b_o)$，$W_o$，$b_o$分别为全连接分类层的参数矩阵与偏置。使用交叉熵损失函数来度量预测$\tilde{y}$和真实标签y之间的损失$\mathcal{L}_{cls}$，并加入MPOA模型(Jiyao et al., 2020)的多极性注意力查询向量的正交损失$\mathcal{L}_{ort}$，如公式(14)所示，其中q_i为第i种极性注意力的查询向量，$q_i/q_j \in \{q_{pos}, q_{neg}, q_{neu}\}$，$q_i \neq q_j$。

$$\mathcal{L}_{ort} = \sum \left| \frac{q_i \cdot q_j}{|q_i| \cdot |q_j|} \right| \tag{14}$$

模型总损失函数定义为公式(15)：

$$\mathcal{L} = \lambda \mathcal{L}_{cls} + (1 - \lambda) \mathcal{L}_{ort} \tag{15}$$

其中，参数λ用于调节分类损失$\mathcal{L}_{cls}$与正交损失$\mathcal{L}_{ort}$的比重。

3 实验

3.1 数据集与评价指标

本文采集、整理并构建了一个用户个性化通用情感分析语料库。原始数据来源于2013-2015年的微博数据，包含有约2000万条原始数据，其中包含了用户id、文本内容等基本信息；性别、地域等用户社会化属性信息；以及用户关注/被关注关系等社会关系信息，所有数据已做脱敏处理。构造数据集时，利用了触发词+情感词典对微博内容进行筛选，存在这样规则的句子前后若有三句及三句以上的句子，视为有效的用户内容知识，保留符合规则的内容知识以及对应用户的各类信息，以保证数据具有较为丰富的内部内容知识，并按句子进行分割标注。标注过程采用众包自动标注+人工校对的方式完成，人工校对率为47.87%。我们基于该语料库构

(c) 2022 中国中文信息学会计算语言学专业委员会

建了2个数据集，D-implicit为隐式情感分析数据集，使用触发词+情感词典筛除了含有显式情感的标注句子，D-general在D-implicit的基础上保留了一部分显式情感句，以验证模型在通用情感分析任务上的泛化能力。两个数据集的标签分布统计如下表2所示。以构建的用户个性化隐式情感分析语料库作为实验数据集，按照5:1:1的比例随机采样对其进行划分得到训练集、验证集和测试集。本文使用各情感倾向标签的精确率(precision)、召回率(recall)、F1值，以及宏平均F1值(F-marco)来评估模型的性能。

	无情感倾向	褒义	贬义	合计标注句子	微博数	用户数	关系边数	平均微博数
D-implicit	4803	3636	4264	12703	11035	3147	5083	3.51
D-general	8763	7272	7791	23836	14714	3546	6049	4.15

Table 2: 用户个性化通用情感分析数据集分析统计表

3.2 实验设置与对比模型

(1) 实验设置 用户社会化表示模块中的堆叠注意力层数(公式(1))与异构知识融合层中的堆叠注意力层数(公式(11))k分别设置为3和3；用户社会化属性最大序列长度(性别，所属地域，个性签名)$[g,\ g,\ m]$=[6, 6, 24]；用户内容信息序列最大长度h=192；token向量维度d=768；Transformer编码器层数设置为6；公式(15)中损失函数系数λ=0.1；使用BERT-Adam优化模型，dropout=0.5，η=2e-5；实验使用pytorch[1] 框架开发，运行环境为Ubuntu22.04+RTX3090*2。

在实验过程中，为了探索不同的异构用户知识学习以及融合方法的效果，我们设置了五组不同的模型组合方案分别做了五种实验，如下表3所示。

模型	社会化属性知识	用户内容知识	社会化关系知识	融合方法
HELENE-1	BERT	BERT+multi-attention	UD-GCN	线性加权融合
HELENE-2	BERT	BERT+Transformer编码器	UD-GCN	stacked-attention
HELENE-3	BERT+stacked-attention	BERT+Transformer编码器	UD-GCN	stacked-attention
HELENE-4	BERT+stacked-attention	BERT+multi-attention	UD-GCN	线性加权融合
HELENE-5	BERT+stacked-attention	BERT+Transformer编码器	UD-GCN	线性加权融合

Table 3: 实验方案模型组合设置

(2) 对比模型 本文研究的用户个性化隐式情感分析任务本质上属于文本分类问题。我们使用在中文文本分类和隐式情感分析中表现出色的流行模型作为基线对比模型以验证本文提出的模型的效果，基线模型主要包括：BERT(Devlin et al., 2018)、RoBERTa(Ott et al., 2019)、BERT-CNN、BERT-RNN、BERT-RCNN、BERT-DPCNN[2]、ERNIE(Sun et al., 2020)、MPOA(Jiyao et al., 2020)、C-MPOA(王素格et al., 2021)、KG-MPOA(Liao et al., 2022)。

3.3 实验结果与分析

用户个性化隐式情感分析任务要求该模型识别包含隐式情感的句子，并对其情感极性进行分类。各实验模型在D-implicit数据集上的结果如表4所示。P，R分别表示准确率和召回率。

本文实验主要从三个方面进行分析，分别是基线模型对比、消融实验对比以及模型参数实验对比。与基线模型对比，我们提出的模型在传统的只对文本情感分类的基础上，融入了用户的知识信息，使具备处理主观差异性的能力。在消融实验中，将组成模型的各个组块分别去掉，探究各类型异构用户知识对整体的影响。在参数实验对比中，分别探索Transformer不同层数与多头自注意机制、堆叠的多头注意力不同层数对结果的提升效果，使得模型能够达到最优。

表4结果表明，与各基线模型相比，我们提出的模型通过融入了用户的知识信息，从内部外部两个维度对其进行画像建模并与隐式情感文本进行有效地融合表示，使得隐式情感分析模型能够充分利用用户信息，具备处理主观差异性的能力。其中，HELENE-5模型整体取得了最优的个性化隐式情感分析效果，相比于各基线模型在F-marco指标上取得了1.9%-9.8%的性能提

[1] https://pytorch.org/

[2] https://github.com/649453932/BERT-Chinese-Text-Classification-Pytorch

(c) 2022 中国中文信息学会计算语言学专业委员会

模型	无情感倾向			隐式褒义			隐式贬义			F-macro
	P	R	F1	P	R	F1	P	R	F1	
BERT	0.491	0.630	0.552	0.667	0.356	0.464	0.579	0.686	0.628	0.548
RoBERTa	0.541	0.414	0.469	0.564	0.500	0.530	0.555	0.748	0.637	0.545
BERT-CNN	0.563	0.482	0.519	0.572	0.524	0.547	0.581	0.714	0.641	0.569
BERT-RNN	0.469	0.622	0.535	0.627	0.242	0.349	0.554	0.714	0.624	0.503
BERT-RCNN	0.629	0.414	0.499	0.558	0.552	0.539	0.559	0.786	0.653	0.564
BERT-DPCNN	0.579	0.430	0.494	0.592	0.446	0.509	0.537	0.808	0.645	0.549
ERNIE	0.567	0.482	0.521	0.590	0.448	0.509	0.577	0.802	0.671	0.567
MPOA	0.588	0.496	0.538	0.600	0.450	0.514	0.566	0.796	0.662	0.571
KG-MPOA	0.441	0.614	0.513	0.564	0.300	0.392	0.556	0.598	0.576	0.498
C-MPOA	0.496	0.652	0.563	0.637	0.428	0.512	0.632	0.682	0.656	0.577
HELENE-1	0.485	0.334	0.396	0.524	0.550	0.537	0.604	0.762	0.674	0.535
HELENE-2	0.530	0.570	0.549	0.572	0.506	0.537	0.631	0.656	0.643	0.576
HELENE-3	0.529	0.592	0.558	0.593	0.496	0.540	0.630	0.658	0.644	0.581
HELENE-4	0.524	0.454	0.487	0.575	**0.586**	**0.580**	0.625	0.696	0.658	0.575
HELENE-5	0.565	0.480	0.519	0.591	0.564	0.577	**0.635**	0.760	**0.692**	**0.596**

Table 4: 用户个性化隐式情感分析任务实验结果对比(D-implicit数据集)

升。说明用户内容知识、社会化属性和关系知识对于隐式情感的识别和倾向性分析具有重要的作用。考虑各基线模型的性能，预训练模型(BERT、RoBERTa、ERNIE等)能为深层语义分析提供基础的语义知识表示。在此基础上，通过对情感极性差异性特征建模(MPOA)并引入上下文信息(C-MPOA)能够显著提升模型的识别隐式情感的能力。对于KG-MPOA模型，由于本文标注的语料库句子相对较短，导致能够引入的外部常识知识规模远小于原始文献中引入知识，导致模型未能充分利用外部常识扩展文本语义。这也说明针对用户个性化隐式情感分析研究，高效地引入外部情感常识知识对于隐式情感分析性能提升具有巨大的挖掘潜力。

对比5种不同的HELENE模型，通过BERT+堆叠注意力，BERT+Transformer编码器以及UD-GCN分别对用户社会化属性、内容和社会化关系进行表示(HELENE-3与HELENE-5)能够达到最优的建模效果。而在异构用户数据融合层，使用线性加权融合(HELENE-5)相对于使用堆叠注意力机制(HELENE-3)效果更佳，分析原因我们认为，线性加权融合相对具有更少的信息损失，同时堆叠注意力模型过高地强调了隐式文本表示h_{mpoa}的重要性，导致融合过程损失了较多的用户知识信息。

我们同时在D-general数据集上开展了相似实验，以验证模型在用户个性化情感分析任务上的泛化性能。实验结果如下表5所示。

模型	无情感倾向			褒义			贬义			F-macro
	P	R	F1	P	R	F1	P	R	F1	
BERT	0.524	0.698	0.599	0.723	0.479	0.577	0.660	0.667	0.664	0.615
RoBERTa	0.514	0.750	0.610	0.728	0.511	0.600	0.680	0.570	0.620	0.610
BERT-CNN	0.529	0.699	0.602	0.669	0.614	0.641	0.717	0.547	0.621	0.621
BERT-RNN	0.538	0.704	0.610	0.674	0.578	0.622	0.696	0.580	0.633	0.622
BERT-RCNN	0.598	0.579	0.588	0.619	0.683	0.649	0.681	0.631	0.655	0.631
BERT-DPCNN	0.510	0.701	0.590	0.737	0.506	0.600	0.645	0.605	0.624	0.605
ERNIE	0.541	0.708	0.613	0.749	0.512	0.608	0.680	0.685	0.683	0.635
KG-MPOA	0.599	0.423	0.496	0.617	0.716	0.663	0.623	0.706	0.662	0.607
MPOA	0.568	0.677	0.618	0.697	0.631	0.662	0.701	0.634	0.666	0.648
C-MPOA	0.655	0.512	0.575	0.659	0.714	0.685	0.666	0.755	0.708	0.656
HELENE-1	0.615	0.488	0.544	0.686	0.665	0.675	0.640	0.791	0.708	0.642
HELENE-2	0.604	0.626	0.615	0.684	0.640	0.661	0.682	0.702	0.692	0.656
HELENE-3	0.592	0.647	0.618	0.688	0.636	0.661	0.703	0.691	0.697	0.659
HELENE-4	0.631	0.629	**0.630**	0.709	0.659	0.683	0.694	0.745	0.718	0.677
HELENE-5	**0.676**	0.580	0.624	0.681	0.695	**0.688**	0.683	**0.766**	**0.722**	**0.678**

Table 5: 用户个性化情感分析任务实验结果对比(D-general数据集)

从表5结果可知，相较于隐式情感，更简单的显式情感整体提高了所有模型的分类性能。所提出的模型除HELENE-1外均优于所有的基线模型，说明用户个性化知识对于显式情感建模同样具有重要的意义。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.4 消融实验与分析

HELENE-5与HELENE-3通过使用BERT+堆叠注意力，BERT+Transformer编码器以及UD-GCN分别对用户社会化属性、内容和社会化关系进行表示，在上节的用户个性化隐式/通用情感分析任务中分别取得了最优和次优的结果。我们针对2个模型各异构用户知识构成进行了消融分析，以进一步探究各类型异构用户知识对整体的影响。在D-implicit和D-general数据集上的消融实验结果如下表6和7所示，-f、-c、-r分别表示原HELENE-3/5模型去除了用户社会化属性知识的模型、用户内容知识的模型以及用户社会化关系知识的模型。

模型	无情感倾向			隐式褒义			隐式贬义			F-macro
	P	R	F1	P	R	F1	P	R	F1	
HELENE-3	0.529	0.592	0.558	0.593	0.496	0.540	0.630	0.658	0.644	0.581
HELENE-3-f	0.530	0.570	0.549	0.572	0.506	0.537	0.631	0.656	0.643	0.576
HELENE-3-c	0.525	0.580	0.551	0.587	0.506	0.544	0.621	0.642	0.631	0.575
HELENE-3-r	0.524	0.586	0.553	0.583	0.498	0.537	0.634	0.652	0.643	0.578
HELENE-5	0.565	0.480	0.519	0.591	0.564	0.577	0.635	0.760	0.692	0.596
HELENE-5-f	0.551	0.456	0.499	0.546	0.582	0.563	0.640	0.708	0.672	0.578
HELENE-5-c	0.620	0.394	0.482	0.541	0.622	0.579	0.580	0.798	0.672	0.577
HELENE-5-r	0.554	0.454	0.499	0.579	0.566	0.572	0.607	0.730	0.663	0.578

Table 6: HELENE-3/5模型在隐式情感分析任务上消融实验对比(D-implicit数据集)

模型	无情感倾向			褒义			贬义			F-macro
	P	R	F1	P	R	F1	P	R	F1	
HELENE-3	0.592	0.647	0.618	0.688	0.636	0.661	0.703	0.691	0.697	0.659
HELENE-3-f	0.579	0.631	0.604	0.675	0.612	0.642	0.686	0.689	0.687	0.644
HELENE-3-c	0.591	0.632	0.611	0.680	0.635	0.657	0.694	0.691	0.692	0.653
HELENE-3-r	0.588	0.629	0.608	0.684	0.627	0.654	0.684	0.694	0.689	0.650
HELENE-5	0.676	0.580	0.624	0.681	0.695	0.688	0.683	0.766	0.722	0.678
HELENE-5-f	0.652	0.591	0.620	0.690	0.671	0.680	0.679	0.761	0.718	0.673
HELENE-5-c	0.582	0.668	0.622	0.689	0.633	0.660	0.699	0.652	0.675	0.652
HELENE-5-r	0.661	0.562	0.608	0.678	0.681	0.680	0.667	0.764	0.712	0.667

Table 7: HELENE-3/5模型在情感分析任务上消融实验对比(D-general数据集)

由表6和7可知，在去掉异构用户知识各组成部分后，在隐式情感或通用情感分析任务上模型的性能均有所降低。以HELENE-5模型为例，在去除了用户社会化属性知识、用户内容知识、用户社会化关系知识后，在两个数据集上的F-macro分别降低了1.8%，1.9%，1.8%(D-implicit)和0.5%，2.6%，1.1%(D-general)，验证了通过用户发布历史内容知识，可以从中分析出大量的用户潜在信息，对于精准用户画像建模具有重要的作用，这可为后续相关工作提供借鉴。除此之外，用户的社会化关系准确反映了“人以类聚”的特征，通过用户有向图关系建模可以较好的学习用户的群体类别特征。用户社会化属性知识由于所含内容相对较少，能够提供的用户信息有限，但仍能较准确地为涉及地域、性别身份信息的用户画像提供支持。

3.5 案例分析

本文对模型预测结果进行案例分析，以表8中如下两个数据为例分析解释模型预测结果，表中内容知识*IS*表示待分析隐式情感句。

在表8中，案例1中待识别的句子为“边走湿透的裤子里面水直往鞋子里流！”。根据待识别句子所在的用户内容信息中“雨”、“真变态”、“不像样”，对应用户的社会化属性信息中“得不到”、“骚动”此类偏消极词信息，以及该用户社会化关注关系知识，与待识别的句子信息融合，我们的模型将其判别为贬义类型。而在案例2中，待识别的句子为“夏天来了，去年蝉蜕下的壳还会在吗。”，本身带了一种因怀念过去而感到悲凉的气息，却被判别为褒义。我们认为由于在对应的内容信息中的“鸽子”、“放飞蓝天”、“过得更好”等偏积极的信息在识别情感时对模型产生了误导，因此导致模型出现错误预测。

4 结论与展望

本文提出了一种基于异构用户知识融合的隐式情感分析的模型HELENE，从用户的内部外部两个维度对其进行画像建模，挖掘了用户的内容、社会化属性和社会化关系三种不同的异构

(c) 2022 中国中文信息学会计算语言学专业委员会

案例	隐式情感句	内容知识	社会化属性知识	真实标签	模型预测
1	边走湿透的裤子里面水直往鞋子里流！	这雨下的真变态！*IS* 上身更不用说的惨全身湿的不像样了！，赶紧热水澡搞起.....	[“女”, “湖南长沙”, “得不到的永远在骚动！”]	贬	贬
2	夏天来了，去年蝉蜕下的壳还会在吗。	这种天气就像鸽子成群放飞蓝天，齐齐拍动翅膀的爽朗声响，心情也被太阳晒得明晃晃的。*IS* 我相信，怀念只是为了让我们过得更好。	[“男”, “福建厦门”, “时间它只负责流动，不负责育你成长。”]	贬	褒

Table 8: 模型预测结果案例分析

知识，并将其与隐式情感文本进行有效地融合表示，使得隐式情感分析模型能够充分利用用户个性化信息，具备处理主观差异性的能力。本文同时构建了一个具有覆盖用户属性-关系-文本信息的较大规模通用情感分析语料库，可同时满足面向用户个性化建模的隐式或显式情感分析相关研究任务的需要。在构建数据集上的实验结果表明，文本提出的方法相比于基线模型在个性化隐式情感分析任务上F-marco取得了1.9%-9.8%的效果提升。

未来的工作将主要集中于针对多视角用户建模的隐式情感分析方面，以期实现更精细化的用户表示与知识融合。

参考文献

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. 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 1684–1695, Hong Kong, China, November. Association for Computational Linguistics.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. 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 340–350, Online, August. Association for Computational Linguistics.

Huan-Yuan Chen and Hsin-Hsi Chen. 2016. Implicit polarity and implicit aspect recognition in opinion mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

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.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Liao Jian, Li Yang, and Wang Suge. 2016. The constitution of a fine-grained opinion annotated corpus on weibo. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 227–240. Springer.

Liao Jian, Wang Suge, and Li Deyu. 2019. Identification of fact-implied implicit sentiment based on multi-level semantic fused representation. *Knowledge-Based Systems*, 165:197–207.

Wei Jiyao, Liao Jian, Yang Zhenfei, Wang Suge, and Zhao Qiang. 2020. Bilstm with multi-polarity orthogonal attention for implicit sentiment analysis. *Neurocomputing*, 383:165–173.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. 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.

第二十一届中国计算语言学大会论文集，第523页-第534页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

Jian Liao, Min Wang, Xin Chen, Suge Wang, and Kai Zhang. 2022. Dynamic commonsense knowledge fused method for chinese implicit sentiment analysis. *Information Processing & Management*, 59(3):102934.

Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. 2021. Interest-aware message-passing gcn for recommendation. In *Proceedings of the International Conference of World Wide Web*, pages 1296–1305.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Chenyang Lyu, Jennifer Foster, and Yvette Graham. 2020. Improving document-level sentiment analysis with user and product context. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6724–6729, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

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. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Y. Samih and K. Darwish. 2021. A few topical tweets are enough for effective user-level stance detection. In *The 16th conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Chen Shiyun, Lin Xin, Xiao Yanghua, and He Liang. 2019. Sentiment commonsense induced sequential neural networks for sentiment classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1021–1030.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

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.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Tao Qi, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with heterogeneous user behavior. 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 4874–4883.

J. Zheng, Q. Li, and J. Liao. 2021a. Heterogeneous type-specific entity representation learning for recommendations in e-commerce network. *Information Processing & Management*, 58(5):102629.

J. Zheng, Q. Li, J. Liao, and S. Wang. 2021b. Explainable link prediction based on multi-granularity relation-embedded representation. *Knowledge-Based Systems*, 230(15):107402.

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.

Enguang Zuo, Hui Zhao, Bo Chen, and Qiuchang Chen. 2020. Context-specific heterogeneous graph convolutional network for implicit sentiment analysis. *IEEE Access*, 8:37967–37975.

廖健. 2018. 基于表示学习的事实型隐式情感分析研究. Ph.D. thesis, 山西大学.

潘东行, 袁景凌, 李琳, and 盛德明. 2020. 一种融合上下文特征的中文隐式情感分类模型. 计算机工程与科学, 42(2):10.

王素格, 王敏, 廖健, and 陈鑫. 2021. 融合上下文信息的隐式情感句判别方法. 山西大学学报*(*自然科学版*)*, 165:1–7.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于主题提示学习的零样本立场检测方法

陈子潇*, 梁斌, 徐睿峰*
(哈尔滨工业大学（深圳）计算机科学与技术学院, 广东省 深圳市 518071
{chenzixiao, bin.liang}@stu.hit.edu.cn
xuruifeng@hit.edu.cn)

摘要

零样本立场检测目的是针对未知目标数据进行立场极性预测。一般而言，文本的立场表达是与所讨论的目标主题是紧密联系的。针对未知目标的立场检测，本文将立场表达划分为两种类型：一类在说话者面向不同的主题和讨论目标时表达相同的立场态度，称之为目标无关的表达；另一类在说话者面向特定主题和讨论目标时才表达相应的立场态度，本文称之为目标依赖的表达。对这两种表达进行区分，有效学习到目标无关的表达方式并忽略目标依赖的表达方式，有望强化模型的可迁移能力，使其更加适应零样本立场检测任务。据此，本文提出了一种基于主题提示学习的零样本立场检测方法。具体而言，受自监督学习的启发，本文为了零样本立场检测设置了一个代理任务框架。其中，代理任务通过掩盖上下文中的目标主题词生成辅助样本，并基于提示学习分别预测原样本和辅助样本的立场表达，随后判断原样本和辅助样本的立场表达是否一致，从而在无需人工标注的情况下判断样本的立场表达是否依赖于目标的代理标签。然后，将此代理标签提供给立场检测模型，对应学习可迁移的立场检测特征。在两个基准数据集上的大量实验表明，本文提出的方法在零样本立场检测任务中相比基线模型取得了更优的性能。

关键词： 零样本立场检测；提示学习；代理任务

A Topic-based Prompt Learning Method for Zero-Shot Stance Detection

Chen Zixiao*, Liang Bin, Xu Ruifeng*
(Harbin Institude of Technology (Shenzhen)
School of Computer Science and Technology, Guangdong Shenzhen 518071
{chenzixiao, bin.liang}@stu.hit.edu.cn
xuruifeng@hit.edu.cn)

Abstract

Zero-shot stance detection (ZSSD) aims to detecting the stance of previously unseen targets during the inference stage. It is generally believed that the stance expression in a sentence is closely related to the stance target and topics discussed.We divide stance expressions of speakers into two categories: target-invariant and target-specific categories. Target-invariant stance expressions carry the same stance polarity regardless of the targets they are associated with. On the contrary, target-specific stance expressions only co-occur with certain targets. As such, it is important to distinguish these two types of stance features to boost stance detection ability.In this paper,we develop an effective approach to distinguish the types of target-related stance expressions to better

* 通讯作者

第二十一届中国计算语言学大会论文集，第535页-第544页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

learn transferable stance features.To be specific,inspired by self-supervised learning,we frame the stance-feature-type identification as a pretext task in ZSSD.We apply prompt learning to predict changing relationship between stance polarity labels and topic information in pretext task.This essentially allows the model to learn transferable stance features. Extensive experiments on two benchmark datasets show that the proposed method obtains an improved performance than the baseline in ZSSD.

Keywords: Zero-shot Stance Detection , Prompt Learning , Pretext Task

1 引言

立场检测(Stance Detection)是自然语言处理(Natural Language Processing)领域(Kachuee et al., 2021)的一个重要任务。立场检测的目的是识别说话者面向特定目标、主题和主张时表达的立场与态度(Somasundaran et al., 2010; Augenstein et al., 2016; Mohammad et al., 2016)。在以往研究聚焦的目标集合内部的立场检测任务中(Gunel et al., 2020)，训练集和测试集共享可见的目标集合。然而在现实生活中，存在大量目标相对于现有立场检测模型是未知的样例(Allaway et al., 2020)。为了解决这个问题进而出现了零样本立场检测(Zero-shot Stance Detection)任务，旨在面向未知目标进行立场检测。

现有方法引进了注意力机制(Allaway et al., 2020)和额外知识在已知目标和未知目标间捕捉可迁移立场特征。但在实践中，这些方法的捕捉能力不强，对外部能力依赖大，缺少对数据集本身信息的充分利用。一般认为，一句话的立场表达是与所讨论的目标和主题紧密联系的，本文将说话者的立场表达划分为两种类型：一类在说话者面向不同的主题和讨论目标时表达相同的立场态度，本文称之为目标无关的表达；另一类在说话者面向特定主题和讨论目标时才表达相应的立场态度，本文称之为目标依赖的表达。在立场检测任务中，区分说话者表达的立场是目标依赖还是目标无关是十分重要的。前者随着面向目标的改变有可能改变立场，而后者则不会。这一思路仅从数据本身挖掘信息，能够有效加强模型的可迁移学习能力，且无需依赖外部知识或特殊网络结构。为了说明本文提出的方法，表1给出了一些目标依赖和目标无关的立场表达的样例。在目标及其相关主题词被隐藏后，对于目标无关的立场表达，样例的立场态度并未发生改变；然而对于目标依赖的立场表达，失去与目标间的联系后立场态度发生了改变。因此，在零样本立场检测任务中，当寻找可迁移特征时，模型需要强化对目标无关的立场表达特征的学习并排除目标相关的立场表达特征的干扰。

目标无关的立场表达
目标：Feminist Movement 立场极性：Against 样例句：feminist only want the same benefiting right as men not those harmful ones 隐藏目标和主题词后的句子：$[MASK]$ only want the same benefiting $[MASK]$ as $[MASK]$ not those harmful ones 隐藏信息后样例句的立场表达：Against
目标依赖的立场表达
目标：Donald Trump 立场极性：Against 样例句：white terrorism is alive and well 隐藏目标和主题词后的句子：$[MASK]$ is alive and well 隐藏信息后样例句的立场表达：Favor

Table 1: 立场检测中立场表达类型样例

本文的主要贡献概括如下：

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

- 本文基于提示学习，从一种新的角度探讨了零样本立场检测问题。该方法通过主题词的保留和掩盖，以提示学习来自动学习立场表达是否依赖目标，进而将目标无关的立场表达特征用于未知目标的立场检测任务。
- 本文提出了一种创新的特征生成方法。该方法通掩盖训练样本的目标词及其相关主题词来生成辅助训练样本，并使用提示学习，通过自监督学习的方式利用预训练语言模型的先验知识判断立场表达特征与目标的关联性。
- 在两个公开数据集上的实验结果表明，本文提出的方法在零样本立场检测任务中取得了比基线模型更优的性能。

2 相关工作

早期对零样本立场检测方法的研究多集中在目标集合内部的立场检测，也就是训练集和测试集共享目标的检测任务(Du et al., 2017; Sun et al., 2018; Li et al., 2019; Siddiqua et al., 2019; Kawintiranon et al., 2021)。跨目标的立场检测是一种和零样本立场检测相似的任务，该任务基于一个已知目标训练分类器对一个未知目标的数据进行立场预测(Xu et al., 2018; Wei et al., 2019; Zhang et al., 2020; Liang et al., 2021)。现存跨目标的立场检测研究通常使用了基于注意力机制(Xu et al., 2018; Wei et al., 2019)或图网络(Zhang et al., 2020; Liang et al., 2021)的模型，根据训练集的目标学习目标关联特征，然后用于与目标数据集相近的测试集的预测。不同于跨目标的立场检测任务，零样本立场检测希望能够自动判断各种未知目标数据的立场结果。在这一任务要求下，Conforti等(Conforti et al., 2020)搭建了一个大范围专家标注的立场检测数据集，其中测试集的目标相对于训练集是不可见的。Allaway等(Allaway et al., 2020)搭建了一个零样本立场检测数据集，该数据集拥有大量的主题，相关话题类别十分广阔。Allaway等(Allaway et al., 2020)还在该数据集的基础上提出了一个主题分组的注意力模型来捕捉目标和通用主题表示间的关系。在另一项研究中，Allway等(Allaway et al., 2021)将一个用于目标内部立场检测的数据集(Mohammad et al., 2016)应用到零样本立场检测中，并使用了对抗学习来提取样本无关的可迁移特征。此外，Liu等(Liu et al., 2021)同时从结构层面和语义层面引入相关的外部知识，提出了一种基于BERT(Devlin et al., 2019)的常识增强图模型来解决零样本立场检测任务。

在特征学习的监督信号是从数据自动生成的情况下，自监督学习有着良好的表现。近年来自监督学习有一种发展方向就来自自动设计的预测任务，或者常被称作代理任务(Zhang et al., 2016; Gidaris et al., 2018; Chen et al., 2020)。很多现有的计算机视觉研究领域的方法，在包括拼图问题(Noroozi et al., 2016)、旋转预测(Gidaris et al., 2018)等任务上都设计了启发式的无注释代理任务，以便为目标问题提供特征学习的替代监督信号(Zhang et al., 2016; Gidaris et al., 2018; Chen et al., 2020; Larsson et al., 2016; Simard et al., 2021)。由此，受现有的自监督方法的启发，本文设计了一个代理任务来挖掘立场表达是否关联目标这一重要特征并用于可迁移学习任务。

通常定义下，提示学习是一种通过在文本输入部分增加提示信息，将下游学习任务转化为文本生成任务的一种学习方法。Fabio等(Petroni et al., 2019)提出了LAMA，一种将关系抽取任务转化为填空任务的提示学习方法，并取得了比基于外部知识的方法更好的性能表现。Shin等(Shin et al., 2020)将提示学习方法应用到文本分类和文本蕴含识别任务上，在没有改变预训练语言模型的情况下完成了这些问题的良好预测。Timo等(Schick et al., 2021)提出了PET，一种通用的半监督训练模式，适用于一系列自然语言处理问题。PET尤其在自动标注数据和扩充训练集的任务上取得了优异表现。受这些方法的启发，本文设置了一个运用提示学习方法自动标注文本立场极性的代理任务，以有效判断立场主题信息改变是否会同样改变立场表达。

3 模型方法

本章将详细描述本文提出的一种基于主题提示学习的零样本立场检测方法（Data Augmentation for **S**tance **T**opic features via **P**rompt **L**earning Pretext Task，ST-PL），其总体结构如图1所示。本方法的核心是在模型训练阶段之前设置一个代理任务，该代理任务应用提示学习在自监督立场主题特征这一维度上进行了数据增强，以帮助训练阶段更好地学习可迁移的立场特

征。本方法主要分为四步：1）对每一个目标使用主题模型挖掘需要隐藏的相应目标和主题信息；2）使用训练后的提示学习模型对隐藏主题信息后的训练集数据判断立场极性，判断结果与原立场极性不同的数据则认为包含目标依赖的特征；3）通过移除训练集数据里包含目标依赖特征的数据进行数据增强；4）在增强后数据上训练立场检测主任务模型，以加强模型对零样本立场检测主任务的预测能力。经过本文提出的ST-PL方法处理后，后续零样本立场检测模型的训练和预测阶段的表现得到了提升。

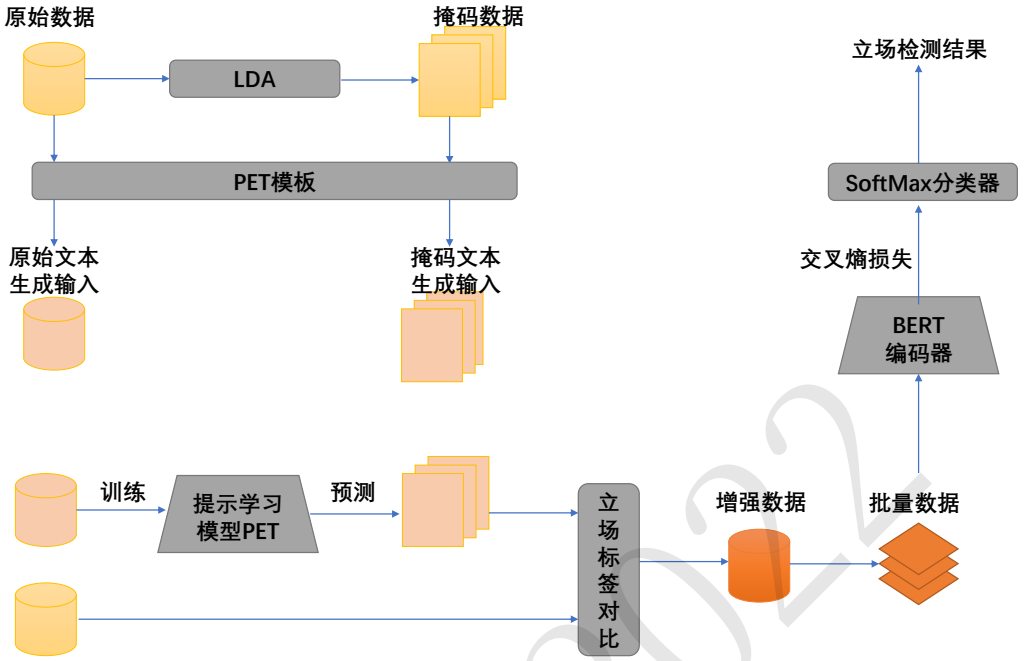


Figure 1: ST-PL方法的总体架构

3.1 任务定义

为不失一般性，假设有一个面向已知目标的已标注实例集合$\mathcal{D}_s = \{(r_s^i, t_s^i, y_s^i)\}_{i=1}^{N_s}$和一个面向未知目标的未标注实例集合$\mathcal{D}_d = \{(r_d^i, t_d^i)\}_{i=1}^{N_d}$。其中$y_s^i$是一个面向已知目标$t_s^i$的已标注实例的立场标签。$N_s$和$N_d$是已知目标数据集和未知目标数据集的数据量。$\mathcal{D}_s$和$\mathcal{D}_d$间没有重合的立场目标。零样本立场检测的目标是对每个来自数据集$\mathcal{D}_s$的面向已知目标t_s^i的句子r_s^i，训练一个模型能够在来自数据集$\mathcal{D}_d$的面向未知目标t_d^i的句子r_d^i上具有泛用的预测能力。

3.2 基于代理任务的数据增强方法

一句话的立场表达是与所讨论的目标和主题紧密联系的。一部分立场表达是较为通用的，可以出现在各种所讨论主题不同的场合而不改变其立场态度，而另一部分则大多局限于某些特定目标和议题。因此本文将立场表达划分为目标无关的和目标依赖的两类。受现有方法(Zhang et al., 2016; Gidaris et al., 2018; Chen et al., 2020; Simard et al., 2021; Schick et al., 2021)启发，为了区分这两种立场特征类型以更好地学习到可用于在零样本立场检测任务中进行迁移学习的立场数据，本文探索了一个结合自监督代理任务与主题提示学习的数据增强策略，用于为下游任务学习提供新的监督信号。

3.3 基于主题提示学习的代理任务框架

对于每一个由句子r_s^i和目标t_s^i组成的实例，本文使用提示学习方法PET(Schick et al., 2021)训练立场检测模型$\mathcal{M}$。PET通过预训练语言模型直接预测被转化为文本生成问题的立场检测任务。这里PET使用的训练模板如表3所示。随后，本文对于每一个目标所对应的训练实例，隐藏这些实例中的目标词和相关主题词，由此得到了隐去部分信息的候选实例。这种候选实例生成方式的目的是运用后续训练的模型学习得到立场表达和目标是否具有关联性。在这

里，提取目标相关主题词所用的是LDA（Latent Dirichlet Allocation），一种基于隐含狄利克雷分布的主题模型。相关样例由表2所示。然后，使用训练后的$\mathcal{M}$预测被隐去目标信息和相关主题词的句子所组成的新训练集的立场极性，记录预测立场极性与未隐去信息的原数据的立场极性的异同。此处使用提示学习这一并非直接适用于立场检测的文本分类任务的方法，其目的在于考虑到提示学习使用文本提示信息帮助模型学习的特性，可以通过对主题词信息的提示，有效挖掘样例中的主题词和其他词的关联，能够捕捉隐藏信息前后对句子带来的变化，以提高本文设置的代理任务场景对立场目标相关表达的学习效果。

目标	主题词
Donald Trump	right,vote,obama,president,america
Hillary Cliton	women,president,right,campaign,wakeupamerica

Table 2: 主题词样例

目标无关的立场表达
目标：Feminist Movement 立场极性-生成用词：Against-No 样例句：Feminist only want the same benefiting right as men not those harmful ones. 提示学习模板：[目标]，[生成用词]。[样例句]。 模板生成句：Feminist Movement , [?]. Feminist only want the same benefiting right as men not those harmful ones. 模板预测结果：Feminist Movement , **No**. Feminist only want the same benefiting right as men not those harmful ones.

Table 3: 提示学习方法PET的训练模板

3.4 生成增强数据

为了区分目标依赖和目标无关的立场特征，以便更好地学习零样本立场检测中的可迁移立场特征，本文通过自监督学习设计的代理任务自动为训练数据生成辅助监督信号。首先，用一个特殊标记$[MASK]$替代被隐去的目标词和相关主题词，如表1所示。然后，将被隐去信息的句子输入PET模型$\mathcal{M}$以重新预测该实例的立场极性标签。此处参照表3样例，$\mathcal{M}$所预测的文本生成任务句形式为：$[MASK]$, [?]. $[MASK]$ only want the same benefiting right as $[MASK]$ not those harmful ones。[?]处为文本生成任务需要预测的生成用词，对应立场检测任务的立场极性。如果重新预测的标签正确，说明该立场表达不依赖于目标，该实例的立场表达是目标无关的，被标记为标签"*target-invariant*"；否则该实例的立场表达是目标依赖的，被标记为标签"*target-specific*"。训练集中目标依赖的数据将会被剔除以生成增强数据。数据增强后，训练集可以从形式上表示为$\mathcal{D}_s = \{(r_s^i, t_s^i, y_s^i, p_s^i)\}_{i=1}^{N_s}$。

3.5 训练架构

3.5.1 编码器模块

给定一个词语序列$r = \{w_i\}_{i=1}^{n}$ 和对应的目标t，n是文本r的长度。这里使用r和t来表示训练实例的句子和目标。在输入实例时，模型将忽略具有目标依赖标签的实例，以便在训练过程中偏好训练可迁移的立场特征。然后，本方法采用预训练的BERT (Devlin et al., 2019)作为编码器模块并将" $[CLS]r[SEP]t[SEP]$"作为输入以获取每个输入样例的标记$[CLS]$ 的d_m维隐藏表示$\boldsymbol{h} \in \mathbb{R}^{d_m}$:

$$\boldsymbol{H} = \text{BERT}([CLS]r[SEP]t[SEP]), \boldsymbol{h} = \boldsymbol{H}_{[CLS]} \tag{1}$$

对于一个批量的数据用例，用例的隐藏表示可以定义为$\mathcal{B} = \{\boldsymbol{h}_i\}_{i=1}^{N_b}$，其中$N_b$ 是批量的大小。

第二十一届中国计算语言学大会论文集，第535页-第544页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

数据集	目标	Favor	Against	Neutral	Unrelated
SEM16	DT	148	299	260	-
	HC	163	565	256	-
	FM	268	511	170	-
	LA	167	544	222	-
	A	124	464	145	-
	CC	335	26	203	-
WT-WT	CA	2469	518	5520	3115
	CE	773	253	947	554
	AC	970	1969	3098	5007
	AH	1038	1106	2804	2949

Table 4: SEM16和WT-WT数据集的数据分布

3.5.2 立场分类器

该模块将批量中数据的隐藏向量$\mathcal{B} = \{\boldsymbol{h}_i\}_{i=1}^{N_b}$输入到softmax函数的分类器中，以生成立场的预测分布：

$$\hat{\boldsymbol{y}}_i = \text{softmax}(\boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{b}) \tag{2}$$

其中$\hat{\boldsymbol{y}}_i \in \mathbb{R}^{d_y}$是输入实例$x_i$的立场预测概率，$d_y$是立场标签的维度$\boldsymbol{W} \in \mathbb{R}^{d_y \times d_m}$和$\boldsymbol{b} \in \mathbb{R}^{d_y}$是可训练的参数。基于立场预测概率，我们采用实例$\boldsymbol{y}_i$的预测分布和真实分布$x_i$间的交叉熵损失来训练分类器：

$$\mathcal{L}_{class} = -\sum_{i=1}^{N_b}\sum_{j=1}^{d_p} y_i^j \log \hat{y}_i^j \tag{3}$$

3.6 模型训练与预测

在此阶段，本文提出的模型的学习目标是在增强训练集上优化立场检测的监督损失函数$\mathcal{L}_{class}$。总的损失函数$\mathcal{L}$通过将监督损失函数和正则项相加得到：

$$\mathcal{L} = \gamma_c \mathcal{L}_{class} + \lambda ||\Theta||^2 \tag{4}$$

其中γ_c是可调整的超参数，Θ表示模型所有可训练参数，λ表示L_2正则化的系数。

4 实验

4.1 实验数据集

本文在以下两个数据集上进行实验：

- SemEval2016(Mohammad et al., 2016)。该数据集包含在多领域的六个预定义的目标：Donald Trump(DT)，Hillary Clinton (HC)，Feminist Movement(FM)，Legalization of Abortion (LA)，Atheism (A)，Climate Change(CC)。参考Allaway等(Allaway et al., 2021)的做法，本文将其中一个目标的数据作为测试集，其余目标数据作为训练集，并随机选择训练集中15%的数据作为验证集来调整超参。数据集的详情在表4中显示。

- WT-WT(Conforti et al., 2020)。该数据集主要讨论公司之间的并购业务，包含4个目标：CVS_AET (CA)，CI_ESRX (CE)，ANTM_CI (AC)，AET_HUM (AH)。每个实例可能包含如下立场标签：Support（即Favor），Refute（即Against），Comment（即Neutral），Unrelated。参考Conforti等(Conforti et al., 2020)的做法，本文将每个目标都作为测试集并在其余三个目标数据作为训练集，并随机选取15%训练集数据作为验证集。数据集详情在表4中显示。

CCL 2022

4.2 训练设置

本文使用预训练好的768维词嵌入的uncased BERT-base(Devlin et al., 2019)作为编码器。[0]，其中学习率设置为0.000005；根据Xu等(Xu et al., 2018)的做法将L_2正则化系数λ设置为0.00001；优化器使用Adam；批量大小设置为16。本文使用gensim库中的LDA获取每个目标的关联主题词，根据实际数据分布设置T=10，K=10，并去除了重复的主题词。本文训练过程中设置在超过5轮迭代仍未能提升性能时模型将提前停止。下文记录的实验数据均是10次运行试验结果的平均值，以获得统计学上的稳定结果。

4.3 评价指标

对于SemEval2016数据集本文参考Allaway等(Allaway et al., 2021)的做法使用F_{avg}衡量性能，也就是在Favor和Agianst两种立场标签上的F1的平均值。对于WT-WT数据集本文参考Conforti等(Conforti et al., 2020)的做法使用每个目标的Macro F1衡量性能。

4.4 主要结果

在表5中记录了在两个基准数据集上进行零样本立场检测的主要实验结果。BiCond方法来自 (Augenstein et al., 2016),CrossNet方法来自 (Xu et al., 2018)，BERT方法来自 (Devlin et al., 2019)。可以观察到本文提出的ST-PL方法在两个数据集上的性能大都优于传统模型，也优于直接使用提示学习对主任务进行学习，这验证了本文提出的方法在零样本立场检测中的有效性。这表明使用提示学习进行自监督学习代理任务以获取目标无关的特征的监督信号，对于学习相对模型不可见的目标的可迁移立场特征是有效的，并可凭借此特征提高零样本立场检测的性能。

Model	SEM16 (%)						WT-WT (%)			
	DT	HC	FM	LA	A	CC	CA	CE	AC	AH
BiCond	30.5‡	32.7‡	40.6‡	34.4‡	31.0‡	15.0‡	56.5♯	52.5♯	64.9♯	63.0♯
CrossNet	35.6	38.3	41.7	38.5	39.7	22.8	59.1♯	54.5♯	65.1♯	62.3♯
BERT	40.1‡	49.6‡	41.9‡	44.8‡	**55.2**‡	37.3‡	56.0♭	60.5♭	67.1♭	67.3♭
PET	45.6	50.9	49.3	46.7	45.8	32.3	68.6	63.7	70.7	71.5
ST-PL (ours)	**48.4**	**53.7**	**51.2**	**48.1**	52.2	**35.2**	**71.2**	**68.6**	**73.5**	**75.7**

Table 5: ST-PL在两个零样本立场检测数据集上的实验结果。带‡ 符号的结果取自文献 (Allaway et al., 2021)，带♯ 符号的结果取自文献 (Conforti et al., 2020)，带♭ 符号的结果取自 (Liang et al., 2021)

.

Model	SEM16 (%)						WT-WT (%)			
	DT	HC	FM	LA	A	CC	CA	CE	AC	AH
ST-PL (ours)	**48.4**	**53.7**	**51.2**	**48.1**	**52.2**	**35.2**	**71.2**	**68.6**	**73.5**	**75.7**
PET	45.6	50.9	49.3	46.7	45.8	32.3	68.6	63.7	70.7	71.5
w/o Topic Information	38.4	45.0	38.6	41.0	46.4	33.3	67.2	63.2	68.6	71.5
w/o Prompt Learning	46.2	51.6	47.8	46.6	54.4	37.8	69.9	67.1	74.2	73.0
w/o Pretext Task	40.1	49.6	41.9	44.8	55.2	37.3	56.0	60.5	67.1	67.3

Table 6: 消融实验结果

.

4.5 消融研究

在表6中记录了本方法的消融实验结果，其中PET表示直接使用提示学习进行零样本立场检测，w/o Topic Information表示在代理任务中不对主题词进行隐藏，w/o Prompt learning表

[0]由于BERT-base也是本文用于比较的基线模型，故本方法也是基于BERT-base进行构建以保证公平的比较。

(c) 2022 中国中文信息学会计算语言学专业委员会

示在代理任务中不使用提示学习仅使用普通过拟合模型BERT对隐藏主题词后的样例进行立场极性预测，w/o Pretext task表示不使用代理任务直接使用主模型进行零样本文本立场检测。可以看出，ST-PL方法构建代理任务学习主题相关的立场可迁移信息是有效的，因为直接使用PET进行零样本文本检测并不能取得比ST-PL的代理任务+主任务架构更好的成果。该结果也验证了提示学习的提出动机，即提示学习更适合对样例中的主题词和其他词的关联进行学习，并且能够捕捉隐藏信息前后对句子带来的变化，而非直接应用于立场检测这一本质文本分类的任务。此外，对主题词隐藏这一步骤的消融带来了最明显的性能下滑，验证了本方法挖掘主题词信息以关联目标信息这一动机；而其他步骤的任何一个缺失也都会带来性能下降，说明这些设置对零样本文本立场检测的性能提升都是必要的。

4.6 案例分析

在表7中，样例1在本文基于提示学习的代理任务下其可迁移与否的标签均判断正确。可以观察到方法在短句中表现良好，因为短句的立场表达词少且语法结构简单。在样例1中，提示学习模型预测隐去信息后数据的立场极性与原样例的立场极性相同，意味着该样例的立场表达是零样本立场检测中的一个可迁移表达，所以模型为这个样例生成一个目标无关标签。然而，在样例2中，模型对其可迁移与否的标签均判断错误，提示学习模型预测了与原样例不同的立场极性，判定该样例的倾向是随着目标信息的变化而会发生改变的，因此模型为这个样例生成一个目标依赖标签。根据类似样例2的复杂句数据的预测准确度相对偏低现象，可以推测出本课题所提出的方法在处理具有复杂语法结构、相异立场态度词表达和复杂语言现象（如讽刺、俗语等）的长句子时，代理任务由于当前模型立场极性预测能力的客观瓶颈产生了错误的可迁移表达预测，这一问题待未来研究进行进一步改善。

样例1
目标：Donald Trump 立场极性：Favor 样例句：one of the key promblems today is that policis is such a disagrace, good people don't go to government 隐藏目标和主题词后的句子：one of the key promblems today is that policis is such a $[MASK]$,$[MASK]$ people don't go to government 正确的可迁移与否标签：目标无关的 模型预测的可迁移与否标签：目标无关的
样例2
目标：Donald Trump 立场极性：Against 样例句：donald trump did not apply to immigrants one of the trade basis , win to win . ignorance can not be excuse 隐藏目标和主题词后的句子：$[MASK]$ did not apply to immigrants one of the trade basis,win to win . $[MASK]$ can not be excuse 正确的可迁移与否标签：目标无关的 模型预测的可迁移与否标签：目标依赖的

Table 7: 两个典型样例

5 结论

本文提出了一种基于主题提示学习的零样本立场检测方法(ST-PL)，该方法通过结合提示学习和代理任务生成增强数据，能有效帮助主任务模型完成零样本立场检测。具体地，ST-PL方法利用了面向目标不变和依赖于目标的立场特征之间的差异来学习可迁移的立场特征，从而能显式地将可迁移的立场特征用于零样本立场检测中，并有效提升零样本立场检测的性能。最终在两个零样本立场检测基准数据集上进行的实验结果表明，本文提出的ST-PL方法性能表现全面优于基线模型。具体的案例分析表明，该方法在语法结构相对简单、立场词态度表达一

第二十一届中国计算语言学大会论文集，第535页-第544页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

致的数据上表现优异，对于复杂语法、相异立场态度词表达与复杂语言现象句的可迁移预测受限于现有模型的预测瓶颈表现一般，有待未来改进。

参考文献

Allaway E, Mckeown K. 2020. *Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*:8913–8931.

Allaway E, Srikanth M, Mckeown K. 2021. *Adversarial Learning for Zero-Shot Stance Detection on Social Media. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*:4756–4767.

Augenstein I, Rocktäschel T, Vlachos A, et al. 2016. *Stance Detection with Bidirectional Conditional Encoding. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*:876–885.

Chen T, Kornblith S, Norouzi M, et al. 2020. *A simple framework for contrastive learning of visual representations. International conference on machine learning. PMLR, 2020*:1597–1607.

Conforti C, Berndt J, Pilehvar M T, et al. 2020. *Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*:1715–1724.

Devlin J, Chang M W, Lee K, et al. 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*:4171–4186.

Du J, Xu R, He Y, et al. 2017. *Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence*:3988–3994.

Gidaris S, Singh P, Komodakis N. 2018. *Unsupervised Representation Learning by Predicting Image Rotations. International Conference on Learning Representations,2018.*

Gunel B, Du J, Conneau A, et al. 2021. *Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. International Conference on Learning Representations,2021.*

Kachuee M, Yuan H, Kim Y B, et al. 2021. *Self-Supervised Contrastive Learning for Efficient User Satisfaction Prediction in Conversational Agents. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*:4053–4064.

Kawintiranon K, Singh L. 2021. *Knowledge enhanced masked language model for stance detection. roceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*:4725–4735.

Larsson G, Maire M, Shakhnarovich G. 2016. *Learning representations for automatic colorization. European conference on computer vision. Springer, Cham, 2016*:577–593.

Liang B, Fu Y, Gui L, et al. 2021. *Target-adaptive graph for cross-target stance detection. Proceedings of the Web Conference 2021*:3453–3464.

Liu R, Lin Z, Tan Y, et al. 2021. *Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph.. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*:3152–3157.

Li Y, Caragea C. 2019. *Multi-task stance detection with sentiment and stance lexicons. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*:6299–6305.

Mohammad S, Kiritchenko S, Sobhani P, et al. 2016. *Semeval-2016 task 6: Detecting stance in tweets. Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*:31–41.

Noroozi M, Favaro P. 2016. *Unsupervised learning of visual representations by solving jigsaw puzzles. European conference on computer vision. Springer, Cham, 2016*:69–84.

(c) 2022 中国中文信息学会计算语言学专业委员会

Petroni F, Rocktäschel T, Riedel S, et al. 2019. *Language Models as Knowledge Bases?. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*:2463–2473.

Schick T, Schütze H. 2021. *Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*,Main Volume 2021:255–269.

Shin T, Razeghi Y, Logan IV R L, et al. 2020. *AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*:4222–4235.

Siddiqua U A, Chy A N, Aono M. 2019. *Tweet stance detection using an attention based neural ensemble model. 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):1868–1873.

Simard N, Lagrange G. 2021. *Improving Few-Shot Learning with Auxiliary Self-Supervised Pretext Tasks. arXiv preprint arXiv:2101.09825, 2021.*

Somasundaran S, Wiebe J. 2010. *Recognizing stances in ideological on-line debates. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*:116–124.

Sun Q, Wang Z, Zhu Q, et al. 2018. *Stance detection with hierarchical attention network. Proceedings of the 27th international conference on computational linguistics*:2399–2409.

Wei P, Mao W. 2019. *Modeling transferable topics for cross-target stance detection. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*:1173–1176.

Xu C, Paris C, Nepal S, et al. 2018. *Cross-Target Stance Classification with Self-Attention Networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*:778–783.

Zhang B, Yang M, Li X, et al. 2020. *Enhancing cross-target stance detection with transferable semantic-emotion knowledge. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*:3188–3197.

Zhang R, Isola P, Efros A A. 2016. *Colorful image colorization. European conference on computer vision.Springer, Cham, 2016*:649–666.

CCL 2022

标签先验知识增强的方面类别情感分析方法研究

吴任伟[1], 李琳[1], 何铮[2], 袁景凌[1]
[1]武汉理工大学 计算机与人工智能学院,武汉 430070
[2]德勤有限公司,上海 510623
{wrw, cathylilin}@whut.edu.cn, zhhe@deloitte.com.cn, yuanjingling@126.com

摘要

当前，基于方面类别的情感分析研究旨在将方面类别检测和面向类别的情感分类两个任务协同进行。然而，现有研究未能有效关注情感数据集中存在的噪声标签，影响了情感分析的质量。基于此，本文提出一种标签先验知识增强的方面类别情感分析方法（AP-LPK）。首先本文为面向类别的情感分类构建了自回归提示训练方式，可以激发预训练语言模型的潜力。同时该方式通过自回归生成标签词，以期获得比非自回归更好的语义一致性。其次，每个类别的标签分布作为标签先验知识引入，并通过伯努利分布对其进行进一步精炼，以用于减轻噪声标签的干扰。然后，AP-LPK将上述两个步骤分别得到的情感类别分布进行融合，以获得最终的情感类别预测概率。最后，本文提出的AP-LPK方法在五个数据集上进行评估，包括SemEval 2015和2016的四个基准数据集和AI Challenger 2018的餐厅领域大规模数据集。实验结果表明，本文提出的方法在F1指标上优于现有方法。

关键词： 基于方面类别的情感分析 ； 提示学习 ； 标签先验知识

Aspect-Category based Sentiment Analysis Enhanced by Label Prior Knowledge

Renwei Wu[1], Lin Li[1], Zheng He[2], Jingling Yuan[1]
[1]School of Wuhan University of Technology, Wuhan 430070, China
[2]Department of Deloitte Limited, Shanghai 510623, China
{wrw, cathylilin}@whut.edu.cn, zhhe@deloitte.com.cn, yjl@whut.edu.cn

Abstract

Current aspect-category based sentiment analysis researches aim at performing joint aspect category detection and category-oriented sentiment classification. However, most of existing studies have not paid much attention on the noisy labels that often occur in sentiment datasets. To cope with this problem, we propose an aspect-category based sentiment analysis approach with Label Prior Knowledge(AP-LPK). Specifically, we firstly construct an Autoregressive Prompting training that can stimulate the potential of pre-trained language models. And then label words are generated through autoregression for better semantic consistency than non-autoregression. Secondly, we introduce the label distribution of each category as label prior knowledge that is refined through Bernoulli distribution to mitigate the interference of noisy labels. And then, the outputted labels from the autoregressive prompting and label prior knowledge refined work together based on the distributions of sentiment polarities to obtain final predictions. Finally, our AP-LPK approach is evaluated on five datasets that include the four benchmark datasets from SemEval 2015 and 2016 and the Restaurant-domain dataset from AI Challenger 2018. Experimental results demonstrate that our approach outperforms existing ones in terms of F1.

Keywords: Aspect-Category based Sentiment Analysis , Prompt learning , Label prior knowledge

第二十一届中国计算语言学大会论文集，第545页-第556页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

随着互联网的发展，网络生活中衣食住行等服务或产品已经逐渐融入人们的日常生活，人们可以在网络生活中浏览、购买和使用这些产品，并分享自己对产品的看法或评论。以文本为主的评论涉及产品的不同方面，用户从产品的不同角度进行描述，包含有非常丰富的和有价值的多方面信息。因此文本评论具有广泛的研究场景，比如基于方面类别的消费评论（外卖、电商等）的情感分析，而对评论细粒度的情感分析有助于相关用户从中高效快捷地获取各方面的信息，为后续决策提供支持。

基于方面类别的情感分析(Aspect-Category based Sentiment Analysis(ACSA))在实际应用中逐渐变得越来越流行 (Schmitt et al., 2018; Dai et al., 2019; Guo et al., 2020; Cai et al., 2020; Fu et al., 2021; Hu et al., 2019; Liang et al., 2021)。ACSA旨在识别评论句子中的多个方面类别，并共同预测每个已识别类别的情感极性。

已有的研究重点在于如何建立类别-情感联合的神经网络模型，通过在情感标签空间中添加一个维度来指示每个类别的出现，例如Schmitt等人 (2018)，Dai等人 (2019)和Guo等人 (2020)。近年来，ACSA 转向预训练和微调范式。Cai等人 (2020)介绍了几种基于微调的方法，即Cartesian-BERT、Pipeline-BERT和AddOneDim-BERT。此外，他们将ACSA重新形式化为类别-情感层次预测问题，他们的方法可以对多个类别之间的内在关系以及类别与情感标签之间的相互关系进行建模，即Hier-BERT、Hier-Transformer-BERT和Hier-GCN-BERT。不过，尽管这些基于微调范式的模型稳定地优于基于传统神经网络的工作，但微调范式仍然存在一个主要障碍，即无法充分激发预训练语言模型的潜力，甚至会导致灾难性的遗忘问题 (Liu et al., 2021)。

在情感数据集中不可避免地会出现噪声标签，这是由于数据标注者的能力以及他们对标注标准的不同理解 (Zhou, 2018; Li et al., 2021)。以图 1中的两条评论为例。 虽然两条评论关于类

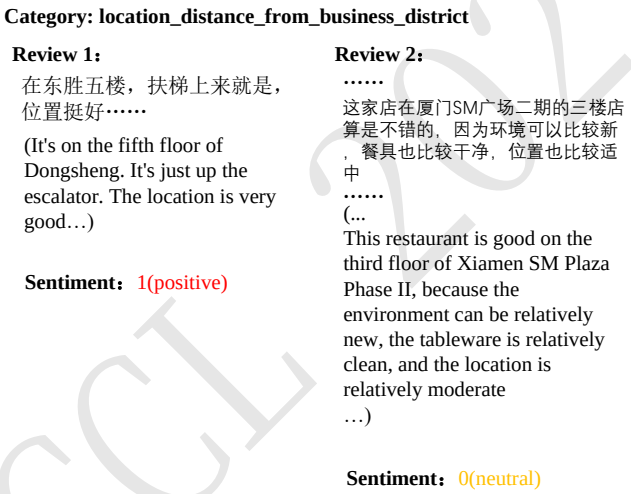


Figure 1: 两条评论样例

别*location_distance_from_business_district*的描述基本相同，但标注的情感极性不同。事实上，两者的情感极性都应该是积极的。更重要的是，这种噪声标签的情况在每个类别中并不少见。由于这些噪声标签会干扰模型性能 (Pan et al., 2022)，并且提示训练在各种NLP任务中表现出有效的性能，因此在ACSA任务中同时考虑它们是一个挑战。

针对该问题，本文提出了一种基于方面类别的带有标签先验知识的情感分析方法，即带有标签先验知识的自回归提示(Autoregressive Prompting with Label Prior Knowledge(AP-LPK))。在本文的自回归提示中，完形填空式的提示模板是一个包含类别的句子，含有多个连续的汉字掩码。由此，通过自回归语言模型 (Yang et al., 2019)，答案生成是自回归的。为了解决噪声标签的问题，本文通过结合伯努利分布和标签先验知识来校准自回归提示训练的输出标签，以生成最终预测。

本文的实验是在五个公开的数据集上进行，其中包括四个基准数据集和一个具有超过100k条带标签样本的中文大规模数据集。实验结果验证了提示训练在ACSA任务中的有效性。在此基础上，本文提出的AP-LPK方法可以通过处理标签噪声始终能促进预测质量的提升。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

2 相关工作

近年来，基于方面类别的情感分析已经取得了一定的进展。现有研究按近代自然语言技术发展的范式 (Liu et al., 2021)大致分为两类，传统的神经网络模型和基于预训练和微调范式的模型

2.1 基于传统神经网络的模型

基于传统神经网络的工作采用的是基于神经网络的完全监督学习范式 (Liu et al., 2021)。在该范式下，基于方面类别的情感分析研究经历了两个阶段。

研究初期，学者们对于ACSA任务多采用Pipeline方法/多任务方法，即将ACSA任务分为方面类别提取(aspect category detection(ACD))任务和方面级情感分类(aspect level sentiment classification(ALSC))任务。Ruder等人 (2016)提出了分层神经网络模型来进行方面级情感分类，其中ACD假定基于Pipeline框架中的一些其他系统（如，SVM），而ALSC是模型的主要任务。

类似的，Hu等人 (2019)使用多任务的方式研究。在辅助任务ACD和主任务ALSC的设定下，他们提出了一种注意力网络CAN，以约束注意力权重分配，帮助学习更好的具体方面的句子表示。

上述的两项工作在公开数据集上都具有不错的表现。但是，Pipeline方法/多任务方法的任务分离会带来误差累积/传播问题 (Guo et al., 2020)，在一定程度上影响着模型效果。

因此，为了解决这个问题，近年来也有一些其他研究在探索联合学习ACD任务和ALSC任务。Schmitt等人 (2018)提出了一种联合模型，被Cai等人 (2020)称为AddOneDim-LSTM。该模型通过标签扩维，即在情感标签空间中添加一个维度来指示每个类别的出现，从而将ACD任务融入ALSC任务中，达到联合建模的目的。这代表了当时最先进的工作。

之后，一些其他的学者也以标签扩维为基础对ACSA进行研究。

Dai等人 (2019)为了捕获文本中的多共享特征以及特定类别的特征，提出了模型MMAM。该模型采用一个多头文档注意力机制作为记忆单元以编码共享的文档特征，并且采用一个多任务注意力机制来提取特定类别的特征。在两个真实数据集上的实验结果表明，MMAM具有良好的预测效果。

Guo等人 (2020)提出了一个改良的多路匹配深度神经网络模型以进行细粒度的情感分析，模型通过直接在多轮校准结构中捕获过去的注意力来改善现有的注意力，以预防误差传播和注意力缺失。

另外，Fu等人 (2021)基于双BiLSTM的多角度注意力，并在模型中引入特定方面类别的信息，提出了MPADB和MPADB_joint结构。MPADB通过丰富的上下文表示和多角度注意力机制避免方面和相应情感极性的错误匹配，而MPADB_joint是为了解决传统Pipeline方法的误差累积问题以及联合模型的注意力权重重叠问题，并提升模型的可解释性。在两个真实数据集上的实验结果表明了这项工作在准确性和可解释性方面的有效性。

然而，从2017年到2019年，自然语言处理模型的学习发生了巨大的变化，这种完全监督的范式现在发挥的作用越来越小 (Liu et al., 2021)。

2.2 基于预训练和微调范式的模型

近年来，随着BERT等一系列大规模预训练语言模型在NLP领域大放异彩，预训练和微调范式成为学术界和工业界广泛关注的重点。因此有学者开始应用该范式来研究基于方面类别的情感分析。Cai等人 (2020)基于预训练BERT和微调，提出了一系列方法，同时结合图卷积网络（GCN）来进行ACSA的研究。

- Cartesian-BERT：以BERT为句子编码器的笛卡尔法。
- Pipeline-BERT：ACD和ALSC都以BERT为编码器进行建模。
- AddOneDim-BERT：将AddOneDim-LSTM (Schmitt et al., 2018)中的LSTM替换为BERT。
- Hier-BERT：以BERT为句子编码器的类别-情感层次预测方法。
- Hier-Transformer-BERT：在Hier-BERT的基础上，Transformer被用来建模类别间的内在关系和类别与情感之间的相互关系。
- Hier-GCN-BERT：基于Hier-BERT，以层次图卷积网络（Hier-GCN）进行关系学习。

在四个公开数据集上的大量实验表明了Cai等人 (2020)工作的有效性，并且上述基于预训练BERT和微调的方法均在F1值方面优于Schmitt等人 (2018)的工作。

(c) 2022 中国中文信息学会计算语言学专业委员会

此外，Liang等人 (2021)通过探索基于外部知识的Beta分布引导的方面感知图构建，从一个新颖的角度研究ACSA任务。他们提出了AAGCN-BERT、AAGCN-BERT-c等模型，并且在六个基准数据集上的实验表明，他们的方法显著优于最先进的基线方法。

综上所述，基于方面类别的情感分析领域上的研究有一定的进展，具有非常好的理论指导作用。但几乎所有的ACSA工作都对数据集中经常出现的噪声标签缺乏关注，同时针对现有工作对提示学习的研究还不够丰富，因此我们的研究主要基于自回归提示学习，采用标签先验知识修正噪声标签问题，提出了一种方面类别情感分析方法AP-LPK，对用户评论文本进行预测。

3 任务定义

本文的任务是对于给定的待测评论文本和m个预定义的方面类别，能够提取所有提及的方面类别，并且识别每个被检测到的类别的情感。

给定一段具有n个字符的评论文本$\boldsymbol{r} = [w_1, ..., w_n]$，令$\boldsymbol{C} = \{c_1, ..., c_m\}$为$m$个预定义方面类别的集合，并且$s = \{negative, neutral, positive, not\ mentioned\}$为情感极性的标签集。因此，对于每一个输入$\boldsymbol{r}$，本项工作中ACSA的目标是生成情感极性集$y = \{..., \hat{y}_i, ..., \hat{y}_m\}$，其中$\hat{y}_i$表示评论文本$\boldsymbol{r}$中第i个方面类别对应的情感极性。

根据上述的定义，我们所提出的AP-LPK方法，其映射关系表示如下：

$$[\boldsymbol{r}^1, ..., \boldsymbol{r}^I, ..., \boldsymbol{r}^N] \xrightarrow{f_i(\cdot)} [\hat{y}_i^1, ..., \hat{y}_i^I, ..., \hat{y}_i^N] \tag{1}$$

其中，N表示数据集的大小，则$\boldsymbol{r}^N$表示第N条评论文本；$f_i(\cdot)$表示第i个类别对应的模型；$\hat{y}_i^I$表示第I条评论文本中第i个方面类别的情感极性。

4 带有标签先验知识的自回归提示

针对上述研究现状和任务定义，本文提出了AP-LPK方法，该方法结合提示学习、自回归模型与标签先验知识对输入的用户评论文本进行模型学习。AP-LPK能够构建ACSA任务的提示学习训练，得到面向特定方面类别的自回归语言模型，并且将提示学习的答案工程与标签先验知识相结合来实现噪声标签干扰的减轻。

4.1 AP-LPK框架

本文们提出方法的整体框架如图 2所示，该方法由两部分组成，自回归提示学习和标签先验知识的引入。 在我们的提示训练中，提示模板工程和答案工程是手动设计的。自回归语言模型（例

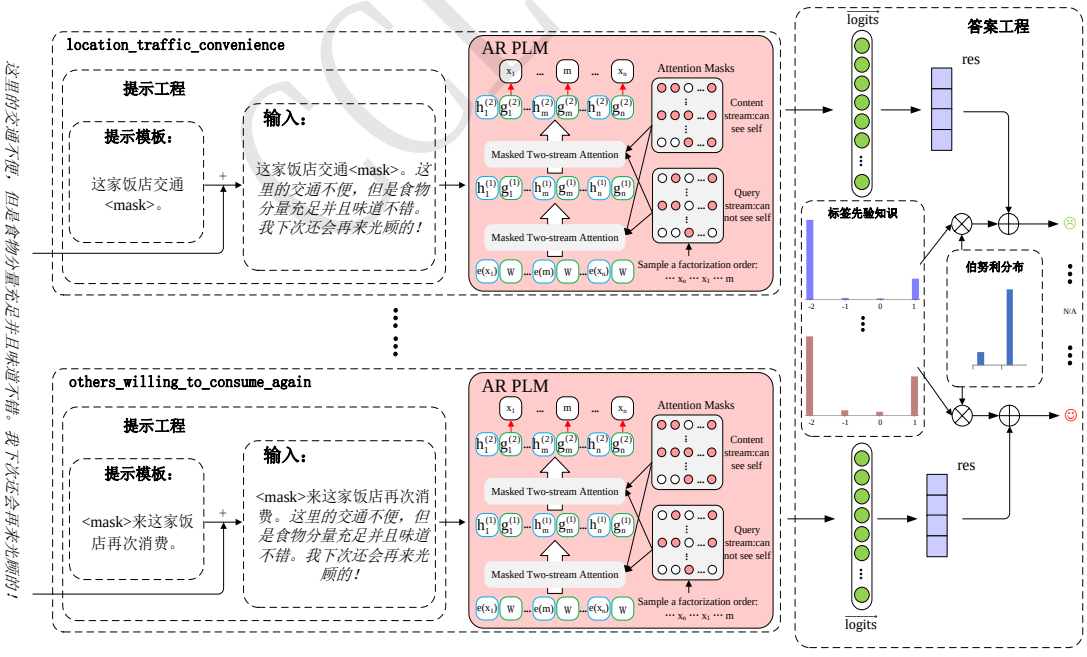


Figure 2: AP-LPK框架图

如，XLNet (Yang et al., 2019)）被引入来训练每个类别，以构建多类别提示训练（第 4.2节）。

(c) 2022 中国中文信息学会计算语言学专业委员会

此外，标签先验知识和伯努利分布的结合协助精炼提示训练的输出，以进一步获得最终预测（第 4.3节）。

4.2 自回归提示

自回归提示学习部分主要如图 2左侧所示，主要分为：提示模板工程与答案工程的设计和自回归预训练语言模型的引入。

4.2.1 提示模板工程与答案工程的设计

现今已有一些关于设计提示模板工程与答案工程的研究，其中设计方式主要包含手动设计和自动化设计 (Liu et al., 2021; Chen et al., 2022; Schick et al., 2020)。我们采用手动方式来设计提示学习中的提示模板工程和答案工程，这种方式策略直观而且已经被证明在提示任务上可以实现稳定的性能 (Schick and Schütze, 2021; Hu et al., 2021; Gao et al., 2021)。

对于每个方面类别，基于完形填空式的提示学习，手动设计的提示模板需要贴合评论文本的上下文表达，以进一步保证模型输入的语义质量。另外，提示模板还必须对应待测的特定方面类别信息，如图 2中所示的**提示模板**。同时，提示模板应尽量简洁，避免引入其他干扰信息。

由此，答案空间以及到输出空间的映射可根据提示模板和评论文本的上下文语境手动设计。在每个方面类别下，答案空间中的标签词是与情感极性对应的具有正确语义的中文字词，从而也构建起从答案空间到输出空间的映射。当这些字词应用到模板中的掩码位置（即，标记<**mask**>）时，它们可以和对应的提示模板构成具有合理语义的评论文本。

提示模板拼接在每个类别的每段评论文本的开头，以生成模型输入，如图 2中所示的**输入**，其中标记<**mask**>是模型的预测对象。

4.2.2 自回归预训练语言模型的引入

基于完形填空式的提示学习一般采用Masked LM作为预训练模型，比如BERT、RoBERTa (Schick and Schütze, 2021; Hu et al., 2021)。BERT、RoBERTa作为自编码语言模型，对于上述**输入**，在训练过程中可以根据上下文，预测标记<**mask**>处被掩盖的字词。BERT等自编码语言模型虽然性能不错，但是对于被掩盖的字符的数目多于1且连续时，它的预测效果便无法保证。特别的，对于我们实验中基于中文字词的待测标签词（一般是由多个连续的中文字符组成，比如中性情感词“一般”等），这些模型还可能生成不合语义的答案词（比如“不般”等），从而干扰后续的标签映射。这是因为BERT等自编码语言模型对于被掩盖的多个字符是相互独立预测的，即非自回归生成。

因此，预训练中文XLNet (Cui et al., 2020)被选择作为我们提示学习的预训练语言模型，它将自回归语言模型和自编码语言模型的优点进行了巧妙的结合。具体来说，对于我们的提示训练，每个方面类别将上述的模型**输入**馈送给XLNetLMHeadModel[0]，并使用答案空间中对应的标签词作为训练标签。该模型在训练过程中使用的损失函数是交叉熵。

如前所述，自回归提示的提出是为了激发预训练语言模型的潜力，并生成具有更好语义一致性的标签词。例如，在图 1中，提示学习会使得下游任务去适应预训练语言模型。在这种情况下，预训练语言模型可以充分利用预训练时的数据信息，其中可能包含类似于图 1的描述；而微调需要预训练语言模型来适应下游任务，可能会导致这些信息的遗忘。此外，生成的标签词也为第4.3节奠定基础。

4.3 标签先验知识

动机：如前所述，本文提出使用每个类别的标签分布作为标签先验知识，来帮助减轻噪声标签的干扰。对于数据标注来说，尽管由于标注不当而产生了少量噪声标签，但数据集的整体标注准确率和习惯是趋于稳定。同样以图 1为例。数据集中具有相似描述的评论基本上被标记为类别*location_distance_from_business_district*上的积极情感极性，并且该知识可以反映在标签分布中，用以校准噪声标签的问题。

具体来说，对于每个类别，本文将训练集中四种情感极性（即消极、中性、积极、未提及）的频率统计为标签先验知识$\boldsymbol{K} \in R^4$。然而，该先验知识$\boldsymbol{K}$不一定是完全有益的。受BERT的启发，BERT在进行基于掩盖的语言模型任务时，使用伯努利分布来决定随机对输入序列中的某些位置进行遮罩 (Devlin et al., 2019)，因此这里伯努利分布被采用来决定$\boldsymbol{K}$可用的概率，该概率在BERT中为0.15，而在这里是可学习的。

[0]https://huggingface.co/docs/transformers/v4.17.0/en/model_doc/xlnet

(c) 2022 中国中文信息学会计算语言学专业委员会

基于自回归模型对标记<**mask**>的预测$\overrightarrow{logits}$，我们提取对应于四个情感极性的标签词的logits，以生成原始输出$\boldsymbol{res}$。因此，进行噪声标签干扰的缓解如下：

$$\boldsymbol{F} = bernoulli(\boldsymbol{K}) \bigoplus \boldsymbol{res} \tag{2}$$

其中，$\bigoplus$表示基于四种情感极性的对应相加。

最后，可以得到最终输出如下：

$$\hat{y}_i^I = argmax(softmax(\boldsymbol{F})) \tag{3}$$

4.4 算法描述

综上所述，AP-LPK的主要算法描述如算法 1所示。

在算法中，每个方面类别下的模型是相互独立的，因此本文对所有预定义方面类别进行循环（第2行）。在每个轮次的循环（第5行）中，算法首先通过训练数据集进行XLNet训练（第6行），并将本轮次训练所得的XLNet用于验证集并得到验证结果（第7行）。因为实验主要以Macro-F1值进行评估，因此判断验证结果中的F1值是否在本轮次中有所提升（第8，13行）。若验证结果中的F1值相比上一轮次有提升，则保存本轮次训练所得XLNet并利用该模型进行测试集和标签先验知识上的评估（第11，12行）；否则，若验证集上的F1值在3个epoch内没有进一步提升，当前方面类别的训练过程才结束（第14，15，16行）。

算法 1 AP-LPK算法

输入： 带有提示模板的训练集评论文本$train_input$，对应$train_input$的标签词$train_label_word$;
带有提示模板的验证集评论文本$valid_input$，对应$valid_input$的标签词$valid_label_word$;
带有提示模板的测试集评论文本$test_input$，对应$test_input$的标签词$test_label_word$;
标签先验知识lpk

输出： 评价指标($test\_result['macro\_f1']$)，每个方面类别的最终预测模型model

```
1: function AP-LPK
2:     for each category do
3:         max_macro_f1 ← −1.0
4:         early_stop ← 0
5:         for each epoch do
6:             training: XLNet(train_input, train_label_word)
7:             validating: valid_result['macro_f1'] ← XLNet(valid_input, valid_label_word)
8:             if valid_result['macro_f1'] > max_macro_f1 then
9:                 early_stop ← 0
10:                max_macro_f1 ← valid_result['macro_f1']
11:                saving model
12:                testing: test_result['macro_f1'] ← XLNet(test_input, test_label_word, lpk)
13:            else
14:                early_stop ← early_stop + 1
15:                if early_stop >= 3 then
16:                    break
17:                end if
18:            end if
19:        end for
20:    end for
21: end function
```

5 实验

5.1 数据集

本文的方法在5个数据集上进行评估，数据集相关的统计信息如表 1所示。 其中，4个基准数据集来自SemEval 2015和2016 (Pontiki et al., 2015; Pontiki et al., 2016)。REST-AI Challenger 2018为

	Restaurant-15	Laptop-15	Restaurant-16	Laptop-16	REST-AI Challenger 2018
训练集样本数	1102	1397	1680	2037	105k
测试集样本数	572	644	580	572	15k
预定义类别数	30	198	30	198	20

Table 1: 数据统计

中文数据集，即在线用户评论数据集的细粒度情感分析2018 (AI Challenger 2018)[1]。原REST-AI Challenger 2018虽然包含测试集A和测试集B，但由于原测试集无法下载且未被标注，因此本文将AI Challenger竞赛中的验证集作为本次实验的测试集。

REST-AI Challenger 2018在数据量上远大于4个基准数据集，因此本文选择它作为实验的主要数据集。REST-AI Challenger 2018中的评价对象按照粒度不同划分为两个层次，第一层为粗粒度的评价对象，例如评论文本中涉及的环境、价格等要素；第二层为细粒度的情感对象，例如“环境”属性中的“装修情况”、“嘈杂情况”等要素。具体情况如表 2所示。

粗粒度层面	细粒度层面
位置(location)	交通是否便利(traffic convenience)
	距离商圈远近(distance from business district)
	是否容易寻找(easy to find)
服务(service)	排队等候时间(wait time)
	服务人员态度(waiter's attitude)
	是否容易停车(parking convenience)
	点菜/上菜速度(serving speed)
价格(price)	价格水平(price level)
	性价比(cost-effective)
	折扣力度(discount)
环境(environment)	装修情况(decoration)
	嘈杂情况(noise)
	就餐空间(space)
	卫生情况(cleaness)
菜品(dish)	分量(portion)
	口感(taste)
	外观(look)
	推荐程度(recommendation)
其他(others)	本次消费感受(overall experience)
	再次消费的意愿(willing to consume again)

Table 2: 评价对象的具体划分

每个细粒度要素有4种情感倾向，如表 3所示。

情感标签(labels)	-1	0	1	-2
情感倾向	消极情感(Negative)	中性情感(Neutral)	积极情感(Positive)	情感倾向未提及(Not mentioned)

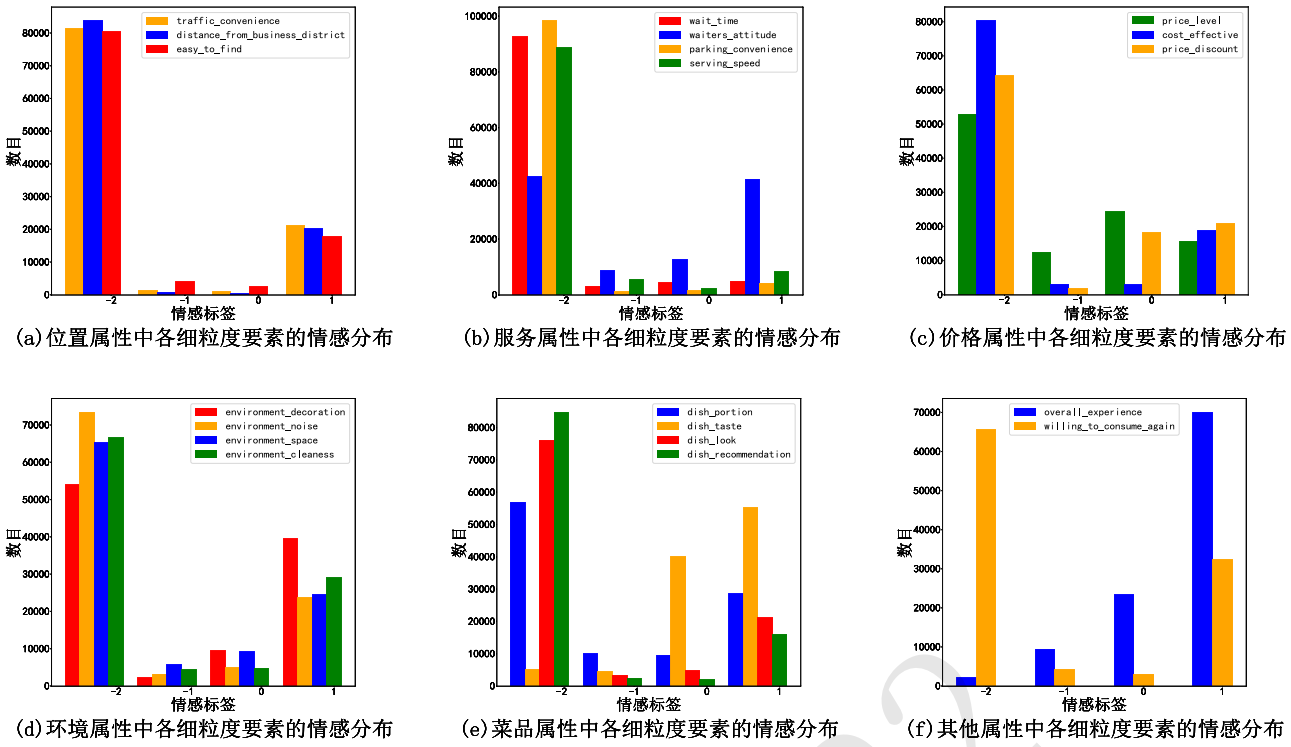
Table 3: 情感倾向

为了直观了解数据情况、探究数据的内在特征，实验中统计了REST-AI Challenger 2018中六种粗粒度方面下每个细粒度要素在四种情感倾向上的评论文本数目分布，数据统计情况如图 3所示。

从统计结果中可以观察到REST-AI Challenger 2018的分布方式极不平衡，总体上最多的情感倾向为Not mentioned，并且消极和中性的情感倾向普遍相对偏少。这种数据失衡给模型的训练带来了一定挑战，但也启发了后续的标签先验知识的引入。

[1]https://challenger.ai/dataset/fsaouord2018.

(c) 2022 中国中文信息学会计算语言学专业委员会



(a)位置属性中各细粒度要素的情感分布 (b)服务属性中各细粒度要素的情感分布 (c)价格属性中各细粒度要素的情感分布

(d)环境属性中各细粒度要素的情感分布 (e)菜品属性中各细粒度要素的情感分布 (f)其他属性中各细粒度要素的情感分布

Figure 3: 六种方面的数据统计情况

5.2 实验设置

对于REST-AI Challenger 2018：原训练集随机以9:1的比例分成训练集、验证集，并且构建的提示模板、答案空间及到输出空间的映射、标签先验知识均以字典形式引入。XLNetLMHeadModel加载了Chinese-XLNet-base的预训练权重，具体设置请参考相应网址[2]。

实验采用AdamW优化器，其学习率和批量大小（batch size）分别设置为1e-5和6。最大文本长度设置为512，通过统计分析训练集中的大部分文本长度主要集中在500附近及其之前的区间；另外，长度在250附近区间的文本最多，并随着文本长度的增加，分布逐渐减小，到文本长度超过1000时，如此长度的文本在数据集中就更稀少了。最后，直到模型在验证集上的性能在3个epoch内没有变化，训练过程才结束。

准确率（Precision）、召回率（Recall）和Macro-F1值是评估指标，且Macro-F1值是实验的主要比较指标。在实验中，我们分别计算每个类别下的三种评估指标，并取所有类别的三种指标均值作为最终的评估结果。

对于四个基准数据集：XLNet使用xlnet-base-cased的预训练权重，具体设置请参考相应网址[3]。其他实验设置参照Cai等人 (2020)的工作。

5.3 基线模型

AddOneDim-LSTM (Schmitt et al., 2018)、Hier-GCN-BERT (Cai et al., 2020)是最近关于ACSA任务的两项工作，并被选为我们的主要基线。

AddOneDim-LSTM: 在标签扩维的基础上，词嵌入模型FastText被用来进行评论文本的词嵌入工作，然后通过双向LSTM进行编码，最后编码结果被用于多个分类器，其中分类器的个数取决于给定方面类别的数目，且分类器之间相互独立。为了得到较好的结果，模型参数也进行了以下调整：词嵌入维度设置为100，双向LSTM隐层的维度设置为300，dropout和学习率设为0.5和0.001。另外，训练优化器采用Adam算法，损失函数为交叉熵。

Hier-GCN-BERT: 近年来最先进的工作之一。模型利用BERT进行特征提取，捕获全局情感，并通过多头自注意力进行方面类别的表示；而后，基于类别之间的关联性和类别与情感之间的关联性，图卷积网络被用来建模其中的关系；最后，在类别-情感层次预测结构的基础上，方面类别提取和情

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

(c) 2022 中国中文信息学会计算语言学专业委员会

感分类分别进行，并且方面类别的高优先级被用来进行类似剪枝的操作，以改善模型性能。

全局上，所有基线模型都与AP-LPK的算法流程保持基本一致，最大文本长度也都设置为512。其他未提及的参数，均与模型原始论文中的参数保持基本一致。因此，实验通过复现以上两项工作，得到在REST-AI Challenger 2018数据集上的实验结果，与我们提出的AP-LPK方法进行对比。

5.4 实验结果

本部分将从总体分析、消融分析以及案例分析对实验结果进行讨论。实验结果见表 4、表 5。

Method	P	R	F1
AddOneDim-LSTM	67.73	65.28	65.88
Hier-GCN-BERT	70.83	68.04	69.04
AP-LPK(**Ours**)	72.74	70.15	**71.00**
w/o Label Prior	72.58	70.24	70.77
w/o Label Prior & AutoRegression	74.14	68.38	69.48

Table 4: 在REST-AI Challenger 2018上的实验结果(%)

Method	Restaurant-15			Laptop-15			Restaurant-16			Laptop-16		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AddOneDim-LSTM	54.33	28.44	37.32	-	-	-	61.56	42.82	50.50	-	-	-
Hier-GCN-BERT	71.93	58.03	64.23	71.90	54.73	62.13	76.37	72.83	74.55	61.43	48.42	54.15
AAGCN-BERT	-	-	71.75	-	-	72.39	-	-	80.77	-	-	69.68
AP-LPK(**Ours**)	77.70	77.28	**77.33**	83.03	82.76	**82.72**	77.26	77.60	77.32	80.33	80.22	**80.10**

Table 5: 在四个基准数据集上的主要结果(%)

总体分析：从表 4中可知，在REST-AI Challenger 2018数据集上，AddOneDim-LSTM的Macro-F1值为0.6588。基于类别-情感层次预测结构的Hier-GCN-BERT的Macro-F1值可以达到0.6904。这也论证了Cai等人 (2020)的工作。而相比于通过微调训练的方法，我们提出的带有自回归提示训练和标签先验知识的方法可以达到0.7100的Macro-F1值。这表明我们的提示建模在这个任务中比微调有更好的性能改进，以及标签先验知识的有效性。

在表 5中，AddOneDim-LSTM、Hier-GCN-BERT的结果取自Cai等人 (2020)，AAGCN-BERT的结果取自Liang等人 (2021)。我们的方法AP-LPK在四个基准数据集上的F1值分别可以达到0.7733、0.8272、0.7732、0.8010。其中，因为Restaurant-16预定义的类别较为宽泛，比如类别“FOOD#QUALITY”会涉及更细粒度的方面类别（例如，口感、新鲜度、质地、温度等），提示模板的设计很难精确，由此影响了提示学习的效果，因而我们提出的AP-LPK在F1值方面低于AAGCN-BERT。针对这样的问题，如何进行更合适的提示模板工程，也是我们后续研究、改进的重要内容。

消融分析：在表 4中，我们进一步展示了AP-LPK消融研究的结果。在无标签先验知识的情况下，F1值降低到0.7077。对于无标签先验和自回归的情况，我们将预训练语言模型XLNet替换为BERT (Devlin et al., 2019)，以完成提示训练。它的设置与我们的自回归提示中的设置保持一致。它的F1值从0.71降低到0.6948。从中可见自回归模型在生成高质量标签词方面的能力更强。

此外，在无标签先验和自回归的情况中，基于非自回归的BERT会出现未知、错误标签词的生成。为了解决这个问题，我们使用这些标签词的最后一个字符作为映射参考，以在预定义的字典中寻找最可能的一个答案。由于字典由类别标签组成，因此大量的占比较大的情感极性的映射可以被正确获得，例如情感极性“未提及”。因此，这种情况下平均精度更高，但召回率要低得多。同时，即使使用我们手动添加的字典，在非自回归条件下的Macro-F1值仍然是最低的。

案例分析：如表 6所示， Review 1和Review 2关于装修风格的描述是相似的，但相应的情感标签不同。通过阅读它们，我们可以很容易地知道Review 2在类别*environment_decoration*上的情感标签是训练集中的噪声。而我们的方法AP-LPK通过利用标签先验知识，可以正确预测Review 1关于该方面类别的情感极性为积极。

同时，在图 1中提到的Review 3为训练集带来了噪声标签，其在类别*location_distance_from_business_district*上的真实情感标签应该是积极的。当使用Review 3 作

(c) 2022 中国中文信息学会计算语言学专业委员会

Review text	Category	Label	w/o Label Prior	AP-LPK
No.1 金殿水库边，靠近云南飞虎队博物馆，位置算很好找的，在一个院子里，整个装修风格很民族风，算是有特点的店.....(**from testing set**)	*environment_decoration*	positive	neutral	positive
No.2 早就听说了这家店，今天在凯德广场转，于是就来尝尝。店铺的装修风格很工厂感，黑色的铁丝网和灰色的墙壁，感觉特别.....(**from training set**)	*environment_decoration*	neutral	-	-
No.3 .....这家店在厦门SM广场二期的三楼店算是不错的，因为环境可以比较新，餐具也比较干净，位置也比较适中.....(**from training set**)	*location_distance_from_business_district*	neutral	-	positive

Table 6: 关于AP-LPK的案例分析

为测试样本时，我们的方法仍然可以正确识别它，这意味着我们的方法可以用来修正训练数据集中的那些噪声标签。这将是未来一项有趣的工作。

另外，从表 7中， 我们也可以直观验证实验中自回归生成与非自回归生成的研究。Review 4在测

Review text	Category	Label	w/o Label Prior & AutoRegression	w/o Label Prior
No.4感谢大众点评，感谢又让我中试吃.....环境，简单干净正经，最值得一题的就是无论你坐在哪里都不会离菜品特别远，不像凯德，每次拿东西都要走十万八千里！！值得表扬(〉ˆωˆ〈).....(**from testing set**)	*environment_decoration*	positive	neutral	positive

Table 7: 关于自回归生成与非自回归生成的案例分析

试集中关于类别*environment_decoration*的真实标签为积极的，从主观分析也可以确认其标注的合理性。而两个消融实验中，只有无标签先验和自回归的实验无法正确预测Review 4在这一方面类别上的情感倾向。这直观地反映了自回归生成的引入对实验是有积极作用的。

通过以上分析，这些案例表明：在ACSA任务中自回归提示和标签先验知识的引入是有效的。

6 结束语

本文着重于缓解基于方面类别的情感分析中噪声标签对分类质量的影响问题，提出了一种自回归提示训练的生成式情感分析方法，从而生成具有更好语义一致性的标签词，并通过伯努利分布引入标签先验知识，以减轻噪声标签的干扰。在五个数据集上的实验结果表明，本文的方法在F1值方面优于最先进的方法。今后的研究将在提示模板的学习上考虑引入连续模式，增加情感分离方法的质量。

参考文献

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics2020*, pages 833–843, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022*, pages 2778–2788, Virtual Event, Lyon, France. ACM.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained

models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Zehui Dai, Wei Dai, Zhenhua Liu, Fengyun Rao, Huajie Chen, Guangpeng Zhang, Yadong Ding, and Jiyang Liu. 2019. Multi-task multi-head attention memory network for fine-grained sentiment analysis. In *Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing*, pages 609–620, Cham. Springer International Publishing.

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. Association for Computational Linguistics.

Yujie Fu, Jian Liao, Yang Li, Suge Wang, Deyu Li, and Xiaoli Li. 2021. Multiple perspective attention based on double bilstm for aspect and sentiment pair extract. *Neurocomputing*, 438:302–311.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Xin Guo, Geng Zhang, Suge Wang, and Qian Chen. 2020. Multi-way matching based fine-grained sentiment analysis for user reviews. *Neural Computing and Applications*, 32(10):5409–5423.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. CAN: Constrained attention networks for multi-aspect sentiment analysis. 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 4601–4610, Hong Kong, China. Association for Computational Linguistics.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. Computation and Language, *arXiv preprint arXiv:2108.02035*. Version 2.

Yu-Feng Li, Lan-Zhe Guo, and Zhi-Hua Zhou. 2021. Towards safe weakly supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):334–346.

Bin Liang, Hang Su, Rongdi Yin, Lin Gui, Min Yang, Qin Zhao, Xiaoqi Yu, and Ruifeng Xu. 2021. Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 208–218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. Computation and Language, *arXiv preprint arXiv:2107.13586*. Version 1.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. Computation and Language, *arXiv preprint arXiv:2103.10385*. Version 1.

Weiran Pan, Wei Wei, and Feida Zhu. 2022. Automatic noisy label correction for fine-grained entity typing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4317–4323. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris 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, Denver, Colorado. Association for Computational Linguistics.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, Austin, Texas. Association for Computational Linguistics.

(c) 2022 中国中文信息学会计算语言学专业委员会

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.

CCL 2022

面向话题的讽刺识别：新任务、新数据和新方法

梁斌[1]，林子杰[1]，秦兵[2,3]，徐睿峰[1,3]*
[1]哈尔滨工业大学（深圳），计算机科学与技术学院，深圳，518055
[2]哈尔滨工业大学，社会计算与信息检索研究中心，哈尔滨，150006
[3]鹏城实验室，深圳，518055
bin.liang@stu.hit.edu.cn , lzjjeffery@163.com
qinb@ir.hit.edu.cn, xuruifeng@hit.edu.cn

摘要

现有的文本讽刺识别研究通常只停留在句子级别的讽刺表达分类，缺乏考虑讽刺对象对讽刺表达的影响。针对这一问题，本文提出一个新的面向话题的讽刺识别任务。该任务通过话题的引入，以话题作为讽刺对象，有助于更好地理解和建模讽刺表达。对应地，本文构建了一个新的面向话题的讽刺识别数据集。这个数据集包含了707个话题，以及对应的4871个话题-评论对组。在此基础上，基于提示学习和大规模预训练语言模型，提出了一种面向话题的讽刺表达提示学习模型。在本文构建的面向话题讽刺识别数据集上的实验结果表明，相比基线模型，本文所提出的面向话题的讽刺表达提示学习模型取得了更优的性能。同时，实验分析也表明本文提出的面向话题的讽刺识别任务相比传统的句子级讽刺识别任务更具挑战性。

关键词： 面向话题的讽刺识别 ； 讽刺识别 ； 提示学习

Topic-Oriented Sarcasm Detection: New Task, New Dataset and New Method

Bin Liang[1], Zijie Lin[1], Bing Qin[2,3], Ruifeng Xu[1,3]*
[1]School of Computer Science and Technology,
Harbin Institute of Technology, Shenzhen, 518055
[2]Research Center for Social Computing and Information Retrieval,
Harbin Institude of Technology, Harbin, 150006
[3] Peng Cheng Laboratory, Shenzhen, 518055
bin.liang@stu.hit.edu.cn , lzjjeffery@163.com
qinb@ir.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Existing research on sarcasm detection generally attempts to recognize the sentence-level sarcastic expression from the context, which lacks of the consideration of the satirical object on the sarcastic expression. Therefore, This paper proposes a new topic-oriented sarcasm detection task, which can better understand and model the sarcastic expression by introducing the topics as the satirical objects. Correspondingly, this paper constructs a new dataset for topic-oriented sarcasm detection, which consists of 707 topics and 4871 topic-comment pairs. Based on this dataset, a topic-based prompt learning model is proposed to deal with the topic-oriented sarcasm detection based on the large-scale pre-trained language model and prompt learning. Experimental results on the proposed topic-oriented sarcasm dataset show that our proposed topic-based prompt learning model outperforms the baseline models. Simultaneously, the in-depth analysis show that the proposed topic-oriented sarcasm detection task is more challenging compared to the traditional sentence-level sarcasm detection.

Keywords: Topic-oriented sarcasm detection , Sarcasm detection , Prompt learning

* 通讯作者

第二十一届中国计算语言学大会论文集，第557页-第568页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

讽刺是一种常见的语言现象，通常使用比喻、夸张等手法对人或事进行揭露、批评或嘲笑，在语言学、心理学和认知科学等领域都得到了广泛关注 (Gibbs, 1986; Gibbs, 2007; Kreuz and Glucksberg, 1989; Kreuz and Caucci, 2007)。韦氏词典（Merriam Webster）[0]将讽刺定义为“使用与你真正想说的意思相反的词语，尤其是为了侮辱某人、表示愤怒或搞笑情绪。”（*the use of words that mean the opposite of what you really want to say especially in order to insult someone, to show irritation, or to be funny.*）。从讽刺的定义可以看出，讽刺表达通常是针对人、事物等讽刺对象而做出的语言表达。但目前大多数文本讽刺检测的研究局限于句子级别的讽刺识别和分类 (Joshi et al., 2015a; Kumar Jena et al., 2020; Xiong et al., 2019; Lou et al., 2021)，而忽略了讽刺对象对讽刺表达的影响。随着社交媒体平台的飞速发展，越来越多网络用户会对热点事件发表想法和评论，包括大量讽刺表达。其中大量评论都是基于特定事件产生的。因此，仅从评论本身出发分析其中的讽刺信息，不足以准确全面地理解用户对特定事件的实际情感。

针对这一问题，本文从一种新的角度观察讽刺表达，提出一个面向话题的讽刺识别任务。由于目前尚未有面向话题的讽刺识别公开数据集，本文设计构建了一个面向话题的讽刺识别新数据集。该数据集包含707个话题以及对应的4871个样本。其中，每一个样本由一个话题和一个评论组成。讽刺识别模型需要针对特定话题从句子的上下文中判断该评论是否为讽刺表达（讽刺或非讽刺）。针对这一问题，本文基于提示学习（prompt learning），提出一种面向话题的讽刺表达提示学习（Topic-Oriented Sarcasm Prompt Learning，TOSPrompt）模型。这一模型通过针对话题设计提示模板，可以更好地从大规模预训练语言模型（pre-trained language model，PLM）中理解句子对于话题的讽刺表达信息，从而判断该句子是否为讽刺句子。

本文的主要贡献如下：

（1）本文首次以一种新的角度观察讽刺表达，并提出一种面向话题的讽刺识别任务。

（2）本文构建了一个新的数据集用于评估面向话题的讽刺识别任务。通过数据的开源，更好地推动这一问题的研究。

（3）本文提出了一种面向话题的讽刺表达提示学习模型。该模型能有效建模面向话题的讽刺识别任务，并取得比基线模型更优的性能。

2 相关工作

2.1 讽刺识别

句子级的文本讽刺识别旨在从句子中识别上下文的讽刺表达，判断句子是否为讽刺句 (Joshi et al., 2015b)。一些早期的研究工作使用特征工程方法来.提取了句子上下文中不一致的情感表达，例如在上下文中搜索一组积极情感的动词和消极情感的表达 (Riloff et al., 2013; Bamman and Smith, 2015)、或构建词汇特征来确定不一致性 (Davidov et al., 2010; González-Ibáñez et al., 2011; Lunando and Purwarianti, 2013)，从而识别上下文的讽刺表达。

随后，基于深度学习的方法被广泛应用于句子级讽刺识别任务中。例如 (Poria et al., 2016; Majumder et al., 2019)采用预先训练的卷积神经网络（convolutional neural networks，CNN）架构来提取上下文的情感和个性特征，用于文本讽刺识别。Zhang等人 (2016)利用双向门控递归神经网络和池化神经网络来捕获推特内容和上下文信息，从而识别推特的讽刺表达信息。孙等人 (2016)提出一种融合多特征的混合神经网络判别模型。该模型融合了CNN和长短期记忆网络（long short-term memory，LSTM），有效挖掘文本中深层次的语义信息，完成讽刺识别。Tay等人 (2018)引入注意力机制，结合神经模型对上下文的情感对比和不协调情感表达进行建模，识别上下文的讽刺信息。Xiong等人 (2019)提出一种结合自匹配网络、双向LSTM网络和低阶双线性池方法的神经网络模型，从而学习单词对之间的不协调情感表达。在基于图网络模型的讽刺识别研究中，Lou等人 (2021)提出了一个基于依赖树和情感知识的情感依赖图网络模型，可以学习上下文中复杂的情感依赖信息，挖掘讽刺表达。

此外，随着大规模预训练模型在自然语言处理任务中取得的成功，也有研究工作将强大的预训练模型应用于讽刺识别任务中。Babanejad等人 (2020)利用情感信息和上下文特征来改

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

[0]http://www.merriam-webster.com/dictionary/sarcasm

进BERT (Devlin et al., 2019)结构，使其可以有效识别文本中的讽刺表达。樊等人 (2021)针对讽刺识别中较少利用上下文语境信息和修辞表达信息的不足，通过结合大规模预训练模型ELMo (Peters et al., 2018)，提出了基于多语义融合的反讽识别方法。上述研究工作在句子级的讽刺识别任务中取得了较好性能，但是这些研究工作都是局限于句子级的文本讽刺识别，忽略了讽刺表达中的讽刺对象和话题信息。

2.2 提示学习

随着ELMo (Peters et al., 2018)、BERT (Devlin et al., 2019)、GPT(Radford et al., 2018; Radford et al., 2019; Brown et al., 2020)等大规模预训练语言模型的发展，自然语言处理的研究和应用很多趋向于以预训练语言模型为中心，在下游任务进行微调的方式来解决问题。近年来，基于提示学习（prompt learning）的研究也受到了越来越多学者的关注，其主要思想是通过改造下游任务、增加专家知识，使任务输入和输出适合原始语言模型，从而获得更好的任务效果。Petroni等人(2019)将关系抽取任务修改为填空题，在不修改预训练语言模型的情况下，得到了比融合知识库更好的关系抽取性能。Shin等人(2020)提出了一种基于梯度引导搜索的自动方法：AUTOPROMPT。该方法可以为一组不同的任务自动创建提示。与手动创建提示的方法相比，AUTOPROMPT的提示可以从掩码语言模型（MASKed language model，MLM）中获得更准确的事实知识。Schick和Schutze(2021)基于大规模预训练语言模型，引入模式利用训练（pattern-exploiting training，PET）方法。该方法是一种半监督的训练方法，将输入样本重新设置为完形填空式短语，以帮助语言模型理解给定的任务。然后使用这些短语为大量未标记的示例指定软标签。最后，在结果训练集上执行有监督训练。这类基于提示学习的方法通过模板的引入，使得下游任务更好地匹配大规模预训练语言模型，在很多自然语言处理任务中都取得了令人瞩目的效果。受这些工作的启发，本文提出一种面向话题的讽刺表达提示学习模型，该模型通过设计面向话题的提示模板，更好地解决面向话题的讽刺识别问题。

3 面向话题的讽刺识别任务

与句子级讽刺识别不同，本文提出了一种面向话题的讽刺识别任务。这一任务通过话题的引入，以话题为讽刺背景/对象，判断文本是否为讽刺表达。例如表1给出的讽刺表达示例。对于同一个句子“真的很优秀”，样例1是一个没有话题信息的文本句子，可以看出，单纯从句子的上下文信息，难以判断该句子是否为讽刺表达。样例2是带有话题的句子。可以看出，结合话题信息可以很容易地判断该句子为讽刺句。而在样例3给定的话题场景下，可以判断样例3不是讽刺句。这就意味着在面向话题的讽刺检测任务中，当针对的话题不同时，同一个句子也有可能是不一样的讽刺标签。这就意味着必须深入结合话题信息，才能更好地判断句子是否为讽刺表达。可以看到，与传统的句子级讽刺识别相比，面向话题的讽刺识别任务更贴切真实场景。

形式化定义：对于给定的输入$x = (t, s)$，面向话题的讽刺识别旨在从评论s中挖掘针对话题t的讽刺表达信息，从而判断s针对t的类别y为“讽刺”或“非讽刺”。

样例	话题	句子	标签
1	-	真的很优秀	?
2	美国两党被曝都曾花钱挖特朗普黑料	真的很优秀	讽刺
3	男孩曾多次获得国家级竞赛奖项	真的很优秀	非讽刺

Table 1: 与话题相关的讽刺表达示例

4 面向话题的讽刺识别数据集

目前尚未有公开的面向话题的讽刺识别标注数据。为此，本文基于中文社交媒体文本，设计并构建了一个新的面向话题的讽刺识别数据集。具体地，为了收集带有话题的讽刺表达文本数据，本文从“观察者”[1]爬取带有话题的中文评论文本，形成初始数据。“观察者”是一个集新闻传播、人文社会科学研究于一体的新闻与评论一体化网站，反映了当前中国与世界各种思潮的对抗，该网站具有新闻内容更新快、活跃用户多、用户对新闻事件评论多、用户之间讨论活跃等特点。接下来本文详细介绍数据集的处理和标注过程。

[1] https://www.guancha.cn/

(c) 2022 中国中文信息学会计算语言学专业委员会

4.1 数据处理

考虑到数据集的规范性、通用性以及可扩展性，本文根据以下方面筛选待标注数据：

- 为确保数据的规范性，在话题选取过程中屏蔽掉包含敏感词语、讽刺表达比例低、攻击性强等话题。
- 因为过长的句子会影响模型对讽刺表达的学习能力，因此过滤掉长度超过200个词语的长文本数据。
- 以话题-评论对来构建一个数据样本，并过滤掉重复的话题-评论对。
- 过滤掉数据中的特殊符号、网页地址等与语义信息无关的信息。

这样可以得到面向话题的讽刺识别初始数据。其中，每一条数据样例由一个评论和对应的话题组成。

4.2 数据标注

在数据标注过程中，针对每一条数据样例，由5名标注者进行独立标注。整个标注过程持续了4个月。对于标注结果不一致的样例，通过多数投票机制获得最终的类别标签。同时，在标注过程中，为了提升数据集的质量，标注者丢弃了原始数据中约20%的噪音数据。包括：

- 评论和话题内容不相关的数据。
- 评论带有敏感词语、攻击性词语等表达的数据。
- 5名标注者都难以通过话题和评论内容判断类别标签的数据。

此外，因为话题通常是有明确的主题，并且有规范的信息表达，而从社交平台中爬取到的某些原始话题会存在噪音。因此，本文在标注过程中对原始话题进行了修正，包括：

- 删除不合适的断句，例如表2话题1中的“外交部回应”。
- 删除话题中的冗余信息，例如表2话题2中的“又一个！”。
- 重新组织话题表述，使话题更通顺，例如表2话题3中前后两个分句缺少的因果关系连接词“造成”。

Id	原话题	修改后话题
1	新西兰禁止运营商使用华为5G技术 外交部回应	新西兰禁止运营商使用华为5G技术
2	又一个！萨尔瓦多与台湾“断交”	萨尔瓦多与台湾“断交”
3	英国曼彻斯特发生恐怖爆炸袭击 22死59伤	曼彻斯特发生恐怖爆炸袭击造成22死59伤

Table 2: 话题处理示例

样例	话题	评论	类别
1	美国驱逐舰又撞船了	真会玩，在海上开“碰碰船”	讽刺
2	中国首款RISC-V高性能家电芯片在青岛诞生	加油	非讽刺

Table 3: 讽刺标注结果示例

受现有的句子级讽刺识别研究工作的启发 (Joshi et al., 2015b; Xiong et al., 2019; Lou et al., 2021)，本文针对每一个话题-评论对组成的样例标注为“讽刺”或“非讽刺”类别，类别标签标注举例如表3所示。因此，面向话题的讽刺识别任务中每一条标注样例可以表示为(t, s, y)，其中y为该样例的类别标签。

为了保证面向话题的讽刺识别任务具有更合理的评估结果，本文尽可能使每一个话题都包含讽刺样本和非讽刺样本。同时，为了避免因样本类别分布过于不平衡而导致无效的模型训练，本文从标注好的数据集中随机挑选样本，构建平衡“讽刺”和“非讽刺”两个类别数据分布的数据集。最终得到一个面向话题的讽刺识别（Topic-Oriented Sarcasm Detection，ToSarcasm）标注数据集。该数据集包含4871个由话题-评论对组成的样本，其中话题有707个。数据集在各个类别下的样本分布情况如表4所示。

(c) 2022 中国中文信息学会计算语言学专业委员会

	讽刺样本	非讽刺样本	总计
样本数量	2436	2435	4871
样本占比（%）	50.01	49.99	100

Table 4: ToSarcasm数据集的数据分布

5 面向话题的讽刺表达提示学习模型

面向话题的讽刺识别任务的主要难点是需要结合特定话题，判断评论针对该话题是否为讽刺表达。因此，本文借助大规模预训练语言模型的优势，设计面向话题的讽刺表达提示模板，目的是让模型更好地从预训练语言模型中学习面向话题的讽刺表达知识。如图1所示，本文提出的面向话题的讽刺表达提示学习（TOSPrompt）模型主要由三个模块组成：1）面向话题的讽刺表达模板构造。根据输入的话题和评论构造面向话题的讽刺表达提示模板；2）标签词预测。通过预训练语言模型给模板中的“$[MASK]$”位置预测标签词；3）模型训练。根据训练数据对模型进行训练优化。接下来，本节将详细介绍TOSPrompt模型的各个模块。

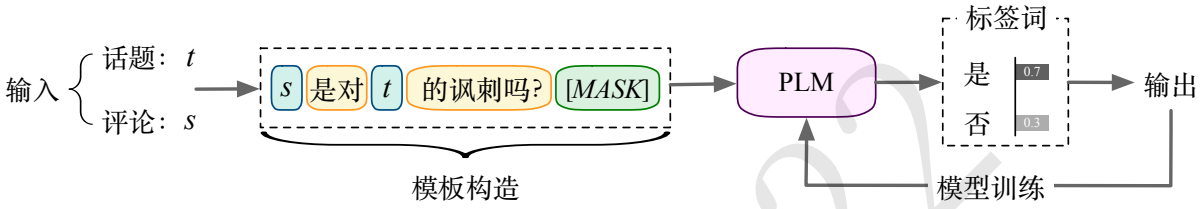


Figure 1: 面向话题的讽刺表达提示学习（TOSPrompt）模型框架图

5.1 面向话题的讽刺表达模板构造

为了使面向话题的讽刺识别任务更适合预训练语言模型，本文基于 (Shin et al., 2020; Schick and Schütze, 2021)等工作提出的基于提示学习的模型，以前缀提示（prefix prompt）模板的形式针对输入样本构造面向话题的讽刺表达模板。因此，本文基于预训练语言模型的掩码语言模型（masked language model）来对“$[MASK]$”标记位置进行词语填补。使用掩码语言模型的优势在于，通过掩码语言模型，可以基于大规模的预训练语料，利用非掩码区域的特征来为掩码位置“$[MASK]$”预测出合适的词语，从而预测出合适的类别标签。所构造的面向话题的讽刺表达模板定义如下：

$$\mathbf{x}_{prompt} = s\text{是对}t\text{的讽刺吗? } [MASK] \tag{1}$$

基于此，可以得到输入样例的面向话题的讽刺表达模板。接下来，需要借助预训练语言模型（PLM）对“$[MASK]$”位置进行类别标签词预测，从而识别该样例是否为讽刺表达。

5.2 标签词预测

针对给定输入(t, s)的模板“$\mathbf{x}_{prompt} = s$是对t的讽刺吗? $[MASK]$”，本文使用BERT中文预训练语言模型（BERT-base Chinese）(Devlin et al., 2019)作为预训练语言模型对输入样例进行建模。模型的输入表示为：

$$\boldsymbol{x} = [CLS]\mathbf{x}_{prompt}[SEP] \tag{2}$$

随后，将输入表示输入到BERT预训练语言模型$\mathcal{M}$中，以掩码语言模型的方式通过语言模型$\mathcal{M}$预测“$[MASK]$”位置的类别标签词分布：

$$P_{\mathcal{M}}([MASK]|\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}) \tag{3}$$

这里，为了使提出的TOSPrompt模型简单且具有更强的通用性，本将词语“是”和“否”作为模型的类别标签词。即，标签词集合$\mathcal{V} = \{\text{是}, \text{否}\}$，分别对应“讽刺”类别和“非讽刺”类别。

(c) 2022 中国中文信息学会计算语言学专业委员会

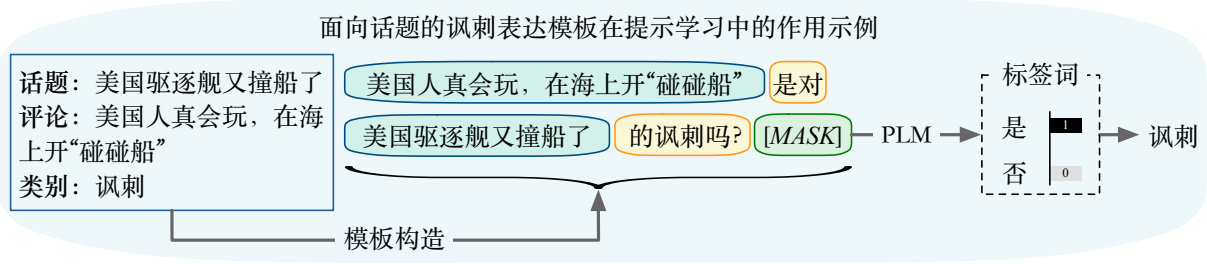


Figure 2: 面向话题的讽刺表达模板在提示学习中的作用示例

图 2给出了一个面向话题的讽刺表达模板在提示学习中的作用示例。如图2所示，借助面向话题的讽刺表达模板，模型针对该输入样本，可以更好地从预训练语言模型中学习到“$[MASK]$”位置的标签词为“是”，对应“讽刺”类别。

5.3 模型训练

基于上述的预训练语言模型$\mathcal{M}$，可以得到$[MASK]$位置预测为标签词集合中每一个标签词v的概率分布。这里，为了将单词的概率映射到标签的概率上，本文定义了一个映射器（verbalizer）将标签词集合$\mathcal{V}$中的词语映射到类别分布空间$\mathcal{Y}$中，即$f: \mathcal{V} \mapsto \mathcal{Y}$。因此，对于输入样本$\boldsymbol{x}$预测出标签词$v \mapsto y$的类别分布$P(y|\boldsymbol{x})$计算如下：

$$P(y|\boldsymbol{x}) = g\big(P_{\mathcal{M}}([MASK] = v|\boldsymbol{x})\big) \tag{4}$$

其中，$g(\cdot)$为将标签词的概率转换为类别标签概率的函数。因此，对于本文设计的标签词集合$\mathcal{V}$，输入样本x的预测类别标签定义为：

$$\text{argmax } \hat{\boldsymbol{y}} = \underset{y \in \mathcal{Y}}{\text{argmax}}\ P_{\mathcal{M}}([MASK] = v_y|\boldsymbol{x}) \tag{5}$$

其中，$\hat{\boldsymbol{y}}$为预测类别分布。v_y为类别标签y对应的标签词。随后，基于训练数据，通过最小化交叉熵损失对提出的TOSPrompt模型进行训练和优化：

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{L} y_i^j \log \hat{y}_i^j + \lambda ||\Theta||^2 \tag{6}$$

其中，N为训练集大小。L为类别数量。$\boldsymbol{y}_i$和$\hat{\boldsymbol{y}}_i$分别代表训练样本i的真实类别和预测类别分布。Θ为模型中所有的可训练参数。λ为L_2正则化系数。

6 实验

本文在构建的面向话题的讽刺识别新数据集（ToSarcasm数据集）上进行实验和分析。通过与现有的讽刺识别基线模型和大规模预训练语言模型进行对比实验，评估本文提出的TOSPrompt模型在面向话题的讽刺识别任务的性能。同时，通过实验和分析，评估本文提出的面向话题的讽刺识别新任务的研究价值和挑战性。

6.1 实验数据与参数设置

为了评估本文提出的TOSPrompt模型在面向话题的讽刺识别任务中的性能，本文首先对ToSarcasm数据集进行数据划分。对于每一条标注数据$x_i = (s_i, t_i, y_i)$，文本将其随机地分配给训练集、验证集或测试集。其中，训练集、验证集和测试集的比例为：6:2:2。基于此，可以得到ToSarcasm数据集的在训练集、验证集和测试集上的数据集合，数据统计如表5所示。

在实验中，本文采用JIEBA分词工具对文本进行中文分词处理[2]。对于本文提出的TOSPrompt模型，使用BERT中文预训练语言模型（BERT-base Chinese）(Devlin et al.,

[2] https://github.com/fxsjy/jieba

(c) 2022 中国中文信息学会计算语言学专业委员会

标签	训练集样本数	验证集样本数	测试集样本数
讽刺	1464	486	486
非讽刺	1461	487	487
总计	2925	973	973

Table 5: ToSarcasm数据集的数据统计信息

2019)作为预训练语言模型编码器，将输入样本编码成768维的向量。L_2正则化系数λ设置为0.00001。Dropout设置为0.1。模型采用Adam算法作为参数优化器，学习率设置为0.00002，权重衰减系数设置为0.002。批量大小Mini-batch设置为32。

6.2 评估指标

在模型评估中，本文采用精确率（Precision）、召回率（Recall）、F1值（F1-score）以及准确率（Accuracy）综合评估模型在面向话题的讽刺识别分析任务中的性能。因为面向话题的讽刺识别任务主要关注模型是否能正确识别出样本中的讽刺表达，同时区分非讽刺表达。基于此，本文设定“讽刺”类别为正例（Positive），“非讽刺”类别为反例（Negative）。因此，准确率表示预测正确的样本占所有样本的比例；精确率表示预测为“讽刺”的样本中有多少是真正的“讽刺”样本；召回率表示针对原来的样本而言，样本中的“讽刺”类别样本有多少被预测正确了；F1值则综合精确率和召回率的结果。各评估指标计算公式如下：

$$\text{准确率} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{精确率} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{召回率} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1值} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{10}$$

其中，TP（True Positive）表示预测类别和实际类别都为“讽刺”。FP（False Positive）表示预测类别为“讽刺”，而实际类别为“非讽刺”。FN（False Negative）表示预测类别为“非讽刺”，而实际类别为“讽刺”。TN（True Negative）表示预测类别和实际类别都为“非讽刺”。

6.3 对比模型

在实验中，本文选取以下基线模型作为本文提出的TOSPrompt模型的对比模型：

- **BiLSTM**：使用一个双向的LSTMs (Hochreiter and Schmidhuber, 1997)分别学习句子和目标的隐藏层表示，最终将隐藏表示拼接作为讽刺识别的分类特征。
- **MIARN** (Tay et al., 2018)：句子级别的讽刺识别模型。使用基于注意的神经网络模型，分别对评论和话题进行讽刺表达的不一致性学习，最终将隐藏层表示拼接作为讽刺识别的分类特征。
- **ADGCN** (Lou et al., 2021)：句子级别的讽刺识别模型。基于外部情感知识的图网络模型，通过学习情感的不一致性挖掘上下文的讽刺表达。本文使用ADGCN分别对话题和评论进行建模，最终将隐藏表示拼接作为讽刺识别的分类特征[3]。
- **BERT** (Devlin et al., 2019)：原始的BERT中文预训练语言模型BERT-base Chinese，该模型使用“$[CLS]s[SEP]t[SEP]$”作为模型输入。
- **ADGCN-BERT** (Lou et al., 2021)：ADGCN的变种，编码器使用BERT-base Chinese。
- **PET** (Schick and Schütze, 2021)：基于Schick和Schutze (2021)的提示学习研究工作，使用“$[CLS]s[SEP]t[SEP]$这是$[MASK]$”构建模板输入到BERT-base Chinese中，预测$[MASK]$位置是“讽刺”或“非讽刺”。

[3]其中中文情感词得分来自BosonNLP：http://bosonnlp.com/

(c) 2022 中国中文信息学会计算语言学专业委员会

其中，对于BiLSTM、IAN、MIARN和ADGCN这四种非BERT基线模型，本文使用Shen等人 (2018)提出的Chinese Word Vectors中文词向量对词语进行词向量初始化。

6.4 面向话题的讽刺识别实验结果

为了评估本文提出的TOSPrompt模型在面向话题的讽刺识别任务中的有效性，本文在ToSarcasm数据集上与基线模型进行了对比实验，实验结果如表6所示。从表中结果可以看出，本文提出的TOSPrompt模型在所有的评估指标上都取得了最佳性能。这显示了本文提出的TOSPrompt模型在面向话题的讽刺识别任务中的有效性。与句子级讽刺识别对比模型（MIARN、ADGCN和ADGCN-BERT）相比，本文提出的TOSPrompt模型在所有评估指标上都取得了大幅度的提升。这表明，单纯的句子级讽刺识别模型并不能很好地处理面向话题的讽刺识别任务，而本文提出的TOSPrompt模型借助话题信息的建模，可以更好地解决面向话题的讽刺识别任务。另一方面，基于BERT的模型相比基于传统词向量的模型总体上取得更优的性能。这说明在面向话题的讽刺识别任务中，使用更强大的预训练语言模型在学习讽刺表达时能取得更好的学习效果。此外，相比现有基于提示学习的模型（PET），本文提出的TOSPrompt模型在所有评估指标上的性能都取得了提升，其中准确率提升了1.06%，F1值提升了1.76%。这说明，在面向话题的讽刺识别任务中，通过面向话题来设计讽刺表达的提示学习模板，可以更好地针对话题来学习评论上下文中的讽刺表达信息，从而取得更优的性能。

模型	准确率	精确率	召回率	F1值
Bi-LSTM	63.72	61.65	72.55	66.65
MIARN (Tay et al., 2018)	65.32	63.25	74.12	68.25
ADGCN (Lou et al., 2021)	65.90	63.16	76.50	69.19
BERT (Devlin et al., 2019)	69.66	67.26	74.79	70.83
ADGCN-BERT (Lou et al., 2021)	70.57	68.93	75.10	71.88
PET (Schick and Schütze, 2021)	70.70	67.57	75.78	71.44
TOSPrompt	**71.76**	**70.02**	**76.68**	**73.20**

Table 6: 模型在面向话题的讽刺识别任务上的性能（%）

6.5 不同模板对实验性能的影响

为了分析本文提出的TOSPrompt模型在使用不同模板的提示学习对性能带来的影响，本文以非面向话题和面向话题两个角度来构建模板，针对TOSPrompt模型设计了以下变种。

1）非面向话题的模板：

- $\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是$[MASK]$表达：该模板采用完形填空提示（cloze prompt）模板，预测$[MASK]$位置是“讽刺”或“非讽刺”。
- $\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是讽刺吗？$[MASK]$：该模板采用前缀提示模板，预测$[MASK]$位置是“是”或“否”。

2）面向话题的模板：

- $\mathbf{x}_{prompt} =$ 针对t的评论s是$[MASK]$表达：该模板采用完形填空提示模板，预测$[MASK]$位置是“讽刺”或“非讽刺”。
- $\mathbf{x}_{prompt} = s$针对t是$[MASK]$表达：该模板采用完形填空模板，预测$[MASK]$位置是“讽刺”或“非讽刺”。

表 7给出了不同模板在面向话题的讽刺识别任务上的实验性能。从表中结果可以看出，针对话题构建面向话题的讽刺表达提示模板在所有评估指标上都取得了比非面向话题的模板更好的性能。这说明，在面向话题的讽刺识别任务中，针对话题来设计讽刺表达提示模板可以更好地从预训练语言模型中挖掘出关于话题的语言知识，从而能更好地根据评论文本学习面向该话题的讽刺识别特征信息，取得更优的性能。

(c) 2022 中国中文信息学会计算语言学专业委员会

模板	准确率	精确率	召回率	F1值
$\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是$[MASK]$表达	70.86	68.62	76.43	72.31
$\mathbf{x}_{prompt} = s[SEP]t[SEP]$这是讽刺吗？$[MASK]$	70.92	69.49	75.61	72.42
$\mathbf{x}_{prompt}$ = 针对t的评论s是$[MASK]$表达	71.37	69.93	76.22	72.94
$\mathbf{x}_{prompt} = s$针对t是$[MASK]$表达	71.60	**70.35**	76.13	73.13
TOSPrompt	**71.76**	70.02	**76.68**	**73.20**

Table 7: 不同模板的性能（%）对比

6.6 不同比例训练数据的性能分析

为了评估训练数据样本数量对本文提出的TOSPrompt模型的性能影响，基于BERT、PET和所提出的TOSPrompt模型使用不同比例的训练数据集合进行了对比实验，结果如图3所示。从图中结果可以看出，相比于原始的BERT模型，基于提示学习的PET和本文提出的TOSPrompt模型在各个比例的训练数据下都取得了更优的性能。这也进一步显示了提示学习在面向话题的讽刺识别任务中的有效性。此外，本文提出的TOSPrompt模型在不同大小的训练数据下性能都始终优于BERT和PET，特别是在仅使用20%-60%的训练数据时提升尤为显著。这说明本文提出的TOSPrompt模型由于设计了面向话题的讽刺表达提示模板，能更好地从预训练语言模型中针对话题学习上下文中的的讽刺表达信息，因此在缺少训练数据时也能取得令人满意的性能。

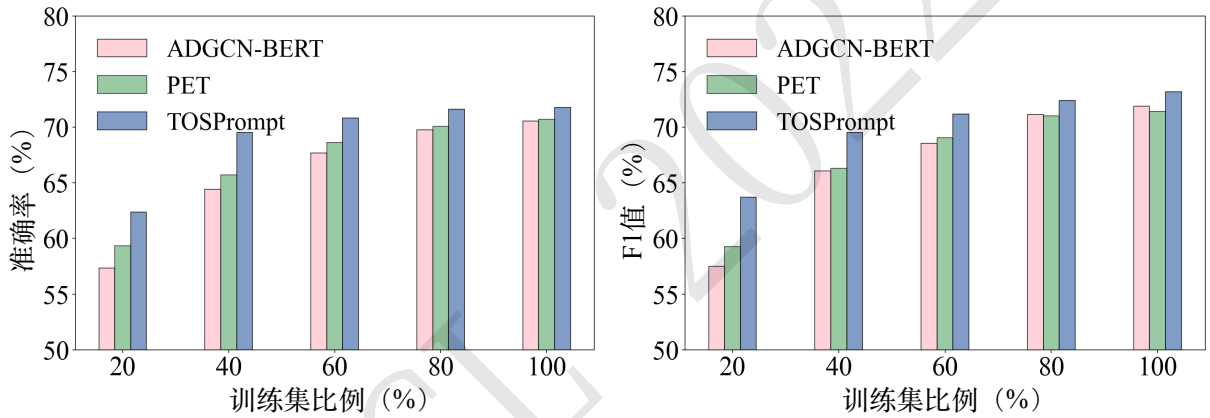


Figure 3: 使用不同比例训练数据的实验结果

样例	话题	评论
1	三位 大妈偷走巨型烧烤炉 ，被抓时正在买食材	中国 大妈们真厉害 ！
2	马克龙访美与特朗普 互动存在分歧	好甜蜜 ，羡慕这样的爱情。
3	特朗普当着一屋子 大学生 的面说： 我超爱贷款 ！	真棒 ， 好潇洒 的美国人！

Table 8: 典型的面向话题的讽刺样本示例

6.7 话题样例分析

本文提出的面向话题的讽刺识别分析任务相比传统的句子级讽刺识别任务，引入了讽刺表达中的话题对象。因此，本文进一步通过典型的讽刺样例来分析话题对讽刺识别的作用。

表 8给出三个典型的面向话题的讽刺表达样例，可以看出，仅从评论的内容都难以判断这三个样例是否为讽刺表达，因为评论中的上下文只表达了单向的积极情感，没有体现出讽刺表达中的情感不一致描述。而当结合话题信息后，可以看出，这三个样例都是讽刺表达样例。对

于样例1，话题中的“大妈偷走巨型烧烤炉”是负面的。因此，结合评论中的“大妈们真厉害”可以判断出该样例是讽刺样本。同样地，对于样例2和样例3，话题中的阴影部分内容都是带有负面情绪的表达，通过结合评论中的描述，可以推断出样例2和样例3也是讽刺样本。可以看出在面向话题的讽刺识别任务中，话题的内容对于判断样本是否为讽刺表达是至关重要的。这充分显示了本文提出的面向话题的讽刺识别任务的合理性和研究价值。此外，从典型样例也可以看出，解决面向话题的讽刺识别任务不仅仅需要针对话题和评论挖掘上下文的语义信息，还需要对话题和评论的上下文信息进行匹配，挖掘话题和评论中重要内容的联系（图8中对应背景颜色的文字描述）。这也意味着面向话题的讽刺识别任务相比句子级的讽刺识别任务具有更强的挑战性。

6.8 错误样例分析

从表 6中结果可以看出，所有的模型在面向话题的讽刺识别任务中的性能指标都没能超过80%，这也侧面反映了面向话题的讽刺识别任务的挑战性和研究价值。为了进一步分析任务数据中的挑战性，本文对提出的TOSPrompt模型的错误样本进行了分析，并将错误分类的样本大致归类为以下类型：

1）需要一定的背景知识才能了解话题和评论所表达的内容。例如以下例子，其正确类别标签为“讽刺”：

话题：又一国际巨头将撤离深圳！留下超10万平米土地谁接盘?

评论：你说的很有道理，深圳只是个小县城

该例子中“深圳”是一个大都市，而不是“小县城”，因此模型需要加入额外的背景知识才能更好地对其内容进行学习，得出正确的分类结果；

2）评论中带有缩写或非正规用词。例如以下例子，其正确类别标签为“讽刺”：

话题：萨尔瓦多与台湾“断交”

评论：恭喜菜菜，真的快“独”了！

该例子中的“菜菜”指的是“蔡英文”。因此，模型需要将这些词语映射为正规用词才能准确理解评论中的上下文信息表达，得出正确的分类结果；

3）评论中带有隐喻表达的词语。例如以下例子，其正确类别标签为“讽刺”：

话题：因缺少备件，德国海军潜艇全部趴窝

评论：“工匠”们累了，要休息啦！

该例子中评论内容将“德国海军”比喻成“工匠”，因此需要将隐喻表达跟话题中的事物对应起来才能识别样本的讽刺表达信息，得出正确的分类结果。

因此，针对上述的错误样例分析，在未来的研究中，可以考虑探索在模型中融入话题和评论中所讨论事物的背景知识、对评论文本中涉及的实体进行识别和共指消解、对样本中的隐喻表达进行识别和对齐等技术，以进一步提升面向话题的讽刺识别任务的性能。

7 结论

针对现有的讽刺识别研究通常只针对句子级别来挖掘上下文的讽刺表达信息，但忽略了讽刺表达的话题背景或讽刺对象的不足，本文提出一个新的面向话题的讽刺识别任务。该任务通过引入话题信息作为讽刺表达的对象，使讽刺识别的研究更贴切真实场景且更具挑战性。为此，本文构建了一个面向话题的讽刺标注数据集，以推动这一研究的开展。此外，为了解决面向话题的讽刺识别任务，本文基于提示学习，提出了一种面向话题的讽刺表达提示学习（TOSPrompt）模型。与一系列基线模型的对比实验结果表明，本文提出的TOSPrompt模型在面向话题的讽刺识别任务中取得了最佳性能。

参考文献

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth international AAAI conference on web and social media*, pages 574–577.

(c) 2022 中国中文信息学会计算语言学专业委员会

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. 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. Association for Computational Linguistics.

Raymond W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1):3.

Raymond W Gibbs. 2007. On the psycholinguistics of sarcasm. *Irony in language and thougt: A cognitive science reader*, pages 173–200.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015a. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015b. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.

Roger J Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of experimental psychology: General*, 118(4):374.

Amit Kumar Jena, Aman Sinha, and Rohit Agarwal. 2020. C-net: Contextual network for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 61–66, Online. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. 2021. Affective dependency graph for sarcasm detection. In *the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pages 1844–1849.

Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 195–198. IEEE.

(c) 2022 中国中文信息学会计算语言学专业委员会

Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612, Osaka, Japan. The COLING 2016 Organizing Committee.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2115–2124. ACM.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

孙晓, 何家劲, and 任福继. 2016. 基于多特征融合的混合神经网络模型讽刺语用判别. 中文信息学报, 30(6):215–223.

樊小超, 杨亮, 林鸿飞, 刁宇峰, 申晨, and 楚永贺. 2021. 基于多语义融合的反讽识别. 中文信息学报, 35(6):103–111.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于相似度进行句子选择的机器阅读理解数据增强

聂双， 叶正， 覃俊， 刘晶
中南民族大学　计算机科学学院　湖北省　武汉市　430074
819258834@qq.com　yezheng@scuec.edu.cn

摘要

目前常见的机器阅读理解数据增强方法如回译，单独对文章或者问题进行数据增强，没有考虑文章、问题和选项三元组之间的联系。因此，本文探索了一种利用三元组联系进行文章句子筛选的数据增强方法，通过比较文章与问题以及选项的相似度，选取文章中与二者联系紧密的句子。同时为了使不同选项的三元组区别增大，我们选用了正则化Dropout的策略。实验结果表明，在RACE数据集上的准确率可提高3.8%。

关键词： 多项选择 ； 长文本数据增强 ； 正则化Dropout

Machine reading comprehension data Augmentation for sentence selection based on similarity

Shuang Nie,　Zheng Ye ,　Jun Qin,　Jing Liu
School of Computer Science, South-Central University for Nationalities, Wuhan 430074
819258834@qq.com　yezheng@scuec.edu.cn

Abstract

At present, the commonly used data augmentation methods for machine reading comprehension, such as back translation, enhance the data of articles or questions alone, without considering the relationship among articles, questions and option triples. Therefore, this paper explores a data augmentation method for article sentence screening by using triplet connection. By comparing the similarity among the article and the questions and options, and select the sentences closely related to the two in the article. At the same time, in order to increase the difference between triples of different options, we use the strategy of regularizing dropout.The experimental results show that the accuracy can be improved by 3.8%.

Keywords: Multiple choice , Long text data augmentation , Regularized Dropout

1 引言

机器阅读理解是自然语言处理中重要的一环，与其他领域息息相关。机器阅读理解的目的是为了让计算机像人一样对文本进行理解 (Seo M et al., 2016)，进而能够实现阅读和推理。为

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第569页-第579页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

了了解计算机对文本的掌握能力，就需要进行计算机对问题回答的测试 (Tang M et al., 2019)。计算机能够回答的问题的难度和回答问题的正确率在一定程度上能够反映出计算机对文本的了解程度。

阅读理解可分为四类，填空式、选择式、抽取式和生成式 (Liu S et al., 2019)。多项选择式的阅读理解既有需要精读的细节题，也有需要总结的概括题，如图1所示，加粗的选项为问题的正确答案，其中的细节题需要找到答案位置，而总结题需要纵观全文，因此，多项选择阅读理解同时测试了计算机关注细节与整体推理能力，是一个综合性的任务，更具有挑战性。

随着预训练模型 (DEVLIN J et al., 2018) 的出现，机器阅读理解任务得到了快速发展，但是这种大型神经网络为了能够得到充分的训练，需要大量的数据来训练，训练的数据越多，往往训练得到的效果也就更好。因此如何得到更多高质量的数据就成为了关注点，然而多项选择式的阅读理解进行数据增强有两个挑战，一是阅读理解的目的是为了考察学生的快速阅读能力，所以文章设置多数很长，二是为了测试学生是否真正理解了文章而不是基于表面，设置的答案不一定是原文片段 (Zhu H et al., 2018)。这两个难点使得进行人工扩充的人员需要一定的知识储备，人工构建难度高。所以大家将目光放在了自动数据增强方法上，且预训练模型输入有长度限制，长文章无法全部输入模型，而目前常用的数据增强方法主要是针对能够输入模型的部分进行的。这种方法的缺点是忽略了被删掉文章部分，会导致有的问题找不到答案，因为通常设置的问题及其对应答案在文章中是均匀分布的。

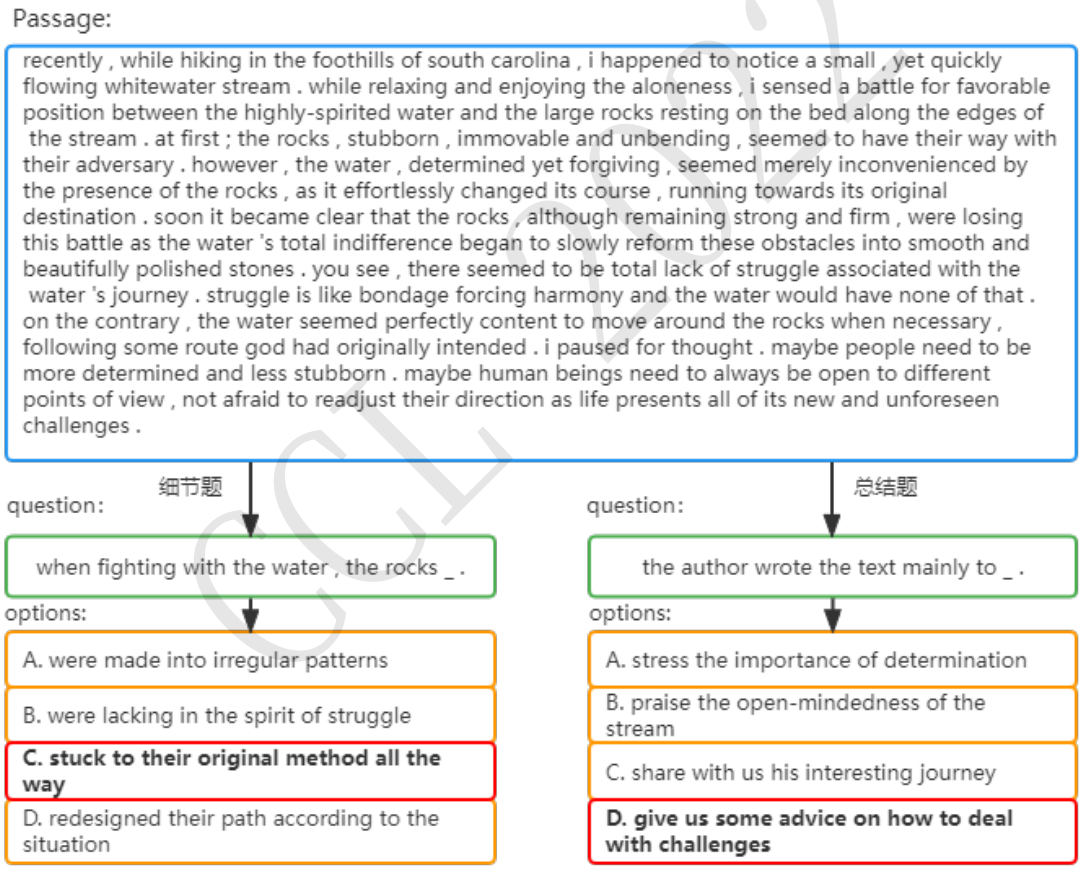


Figure 1: 机器阅读理解中的细节题与总结题比较

为了解决多项选择机器阅读理解数据集人工扩充难度大的问题，同时充分利用全部文本信息，本文提出了一个基于句子选择的自动数据增强方法。该方法首先利用文本相似度计算文章与问题以及选项的相似度，挑选出文章中与问题以及选项相关的句子，保持相对顺序不变的前提下形成新的文章。这种数据增强方法不仅能扩充数据集，而且能在保证原文语义不变的条件下让模型学到更多的内容。同时大型神经网络会采用Dropout来防止过拟合，但是Dropout在训

练时和推理时使用的策略不一样，即训练时会以概率P随机删除一些神经元，而在推理时将所有的神经元加入，因此会导致训练和推理的模型差异。为了消除两者之间的差异，我们引入了简单的R-Drop正则化 (Wu L et al., 2021)。

2 相关工作

机器阅读理解是机器通过阅读文章来回答问题的技术，机器阅读理解任务可分为基于规则的传统方法 (Smith E et al., 2015) 和基于深度学习的方法 (Seo M et al., 2016)。基于规则的传统方法由于受到数据集的限制和需要特征构建而不能取得较好的效果。随着CNN/DailyMail (Hermann K M et al., 2015)大型数据集的出现，深度学习在机器阅读理解任务得到了迅速的发展，已经成为当前的主流方法。基于深度学习的机器阅读理解要求训练数据足够多，训练的数据越多，往往训练得到的效果也就更好。而数据增强是可以解决数据匮乏的一种有效方法，且能够提高模型的准确率。因此许多学者开始研究基于文本的数据增强方法。

传统的文本数据增强方法主要可分为三类，基于字符层面的数据增强、基于词层面的数据增强和基于句子层面的数据增强。基于字符层面的数据增强一般是在文本中加入噪声，常用的有拼写错误注入、键盘错误注入，使所训练的模型对扰动具有鲁棒性。基于词层面的数据增强方法最常见的是词汇替换，词汇替换方法可分为基于词典的替换 (Zhang et al., 2015) 、基于词向量的替换 (Jiao X et al., 2020)和基于TF-IDF的词替换 (Xie Q et al., 2019)。这些方法都是将文本句子中的某个词替换为另一个相近词，基于词典的替换是随机将句子中的一个单词使用同义词词典替换为同义词，基于词向量的替换使用预先训练好的单词嵌入，如Word2Vec (Mikolov T et al., 2013)、GloVe等，并使用嵌入空间中最近的相邻单词替换句子中的某些单词，以此来提高语言模型在下游任务上的泛化能力。而基于TF-IDF的词替换的思想是分数较低的单词不能提供信息，因此可以在不影响句子的基本真值标签的情况下替换它们。随着预训练语言模型的发展，研究发现MLM（遮盖语言模型）通过预训练也能进行词的替换 (JM Tapia-Téllez and Escalante H J, 2020)，由于MLM是基于上下文来推测出遮盖词，因此替换的词拥有上下文语境，同一个词在不同的语境中可能会生成不同的同义词，从而对解决歧义问题有帮助。基于句子层面的数据增强比较常用的是回译方法 (Lee S et al., 2021)，回译的方式就是将句子翻译成另一种语言，然后再翻译成原来的语言。这种数据增强方式的优点是尽量保证了在原文意思不变的基础上生成了新的补充版本。还可以同时使用多种不同的语言来进行回译以生成更多的文本变体。最近通过对语言的语义分析，出现了在不改变语义的情况下进行语态转变的数据增强方法 (Dehouck M and Gómez-Rodríguez C, 2020)，这主要是通过语法分析建立依赖树，转换依赖树后生成意思相同的句子。

还有介于字符与词之间的数据增强方法，即使用正则表达式的简单的模式匹配的转换，文本表面转换 (Coulombe C, 2018)是其中常见的一种。是将动词形式由简写转化为完整形式或者反过来的方法。这种方法在扩展模棱两可的动词形式时可能会出现错误，为了避免出现这种问题，提出了允许模糊收缩，但跳过模糊展开的方法。

以上是自然语言处理领域通用的数据增强方法，在机器阅读理解任务中，常用的数据增强方法也可以分为以下几种:

(1)传统的简单数据增强方法EDA (Wei J and Zou K, 2019)，也是基于词层面的方法，利用对词的简单操作，同义词替换，随机插入，随机互换和随机删除来扩充数据，然而对于是使用预训练的模型来说，效果并没有什么提高。

(2)基于问题生成的数据增强，问题生成又进一步可以分成基于规则和模板 (Mitkov R, 2003)以及基于深度学习 (Mirshekari M et al., 2021; Yu A W et al., 2018)两种方式，前者是使用设定的规则或者模板来生成问题，效率低且泛化能力差。后者是通过将文章和答案放入生成模型中训练来形成新问题。这种数据增强方法尽量使生成的问题贴近原问题，以此达到原义相近的条件下问题增多的目的。然而这种方法当答案在文章中多次出现时，无法判断是哪个位置，因此可能会出现与原问题背道而驰的情况。而后者仅依赖文章来生成问题-答案对，这种生成新数据的方法生成的问题-答案对质量不高，会造成冗余的数据。

(3)无监督数据增强，在英语机器阅读理解中常使用的是英法回译方法 (Fabbri A R et al., 2021)，这样可以保留原义而生成不同的意译。然而对于问题中的关键字并不一定能保留下来，影响找答案的位置。

以上的数据增强方法遇到输入序列过长时，采取的是简单的截断处理。这么做的缺点是忽

(c) 2022 中国中文信息学会计算语言学专业委员会

视了截断文章部分，对于多项选择型的阅读理解任务来说，文章的问题在文章中是平均分布的，只截取前半部分文章会导致后半部分问题在输入的文本中找不到答案，因此生成新的数据也难以提高模型的准确性。

针对以上缺点，本文引出了一种同时兼顾机器阅读理解的长文本与数据量不够的数据增强方法，该方法是对数据集中的文章进行扩充，以增加文章样本多样性。但是本方法也考虑了长文本问题。通过机器阅读理解中的三元组信息，比较多个计算文本相似度将文章总与问题以及选项相关的重要信息提取出，构造出既与输入文本不完全一致，能够起到补充作用，又能不改变原文意思的数据增强方法。

3 方法

多项选择的机器阅读理解过程可以化为以下的三元组形式$< P, Q, A >$：给定的一篇文章P，根据文章内容提出的问题Q以及相对应的几个选项A。多个答案选项中只有一个选项为正确答案，其余的为迷惑的错误答案，目的就是为了选出正确答案。其中问题$Q = \{Q_1, Q_2, Q_3...Q_m\}$（m表示文章的问题数量），选项$A = \{A_1, A_2, A_3...A_n\}$（n表示每个问题的选项数量）。本文将运行过程分成了三个部分：第一部分是利用三元组中三者关系的相似性，计算文本相似度来抽取重要的句子生成新的文章来进行数据增强。第二部分将原来的数据和新生成的数据一起放入基线模型即双向匹配模型中进行训练。第三部分是利用自身的数据来进行样本间对比的R-Drop正则化来选出最终的答案。模型的整体架构如图2所示。

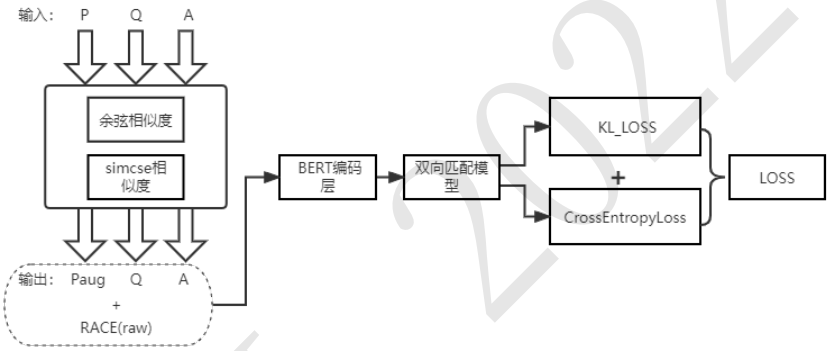


Figure 2: 结构图

3.1 数据增强模块

目前多项选择的机器阅读理解任务的训练数据量相对较少且人工扩充数据集较难，因此找到有效的数据增强方法很重要。多项选择阅读理解任务中问题取材于文章，答案也在文章中寻找。所以受解题信息来源于文章的启发，本文提出建立以文章为中心的数据增强方法。考虑到预训练模型长度限制，就首要挑选文章最重要的内容来进行数据增强。因此，本文研究了两种短文本相似度的方法来获取文章的核心句子，分别是余弦相似度和SimCSE相似度 (Gao T et al., 2021)。余弦距离是通过算出两个向量的夹角余弦值来衡量两者相似程度的。对于文本来说余弦距离是通过利用两个短文本词频向量来计算相似性的。余弦相似度由于通常用于正空间，因此规定夹角余弦取值范围为[0,1]。通过余弦值可以判断两个向量指向相似度。0度为重合，余弦值为1，90度余弦值为0。两个向量越相似，余弦值越大。短文本余弦相似度的计算方法为：

$$similarity = Cos\theta = \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{1}$$

其中的A_i,B_i分别表示两个句子的第i个词频向量。文章中每个句子与问题、每个句子与选项的相似度可计算出：

$$S^{pq} = Cosine(p_i, q) \tag{2}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

$$S^{pa} = Cosine(p_i, a) \tag{3}$$

$$Score^{pq} = W_1 S^{pq} \tag{4}$$

$$Score^{pa} = W_2 S^{pa} \tag{5}$$

$$M_p = K_{max}(Score^{pq}, Score^{pa}) \tag{6}$$

其中，S^{pq}是问题和文章中第i个句子余弦相似度，S^{pa}是每个选项和文章中第i个句子的余弦相似度。P_i表示文章中的第i个句子，q表示问题，a表示选项。其中W_1和W_2表示可调节的参数。$Score^{pq}$，$Score^{pa}$分别表示文章与问题，文章与选项之间的相似度倒序排列。K_{max}表示将$Score^{pq}$，$Score^{pa}$序列的前K个值，M_p是从原文章中按照K的取值抽取的句子。

SimCSE相似度是利用对比学习的方式来进行模型训练，通过拉近相似的数据推远不相似的数据的方式来进行对比。

$$l_i = -log\frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(h_i, h_j)/\tau}} \tag{7}$$

其中，τ表示温度超参数，h_i，h_i^+，h_j表示第i个样本、其对应的正样本以及其对应的负样本经过模型编码得到的向量。sim表示相似度，l_i表示第i个样本的损失函数。

本文采取的相似度方法是分别提取文章句子与问题，文章句子与每个选项之间的相似度的前k个句子，这是为了让挑选出的句子尽可能与问题和每个选项有关联。同时为了最大保持原始文本的语义信息，我们将抽取出来的句子仍然按照原始文本的顺序采集出来，不破坏原文的相对位置。

3.2 多项选择的基础模型

本小节将介绍多项选择型的阅读理解模型的基本架构，本文选用DCMN(Dual Co-matching network) (Zhang S et al., 2020)作为基线模型，该基线模型主要由三部分组成，首先使用预训练模型对三元组进行编码，获得编码信息后将三元组信息通过双向匹配模块相互进行融合和交互，最后再进行答案选择的输出。同时在此模型上采用R-Drop作为辅助策略。

3.2.1 三元组编码

对三元组的编码需要获取上下文信息，所以我们选择使用预训练模型（BERT）作为编码器。BERT把三元组即文章P,问题Q,选项A中的每一个词编码成固定长度的向量，编码之后可以获得:

$$H^p = Encode(M^p) \tag{8}$$

$$H^q = Encode(M^q) \tag{9}$$

$$H^a = Encode(M^a) \tag{10}$$

H^P,H^q,H^a分别表示文章、问题、选项在预训练模型中最后一层的输出表示。

3.2.2 双向匹配模块

为了充分提取三元组两两之间的信息，我们需要用到前面预训练模型作为编码器的输出表示H^p,H^q,H^a。注意力机制方式本文采用了双向匹配机制，分别获得文章-问题表示，文章-选项表示，问题-选项表示。

$$C = [M^{pq}; M^{pa}; M^{qa}] \tag{11}$$

M^{pq}、M^{pa}、M^{qa}分别是文章-问题对、文章-选项对、问题-选项对的匹配表示，C是最终每个问题对应的三元组表示。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.2.3 正则化Dropout思想

第一个句子选择模块对多项选择的机器阅读理解中的文章提取了多方面且精准的内容。而模型的双向匹配机制提供了两两配对，将三元组的内容通过提取融合在了一起。多项选择机器阅读理解采用是交叉熵损失函数，Dropout两次就能得到两个不同的子模型分布，即：

$$L^{E}_{P,Q,A_i} = -logp_1(y_i|x_i) - logp_2(y_i|x_i) \tag{12}$$

正则化Dropout就是控制使用两个不同的子模型预测的结果能尽量保持相同，达到模型优化的目的。通过最小化两个分布之间的双向KL散度，减小Dropout带来的训练和测试时带来的不同。

$$L^{KL}_{P,Q,A_i} = \frac{1}{2}[KL(P_2(y_i|x_i)|P_1(y_i|x_i)) + KL(P_1(y_i|x_i)|P_2(y_i|x_i))] \tag{13}$$

所以最终的损失函数为两者的加权和：

$$L = L^{E}_{P,Q,A_i} + \alpha.L^{KL}_{P,Q,A_i} \tag{14}$$

4 实验与分析

4.1 实验数据集及评价指标

本文研究的是机器阅读理解型的特定任务的数据增强方法，本文选择在多项选择型的机器阅读理解任务数据集RACE (Lai G et al., 2017)上进行实验设计并进行分析，来验证本文所论述的数据增强方法的有效性。RACE 是一个来源于初高中学生英语考试题目的大规模多项选择型的阅读理解数据集，RACE的形式是给定一个三元组<文章，问题，选项>，通过阅读并理解文章，对提出的问题从四个选项中选择正确的答案。该数据集由初中阅读RACE-MIDDLE和高中阅读RACE-HIGH组成，表1显示了这两个子集的训练集中文章长度分析，在5个长度区间统计了包含的文章数量以及占的总比、文章长度的平均值。如表1所示，训练集中文章长度在500以下的仅占全部数据集的2.1%,有98%左右的文章是不能全部放入模型中的。

passage_len	**0-500**	**500-1000**	**1000-1500**	**1500-2000**	>2000
RACE_M	493	2453	2519	816	128
RACE_H	34	608	3873	9829	4384
SUM	527	3061	6392	10645	4512
% ON RACE	2.10%	12.18%	25.43%	42.34%	17.95%

Table 1: RACE训练集文章长度统计

多项选择机器阅读理解的评价指标采用的是准确率指标ACC(Accuracy):

$$ACC = \frac{right}{all} \tag{15}$$

right表示的是模型预测正确的数量，all表示问题总数。

4.2 实验参数设置

为了证明该阅读理解任务数据增强方法的有效性，本文选择了双向匹配策略模型(DCMN)作为本实验的基线模型，本文采用深度学习框架PyTorch对相关内容进行编码实现，并在Ubuntu系统上采用GPU进行模型的训练和测试。本文采用12层的预训练模型BERT作为三元组的编码器，其中batch_size为16，gradient_accumulation_steps设置为2，训练的epochs的值是10，learning_rate为1e-05，优化函数采用Adam。

(c) 2022 中国中文信息学会计算语言学专业委员会

4.3 实验结果及分析

4.3.1 不同数据增强方法与基线模型的实验结果比较

本文在RACE数据集上进行了数据增强实验，将DCMN基线模型和四种不同的数据增强方法进行比较。在此实验中，回译数据增强选用的方法是英语-德语-英语的方式，词向量替换选用词向量Word2Vec。而生成问题的数据增强方法，使用unilm生成模型，以RACE数据集中文章截断部分作为生成模型输入部分的文章，将正确答案在此截断文章中的数据提取出来，正确答案所在片段截取出来作为生成模型输入部分的答案，这两者一起来生成问题。对于选项中的错误选项部分，本文采取与正确答案相似的处理，将错误选项所在文章片段提取出来作为迷惑选项。而我们的方法选择的是用SimCSE相似度计算方法来选择与问题选项相关TOP2的句子，并且其中R-Drop的α值设置为1。

MODEL	RACE
DCMN	65.05
DCMN+回译	63.99
DCMN+词向量替换	64.56
DCMN+生成问题	66.24
OURS	68.84

Table 2: 不同的数据增强方法实验结果对比

如表2结果所示，DCMN基线模型结果为65.05%，相比于DCMN的基线模型，对多项选择的机器阅读理解的长文本文章进行截断处理，对截断部分使用回译和词向量替换的数据增强方法，其准确率反而下降。原因之一是预训练模型本身强大，对模型中的文章部分进行数据增强学到的新内容有限，而且这两种数据增强的方法学习到了很多与问题无关的内容。原因之二是回译与词向量替换的数据增强方法有过多的替换，无法保留解决问题的关键词，因此影响在文章中找到问题的答案。基于生成问题的数据增强方法所得到的效果相比基线模型有所提升，是由于挑选的是能够在截断部分找到答案的数据来进行数据增强，因此生成的新数据中问题也能找到答案，是有效数据。而我们的数据增强方法使模型学习了截断之外的文章内容，其次，我们的方法并没有改变原始的文章内容，即保留了指示答案位置关键词。从实验结果来看，对于长文本的数据增强方法来说，我们的数据增强方法更能提升模型的准确性。

为了能够更加直接的感受各种数据增强方法之间的不同，我们将基线模型与回译、数据增强以及本文对文章的不同选择来进行对比。如图3所示，基线模型采取的是截断方式，基线模型中的文章部分为画线部分。而回译和词向量替换是基于基线模型的文章进行的。而本文数据增强的方法提取的句子为红色字体部分。可以看出，当相对应的问题是涉及到截断部分之外的内容，基线模型和其他两种数据增强方法都无法找到对应的答案，会造成冗余数据。

4.3.2 文本相似度对比实验

4.3.3 消融实验

为了进一步分析本文的数据增强方法中各部分的作用，研究利用文本相似度进行的句子选择数据增强模块和R-Drop模块各自的作用，本文设置了保留其中一个模块，去掉另一个模块的消融实验的对比。

MODEL	ACC
DCMN	65.05
DCMN_SENAUG	67.34
DCMN_RDROP	67.23
DCMN_SENAUG_RDROP	68.86

Table 3: 消融实验比较

表3中DCMN表示数据增强模块和R-Drop模块都去除。DCMN_SENAUG表示只保留数据增强模块，去除R-Drop模块。DCMN_RDROP表示保留R-Drop模块，去除数据增强模块。表

(c) 2022 中国中文信息学会计算语言学专业委员会

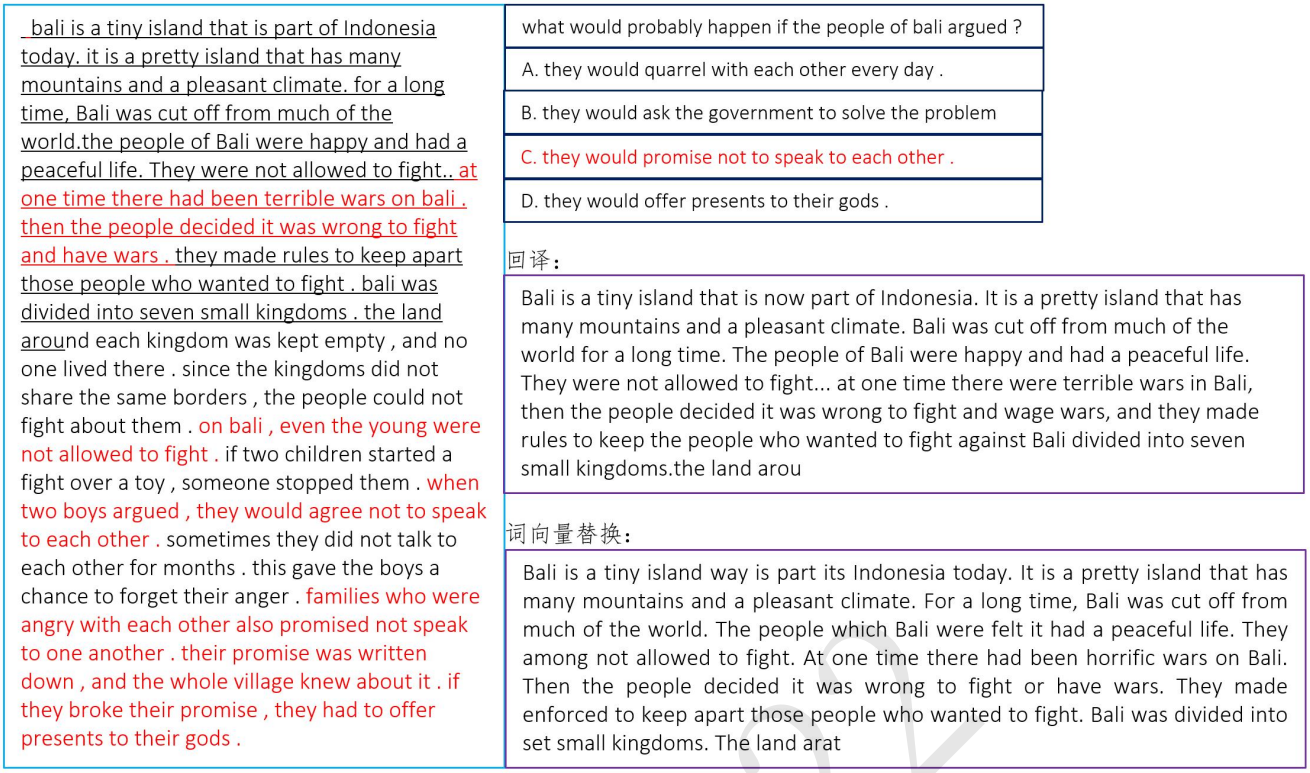


Figure 3: 不同数据增强的文章部分

中最后一行表示两个模块都保留。如表3实验结果所示，基于句子选择的部分做数据增强的方法在DCMN基线模型上提高了2.3%，而R-Drop模块在基线模型上提高了近2.2%，这表明这两部分都使模型得到了正确的数据训练，都是有效的，而最后组合在一起的结果比在R-Drop的基础上得到的结果提高了1.63%，说明了两者学习到的内容并不是完全一样的，都是分别有效果的。

4.3.4 文本相似度对比实验

本文选择通过数据增强的方式提高模型的准确率，为了证明通过文本相似度的方式选择的句子质量对实验结果的影响。我们选择了两种文本相似度方法进行句子的挑选。为了比较余弦相似度和simcse相似度挑选句子的效果，本文设置了两者的两组比较值，分别是在不同K值下和两种相似度方法在RACE数据集及其子集上的对比实验。为了使其不受R-Drop的影响，通过设置α=0使得R-Drop模块不起作用。

如图4结果所示，两种余弦相似度的方法做数据增强得到的结果趋势都是先上升然后下降迅速，再缓慢上升。这种准确率振荡的原因有以下几点：第一是由于本文采用的是对问题和每个选项与文章的相似度赋予同等权重，并且取各自相似度的Top k按照相对顺序进行组合，SimCSE相似度在k为2时取得最大值，表明SimCSE相似度能够在各自挑选2个句子的时候找到最关键的句子，后面k增加1急速下降说明各自新增加的句子具有干扰作用，学习到了无关的句子。而后慢慢上升表示挑选出的句子增多时，由于句子内容逐渐丰富，能逐渐消除带来的歧义。因为SimCSE相似度更敏锐所以相对余弦相似度曲线表现得更明显。使用SimCSE相似度得到的结果比使用余弦相似度得到的结果效果更好，表明不同相似度挑选出的句子是有很大区别的。挑选出更相关的句子进行训练对模型提升准确度更有帮助。

4.3.5 R-Drop中α值的选择

为了比较R-Dropα值的效果，我们设置了四组实验，分别设置为α为1到4，α等于1时得到的结果是最好的。α是KL散度损失函数的权重系数，值越大表示越重要，在α为1时取得最好的效果，说明我们不需要太关注KL散度。只需要给一点KL散度正则化就可以。

(c) 2022 中国中文信息学会计算语言学专业委员会

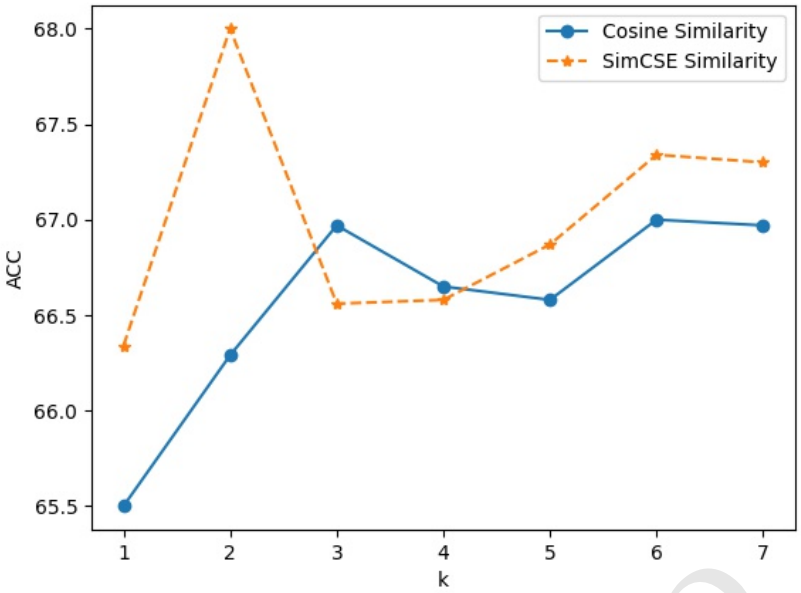


Figure 4: 两种相似度方法在k值不同时的准确率

α	**1**	**2**	**3**	**4**
ACC	68.86	68.59	68.43	68.08

Table 4: α值的选择

4.3.6 长文本文章长度影响

为了探索不同文章长度进行基于句子选择数据增强方法对于结果的影响，我们将以4.1数据集的划分方式将RACE_HIGH数据集分为四段分别进行数据增强。比较不同长度段之间的效果。

passage_len	**ACC**
DCMN	61.69
500-1000	62.41
1000-1500	63.22
1500-2000	63.66
>2000	63.14

Table 5: RACE_HIGH数据集不同长度区间进行数据增强后的结果比较

从表5结果可以看出，这四个区间段进行分别的数据增强都是有效果的，但是效果是有些微差别的。原因一是RACE_HIGH数据集每个区间段数量相差较大，500-1000区间仅有2028条数据，而1000-1500为11422条，1500-2000数据量最多，达到了31481，而大于2000以上的有17410条，数量差别影响了带来效果提升的区别。原因二是不同长度的难度不一样，这也会导致结果的差异。

5 总结

本文通过对多项选择的机器阅读理解进行简单有效的数据增强，比较两种相似度的方法来提取长文本文章中对问题和选项有用的句子来进行数据增强，又使用了R-Drop来消除训练和推理时的差异。实验结果表明使用不同的相似度方法挑选出的句子对提高模型效果是不一样的，

如何能够找到完全切合问题的句子有待进一步探究，甚至可以不再拘泥于挑选整个句子，而是将挑选的句子关键字进行整合，形成新的句子再组成新的文章进行数据增强。

参考文献

Coulombe C. 2018. *Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs.* Prentice-Hall, Englewood Cliffs, NJ.

Dehouck M and Gómez-Rodríguez C. 2020. *Data augmentation via subtree swapping for dependency parsing of low-resource languages.* Proceedings of the 28th International Conference on Computational Linguistics. 2020: 3818-3830.

DEVLIN J, CHANG M-W, LEE K, et al. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Computation and Language (cs.CL).

Fabbri A R, Han S, Li H, et al. 2021. *Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation.* Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 704-717.

Gao T, Yao X, Chen D. 2021. *SimCSE: Simple Contrastive Learning of Sentence Embeddings.* Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 6894-6910.

Hermann K M , Tomávs Kovcisk, Grefenstette E , et al. 2015. *Teaching Machines to Read and Comprehend.* NIPS.

Jiao X , Y Yin, Shang L , et al. 2020. *TinyBERT: Distilling BERT for Natural Language Understanding.* Findings of the Association for Computational Linguistics: EMNLP 2020.

JM Tapia-Téllez, Escalante H J . 2020. *Data Augmentation with Transformers for Text Classification.*

Lai G, Xie Q, Liu H, et al. 2017. *RACE: Large-scale ReAding Comprehension Dataset From Examinations.* Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 785-794.

Lee S, Lee D B, Hwang S J. 2021. *Contrastive Learning with Adversarial Perturbations for Conditional Text Generation.* Ninth International Conference on Learning Representation, ICLR 2021. The International Conference on Learning Representations.

Liu S , Zhang X , Zhang S , et al. 2019. *Neural Machine Reading Comprehension: Methods and Trends.* Applied Sciences.

Mikolov T, Sutskever I, Chen K, et al. 2013. *Distributed representations of words and phrases and their compositionality.* Advances in neural information processing systems, 2013, 26.

Mirshekari M, Gu J, Sisto A. 2021. *ConQuest: Contextual Question Paraphrasing through Answer-Aware Synthetic Question Generation.* Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). 2021: 222-229.

Mitkov R. 2003. *Computer-aided generation of multiple-choice tests.* Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing. 2003: 17-22.

Seo M , Kembhavi A , Farhadi A , et al. 2016. *Bidirectional Attention Flow for Machine Comprehension.*

Smith E , Greco N , Bosnjak M , et al. 2015. *A Strong Lexical Matching Method for the Machine Comprehension Test.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Tang M , Cai J , Zhuo H H . 2019. *ulti-Matching Network for Multiple Choice Reading Comprehension.* Proceedings of the AAAI Conference on Artificial Intelligence. 33:7088-7095.

Wei J, Zou K. 2019. *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks.* 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: 6382-6388..

(c) 2022 中国中文信息学会计算语言学专业委员会

Wu L, Li J, Wang Y, et al. 2021. *R-drop: regularized dropout for neural networks.* Advances in Neural Information Processing Systems.

Xie Q , Dai Z , Hovy E , et al. 2019. *Unsupervised Data Augmentation for Consistency Training.*

Yu A W, Dohan D, Luong M T, et al. 2018. *QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension.* nternational Conference on Learning Representations.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. *Character-level convolutional networks for text classification.* Advances in neural information processing systems 28 (2015): 649-657.

Zhu H, Wei F, Qin B, et al. 2018. *Hierarchical attention flow for multiple-choice reading comprehension.* Proceedings of the AAAI Conference on Artificial Intelligence. 32(1).

Zhang S , Zhao H , Wu Y , et al. 2020. *DCMN+: Dual Co-Matching Network for Multi-Choice Reading Comprehension.* Proceedings of the AAAI Conference on Artificial Intelligence 34(5):9563-9570.

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

一种非结构化数据表征增强的术后风险预测模型*

王亚强[1,2,3†]，杨潇[1,2,3]，郝学超[4]，舒红平[1,3]，陈果[4†]，朱涛[4†]
[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]四川大学华西医院麻醉手术中心
[†]通讯作者：yaqwang@cuit.edu.cn, grace_chenguo@hotmail.com, 739501155@qq.com

摘要

准确的术后风险预测对临床资源规划和应急方案准备以及降低患者的术后风险和死亡率具有积极作用。术后风险预测目前主要基于术前和术中的患者基本信息、实验室检查、生命体征等结构化数据，而蕴含丰富语义信息的非结构化术前诊断的价值还有待验证。针对该问题，本文提出一种非结构化数据表征增强的术后风险预测模型，利用自注意力机制，精巧的将结构化数据与术前诊断数据进行信息加权融合。基于临床数据，将本文方法与术后风险预测常用的统计机器学习模型以及最新的深度神经网络进行对比，本文方法不仅提升了术后风险预测的性能，同时也为预测模型带来了良好的可解释性。

关键词： 文本数据；术后风险预测；自注意力机制；信息融合

An Unstructured Data Representation Enhanced Model for Postoperative Risk Prediction

Yaqiang Wang[1,2,3†], Xiao Yang[1,2,3], Xuechao Hao[4], Hongping Shu[1,3], Guo Chen[4†], Tao Zhu[4†]
[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 Anesthesiology, Sichuan University
[†]Corresponding author: yaqwang@cuit.edu.cn, grace_chenguo@hotmail.com, 739501155@qq.com

Abstract

Postoperative risk prediction has a positive effect on clinical resource planning, emergency plan preparation and reducing postoperative risk and mortality. Postoperative risk prediction is mainly based on patient's basic information, laboratory tests, vital signs and other structured data, while the value of unstructured preoperative diagnosis with rich semantic information remains to be verified. Aiming at attempting this problem, an unstructured data representation enhanced postoperative risk prediction model is proposed in this paper. The model utilizes self-attention to fuse structured data with preoperative diagnosis. Through comparing with the commonly used statistical machine learning models and the state-of-the-art deep neural networks, the proposed model has not only better prediction performance, but also better interpretability.

Keywords: Text Data , Postoperative Risk Prediction , Self-Attention Mechanism , Information Fusion

*四川大学华西医院1·3·5项目(ZYJC21008)和国家重点研发计划项目(2018YFC2001800)资助
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

由于术后并发症（如肺部并发症(魏娟 et al., 2021)、心血管不良(Lundberg et al., 2018)、ICU入室(Chiew et al., 2020)等）风险（简称术后风险）所导致的术后30天内死亡，已成为全球排名第三位的人群死亡原因(Li et al., 2022)。准确的术后风险预测对医生进行合理的临床资源规划、应急方案准备具有重要的辅助作用，对降低患者的术后风险和死亡率具有积极意义(Xue et al., 2021; Bonde et al., 2021)。

目前，术后风险预测主要基于术前和术中的患者基本信息（如体温、血压、体重等）、实验室检查（如氧分压、氧饱和、蛋白等）、生命体征（如手术时长、出血量等）等结构化数据，利用XGBoost、逻辑回归、随机森林、人工神经网络等模型实现(Lundberg et al., 2018; Xue et al., 2021)。

近年来深度神经网络在各领域的预测任务中表现优秀，受到研究者的广泛关注，也被引入到术后风险预测任务中(Bonde et al., 2021)。Fritz(2019)等人构建了一种多路径卷积神经网络，提取和融合患者基本信息、共病情况、术前实验室检查和术中生命体征等结构化数据中的特征，用于患者术后死亡风险预测。Barbieri(2020)等人利用双向门控循环单元，将结构化数据之间的时间信息以拼接的方式融入数据表征，采用注意力机制提取重要特征，用于患者术后ICU入室风险预测。现有方法的核心是如何将结构化数据中的离散型和连续型特征向量化，形成输入基于深度神经网络的术后风险预测模型的数据表征。

图 1. 包含患者基本信息、实验室检查和术前诊断的术前数据示例

术前数据中除结构化数据之外，还包含语义丰富的非结构化术前诊断数据。术前诊断中不仅包含医生基于医学知识，根据局部的结构化数据，对患者病情的总结信息，还包含医生以整体的结构化数据为依据，利用经验知识，对患者病况的推断信息。如图1中患者1的术前数据所示，根据结构化数据收缩压156mmHg（毫米汞柱）与舒张压76mmHg，基于医学知识，成人的收缩压和舒张压正常范围分别在90至120mmHg，因此，医生在术前诊断中总结该患者有“高血压病”，且属于“3级很高危”。此外，依据目前患者整体的结构化数据，医生根据经验知识，推断患者是“肺部感染”。更进一步地，术前诊断的整体描述，反应了当前患者的全局状态。这些语义信息能够丰富术后风险预测的特征，有助于增强预测模型的性能。

然而，术前诊断数据尚未在术后风险预测任务中被有效利用，如何充分地利用非结构化的术前诊断数据，形成有效的术后风险预测数据表征，尚有待进一步探索。因此，本文围绕非结构化的术前诊断数据如何增强术后风险预测任务展开研究，主要的贡献包括以下三个方面：

1. 与围术期医学专家合作，经过清洗、处理、转换和去隐私过程后，构建了一份包含12240个实例、面向术后风险预测任务的数据集。该数据集的结构化数据部分包含了95列离散型变量、61列连续型变量、一列非结构化的术前诊断变量，以及三列二元的术后风险标签变量，分别表示肺部并发症、心血管不良和ICU入室风险的发生情况。

2. 为充分地利用非结构化的术前诊断数据，本文提出一种非结构化数据表征增强的术后风险预测模型，利用自注意力机制，将结构化数据与局部的细粒度实体信息及全局的粗粒度文本语义加权融合，有效地将非结构化数据用于增强术后风险预测性能。

3. 本文提出的基于自注意力机制融合结构化与非结构化数据的模型结构，为术后风险预测带来了良好的可解释性。细节实验结果分析发现，利用自注意力机制获得的关系权重矩阵，可以解释和展示出非结构化数据不仅增强了重要的结构化数据的贡献度，还补充了风险预测信息。

实验结果表明，本文提出的非结构化数据表征增强的术后风险预测模型明显优于所对比的常用统计机器学习模型和最新的深度神经网络，在三种重要的术后风险预测任务上，本文

提出的模型均取得了最优的结果，F1-Score分别达到了0.669、0.558和0.608。此外，通过消融实验，进一步验证了本文提出的模型有效地加权融合了局部的细粒度实体信息和全局的粗粒度文本语义信息。利用非结构化术前诊断数据表征增强术后风险预测模型后，肺部并发症风险预测性能提升了6.878%，心血管不良风险预测性能提升了9.541%，ICU入室风险预性能提升了7.641%。

2 相关工作

术后风险预测是医学信息学领域的研究热点问题，目前的研究主要集中于机器学习模型应用于术后风险预测的有效性验证及面向特定类型的术后并发症风险特征分析两个层面。Canet(2010)等人利用逻辑回归模型，确定了7个独立且具有良好鉴别能力的危险因素，构建了术后肺部并发症风险预测指标，用于评估和预测术后肺部并发症的个体风险。Hill(2019)等人采用随机森林模型，自动地发现重要的术前特征，将结构化的美国麻醉医师协会身体状况特征与术前特征结合，提升术后死亡风险的预测性能。与先前工作不同，本文提出了一种非结构化数据表征增强的术后风险预测模型，该模型基于自注意力机制，在预测中有效地融合结构化数据和非结构化语义信息，并提供可解释性。

术后风险预测目前的主要研究对象是术前和术中的结构化数据，其中包含两种类型的变量，一种是离散型变量，另一种是连续型变量。通常会将连续型变量进行离散化后，与离散型变量一起共同构建特征向量，作为术后风险预测模型的输入(Arik and Pfister, 2021)。本文实验主要基于术前患者基本信息和实验室检查等结构化数据，采用与先前工作相同的连续型变量的基本处理方法。不同的是，借鉴Fritz(2019)等人的思想，本文将离散型变量和离散化的连续型变量一并构建离散特征词典，并基于深度神经网络学习离散特征的嵌入表征。

术后风险预测除可利用术前和术中的结构化数据作为特征之外，通过观察发现，包含医学语义信息的非结构化术前诊断数据可用于增强术后风险预测。Zhang(2020)等人提出将英文临床文本利用Doc2Vec模型(Le et al., 2014)直接形成数据表征，然后与结构化数据合并的方式，将非结构化数据与结构化数据融合，应用于住院死亡率、住院时间长短和术后30天再入院的预测任务，该方法在英文临床数据MIMIC-III(Johnson et al., 2016)上进行了实验验证。然而，与该工作不同，本文首次探索了将中文非结构化临床文本引入术后风险预测的方法。

此外，本文通过观察发现，非结构化的术前诊断中，既包含全局的粗粒度文本语义信息，还包含局部的细粒度实体信息，它们均可为术后风险预测提供医学语义特征（如图1所示）。为将这些信息与离散特征的嵌入表征相融合，本文首先基于目前常用的中文MedBERT[1]获得实体的嵌入表征，并将术前诊断视为句子后，采用词嵌入平均池化的方法将其向量化。然后利用自注意力机制(Bahdanau et al., 2014)，将离散特征的嵌入表征与实体的嵌入表征以及向量化的术前诊断进行加权融合，在充分地综合利用全局和局部的文本语义信息的基础上，还为模型带来良好的可解释性(Hao et al., 2021)。

3 术后风险预测

3.1 任务定义

本文将术后风险预测定义为一项二分类任务，采用有监督学习方法解决。定义$(\boldsymbol{x}, y)$为一个训练实例，$\boldsymbol{x}$中包含$\boldsymbol{x}_{num}$、$\boldsymbol{x}_{cat}$和$\boldsymbol{x}_{PD}$三种特征数据。其中，$\boldsymbol{x}_{num}$表示表格数据中的连续型特征数据，共m列，$\boldsymbol{x}_{cat}$表示表格数据中的离散型特征数据，共n列，$\boldsymbol{x}_{PD}$表示非结构化术前诊断文本数据，y表示术后风险的发生情况，用1或0分别表示风险的发生或未发生。

3.2 表格数据的向量表示

本文提出的术后风险预测模型主要利用结构化表格数据和非结构化术前诊断文本数据对术后风险进行预测（如图2所示）。结构化表格数据由$\boldsymbol{x}_{num}$和$\boldsymbol{x}_{cat}$组成。本文采用分类与回归树算法(Loh, 2011)，先将连续型特征转换为离散型特征，在引入医学语义信息的同时，降低数据的复杂度。转换之后的连续型特征不仅能够表达医学语义信息，还被统一成离散型特征。转换后的连续型变量表征被定义为$\boldsymbol{x}_{n2cat}$：

$$\boldsymbol{x}_{n2cat} = discretized(\boldsymbol{x}_{num}) \tag{1}$$

[1]https://code.ihub.org.cn/projects/1775

(c) 2022 中国中文信息学会计算语言学专业委员会

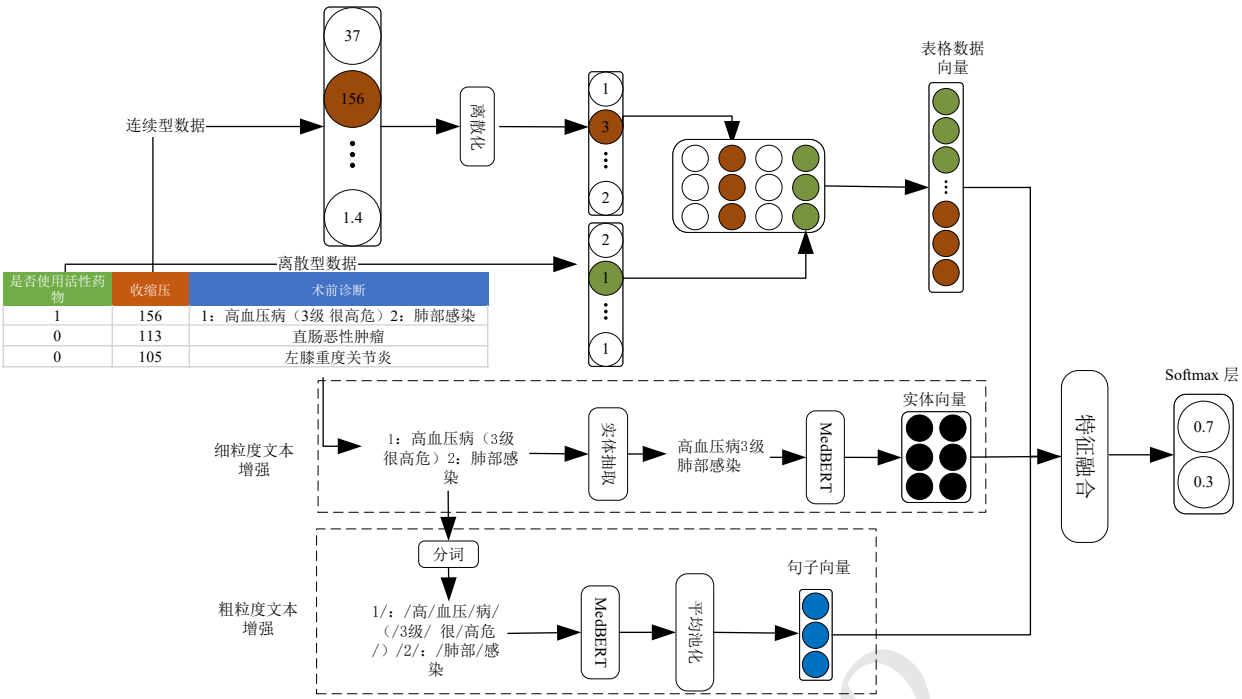


图 2. 模型结构图

处理离散型变量表征的常用方式是采用实体嵌入(Guo and Berkhahn, 2016)的方式，即为每一个离散型变量构建一个特征词表，词表大小为当前离散型变量的不同取值的数量。然而该方法在建模的过程中仅考虑了单一变量下的不同取值之间的语义关联性，而不同的变量之间的相关性未被考虑其中。为引入全局不同变量之间的语义关联性，本文改进了原始的实体嵌入方式，让所有的离散型变量共用相同的特征词表。每一个离散型变量（包括$\boldsymbol{x}_{cat}$和$\boldsymbol{x}_{n2cat}$）的不同取值，都会赋予唯一的索引值x_i，其中$i \in [0, |V|]$，$|V|$是所有离散型变量的所有不同取值的数量总和，即共用的特征词表的词表大小。每个x_i都将通过学习过程被映射为一个维度为d的向量，定义为$\boldsymbol{e}_{tabular}$，其中d为超参数。通过构建全局共用的特征词表，原始的离散型变量转换为语义向量之后，不仅扩充了语义信息，而且不同离散型变量之间也产生了语义关联，相比原始的实体嵌入方式，解决了不同离散型变量之间语义关联缺失的问题。最后，将所有的$\boldsymbol{e}_{tabular}$拼接形成表格数据的向量表示$\boldsymbol{E}_{tabular}$。

3.3 文本数据的向量表示方法

术前诊断文本$\boldsymbol{x}_{PD}$主要包含医生总结的病人身体症状和初步推断的病情描述，两者可统一定义为病症实体。因此，术前诊断文本可以归纳为由多个病症实体、连接词以及标点符号构成的集合，每个实例$\boldsymbol{x}_{PD}$包含l_{max}项的病症实体，l_{max}表示数据集中，$\boldsymbol{x}_{PD}$最多可饱含的病症实体数量。

术前诊断文本可以有两种向量表示方式，一种是形如利用Doc2Vec模型(Le et al., 2014)得到的全局语义向量，获取该类向量表示的方法我们称之为粗粒度文本的向量表示方法；另一种是直接将病症实体对应的语义向量拼接，形成细粒度文本的向量表示方法。后文将具体介绍两种方式获取术前诊断文本粗粒度语义信息和细粒度语义信息的方法。

3.3.1 粗粒度文本的向量表示方法

为获取术前诊断文本的粗粒度语义向量，先将文本进行分词[1]，得到分词列表$\{token_0, token_1, ..., token_p\}$，其中$p$表示文本分词后得到的词的数量。将分词列表输入领域微调后的预训练模型MedBERT中，生成维度为768的动态词向量列表$\{\boldsymbol{e}_0^{768}, \boldsymbol{e}_1^{768}, ..., \boldsymbol{e}_p^{768}\}$，768是MedBERT的词向量维度。为进一步获取句子向量，选择采用快速且高效的平均池化方法整合词向量的语义信息。对词向量矩阵中的每一列取均值，将词向

[1]本文实验中直接采用了https://huggingface.co/hfl/chinese-macbert-base的内置分词工具

(c) 2022 中国中文信息学会计算语言学专业委员会

量矩阵压缩为包含整个术前诊断语义信息的粗力度文本的向量表示$\boldsymbol{e}_{sentence}$。

$$\boldsymbol{e}_{sentence} = MeanPooling(\{\boldsymbol{e}_0^{768}, \boldsymbol{e}_1^{768}, ..., \boldsymbol{e}_p^{768}\}) \tag{2}$$

3.3.2 细粒度文本的向量表示方法

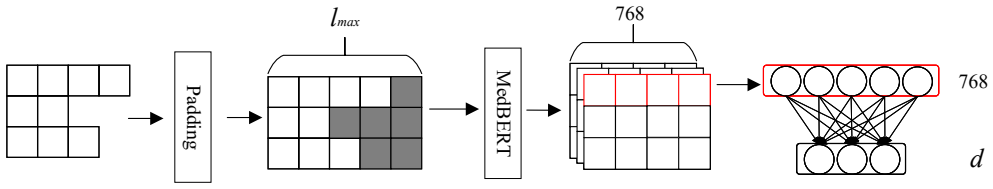


图 3. 细粒度文本增强

将术前诊断文本分词后通过MedBERT生成的词向量简单压缩为一个向量的过程中可能会造成局部语义信息的丢失，且无法明确术前诊断文本中哪些信息在术后风险预测过程中起到了关键作用。为保留术前诊断文本的局部细粒度实体语义信息，首先利用医学领域数据集基于BERT+BiLSTM+CRF模型训练得到实体抽取模型(Dai et al., 2019)，然后利用该模型抽取$\boldsymbol{x}_{PD}$中的病症实体，形成病症实体集合$\{w_0, ..., w_k, ..., w_K\}$，其中，$K$表示当前$\boldsymbol{x}_{PD}$中抽取得到的病症实体数量。

因$\boldsymbol{x}_{PD}$中包含的实体数量可能不一致，为后续处理统一，本文将病症实体数量达到l_{max}的集合，通过补全特殊字符$[PAD]$的方式，形成数量均为l_{max}的实体集合（如图3所示）。然后，每一个实体w_k将通过MedBERT转换为蕴含医学语义的向量$\boldsymbol{e}_k^{768}$。为后续与表格数据的向量表示进行融合，细粒度文本语义向量将进一步通过全连接层降维，从768维降至d维，得到降维后的细粒度语义向量集合$\{\boldsymbol{e}_0^d, \boldsymbol{e}_1^d, ..., \boldsymbol{e}_{l_{max}}^d\}$。最后，将含有全局语义信息的粗粒度文本的向量表示和含有局部语义信息的细粒度文本的向量表示组合，得到最终的文本数据的向量表示$\boldsymbol{E}_{text}$。

$$\boldsymbol{E}_{text} = \{\boldsymbol{e}_0^d, \boldsymbol{e}_1^d, ..., \boldsymbol{e}_{l_{max}}^d, \boldsymbol{e}_{sentence}\} \tag{3}$$

3.4 特征融合

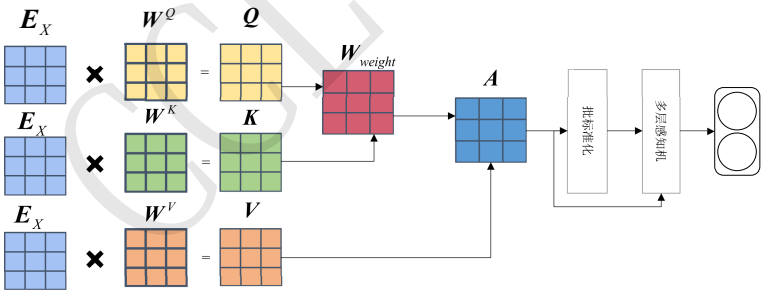


图 4. 特征融合

在特征融合层，本文选择使用Self-Attention(Bahdanau et al., 2014)将表格数据和文本数据进行特征融合（如图4所示）。首先，将表示表格数据信息的特征向量$\boldsymbol{E}_{tabular}$与表示文本语义信息的特征向量$\boldsymbol{E}_{text}$拼接，形成新的特征向量集合$\boldsymbol{E}_X$。并将$\boldsymbol{E}_X$通过三个参数矩阵$\boldsymbol{W}^Q$、$\boldsymbol{W}^K$和$\boldsymbol{W}^V$映射为三个不同的矩阵$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$。然后对$\boldsymbol{Q}$和$\boldsymbol{K}^T$执行点积并利用d_k放缩结果，以保证训练过程中梯度的稳定性，d_k是指矩阵$\boldsymbol{K}$的维度，计算方法如公式(4)$softmax$函数的输入所示。随后执行softmax操作进行归一化，得到不同的特征向量之间（包含文本数据向量和表格数据向量）的注意力权重$\boldsymbol{W}_{weight}$（计算公式如公式(4)所示）。

(c) 2022 中国中文信息学会计算语言学专业委员会

$$\boldsymbol{W}_{weight} = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}) \tag{4}$$

最后将$\boldsymbol{W}_{weight}$与$\boldsymbol{V}$相乘得到增强后的特征表示$\boldsymbol{A}$。具体计算过程如下公式所示。

$$\boldsymbol{E}_X = \boldsymbol{E}_{tabular} \bigoplus \boldsymbol{E}_{text} \tag{5}$$

$$\boldsymbol{Q} = \boldsymbol{E}_X \boldsymbol{W}^Q, \boldsymbol{K} = \boldsymbol{E}_X \boldsymbol{W}^K, \boldsymbol{V} = \boldsymbol{E}_X \boldsymbol{W}^V \tag{6}$$

$$\boldsymbol{A} = Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{W}_{weight} \boldsymbol{V} \tag{7}$$

通过注意力机制，模型可以自动地学习到特征在推理过程中的重要性或贡献度。因此，在模型推理过程中，可以通过提取并分析注意力权重矩阵，来探究在模型预测过程中，各特征发挥作用的重要程度。

为了解决梯度消失的问题，受到(He et al., 2016; Ba et al., 2016)的启发，新的特征矩阵$\boldsymbol{A}$在输入前馈神经网络之前，还经过了残差网络和层标准化操作。接着将向量输入到带有sigmoid激活函数的前馈神经网络中，计算预测术后风险的发生概率P，其中$\boldsymbol{W}$和$\boldsymbol{b}$均是前馈神经网络将学习得到的参数。

$$P = sigmoid(\boldsymbol{W}^T \boldsymbol{A} + \boldsymbol{b}) \tag{8}$$

最后，模型的损失定义为：

$$loss = -\frac{1}{M}\sum_{i=0}^{M}(y_i log P_i + (1-y_i) log(1-P_i)) \tag{9}$$

其中M指批量包含的实例数量。

4 实验

4.1 实验数据

本文实验采用了从医院的临床数据系统中获取的数据，其中包含患者的基本身体状况信息、实验室检查数据和术前诊断，以及病人术后发生的肺部并发症、心血管不良和ICU入室结局。该数据经过了如下基本的预处理过程：

- 删除了有关患者身份的个人信息。
- 删除了缺失率高于50%的变量。

最终得到包含12240个实例的术后风险预测任务数据集，包含有95项离散型变量和61项连续型变量以及1项术前诊断变量。该数据集包含的三种术后风险的标签分布如图5所示，阳性率分别是15.93%、6.25%和3.02%。实验将数据集按照7:1:2的比例划分得到训练集、验证集和测试集。

4.2 评估指标

为了评估模型的效果，本文采用精确率（Precision）、召回率（Recall）和F1-Score作为主要的评估指标。计算公式如下：

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

第二十一届中国计算语言学大会论文集，第580页-第590页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

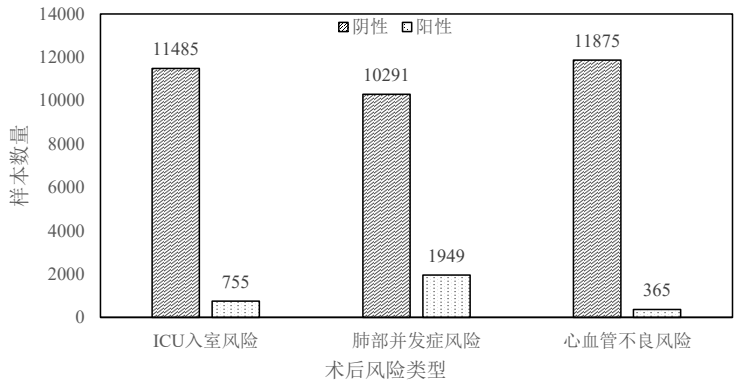


图 5. 数据集标签分布

$$F1-Score=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{12}$$

其中，*TP*表示在阳性实例中，模型预测为阳性的实例数量，*FP*表示在阴性实例中，模型预测为阳性的实例数量，*FN*表示在阳性实例中，模型预测为阴性的实例数量。

4.3 参数设置

模型训练采用了Adam优化器，初始学习率设置为3e-4，$\beta_1=0.9$，$\beta_2=0.999$，批量大小设置为128，训练轮次设置为100，超参数d设置为32，dropout参数设置为0.5。在以上超参数设置条件下，本文提出的术后风险预测模型达到了收敛。

为验证本文提出的模型在术后风险预测任务上的有效性，选择了两种常见的统计机器学习模型LR和XGBoost以及两种用于表格数据的最新的深度神经网络Wide&Deep(Cheng et al., 2016)和Tabtransformer(Huang et al., 2020)作为对比模型。LR和XGBoost采用scikit-learn框架(Pedregosa et al., 2011)实现，Wide&Deep和Tabtransformer分别采用开源的代码库实现[1]。

4.4 实验分析

首先在三项术后风险预测任务上对比了模型的预测性能，实验结果如表1所示。

Model	**肺部并发症风险**			**ICU 入室风险**			**心血管不良风险**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
LR	71.508	31.068	43.316	72.840	36.875	48.963	45.833	13.415	20.755
XGBoost	69.965	48.293	57.143	74.118	39.375	51.429	41.176	8.537	14.141
WideDeep	73.214	54.089	62.215	74.603	37.600	50.000	52.778	29.788	37.255
Tabtransformer	68.563	60.422	64.236	75.385	39.200	51.579	55.556	30.303	39.216
Our	68.378	65.723	**66.909**	65.088	57.664	**60.833**	77.395	44.260	**55.888**

表 1. 实验整体结果

从表1中可以观察得到，Wide&Deep和Tabtransformer在三项术后风险的预测任务上均优于LR和XGBoost，特别是在阳性率较低的心血管不良风险预测任务上，Wide&Deep和Tabtransformer的表现远优于LR和XGBoost。该结果说明，深度神经网络在术后风险预测任务上的性能优于统计机器学习模型，这与(Bonde et al., 2021; Fritz et al., 2019)报告的结果保持一致。

此外，从表1还可以看出，通过引入了术前诊断文本数据表征，本文提出的模型在肺部并发症、心血管不良和ICU入室三个风险预测任务上均取得了最优的效果，F1-Score分别达到了66.909%、55.888%和60.833%。该结果证明，本文提出的文本数据表征增强的术后风险预测模型是有效的。

进一步观察表1中的结果发现，相比于其它模型，本文提出的模型是在保持了良好的精确率的条件下，召回率得到了大幅提升，从而提升F1-Score结果。该结果说明，当模型预测引入

[1]https://github.com/jrzaurin/pytorch-widedeep/tree/pytorch_widedeep

(c) 2022 中国中文信息学会计算语言学专业委员会

术前诊断数据表征后，进一步丰富了特征的语义信息，对阳性实例的预测带来了额外的语义信息补充，从而帮助模型将之前无法判断的阳性实例准确地预测为阳性，进而提高了模型的召回率。

4.5 消融实验

为进一步验证文本数据表征对模型预测效果增强的作用，并探究文本的粗粒度文本语义信息和细粒度文本语义信息对预测任务的影响，本文还设计了不加入文本以及分别加入粗粒度和细粒度的文本语义信息的对比消融实验，结果如表2所示，其中-E是指去除细粒度文本的语义向量，-S是指去除粗粒度文本的语义向量，-E-S是指去除所有的文本数据。

Model	肺部并发症风险			ICU 入室风险			心血管不良风险		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Our-E-S	60.842	59.444	60.031	53.878	52.598	53.192	52.854	42.687	46.347
Our-E	62.547	62.394	62.754	56.365	53.670	54.913	53.551	42.663	46.905
Our-S	68.089	66.010	66.883	61.129	58.152	59.570	79.697	43.029	55.577
Our	68.378	65.723	**66.909**	65.088	57.664	**60.833**	77.395	44.260	**55.888**

表 2. 消融实验结果

观察表2发现，在加入文本数据后，本文提出的模型在肺部并发症风险预测任务上的性能提高了6.878%，在ICU入室风险预测中提高了7.641%，在心血管不良风险预测中提高了9.541%，并且无论是单独加入粗粒度文本的语义向量还是细粒度文本的语义向量，模型的预测性能均得到明显改善。该结果说明，非结构化术前诊断中的信息对术后风险预测具有积极的作用，为术后风险预测提供了额外的决策信息，有效地增强了模型的预测能力。

此外，观察表2还可以发现，阳性率越低的术后风险，通过引入非结构化术前诊断数据表征后，模型的预测性能提升越高。该结果说明，对于阳性实例更少的术后风险，模型需要更多的特征才能更准确地预测阳性病例，术前诊断的引入能够为模型带来更丰富的特征，从而使得本文提出的模型在阳性率越低的术后风险预测中表现得越出色。

从表2结果还能够看出，相比于全局的粗粒度文本语义向量的缺失，模型对于局部的细粒度文本语义向量的缺失更加敏感，该结果说明，在术后风险预测的过程中引入围术期医学领域知识，对模型的预测性能提升具有重要的作用。这也进一步说明了本文提出的非结构化数据表征增强的术后风险预测模型的有效性和应用价值。

更进一步地，从表2还可以看出，当模型同时引入粗粒度文本的向量表示和细粒度文本的向量表示时，模型的预测性能达到最优。该结果说明，在非结构化数据表征增强术后风险预测模型时，既需要引入粗粒度文本向量携带的全局语义信息，又需要引入细粒度文本向量携带的局部语义信息。

4.6 细节分析

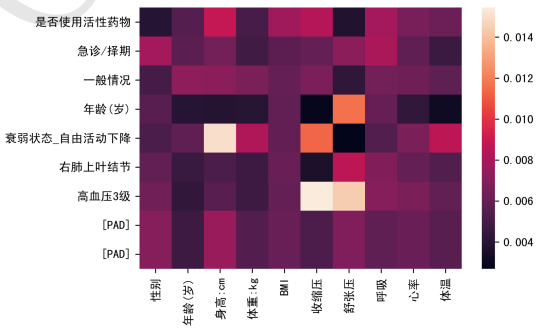


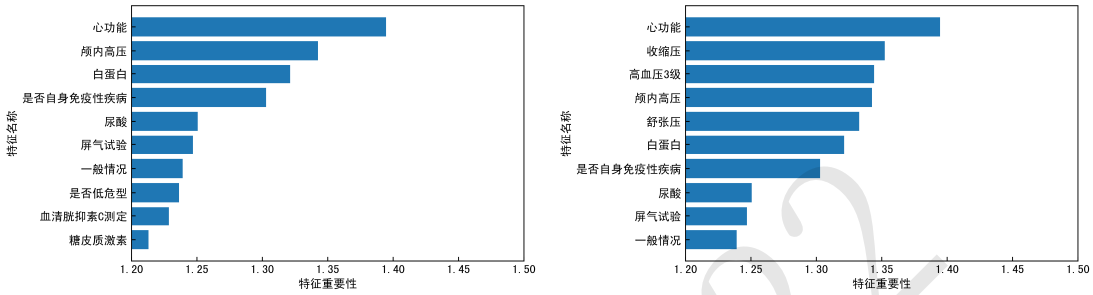
图 6. 心血管不良实例的注意力权重热力图

本文提出的模型通过自注意力机制为术后风险预测模型带来了可解释性，为验证和说明该效果，本文选取了一个发生了术后心血管不良的病人实例，该实例的术前诊断是“右肺上叶结节，高血压3级”。本文提出的模型准确地预测得到该病人实例将发生术后心血管不良风险。通

过提取模型中的注意力权重矩阵$\boldsymbol{W}_{weight}$，画出热力图，见图6。在图6中，纵轴的中“右肺上叶结节”和“高血压3级”是指术前诊断中的实体病症，“PAD”指补全的字符，其余的特征均是表格数据的部分特征，横轴也是部分表格数据的特征。

从图6中可以看到，在术前诊断描述中，“高血压3级”显著地与表格数据中的收缩压和舒张压变量具有强关联。该强关联预示着模型通过训练，学习到了数据集中包含的医学领域知识关联信息，该关联信息保存在了$\boldsymbol{W}_{weight}$中，在术后风险预测中起到了重要的预示作用。另一个方面，该结果还说明，利用自注意力机制为术后风险预测模型还带来了可解释性。总体地，实验结果验证了本文提出的模型在增强术后风险预测性能方面的鲁棒性。

通过对比引入非结构化数据表征前后，$\boldsymbol{W}_{weight}$中包含的权重值按列求和后得到每个变量在术后风险预测中的权重比率排序，进一步观察在风险预测中，起重要作用的变量与术后风险结局是否存在医学语义相关性，对比结果如图7所示。



(a) 未引入非结构化数据表征　　(b) 引入非结构化数据表征

图 7. 变量在模型术后风险预测中的重要性排序

在图7中，权重比率越高，说明变量在预测中具有更高的重要性。从结果可以看出，在引入非结构化数据表征后，与术后心血管不良风险强相关的医学变量收缩压与舒张压的权重比率排序更加靠前。该结果进一步证实了本文提出的模型在提升术后风险预测性能的同时，还学习到了医学领域知识，并且具有更好的可解释性。

从图7中还可以看到，在引入非结构化数据表征后，临床医生根据经验总结和/或推断得到的额外的重要医学语义信息，也在风险预测中起到了更重要的作用，例如临床医生基于收缩压和舒张压总结并记录下的“高血压3级”疾病。一方面，该结果进一步证明了，本文提出的模型学习到了医学领域知识，并对术后风险预测起到了积极的作用。另一方面，该结果也说明，本文模型的直觉观察是正确的，术前诊断中包含了大量的医学领域信息，这些信息既包含表格数据中已有的信息，还包含了大量可以用于丰富原始表格数据的额外的医学领域知识信息，这些信息会对模型的预测性能提升起到积极的作用。更进一步地，该结果也说明，本文提出模型在提升了术后风险预测性能的同时，还具有良好的模型可解释性。

5　结束语

术后风险预测在临床医学中具有重要意义，基于表格数据构建统计机器学习模型和深度神经网络，实现术后风险预测是常见的方式。非结构化术前诊断数据中蕴含了大量额外的医学领域知识，可为术后风险预测提供丰富的信息，然而它们尚未被有效利用。针对这个问题，本文提出了一种新的模型，将非结构化数据表征增强术后风险预测，并在模型中引入自注意力机制，在有效融合表格数据和非结构化数据的同时，为模型带来良好的可解释性。实验结果表明，本文提出的非结构化数据表征增强的术后风险预测模型的性能显著高于其他比较的基线模型。通过消融实验，验证了在术后风险预测中引入非结构化术前诊断数据的重要性，证明了本文提出的模型的有效性。此外，通过对模型的注意力权重的细节分析，发现通过自注意力机制，将表格数据与非结构化的术前诊断数据融合到术后风险预测中，为模型带来了良好的模型可解释性。

参考文献

魏娟, 邓惠民, 吕欣. 2021. 术后肺部并发症围手术期风险因素及防治策略. 同济大学学报*(*医学版*)*, 42(06):736-743.

Alexander Bonde, Kartik M. Varadarajan, Nicholas Bonde, Prof Anders Troelsen, Orhun K. Muratoglu, Henrik Malchau, Anthony D. Yang, Prof Hasan Alam, Martin Sillesen. 2021. Assessing the utility of deep neural networks in predicting postoperative surgical complications: a retrospective study. *The Lancet Digital Health*, 3(8):e471-e485.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bing Xue, Dingwen Li, Chenyang Lu, Christopher R. King, Troy Wildes, Michael S. Avidan, Thomas Kannampallil, Joanna Abraham. 2021. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. *JAMA Network Open*, 4(3):e212240-e212240.

Bradley A. Fritz, Zhicheng Cui, Muhan Zhang, Yujie He, Yixin Chen, Alex Kronzer, Arbi Ben Abdallah, Christopher R. King, Michael S. Avidan. 2019. Deep-learning model for predicting 30-day postoperative mortality. *British Journal of Anaesthesia*, 123(5):688-695.

Brian L. Hill, Robert Brown, Eilon Gabel, Nadav Rakocz, Christine Lee, Maxime Cannesson, Pierre Baldi, Loes Olde Loohuis, Ruth Johnson, Brandon Jew, Uri Maoz, Aman Mahajan, Sriram Sankararaman, Ira Hofer, Eran Halperin. 2019. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *British Journal of Anaesthesia*, 123(6):877-886.

Cheng Guo, Felix Berkhahn. 2016. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.

Calvin J. Chiew, Nan Liu, Ting Hway Wong, Yilin E. Sim, Hairil R. Abdullah. 2020. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Annals of Surgery*, 272(6):1133-1139.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825-2830.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, Hemal Shah. 2016. Wide & deep learning for recommender systems. *In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp.7-10.

Yaru Hao, Li Dong, Furu Wei, Ke Xu. 2021. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963-12971.

Jaume Canet, Lluís Gallart, Carmen Gomar, Guillem Paluzie, Jordi Vallès, Jordi Castillo, Sergi Sabaté, Valentín Mazo, Zahara Briones, Joaquín Sanchis. 2010. Prediction of postoperative pulmonary complications in a population-based surgical cohort. *Anesthesiology*, 113(6):1338-1350.

Jimmy Lei Ba, Jamie Ryan Kiros, Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 2016. Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778.

Wei-Yin Loh. 2011. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, 1(1):14-23.

Peiyi Li, Yunmei Luo, Xuexin Yu, Elizabeth Mason, Zhi Zeng, Jin Wen, Weimin Li, Mohammad S. Jalali. 2022. Readiness of Healthcare Providers for e-Hospitals: A Cross-sectional Analysis in China before COVID-19. *BMJ Open*, 12(2):e054169.

Sercan O. Arik, Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679-6687.

(c) 2022 中国中文信息学会计算语言学专业委员会

Sebastiano Barbieri, James Kemp, Oscar Perez-Concha, Sradha Kotwal, Martin Gallagher, Angus Ritchie, Louisa Jorm. 2020. Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-risk. *Scientific Reports*, 10(1):1111.

Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749-760.

Xin Huang, Ashish Khetan, Milan Cvitkovic, Zohar Karnin. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. *In the 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp.1-5.

Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan, and Ping Zhang. 2020. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Medical Informatics and Decision Making*, 20(1):280.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *In Proceedings of the 31st International Conference on Machine Learning*, pp.1188-1196.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

CCL 2022

融合外部语言知识的流式越南语语音识别

王俊强[1,2]，余正涛*[1,2]，董凌[1,2]，高盛祥[1,2]，王文君[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
814220330@qq.com, ztyu@hotmail.com, 46761956@qq.com
gaoshengxiang.yn@foxmail.com, 175360805@qq.com

摘要

越南语为低资源语言，训练语料难以获取；流式端到端模型在训练过程中难以学习到外部大量文本中的语言知识，这些问题在一定程度上都限制了流式越南语语音识别模型的性能。因此，本文以越南语音节作为语言模型和流式越南语语音识别模型的建模单元，提出了一种将预训练越南语语言模型在训练阶段融合到流式语音识别模型的方法。在训练阶段，通过最小化预训练越南语语言模型和解码器的输出计算一个新的损失函数L_{AED-LM}，帮助流式越南语语音识别模型学习一些越南语语言知识从而优化其模型参数；在解码阶段，使用Shallow Fusion或者WFST技术再次融合预训练语言模型进一步提升模型识别率。实验结果表明，在VIVOS数据集上，相比基线模型，在训练阶段融合语言模型可以将流式越南语语音识别模型的词错率提升2.45%；在解码阶段使用Shallow Fusion或WFST再次融合语言模型，还可以将模型词错率分别提升1.35%和4.75%。

关键词： 流式语音识别；越南语；语言模型；预训练；端到端模型

Streaming Vietnamese Speech Recognition Based on Fusing External Vietnamese Language Knowledge

Junqiang Wang[1,2], Zhengtao Yu*[1,2], Ling Dong[1,2], Shengxiang Gao[1,2], Wenjun Wang[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
814220330@qq.com, ztyu@hotmail.com, 46761956@qq.com
gaoshengxiang.yn@foxmail.com, 175360805@qq.com

Abstract

Vietnamese is known as a low-resource language with few available corpora, and for end-to-end speech streaming recognition, it's difficult to fuse external knowledge from large-scale text corpora, which limits the performance. Given that, we proposed a method that fuses a pre-trained Vietnamese transformer language model into the streaming Vietnamese speech recognition model at the training stage both using the token level of Syllable. At the training stage, a novel loss function called L_{AED-LM} was introduced to optimize the parameters, learning the language knowledge by minimizing the difference between the output of pre-trained Vietnamese transformer language model and decoder. At the inference stage, we applied the Shallow Fusion or WFST technology to enhance the performance further. Experiments on the Vivos dataset show that, compared with the baseline model, the WER of streaming Vietnamese speech recognition can be improved by 2.45% by fusing the pre-trained language model during training; At the inference stage, Shallow Fusion or WFST improved the WER by 1.35% and 4.75% respectively.

Keywords: streaming speech recognition , Vietnamese , language model , pre-trained , end-to-end model

*余正涛(通信作者):ztyu@hotmail.com
国家自然科学基金（61732005, 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, 202002AD080001-5）；云南省基础研究计划（202001AS070014）；云南省学术和技术带头人后备人才（202105AC160018）

(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

越南作为“一带一路”沿线的重要国家，中越沟通合作交流越来越频繁。越南语语音识别可以提高中越双方沟通交流效率，开展越南语语音识别研究对促进中越贸易、政策沟通以及企业合作具有重要意义。

近几年来，端到端模型在语音识别任务中受到了广泛关注。端到端模型将传统语音识别系统的声学模型、发音词典和语言模型融合成一个模型，极大地减少了语音识别模型的训练流程和复杂性。目前，主流的端到端模型有Connectionist temporal classification (CTC)(Graves et al., 2006)、Recurrent Neural Network Transducer (RNN-T)(Rao et al., 2017)、Attention Based Encoder-Decoder(AED) (Chorowski et al., 2014; Chorowski et al., 2015; Chan et al., 2015)和Hybrid CTC/Attention(Kim et al., 2017; Hori et al., 2017)等模型。虽然这些端到端模型在多资源语种上取得了很好的效果，但是在训练过程中端到端模型难以利用外部大量文本中的语言知识(Gulcehre et al., 2015)。因此，一些研究者针对此问题，提出了一些在训练阶段将语言模型融合到语音识别模型的方法(Deep Fusion(Gulcehre et al., 2015)、Cold Fusion(Sriram et al., 2018)和Component Fusion(Shan et al., 2019))。实验结果表明，在训练阶段，将预训练语言模型融合到语音识别模型可以有效地帮助语音识别模型学习到一些语言知识，并弥补端到端模型在训练过程中难以利用外部大量文本语言知识的缺陷，同时提升语音识别模型的识别准确率。但是Deep Fusion、Cold Fusion和Component Fusion方法都需要语音识别模型增加额外的参数来融合语言模型，因此导致语音识别模型参数量增加的问题。并且这三种方法都采用RNN作为语言模型，在训练过程中不能像Transformer(Vaswani et al., 2017)模型一样并行训练，在一定程度上增加了语音识别模型的训练时间。

在越南语标注语音语料缺失的情况下，越南语语音识别模型的性能难以提升。相比获取越南语语音语料，获取越南语文本语料要容易得多，但目前的越南语语音识别模型并没有利用外部大量越南语文本中的语言知识来提升语音识别模型的识别率。同时，国内外针对流式端到端越南语语音识别模型的研究还很有限，大部分流式越南语语音识别模型仅在解码阶段使用了Shallow Fusion(Chorowski and Jaitly, 2016)方法融合语言模型，并没有在训练阶段融合语言模型的方法研究。

因此，本文针对以上问题，提出了一种将预训练Transformer越南语语言模型在训练阶段融合到流式越南语语音识别模型的方法。在训练阶段，仅通过预训练越南语语言模型和解码器的输出计算一个新的损失函数L_{AED-LM}，不会额外增加模型参数，还可以帮助流式越南语语音识别模型在训练过程中学习到一些越南语语言知识从而优化其模型参数；在解码阶段，使用传统的Shallow Fusion或者WFST(Wang et al., 2021)技术再次融合语言模型来纠正流式越南语语音识别模型的识别结果进一步提升模型性能。

本文的贡献如下：

(1)在训练阶段，将预训练越南语语言模型融合到流式越南语语音识别模型中，提升了流式越南语语音识别模型的识别率。

(2)在解码阶段，使用Shallow Fusion和WFST方法再次融合越南语语言模型进一步提升了流式越南语语音识别模型的识别率。

(3)本文在开源越南语数据集VIVOS上进行实验，在解码阶段不融合语言模型的情况下，相比基线模型，将流式越南语语音识别模型的词错率从31.03%降到了28.58%。在解码阶段，使用Shallow Fusion融合方法融合Transformer语言模型，词错率能提升到27.23%；使用WFST方法融合3元语言模型，词错率能提升到23.83%。

2 相关工作

近年来，虽然端到端语音识别受到了广泛关注，但目前针对越南语语音识别研究还比较少。Nguyen等人(2018)构建了500小时的越南语数据集并使用TDNN和BLSTM神经网络构建声学模型，在解码阶段融合了4元语言模型。为了提升模型性能，它将4元语言模型替换为RNN语言模型，在3小时测试集数据上进行测试，词错率达到6.9%。Nguyen和Huy(2019)使用CTC损失函数将TDNN和BLSTM模型结合一起联合训练越南语语音识别模型，在FPT测试数据集上，词错率达到14.41%。刘佳文(2020)提出了一种基于Transformer模型的越南语语音识模型，在VIVOS数据集上，字符错率达到40.4%。ESPNET(2021)基于不同的Transducer(Graves, 2012)模型在VIVOS数据集上做了不同实验，RNN-T词错率达到36.6%，Conformer(Gulati et al., 2020)/RNN-T词错率达到26%。为了提升模型识别率，这些模型都在解码阶段融合了语言模型，但在解码阶段融合语言模型只能影响模型的识别结果，并不能利用语言模型来优化语音识别模型的参数。因此，本文在流式越南语语音识别模型的训练阶段和解码阶段都融合了语言模型。在训练阶段融合语言模型可以帮助流式语音识别模型学习一些越南语语言知识优化其模型参数；在解码阶段融合语言模型可以帮助流式越南语语音识别模型纠正识别错误进一步提升其模型的识别率。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

3 融合外部语言知识的流式越南语语音识别

为了解决流式越南语语音识别模型难以学习到大量外部语言知识的问题，本文使用Hybrid CTC/Attention模型架构作为流式越南语语音识别模型的基线模型，在此基础上使用越南语单语文本语料预训练Transformer-xl(Dai et al., 2019)语言模型，并将其在训练阶段融合到流式越南语语音识别模型中，以提升模型识别效果，具体方法如下所述。

3.1 模型架构

语音识别Hybrid CTC/Attention模型架构由三个部分组成：共享编码器、CTC解码器和Attention-Based解码器。共享编码器由多层Transformer编码器构成；CTC解码器由一个线性层、log softmax层构成；Attention-Based解码器由多层Transformer解码器构成。在Hybrid CTC/Attention模型架构的基础上，本文将预训练越南语Transformer-xl语言模型与Hybrid CTC/Attention模型中的Transformer解码器进行了融合，如图1所示。

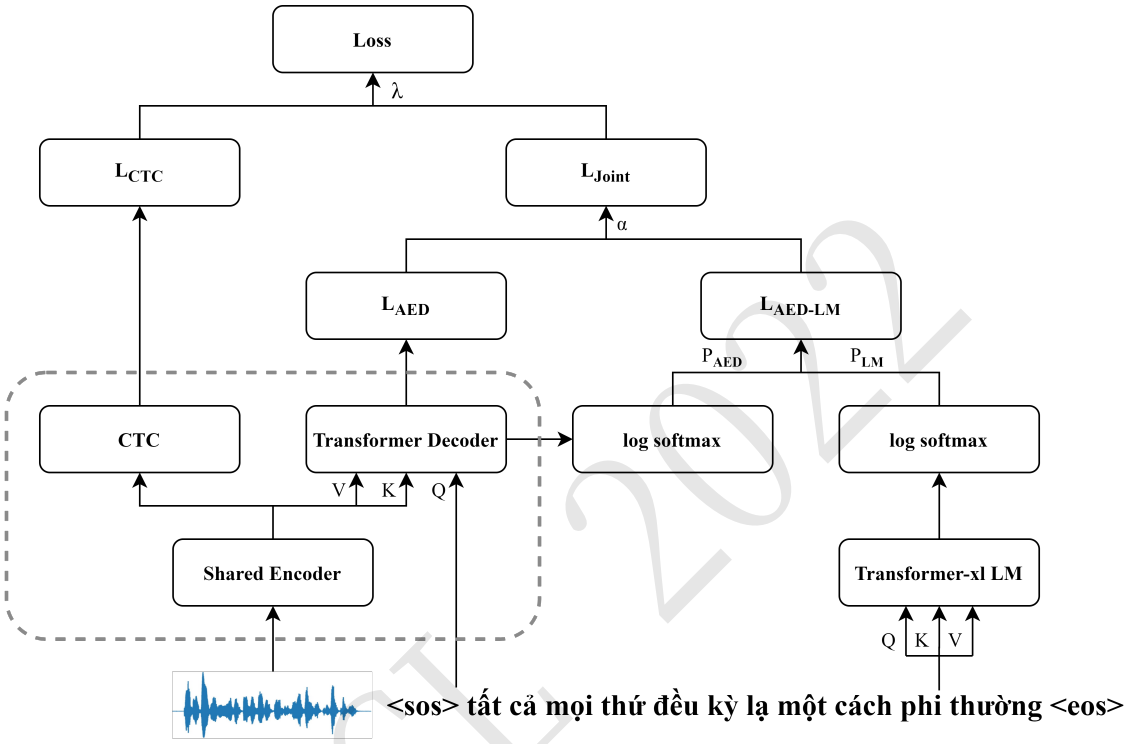


图 1. Hybrid CTC/attention模型融合越南语语言模型架构图

音频特征X=(x_t|t= 1,2,...,T)经过共享编码器编码生成高维音频特征向量H=(h_t|t= 1,2,...,T)，然后将音频特征向量H和标签送入CTC解码器和Transformer解码器进行解码。利用Transformer解码器和预训练Transformer-xl语言模型的输出经过log softmax函数计算出两个输出概率P_{AED}和P_{LM}。使用P_{AED}概率计算Transformer解码器的损失函数L_{AED}；使用P_{AED}和P_{LM}两个输出概率计算越南语Transformer-xl语言模型和Transformer解码器的融合损失函数L_{AED-LM}。再通过语言模型融合超参数$\alpha(0\leq\alpha\leq1)$联合P_{AED}和L_{AED-LM}生成L_{Joint}损失函数，最终经过CTC权重超参数λ联合L_{CTC}和L_{Joint}损失函数训练流式越南语语音识别模型。

3.2 越南语语言模型

在构建语言模型时，虽然越南语是一种以单音节为主的语言(Haudricourt, 2010; Alves, 2006; Hwa-Froelich et al., 2002; Thompson, 1991)，但是少部分词会包含多个音节。在包含多音节词的句子中，上下文音节之间的依赖长度较长会导致模型出现长期依赖丢失的问题，并且在模型编码越南语音节时，句子长度过长会使音节丢失在句子中的位置信息。因此，本文使用Transformer-xl作为越南语语言模型，可以解决越南语音节长期依赖和位置编码丢失的问题，从而使越南语语言模型更好地表征越南语语言知识。在融合过程中能让语音识别模型从越南语语言模型更好地学习到越南语语言知识，提升语音识别模型的识别率。

本文使用33万句越南语文本作为训练语料，在训练阶段，Transformer-xl使用片段递归的方法将当前隐藏层状态作为Q，将之前的隐藏层状态与当前隐藏层状态拼接后作为K和V，最终使用Q、K、V来计算attention的输出。这种方法使得越南语语言模型具有更强的长期依赖能力，从而更好地表征越南语语

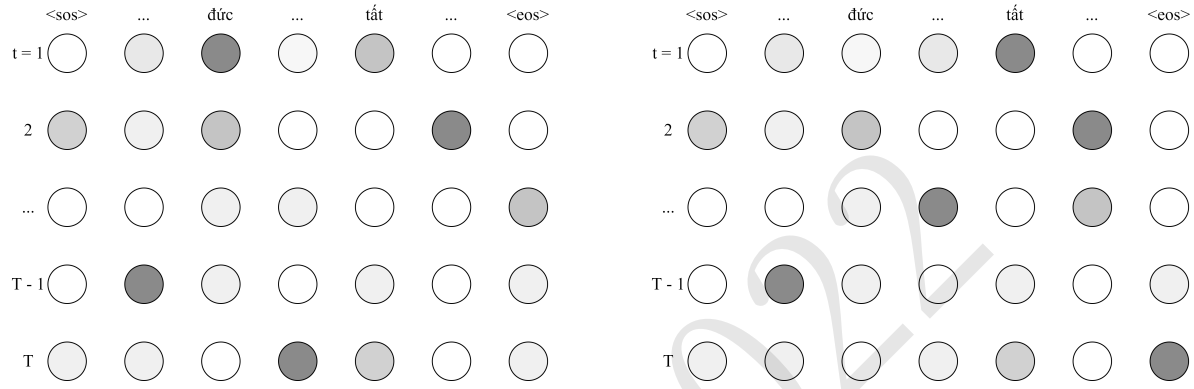
(c) 2022 中国中文信息学会计算语言学专业委员会

言知识。但由于语音识别任务句子之间没有上下文关系，因此在解码阶段，本文并没有使用片段递归的方法来保存之前句子的隐藏层状态，而是直接将目标句子作为Transformer-xl的输入经过解码后得到预测每一个音节的概率分布，并在语音识别模型中使用此概率来融合越南语语言知识。

3.3 越南语语言模型融合方法

在训练阶段将语言模型和流式越南语语音识别模型融合，本文使用KL散度来计算Transformer解码器和越南语Transformer-xl语言模型输出之间的融合损失函数。其目的是为了让Transformer解码器的输出概率分布向越南语语言模型的输出概率分布靠近，从而帮助语音识别模型从越南语语言模型中学习到越南语语言知识。具体融合方法如下所述。

假设，目标序列长度为L，0表示开始符<sos>，L表示结束符<eos>，P_{AED}=P($Y_{1\sim L}$|H,$Y_{0\sim L-1}$)表示Transformer解码器在给定共享编码器输出特征向量H和输入目标序列$Y_{0\sim L-1}$的条件下，预测出目标序列$Y_{1\sim L}$的输出概率分布；P_{LM}=P($Y_{1\sim L}$|$Y_{0\sim L-1}$)表示越南语语言模型在给定输入目标序列$Y_{0\sim L-1}$的情况下输出目标序列$Y_{1\sim L}$的输出概率分布。本文将越南语语言模型输出的P_{LM}作为真实分布，Transformer解码器输出的P_{AED}作为理论数据分布，如图2所示。



(a) Transformer解码器输出(P_{AED})　　(b) 越南语语言模型输出(P_{LM})

图 2. 越南语语言模型融合方法。每一个节点代表Transformer解码器和越南语语言模型在给定输入后输出对应越南语音节的概率值（颜色深浅代表概率值大小）。将(a)作为理论数据分布，(b)作为真实分布。

然后使用KL散度来计算Transformer解码器与越南语Transformer-xl语言模型的融合损失函数，如公式(1)所示：

$$D_{KL}(P_{LM}||P_{AED}) = \sum_{i=0}^{L} P_{LM} \log \frac{P_{LM}}{P_{AED}} \tag{1}$$

最后令L_{AED-LM}损失函数等于$D_{KL}(P_{LM}||P_{AED})$并使其最小化。

在融合过后，本文引入了一个语言模型融合超参数$\alpha(0\leq\alpha\leq1)$，用来调节Transformer解码器L_{AED}损失函数和L_{AED-LM}损失函数。令联合损失函数为L_{Joint}，其计算过程如下：

$$L_{Joint} = (1-\alpha)L_{AED} + \alpha L_{AED-LM} \tag{2}$$

最终训练的联合损失函数与传统Hybrid CTC/Attention模型损失函数类似，不同的是将Attention损失函数修改为Transformer解码器融合了越南语语言模型的损失函数。如公式(3)所示，其中$\lambda(0\leq\lambda\leq1)$参数和传统Hybrid CTC/attention模型一致。

$$Loss = \lambda L_{CTC} + (1-\lambda)L_{Joint} \tag{3}$$

流式越南语Hybrid CTC/Attention模型最后使用Loss损失函数来训练。这种训练方式可以让融合损失函数L_{AED-LM}在训练阶段同时优化CTC解码器和Transformer解码器的参数，帮助CTC和Transformer解码器联合学习到越南语语言知识从而提升流式越南语Hybrid CTC/Attention模型的识别率。

3.4 解码

传统Hybrid CTC/Attention模型使用自回归的方式解码，解码速度慢，所以本文采用了two-pass(Sainath et al., 2019)模型架构中的二次评分模式实现非自回归方式解码加快解码速度。首先，

(c) 2022 中国中文信息学会计算语言学专业委员会

使用CTC解码器生成N个最好的序列，然后再将这N个序列和编码器的输出送入到Transformer解码器以Teacher-Forcing的方式解码，最终选取评分最高的序列输出。

由于在训练阶段融合语言模型后，可以在解码阶段选择性地使用传统语言模型再次融方法进一步提升模型识别率，因此，本文在解码阶段分别使用了Shallow Fusion和WFST两种方法再次融合语言模型进一步提升模型识别率。

4 实验

4.1 数据集

本文使用开源越南语数据集VIVOS进行实验。VIVOS数据集的训练数据集由46个说话人在安静环境下录制了15个小时的音频，其中包含11660个句子；测试集由19个说话人在相同环境下录制了45分钟的音频，其中包含了760个句子。所有音频都是以16kHz的采样率采样，并且以小端WAV格式存储。

本文爬取了33万句越南语文本语料来训练语言模型。在语料的预处理过程中，去除了标点符号、数字。由于部分网站使用不同的越南语编码格式，在构建语料库时，需要将越南语文本语料统一编码为utf-8编码格式，确保在识别过程中不会因为编码问题而导致同一个音节识别错误和同一个音节在词典中出现多次的问题。

4.2 参数设置

本文使用80维log梅尔滤波器组(FBANK)在窗口大小为25ms帧移为10ms的条件下对VIVOS数据集的音频文件进行特征提取；使用SpecAugment(Park et al., 2019)技术对提取的音频特征进行数据增强；在数据前处理阶段，使用卷积核大小为3*3，步长为2的两个卷积层对音频数据进行做下采样处理；使用12层带有4头注意力的Transformer作为编码器，令编码器的输出维度为256；为了使编码器支持流式编码，使用固定chunk和动态chunk(Zhang et al., 2020)两种方式进行编码，分别对比了不同编码方式对流式越南语语音模型的影响；使用6层带有4头注意力的Transformer联合CTC作为解码器；为了防止过拟合，在解码器和编码器的每一层都设置了dropout，并且将dropout比率设置为0.1；使用Adam优化器，学习率设置为0.002，学习率预热设置为25000步；使用标签平滑技术来计算损失函数，标签平滑率设置为0.1。最后，选取30个最好的模型进行参数平均得到最终模型，提升模型的泛化能力。在解码阶段，使用前缀束搜索算法对CTC解码器的输出进行搜索，束搜索宽度大小设置为16，束搜索产生束宽度大小个first-pass结果，然后再将这些结果送入Transformer解码器中进行二次评分。在二次评分阶段使用CTC权重超参数λ来控制CTC解码器的输出权重和Transformer解码器输出权重。

本文使用Transformer-xl作为语言模型，它由12层带有4头注意力的Transformer编码器构成并引入了相对位置编码和片段递归机制。本文使用自己构建的33万句越南语单语文本语料训练越南语语言模型并采用Adam优化器优化模型参数并设置学习率为0.00025，学习率预热为20000步。

模型词表中共9078个词，其中包含4个特殊标签，<blank>表示CTC的空标签，<unk>表示未登录词，<sos>和<eos>表示句子的开始和结束。最终CTC解码器和Transformer解码器的输出维度为9078。本文所有实验在一张NVIDIA Tesla T4上完成。

4.3 实验结果及分析

4.3.1 不同chunk方法训练对流式越南语语音识别模型性能的影响

本文使用Hybrid CTC/Attention模型架构作为越南语语音识别的基线模型。为了使模型能够流式输出，本文修改了Hybrid CTC/Attention模型的编码方式，使用不同chunk大小和动态chunk的编码方式训练模型，并对比了不同编码方式对模型性能的影响。

在对比不同chunk方法对模型的影响时，本文将CTC权重超参数λ固定设置为0.3，语言模型融合权重超参数α固定设置为0，然后将chunk大小分别设置为8/16/动态chunk进行对比实验，实验结果见表1。

表 1. 不同chunk方法对模型性能的影响

chunk大小	解码chunk大小	词错率(WER%)
8	8	36.69
16	16	34.48
动态chunk	16	31.03

根据实验结果数据显示，当chunk大小设置为8/16时，流式越南语语音识别模型的词错率分别为36.39%和34.48%。当chunk大小设置为动态chunk时，流式越南语语音识别模型识别性能达到最佳31.03%。

对于越南语语音识别任务而言，动态chunk方式训练的模型效果明显优于以固定chunk大小训练的模型识别性能。主要是因为越南语自身是一种以单音节为主的语言，但有一些词包含多个音节，因此使用动态chunk的方式编码更符合越南语由不同音节个数构成词的特点从而使得模型识别性能更佳。

(c) 2022 中国中文信息学会计算语言学专业委员会

在接下来的实验中，我们均采用动态chunk编码方式训练的Hybrid CTC/Attention语音识别模型作为流式越南语语音识别模型的基线模型。

4.3.2 融合语言模型对流式越南语语音识别模型性能的影响

为了验证本文提出的方法对流式越南语语音识别模型性能有提升，本文将流式越南语语音识别模型的CTC超参数λ和语言模型融合超参数α分别设置为不同值，对比在训练阶段融合语言模型前后和不同超参数对流式越南语语音识别模型性能的影响，实验结果见表2和表3。

表 2. 当CTC权重为0.3时，融合语言模型权重α对流式越南语语音识别模型的影响

语言模型融合权重α	词错率(WER%)
0(baseline)	**31.03**
0.3	**28.58**
0.5	33.22
0.7	29.15

表 3. 当CTC权重为0.5时，融合语言模型权重α对流式越南语语音识别模型的影响

语言模型融合权重α	词错率(WER%)
0(baseline)	**30.30**
0.3	29.41
0.5	29.54
0.7	29.60

实验结果数据显示，当CTC权重参数设置为0.3时，在不融合语言模型(融合语言模型权重参数α为0)的情况下，流式越南语语音识别模型词错率为31.03%(baseline)。当以0.3的权重融合语言模型时，性能有明显提升，词错率达到了28.58%。但当语言模型融合权重设置为0.5时，性能相比基线模型有一定下降。当语言模型融合权重设置为0.7时，性能相比基线模型又有一定提升，达到29.15%。当CTC权重参数设置为0.5时，在不融合语言模型的情况下，流式越南语语音识别模型词错率为30.30%(baseline)。当语言模型融合权重参数分别设置为0.3/0.5/0.7时，流式越南语语音识别模型的识别性能相比基线模型都有所提升，但语言模型融合权重参数对流式越南语语音识别模型的识别词错率影响不怎么明显，词错率保持在29%左右。

当CTC权重参数为0.3，语言模型融合权重参数为0.5时，性能相比基线模型有一定下降。主要是因为当语言模型融合权重设置为0.5时，解码器和语言模型的输出比重相同，语音识别模型不能抉择解码器和越南语语言模型输出的重要性，从而导致模型混乱，识别性能下降。但是当语言模型融合权重设置为其他值时，性能相比基线模型都有一定提升。这说明了流式越南语语音识别模型可以从越南语语言模型中学习到越南语语言知识从而优化其模型参数，达到识别性能提升的效果。

4.3.3 识别结果示例分析

本文将流式越南语语音识别模型的CTC权重参数设置为0.3，语言模型融合权重参数分别设置为0/0.3进行了对比实验，并通过对测试集识别结果示例的分析来说明融合越南语语言模型可以提升模型的识别效果。实验结果见4。

表 4. 识别结果示例分析

识别结果α	词错率(WER%)
tất cả mọi thứ đều kỳ lạ một cách phi thường (原标签)	-
tất cả mọi thứ đều kỳ **lạ** một cách phi thường (融合语言模型识别出的标签)	**0**
đức cả mọi thứ đều kỳ **là** một cách phi thường (未融合语言模型识别出的标签)	18.18

实验结果表明，融合了语言模型的流式越南语语音识别模型识别结果完全正确，而未融合语言模型的流式越南语语音识别模型识别结果词错率为18.18%。

未融合语言模型的流式越南语语音识别模型识别错了两个音节đức和là，主要原因是đức和tất、là和lạ音节的发音非常相似，提取出来的语音特征也非常接近，从而导致语音识别模型不能辨别。而融合了语言模型的流式越南语语音识别模型可以从语言模型中学习到tất cả和kỳ lạ可以组

(c) 2022 中国中文信息学会计算语言学专业委员会

成一个词，而đức cả和kỳ là不能组成词，从而tất cả和kỳ lạ的输出概率高于đức cả和kỳ là，因此语音识别模型选择tất cả和kỳ lạ输出。

实验结果表明，在训练阶段融合语言模型确实可以优化流式越南语语音识别模型参数从而纠正一些将越南语音节识别错误的情况。

4.3.4 二次融合语言模型对流式越南语语音识别模型性能的影响

当流式越南语语音识别模型的CTC权重参数设置为0.3，语言模型融合权重参数设置为0.3时性能最佳，因此在解码阶段二次融合语言模型也使用此参数配置。为了验证二次融合语言模型对流式越南语语音识别模型识别率的影响。本文在解码阶段使用Shallow Fusion和WFST方法分别对Transformer-xl语言模型和3元语言模型进行融合。实验结果如表5和表6所示。

表 5. 使用Shallow Fusion融合方法对流式越南语语音识别模型性能的影响

模型	词错率(WER%)
Hybrid CTC/Attention(baseline)	**31.03**
Hybrid CTC/Attention +训练阶段融合语言模型	28.58
Hybrid CTC/Attention + Shallow Fusion	29.83
Hybrid CTC/Attention + 训练阶段融合语言模型+ Shallow Fusion	**27.23**

表 6. 使用WFST方法对流式越南语语音识别模型性能的影响

模型	词错率(WER%)
Hybrid CTC/Attention(baseline)	**31.03**
Hybrid CTC/Attention +训练阶段融合语言模型	28.58
Hybrid CTC/Attention + WFST	24.32
Hybrid CTC/Attention + 训练阶段融合语言模型+ WFST	**23.83**

实验数据结果显示，在训练阶段融合语言模型后，在解码阶段使用Shallow Fusion方法再次融合Transformer-xl语言模型还可以将模型的识别率提升1.35%；在训练阶段融合语言模型后，在解码阶段使用WFST融合3元语言模型，性能达到最佳23.83%。

虽然使用Shallow Fusion或WFST方法进行解码，模型识别率会有所差距，但实验数据结果显示，在训练阶段融合语言模型后，在解码阶段再次融合语言模型确实可以进一步提升流式越南语语音识别模型的识别率。同时，在训练阶段和解码阶段都融合语言模型，模型的识别率要明显高于在解码阶段单独融合语言模型的识别率。

4.3.5 和其他模型的性能比较

本次实验对比了本文使用的流式模型和ESPNET使用RNN-T、Conformer/RNN-T模型在VIVOS测试数据集上的结果。实验结果如表7所示。

表 7. 和其他模型识别效果对比

模型	词错率(WER%)
RNN-T	36.6
Conformer/RNN-T	26.0
Hybrid CTC/Attention + 训练阶段融合语言模型+ WFST	**23.83**

实验结果数据显示，本文使用的流式模型词错率达到23.83%，RNN-T和Conformer/RNN-T模型的词错率分别为36.6%和26.0%。

本文在训练阶段融合语言模型后，再使用WFST在解码阶段融合3元语言模型的识别率达到最佳。其主要原因是本文同时在训练阶段和解码阶段都融合了语言模型。在训练阶段融合语言模型可以优化模型的参数；在解码阶段融合语言模型可以纠正语音识别模型识别结果。而ESPNET仅在解码阶段融合了语言模型，只影响了语音识别模型的识别结果，并不能优化模型的参数。

5 总结

由于越南语标注语音语料难以获取，流式越南语语音识别模型难以在训练阶段利用外部文本语言知识的问题，本文提出了一种在训练阶段将预训练越南语Transformer-xl语言模型融入到流式越南

(c) 2022 中国中文信息学会计算语言学专业委员会

语Hybrid CTC/Attention模型的方法。实验表明，这种融合方法可以提升流式越南语语音识别模型的识别率并弥模型在训练过程中难以学习外部语言知识的缺陷。另外，在解码阶段再使用一些传统语言模型融合方法还可以进一步提升语音识别模型的识别率。

参考文献

Mark Alves. 2006. Linguistic research on the origins of the vietnamese language: An overview. *Journal of Vietnamese Studies*, 1(1-2):104–130.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

Jan Chorowski and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *ArXiv*, abs/1211.3711.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, and Yuekai Zhang. 2021. Recent developments on espnet toolkit boosted by conformer. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878.

André-Georges Haudricourt. 2010. The origin of the peculiarities of the vietnamese alphabet. *Mon-Khmer Studies*, 39:89–104.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. 2017. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. *arXiv preprint arXiv:1706.02737*.

Deborah Hwa-Froelich, Barbara W Hodson, and Harold T Edwards. 2002. Characteristics of vietnamese phonology.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

Quoc Bao Nguyen, Van Tuan Mai, Quang Trung Le, Ba Quyen Dam, and Van Hai Do. 2018. Development of a vietnamese large vocabulary continuous speech recognition system under noisy conditions. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, pages 222–226.

Huy Nguyen. 2019. An end-to-end model for vietnamese speech recognition. pages 1–6, 03.

(c) 2022 中国中文信息学会计算语言学专业委员会

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Kanishka Rao, Haşim 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)*, pages 193–199. IEEE.

Tara N Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, et al. 2019. Two-pass end-to-end speech recognition. *arXiv preprint arXiv:1908.10992*.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Component fusion: Learning replaceable language model component for end-to-end speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5361–5635. IEEE.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. Cold fusion: Training seq2seq models together with language models. pages 387–391, 09.

Laurence C Thompson. 1991. A vietnamese reference grammar (revised edition).

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.

Zhichao Wang, Wenwen Yang, Pan Zhou, and Wei Chen. 2021. Wnars: Wfst based non-autoregressive streaming end-to-end speech recognition. *arXiv preprint arXiv:2104.03587*.

Binbin Zhang, Di Wu, Zhuoyuan Yao, Xiong Wang, Fan Yu, Chao Yang, Liyong Guo, Yaguang Hu, Lei Xie, and Xin Lei. 2020. Unified streaming and non-streaming two-pass end-to-end model for speech recognition. *arXiv preprint arXiv:2012.05481*.

刘佳文, 屈丹, 杨绪魁, 张昊, and 唐君. 2020. 基于transformer的越南语连续语音识别. 信息工程大学学报, 21(2):129–133,152, 4.

CCL 2022

针对古代经典文献的引用查找问题的数据构建与匹配方法

李炜，邵艳秋
北京语言大学
信息科学学院
北京市海淀区学院路15号，100083
liweitj47@blcu.edu.cn，yqshao163@163.com

毕梦曦*
复旦大学
哲学学院
上海市杨浦区邯郸路220号，200433
1207950557@qq.com

摘要

中国古代思想家的思想建构往往建立在对更早期经典的创造性诠释中，将这些诠释中包含的引用查找出来对思想史研究意义重大。但一些体量较大的文献如果完全依靠手工标记引用将耗费大量时间与人力成本，因此找到一种自动化的方法辅助专家进行引用标记查找非常重要。以预训练语言模型为代表的自然语言处理技术的发展提升了计算机对于文本处理和语义理解的能力。据此，本文提出多种利用专家知识或深度学习语义理解能力的无监督基线方法来自动查找古代思想家著作中对早期经典的引用。为了验证本文提出的方法的效果并推动自然语言处理技术在数字人文领域的应用，本文以宋代具有重大影响力的理学家二程（程颢、程颐）对早期儒家经典的引用为例进行研究，并构建和发布相应的引用查找数据集[1]。实验结果表明本文提出的基于预训练语言模型和对比学习目标的复合方法可以较为准确地判断是否存在引用关系。基于短句的引用探测ROC-AUC值达到了87.83，基于段落的引用探测ROC-AUC值达到了91.02。进一步的分析表明本文的方法不仅有利于自动化找到引用关系，更能够有效帮助专家提高引用查找判断效率。本方法在注释整理、文本溯源、重出文献查找、引用统计分析、索引文献集制作等方面具有广阔的应用前景。

关键词： 引用查找；数字人文；古代文献

Data Construction and Matching Method for the Task of Ancient Classics Reference Detection

Wei Li, Yanqiu Shao
Beijing Language and Culture University
School of Information Science
15 Xueyuan Rd., HaiDian District,
Beijing, 100083
liweitj47@blcu.edu.cn，yqshao163@163.com

Mengxi Bi
Fudan Univeristy
School of Philosophy
Handan Rd., Yangpu District
Shanghai, 200433
1207950557@qq.com

Abstract

The idea construction of ancient Chinese ideologists tend to be built on the basis of explaining early ideological claims. Therefore, finding out the references owes great significance for the research on ideological history. However, it would be much too expensive for both time and man power if we only fully depend on human experts to label the literature especially when the literature is of large amount. Hence force, it is of great importance to develop an automatic method to facilitate experts looking for the reference items. With the development of natural language processing technologies typified by pre-trained language models, the ability for processing and understanding

* 通讯作者 Corresponding Author
[1]数据可在https://github.com/liweitj47/classic_reference_detection找到

第二十一届中国计算语言学大会论文集，第600页-第610页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

natural language has improved a great deal. Based on these observations, we propose several unsupervised baseline methods to automatically detect the references to early literature, which use either expert knowledge or deep language understanding technologies. To testify the effectiveness of our proposed method as well as promote the application of natural language processing techniques to the field of Digital Humanities, we take the literature of Ercheng referencing early Confusion classics as example, and construct the corresponding labelled dataset for reference detection. The experiment results show that our ensemble method based on pretrained language model and contrastive objective can accurately detect whether there exists reference. Sentence level reference detection achieves 87.83 on ROC-AUC, while paragraph level reference detection achieves 91.02 on ROC-AUC. Further analysis show that this work can not only help automatically find reference, but also improve the efficiency for the expert finding references. This model has great prospects on organizing annotation, text tracing, duplicate literature detection, reference statistical analysis and reference index generation.

Keywords: Reference Detection , Digital Humanity , Ancient Classics

1 引言

中国古代思想家的思想建构往往不是从零开始自己创建一套思想体系，而是在诠释早期经典的过程中完成自己思想的构建，中国古代思想的发展演变在这一方面具有一定的特殊性。黄俊杰先生就曾提出(黄俊杰, 2018)，“儒家经典解释者每一次所提出的新解释，都是一次通过解释者个人的思想系统和生命体验而完成的再创造。但同时解释者也不是完全自由的，诠释性的发挥需要在原典文本的印可范围之内。”所以在中国古代思想研究这一领域中，无论是哲学研究、思想史研究还是经学研究，思想家们对早期经典的引用与诠释都是研究中重要的问题。宋元明清思想史研究中，材料体量普遍比较大，给查找工作带来了难度。

传统人文研究者面对体量比较大的文献材料时，引用的查找统计就会变得非常困难。以本文所举的二程文献为例，中华书局点校本《二程集》总体字数约50万字左右，按照点校者的分段计数，共有语录、文章、书信等不同体裁的文字6000多段，在宋元明清思想研究中属于中等的文献量。即便是在准确识别引用的前提下，统计和标注也需要耗费比较大的工作量，这些重复工作会给人文领域的研究带来巨大的时间和人力成本。

利用自然语言处理技术来处理规模较大的文本具有天然的优势，对处理经典古籍等文本也具有很强的借鉴意义，为数字人文领域提供了重要的技术基础。但是传统的基于字符匹配的方法存在依赖专家经验，灵活度不足，覆盖范围不全，难以准确把握上下文语义内涵等问题。这些问题使得人文领域学者难以在实际中使用这类工具得出可靠的结论，限制了相关技术在数字人文领域的应用。

随着预训练模型等深度学习方法的发展(Devlin et al., 2019; Qiu et al., 2020)，其被应用在了自动问答(Yang et al., 2019)、阅读理解(Zhang et al., 2020)、机器翻译(Conneau et al., 2020)等众多任务中，并取得了巨大的成功，而这与预训练语言模型能够在一定程度上能够更好地把握上下文语义有关。此外，基于对比学习提出的SIMCSE(Gao et al., 2021)在预训练语言模型的基础上，通过对句中字符表示随机加入dropout(Srivastava et al., 2014)噪音的方式来获得对比学习(Jaiswal et al., 2020)中的正例，因而能够通过自监督方式学习到包含语义信息的更有区分度的向量表示。

在本文中，我们结合人文领域专家对文本本身及引用现象的观察和经验以及预训练语言模型和对比学习目标，提出了三种基础的引用查找和探测方法，并基于这三种基础方法，组合出不同的复合方法作为此任务的基线模型。为了能够对相关任务和方法做出可以量化的评价，我们首先给出了引用问题的明确定义，之后筛选出二程文本中一些较有代表性的文本，并标注了它们中对古代儒家最重要的十三经文本是否存在引用关系。最终得到了含822句的开发集

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

和2484句的测试集。我们在构建的数据集上进行了大量实验。结果表明，我们提出的利用预训练语言模型和对比学习目标的字符片段和句子语义相结合的复合引用查找方法在测试集上的效果最好，基本达到了人文领域实际应用的要求，并且泛化性也相较于基于规则的方法更好。

我们将本文的贡献总结如下：

- 我们提出一种考虑人文领域实际研究应用需求的古代经典引用定义。并据此以二程文本为例，构建并发布二程对早期儒家经典文本（十三经）的引用关系数据集；
- 我们结合人文领域专家经验和预训练语言模型以及对比学习目标，提出了多种引用查找和探测的基线方法，并进行了大量的实验和分析。结果表明，基于预训练语言模型目标和对比学习的字符片段和句子两种级别的复合方法可以较为有效地探测出是否存在引用关系，并能为人文学者的实际研究提供有力支持。

2 相关工作

国内学者对于古文引用相关的研究已经取得了一些成果。黄水清等人(黄水清et al., 2021)和周好等人(周好et al., 2021)针对规整古代文献中出现的论著名（明引）采用基于序列标注的方法进行了识别和统计学分析。尽管该方法可以识别带有明确书名的相关表述，后世文献（如宋代及之后）对早期经典文献的引用大多只引用早期经典的只言片语，并且是基于语义的引用，因此该方法并不适用于广泛存在于后世文献中对早期经典文献的引用识别（暗引）。本文的研究方法可以同时识别有书名标引的**明引**和没有书名标引而只有语义关联的**暗引**情况，因此具有更广泛的应用场景以及对人文领域更实际的应用价值。

对于在数字人文领域利用预训练语言模型来说，也有许多工作进行了尝试。耿云冬等人(耿云冬et al., 2022)、刘江峰等人(刘江峰et al., 2021)、胡昊天等人(胡昊天et al., 2022)和徐润华等人(徐润华et al., 2022)利用在四库全书语料上训练的siku-bert(王东波et al., 2022)模型分别尝试了词性标注、实体识别、文本分类和自动摘要的任务，俞敬松等人(俞敬松et al., 2019)将自行训练的古文BERT应用于古文的自动断句，均取得了良好的效果。然而这些工作大多仍然局限在自然语言处理的传统任务范式中，与直接的人文领域研究有一定距离。本文将siku-bert作为基础的预训练模型来辅助进行语义匹配，进而和专家知识、对比学习等相结合，以查找后世（以二程为例）文献对早期儒家经典论述（十三经）的引用。

3 引用查找数据构建

3.1 引用的定义

这里我们对何为引用进行定义。所谓引用需要与原文至少有两个字的重叠，作为引用的标识，而且重叠的部分具有辨识度，对原文的索引有提示作用，提示的作用指能够根据重合的文字找到原文出处之外，重合的文字还不能属于古文中反复出现的字词。以二程文献来说，首先二程文献如果与早期儒家经典有整段的连续重合，可以视为（规则的）引用：

不**甚，则身危国削**，名之曰幽厉，虽孝子慈孙，百世不能改也（《程氏遗书·卷23》）
不甚，则身危国削。（《孟子》）

此外，还存在大量不规则的引用。不规则的引用有多种情况，有的时候二程提及的字数比较少，只提到了关键的两三个字，或者在引用中调换了字词的顺序等，在没有使用模型判断之前，无论是人工判断还是在数据库中搜索，这一部分引用都是查找的难点：

既为先觉之民，岂可不觉未觉者（《程氏遗书·卷1》）
予，**天民之先觉者也。**（《孟子》）

无论是在整段的连续引用还是在不规则的引用中，“引用”都不只是字符的简单重复，还要求重合的字符需要有一定的特殊性和辨识度，能够辨认出二程文献中所提及的词语是来自哪部经典。以二程文献和《孟子》举例说明，比如二程提到了“浩然之气”，这种说法就具有一定的辨识度，可以认为是在用《孟子》中的典故。而“仁义”这样的说法就比较宽泛，多种早期经典中都有出现，虽然《孟子》当中也有，但是无法根据语境认为这个说法是在诠释或者引用《孟子》，这样的重复就不算引用，而应视为二程对“仁”或者“义”这些概念的发挥诠释。

(c) 2022 中国中文信息学会计算语言学专业委员会

3.2 数据整理和标注

本研究采用中华书局2004年版由王孝鱼先生点校的《二程集》作为数据来源，是较为可靠全面的二程文献集。为了后续导出的结果可以有更多的分析维度，本研究在数据构建阶段即对材料进行了更细致的标注。

首先是书名和卷数的标记。《二程集》是二程的几种文献集合在一起形成的，一共分为63卷，包括《程氏遗书》25卷，《程氏外书》12卷，《程氏文集》12卷，《伊川易传》4卷，《经说》8卷，《粹言》2卷，汇集了二程的全部语录和著作。首先需要对每段文字所在的书名以及卷数进行标记，这样引用查找结果导出后就可以进一步分析引用在二程文献中的分布情况，展开下一步的研究。

其次是作者的标记，即该段文字属于程颐或属于程颢。《程氏文集》《伊川易传》等著作文献作者归属是明确的，根据署名标记作者即可。但二程文献的语录部分（《程氏遗书》《程氏外书》《粹言》）有一部分记录和整理者已经标注了语录属于程颐或程颢，有一部分语录尚不明归属，学生记录的时候没有标明作者。所以需要以程颐、程颢、未知三类对每一段语录的作者进行标记。

第三是对注释性质的文献进行清理。《伊川易传》将《周易》原文分成了1253段，并且按照原文的顺序逐一进行了注释，体例是一段原文，一段注释，所引用的经典原文独自成段。这1253处可以直接进行统计，不需要再进行匹配。在程颐《伊川易传》注释工作中使用的其他儒家经典，则需要通过匹配进一步统计。进行相同处理的还有《经说》当中的《春秋传》和《书解》。文献中的祭文、年谱，文献编辑者所写的序文等由他人完成的，程颐程颢自身没有参与的文献内容需要删除。诗歌部分引用不明显，而且与语录、书信等文献的语言情况差异比较大，所以也进行删除处理。《易序》因为作者归属尚且存疑，进行删除处理。《经说》部分的伊川与明道先生改正《大学》，因为不是二程自己的表述，所以也进行删除。根据葛瑞汉先生的考证(葛瑞汉, 2000)，《经说》当中《春秋解》程颐亲作的部分应当到桓公九年止，所以桓公九年之后的部分予以删除。《经说》部分《中庸解》作者存疑，删除处理。

因为原文本中的许多标点可能会对模型的自动判断起到误导作用，因此本研究首先去除了原文本中诸如“【】『』”等标点符号。为了保证匹配的粒度相对统一，减少句长差异带来的干扰，本研究将文本大致限定在8~30字的范围内。为了达到这样的目的，如果长度超过30，那么本研究按照较大语义停顿的标点符号（如，句号、感叹号、问号等）进行分句。经过分句处理后，部分句子长度会非常短，以致难以提供足够的信息进行匹配。因此本研究对长度太短（小于8字）的句子进行向后合并（与后一句合并）。因为段落中的匹配是建立在内部句子级别的匹配上的，因此本研究中经段落拆分后的句级文本可以较好地用于后续的引用判断。

需要在二程文献中进行查找的经典原文指早期儒家经典十三经，使用古籍文献ctext网站[1]的版本。其中《春秋》三传当中都对《春秋》原文有引用，其体例都是原文引用+注释，为了清楚地查找二程所引的文献是出自《春秋》原文还是传文，将三传中的《春秋》原文清理出去，只留下传文。同时将《春秋》原文单列为一个文件。因为《中庸》《大学》在中国古代的经典注释和传播中往往是独立于《礼记》而进行的，所以将这两篇文章从《礼记》里摘出来单列为两个文件，共有《论语》《孟子》《大学》《中庸》《诗经》《尚书》《周易》《春秋》《仪礼》《礼记》《周礼》《谷梁》《公羊》《左传》《尔雅》《孝经》16个经典文本。

有时候二程讨论早期儒家经典没有直接引述经典内容，而是只提及了书名（如《论语》）或者某个章节的名称（如《离娄》），所以十三经本身的标题以及经典原文当中的各个章节的名称，比如《论语》等，也需要纳入查找的范围。本研究将这些书名和章节名单独列为一个清单，加入所属的经典文件中。《谷梁传》常简称《谷梁》，两个说法都需要列出，其他经典的别称也做同样的处理。对早期经典文献的处理与二程文本基本相同。不同点在于早期经典相较于二程文本来说的表意普遍较为集中、凝练，因此本研究只对早期经典按照标点符号切分成小句，而不对小句进行聚合。

3.3 停用词集构建

停用词主要是指文本中基本不承担实际语义的字或词，在现代汉语中一般以虚词的成分出现。停用词主要用于检索系统中，在查询语句与目标语句进行匹配时，停用词不计入匹配结果或者给与停用词一定惩罚。

[1] https://ctext.org/zhs

(c) 2022 中国中文信息学会计算语言学专业委员会

古代汉语的虚词与实词的分界本身是一个比较复杂的问题，有些词语介于实词和虚词之间；同时古汉语中也会出现这样的情况：虽然是虚词，但是在很多句子中可以作为查找的标记。筛选标准过分严格会影响召回率。所以停用词的构建不能单纯以实词和虚词来进行区分，而是需要综合考虑该词是否可以作为查找的具有辨识度的依据。同时停用词构建不能只考虑二程的文献情况，而是需要具有一定的普适性，希望在通过少量修改后即可在其他宋元明清思想史文献中也能发挥比较好的效果。

4 引用查找方法

从对问题本身的观察和人文领域专家经验出发，本文认为古文中的大部分引用具有较为鲜明的规律和模式，最突出的特点就是候选文本和参考文本中匹配片段的长度越长或者非连续匹配的字词个数越多，那么候选文本中包含对参考文本中内容的引用的可能性就越大。另一方面，本研究观察到对于匹配片段长度较短且数量较少的情况下，仍有许多引用存在，并且对是否包含引用的判断应该结合具体语境中的上下文信息，通过语义级别的匹配进行判断。对于结合语义的匹配，近年来在许多场景中取得了巨大突破的基于预训练语言模型的深度学习方法可以起到良好的作用。基于以上两个观察，本文提出三种分别利用专家规则、字符片段粒度语义匹配和句子粒度语义匹配的引用查找基础方法，并对三种方法结合得到不同的复合判断方法。

4.1 结合专家知识的规则方法

在本小节中首先介绍结合专家知识的规则判断方法。本研究将候选文本（比如二程集）称为源文本，将参考文本（比如十三经）称为目标文本，源文本和目标文本中连续的字符字面匹配称为直接匹配。即，设源文本为S (source)，包含的字符串为$s_1, s_2, ..., s_n$，目标文本为T (target)，包含的字符串为$t_1, t_2, ..., t_m$ 。其中n和m分别为源文本和目标文本的长度。如果S和T中存在长度为k的连续片段$s_{i+1}, \cdots, s_{i+k} = t_{j+1}, \cdots, t_{j+k}$, 那么我们认为这是一个$k$元组的直接匹配。基于人文领域专家对数据集中划定的开发集的研究，本文认为是否存在引用的关键信息来自于两个方面：

- 源文本（候选文本）和目标文本（参考文本）片段中直接匹配连续文本的长度，匹配片段的长度越长，则包含对目标引用的概率越大。比如出现了匹配的四元组文本片段，则两句有引用的概率较大；
- 源文本（候选文本）和目标文本（参考文本）片段中直接匹配的k元组个数越多，则包含对目标引用的概率越大，如匹配的文本片段个数较多，则两句有引用的概率较大。

除了以上两个原则外，本研究还发现直接匹配的文本片段中如果出现了在古文文本中大量出现的包含较少实际语义的辅助性字词，会对结果产生较大干扰。本研究将此类字词归为停用词（具体的停用词构建参见3.3小节）。

由于古文经常以单字成词，因此本文不对文本进行分词处理，而直接使用字作为单位进行匹配，匹配片段长度范围定义在$1 \sim 4$之间。本研究定义单字匹配的个数为unigram（一元组），双字匹配的个数为bigram（二元组），三字匹配的个数为trigram（三元组），四字匹配的个数为quadgram（四元组）。对这些匹配本文均只考虑不重复的组合，也就是说如果一个字符片段在源文本或目标文本中出现多次，我们只认为它起一次匹配作用。此外，对于较长片段中所包含的较短匹配（比如四元组内部一定会包含三元组），我们不进行重复计数。

以图1所举例子加以说明，源文本S为“既为先觉之民”，目标文本T为“天民之先觉者也”。S和T之间最长的匹配为双字匹配“先觉”，此时trigram和quadgram皆为0，而$bigram = 1$。单字匹配的个数除去“先”和“觉”（已经在二元组中体现）外，还有“民”和“之”，$unigram = 2$。根据本研究的经验和对开发集的观察，本研究通过以下规则来对源文本和目标文本的匹配度进行打分：

$$score_{rule} = unigram \times 0.4 + bigram \times 0.6 + trigram \times 1.4 + quadgram \times 2 \quad (1)$$

$$k-gram = k-gram \times 0.5 \quad if \quad words_{stop} \ \ in \ \ k-gram, \ n = 1, 2, 3, 4 \quad (2)$$

其中，k-gram代表连续片段的长度可以为1,2,3,4，如果k-gram中出现了停用词，那么该匹配片段的实际分数减半。如果多个停用词出现，本研究只对该k-gram分数做一次减半惩罚。

(c) 2022 中国中文信息学会计算语言学专业委员会

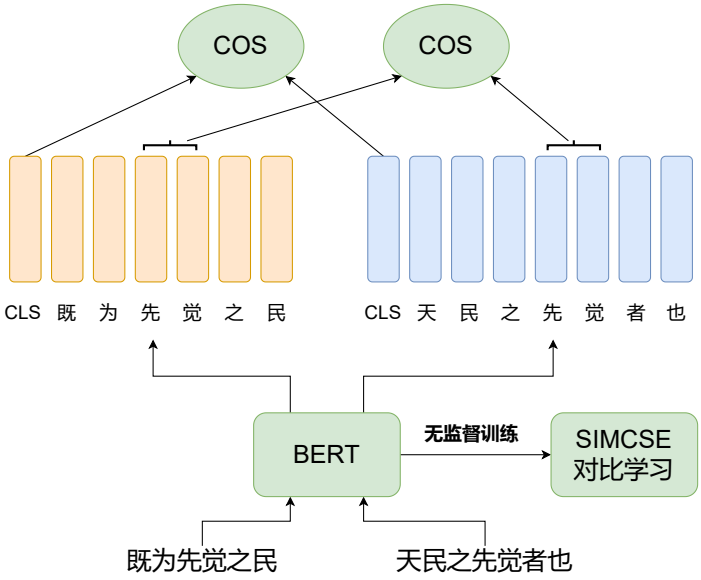


Figure 1: 基于预训练语言模型的字符片段语义和句子语义匹配方法示意图。源文本和目标文本分别经过SIMCSE模型（BERT经过对比学习无监督训练）编码后取重合的片段（字符片段）或CLS（句子）对应的向量表示进行余弦相似度计算

在上面的例子中，unigram中的“之”为停用词，也就是“之”只能算半个匹配，因此$unigram = 1.5$。这样得到$score_{rule} = 1.5 \times 0.4 + 1 \times 0.6 = 1.2$。在实际的匹配判断中，只要score大于等于1，模型就判定引用关系成立。

4.2 基于预训练语言模型的字符片段语义匹配方法

基于预训练语言模型的深度学习方法可以对上下文信息进行更深入地建模，从而更好地把握文本的语义信息。本文提出利用基于预训练语言模型和对比学习目标的深度学习方法来学习源文本和目标文本中匹配文本片段的语义表示，进而通过带有上下文语义的文本语义匹配来判断是否存在引用。相比于上文提到的基于规则的字符匹配判断方法，基于预训练语言模型的深度学习方法可以将上下文信息引入到每个共现k元组的匹配中，从而在匹配时考虑到带有上下文语义的信息。

本文采用基于对比学习的SIMCSE模型，以在古文语料上进行预训练的Sikubert（base）(王东波et al., 2022)为基础，在源文本和目标文本集上均采用掩码语言模型（masked language model）目标和SIMCSE中的对比学习目标进行适应性训练，来学习得到更贴合研究对象文本的模型参数。

如图1中所示，对任意一对待判断的源文本S和目标文本T，本研究首先在S和T的句首加入特殊标签CLS，得到$S' = [CLS, s_1, s_2, ..., s_n]$和$T' = [CLS, t_1, t_2, \ldots, t_m]$作为输入，分别通过SIMCSE模型得到S’和T’各自对每个字的表示$h^s_{CLS}, h^s_1, h^s_2, \cdots, h^s_n$和$h^t_{CLS}, h^t_1, h^t_2, \cdots, h^t_m$。其中，CLS为插入在句首的特殊符号，其对应的向量h_{CLS}代表整个句子的语义。以上图1为例加以解释，源文本S为“既为先觉之民”，目标文本T为“天民之先觉者也”。送给SIMCSE模型的输入分别为S’=[CLS，既，为，先，觉，之，民]，T’=[CLS，天，民，之，先，觉，者，也]。经过SIMCSE模型的编码，对于S’和T’中任意一个字符，都会得到一个768维（维度与所选预训练模型设定有关）的向量。该向量的每一维度均以一个实数表示。

之后，对于S和T中直接匹配的k-gram（k元组，长度为k的片段），本研究通过并联的方式得到k元组在S和T中分别的向量表示$h^s[i:i+k-1]$和$h^t[j:j+k-1]$，并计算$h^s[i:i+k-1]$和$h^t[j:j+k-1]$之间的余弦相似度。同直接匹配判断中的计算方法一样，如果k元组中出现了停用词，那么本研究对该k元组计算得到的相关度分数做减半惩罚。为了避免因为目标文本长度长而带来的匹配概率过高的问题，本研究对所有的k元组的余弦相似度求和并通过目标文本k元组数量

（$m-k-1$）进行规范化得到匹配的分数。具体计算方法如下：

$$score_k = \sum_i \frac{cos([h_i^s, h_{i+1}^s, \cdots, h_{i+k-1}^s] : [h_j^t, h_{j+1}^t, \cdots, h_{j+k-1}^t]) + 1}{2 \times (m-k-1)} \quad if\ s[i:i+k-1] = t[j:j+k-1] \tag{3}$$

注意，这里m指的是目标文本的长度，$m-k-1$代表目标文本中k元组的个数。

在上面的例子中，“先觉”是直接匹配的最长片段，此时本研究从S’和T’对应的向量中取出各自“先”和“觉”对应的向量h_3^s，h_4^s和h_4^t，h_5^t。本研究把维度均为768的两个向量h_3^s，h_4^s合并成一个1536维的向量，把同样维度均为768的两个h_4^t，h_5^t也合并成1536维的向量（本研究称这个操作为向量并联）。这样“先觉”在S’和T’两端均有了一个1536维的向量。这两个向量尽管均是代表“先觉”一词的含义，但是因为结合了S和T中不同的上下文，因此它们的向量数值并不相同。为了衡量在两种语境下“先觉”代表的语义的相关性，本研究采用余弦相似度来对这两个向量进行比较。为了使余弦相似度的范围保持在0 ∼ 1这个范围内，本研究对余弦相似度数值做了加1后除以2的规范化操作。因为只有一个二元组的匹配，这样就得到了$score_2$为上面计算得到的分数除以6，其中6为目标文本T中二元组的个数。

最后，对于不同长度的k元组得到的$score_k$本研究根据经验设定权重并进行加权求和，得到最终的基于预训练模型的字符片段语义方法判断是否存在引用的分数。在实际的设定中，一元组对应的权重为0，即因为一元组匹配带来的噪声太大，本研究不考虑一元组的语义匹配，二元组的权重为0.2，三元组的权重为0.3，四元组的权重为0.5:

$$score_{ngram} = 0 \times score_1 + 0.2 \times score_2 + 0.3 \times score_3 + 0.5 \times score_4 \tag{4}$$

4.3 基于预训练语言模型的句子语义匹配方法

在本节中，本文提出直接使用代表句子粒度语义含义的CLS特殊符号对应的经SIMCSE模型编码后的隐向量h_{CLS}匹配的方式来判断是否存在引用关系（如图1所示）。Gao等人(Gao et al., 2021)指出，通过对比学习目标获得的句子级别的表示能够更好地区分和表达不同句子之间的语义关系。具体来说，我们使用在4.2节中提到的h_{CLS}^s和h_{CLS}^t表示，并计算它们之间的余弦相似度作为句子级别的语义相似度分数:

$$score_{sentence} = \frac{cos(h_{CLS}^s, h_{CLS}^t) + 1}{2} \tag{5}$$

4.4 复合判断方法

为了结合前面提到的三种基本判断方法的优点，本文提出使用基于三种判断方法的复合判断方法。按照组合关系，我们共得到四种复合模型：规则+字符，规则+句子，字符+句子，规则+字符+句子。其中，“规则”指4.1节中提到的结合专家知识的文本匹配规则方法，“字符”指4.2节中提到的基于预训练语言模型的字符片段语义匹配方法，“句子”指4.3节中提到的基于预训练语言模型的句子语义匹配方法。在复合判断方法中，本研究将三种方法得到的判断分数进行加权平均。在实际的计算中，为了简单起见，本研究将不同方法各自的分数均按照0.5的权重进行平均。并根据在开发集上的实验结果设定一个阈值来实际判断是否存在引用关系。

5 实验

5.1 实验设定

本文采用的BERT模型是在四库全书古文语料上进行预训练的Sikubert[2]（base）。其中的隐藏层向量维度是768，模型层数是12层，多头注意力机制（multi-head attention）的头数是12。在本文相关数据集上继续训练（SIMCSE方法和掩码语言模型）的轮数是5轮。

[2]https://github.com/hsc748NLP/SikuBERT-for-digital-humanities-and-classical-Chinese-information-processing

(c) 2022 中国中文信息学会计算语言学专业委员会

5.2 评价指标

本文采用正确率（accuracy）、准确率（precision）、召回率（recall rate）、F1值以及ROC-AUC值作为实验结果的评价指标。并以ROC-AUC作为主要评价标准。

正确率（accuracy）是衡量模型表现的一个常用指标，它的定义为：

$$ACC = \frac{right\ number}{total\ number}$$

F1值的计算依赖于准确率和召回率，F1值是对准确率和召回率的几何平均，反映了在某个阈值下相对平衡的模型表现。这三个指标的计算方式如下：

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

其中，TP代表True Positive，FP代表False Positive，FN代表False Negative。

但是，无论是正确率还是F1值等都会受到分类器中阈值选取的影响，因而，ROC-AUC指标常被用来衡量模型在不同阈值下的实际表现。ROC-AUC值根据预测分数计算接收器工作特性曲线下的面积。这个计算指标可以对于二值分类中不同的阈值取值下，得到不受阈值影响的分类器性能，本研究选取ROC-AUC这个指标来作为衡量模型实际分类性能的主要指标。理论上，以上指标都是越高越好。

5.3 判断的两种粒度：切分后的单句和自然段落

考虑到人文领域研究者对于结果的实际需求，除了上面介绍的对拆分后的单句级别文本进行是否包含引用的判断，本研究还提出按照通行本文献的分段方式进行引用判断。为了避免段落长度对判断模型的干扰，本研究采取一种简单直接的方法来判定段落中是否含有引用，即如果段落中任意一个单句级别的文本被判断为包含引用，那么本研究判定整段文本包含引用。

5.4 实验结果

在表1中，本文给出单一基线方法在测试集上以单句和段落为单位的模型表现。段落级别也就是判断一段中至少存在一处引用，这种设定更加接近人文领域的实际需求。从表中数据可以看出，基于预训练语言模型不同语义粒度的两种方法（字符、句子）单独使用均好于基于规则的方法，且“字符”和“句子”两种关注不同语义粒度的方法表现大体相当。在以段落为单位进行判断时，各项指标相比以单句为单位时均有所上升，尤其是F1值上升更为明显。单一方法的F1值即可达到80左右。

在表2中，本研究进一步给出使用复合方法得到的结果。当将三种单一基线方法结合，得到复合基线方法后，可以看到，效果均有较大幅度提升。说明三种单一方法之间关注的语义层次均具有互补之处。但是将三种方法同时结合在一个复合模型中的效果则不如只使用两种基于预训练语言模型的方法。此外可以看出，在以段落为单位时，将两种基于预训练语言模型的方法复合起来可以获得最好的效果，达到超过90的ROC-AUC值（91.02），且正确率达到了83.31，F1也超过了86，达到了能够实际帮助人文学者得出可靠结论的水平。而再结合规则后，效果反而会有所下降，这与基于预训练语言模型的方法具有更好的泛化性和迁移性有关。

5.5 实验分析

5.5.1 使用对比学习目标的效果

在图5.5.1中我们展示了使用不同模型在测试集上两种粒度（单句、段落）下的ROC-AUC值。其中，Sikubert指直接使用原始Sikubert预训练语言模型得到两种粒度的表示，Fine Tune指在领域内（二程和十三经文本）使用掩码语言模型进行领域适应性训练5轮的模型，SIMCSE指同时使用掩码语言模型和基于dropout的对比学习在领域内文本进行适应性训练5轮的模型。目标匹配方法使用表现最好的“字符+句子”方法。

从图中可以看出，在领域内进行适应性训练在单句级别上的效果提升尤其明显，因为这种复合方法中的字符片段级别匹配依赖于每个字符的正确学习和表示，而在领域内进行继续训练可以使得对字符的表示更贴近领域真实的分布。此外，可以看出，使用带有对比学习目标的SIMCSE模型相比单纯使用掩码语言模型作为目标的模型又有显著提高，这也验证了对比学习目标的有效性。

(c) 2022 中国中文信息学会计算语言学专业委员会

单句	Acc	Precision	Recall	F1	ROC-AUC
规则	56.40	45.20	90.75	60.34	64.01
字符	**76.49**	**68.79**	65.31	**67.01**	**81.65**
句子	72.83	60.26	**75.33**	66.96	81.57
段落	Acc	Precision	Recall	F1	ROC-AUC
规则	64.93	63.47	96.92	76.70	57.41
字符	**76.72**	**79.70**	81.75	80.71	**84.93**
句子	76.11	75.83	**87.92**	**81.43**	83.52

Table 1: 单句和段落级别**单一**基线方法对引用判断的结果。字符代表基于预训练语言模型字符片段语义的判断方法，句子代表基于预训练语言模型句子语义的判断方法。

单句	Acc	Precision	Recall	F1	ROC-AUC
规则+字符	75.56	64.43	74.01	68.89	82.54
规则+句子	75.64	63.37	79.07	70.36	83.61
字符+句子	**80.64**	**72.40**	75.99	**74.15**	**87.83**
规则+字符+句子	79.15	68.61	**79.19**	73.52	85.99
段落	Acc	Precision	Recall	F1	ROC-AUC
规则+字符	77.64	76.59	89.97	82.74	86.18
规则+句子	77.18	75.53	91.26	82.65	85.63
字符+句子	**83.31**	**82.56**	91.26	**86.69**	**91.02**
规则+字符+句子	81.01	79.51	**91.77**	85.20	89.99

Table 2: 单句和段落级别**复合**方法对引用判断的结果。字符代表基于预训练语言模型字符片段语义的判断方法，句子代表基于预训练语言模型句子语义的判断方法。

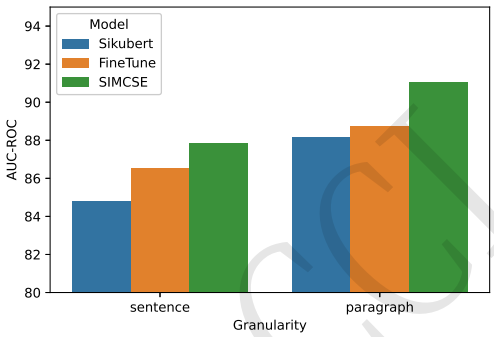


Figure 2: 不同模型在句子级别和段落级别的ROC-AUC值，SIMCSE为加入对比学习训练目标后的方法，匹配方法使用表现最好的“字符+句子”方法。

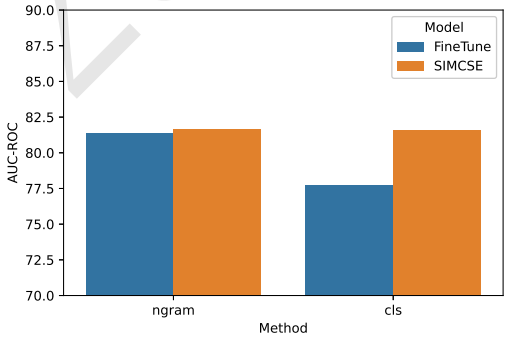


Figure 3: 使用对比学习目标在两种方法下的差别。ngram代表基于预训练语言模型的字符片段语义匹配方法，cls代表基于预训练语言模型的句子语义匹配方法。

从图5.5.1中可以进一步看出，加入对比学习目标后，相比于只做掩码语言模型训练，提升主要体现在句子级别的语义提升上。而在字符片段级别的语义上基本没有改变。这是因为SIMCSE中的对比学习目标主要针对的是整句文本表示的学习，而对字符级别影响不大。我们还可以看出，加入对比学习目标前，使用句子级别的语义进行判断效果与使用字符片段相比有较大差距，而在SIMCSE加入对比学习目标后，使用两种级别的语义判断效果基本相当。

5.5.2 使用停用词的效果

在图5.5.2中，我们展示了是否使用停用词惩罚的ROC-AUC效果差别。可以看出，尽管差距不大，但是使用停用词惩罚后，无论在单句粒度上还是段落粒度上均有稳定的提升。这说明引入带有人文专家经验的停用词可以在一定程度上提升引用查找的效果。在图5.5.2中，我们进

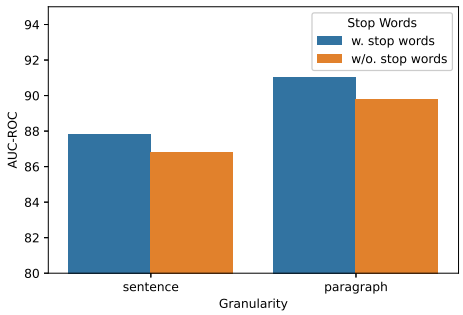


Figure 4: 是否使用停用词惩罚的ROC-AUC效果。引用查找方法选择的是基于预训练语言模型的“字符+句子”方法。

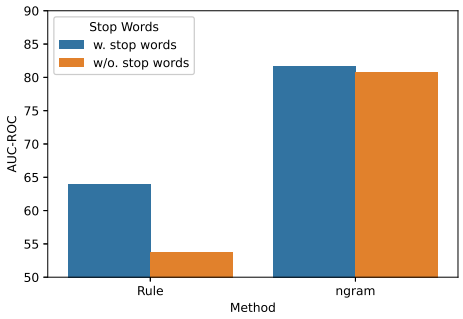


Figure 5: 对于字符片段方法和规则方法是否使用停用词ROC-AUC效果。

二程（正确）	得此道而不者，仁者之事也；因其不，故曰此仁也
经典（论语）	子曰：君子道者三，我能焉：仁者不，知者不惑，勇者不
二程（错误）	古人祭祀用尸，有深意，不可不深思
经典（礼记）	子云：祭祀之有尸也，宗之主也，示民有事也
二程（错误）	畏天命，可以不失付畀之重
经典（论语）	孔子曰：君子有三畏：畏天命，畏大人，畏人之言

Table 3: 字符片段和句子语义复合引用方法判断的代表性样例

一步给出了规则方法（Rule）中和基于预训练语言模型的字符片段语义匹配方法（ngram）中是否使用停用词的差别，从图中可以看到，规则方法对停用词更加敏感，而尽管影响较小，字符片段匹配方法也会在一定程度上受到是否停用词的影响。

5.6 样例分析

在表3中，我们给出了几个具有代表性的例子。在第一个例子中，经典（论语）原文是“仁者不（忧）”，而二程的解释里并没有出现这四个字，而是分开解释了“忧”与“仁者”之间的关系。单看“仁”这个概念的话，实际上会在经典文献和二程文本中大量出现，属于儒家思想的核心概念。无论对于专家人工检索（关键词“仁”和“忧”均存在大量干扰项）还是使用规则方法均存在很大的挑战。我们提出的基于预训练语言模型和对比学习目标的方法因为能够对整个句子的语义进行建模而很好地解决了这个问题。

对于后面两个例子来说，模型给出了错误的预测。对于二程文献中提到的“祭祀用尸”在先秦两汉文章中多次提到过，这里二程可能是针对某一处经典所发的议论，也有可能泛泛地使用“祭祀用尸”这个说法。即便在人工的引用查找过程中，这类句子也要结合上下文来综合判断，对模型来说难度也非常大。这部分引用查找模型只能给出可能存在的引用，而需要继续由专家校对完成判断。对于第三个例子来说，关键信息“天命”在多数句子中都能担任辨识的依据，所以不能作为停用词处理，但是“天命”这种说法在二程文献和经典文献中都非常多，在很多情境中其存在可以视为噪音，也容易给判断造成干扰。

6 结论

本文主要研究了如何自动化探测中国古代思想家对早期经典文献的引用，并给出了引用的明确定义。本文提出了多种结合专家知识和基于预训练语言模型和对比学习目标的无监督基线方法来自动查找中国古代思想史中思想家在阐发思想时对早期文献的引用。为了验证方法的有效性，本文以二程对早期儒家经典文献的引用为例，构建并发布二程对早期儒家经典引用的数据集并在该数据集上进行了大量实验。实验结果表明本文提出的基于预训练语言模型的字符片段和句子复合方法可以有效地找到大多数引用，并且能够为提高专家人工精确查找效率提供有效帮助。本文的研究成果在集注集释整理、文本生成溯源、重出文献查找、引用统计分析、索引文献集制作等方面具有广阔的应用前景。

致谢

本成果受国家自然科学基金项目（61872402）,教育部人文社科规划基金项目(17YJAZH068)，中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)，模式识别国家重点实验室开放课题基金资助。

参考文献

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 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 *Empirical Methods in Natural Language Processing (EMNLP)*.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *CoRR*, abs/2011.00362.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

俞敬松, 魏一, and 张永伟. 2019. 基于bert的古文断句研究与应用. 中文信息学报, 33(11):7.

刘江峰, 冯钰童, 王东波, 胡昊天, and 张逸勤. 2021. 数字人文视域下sikubert增强的史籍实体识别. 图书馆论坛.

周好, 王东波, and 黄水清. 2021. 古籍引书上下文自动识别研究——以注疏文献为例. 情报理论与实践, 44(9):7.

徐润华, 王东波, 刘欢, 梁媛, and 陈康. 2022. 面向古籍数字人文的《资治通鉴》自动摘要研究——以sikubert预训练模型为例. 图书馆论坛.

王东波, 刘畅, 朱子赫, 刘江峰, 胡昊天, 沈思, and 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.

耿云冬, 张逸勤, 刘欢, and 王东波. 2022. 面向数字人文的中国古代典籍词性自动标注研究——以sikubert预训练模型为例. 图书馆论坛.

胡昊天, 张逸勤, 邓三鸿, 王东波, 冯敏萱, 刘浏, and 李斌. 2022. 面向数字人文的《四库全书》子部自动分类研究——以siku bert和siku ro berta预训练模型为例. 图书馆论坛.

葛瑞汉. 2000. 二程兄弟的新儒学:中国的两位哲学家. 二程兄弟的新儒学:中国的两位哲学家.

黄俊杰. 2018. 东亚儒家经典诠释史中的三个理论问题. 山东大学学报：哲学社会科学版, (2):8.

黄水清, 周好, 彭秋茹, and 王东波. 2021. 引书的自动识别及文献计量学分析. 情报学报, 40(12):13.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于批数据过采样的中医临床记录四诊描述抽取方法

王亚强[1,2,3†]，**李凯伦**[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

中医临床记录四诊描述抽取对中医临床辨证论治的提质增效具有重要的应用价值，然而该抽取任务尚有待探索，类别分布不均衡是该任务的关键挑战之一。本文围绕该任务展开研究，构建了中医临床四诊描述抽取语料库；基于无标注中医临床记录微调通用预训练语言模型实现领域适应；利用小规模标注数据，采用批数据过采样算法，实现中医临床记录四诊描述抽取模型的训练。实验结果表明，本文提出方法的总体性能均优于对比方法，与对比方法的最优结果相比，本文提出的方法将少见类别的抽取性能F1值平均提升了2.13%。

关键词： 四诊描述抽取；类别分布不均衡；批数据过采样；临床记录；中医

Four Diagnostic Description Extraction in Clinical Records of Traditional Chinese Medicine with Batch Data Oversampling

Yaqiang Wang[1,2,3†], **Kailun Li**[1,2,3], **Yongguang Jiang**[4], **Hongping Shu**[2,3]
[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

Four diagnostic description extraction in clinical records of traditional Chinese medicine (TCM) has important application value in improving the quality and efficiency of TCM clinical syndrome differentiation and treatment. However, the extraction task is yet to be explored, and imbalanced class distribution is one of the key challenges of this task. As a first exploration of this task, we firstly constructed a TCM clinical four diagnostic description extraction corpus and then solved the domain adaptation by fine-tuning the general domain pre-trained language model based on unlabeled TCM clinical records. At last, we trained our proposed four diagnostic description extraction model by utilizing a small labeled dataset through a well-designed batch data oversampling algorithm. The experimental results show that the performance of the proposed method in this paper is better than that of the compared methods, and the proposed method improves the extraction performance F1 score of the rare class by 2.13% on average.

Keywords: Four diagnostic description extraction , Imbalanced class distribution , Batch data oversampling , Clinical records , Traditional Chinese medicine

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

1 引言

辨证论治，又称辨证施治，是中医特有的一种对疾病研究、处理、认知和治疗的基本原则与方法(印会河, 2005)。辨证是论治的前提和依据，四诊（即望、闻、问、切）信息是中医专家综合分析病人的病情，认知疾病，最终辨清证型的重要参考(李红岩 et al., 2022)。快速、准确地获取中医临床记录中的四诊信息，对提升中医专家辨证和诊疗的效率与质量以及为中医临床辅助辨证提供更丰富的医学语义信息具有重要的价值(屈丹丹 et al., 2021)。

在四诊信息中，局部的、具体的疾病、症状、脉象、舌质等实体信息的抽取已有广泛研究。Wang等人(2012)基于条件随机场等统计序列标注模型首次尝试从中医临床记录中抽取症状信息。肖瑞等人(2020)围绕中医临床记录中的疾病和症状信息抽取，采用深度学习模型展开研究。然而，面向全局的、叙述性的中医临床记录四诊描述的抽取还未见相关报道。

中医临床记录中的四诊描述不仅包含局部的、具体的实体修饰信息。如图1所示，实体的“有”或“无”、时间的“长”或“短”、情况的“重”或“轻”等，还蕴含着实体之间的关联信息，如实体之间的因果关系、并列关系等。因此，中医临床记录四诊描述抽取的结果将形成对实体信息抽取研究的补充，为下游任务提供更丰富的医学语义信息。

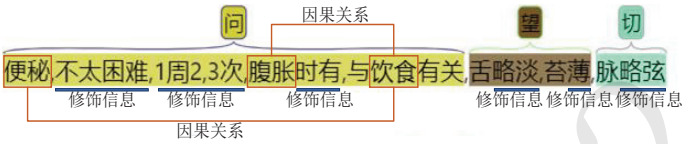


图 1. 中医临床记录中的四诊描述包含的修饰和关联语义信息

与中医临床记录中的实体信息抽取任务不同，中医临床记录四诊描述抽取任务具有其特殊性。首先，与实体的字面值相比，四诊描述的文本长度通常较长，会带来更强的稀疏性。在本文的实验数据集中，每段四诊描述平均包含12个字[1]。此外，如图2所示，通过对不同的四诊描述进行计数发现，四诊描述呈现长尾分布。

其次，由于中医专家的临床实践习惯不同，使得四诊描述天然存在类别分布不均衡的问题。一般地，望诊、问诊、切诊被中医专家更广泛地在临床实践中使用，而闻诊的使用相对较少。基于本文实验数据统计发现（如图3所示），中医临床记录中包含望诊和切诊描述的实例数量少于包含问诊描述的实例数量，而包含闻诊描述的实例数量相较于其他三诊描述格外稀少。

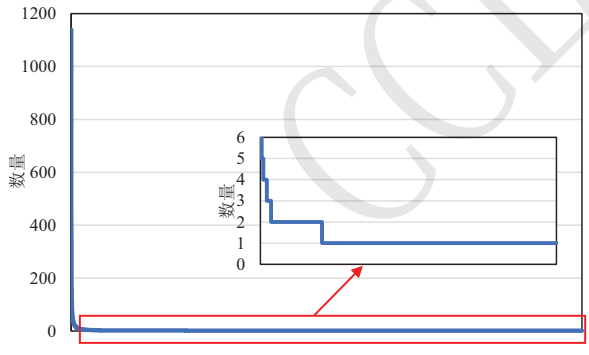


图 2. 不同的中医四诊描述的计数结果

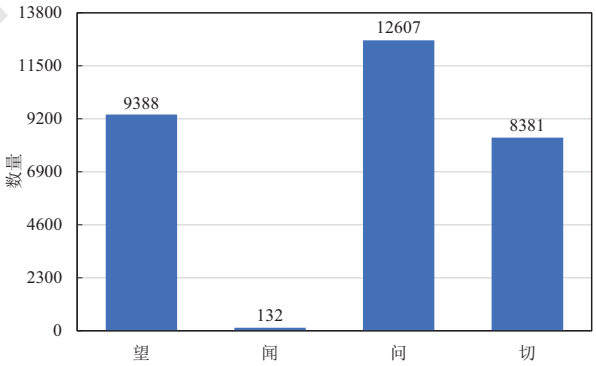


图 3. 中医临床记录中的四诊描述分类计数

因此，本文首次开展了中医临床记录四诊描述抽取任务的探索，针对中医临床记录中的四诊描述的特点，在最新的模型框架基础上，设计并验证了相应的改进策略与算法，取得以下具体成果：

首先，本文将中医临床记录四诊描述抽取定义为基于字的序列标注任务，采用广泛使用的“BIO”标注模式(Irsoy and Cardie, 2014)，提出基于BERT+BiLSTM+CRF(Dai et al., 2019)的中医临床记录四诊描述序列标注模型。在该模型中，利用BERT的动态上下文语义嵌入

[1]本文将中医临床记录中的标点符号也均视为字

(c) 2022 中国中文信息学会计算语言学专业委员会

学习能力和多头注意力机制(Vaswani et al., 2017)，实现对中医临床记录四诊描述的文本语义信息增强，进而在数据稀疏条件下，保证四诊信息抽取的性能。

其次，采用在无标注的中医临床记录数据上微调通用领域BERT的方法(Gururangan et al., 2020)，验证BERT在进行领域适应后对中医临床记录四诊描述序列标注性能的影响。实验结果发现，该方法有助于提升中医临床记录四诊描述的整体标注性能。对于各类描述的标注结果来说，该方法对“I-望”、“I-闻”、“I-问”、“I-切”等标签的标注有更积极的促进作用。

第三，提出基于批数据过采样的模型训练算法，提升模型对少见的四诊描述类别的标注性能。该算法在基于小批量梯度下降算法（Mini-Batch Gradient Descent，MBGD）(Ruder, 2016)的中医临床记录四诊描述序列标注模型训练框架基础上，通过过采样包含少见类别的数据实例，实现在每轮随机划分的批量数据中，策略性地增加对少见类别数据的学习关注。该方法在实现序列标注模型对常见四诊描述类别的标注性能提升的基础上，大幅提升了少见类别的标注性能。

实验结果表明，本文提出的基于批数据过采样的中医临床记录四诊描述抽取方法的效果优于HMM(Rabiner, 1989)、CRF(Lafferty et al., 2001)、BiLSTM和BiLSTM+CRF(Lample et al., 2016)等对比模型。与对比模型在本文任务上的最佳性能相比，本文方法的标注性能F1值平均提升了1.37%。特别地，本文方法大幅提升了少见类别的标注性能F1值，“B-闻”和“I-闻”标签的F1值分别达到了62.22%和61.54%，相比最佳的对比方法平均有2.13%的提升。

2 相关工作

2.1 中医临床记录信息抽取

中医临床记录信息抽取是近年来中医信息化领域广泛研究的课题。Zhang等人(2022)综述了从2010年至今中医文本信息抽取的相关工作，中医临床记录信息抽取是其中的重要任务之一。目前，中医临床记录信息抽取主要针对疾病、症状、体征、诊断、方剂、药物等局部的、具体的实体信息抽取任务展开，针对包含丰富语言学和临床语义信息的中医临床记录四诊描述抽取的研究甚少。因此，本文开展了中医临床记录四诊描述抽取任务的探索研究。

与一般领域的信息抽取任务相同，中医临床记录信息抽取通常采用序列标注方法实现(Wang et al., 2014)。该类方法将抽取任务转换为序列标注任务，通过对中医临床记录中的基本语义单元进行分类实现对连续的基本语义单元构成的目标类别信息的抽取。其中，语义单元一般为中文字，分类标签通常由待抽取的信息分别定义的BIO标签形成，B表示待抽取的语义单元在待抽取信息的开始位置，I表示待抽取的语义单元在待抽取信息的中间和结束位置，O表示非待抽取的语义单元(Irsoy and Cardie, 2014)。作为初步探索工作，本文沿用了该语义单元和分类标签定义方法。

2.2 序列标注模型

HMM、CRF是被广泛使用的统计序列标注模型，在训练数据规模不大的情况下，因模型复杂度相对较低，它们通常能够取得与深度序列标注模型相当的标注性能(Nasar et al., 2021)。作为中医临床记录四诊描述抽取任务的初探，本文在自建数据集上验证了HMM和CRF的性能，并将它们作为基线模型与目前被更广泛应用的深度序列标注模型BiLSTM+CRF进行比较。

当前，深度序列标注模型在各项信息抽取任务（包括中医临床信息抽取任务）上都取得了优秀的性能，BiLSTM+CRF是其中的代表(Lample et al., 2016)。因此，本文将其作为SOTA基线模型应用于中医临床记录四诊描述抽取任务。此外，BERT能够基于上下文信息，利用多头注意力机制，获取当前待标注语义单元的多角度的丰富的语义信息，动态地形成该语义单元的词嵌入，从而提升下游预测任务模型的性能。因此，本文采用BERT+BiLSTM+CRF来解决中医临床记录四诊描述抽取任务由于数据稀疏带来的语义模糊问题。

BERT是利用通用领域大规模数据集训练得到的预训练模型(Devlin et al., 2018)，其生成的词嵌入携带的是通用语义信息。中医临床记录四诊描述抽取任务的待标注语义单元具有中医领域特殊含义，其上下文蕴含中医领域特殊语义。为更好地适应中医领域的特殊语义表达，借鉴Zhang等人(2020)方法的思想，本文利用中医临床记录数据在MC-BERT的基础上进行微调，以期获得能够更好地表达中医临床记录语义的预训练语言模型。

(c) 2022 中国中文信息学会计算语言学专业委员会

2.3 不均衡类别分布学习

数据采样是在不均衡类别分布学习中广泛采用的方法之一(刘树栋 and 张可, 2019)。该方法主要通过设计特殊的采样策略，如过采样、欠采样或过采样与欠采样融合，改变数据集的类别分布，达到数据集类别分布均衡的目标。其中，过采样算法是在数据有限条件下，更多被使用的数据采样方法。中医临床记录四诊描述存在类别分布不均衡问题，通常特定领域任务的数据规模有限，因此，本文将数据过采样方法应用到BERT+BiLSTM+CRF的模型训练过程。

BERT+BiLSTM+CRF的模型训练主要采用MBGD框架完成，该框架的参数学习过程的核心是基于每一组批数据估计梯度(Ruder, 2016)。类别分布不均衡会直接导致各组批数据中包含少见类别数据的可能性低，进而导致少见类别学习不充分。为了让模型在训练的过程中更多地关注少见类别，借鉴数据过采样方法(Lin et al., 2017; Shahee and Ananthakumar, 2018)，本文通过过采样少见类别数据，实现在每轮随机划分的批数据中，策略性地增加对少见类别数据的学习，进而达到模型在训练过程中充分学习少见类别数据的目标。

3 方法

3.1 任务定义

中医临床记录四诊描述抽取任务可归结为序列标注任务，因此任务可形式化定义为：给定一条中医临床记录$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$，目标是训练一个序列分类器，该序列分类器将顺序地预测输入序列$\boldsymbol{x}$中第i个文字x_i对应的标签y_i。本文采用“BIO”标注模式，因此有y_i属于预定义的标签集合$\boldsymbol{L}$ = {O, B-望, I-望, B-闻, I-闻, B-问, I-问, B-切, I-切}。给定训练数据集，中医临床记录四诊描述抽取任务的模型优化目标为：

$$\hat{\boldsymbol{y}} = argmax_{\boldsymbol{y}}\, \mathrm{P}\left(\boldsymbol{y} \mid \boldsymbol{x}\right) \tag{1}$$

3.2 模型

本文以BERT+BiLSTM+CRF模型为基础实现中医临床记录四诊描述抽取，该模型的基础框架如图4所示。

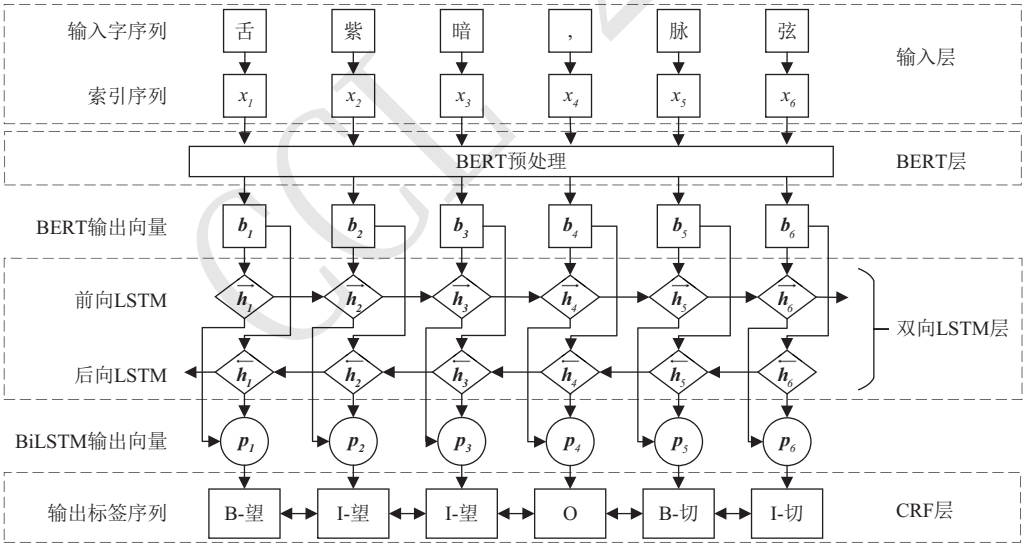


图 4. BERT+BiLSTM+CRF模型基础框架图

如图4的输入层所示，中医临床记录以字为基本单元进行切分，并将切分后的字替换为BERT词表中对应的索引值$x_1 \sim x_6$，形成索引序列。

输入字的索引序列经过图4的BERT层特征提取，得到包含丰富的上下文语义信息的字向量$\boldsymbol{b_1} \sim \boldsymbol{b_6}$。多头注意力机制是BERT模型最关键的部分。在BERT层中，注意力机制实质上是通过字序列的字与字之间的关联程度调整权重系数矩阵，从而获得字序列中所有的字在引入上下文信息后的语义表征向量，其计算公式如(2)所示。

(c) 2022 中国中文信息学会计算语言学专业委员会

$$Attention\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$为BERT的Embedding层输出的所有字向量经过不同的线性变换后得到的加权矩阵，d_k为字向量的维度。多头注意力机制从不同的角度学习输入序列中的上下文语义信息，均衡单一注意力机制可能产生的偏差，给字向量注入更多元的上下文语义信息，其公式如式(3)和式(4)所示。

$$MultiHead\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=Concat\left(\boldsymbol{head_1},\ldots,\boldsymbol{head_n}\right)\boldsymbol{W^O} \tag{3}$$

$$\boldsymbol{head_i}=Attention\left(\boldsymbol{Q}\boldsymbol{W_i^Q},\boldsymbol{K}\boldsymbol{W_i^K},\boldsymbol{V}\boldsymbol{W_i^V}\right) \tag{4}$$

其中，$\boldsymbol{W_i^Q}$、$\boldsymbol{W_i^K}$、$\boldsymbol{W_i^V}$为多头注意力机制中第$\boldsymbol{i}$个注意力头的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$所对应的权重矩阵，$\boldsymbol{W^O}$为多头注意力拼接后的线性变换所需的权重矩阵。

在图4中，BiLSTM层的前向过程和后向过程的LSTM单元可以舍弃当前时刻输入字向量的无用信息，并将当前时刻输入字向量的有用信息传递到下一时刻的LSTM单元。然后，将双向过程中每个时刻对应的输出拼接，如公式(5)所示，得到包含长距离上下文信息的字向量$\boldsymbol{p_1}\sim\boldsymbol{p_6}$。

$$\boldsymbol{p}_t=\left[\overrightarrow{\boldsymbol{h}_t},\overleftarrow{\boldsymbol{h}_t}\right] \tag{5}$$

其中，$\overrightarrow{\boldsymbol{h}_t}$为前向过程的LSTM单元在时刻$t$的输出，$\overleftarrow{\boldsymbol{h}_t}$为后向过程的LSTM单元在时刻$t$的输出。

最后，在图4的CRF层中，CRF模型利用邻近标签之间的依赖关系对BiLSTM层输出的所有字向量进行解码，解码目标如式(6)所示，从而得到最优的预测序列。

$$\boldsymbol{Y}^*=\underset{\widetilde{\boldsymbol{Y}}\in\boldsymbol{Y_X}}{argmax}\ s\left(\boldsymbol{X},\widetilde{\boldsymbol{Y}}\right) \tag{6}$$

在公式(6)中，$\boldsymbol{Y_X}$表示所有可能的标注序列，$\boldsymbol{Y}^*$表示解码后获得最大评分的输出序列，s表示标注序列对应的分数函数。

3.3 模型训练方法

3.3.1 模型训练流程

如章节1中所述，中医临床记录四诊描述任务存在严重的类别分布不均衡问题，闻诊描述的数量远少于其它三诊描述的数量。直接利用具有该特点的训练数据对BERT+BiLSTM+CRF模型进行训练，将使模型对训练数据中较少训练数据对应的类别学习不充分，进而影响该类别的预测性能。为克服上述问题，本文设计了基于批数据过采样的小批量(mini-batch)梯度下降算法训练BERT+BiLSTM+CRF模型，以期在一定程度上缓解类别分布不均衡对中医临床四诊描述抽取模型性能的影响，算法训练模型的流程如图5所示。

在如图5的模型训练流程中，主要包含六个关键的处理步骤。

(1) 批数据过采样：在数据处理过程中，按批量大小M将训练数据集$\boldsymbol{D}$划分为包含$\lfloor|\boldsymbol{D}|/M\rfloor$个批量的批量集合$\boldsymbol{B}$。然后，使用批数据过采样的方式增加批量中闻诊信息的数量，生成批量集合$\boldsymbol{B'}$，用于模型训练，从而提高模型对于闻诊信息的抽取性能（此步骤将在3.3.3节中详细介绍）。

(2) 模型参数$\boldsymbol{\theta}_0$初始化：该步骤完成BERT+BiLSTM+CRF模型的初始化参数$\boldsymbol{\theta}_0$的设置。其中，BERT模型的参数是在一定规模的无标注中医临床记录数据上微调得到，该方法参见3.3.2节，BiLSTM模型和CRF模型的初始化参数为随机生成，服从均匀分布。

(3) 损失计算：该步骤将计算模型在当前批量包含的数据样本上的平均损失值。其中，$f_{\boldsymbol{\theta}_k}(\boldsymbol{x}'_i)$代表模型以当前批量中第$i$个数据样本$\boldsymbol{x}'_i$作为输入，且此时模型的参数为第$k$次迭代的参数$\boldsymbol{\theta}_k$。

(c) 2022 中国中文信息学会计算语言学专业委员会

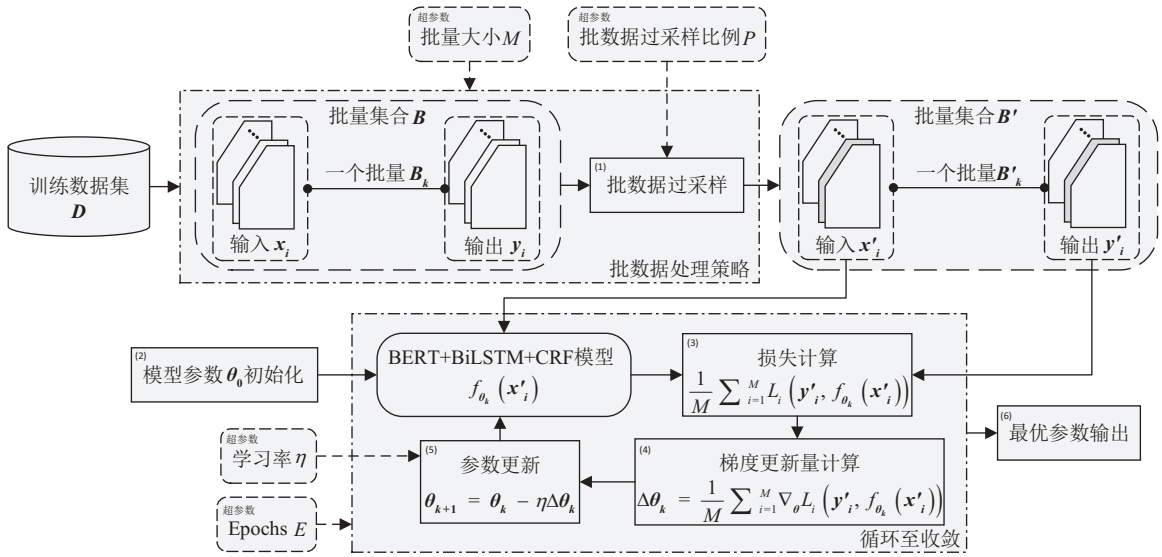


图 5. 基于批数据过采样的小批量梯度下降算法训练BERT+BiLSTM+CRF模型的流程图

(4) 梯度更新量计算：通过误差反向传播算法，当前轮批量$\boldsymbol{B}'_k$中所有数据样本计算梯度的平均值作为模型在第k轮迭代时的梯度更新量$\Delta\boldsymbol{\theta}_k$。

(5) 参数更新：基于当前轮迭代过程中的模型参数$\boldsymbol{\theta}_k$、梯度更新量$\Delta\boldsymbol{\theta}_k$和学习率$\eta$计算第k+1轮迭代过程的模型参数$\boldsymbol{\theta}_{k+1}$

(6) 最优参数输出：步骤(2)到步骤(5)循环执行$\lfloor|\boldsymbol{D}|/M\rfloor * E$轮（$E$为对数据集$\boldsymbol{D}$遍历的轮数），直到模型收敛，输出模型在收敛处的最优参数。

上述的步骤中，M、η、E均为模型训练过程中的超参，它们在本文实验中的取值参见4.3节。

算法 1： 批数据过采样算法

输入： 训练数据集 $\boldsymbol{D}$，闻诊信息数据集 $\boldsymbol{W}$，批数据过采样比例 P，批量大小 M，数据集洗牌函数 *shuffle*，数据集划分函数 *split*，数据样本随机移除函数 *remove*，数据样本随机选取函数 *select*，数据样本添加函数 *append*，批量添加函数 *add*

输出： 过采样闻诊信息后的批量集合 $\boldsymbol{B}'$

$\boldsymbol{S}$ = *shuffle*($\boldsymbol{D}$)　// 对训练数据集 $\boldsymbol{D}$ 进行洗牌操作

$\boldsymbol{B}$ = *split*($\boldsymbol{S}$, M)　// 将洗牌后得到的数据集 $\boldsymbol{S}$ 按批量大小 M 切分为批量集合 $\boldsymbol{B}$

$N = \lfloor|\boldsymbol{D}|/M\rfloor$　// 得到批量集合 $\boldsymbol{B}$ 中批量的数量 N

for $k = 1, 2, ..., N$ **do**

　　$\boldsymbol{B}_k$ = *remove*($\boldsymbol{B}_k$, P, M)　// 从批量 $\boldsymbol{B}_k$ 中随机移除$\lfloor P\times M\rfloor$个数据样本

　　$\boldsymbol{A}_k$ = *select*($\boldsymbol{W}$, P, M)　// 从闻诊信息数据集 $\boldsymbol{W}$ 中有放回地随机选取$\lfloor P\times M\rfloor$个数据样本

　　$\boldsymbol{B}'_k$ = *append*($\boldsymbol{B}_k$, $\boldsymbol{A}_k$)　// 将 $\boldsymbol{A}_k$ 加入批量 $\boldsymbol{B}_k$ 中，得到过采样闻诊信息后的批量 $\boldsymbol{B}'_k$

　　$\boldsymbol{B}'$ = *add*($\boldsymbol{B}'$, $\boldsymbol{B}'_k$)　// 将批量 $\boldsymbol{B}'_k$ 加入到批量集合 $\boldsymbol{B}'$中

end for

图 6. 批数据过采样算法

3.3.2 领域适应方法

为使通用领域的BERT预训练语言模型所生成的词嵌入携带更丰富的中医临床语义信息，使其更适用于中医临床记录四诊描述抽取任务，本文借鉴了关于特定领域BERT的领域适应方法的相关工作(Lee et al., 2020; Gururangan et al., 2020)的基本做法。

在Zhang等人(2020)提出的中文医疗预训练语言模型MC-BERT的基础上，使用领域内的无标注中医临床记录数据集，对MC-BERT的掩码语言模型进行微调，使其可以更好地适应本文

(c) 2022 中国中文信息学会计算语言学专业委员会

任务领域的语义表达。在领域适应的过程中，更新的掩码语言模型$f_{LM}(\cdot;\theta_{enc},\theta_{LM})$的参数包括从MC-BERT模型上初始化的编码器参数θ_{enc}和分类头参数θ_{LM}。

3.3.3 批数据过采样

由于带标注的中医临床记录数据集存在严重的类别分布不均衡问题，如图3所示，数据集中闻诊描述的数量远少于其他三诊描述的数量，这会严重影响模型对于闻诊描述的抽取性能。为解决这个问题，本文提出在利用小批量梯度下降算法训练四诊信息序列标注模型的过程中，采用批数据过采样的方式去增加批量中闻诊信息的数量，从而在一定程度上消除类别分布不均衡问题对模型抽取性能的影响。批数据过采样的伪代码如图6所示，其中批数据过采样比例P为超参数，其取值参见4.3节。

在图6的批数据过采样算法中，闻诊信息数据集$\boldsymbol{W}$由训练数据集$\boldsymbol{D}$中所有包含闻诊信息的数据样本构成，将在4.1节中具体介绍。并且，批数据过采样在模型训练过程中的每个Epochs中都会执行一次。

4 实验

在测试数据集上，本文将所提出的方法与HMM、CRF、BiLSTM、BiLSTM+CRF等模型进行了比较。本章节后续将依次具体介绍实验中使用的数据集、评价指标、实验设置，以及实验得到的结果。

4.1 数据集

数据集	标签数量	样本数量	抽取信息的数量			
			望	闻	问	切
无标注中医临床记录数据集	-	11251	-	-	-	-
带标注中医临床记录数据集	9	10594	9388	132	12607	8381
训练数据集	9	6346	5652	82	7570	5028
验证数据集	9	2124	1881	28	2545	1661
测试数据集	9	2124	1855	22	2492	1692
闻诊信息数据集	9	79	94	82	139	65

表 1. 所有实验数据集的详细信息

本文实验使用的无标注和带标注的中医临床记录数据集均是基于真实的中医临床记录数据创建，该数据由中医专家在日常诊疗疾病的过程中收集，包含11251条中医临床记录。其中，无标注的中医临床记录数据集由此11251条无标注的中医临床记录直接构成。带标注的中医临床记录数据集则是在11251条中医临床记录的基础上，经过一系列处理得到，具体处理步骤如下：

(1) 讨论并定义中医临床记录中的四诊描述，然后制定标注指南，用于指导后续的数据标注。

(2) 中医专家按照制定好的标注指南，利用Zhang等人(2020)论文中所构建的标准化实验语料构建系统[2]，对11251条中医临床记录数据样本进行四诊信息标注。

(3) 中医专家对标注好的所有数据样本反复审查并修改，形成高质量的标注数据。

(4) 将步骤(3)中得到的高质量的标注数据，按照预定义的标签集合$\boldsymbol{L}$，转化为以字为基本标注单元的BIO标注数据。

(5) 将步骤(4)处理后的数据中包含多重标签（即数据样本中的字具有多个不同标签）的数据样本移除，并将剩余数据样本中的空格和\t移除。

经过上述处理过程，最终得到10594条带标注的中医临床记录。实验中将该数据集按6:2:2的比例随机划分为三部分，得到的训练数据、验证数据和测试数据大小分别为6346条、2124条和2124条带标注的中医临床记录。实验中还将训练数据集中所有包含闻诊信息的数据样本单独复制，组成闻诊信息数据集。各类实验数据集具体信息如表1所示。

[2]实验语料构建系统: http://hknlprel.it.sunsheen.cn/HKKSNLP/build/index.html#/LoginRelation

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 评价指标

本文利用F1值和准确率(Accuracy)评价各模型的中医临床记录四诊描述抽取性能，F1和Accuracy的计算公式如下：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

$$Accuracy = \frac{PT}{PT + PN} \tag{8}$$

在公式(7)中，Precision和Recall分别表示模型在测试集上对各类标签预测的精确率和召回率，它们的具体计算方法可参见文献(Wang et al., 2014)。在公式(8)中，PT表示模型预测标签正确的字单元数量，PN表示模型预测标签错误的字单元数量。

4.3 实验设置

在采用领域适应方法微调MC-BERT时，初始学习率被设置为5e-5，批量大小被设置为512，最大句子长度被设置为256。本文提出的模型在训练时，采用了AdamW优化器，初始学习率η被设置为3e-5，β_1被设置为0.9，β_2被设置为0.999。此外，批量大小M被设置为64，最大句子长度被设置为256，批数据过采样比例P被设置为0.4，E被设置为400，Dropout被设置为0.1。

在对比实验中，HMM模型是基于Rabiner等人(1989)的论文实现。CRF模型使用了CRF++开源工具包[3]，其特征定义为在窗口大小为2的上下文中的一元组和二元组。BiLSTM、BiLSTM+CRF等深度神经网络模型是基于Lample等人(2016)论文中的开源代码实现，它们的输入为2451（即实验数据集中包含的字表大小）维的one-hot向量，中间层字向量的维度设置为128。

4.4 实验结果及分析

方法	F1 (%)									Acc (%)
	O	B-望	I-望	B-闻	I-闻	B-问	I-问	B-切	I-切	
HMM(Rabiner, 1989)	73.07	87.78	91.41	22.99	26.15	84.13	96.02	93.07	92.45	92.80
CRF(Lafferty et al., 2001)	80.90	93.96	94.71	61.54	57.97	88.86	97.05	94.82	93.94	94.92
BiLSTM(Lample et al., 2016)	80.22	93.91	94.07	28.57	12.24	87.44	96.85	94.43	92.65	94.49
BiLSTM+CRF(Lample et al., 2016)	80.97	94.08	94.81	32.26	35.09	87.70	96.93	94.48	93.21	94.67
BERT+BiLSTM+CRF	83.46	94.00	94.92	54.55	52.63	89.64	97.45	94.22	93.81	95.37
BERT+BiLSTM+CRF+BDO[1]	83.71	94.58	95.47	60.00	56.34	89.18	97.44	94.47	94.41	95.47
本文方法-DA[2]-BDO[1]	84.49	94.39	95.26	54.05	51.28	89.70	97.41	94.82	94.39	95.49
本文方法-BDO[1]	84.32	94.25	95.37	50.00	53.73	89.11	97.45	94.67	94.57	95.52
本文方法	**85.14**	**94.62**	**95.51**	**62.22**	**61.54**	**89.93**	**97.54**	**94.91**	**94.67**	**95.70**

[1] “BDO”指**B**atch **D**ata **O**versampling，即“批数据过采样”
[2] “DA”指“**D**omain **A**daptation”，即“领域适应”

表 2. 实验结果

表2列出了本文方法和对比方法在测试数据集上获得的最佳F1值和准确率（在表2中以Acc表示）结果。从表2可以看出，本文方法在各标签的预测结果上均优于所对比的方法。本文方法的Acc达到了95.70%，相比所有对比方法有0.78%到2.9%的提升。本文方法相比最优的对比方法，在每种标签的预测F1值上，平均提升了1.37%。上述结果充分验证了本文方法在中医临床记录四诊描述抽取任务上的预测性能。

此外，通过消融实验，本文还进一步验证了所提出方法的各主要部分的有效性。从表2中可以看到，当本文方法移除领域适应和批数据过采样之后，准确率仍优于其它对比方法。具体地，除少见类别“B-闻”和“I-闻”以外的其他标签的F1值均高于其它对比方法。这证明了本文将MC-BERT+BiLSTM+CRF模型应用于中医临床记录四诊描述抽取的有效性。少见类别预测性能较差的主要原因是基于BERT的深度神经网络模型结构复杂，参数量巨大，对训练数据集

[3]CRF++开源工具包：https://taku910.github.io/crfpp/#source

(c) 2022 中国中文信息学会计算语言学专业委员会

中包含的少见类别学习不充分，导致其预测性能低于模型复杂度相对较低的统计机器学习模型CRF。

当本文方法只移除批数据过采样方法时，模型预测的准确率仍然优于所有对比方法。并且，在“I-望”、“I-闻”、“I-问”、“I-切”等标签上的F1值优于同时移除领域适应和批数据过采样的情况。这说明领域适应方法能够有效提升模型抽取四诊描述的整体性能，且对于非边界四诊描述的判别有强的促进作用。当本文方法不移除任何组件时，其性能在准确率以及每个标签的F1值上均优于所有的对比方法，这进一步验证了本文方法的领域适应与批数据过采样的有效性。

此外，本文方法对于存在类别分布不均衡问题的“B-闻”和“I-闻”标签的抽取效果有显著提升，对应的F1值分别达到了62.22%和61.54%。该结果说明，在训练模型的过程中策略地增加批量中的闻诊信息，使模型更充分地学习闻诊描述特征，进而在一定程度上缓解了因类别分布不均衡问题给模型预测性能带来的负面影响。从表2中还可以看出，将批数据过采样应用于通用领域的BERT+BiLSTM+CRF模型时，模型在少见的四诊描述类别标签“B-闻”和“I-闻”上的抽取性能F1值也出现了显著提升，这验证了本文3.3.3节设计的批数据过采样方法的有效性。

为进一步证明批数据过采样的有效性，对模型在测试数据集上的标注结果进行了进一步的分析。发现移除批数据过采样后的模型往往会将闻诊信息错误地标注为问诊信息。例如：将“肠鸣，少腹重坠略有缓解”一起标注为问诊，而“肠鸣”实则应标注为闻诊。这是由于训练数据集中闻诊信息的数据量极少，直接利用小批量梯度下降算法对模型进行训练时，闻诊信息仅出现在少数用于计算更新梯度的批量中，在大多数批量中其出现次数甚至为0。

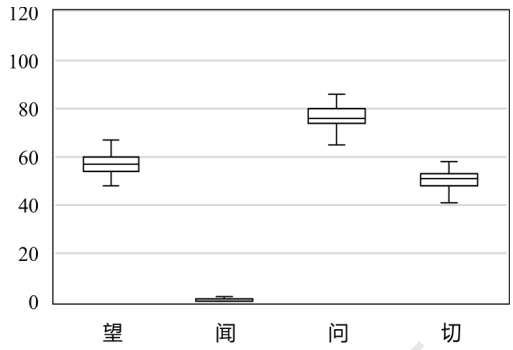


图 7. 批量中的四诊信息数量统计(P = 0)

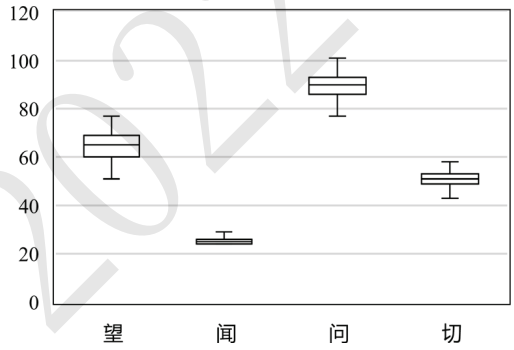


图 8. 批量中的四诊信息数量统计(P = 0.4)

图7是在移除了批数据过采样的模型训练过程中，对训练数据集一轮遍历时，以划分的批量为单位，对批量中包含的四诊描述的出现次数的统计结果。从图7中可以明显地看出，批量中闻诊描述出现的次数极少，近乎为0，与问诊描述在平均出现次数上的差值接近80。这导致模型无法充分地学习到闻诊描述特征，将闻诊描述错误地预测为其它类型的描述。

从图8中可以看出，在不移除批数据过采样且P值被设置为0.4的情况下（采用图7相同的统计方法），批量中包含的闻诊描述的数量大幅提升，这将使模型能够在训练过程中更充分地学习闻诊描述特征，同时使模型对“B-闻”和“I-闻”标签的预测性能显著提升。

4.4.1 批数据过采样比例P对模型抽取性能的影响

为验证不同批数据过采样比例P的设置，对本文所提出的中医临床记录四诊描述抽取方法的影响，本文进一步实验了在P被设置为0、0.2、0.4、0.6、0.8或1时，模型在测试集上，对$\boldsymbol{L}$中的各类标签的预测性能，实验结果如图9所示。

从图9可以看出，当$P = 0.4$时，所有标签的F1值均达到最高，并且相较于其它标签，“B-闻”和“I-闻”的F1值增幅最大。该结果说明，当$P = 0.4$时，本文模型能够最有效地从批量中学习到闻诊描述特征，能够更好地消除类别分布不均衡对模型预测性能的影响。该结果进一步说明，在批量中策略地增加包含闻诊描述的实例，间接地降低其他三诊描述在批量中出现的占比，能够有效地避免模型在训练时过度地拟合望诊、问诊和切诊类别标签，让模型更充分地学习少见的闻诊类别标签，进而增强模型的预测性能和泛化能力。

此外，当P<0.4时，本文模型在“B-闻”和“I-闻”标签上的F1值有明显降低，而在其他标签上的F1值无明显波动。该结果说明，在模型训练过程中，闻诊描述在批量中出现的次数降

(c) 2022 中国中文信息学会计算语言学专业委员会

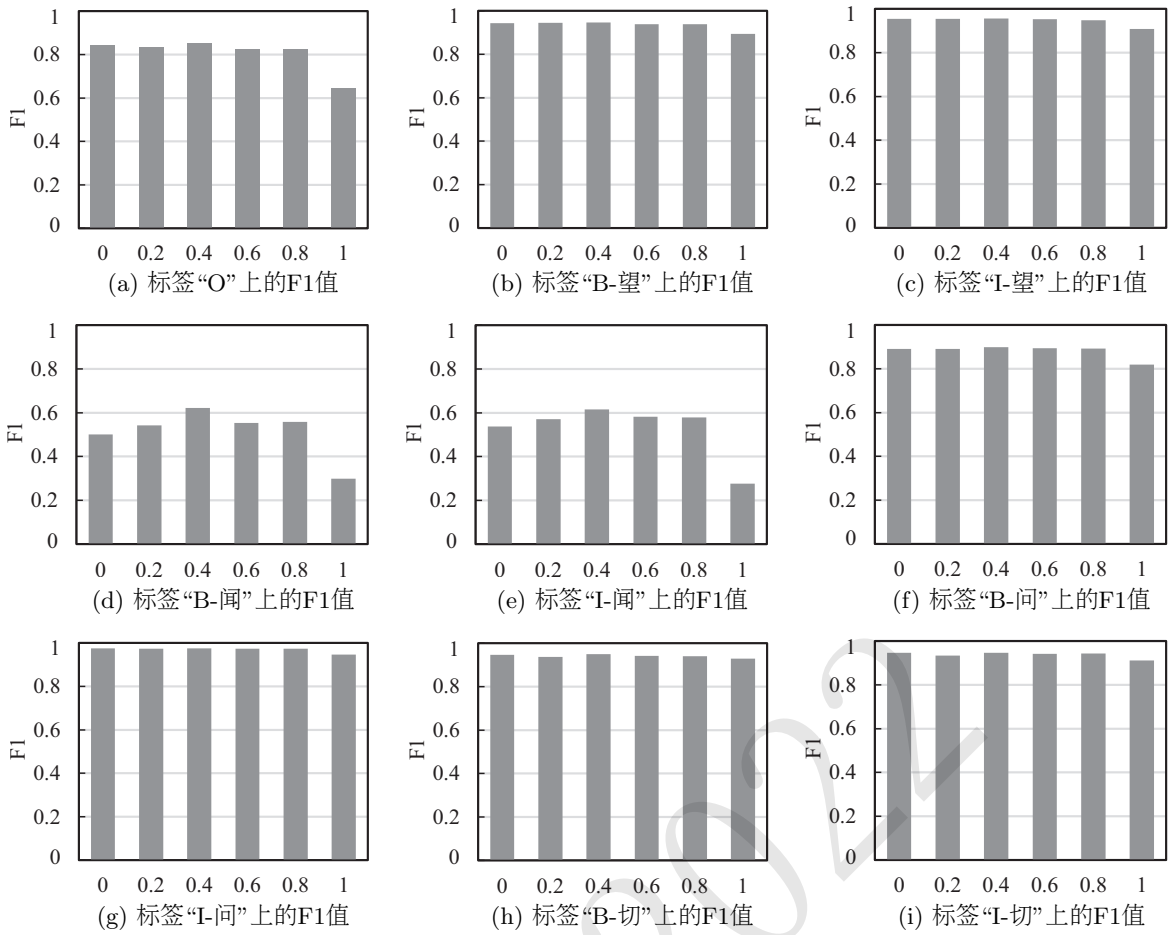


图 9. 不同批数据过采样比例下的F1值

低，导致类别分布不均衡问题加重，使得模型对闻诊类别标签的特征学习不充分，导致模型对“B-闻”和“I-闻”标签的预测能力降低。而其他三诊描述在批量中出现占比变化不大，模型仍能充分学习，因此，它们的F1值并没有明显变化。

最后，当P>0.4时，模型在各类标签上的F1值均有不同程度的降低，特别是在P被设置为1时。这是由于批数据过采样是从闻诊信息数据集中选取数据样本放入批量中导致的。P值越高，模型越近似于在闻诊信息数据集上进行模型的训练，然而，闻诊信息数据集仅包含训练数据集中所有包含闻诊描述的数据样本，数据规模小，包含信息少，这将直接影响模型的训练效果，最终导致所有标签的F1值下降。

5 总结与展望

本文初探了中医临床记录四诊描述抽取任务，以万余条中医临床记录自建了标准实验语料，针对四诊描述类别分布不均衡带来的挑战，提出了一种基于批数据过采样的中医临床记录四诊描述抽取方法。在标准实验语料上，与对比方法相比，本文提出方法取得了最优结果。与对比方法的最优结果相比，本文提出的方法将少见类别的抽取性能F1值平均提升了2.13%。此外，通过多个角度的细致分析，进一步验证了本文所提出方法的有效性。

目前，中医临床记录四诊描述抽取模型的预测性能还有待进一步提升，未来将深入探究中医临床记录四诊描述抽取任务的特点及存在的问题，设计并实践更优的抽取方法，进一步提升中医临床记录四诊描述抽取方法的性能，并将方法应用于实践，达到为中医临床辨证论治提质增效的目标。

参考文献

李红岩, 李灿, 郎许锋, 杨涛, 周作建, 战丽彬. 2022. 中医四诊智能化研究现状及热点分析. 南京中医药大

(c) 2022 中国中文信息学会计算语言学专业委员会

学学报, 38(02):180-186.

刘树栋, 张可. 2019. 类别不均衡学习中的抽样策略研究. 计算机工程与应用, 55(21):1-17.

屈丹丹, 杨涛, 胡孔法. 2021. 基于字向量的BiGRU-CRF肺癌医案四诊信息实体抽取研究. 世界科学技术-中医药现代化, 23(09):3118-3125.

肖瑞, 胡冯菊, 裴卫. 2020. 基于BiLSTM-CRF的中医文本命名实体识别. 世界科学技术-中医药现代化, 22(07):2504-2510.

印会河. 2005. 中医基础理论. 上海科学技术出版社, 上海, 中国.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. *12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1-5.

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*.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 720-728.

John Lafferty, Andrew Mccallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, 282-289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

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.

Chin Teng Lin, Tsung Yu Hsieh, Yu Ting Liu, Yang Yin Lin, Chieh Ning Fang, Yu Kai Wang, Gary Yen, and Nikhil R. Pal. 2017. Minority oversampling in kernel adaptive subspaces for class imbalanced datasets. *IEEE Transactions on Knowledge and Data Engineering*, 30(5): 950-962.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik,. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1): 1-39.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): 257-286.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Shaukat Ali Shahee and Usha Ananthakumar. 2018. An adaptive oversampling technique for imbalanced datasets. *Industrial Conference on Data Mining*, 1-16.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 30.

Yaqiang Wang, Yiguang Liu, Zhonghua Yu, Li Chen, and Yongguang Jiang. 2012. A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features. *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, 223-230.

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.

(c) 2022 中国中文信息学会计算语言学专业委员会

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.

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*.

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.

CCL 2022

篇章级小句复合体结构自动分析

罗智勇* 北京语言大学 luo_zy@blcu.edu.cn
韩瑞昉 北京语言大学 15321103341@163.com
张明明 北京语言大学 15801572558@163.com
韩玉蛟 北京语言大学 chloe_hanyu@163.com
赵志琳 北京语言大学 Zhi_Lin_Zhao@163.com

摘要

话头话身共享关系是小句组合成小句复合体的重要语法手段，也是汉语篇章级句法语义分析的重要基础。本文通过引入窗口滑动机制，将篇章文本及其成分共享关系转换为文本片段及片段内部的成分共享关系预测问题，并针对预测结果合并与选择问题，依据话头话身共享关系的语法限定性，提出了多种候选项消除策略。实验结果表明，本文方法在缺少小句复合体边界信息条件下仍取得了与传统基于NTC的方法可比的实验结果，尤其是在确实缺失共享成分的待预测位置处的召回率提高了约0.4个百分点。

关键词： 小句复合体；边界；滑动窗口；预训练语言模型

Chinese Clause Complex Structure Automatic Analysis on Passage

Luo Zhiyong* 北京语言大学 luo_zy@blcu.edu.cn
Han Ruifang 北京语言大学 15321103341@163.com
Zhang Mingming 北京语言大学 15801572558@163.com
Han Yujiao 北京语言大学 chloe_hanyu@163.com
Zhao Zhilin 北京语言大学 Zhi_Lin_Zhao@163.com

Abstract

The naming-telling relationship is an important grammatical method for combining clauses into clause complexes, and it is also an important basis for Chinese text-level syntactic and semantic analysis. In this paper, by introducing the window sliding mechanism, the text and its component sharing relationship are transformed into the prediction problem of the text segment and the component sharing relationship within the segment. At the same time, in view of the problem of combining and selecting prediction results, this paper proposes a variety of candidate elimination strategies based on the grammatical limitation of the sharing relationship between words and phrases. The experimental results show that the method in this paper still achieves comparable experimental results with the traditional NTC-based methods in the absence of the boundary information of the clause complex, especially the recall rate at the to-be-predicted position where the shared components are indeed missing is improved by about 0.4 points.

Keywords: Clause complex , Boundary , Sliding window , Pre-Training language model

第二十一届中国计算语言学大会论文集，第623页-第633页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

在中文自然语言处理中，通常会先对文本进行分句处理，再进行相关的下游任务。分句处理常见的做法是根据标点符号，如：句号、感叹号、问号、分号等，将文本分割成标点句序列。由于汉语标点句间存在成分共享的现象，这种机械分割方式会使得切分出来的标点句序列结构和意义不完整，从而直接影响阅读理解、机器翻译、信息抽取、搜索推荐等下游任务的性能。以下图1为机器翻译任务中的一个例子，包括中文原文及英语参考译文。

中文原文： 对于公民的申诉、控告或者检举，有关国家机关必须查清事实，负责处理。任何人不得压制和打击报复。
参考译文： The State organ concerned must, in a responsible manner and by ascertaining the facts, deal with the complaints, charges or exposures made by citizens. No one may suppress such complaints, charges and exposures or retaliate against the citizens making them.

Figure 1: 机械分割后标点句序列结构意义不完整示例

这段话引自中华人民共和国宪法第四十一条，中文原文由两个句号句组成，它的英语参考译文则引自全国人大网，同样也是两个句号句。其中，第二个句号句“任何人不得压制和打击报复”的受事是第一个句号句中的“公民的申诉、控告或者检举”。但它前面的句号隔断了这种关系，使“压制”和“打击报复”无法找到被施用者。如果机器翻译系统以句号句为单位进行翻译，则很难译出参考译文中加下划线的部分。由于句号句不一定能表示完整的意义，以句号句为单位进行语言信息处理的工作会受到本质性的影响。

针对标点句间成分共享现象，宋柔(2008)(2013)、尚英(2014)等人提出并完善了关于篇章层级的广义话题理论：小句复合体理论体系。小句复合体是语篇中的标点句序列，它是语篇的上下文语境中最大的紧密逻辑语义结构，也是最小的自足话头结构。自足话头结构即一个标点句序列既没有话头在上下文中，也没有词语可以看作上下文中某标点句的话头(宋柔, 2022)。

依据上述小句复合体的定义，界定小句复合体边界将依赖于标点句间的话头话身共享关系和逻辑语义关系。也就是说，在标点句间话头话身共享关系与逻辑语义关系分析清楚之前，很难界定小句复合体的边界。按照自然语言文本的认知方式，存在先有结构分析，后有边界的天然逻辑顺序。而目前关于小句复合体的结构自动分析的研究，都建立在人工标注好话头话身结构、划分好边界的小句复合体上，即先定好边界，而后进行结构分析。这样的做法颠倒了先结构分析、后有边界的逻辑顺序，不符合常规认知。

本文的研究决定回归人类对自然语言文本的认知方式，遵循先结构分析后边界的逻辑规律。即跳出小句复合体边界的限制，不再以边界完整的小句复合体为单位，而是将文本处理范围扩展到边界清晰的文本篇章上，进行基于滑动窗口的篇章级小句复合体结构自动分析。将问题转换为预测滑动窗口截取出的完整标点句序列上的每个标点句开头和结尾是否缺失成分，并找出缺失成分的开始位置和结束位置。再将所有窗口内部的结构拼接起来，形成整个篇章层级的小句复合体结构。

本文的主要贡献在于：（1）提出了在篇章长文本上进行结构自动分析的三步走策略：首先，将篇章文本及其成分共享关系转换为预训练语言模型可接受的样例级别的文本片段和片段内部的成分共享关系；其次，将样例经过预训练语言模型来学习和预测，得到每个样例对应的共享关系预测结果；最后，将样例级别的预测结果转换回篇章中去，得到篇章中每个标点句句首和句尾的共享成分在篇章中的位置。（2）提出了合并候选项的方法、是否重组答案、是否

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家自然科学基金（62076037）

清除不合规预测答案共三种候选项消除策略，以及为缺失成分超出样例范围的待预测点设置不同标签类型的方法，来解决一个样例中包含多个待预测点时需要对多个预测结果进行合并的问题，以及待预测点缺失的成分不在样例文本范围内的问题。（3）在BERT-base、BERT-wwm、RoBERTa三个预训练语言模型上进行了实验，并与传统基于NTC的方法进行对比，对实验结果和预测错误的样例进行分析，验证了本文方法的有效性。

论文第2节介绍小句复合体结构自动的相关研究工作与进展；第3节介绍篇章级小句复合体自动分析问题的形式化定义；第4节介绍篇章级小句复合体模型的具体实现；论文第5节为实验验证与分析；第6节为论文总结与展望。

2 相关研究

宋柔指出，成分共享是构造小句复合体的基本语法手段。所谓成分共享(宋柔, 2022)，是指由于语言经济性需要，文本中话语片段中会有空缺成分，空缺的信息可能来自文本语境、话语交际场景、知识背景等。文本中某个话语片段字面上出现的一些成分，被另一些话语片段以字面上空缺的方式在语义上使用，这种现象就是成分共享。

汉语小句复合体理论体系中，包括四种成分共享模式(宋柔, 2022)，分别是话头共享模式（分支模式、新支模式、后置模式）、话身尾部共享模式（汇流模式）、超级小句复合体导引模式，以及模板模式。四种共享模式中，模板模式是在标点句中间的位置缺失成分，因此在现阶段的结构自动分析任务中难以形式化。其他三种成分共享模式都可以被形式化定义为话头话身识别任务。目前，已有许多研究者对这方面进行研究，主要可以分为基于传统机器学习的方法和基于深度学习的方法两类。

基于理论的传统方法主要由蒋玉茹等人提出。蒋玉茹(2012)首先根据话头分支模型，开展单个标点句的话头识别任务。其采用穷举策略，根据上一个话头自足句t_{i-1}为当前标点句c_i构造候选话头集合，然后利用编辑距离，计算候选话头和话头实例的相似性，筛选出正确的话头。在实现单个标点句的话头识别任务后，蒋玉茹(2017)又将话头识别任务扩展到标点句序列上。其仍采用穷举方法依次列举出每个标点句的候选话头，用树结构进行存储；再采用适当的策略计算话头候选树中每个节点的值，并计算每个叶节点到根节点的路径值；最后从中找到路径值最大的路径，从而得到标点句序列相对应的话头序列。但由于穷举策略会极大影响系统的执行效率，蒋玉茹(2014)又利用标点句在篇章中的位置特征、话头的语法特征、话头串和说明的邻接性等细粒度特征，指导候选话头的生成过程，尽可能的减少候选话头的个数，提高系统效率，在单个标点句和标点句序列上的话头识别任务的正确率均有提升。

基于深度学习的方法已有多人进行研究。M.Teng (2018)提出基于Attention-LSTM 的深度神经网络模型，进行单个标点句话头识别任务的研究，其实验结果相较于传统方法又有提升。但该研究在小句复合体话头共享关系自动分析方面，仅局限于分支模式，而对于新支模式、汇流模式、后置模式，以及多种模式混合的小句复合体结构的自动分析还未涉及。

针对多种成分共享模式，胡紫娟(2020)通过构建有向无环图NTCGraph结构来对小句复合体话头话身关系进行表示，并在预训练语言模型的基础上，通过Gather层分别取出待预测点向量和NTC向量，通过attention 交互层进行话头、话身的预测，进一步解决了之前只能够处理单一话头模式的问题。但该研究存在不缺失成分与缺失第一个字符冲突的问题，且处理的小句复合体长度小于128，不利于小句复合体结构自动分析的应用扩展。

因此，刘祥(2022)在胡紫娟的基础上开展进一步的研究。其首先引入空锚点机制，在输入的过程中通过插入特定的特殊token解决了因缺失成分位于句首处的标签与不缺失成分的标签重合问题；其次，针对特定插入待预测点的预测方式，提出了NT-MASK局部注意力机制，将全局注意力矩阵切分为了Sentence注意力矩阵和Mask-Sentence注意力矩阵，缓解了因插入待预测点对上下文信息编码带来的噪声干扰，减少不同待预测点之间的噪声干扰。除此之外，还构建了远距离共享成分识别数据集，将小句复合体输入长度由128扩充至512，进一步改善了汉语小句复合体自动分析的性能。

3 篇章级小句复合体自动分析形式化定义

篇章级小句复合体结构自动分析的任务主要可以分解为三个步骤。第一步，将篇章形式的文本处理为预训练语言模型可以接受的最大长度的文本样例；第二步，对样例文本内部的标点句句首和句尾的成分共享关系进行识别；第三步，将每个样例预测出的结果合并为篇章的结

(c) 2022 中国中文信息学会计算语言学专业委员会

果。本节将对第一步和第三步中小句复合体话头话身关系形式化定义、转换和还原，预测结果合并和选择，以及篇章级小句复合体结构自动分析问题描述进行介绍，图2为主要流程。

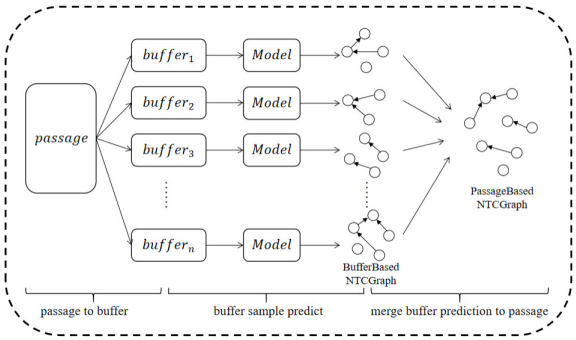


Figure 2: 篇章级小句复合体结构自动分析流程

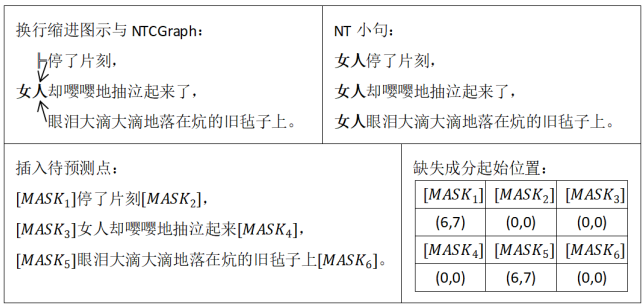


Figure 3: 成分共享关系形式化定义图示

3.1 成分共享关系形式化定义

胡紫娟首次将小句复合体的话头话身结构定义为广义有向无环图NTCGraph(胡紫娟, 2020)，表示为G=(V,E)。V是结点的集合，一个结点表示NTC中的一个标点句。E是小句复合体中的每两个标点句间话头话身成分共享关系的集合。NTCGraph与小句复合体一一对应。篇章级小句复合体话头话身成分共享关系则由一个篇章中多个NTCGraph线性相连而来，表示为Passage-NTCGraph，包含了篇章中每个标点句句首句尾缺失的共享成分位置。如图3所示，该NTC中包含三个标点句，换行缩进图示与NTCGraph中的箭头由标点句句首或句尾指向其所缺失的共享成分，NT小句为补充共享成分后的句子。在每个标点句的句首和句尾各插入一个[MASK]作为待预测点，每个MASK处缺失的共享成分位置则为在文本中的开始和结束位置。

3.2 成分共享关系转换与还原

由于预训练语言模型对输入长度的限制，成分共享关系需要由篇章级转换为预训练语言模型可接受的样例级，以方便预训练语言模型的学习和预测。每个待预测点的成分共享关系转换公式如下：

$$(start_b, end_b) = \begin{cases} (0,0), & start_p = 0, end_p = 0 \\ (0,0), & start_p < buf_s \, or \, end_p > buf_e \\ (start_p - buf_s, end_p - buf_s), & start_p > buf_s \, and \, end_p < buf_e \end{cases}$$

$(start_b, end_b)$表示每个待预测点的共享成分在样例中的相对位置；$(start_p, end_p)$是从Passage-NTCGraph中获取的该待预测点的共享成分在篇章中的绝对位置；buf_s和buf_e分别指样例中第一个字符与最后一个字符在篇章中的位置。

样例经过模型预测之后，再将生成的样例级预测结果转换为篇章级。即将预测得到的样例级成分共享关系还原回篇章级的成分共享关系。每个待预测点的成分共享关系还原公式如下：

$$(start_p, end_p) = \begin{cases} (0,0), & start_b = 0, end_b = 0 \\ (start_b + buf_s, 0), & start_b \neq 0, end_b = 0 \\ (0, end_b + buf_s), & start_b = 0, end_b \neq 0 \\ (start_b + buf_s, end_b + buf_s), & start_b \neq 0, end_b \neq 0 \end{cases}$$

3.3 预测结果合并与选择

由于一个待预测点会出现在多个样例中，对应地会产生多个$(start_b, end_b)$，这些预测结果位置在各自的样例文本中不同，但还原回篇章中的$(start_p, end_p)$后可能相同，因此需要对相同的预测结果进行合并，最终从不同的预测结果中选取一个作为该待预测点处的成分共享位置。

本文将还原后的每一个篇章中的$(start_p, end_p)$及其产生概率作为一个候选项，加入到对应待预测点的所有预测结果中。对于相同的$(start_p, end_p)$，将其对应的产生概率进行取平均、求和、或取最大的运算来进行合并，合并后的概率作为该$(start_p, end_p)$的最终概率。根据最终概率最大的原则从多个候选项中选择最佳的作为该待预测点处的最终预测结果。合并的过程如图4所示。

(c) 2022 中国中文信息学会计算语言学专业委员会

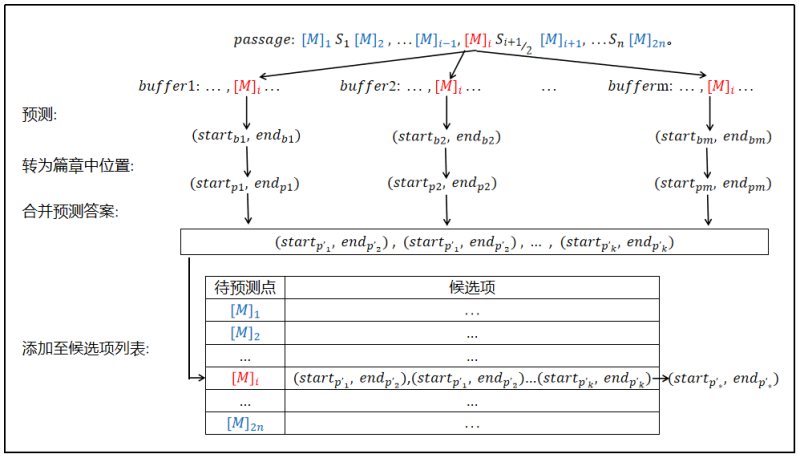


Figure 4: 预测结果合并与选择流程

3.4 篇章级小句复合体结构自动分析问题描述

成分共享关系的转换和还原定义之后，篇章级小句复合体结构自动分析转化为在样例窗口内部进行话头话身关系识别的任务。任务定义为：对于每个标点句的开始和结束位置，在给定的样例文本范围内预测每个字符位置可能成为该待预测位置缺失成分的开始位置和结束位置的概率，也就是预测样例窗口内每个标点句的开头和结尾是否缺失成分，并找出缺失成分在窗口中的开始位置和结束位置。输入输出格式定义如下：

输入的格式为：在标点句两端插入标志[MASK]，形成待预测的信息点，开头插入[CLS]，结尾插入[SEP]。输出的格式则为每个[MASK]位置所缺失的成分在窗口的文本（未插入[MASK]等符号)中开始和结束的索引位置(start,end)。如果不缺失成分，则为(0,0)。图3提到的NTC对应的输入输出形式如图5所示。样例内部的成分共享关系识别即定义为，对于样例中的每个待预测点，在给定的输入文本片段中选择该待预测点缺失成分的开始位置和结束位置，与抽取式机器阅读理解任务的答案形式(Cui et al., 2019b)相同。

Input:

[CLS][MASK]停了片刻[MASK],[MASK]女人却嘤嘤地抽泣起来[MASK],[MASK]眼泪大滴大滴地落在炕的旧毡子上[MASK]。[SEP]

Output:

(6,7), (0,0), (0,0), (0,0), (6,7), (0,0)

Figure 5: 输入输出格式

4 模型方法

4.1 模型架构

小句复合体结构自动分析模型大体仍采用胡紫娟(2020)和刘祥(2022)基于NTC的方法中使用的框架。模型共包括五层：输入层、编码层、收集层、注意力层和输出层。图6是模型的整体框架。

输入层的Embedding表示由Token Embeddings、Segment Embeddings、Position Embeddings三者相加而来。本文的Segment Embeddings不需要区分句子，因此统一用'0'来表示。Position Embeddings则是每个token 的位置向量表示。输入的最终表示为：

$$E = (E_{[CLS]}, E_{m1}, E_{t1}, E_{t2}, E_{m2}, ..., E_{tn}, E_{mm}, E_{[SEP]})$$

编码层的作用是对文本的上下文信息进行编码。该层基于BERT预训练语言模型，将每个token的embedding表示，经过12层Transformer Encoder Blocks来获得结合了上下文语义信息的语义表示，每一层的输出作为下一层的输入，并将Transformer最后一层的输出作为编码层的输出。

(c) 2022 中国中文信息学会计算语言学专业委员会

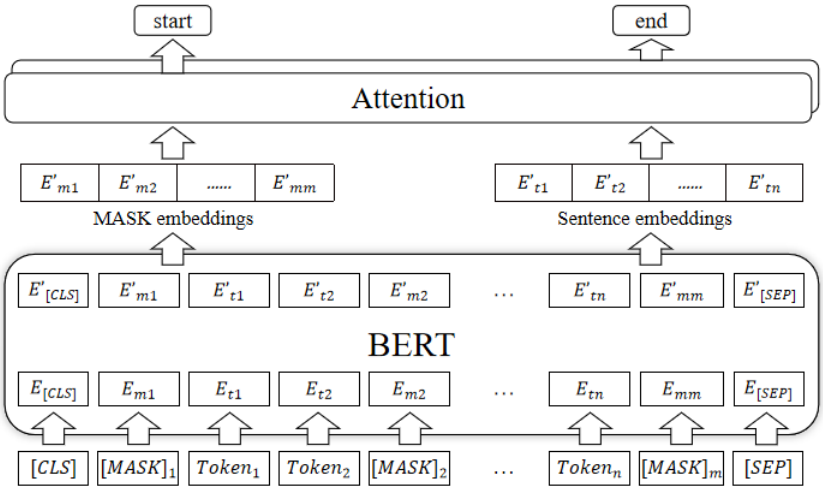


Figure 6: 窗口内标点句序列话头话身关系识别模型整体架构

收集层的作用是，按照在处理语料时记录的[MASK]与非[MASK]的位置，将样例中所有的待预测点[MASK]和文本信息分别抽取出来，得到MASK_embeddings矩阵和Sentence_embeddings矩阵。

注意力层的核心任务是充分结合文本信息，对MASK_embeddings和Sentence_embeddings的相似性进行计算，最终获得每个MASK在每个文本信息字符上能够作为答案的开始位置和结束位置的概率。该层主要依据多头注意力机制(Vaswani et al., 2017)，将上一层的输出MASK_embeddings和Sentence_embeddings 分别看作Quary和Key进行匹配。Attention分为start_attention和end_attention。前者负责预测缺失成分的起点，后者负责预测缺失成分的终点。这样即可在一个[MASK]位置获取两个可以代表start_logits和end_logits的结果。

输出层为最后一层，主要任务是将所有[MASK]对应的在每个文本token处作为答案的开始和结束的概率转化成答案的开始和结束位置，期望输出格式为(start,end)，用于找到其对应的话头话身。start和end结果的维度为[batch_size,MASK_length,1]，即将start_logits中一个[MASK]对应的多个概率值，转化为一个[MASK]对应一个位置索引。

实现的方法是将上一层得到的start_logits和end_logits经过softmax来得到两个使得成为答案的概率值最高的位置索引。计算公式如下：

$$P_s(p_m = i|M) = softmax(H_m^s \cdot H_t^{sT})[i]$$

$$P_e(p_m = i|M) = softmax(H_m^e \cdot H_t^{eT})[i]$$

前者表示MASK处预测的start位置为文本中第i个token的概率，后者表示MASK处预测的end位置为文本中第i个token的概率。$H_m^s, H_t^s, H_m^e, H_t^e$上标的s和e分别表示是对开始和结束位置的预测，下标m和t分别表示MASK和非MASK的token。模型通过端到端的过程来学习。Loss的计算分为对[MASK]处缺失的共享成分开始位置start的计算，和对共享成分的结束位置end的计算，并将二者的平均值作为模型整体的loss表示。计算公式如下：

$$L_s = -\sum_{m=1}^{n} \log P_s(p_m|M)$$

$$L_e = -\sum_{m=1}^{n} \log P_e(p_m|M)$$

$$Loss = \frac{L_s + L_e}{2}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

4.2 候选项消除策略

基于滑动窗口的方法中，同一待预测点会出现在多个窗口样例中，对应地产生多个预测结果，需要将多个窗口预测出的结果（窗口中的相对位置）转换为篇章中的结果（绝对位置）。每个预测结果及其产生概率(start,end,prob)作为该待预测点的一个候选项，问题转换为如何在多个候选项中选出最佳作为最终预测结果。本文从候选项合并方法、是否重组答案，以及是否清除不合规预测答案三个方面来考虑。

候选项合并方法即3.3节中对相同$(start_p, end_p)$产生概率的合并方法，包括概率求平均、求和、求最大。同时还试验了不合并候选项的方法，以及去除同一个待预测点下为零候选项的方法。去除为零候选项指当候选项中有不为(0,0)的候选项时（即预测该位置缺失成分），去除掉为(0,0)的候选项，最终结果从不为(0,0)的候选项中选择概率最大者；否则不去除。

重组答案是指将预测得到的答案对拆分开来进行开始和结束位置的重组的方式。即将同一个待预测点得到的全部开始位置和结束位置重新进行组合。在拆分重组的过程中，开始和结束位置应当遵循答案的合法性原则，即如果开始位置为0，只能和结束位置为0组合，以及答案的开始位置应当小于或等于答案的结束位置等。

清除不合规预测答案是指清除掉模型预测得到的不合法的答案。不合规分为以下几类：答案对中有一个位置为0而另一个位置不为0；答案的开始位置大于结束位置；答案中的共享成分跨越了一个以上标点句；答案中的成对标点符号不匹配……将不合规候选答案从候选项列表中清除之后，有利于减少错误答案的干扰，选出正确答案。

4.3 缺失成分不在样例内部的处理策略

基于滑动窗口的方法使用滑动窗口截取文本中的标点句序列，这样会截断窗口内部和外部的成分共享关系，导致如果窗口内成分不完整的待预测点的共享话头话身在窗口范围外，则无法正确预测到其缺失的内容。

对于这种待预测点，本文采取了以下几种处理策略：第一，赋予(0,0)的标签，即在样例文本内其缺失成分的位置指向[CLS]；第二，赋予其样例末尾最后一个字符的位置的标签，即在样例文本内其缺失成分的位置指向[SEP]的位置；第三，在第一种策略的基础上增加第三位标签来表示该待预测点在篇章中的成分是否完整，‘0’表示该待预测点在篇章中成分完整，不缺失成分；‘1’表示该待预测点在篇章中成分不完整，缺失成分；第四，为每个待预测点单独构造一条样例，并确保该待预测点缺失的成分处于样例文本内部，来避免这个问题。以上策略中除第一种方式和不缺失成分的待预测点标签相重合之外，其他三种方式均将这类缺失成分不在样例内部的待预测点和不缺失成分的待预测点的标签区别开来。

5 实验

5.1 数据集

本文的语料为北京语言大学中文小句复合体标注语料（CBBC）(宋柔, 2017)，包含百科、报告、新闻、小说共4个领域。经过对语料的重新清洗后，共计9种类别，245个篇章，12509个NTC，40379个标点句，80758个待预测点。

为保证训练集、验证集和测试集之间没有重合的篇章，在每种类别下按7:2:1的比例划分篇章，得到训练集、验证集和测试集。按照滑动窗口数据构造格式，以及NTC为单位的数据格式，分割后的数据集样例个数如下表1所示：

数据集样例个数	train	eval	test	总计
滑动窗口样例	21240	4101	6845	32186
NTC样例	8155	1818	2534	12507

Table 1: 不同格式的数据集样例个数

5.2 评估指标

本文采取的评估指标包括篇章中缺失成分的待预测点处的精确率、召回率、F1值，以及篇章中全部待预测点的总正确率。具体评估指标如下：

(c) 2022 中国中文信息学会计算语言学专业委员会

精确率Precision: 缺失成分的待预测点预测正确的个数占预测结果为缺失成分的待预测点总个数的比重（当且仅当开始位置和结束位置均正确时，该待预测点正确）。定义为：$Precision = N_{lost}/N_{predict_{lost}}$。

召回率Recall: 缺失成分的待预测点预测正确的个数占目标答案中缺失成分的待预测点总个数的比例。定义为：$Recall = N_{lost}/N_{target_{lost}}$。

F1值：精确率和召回率的调和平均。定义为：$F1 = (2 * Precison * Recall)/(Precison + Recall)$。

总正确率Accuracy：预测正确的待预测点个数（包括不缺失成分的待预测点和缺失成分的待预测点）占所有待预测点总个数的比例。定义为：$Acc_{total} = (N_{lost} + N_{nlost})/(N_{total_{lost}} + N_{total_{nlost}})$。

5.3 实验结果

本小节使用基于滑动窗口的数据集在BERT-base(Devlin et al., 2018)，BERT-wwm(Cui et al., 2019a)，RoBERTa(Liu et al., 2019)三个预训练语言模型上进行了实验。实验以概率求和合并候选项作为候选项合并方法，并清除不合规候选项。基线方法为刘祥(2021)的传统的限定小句复合体边界的样例构造方式，即利用人工标注的小句复合体边界信息，以边界完整的小句复合体为单位进行训练和测试。二者最终得到的每个样例的预测结果都对应回篇章中的绝对位置，且用相同测试篇章进行评测。下表2为实验结果：

	Precision	Recall	F1	Acc_total
Baseline	0.7345	0.7490	0.7417	0.9103
BERT-base	0.6638	0.7293	0.6950	0.8867
BERT-wwm	0.6622	0.7432	0.7003	0.8859
RoBERTa	0.6842	**0.7530**	0.7169	0.8923

Table 2: 不同预训练语言模型上的实验结果

实验表明，基于滑动窗口的数据集在RoBERTa上，缺失成分的待预测点处精确率、召回率、F1值都达到最佳，分别为0.6842的精确率、0.7530的召回率、0.7169的F1值，在全部待预测点（包括缺失成分的待预测点和不缺失成分的待预测点）处的正确率达到0.8923。其次为BERT-wwm，BERT-base上效果最差。

与baseline基于NTC的方法相比，基于滑动窗口的方法在缺失成分的待预测点处召回率达到历史最佳水平，比baseline基于NTC的方法的召回率0.7490提高0.4个百分点，这体现了基于滑动窗口的方法对缺失成分的识别能力有所提升。而在其他三个指标下，基于滑动窗口的方法表现稍逊于基于NTC的方法，但也在一定程度上达到了可比的性能。这是由于基于滑动窗口的方法难度更大。

第一，基于NTC的方法使用独立的NTC作为输入，没有像基于滑动窗口方法那样带来NTC之外的干扰信息。滑动窗口方法的输入包含了大量的标点句文本信息，文本长度最大程度接近512，模型需要从近512个字符中去学习和预测两个位置，包含了大量的冗余信息。而NTC方法输入文本长度为NTC长度，待预测点缺失成分一定在NTC范围内，不包含NTC之外的干扰信息。

第二，基于NTC的方法每个待预测点只产生一个预测结果，无需对预测结果进行筛选合并。基于滑动窗口的方法同一待预测点出现在多个样例中，相应地会产生多个预测结果。虽然预测结果中会有NTC方法得不到的正确结果，但如何从多个预测结果中筛选出正确答案仍是难点。

第三，基于NTC的方法每个待预测点在样例中的标签具有一致性，而基于滑动窗口的方法同一个待预测点在不同样例中的标签不一致。后者同一待预测点在不同样例中位置不同，答案标签也不同。更有共享成分超出文本范围的情况，使得同一个待预测点的标签既有真正缺失成分的答案位置，又有(0,0)或[SEP]的位置。标签不一致给模型的训练和预测造成一定的困难。

因此，基于滑动窗口的方法与基于NTC的方法相比难度更大，基于NTC的方法占优是在情理之中。

(c) 2022 中国中文信息学会计算语言学专业委员会

5.4 候选项选择策略的研究

基于滑动窗口的方法使得同一待预测点处有多个窗口产生的多个预测结果候选项，针对这一问题，本文从合并候选项的方法、是否重组答案、是否清除不合规预测答案三个方面来考虑。不同的候选项合并方法下缺失成分的待预测点处P、R、F1值如表3所示:

	合并候选项		不合并候选项
	avg-prob	sum-prob	
不去零项	0.7265/0.2573/0.3800	0.7186/0.6959/0.7071	0.8105/0.4409/0.5712
去除零项	0.6788/0.7435/0.7097	0.6886/0.7442/0.7153	0.6832/0.7495/0.7148
	重组答案对		
	avg-prob	sum-prob	max-prob
不去零项	0.7119/0.2648/0.3860	0.7176/0.6978/0.7075	0.8036/0.4547/0.5808
去除零项	0.6733/0.7447/0.7072	0.6774/0.7492/0.7115	0.6767/0.7485/0.7108
	清除不合规答案		
	avg-prob	sum-prob	max-prob
不去零项	0.7327/0.2578/0.3813	0.7250/0.6994/0.7120	0.8147/0.4412/0.5724
去除零项	0.6803/0.7487/0.7129	0.6842/0.7530/0.7169	0.6839/0.7528/0.7167

Table 3: 不同候选项选择策略在测试集上的P、R、F1值

表中1-4行表示采取不同合并候选项方法且不重组答案对时，缺失成分待预测点的P、R、F1值。包括从候选答案中去除零项、将相同开始结束位置候选项的概率值合并、概率求平均、求和、求最大等情况。5-8行表示重组答案对与合并候选项相结合的实验结果。9-12行表示清除不合规答案且不重组答案对时，与合并候选项相结合的实验结果。

实验结果表明，不重组答案对、清除不合规答案，且与概率求和的候选项合并方法结合的情况下，缺失成分的待预测点处的召回率、F1值达到最佳表现。尤其是召回率，达到了0.7530。这是由于概率求和的方法下，某个(start,end)出现的次数越多，最终概率值就越大。即预测到某个(start,end)的窗口个数越多，该候选项就越可能成为正确答案。同时，重组答案对与不重组的方法相比，为无正确答案对的待预测点增加了正确的答案对候选项，有助于将正确答案从候选项列表中被筛选出来，提升了缺失成分待预测点的召回率。而清除不合规答案后，将模型预测得到的不合规的答案对从候选项列表中剔除掉，降低了这类概率高但错误的候选答案对正确答案的影响。也就是在缺失成分的待预测点处，如果候选项中包含正确答案，清除不合规答案的方法有助于将正确答案从候选项列表中被筛选出来，作为最终答案。

5.5 缺失成分不在样例内部的处理策略研究

本小节对4.3节中提到对于缺失成分不在样例内部的四种处理策略进行了实验，实验结果如表4所示。其中CLS_Labels表示第一种指向[CLS]的位置，也就是5.3中Roberta上的实验结果；SEP_Labels表示第二种指向[SEP]的位置；Triple_Labels表示第三种三位标签的策略；Single_Sample表示第四种样例中只包含单个待预测点的策略。

	Precision	Recall	F1	Acc_total
CLS_Labels	0.6842	**0.7530**	0.7169	0.8923
SEP_Labels	0.6754	0.7503	0.7108	0.8902
Triple_Labels	0.7088	0.7335	**0.7209**	**0.8959**
Single_Sample	**0.7171**	**0.6769**	0.6964	0.8878

Table 4: 缺失成分不在样例内部的不同处理策略实验结果

这几种处理策略中，CLS_Labels在缺失成分的待预测点处召回率达到历史最佳水平，为0.7530。其只利用了每个待预测点处缺失成分的位置信息，在对多个预测结果合并时采取了候选项概率求和合并以及清除不合规候选答案的措施，使得真正缺失成分的待预测点的召回率最高。SEP_Labels将缺失成分不在样例内部的待预测点与不缺失成分的待预测点用标签区

(c) 2022 中国中文信息学会计算语言学专业委员会

分开来，帮助模型更好地识别缺失成分与不缺失成分的待预测点，再加上候选项选择策略，达到了稍弱于CLS_Labels的性能。Triple_Labels的实验将每个待预测点在篇章中是否成分完整的信息融合进来，使得最终的答案预测和多个结果合并时候选答案的选取更加合理。其F1值和Acc_{total}最高，综合结果在三者中表现最好。Single_Sample的实验则直接避免了多个预测结果合并这一难点，其实验结果中对于缺失成分的待预测点处Precision在三种方法中最高。

5.6 错误样例分析

从上述实验结果来看，基于滑动窗口的方式综合实验结果很难超越以NTC为样例的实验结果。前一小节中我们对实验结果进行了宏观的对比和分析，本节中对测试集的15262个待预测点中两种方法预测错误的样例进行详细分析。

基于滑动窗口的方法预测错误的待预测点共1643个（缺失成分987个,不缺失成分656个）。基于NTC的方法预测错误的待预测点共1374个（缺失成分1008个，不缺失成分366个）。虽然滑动窗口的方法预测错误的总数大于NTC的方法，但对于缺失成分的待预测点，滑动窗口的方法预测错误的数量小于NTC的方法，即滑动窗口方法对于缺失成分的待预测点上的话头话身识别更加准确。同时，基于滑动窗口预测错误的987个缺失成分待预测点中，有197个（近20%）待预测点的候选答案中包含正确答案，只是其产生概率较低，无法在答案选择环节被选为最终答案，但这仍然表明了滑动窗口方法的潜力所在。

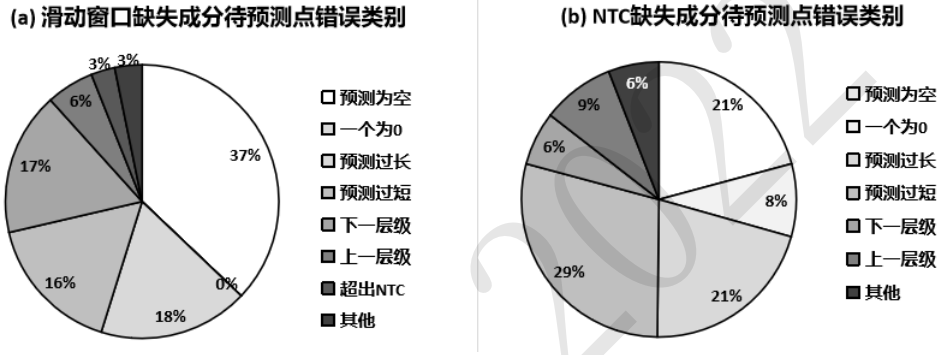


Figure 7: 缺失成分待预测点的错误类型

除此之外，对于缺失成分的待预测点来说，从答案长度、答案的层级、超出NTC范围来对比NTC方法和滑动窗口方法。各错误类型及占比如图7所示。NTC方法错误的类型为预测过长和过短的比例为50%，滑动窗口方法这一项的比例为34%，这说明NTC方法更容易预测为一个答案位置正确，而另一个答案位置错误，使得超出正确答案范围或得到正确答案的一部分，而滑动窗口方法预测一个答案位置正确、另一个答案位置错误的比例较低。NTC方法错误的类型为上一层级和下一层级的比例为15%，滑动窗口方法这一项的比例为23%，说明了滑动窗口方法在层级的学习上表现不如NTC的方法，错误地预测为上一层级或下一层级的情况更多。除此之外，滑动窗口预测错误的一个重要类型是超出NTC范围，这也是滑动窗口方法的难点之一，即滑动窗口的输入范围比NTC方法的输入范围更大，使得预测得到的结果会超出NTC文本。

6 总结与展望

基于NTC的方法存在很大的限制，那就是需要明确的NTC边界信息，这在测试场景中很难满足。基于滑动窗口的方法突破了这种限制，虽然综合结果并未超过NTC方法，但在缺失成分的待预测点处的召回率取得最优，其他指标结果也仍具有可比性，这表明基于滑动窗口的方法客观上是有意义和一定潜力的。除此之外，如何将汉语小句复合体结构自动分析应用到其他下游任务中去，也至关重要。何晓文(2021)等人提出基于小句复合体的句子边界自动识别研究，王瑞琦(2021)和刘祥(2021)等人提出将小句复合体结构自动分析模型与阅读理解任务相结合，在一定程度上提升了阅读理解任务的性能。

未来的工作将对现有的方法进行总结和反思，从当前错误样例总结的规律中继续展开研究，进一步提升小句复合体结构自动分析任务的性能，并探索小句复合体结构自动分析的应用。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu. 2019a. Pre-training with whole word masking for chinese bert.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. A span-extraction dataset for Chinese machine reading comprehension. 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 5883–5889, Hong Kong, China, November. Association for Computational Linguistics.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Y. Jiang and R. Song. 2017. topic structure identification of pclause sequence based on generalized topic theory *.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Xiang Liu, Ruifang Han, Shuxin Li, Yujiao Han, Mingming Zhang, Zhilin Zhao, and Zhiyong Luo. 2021. Shared component cross punctuation clauses recognition in chinese. In Lu Wang, Yansong Feng, Yu Hong, and Ruifang He, editors, *Natural Language Processing and Chinese Computing*, pages 709–720, Cham. Springer International Publishing.

M. Teng, Y. Zhang, Y. Jiang, and Y. Zhang. 2018. Research on construction method of chinese nt clause based on attention-lstm. In *Ccf International Conference on Natural Language Processing & Chinese Computing*.

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*.

Ruiqi Wang, Zhiyong Luo, Xiang Liu, Rui Han, and Shuxin Li. 2021. 基于小句复合体的中文机器阅读理解研究(machine reading comprehension based on clause complex). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 723–735, Huhhot, China, August. Chinese Information Processing Society of China.

何晓文, 罗智勇, 胡紫娟, and 王瑞琦. 2021. 基于小句复合体的句子边界自动识别研究. 中文信息学报, 35(5):8.

刘祥. 2022. 汉语小句复合体结构自动分析改进策略研究. Master's thesis, 北京语言大学.

宋柔. 2008. 现代汉语跨标点句句法关系的性质研究. 世界汉语教学, (2):19.

宋柔. 2013. 汉语篇章广义话题结构的流水模型. 中国语文, (6):12.

宋柔. 2017. 小句复合体的理论研究和应用. https://2011.gdufs.edu.cn/info/1070/2085.htm.

宋柔. 2022. 小句复合体的语法结构.

尚英. 2014. 汉语篇章广义话题结构理论的实证性研究. Ph.D. thesis, 北京语言大学.

胡紫娟. 2020. 汉语小句复合体话头结构分析. Master's thesis, 北京语言大学.

蒋玉茹and 宋柔. 2012. 基于广义话题理论的话题句识别. 中文信息学报, 26(5):114–120.

蒋玉茹and 宋柔. 2014. 基于细粒度特征的话题句识别方法. 计算机应用, 34(5):5.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于话头话体共享结构信息的机器阅读理解研究

韩玉蛟[1] 罗智勇[2] 张明明[3] 赵志琳[4] 张青[5]
北京语言大学
chloe_hanyu@163.com

摘要

机器阅读理解(Machine Reading Comprehension, MRC)任务旨在让机器回答给定上下文的问题来测试机器理解自然语言的能力。目前，基于大规模预训练语言模型的神经机器阅读理解模型已经取得重要进展，但在涉及答案要素、线索要素和问题要素跨标点句、远距离关联时，答案抽取的准确率还有待提升。本文通过篇章内话头话体结构分析，建立标点句间远距离关联关系、补全共享缺失成分，辅助机器阅读理解答案抽取；设计和实现融合话头话体结构信息的机器阅读理解模型，在公开数据集CMRC2018上的实验结果表明，模型的F1值相对于基线模型提升2.4%，EM值提升6%。

关键词： 机器阅读理解 ； 话头话体结构分析 ； 注意力机制 ； 预训练语言模型

Rearch on Machine reading comprehension based on shared structure information between Naming and Telling

han_yujiao[1] luozhiyong[2] zhangmingming[3] zhaozhilin[4] zhang_qing[5]
Beijing Language and Culture University
chloe_hanyu@163.com

Abstract

The machine reading comprehension (MRC) task aims to test the machine's ability to understand natural language by asking the machine to answer questions in a given context. At present, the neural machine reading comprehension model based on a large-scale pre-training language model has made important progress, but the accuracy of answer extraction needs to be improved when it involves the crossing of punctuation sentences and long-distance correlation of answer elements, clue elements and question elements. By analyzing the Naming-Telling structure information of a text, this paper establishes the long-distance relationship between punctuation sentences, complements and shares the missing components, and assists in the extraction of answers in machine reading comprehension; Design and implement a machine reading comprehension model that integrates the Naming-Telling structure information. The experimental results on the public data set CMRC2018 show that the F1 value of the model is increased by 2.4% compared with the baseline model, and EM value is increased by 6%.

Keywords: Machine reading comprehension , Analysis of the Naming-Telling structure , Attention mechanism , Pretraining language model

第二十一届中国计算语言学大会论文集，第634页-第643页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

机器阅读理解任务主要是让机器像人类一样，通过对给定文本（Context）的分析和理解，回答与给定文本相关的问题（Question）。早期的机器阅读理解大多都是基于规则的或是基于机器学习方法的，例如Lehnert (1977)提出的QUALM系统等，但是这些方法需要人工撰写规则，耗时费力且往往专注于某个领域，不具备很好的延展性。之后随着深度学习的崛起和大规模标注数据集（例如CNN/Daily Mail(Hermann et al., 2015)，SQuAD(Rajpurkar et al., 2016)，DuReader(He et al., 2018)）的出现，研究人员开始考虑将二者相结合，利用预训练语言模型辅助阅读理解任务(Otter et al., 2021)，先在大规模无标注语料上训练模型，再将其应用于机器阅读理解，微调模型使其将在大规模无标注语料上学到的信息融合到只有小规模标注数据的具体任务上。

近年来，机器阅读理解研究发展迅速，在部分领域和数据集上已经达到了人类水平(顾迎捷et al., 2020)，但机器阅读理解在处理远距离成分共享的语义关系时还没有取得实质性的进展，主要体现在于机器阅读理解在处理较长的篇章时，问题中的线索要素和此问题对应答案中的答案要素跨越了多个标点句，从而给机器阅读理解任务带来了较大的困难。具体样例如图1所示。

Context：**严长寿**，出生于上海市，祖籍浙江杭州。服兵役退伍后，找不到合适的工作，二十三岁经父母的朋友介绍，进入美国运通担任传达小弟。
Question：**严长寿 23 岁时**做了什么？
Answer：二十三岁经父母的朋友介绍，进入美国运通担任传达小弟。

话头-话体换行缩进图式：
C1：**严长寿**，
C2：　　出生于上海市，
C3：　　祖籍浙江杭州。
C4：　　服兵役退伍后，
C5：　　找不到合适的工作，
C6：　　二十三岁经父母的朋友介绍，
C7：　　进入美国运通担任传达小弟。

Figure 1: 远距离成分共享样例

图1中用换行缩进图示直观展示出Context中的话头-话体关系[11]：其中，Context共包含7个标点句(用C1-C7表示),C2至C7均共享C1中的话头“严长寿”(用缩进表示)；此外，C5共享C1中的“严长寿”与C4中的“服兵役退伍后”，C7共享C1中的“严长寿”与C6中的“二十三岁经父母的朋友介绍”。Question中的线索要素为“严长寿”与“23岁时”(“**加粗**”表示)，问题要素为“做了什么”(“波浪线”表示)。答案要素为“二十三岁经父母的朋友介绍，进入美国运通担任传达小弟。”(“双下划线”表示)。线索要素与答案要素跨越了4个标点句，为远距离成分共享语义关系，机器在回答此类问题时需将C7的缺失话头信息补充完整。根据Wang[9]之前的统计，这种跨标点句类型的问题占CMRC2018[10]训练集的67.89%，且在此类问题上用BERT模型预测出的答案EM值相比于其它类型问题的EM值要小11%，这就说明了跨标点句问答的难度。

针对此问题，本文提出了一种将小句复合体结构自动分析与机器阅读理解任务相融合的方法，将每个标点句补充为话头自足句(NT小句)(宋柔, 2017)，将问题要素与答案要素归置于同一标点句中，从而帮助模型更好地理解远距离成分共享的语义信息，降低任务难度，图1将话头信息补充完整之后如图2所示。

本文利用小句复合体结构自动分析将阅读理解数据中标点句补充为NT小句，从而使得问题要素、答案要素处于同一标点句中，降低跨标点句答案抽取的难度。本文的主要贡献在于验证了小句复合体结构自动分析对机器阅读理解任务处理远距离成分共享问题的辅助分析作用；其次，提出了区别于Wang(2021)的小句复合体与机器阅读理解的融合机制，同时证明了不同的融合机制会对结果造成不同的影响，并且可能是负面影响。

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家自然科学基金（62076037）；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果(22YCX153)

C1: 严长寿，
C2: 严长寿出生于上海市，
C3: 严长寿祖籍浙江杭州。
C4: 严长寿服兵役退伍后，
C5: 严长寿服兵役退伍后找不到合适的工作，
C6: 严长寿二十三岁经父母的朋友介绍，
C7: **严长寿**二十三岁经父母的朋友，介绍进入美国运通担任传达小弟。

Figure 2: 话头自足句样例

本文第二节介绍小句复合体相关概念以及相关研究；第三节介绍了融合话头话体结构信息的机器阅读理解模型；第四节则是对实验结果的分析，验证小句复合体理论在辅助机器阅读理解任务处理远距离成分共享上的有效性；第五节总结本文并探索接下来可能的研究方向。

2 相关研究

2.1 小句复合体自动分析相关研究

成分共享模式主要分为四种：分支模式，新支模式、后置模式以及汇流模式(宋柔, 2017)，如图3所示(左图为四种成分共享模式，右图为补齐之后的话头自足句)：分支模式的特点在于一个话头被多个右置的话体共享，如例子中C2补全之后则变成“19世纪初，美国第二次抗英战争胜利”；新支模式的共享话头位于中间位置而非首部位置，如例子中C2补充完整则是“你就为了一个黄毛丫头”；后置模式共享的话头在后面的标点句中，如例子中C1补充之后为“凤姐下了车”；汇流模式区别于前三种成分共享模式的特点在于其共享的是话体，而非话头，如例子中C1将位于C2的缺失话体补充完整则为“他把饼干匣子全划破了”。

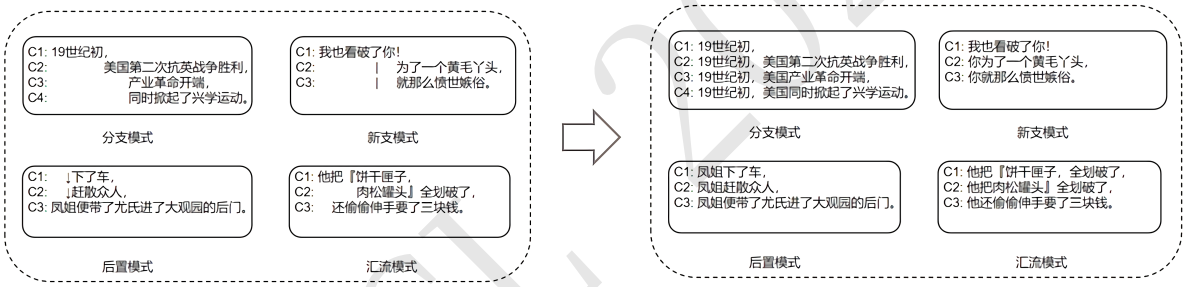


Figure 3: 四种成分共享模式

小句复合体结构自动分析则是对这四种成分共享模式补全其所缺失的话头话体信息，胡紫娟(2020)于2020年结合预训练模型BERT(Devlin et al., 2018)分析汉语小句复合体结构，对标点句是否缺失成分以及缺失成分在篇章中的位置进行预测，准确率达到93.24%，此后，Liu(2021)在此基础上提出了改进策略，在对缺失成分位置的预测上准确率提升了6.89%，总体预测准确率提升了1.22%。

2.2 机器阅读理解相关研究

根据答案形式的不同，机器阅读理解可分为以下四种类型：跨度提取、自由问答、多项选择、完形填空，此外，还有不可回答问题、多跳阅读、会话式回答等类型。

由于深度学习技术在抽取上下文语义信息方面优于基于规则、基于机器学习的方法，现在大多数机器阅读理解模型都采用了深度学习技术：带有CoVe(McCann et al., 2017)的动态协同注意网络在SQuAD数据集上的表现优于原始动态协同注意网络；在将ELMo(Peters et al., 2018)在SQuAD数据集上将最先进的单个模型改进了1.4%，但其受限于LSTM(Hochreiter and Schmidhuber, 1997)的特征提取能力不足；半监督模型GPT[20]在RACE[36]数据集上与SOTA相比实现了5.7%的改进；BERT(Devlin et al., 2018)在SQuAD数据集上F1值高达93.16%。除了对上下文语义信息的抽取以外，还有在答案预测方面的方法，例如Xiong等(2016)提出了一种动

(c) 2022 中国中文信息学会计算语言学专业委员会

态指向解码器，通过多次迭代来选择一个答案跨度。由于预训练语言模型参数过大，训练耗时长，Ren(2020)等人采用蒸馏传统阅读理解模型简化预训练语言模型的方法。然而这些方法多为对模型的修改，并没有考虑远距离成分共享问题。

Wang (2021)于2021年提出了三种将小句复合体结构自动分析与机器阅读理解任务相结合的方法，方法一是将胡紫娟(2020)的小句复合体结构自动分析模型先在中文小句复合体数据集上训练好之后，再将其直接应用于机器阅读理解任务上并对其进行微调，但这就忽略了一个问题：小句复合体结构自动分析任务和机器阅读理解任务是两个不同的任务；方法二采用多任务学习的方式，让小句复合体结构自动分析任务与机器阅读理解任务共享同一个模型的参数，同样，缺点也是未将两个任务区分开来；方法三使用了两个预训练模型，将小句复合体结构自动分析任务与机器阅读理解任务分开处理，最后再将二者得到的向量表示直接相加，此方法不足之处在于没有充分利用到小句复合体结构自动分析模型预测出的标点句缺失成分位置信息。

本文的融合话头话体结构信息的机器阅读理解模型与Wang(2021)提出的融合模型区别有以下几点：第一，本文利用的是小句复合体结构自动分析模型预测出的标点句缺失成分位置信息(start/end logits)，而非Context的向量表示；第二，本文将start logits和end logits转换成注意力矩阵并将其应用于机器阅读理解模型的自注意力机制中。

3 融合话头话体结构信息的机器阅读理解模型

3.1 小句复合体自动分析模型

小句复合体自动分析模型的如图4所示，输入为标点句序列，输入经过预训练模型(BERT-NTC)之后得到整个输入的向量表示(sequence_output)，从中抽取出句子的向量表示(sentence_tensor)和标点句句首句尾插入的MASK的向量表示(mask_tensor)，再将二者取得内积最大值时所对应的序号取出，即标点句缺失成分的起始位置(start_predict)和结束位置(end_predict)，如公式1所示。

$$start\_predict/end\_predict = argmax(\frac{sentence\_tensor * mask\_tensor}{\sqrt{d_k}}) \quad (1)$$

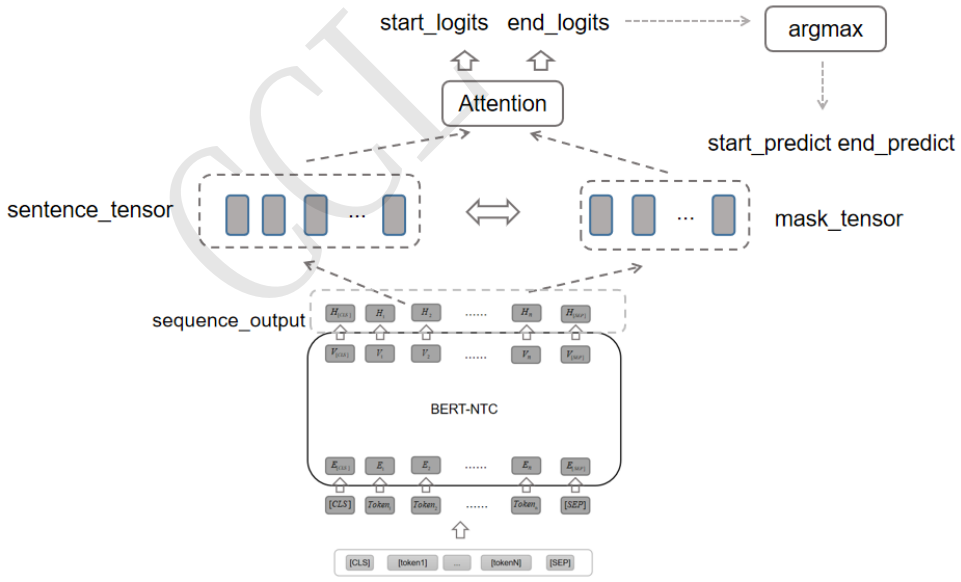


Figure 4: 小句复合体自动分析模型

3.2 机器阅读理解模型

机器阅读理解模型结构如图5所示，训练时输入为问题-答案-文本三元组，同样，输入经过预训练语言模型之后得到sequence_output,再经过线性层调整维度以及经过argmax层取概率

值最大时对应的位置编号，表示为start_predict和end_predict(组成start/end span)，如公式2所示。

$$\begin{aligned} start\_predict &= argmax(start\_logits) \\ end\_predict &= argmax(end\_logits) \end{aligned} \tag{2}$$

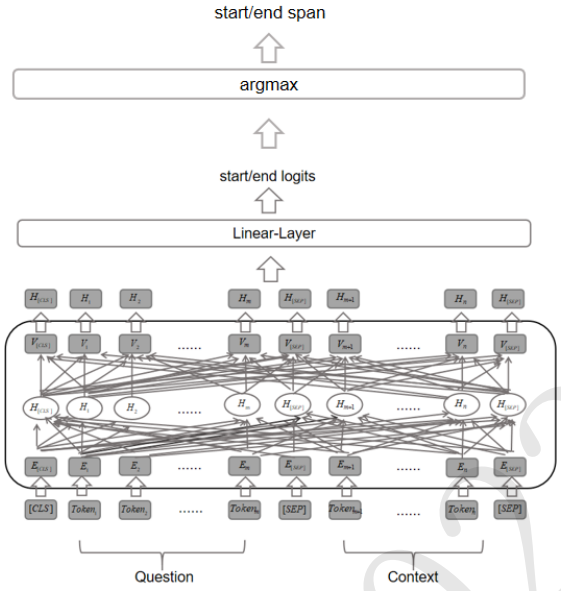


Figure 5: 机器阅读理解模型

3.3 基于finetune的融合模型

Liu (2021)提出的基于NT-MASK的汉语小句复合体自动分析模型是在标点句的首尾位置插入MASK用以预测标点句成分缺失的位置。经过分析，机器阅读理解任务的输入没有在标点句的首尾插入MASK，为了统一输入格式，我们对Liu (2021)的模型做了一点改动：将原来标点句首尾插入的MASK去掉，把待预测的位置直接换成标点句首尾字符的在输入数据中的绝对位置，再在中文小句复合体语料上训练并保存模型。融合模型一就是将此改动之后的模型作为预训练模型直接用于机器阅读理解任务中，并对其进行微调，实验结果将在下一章展示。

3.4 基于NTC_attention_matrix的融合模型

考虑到小句复合体结构自动分析任务与机器阅读理解任务是完全不同的两个任务，直接将在小句复合体结构自动分析任务上训练好的模型用于机器阅读理解任务上效果并不好，于是我们考虑将二者分开处理，使用两个预训练模型分别对两个任务进行训练，并考虑如何将二者相融合。

基于NTC_attention_matrix的融合模型具体结构如图6所示，右半部分是小句复合体自动分析模型(BERT-NTC)，此部分的输入抽取自机器阅读理解任务输入的Context部分。BERT-NTC模型对Context中每个标点句首尾缺失成分的位置进行预测，得到start logits和end logits，再由此得到自注意力矩阵(NTC_attention_matrix)，NTC_attention_matrix初始化为值全为1的大小为lxl的矩阵(l表示输入的长度)，再对标点句首尾位置缺失成分的起始位置以及结束位置区间段加大权值，计算方式如公式3所示(其中，insert_mask_pos表示标点句首尾位置)。将Attention_matrix应用于机器阅读理解模型(BERT-MRC)的自注意力机制中，于是BERT-MRC模型的自注意力机制就演变成公式4,BERT-MRC的输出经过一层线性层就得到最后的预测答案的在上下文中的起止位置。

$$NTC\_atention\_matrix[i,j] = \begin{cases} 2, & j \in start\_logits_{\mathrm{i}}/end\_logits_{\mathrm{i}}, i \in insert\_mask\_pos \\ 1, & otherwise \end{cases} \tag{3}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

$$MRC\_attention\_matrix = Softmax(\frac{QK^T * NTC\_attention\_matrix}{\sqrt{d_k}})V \tag{4}$$

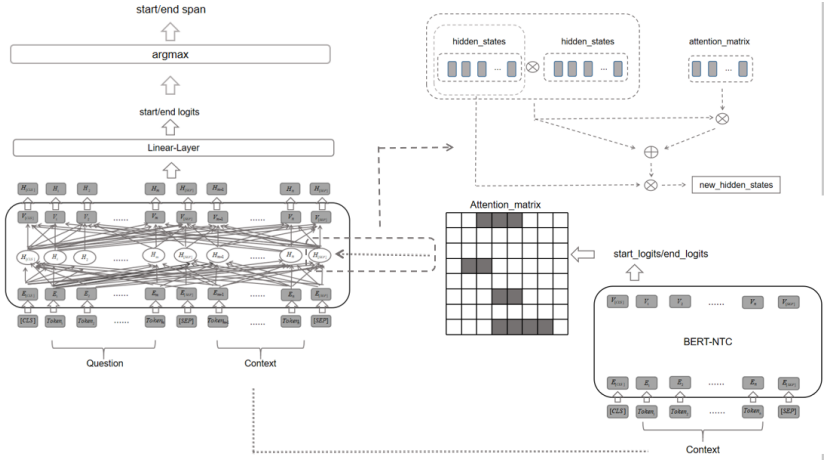


Figure 6: 基于NTC_attention_matrix的融合模型结构

引入NTC_attention_matrix的意义在于让模型更多地关注标点句缺失的成分，也就是将标点句补齐为不缺失成分的NT小句，从而帮助BERT-MRC更好的理解远距离成分共享的语义。同时，因为两个模型是相互独立的，所以可以并发运行，也就意味着加入小句复合体自动分析模型对机器阅读理解模型效率没有影响。

4 实验结果和分析

4.1 数据集

4.1.1 小句复合体语料

本文小句复合体自动分析部分的语料来源于北京语言大学小句复合体语料，其中包括百科、小说、政府报告、新闻四个领域，共有37635个标点句，以及9625个小句复合体。

4.1.2 机器阅读理解语料

本文机器阅读理解部分的语料来源于中文机器阅读理解片段抽取数据集CMRC2018，CMRC2018的初始语料选自维基百科网页dump2的中文部分，将收集到的文章切分成不高于500字的篇章，之后由专家根据这些篇章人工构建问题，每个篇章的问题数不超过5个，同时，每个问题有对应的三个答案。

4.2 评估指标

本文的评估指标为模糊匹配率(F1值)以及EM值。F1值为精确率(precision)与召回率(recall)的调和平均值，如公式7所示，precision的计算方式如公式5所示，lcs_len为模型预测答案与正确答案重合部分的长度，prediction_len为预测答案的长度。Recall的计算方式如公式6所示，其中answer_len表示正确答案的长度。EM值的取值只有两个：1和0，当预测答案和正确答案完全匹配时，EM取值为1，此外一切情况都等于0。

$$precision = \frac{lcs\_len}{prediction\_len} \tag{5}$$

$$recall = \frac{lcs\_len}{answer\_len} \tag{6}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

4.3 实验设置

本文使用到的预训练模型有：BERT-wwm-ext、roberta-base(Liu et al., 2019)、roberta-wwm-ext、roberta-wwm-ext-large，除roberta-wwm-ext-large的batch-size设置为4之外，其余模型的batch-size都为16，Epoch设置为2，learning-rate设置为3e-5。

4.4 结果分析

本文首先复现了CLUE(Xu et al., 2020)中部分预训练基线模型，观察对比其在CMRC2018阅读理解验证集上的表现，如Table1所示，roberta-wwm-ext在F1值上优于其它几个模型，达到87.9%，然而在EM值上表现最好的却是roberta-wwm-ext-large。

Model	F1(%)	EM(%)
bert-wwm-ext	86.526	65.207
roberta-base	86.597	67.661
roberta-wwm-ext-large	87.872	69.028
roberta-wwm-ext	87.912	68.344

Table 1: 复现模型在CMRC2018验证集上的结果

由于BERT-NTC的输入直接抽取的是BERT-MRC输入的Context部分，而之前的设置是：对于每一条输入数据，我们将输入的长度固定(选择BERT输入的最大长度512)，那么Context的长度如公式8所示，是用最大输入长度减去Question的长度以及三个标识符的长度，这就意味着每条输入数据中的Context部分可能包含不完整的标点句，但是小句复合体自动分析任务要求输入应该是完整的标点句，于是在基于NTC_attention_matrix的融合模型的基础上，我们对BERT-NTC以及BERT-MRC的数据预处理部分做了一些改动：利用滑动窗口机制，步长设置为一个标点句的长度，将每条输入数据的Context部分限制为一个个完整的标点句。

$$len_{\text{Context}} = 512 - len_{\text{question}} - 3\,([CLS], [CLS], [SEP]) \tag{8}$$

Model	F1(%)	EM(%)
roberta-wwm-ext	87.912	68.344
model-1	85.5(↓2.412)	65.082(↓3.262)
model-2	89.248(↑1.336)	73.812(↑5.468)
model-3	88.056(↑0.144)	73.501(↑5.157)
bert-wwm-ext	86.778	67.164
model-1	85.5(↓1.278)	65.082(↓2.082)
model-2	87.848(↑1.07)	70.923(↑3.759)
model-3	86.316(↓0.462)	70.985(↑3.821)
roberta-base	86.597	67.661
model-1	84.578(↓2.019)	64.15(↓3.511)
model-2	88.458(↑1.861)	71.917(↑4.256)
model-3	87.721(↑1.124)	70.27(↑2.609)
roberta-wwm-ext-large	87.872	69.028
model-1	82.461(↓5.411)	60.857(↓8.171)
model-2	90.26(↑2.388)	74.992(↑5.964)
model-3	87.484(↓0.388)	70.705(↑1.677)

Table 2: 基于小句复合体的机器阅读理解模型于CMRC2018验证集上的结果

加入小句复合体自动分析模型之后，模型的表现如Table2所示，其中model-1、model-2、model-3分别表示上文提出的基于finetune的融合模型、基于NTC_attention_matrix的融合模型以及加入滑动窗口机制之后的基于NTC_attention_matrix的融合模型。可以看出，model-1在baseline的基础上F1值和EM值分别下降了约2.4%以及3.3%，说明直接将roberta-wwm-ext替

(c) 2022 中国中文信息学会计算语言学专业委员会

换为BERT-NTC不可取，BERT-NTC是预训练模型在小句复合体自动分析任务上微调得来的，所以适用于小句复合体结构的分析，而机器阅读理解任务和小句复合体自动分析任务是两个完全不同的任务，因此，model-1在验证集上的结果甚至还不如baseline。

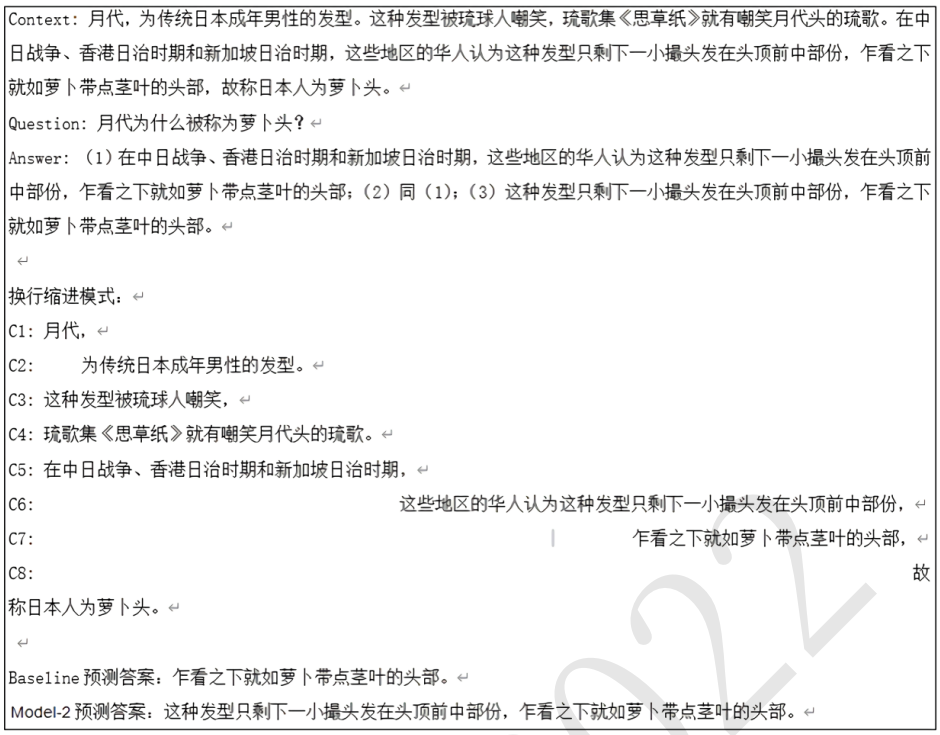


Figure 7: 融合小句复合体自动分析后的机器阅读理解答案预测

从实验结果中可以看出，model-2的表现最佳，F1值最高达到90.26%，相比于baseline提升了约2.4%，EM值最高达到74.992%，相比于baseline提升了约6%，说明小句复合体理论对机器阅读理解任务处理远距离成分共享问题有一定的贡献，具体实例如图7所示，换行缩进模式为小句复合体表现话头话身关系的一种形式，从图中可以看出，标点句C5到C8为处于同一话头结构内，C7共享C6的新支话头“这种发型”，也就是说，按照图示将C7补齐之后就成了“这种发型乍看之下就如萝卜带点茎叶的头部”，未添加小句复合体自动分析信息前，预测答案不完全准确，但在加入此信息后，模型更多关注到远距离成分共享的信息，预测答案完全正确。

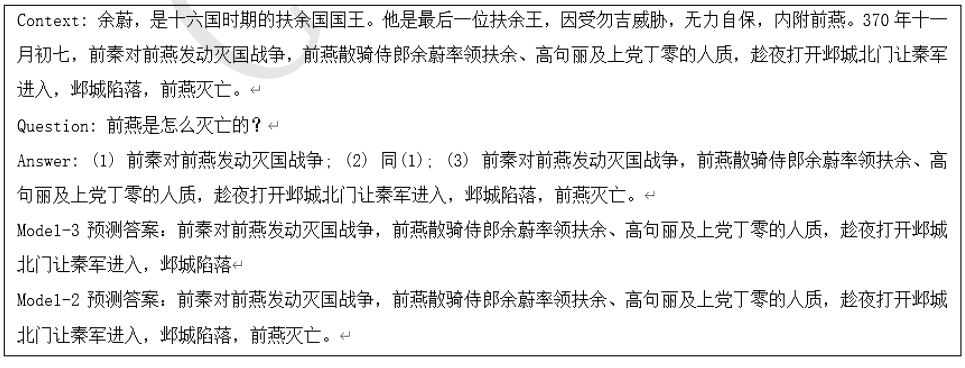


Figure 8: model-3与model-2预测答案对比

在model-2的基础上加入滑动窗口机制之后效果反而有所下降，其中EM值的差距为0.3%，这就说明相比于model-2，model-3预测答案长度过长或过短，具体实例如图8所示，此示例中，model-3预测的答案相较于正确答案过短，而model-2的预测答案和正确答案一致。同时，二者的F1值的差距较大，约为1%，经过分析发现，对于model-3预测不准确的问题，其大部分

(c) 2022 中国中文信息学会计算语言学专业委员会

预测的答案的长度要大于model-2预测答案的长度，而当预测答案过长时就会导致precision下降，从而使得F1值下降。同时，BERT不是以标点句边界为界来训练的，将输入限制为完整的标点句序列与BERT预训练方式不符，因此导致模型性能下降。

5 总结与展望

对于片段抽取式机器阅读理解任务，本文在已有的模型上做出了改进，将小句复合体自动分析与机器阅读理解相结合，增强了模型中的自注意力机制对话头的注意力，从而降低远距离成分共享信息抽取的难度。

从实验结果来看，本文提出的方法具有一定的成效，但仍有提升的空间。首先，小句复合体自动分析模型预测结果的准确率虽然达到了93%，但语料中绝大部分是不缺失成分的，而对于缺失成分的语料的预测结果准确率只达到了70%，因此，进一步提升小句复合体自动分析模型对缺失成分的预测准确率对机器阅读理解任务正确理解远距离成分共享语义信息是必不可少的。其次，对于长文本依赖的语义关系，模型输入长度限制使得模型不能观察到一个完整的紧密逻辑语义结构，适当提高模型输入的长度，使其能观察到一个完整的小句复合体的信息，可能对模型效果的提升有所帮助。

参考文献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. north american chapter of the association for computational linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Liu Xuan, Wu Tian, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. meeting of the association for computational linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. neural information processing systems.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation.

Wendy G. Lehnert. 1977. The process of question answering.

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

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. neural information processing systems.

Daniel W. Otter, Julian Richard Medina, and Jugal Kalita. 2021. A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. north american chapter of the association for computational linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. empirical methods in natural language processing.

Xingkai Ren, Ronghua Shi, and Fangfang Li. 2020. Distill bert to traditional models in chinese machine reading comprehension (student abstract). national conference on artificial intelligence.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. Learning.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and

(c) 2022 中国中文信息学会计算语言学专业委员会

Zhenzhong Lan. 2020. Clue: A chinese language understanding evaluation benchmark. international conference on computational linguistics.

Xiang Liu (刘祥), Ruifang Han, Shuxin Li (李舒馨), Yujiao Han, Mingming Zhang, Zhilin Zhao, and Zhiyong Luo (罗智勇). 2021. Shared component cross punctuation clauses recognition in chinese. international conference natural language processing.

宋柔. 2017. 小句复合体的理论研究和应用.

Ruiqi Wang (王瑞琦), Zhiyong Luo (罗智勇), Xiang Liu (刘祥), Rui Han (韩瑞), and Shuxin Li (李舒馨). 2021. 基于小句复合体的中文机器阅读理解研究(machine reading comprehension based on clause complex). Proceedings of the 20th Chinese National Conference on Computational Linguistics.

胡紫娟. 2020. 汉语小句复合体话头结构分析. 北京语言大学硕士学位论文.

顾迎捷, 桂小林, 李德福, 沈毅, and 廖东. 2020. 基于神经网络的机器阅读理解综述. Journal of Software, 31(7):2095–2126.

CCL 2022

基于神经网络的半监督CRF中文分词

罗智勇
北京语言大学
luo_zy@blcu.edu.cn

张明明
北京语言大学
MingmingZhang_blcu@163.com

韩玉蛟
北京语言大学
chloe_hanyu@163.com

赵志琳
北京语言大学
zhi_lin_zhao@163.com

摘要

分词是中文信息处理的基础任务之一。目前全监督中文分词技术已相对成熟并在通用领域取得较好效果，但全监督方法存在依赖大规模标注语料且领域迁移能力差的问题，特别是跨领域未登录词识别性能不佳。为缓解上述问题，本文提出了一种充分利用相对易得的目标领域无标注文本、实现跨领域迁移的半监督中文分词框架；并设计实现了基于词记忆网络和序列条件熵的半监督CRF中文分词模型。实验结果表明,该模型在多个领域数据集上F-值和ROOV值分别取得最高2.35%和12.12%的提升，并在多个数据集上成为当前好结果。

关键词： 半监督 ； 序列条件熵 ； 跨领域 ； 中文分词

Semi-supervised CRF Chinese Word Segmentation based on Neural Network

Zhiyong Luo
北京语言大学
luo_zy@blcu.edu.cn

Mingming Zhang
北京语言大学
MingmingZhang_blcu@163.com

Yujiao Han
北京语言大学
chloe_hanyu@163.com

Zhilin Zhao
北京语言大学
zhi_lin_zhao@163.com

Abstract

Chinese word segmentation (CWS) is a fundamental task of natural language processing. Currently, CWS model using fully supervised learning technology has achieved good results in the common domain. However, it has the problem of relying on the large-scale annotated corpus and poor domain migration capability, especially the cross-domain OOV word recognition is not effective. In order to alleviate these problems, this paper proposes a semi-supervised CWS framework that uses relatively easy-to-obtain unlabeled texts in the target domain to achieve cross-domain transfer. We design a semi-supervised model based on word memory network and sequence conditional entropy. Our model based on this framework achieves significant improvements in F-scores and ROOV on several datasets,some of them are state-of-the-art.The maximum F-value and ROOV improvements are 2.35% and 12.12%.

Keywords: Semi-supervised learning , Conditional entropy for sequence , Chinese word segmentation

第二十一届中国计算语言学大会论文集，第644页-第655页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

中文文本可视为由汉字字符（包含标点）组成的连续字符串，且词间无明确标记。因此，在中文信息处理任务中，分词通常是词法分析层面的一项重要任务，准确的分词结果有助于提升基于词的深层次语言信息处理任务的性能。

Xue(2003)将中文分词转化为给汉字标注词位的序列标注任务，自此基于字序列标注方法被广泛应用于中文分词。近年来，随着深度学习技术的发展，基于神经网络的全监督序列标注模型逐渐应用于中文分词任务中 (Chen et al., 2015; Cai and Zhao, 2016; Cai et al., 2017)，不仅提升了歧义切分的性能，还在通用领域（新闻领域）中文分词上取得较好的结果。例如，He et al.(2019)基于多准则的通用领域（PKU数据集）中文分词F-值达96%以上。但是，全监督的分词方法依赖大规模人工标注语料，并且模型领域迁移能力差，训练数据和测试数据之间差异对模型效果影响很大。目前，中文分词的标注语料主要来自于新闻领域，特定专业领域（如医学、知识百科、小说等）标注语料稀少。因此，尽管全监督分词模型在通用领域分词任务上的准确率较高(He et al., 2019)，但是受限于其领域迁移能力，跨领域分词准确率和未登录词的识别性能都有待提升；此外，由于专业领域缺乏大规模标注语料，因而也无法通过全监督方式训练有效的领域分词器。但是，较于昂贵的专业领域标注语料，专业领域无标注文本却相对易得。因此，如何充分利用用现有大规模通用领域标注语料和专业领域无标注语料，构建具备较好的分词准确率和未登录词识别性能的半监督分词模型,是近年来中文分词任务关注的问题之一。

邓丽萍and罗智勇(2017)首次提出以条件熵作为正则化项，利用CRF++工具和人工特征模板构建半监督CRF并应用到跨领域分词任务上，使得百科领域分词F-值与未登录词召回率有效提升。但该模型基于统计机器学习的方法，存在需手工定制特征模板的弊端且不适用于神经网络框架；Fu et al.(2020)研究表明用大规模预训练语言模型BERT、ELMo等作为字符特征编码器有助于更好地提升未登录词识别的性能；Tian et al .(2020)构建了词记忆网络，将n-gram信息编码到上下文表示中来，提高了模型对于词语边界的预测能力，但是该方法未能有效利用目标领域无标注文本中的字符共现特征，因而领域迁移能力有限。

在上述研究的基础上，本文提出了一种基于神经网络的半监督跨领域序列标注框架，并实现了两种基于该框架的半监督中文分词模型：以BERT作为特征提取器的半监督CRF模型（记为BERT-semiCRF），以及增加词记忆网络的半监督CRF模型(记为BERT-WM-semiCRF)。上述模型将已标注通用领域分词语料和无标注专业领域文本作为输入，不仅可以通过编码无标注文本上下文信息，提高模型领域迁移能力和未登录词识别性能，还可以通过减小序列条件熵增强模型预测置信度，从而实现跨领域分词准确率与未登录词识别性能的有效提升。实验表明，在特定专业领域（如专利PT、小说ZX/FR、医学DM等数据集上F-值和未登录词召回率达当前最好，相较于基线模型，F-值提升最高达2.35%，未登录词召回率提升最高达12.12%。

本文第2节介绍中文分词的相关研究；第3节介绍基于神经网络的半监督CRF模型；论文的第4节介绍模型的实验验证、消融研究和实验分析；第5节为总结和研究展望。

2 相关研究

2.1 半监督中文分词

中文分词方法主要分为全监督、无监督和半监督三类。全监督分词方法要求训练数据集全部为有标注数据，因此存在依赖大规模标注语料且跨领域迁移能力差的缺陷；无监督分词方法的训练集则全部为无标注数据，主要用于新词发现任务，其分词准确性较低；而半监督分词方法指训练数据集中既包含有标注数据，也包含无标注数据的分词方法。目前，半监督中文分词方法主要分为两类：一类半监督方法通过设计损失函数 (Liu et al., 2014; Zhao et al., 2018)，使其能够利用标注和部分标注的中文分词数据训练，但是无法避免来自网络的自然标注数据带来的歧义、标注准则不一致的问题；另一类半监督方法通过自采样得到伪标注从而扩充训练样本(Liu and Zhang, 2012)，这种方法存在错误累加和未登录词别受限的问题；Wang and Xu (2017)使用训练好的教师模型自采样无标注数据构建词表，然后使用预训练方法获得该词

©2022 中国计算语言学大会根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家自然科学基金(62076037)、北京语言大学研究生创新基金(中央高校基本科研业务费专项资金)项目成果(22YCX173)

表的向量表示，并将该信息以词嵌入方式增加到学生模型的全监督训练中，这是一种通过增加无标注文本特征实现半监督的方法，但是该方法缺乏对未登录词的关注；针对跨领域分词任务，Ding et al.(Ding et al., 2020)等人提出基于远距离标注和对抗训练的分词模型。该方法在F-值上有一定的提升，但需要为每个领域从初始状态训练模型，耗时较长，且分词效果很受无监督词挖掘器影响。

本文区别于上述研究工作，从增加无标注文本特征表示、设计可同时关注标注数据和完全无标注数据的损失函数这两个角度，构建基于神经网络的半监督CRF分词模型。

2.2 基于神经网络序列标注的中文分词

基于神经网络的序列标注模型通常包含字/词表示、特征提取、推理三大模块(He et al., 2020)。采用序列标注方法的中文分词，是以字为单位，因此第一模块为字表示。字表示模块将字映射到对应的表示向量上，该向量蕴含字意信息,早期使用上下文无关的静态向量，例如word2vec(Mikolov et al., 2013)等，现常用包含语境信息的动态向量，例如ELMo(Peters et al., 2018)、BERT(Devlin et al., 2018)等。特征提取模块又称上下文编码模块，用于捕获序列的上下文信息和边界信息，在基于神经网络的序列标注模型中通常将经过特征提取模块的向量表示作为发射状态矩阵，BILSTM常作为特征提取模块(Huang et al., 2015)。BERT既可作为独立的字表示模块，也可以视为字表示与特征提取为一体的表示层。推理模块的功能是给出输入序列的标注结果，常用可关注标签间转移关系的条件随机场(conditional random filed,CRF)。

本文提出的半监督框架建立在字表示、特征提取和推理模块结构上，BERT和加入词记忆网络的BERT作为字表示和特征提取层，基于神经网络的semiCRF作为推理模块。

3 基于神经网络的半监督CRF分词

本节将详细介绍半监督模型框架、模型具体结构和训练策略。首先形式化定义输入、输出并描述整个模型的半监督框架；接着介绍基于神经网络的半监督CRF模型结构和具体实现细节；最后介绍模型的训练策略。

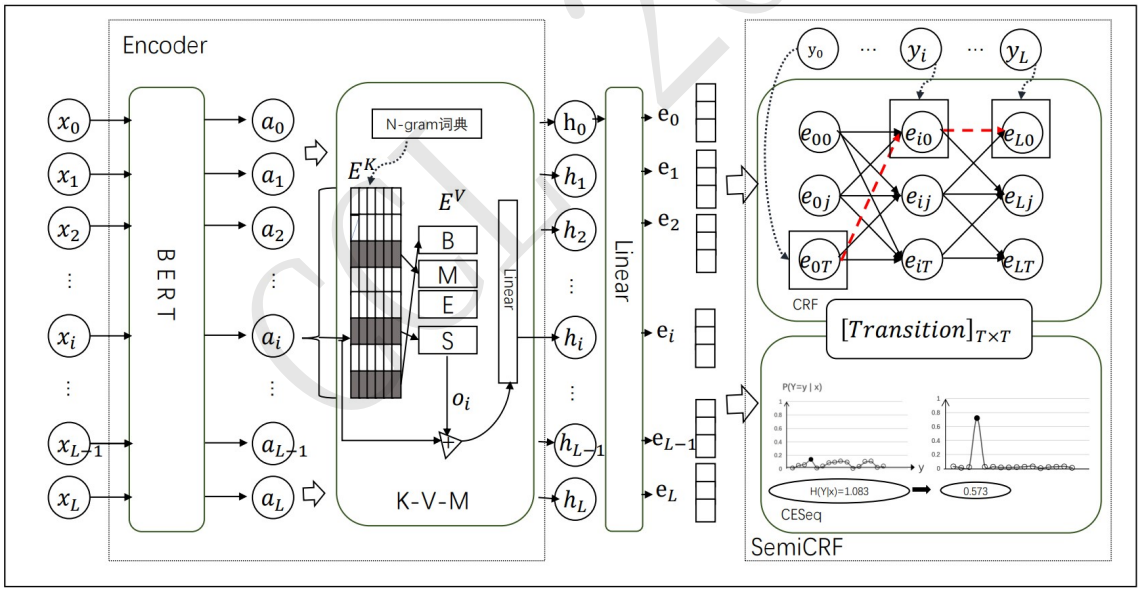


Figure 1: 模型结构

3.1 半监督框架

基于字标注的中文分词任务，通常形式化为：给定长度为n的句子x，其标注分词结果序列为y ,其中：$x = \langle x_1, x_2, \ldots, x_i, \ldots, x_n \rangle$, x_i表示句中第i个字，$y = \langle y_1, y_2, \ldots, y_i, \ldots, y_n \rangle$,且$y_i \in \{B, M, E, S\}$，分别表示词语的开始、中间、结束位置，以及单独成词。例如句子“一/面/鲜艳/的/五星红旗”的切分形式，可以表式为：x = <一，面，鲜，艳，的，五，星，红，旗>，y = $< S, S, B, E, S, B, M, M, E >$。在模型的训练阶段，全监督模型的输入为$(x^i, y^i)$

(c) 2022 中国中文信息学会计算语言学专业委员会

, $x^i \in X = \left\{x^1, x^2, .., x^N\right\}, y^i \in Y = \left\{y^1, y^2, .., y^N\right\}$ 其中X, Y 分别表示标注数据集的字符序列集合和标签序列集合，N 表示标注数据集样本规模。与全监督分词不同的是，半监督分词模型的输入是标注样本(x^i, y^i) 或无标注样本(x^j)，$x^j \in U = \left\{x^1, x^2, .., x^M\right\}$，其中$U$表示无标注数据的样本集。半监督模型的优化目标是最小化模型在训练数据上的损失：

$$\theta^* = \arg\min_{\theta} \left(M_\theta\left(X, Y\right) + \beta * M_\theta\left(U\right)\right) \tag{1}$$

M表示半监督模型，θ表示模型参数，$M_\theta\left(\bullet\right)$ 表示模型在数据集上的损失值，β为控制模型对无标注数据关注程度的超参数，当$\beta = 0$时，将变为全监督模型。

图1给出了本文提出的半监督模型整体结构，主要包括输入层、编码层$Encoder_{\theta_1}$、发射状态编码层和半监督CRF层$semiCRF_{\theta_2}$。

输入层 对于任何一个作为输入样本的句子$x = \langle x_1, x_2, \ldots, x_n\rangle$，$x \in X$ 或$x \in U$。在样本输入模型前，预处理工作将“$[CLS]$”和“$[SEP]$”加入到样本首尾,并将句子填充为所设最大长度L。此时样本表示为$x = \langle x_0, x_1, \ldots, x_L\rangle$。

编码层($Encoder_{\theta_1}$) 编码层主要获得具有上下文和边界信息的特征表示$h = \langle h_0, h_1, \ldots, h_L\rangle$, $h \in R^{L \times d_h}$, d_h为隐藏层维度,θ_1为模型中参数。主要核心模块包括预训练语言模型和词记忆网络（请见3.3节）。

发射状态编码层 该层将每一个字符的上下文编码特征向量，通过全连接层投射到字标注类别的未归一化概率，即发射状态$e \in R^{L \times T}$,如公式(3)所示，$w^T \in R^{d_h \times T}$，$\mathrm{b} \in R^{1 \times T}$，$T$为发射状态维度，即可预测标签数。

半监督CRF层($semiCRF_{\theta_2}$) 该层主要包含一个CRF模块和一个序列条件熵计算模块CEseq，以及两个模块之间共享的标签转移矩阵B，其中θ_2为标签转移矩阵参数。若原输入x来自标注数据集X，模型将通过CRF模块计算对应标签序列y的负对数似然作为模型预测损失值；若x来自无标注数据集U，模型则由序列条件熵模块CESeq计算模型预测的损失值。具体计算方法详见3.2节。

整个网络前向传递过程如公式(2)-(4)所示：

$$h = Encoder_{\theta_1}\left(x\right) \tag{2}$$

$$e = h * w^T + \mathrm{b} \tag{3}$$

$$loss = \begin{cases} semiCRF_{\theta_2}\left(e\right) & x \in U \\ semiCRF_{\theta_2}\left(e, y\right) & x \in X \end{cases} \tag{4}$$

在预测阶段，对于待预测样本,获得发射状态步骤与训练阶段相同，即式(2)、(3)；在获得发射状态后由semiCRF中CRF模块使用维特比算法解码，获得最大预测概率的标签序列，可形式化为式(5)：

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p_\theta(y|x) \tag{5}$$

其中,$\mathcal{Y}$表示输入序列x所有可能标注序列集合。

本文实现的BERT-semiCRF和BERT-WM-semiCRF的Encoder分别对应BERT、加上词记忆网络的BERT(Tian et al., 2020)，即：$h = BERT_{\theta_1}\left(x\right)$ 和$h = BERT_WM_{\theta_1}\left(x\right)$，其中词记忆网络结构在3.3节说明。此处值得注意的是，本文实现的BERT-WM-semiCRF模型在构建n-gram词典时，数据源范围较原方法增加了无标注数据U，用无监督的邻接多样性(Feng et al., 2004a)方法构建,以便获得目标领域文本的字符组合特征表示信息。

3.2 基于神经网络的半监督条件随机场

本文实现的基于神经网络的半监督条件随机场（记为semiCRF）由推理模块CRF和序列条件熵(Conditional Entropy for Sequence,CESeq)两部分组成，它们共享标签间转移矩阵B。该模型具备提取标签之间的转移特征、计算损失值和解码功能。如式(6)所示，有标注样本损失为条件概率的负对数似然，无标注样本的损失为条件熵。

$$loss = \begin{cases} H_\theta\left(\mathcal{Y}|x\right) & x \in U \\ -logp_\theta\left(y|x\right) & x \in X, y \in Y \end{cases} \tag{6}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

CRF (Huang et al., 2015; 李航and others, 2012)模块关注于有标注样本，它描述给定输入序列x产生标注序列y的条件概率为：

$$\begin{aligned} \mathrm{p}\left(y|x\right) &= \frac{\Psi(y|x)}{Z(x)} \\ Z\left(x\right) &= \sum_{y\in\mathcal{Y}} \Psi\left(y|x\right) \end{aligned} \tag{7}$$

其中，$\mathcal{Y}$表示x所有可能标注序列集合，当序列长度为L、标签数为T时，$\mathcal{Y}$集合大小为T^L。$\Psi(y|x)$ 是x标注为y的打分函数：

$$\Psi\left(y|x\right) = \prod_{i=1}^{L} \varphi\left(x, i, y_{i-1}, y_i\right) \tag{8}$$

$$\varphi\left(x, i, y_{i-1}, y_i\right) = \exp\left(e_{i,y_i} + B_{y_{i-1},y_i}\right) \tag{9}$$

其中，发射状态e由式(3)获得，$e \in R^{L\times T}$ ，e_{i,y_i}表示x的i号位置标注为标签y_i的发射状态分值；B_{y_{i-1},y_i}表示由标签y_{i-1} 转移到y_i的转移分值。为避免式(7)中$Z\left(x\right)$ 指数级时间复杂度，通常使用动态规划算法，对$Z\left(x\right)$通过以下递推公式求解：

$$Z\left(x\right) = \sum_{i=1}^{T} \exp\left(\alpha_L\left[i\right]\right) \tag{10}$$

$$\alpha_t = \log \sum_{i=1}^{T} \exp\left(\mathrm{M_t}\left[\mathrm{i}\right]\left[\mathrm{j}\right]\right) \tag{11}$$

$$\mathrm{M_t} = \begin{bmatrix} \alpha_{t-1} & \dots & \alpha_{t-1} \end{bmatrix}_{T\times T} + B + \begin{bmatrix} e_t \\ \dots \\ e_t \end{bmatrix}_{T\times T}, t = 1, 2, \dots, L \tag{12}$$

上式中，e_0为样本经过编码层和线性层后的第一个时序，即式(3)中e的第一个时序，$e_0^T \in R^{T\times 1}$ ，T 表示标签数，L 表示序列长度。

序列条件熵计算模块CESeq，关注于无标注数据的序列标注条件熵，表示在给定输入x的情况下,所有可能标注序列$\mathcal{Y}$的概率分布的不确定性，记为$H(\mathcal{Y}|x)$。当$\mathcal{Y}$的不确定性越小时，对于x的各可标注结果y'间的概率区分度越大，满足低密度划分原则。具体计算方式如公式(13)所示：

$$\mathrm{H}\left(\mathcal{Y}|\mathrm{x}\right) = -\sum_{y'\in\mathcal{Y}} p(y'|x)\log(y'|x) \tag{13}$$

上式中，$p(y'|x)$由式(6)定义。为避免遍历$\mathcal{Y}$所带来的指数级时间代价，邓丽萍and罗智勇(2017)采用子序列条件熵对(13)式进行化简，达到了和CRF相同时间复杂度级别$O(LT^2)$：

$$H\left(\mathcal{Y}|x\right) = -\sum_{y_n} p(y_n|x)\left[\log p(y_n|x) + H^{\alpha}(\mathcal{Y}_{1,2,\dots n-1}|y_n, x)\right] \tag{14}$$

$$H^{\alpha}(\mathcal{Y}_{1,2,\dots,t}|y_{t+1}, x) = \sum_{y_t} p(y_t|y_{t+1}, x)[\log p(y_t|y_{t+1}, x) + H^{\alpha}(\mathcal{Y}_{1,2,\dots t-1}|y_t, x)] \tag{15}$$

式中，n表示序列长度，y_i表示序列i号位置被标注的标签类别，$\mathcal{Y}$表示x可标注序列集合，$t \in \{0, \dots, n\}$，$H^{\alpha}\left(|y_0, x\right) = 0$ 。其中$p\left(y_t|y_{t+1}, x\right)$可有由(16)、(17)式获得：

$$\mathrm{p}\left(y_t y_{t+1}|x\right) = \frac{\sum_{y'\in\{\mathcal{Y}\cap y_t y_{t+1}\}} \Psi(y'|x)}{Z(x)} \tag{16}$$

$$\mathrm{p}\left(y_{t+1}|x\right) = \frac{\sum_{y'\in\{\mathcal{Y}\cap y_{t+1}\}} \Psi(y'|x)}{Z(x)} \tag{17}$$

$\mathrm{p}\left(y_{t+1}|x\right)$表示序列$x$的$t+1$号位置被标注为标签$y_{t+1}$ 的条件概率，$\Psi(\cdot|x)$由式(8)定义。集合$\{\mathcal{Y}\cap y_{t+1}\}$表示$x$的$t+1$号位置被标注为$y_{t+1}$的所有可能标注序列集合，集合$\{\mathcal{Y}\cap y_t y_{t-1}\}$同

CCL 2022

理，表示t、$t+1$ 号位置被标注为$y_t y_{t+1}$ 的所有可能标注序列集合。本文对式(16-17)进一步推导，直接给出计算对数标签转移概率$p\left(y_t|y_{t+1},x\right)$的前向递推公式：

$$\log p(y_t = i|y_{t+1} = j, x) = \mathrm{M_{t+1}\,[i]\,[j]} - \alpha_{t+1}\,[j] \tag{18}$$

式(18)采用动态规划算法，其中$\mathrm{M_{t+1}}$与α_{t+1}由式(11)、(12)定义。

整体来看，建立在神经网络自动编码结构上的半监督条件随机场具备以下优势：能够关注无标注数据的文本特征，可以提高模型的领域迁移能力；CRF模块具备监督模型习得正确识别词边界的能力，CESeq模块能够扩大可能标注的结果序列之间的概率区分度。上述两部分组合，使模型具备对序列正确标注且正误区别明显的能力。从本文4.3.3节中的消融研究结果来看，加入semiCRF后的半监督中文分词的准确率、未登录词识别性能均有提升，可印证上述观点。

3.3 词记忆网络

词记忆网络(Wordhood Memory Network)(Tian et al., 2020)是对基于预训练语言模型的上下文特征编码的增强。通过加入键值网络结构，为每个输入的上下文编码表示a_i，增加了n-gram信息。在给定的上下文情况下，这些n-gram信息将显式地增加或降低当前字符成为某个词位的可能性或概率。具体计算方法如式(19)、(20)所示，其中,c_i表示输入字符i在当前语境下可查找到的m个n-gram,$E_{c_i}^K$为c_i的向量表示，例如，“分”在语境“原子结合成分子”中的可查找到的ngram有：“成分”、“分子”。n-gram词典由输入文本采用邻接多样性(Feng et al., 2004b)的无监督方法构建；$E_{c_i}^V$为当前字符i在各n-gram中对应的可能词位(BMES)的向量表示。

$$o_i = softmax\left(a_i, E_{c_i}^K\right) \cdot E_{c_i}^V \tag{19}$$

$$h_i = Linear\left(o_i + a_i\right) \tag{20}$$

3.4 模型训练

针对本文提出的半监督框架，我们提出了两种训练策略：联合训练和分步训练策略。为避免模型在无标注数据集上过度关注区分度而忽略正确性，设计如式(21)损失函数，与训练目标式(1)相对应，包含标注数据损失和无标注数据损失，意在使模型在关注区分度时，不偏离标注准则。

$$\mathrm{loss} = -\frac{1}{N}\sum_{i=1}^{N} logp_{\theta_1}\left(y^i|x^i\right) + \frac{\beta}{M}\sum_{j=1}^{M} H_{\theta_2}\left(\mathcal{Y}|x^j\right) \tag{21}$$

回顾模型的整体结构(图1)，来自源领域的标注数据或目标域的无标注数据x都会经过共享的上下文特征编码结构和发射状态编码层得到发射状态编码e；随后，在semiCRF层中,模型根据源领域提供的标注y，由CRF模块计算得到其负对数似然损失；属于目标领域的样例则进入CESeq模块，计算得到序列条件熵损失。模型的损失是上述两部分损失之和。值得说明的是，对于通用领域中文分词任务，本模型源领域与目标领域均为新闻领域，对于跨领域分词，不同目标域仅使用该领域对应的无标注数据。

本文实现的分步训练策略指先在源领域有标注训练集上对模型全监督训练，此时式(21)中β设置为0，在模型习得一定分词能力后再设定非零β进行半监督训练。联合训练策略则不再包含全监督训练步骤，直接使用非零β进行半监督训练，其中，每个batch中标注样本和无标注样本比例为1:1。

4 实验

本节介绍实验使用数据集和实验设置，然后展示我们实现的模型在各领域的分词效果、与其它工作的结果比较、消融研究，并进行分析与总结。

4.1 实验数据集

本文实验在四个领域的六个数据集上进行，包含：新闻领域的标注语料PKU（来自SIGHAN CWS BACKOFF 2005），是1998年人民日报1月份新闻文本；无标注语料pku-07，是1998年人民日报7月份新闻文本；专利领域（PT）、医学领域（DM）、小说领域

(c) 2022 中国中文信息学会计算语言学专业委员会

（ZX、DL和FR）5份无标注数据集和标注测试集均来源于(Ding et al., 2020)文献，其中小说领域三个语料分别是网络小说《诛仙》(ZX)、《斗罗大陆》(DL)、《凡人修仙传》(FR)。表1中列出了各语料详细信息，其中PKU的训练集和开发集为SIGHAN CWS BACKOFF 2005的训练集根据9:1划分所得，测试集与原测试集保持一致。为方便比较，其他领域数据集划分均与参考文献中保持一致。因语料存在以段落为间隔的情况，导致一个样本长度过长，因此本实验中对数据进行了以句号、问号、引号、省略号为分割符的分句处理。

数据集		句数	领域	标注
PKU	训练集	39.7K	新闻	是
	开发集	4.4K		
	测试集	4.2K		
pku-07	训练集	57.2K	新闻	否
PT	训练集	17.7K	专利	否
	测试集	1.0K		是
DM	训练集	32.0K	医学	否
	测试集	1.0K		是
ZX	训练集	59.0K	小说	否
	测试集	1.0K		是
DL	训练集	40.0K	小说	否
	测试集	1.0K		是
FR	训练集	148.0K	小说	否
	测试集	1.0K		是

Table 1: 实验数据

参数	
序列最大长度L	300
学习率	0.00001
隐藏层维度d_h	768
标签数T	7
超参数β	{0.1，0.05，0.01}
dropout	0.5
随机种子	42
n-gram词典阈值	3

Table 2: 参数设置

4.2 实验设置

实验中BERT使用"bert-base-chinese"作为初始模型参数，BERT-WM-SemiCRF 中词记忆网络的设置参考WMSeg(Tian et al., 2020)。其他参数设置如表2所示。中文分词的测评指标通常包含准确率（P）、召回率（R）、平衡P和R的F-值、未登录词召回率（ROOV）,本实验中采用F-值和ROOV作为主要评价指标，具体计算方式和前人工作保持一致，这里不再赘述。

4.3 实验结果与消融研究

为验证本文方法的有效性，本文以通用领域语料PKU在全监督方式下训练模型BERT-CRF和WMSEeg，获得通用领域分词器，在PKU和其他领域测试集上进行测评，以上述测评结果作为本文比较的基线。此外，本文方法也与近年其他方法在同数据集上的分词结果进行了比较。本文实现的模型在通用领域PKU数据集上最高F-值达96.76%，ROOV达87.48%，成为当前最好结果；在跨领域分词任务上，模型结果较两个基线模型在F-值和ROOV上均有提升，其中，专利领域F-值最大提升达2.78%、未登录词召回率提升达7.93%。

4.3.1 通用领域

本文实现的半监督模型在通用领域的实验结果如表3所示。我们首先将本文实现模型与全监督方法进行对比，表格1-3行为近年来基于神将网络的全监督中文分词在PKU数据集上的分词结果。接着我们列举了半监督分词模型的实验结果，为表格4-6行，WCC-CWS(Hao et al., 2017)是一种增加上下文特征表示的半监督方法，WE_CONV_SEG(Wang and Xu, 2017)则通过特征蒸馏的自采样实现，BILSTM_LM_PL(Zhao et al., 2018)构建了由交叉熵和语言模型组合的损失函数实现半监督；除WMSeg的实验结果为本文复现该论文结果，其他结果摘自原文献。本文实现的模型结果为表格最后三行，BERT-CRF是本文的基线模型，其它两个半监督模型由有标注的PKU训练集和无标注的人民日报中文文本训练。观察实验结果，可以发现：（1）本文方法实现的两个半监督模型，在未登录词召回率上超出所列全监督模型和半监督模型；其中，BERT-WM-semiCRF模型的实验结果达到当前最好，为96.76%和87.48%；（2）本文实现的两个模型较基线模型F-值分别提升0.12%、0.19%，未登录词召回率提升1.12%、1.34%，说明本文的模型不仅能提升分词准确率，在识别未登录词识方面提升更为显著；（3）与BILSTM_LM_PL使用部分标注语料实现半监督分词的结果比较，本文实现的半监督模型

(c) 2022 中国中文信息学会计算语言学专业委员会

MODEL	F	ROOV
BILSTM_CWS(2018)	96.10%	78.80%
MC_CWS(2019)	96.41%	78.91%
WMSeg(BERT)(2020)	96.73%	86.90%
WCC_CWS(2017)	96.00%	-
WE_CONV_SEG(2017)	96.50%	-
BILSTM_LM_PL(2018)	95.50%	-
our model		
BERT-CRF(BASE)	96.56%	86.13%
BERT-semiCRF(step)	96.69%	87.25%
BERT-WM-semiCRF	**96.76%**	**87.48%**

Table 3: 通用领域PKU数据集分词结果

高约1个百分点，一方面由于本文的特征提取层较BILSTM能更好的提取无标注文本上下文特征，另一方面，相较于构建语言模型和交叉熵损失的半监督方法，本文的semiCRF方法既能关注无标注文本，又避免了部分标注文本带来的噪音。

4.3.2 跨领域分词

	PT		ZX		DM		FR		DL	
	F	ROOV	F	ROOV	F	ROOV	F	ROOV	F	ROOV
Partial-CRF(2014)	85.00%	-	83.90%	-	82.80%	-	90.20%	-	92.50%	-
WCC_CWS(2017)	-	-	89.10%	70.40%	-	-	-	-	-	-
WEB_CWS(2019)	85.10%	-	89.60%	-	82.20%	-	89.60%	-	93.50%	-
DAAT(2020)	89.60%	-	90.90%	-	85.00%	-	93.10%	-	**94.10%**	-
WSEeg(2020)	91.02%	78.64%	90.48%	69.42%	88.55%	75.48%	91.21%	73.74%	92.22%	65.03%
ours:										
BERT-CRF	90.58%	76.36%	90.20%	69.00%	88.39%	74.67%	90.82%	72.78%	92.21%	65.82%
BERT-semiCRF	93.30%	81.79%	91.02%	71.01%	89.72%	78.90%	**93.15%**	83.89%	92.58%	**66.48%**
BERT-WM-semiCRF	**93.37%**	**84.29%**	**91.11%**	**72.98%**	**90.43%**	**80.73%**	93.01%	**85.86%**	92.64%	66.22%

Table 4: 跨领域分词实验结果

表4给出了本文在专利(PT)、医学(DM)、小说（ZX、FR、DL）三个领域5个数据集上的跨领域分词实验结果。表中最后三行列出了我们实现的全监督基线模型（BERT-CRF）、两个半监督模型在各领域数据集上的实验结果，其中半监督模型均由有标注的新闻领域PKU训练集、无标注的目标领域中文文本联合训练得到。表格1-4行，列出了近年使用半监督方法进行跨领域分词在相同的数据集上的实验结果。其中，DAAT模型在各跨领域分词上的F-值为最高(Ding et al., 2020)，本文与该模型训练数据集的设置一致。表格第5行WSEeg(Tian et al., 2020)是用PKU训练的全监督模型的跨领域分词结果。观察实验结果，可以发现：（1）本文实现的半监督模型除DL数据集F值外，其它评估值均达到最好结果，其中BERT-semiCRF在FR小说数据集的F值上最好，达93.15%，其余则是加上词记忆网络的BERT-WM-semiCRF有更高的结果；（2）DAAT为近年来半监督方法在以上数据集最好的结果，本文在各数据集上的分词准确率提升分别为：从89.60%到93.37%，从90.90%到91.11%，从85.00%到90.43%，从93.10%到93.15%，最高提升达5.43%，对应医学数据集；（3）本文实现的半监督模型较全监督的基线模型，在分词准确率和未登录词召回率上都有提升，这一点表5展示的更清晰，专利领域F-值提升最大，为2.72%，FR小说领域未登录词识别提升最大，达12.12%；

4.3.3 消融研究

为更清晰展示本文方法的效果，表5列出了序列条件熵和词记忆网络模块在新闻、医学、专利、小说各领域数据集上的消融实验结果比较。BERT-CRF 、WMSeg(Tian et al., 2020)为两个对比的基线模型，均使用用PKU训练集全监督训；BERT-semiCRF相较于基线模型1，使用了semiCRF和与测试集领域对应的无标注数据U，而semiCRF相较于CRF增加了计算序

列条件熵的辅助推理模块CESeq,因此信息列BERT-semiCRF与对比基线模型BASE1的区别记为：+U+CESeq。BERT-WM-semiCRF在BERT-semiCRF基础上增加了词记忆网络，因此较基线模型BASE1的区别为增加了辅助训练的无标注数据、词记忆网络、辅助推理模块，因此记为：+U+K-V-M+CESeq。分析表格数据，可以发现：

（1）本文提出的半监督框架下的两个分词模型，在同领域（PKU）和跨领域（DM、PT、FR）上的分词F值和未登录词召回率较全监督基线模型均有提升，说明本文的半监督方法在提高模型分词性能和未登录词识别能力上有效；

（2）表中第3、4行，增加无标注数据和模块CESeq模型在各领域结果均有提升，其中专利领域F值提升达2.72%，小说FR未登录词召回率提升达12.12%，说明增加模块CESeq的semiCRF的半监督方法有效；对比3、4行的提升幅度，发现基线模型未登录词识别相对较低的数据集，未登录词召回率提升更明显，说明模型有较强泛化能力和迁移能力；

（3）对比表中4、5行的提升幅度，第5行提升值大于第4行，说明在半监督框架下增加词记忆网络也有助于分词性能提升，但是它的提升效果不如增加模块CESeq；

（4）与通用领域的提升幅度相比较，跨领域分词提升幅度更明显，一方面由于跨领域分词任务难大，基线模型待提升空间大；另一方面由于本文的最小化无标注数据序列条件熵的辅助模块（CESeq）使得模型可以学习目标领域的信息，平衡模型"看到"的目标域和源领域的数据分布，从而达到提升效果。

			PKU		DM		PT		FR	
模型名称	比较	信息	F	ROOV	F	ROOV	F	ROOV	F	ROOV
BERT-CRF	-	BASE1	96.56%	86.13%	88.39%	74.67%	90.58%	76.36%	90.82%	72.78%
WMSeg	-	BASE2	96.73%	86.90%	88.55%	75.48%	91.02%	78.64%	91.21%	73.74%
BERT-semiCRF	BASE1	+U+CESeq	+0.12%	+1.12%	+1.34%	+4.23%	+2.72%	+5.44%	+2.33%	+11.12%
BERT-WM-semiCRF	BASE2	+U+CESeq	+0.03%	+0.58%	+1.87%	+5.25%	+2.35%	+5.65%	+1.80%	+12.12%
	BASE1	+U+K-V-M+CESeq	+0.19%	+1.34%	+2.04%	+6.06%	+2.78%	+7.94%	+2.19%	+13.09%

Table 5: 消融实验

4.4 实验分析

4.4.1 模型预测结果的置信度

数据集	BERT_CRF	BERT_semiCRF	BERT_WM_semiCRF
PKU	0.542	0.577	0.569
DM	0.649	0.694	0.719

Table 6: 模型平均置信度

为进一步说明semiCRF的效果并证明该方法的有效性，我们分别使用基线模型（BERT-CRF）、两个半监督模型为PKU、DM测试集样本的正确标注结果打分，表6给出评分均值，分值取值是[0,1]的概率表示。模型为正确标注序列打出的分值越高，说明模型对正确标注结果的置信度越大。从表中可以看出，增加辅助推理模块的半监督模型较全监督模型的平均分值均有提升，说明增加序列条件熵作为semiCRF的辅助推理模块，模型预测置信度会更高，从而提高了分词效果。

4.4.2 训练策略与超参数的选择

本文根据半监督结构提出了两种训练策略：分步训练策略和联合训练策略。表8给出半监督模型分步、联合两种策略下的实验比较，同领域（PKU）分词，采用联合训练策略有较好的结果，而在跨领域分词任务中，使用分步训练会取得相对较好的实验结果。由于联合训练需要从初始状态训练模型，而分步训练可在半监督训练时，加载同一个已训练好的全监督模型再训练，会节省训练耗时。

(c) 2022 中国中文信息学会计算语言学专业委员会

	PKU		FR	
joint	96.76%	87.48%	92.58%	83.39%
stepwise	96.69%	87.25%	93.15%	83.89%

Table 7: 不同训练策略下的实验结果

本文在3.2节半监督框架中指出，式(21)中的超参数β是模型对无标注文本的关注程度的调节，表给出了模型在不同β值下在专利领域PT上的评测结果，β的选择对结果存在一定幅度的影响，根据实验经验，选择较大β值会使得模型过分关注无标注文本，忽略分词标准的学习，导致结果降低，因此，在选择β值时应从较小值选取。

MODEL	β	F	ROOV
BERT-semiCRF	0.1	91.77%	83.12%
	0.05	93.01%	83.30%
	0.01	93.30%	81.80%
BERT-WM-semiCRF	0.1	93.00%	82.39%
	0.05	92.93%	83.04%
	0.01	93.37%	84.29%

Table 8: 不同β的影响

4.4.3 未登录词分析

本文对BERT-CRF基线模型和BERT-semiCRF在FR数据集上的未登录词识别情况进行了详细分析。图2为半监督方法较基线模型新增识别的未登录词的分布情况，有新增识别的未登录词种类数87种(图中第一列)，共计328频次（图中第二列）。其中，新发现未登录词57种，占有新增识别的未登录词总种数的66%，主要为“御剑”、“剑诀”、“五行环”一类的低频未登录词，占总增加频数的22%。基线模型已发现未登录词，但是在某些句子中未能正确识别，而半监督模型正确识别，这类有30种，占种数的34%，但是占总频数的78%，为“玄骨”、“冰焰”、“修士”等高频未登录词。上述现象说明，本文实现的半监督模型不仅具备简单高频未登录词在复杂语境下识别能力，还有更强的未登录词发现能力，能够发现低频未登录词。

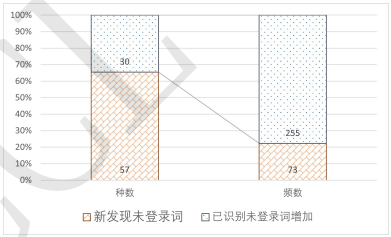


Figure 2: 新增识别的未登录种类和频数分布情况

4.5 总结与展望

词是中文信息处理的基本任务之一，分词结果的准确性将影响基于词的深层次语言信息处理任务的性能。本文提出了一种充分利用相对易得的目标领域无标注文本、实现跨领域迁移的半监督中文分词框架；通过引入词记忆网络和序列条件熵方法，不但提高了跨领域分词任务的准确性，特别是提升了跨领域未登录词识别的召回率。实验结果表明，本文提出的方法增强了分词模型的泛化能力和跨领域迁移能力。

本文还有一些尚未开展的工作和不足之处，包括如何克服不同标注语料在标注准则上的不一致问题，下一步研究可以增加模型多准则学习能力。此外，本文实现的半监督框架是基于神经网络的半监督序列标注框架，在中文分词上较好的提升效果，理论上，该半监督框架可应用于任何序列标注任务，因此，可尝试将该框架应用于命名实体识别、方面级情感分析等序列标注任务。

(c) 2022 中国中文信息学会计算语言学专业委员会

参考文献

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. *arXiv preprint arXiv:1606.04300*.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for chinese. *arXiv preprint arXiv:1704.07047*.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuan-Jing Huang. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1197–1206.

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

N. Ding, D. Long, G. Xu, M. Zhu, and H. T. Zheng. 2020. Coupling distant annotation and adversarial training for cross-domain chinese word segmentation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

H. Feng, C. Kang, X. Deng, and W. Zheng. 2004a. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004b. Accessor variety criteria for chinese word extraction. *Computational linguistics*, 30(1):75–93.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. Rethinkcws: Is chinese word segmentation a solved task? *arXiv preprint arXiv:2011.06858*.

Z. Hao, Z. Yu, Z. Yue, S. Huang, and J. Chen. 2017. Word-context character embeddings for chinese word segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Han He, Lei Wu, Hua Yan, Zhimin Gao, Yi Feng, and George Townsend. 2019. Effective neural solution for multi-criteria word segmentation. In *Smart Intelligent Computing and Applications*, pages 133–142. Springer.

Z. He, Z. Wang, W. Wei, S. Feng, X. Mao, and S. Jiang. 2020. A survey on recent advances in sequence labeling from deep learning models.

Z. Huang, X. Wei, and Y. Kai. 2015. Bidirectional lstm-crf models for sequence tagging. *Computer Science*.

Yang Liu and Yue Zhang. 2012. Unsupervised domain adaptation for joint segmentation and pos-tagging. In *Proceedings of COLING 2012: Posters*, pages 745–754.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for crf-based chinese word segmentation using free annotations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 864–874.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art Chinese word segmentation with Bi-LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4908, Brussels, Belgium, October-November. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Matthew Peters, M. Neumann, M. Iyyer, M. Gardner, and L. Zettlemoyer. 2018. Deep contextualized word representations.

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.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for Chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 163–172, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

(c) 2022 中国中文信息学会计算语言学专业委员会

Nianwen Xue. 2003. Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

Yuxiao Ye, Yue Zhang, Weikang Li, Likun Qiu, and Jian Sun. 2019. Improving cross-domain chinese word segmentation with word embeddings. *CoRR*, abs/1903.01698.

Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018. Neural networks incorporating unlabeled and partially-labeled data for cross-domain chinese word segmentation. In *IJCAI*, pages 4602–4608.

李航et al. 2012. 统计学习方法. Qing hua da xue chu ban she.

邓丽萍and 罗智勇. 2017. 基于半监督crf 的跨领域中文分词. 中文信息学报, 31(4):9–19.

CCL 2022

数字人文视角下的《史记》《汉书》比较研究

邓泽琨[1,2], 杨浩[2], 王军(✉)[1,2]
[1] 北京大学信息管理系/ 北京市
[2] 北京大学数字人文研究中心/ 北京市
{dzk,yanghao2008,junwang}@pku.edu.cn

摘要

《史记》和《汉书》具有经久不衰的研究价值。尽管两书异同的研究已经较为丰富，但研究的全面性、完备性、科学性、客观性均仍显不足。在数字人文的视角下，本文利用计算语言学方法，通过对字、词、命名实体、段落等的多粒度、多角度分析，开展对于《史》《汉》的比较研究。首先，本文对于《史》《汉》中的字、词、命名实体的分布和特点进行对比，以遍历穷举的考察方式提炼出两书在主要内容上的相同点与不同点，揭示了汉武帝之前和汉武帝到西汉灭亡两段历史时期在政治、文化、思想上的重要变革与承袭。其次，本文使用一种融入命名实体作为外部特征的文本相似度算法对于《史记》《汉书》的异文进行自动发现，成功识别出过去研究者通过人工手段没有发现的袭用段落，使得我们对于《史》《汉》的承袭关系形成更加完整和立体的认识。再次，本文通过计算异文段落之间的最长公共子序列来自动得出两段异文之间存在的差异，从宏观统计上证明了《汉书》文字风格《史记》的差别，并从微观上进一步对二者语言特点进行了阐释，为理解《史》《汉》异文特点提供了新的角度和启发。本研究站在数字人文的视域下，利用先进的计算方法对于传世千年的中国古代经典进行了再审视、再发现，其方法对于今人研究古籍有一定的借鉴价值。

关键词： 数字人文；《史记》；《汉书》；命名实体；异文；文本相似度

A Comparative Study of *Shiji* and *Hanshu* from the Perspective of Digital Humanities

Zekun Deng[1,2], Hao Yang[2], Jun Wang(✉)[1,2]
[1] Department of Information Management, Peking University / Beijing
[2] Research Center for Digital Humanities of PKU / Beijing
{dzk,yanghao2008,junwang}@pku.edu.cn

Abstract

Shiji and *Hanshu* have been studied extensively throughout the past centuries. Although the similarities and differences of the two works have been researched in numerous literatures, these studies are limited in their collectiveness, comprehensiveness, objectiveness and rigor. Under the sight of digital humanities, this paper attempts to adopt computational linguistic methods to compare *Shiji* and *Hanshu* in a multi-scale and multi-perspective way by analyzing the characters, words, named entities and paragraphs in the books. Firstly, this paper compares the distribution and characteristics of the characters, words and named entities in *Shiji* and *Hanshu*, finding out the major similarities and differences of their contents by an exhaustive enumeration, revealing the significant political, cultural and ideological transformations from pre-2th

第二十一届中国计算语言学大会论文集，第656页-第670页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

century B.C. to the rest of Western Han dynasty. Secondly, this paper adopts a text similarity metric incorporating named entity as an external feature to automatically probe variant readings in *Shiji* and *Hanshu*. We manage to discover variant readings that have not been found by past researchers who rely solely on manual approaches, obtaining much richer knowledge of *Hanshu*'s inheritance of *Shiji*. Thirdly, this paper derives the differences between the variant readings of *Shiji* and *Hanshu* automatically by computing their longest common subsequences. Based on the results, this paper rigorously proves the writing style discrepancies of the two books through a macroscopic statistical analysis and interprets their linguistic features respectively with microscopic example texts, providing new perspective and insights on the variant readings of the two books. In conclusion, under the sight of digital humanities, this paper employs advanced computational methods to reexamine and reexplore centuries-old ancient Chinese classics, bringing enlightenment to moderners about new approaches for studying ancient literatures.

Keywords: digital humanities , *Shiji* , *Hanshu* , named entity , variant reading , text similarity

1 引言

《史记》和《汉书》是中国古代史籍的经典之作，在文学、历史学、语言学等领域具有宝贵且不可替代的研究价值。《史记》和《汉书》有诸多相似之处。从二者记载的历史时段上看，《史记》记载的历史横跨三皇五帝到汉武帝时期，而《汉书》则写的是整个西汉的历史，同时也包括秦朝末年至西汉建立之前的部分事件，因此两书记录的历史在时间上存在大幅重合。从二者的体例上看，二者都是以人物传记为主的纪传体史书。当然，《史记》和《汉书》也有很多方面的差异。两书作者不同的写作动机、行文风格，两书所载历史时期的差异，后世文学家和历史学家对两书的不同看法等等，也都是有价值的研究主题，使得两书的相同和相异之处交织，让《史记》和《汉书》的比较研究成为了有意义的研究问题。

尽管前人对这两部经典的研究已经相当丰富，但是，这些研究仍然存在一些不足。一方面，过往的研究在对两书文本进行分析时，往往通过举例的方法进行论证，采用"以点带面"的模式，用个别的例子来分析得出结论。这种研究方法虽然能够得到可信的结论，但是在全面性和完备性上有所欠缺，有可能忽略某些重要的文本细节。另一方面，人文学者在研究《史记》和《汉书》时采用的定性方法往往具有较强的主观性，其结论往往受制于学者自身的知识储备，并且其分析过程时常具有随意性和偶然性，研究的客观性和科学性有所不足。

近年来，数字人文对于人们阅读古代经典文献的方式产生了巨大的改变。一方面，数字人文使得我们可以采用量化计算手段对于大量文本进行提炼、抽象和概括，从而使得人们可以在不完整阅读全文的情况下把握一本书的主线和要旨。另一方面，数字人文将信息抽取、信息检索等现代计算机科学技术引入人文研究，使得机器可以自动发现卷帙浩繁的古代文献之间潜藏的关系和知识，可以帮助我们做出以前人的能力无法达成的新发现、新洞察。

在数字人文的视域下，本文试图利用计算语言学方法，通过对字、词、命名实体、段落等的多粒度、多角度分析，开展对于《史记》和《汉书》的比较研究。本研究的整体流程如Figure 1所示。

本文的主要贡献如下：

1. 本文利用基于深度学习的古汉语分词和命名实体识别模型对《史记》《汉书》进行处理，对于《史记》《汉书》中的字、词、命名实体的分布和特点进行对比，以遍历穷举的考察方式提炼出两书在主要内容上的相同点与不同点，通过对典型实例的深入分析，揭示了汉武帝之前和汉武帝到西汉灭亡两段历史时期在政治、文化、思想上的重要变革与承袭。

2. 本文利用以命名实体作为外部特征的文本相似度计算方法对于《史记》《汉书》的异文进行自动发现，其结果表明本算法的发掘结果不但与过往学者人工发现的结果相吻合，并且能

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
本研究得到国家自然科学基金国际重点合作项目" 中国儒家学术史知识图谱构建研究"(项目号:72010107003)的支持

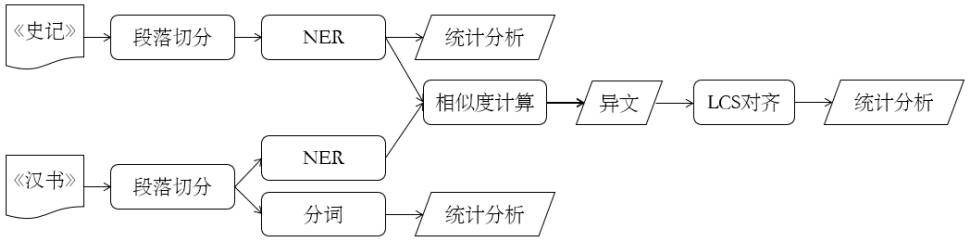


Figure 1: 研究整体流程

够发现过去研究者通过人工手段没有发现的袭用段落。这使得我们对于《史记》《汉书》的承袭关系形成更加完整和立体的认识。

3. 本文通过计算异文段落之间的最长公共子序列来自动得出两段异文之间存在的差异，首先从统计上证明了《汉书》文字风格相比于《史记》的差别，随后选取了《史记》《汉书》的若干典型异文段落进一步对二者的语言特点进行了阐释，为理解《史记》《汉书》的异文特点提供了新的角度和启发。

2 文献综述

2.1 《史记》《汉书》比较研究

《史记》与《汉书》的比较研究具有悠久的历史。关于《史》《汉》二书的优劣问题，早在汉朝就有文人论述。东汉的王充[0]、晋人张辅(朱一玄, 刘毓忱, 2012)、南宋郑樵[1]、邵博[2]等皆从不同角度对于《史》《汉》做了比较。在清朝以前，《史》《汉》研究多停留在争论二者"孰优孰劣"的程度，而到了清及以后，《史》《汉》研究逐渐走向更加客观和科学的视角，"互有得失论"占据上风，且大量学者开始进行仔细的考据，着重于对文本进行更细致的分析。清代涌现出大量专门论述马、班异同的著作或著作章节，包括周中孚《补班马异同》、王筠《史记校》、蒋中和《马班异同议》等(曾小霞, 2009)。到了20世纪，吴福助(1975)著《史汉关系》一书对二者进行全面探讨。朴宰雨(1994)著《〈史记〉〈汉书〉比较研究》一书，详尽系统地对《汉书》袭用《史记》各卷的情况用表格进行了一一列举说明，具有很高的参考价值。

21世纪以来，《史》《汉》对比研究迈上了新的台阶。在文献学、语言学领域，沙志利(2005)遵循文献学方法，以《史》《汉》的详细比勘为依据，从史料、文字、思想等多个方面对两书进行了分析，从形式和内容两个角度分析了《史》《汉》异同产生的内因和外因。王海平(2003)总结了《史》《汉》异文在字、词、句三个层次上的表现形式，指出了其在文献学和语言学领域的运用。张明月(2021)研究了《史》《汉》重合篇章的文字差异和文学解读。张添雅(2021)梳理了《汉书》八表对于《史记》十表的承袭和创新之处。在文学领域，也有研究者(曾小霞, 2012; 诸雨辰, 2016; 夏德靠, 2019)从叙事、人物塑造、思想等角度对《史》《汉》进行了比较。

2.2 文本相似度算法和异文发现

在古典文献学中，"异文"有多种含义，一种是指一本书中的某一段文字因为传抄而产生的不同版本，另一种则是指记载同一件事但措辞有差异的字句。由于异文存在字词和语义上的相似性，因此我们可以利用文本相似度算法对古书中的异文进行自动发现。传统上的语义文本相似（Semantic Textual Similarity，STS）(Agirre et al., 2013)任务采用词袋模型（Bag of words）或者TF-IDF(Ramos, 2003)方法将文本转化成实值向量，通过计算向量之间的接近程度来判断文本语义的相似性。例如，肖磊和陈小荷(2010)用bigram计算句子相似度在春秋三传中进行异文寻找。李越(2014)利用改进的编辑距离算法与事件信息标注相结合，利用事件数据库对语料进行人物、地点、时间标注，加权计算文本相似度，实现《左传》《史记》异文发现。然而，该论文并未明确给出事件相似度的计算公式，且其方法对于数据库的依赖过强。近年来，神经网络开始成为STS的主流方法。梁媛等(2021)用《春秋》和春秋三传建立了异文平

[0] 《论衡》
[1] 《通志》
[2] 《邵氏闻见后录》

(c) 2022 中国中文信息学会计算语言学专业委员会

行语料，训练了BERT模型判断两句话是否为异文，但该方法的语料切割粒度过小，准确率不高，且需要大量人工标注数据。

本文提出以命名实体为外部特征的文本相似度计算方法，该方法不需要人工监督数据，且不存在输入长度限制，能够适应长文本比较的需要。

3 《史》《汉》字、词和命名实体的统计对比

3.1 《史》《汉》字频分布对比

欲全面了解《史记》和《汉书》的异同，两本书在用字和用词上的差异是必须注意的。本文首先对两书用字总数进行了统计。在将异体字合并后，本文统计得到《史记》全书共使用了4619个不同的字，而《汉书》则用了5343个，与《史记》相比增加了15.7%。

本文对两书中频率最高的15个字进行了统计和对比，其结果见附录A Table 7。本文发现，在这些字中，有3个字在两书中的频率差异较大，它们分别是“王”、“子”和“公”。具体而言，三个字在《汉书》中的频率与《史记》中相比分别低了6.28‰、5.13‰和5.50‰。Table 1展示了两书中包含这三个字的最常见词语（分词的方法见3.2节），按照出现频率由高到低排序。综合这些数据可以看出，以上三个字在两书中的频率的差异显然不只是偶然造成的。为了更充分地理解这种差异出现的原因，本文进一步统计了这三个字在命名实体内出现的频率（即在命名实体内出现的次数除以全书总字数），如Table 2所示。可以看出，这三个字的频率差异很大程度上可以由包含这三个字的命名实体占全文比例的降低来解释。结合以上信息，本文猜测，这一变化可能是来源于两书所叙历史时期的政治形势、政治制度和所涉人物的差异。

字	包含该字的最常见的词	
	《史记》	《汉书》
王	王 汉王 大王 秦王 赵王 齐王 项王 楚王	王 汉王 王莽 大王 王者 齐王 赵王 淮南王
子	子 太子 天子 孔子 公子 君子 子孙 弟子	子 天子 太子 孔子 子孙 父子 君子 弟子
公	公 公子 沛公 太史公 桓公 文公 周公 景公	公 公卿 沛公 周公 公主 三公 安汉公

Table 1: 《史记》《汉书》中包含“王”、“子”、“公”三个字的最常见词

字	在全文中			在命名实体中		
	《史》	《汉》	Δ	《史》	《汉》	Δ
王	16.36	10.08	-6.28	11.09	6.65	-4.43
子	12.96	7.83	-5.13	6.22	3.00	-3.22
公	9.62	4.12	-5.50	7.85	2.91	-4.94

Table 2: 《史记》《汉书》中“王”、“子”、“公”在全文和在命名实体中的词频(‰)及差值(Δ)

3.2 《史》《汉》词频分布对比

本研究对《史记》和《汉书》的词频分布进行了对比。本文中《史记》的词频统计利用的是台湾“中央研究院”的《史记》人工标注语料，该语料已经由专家进行分词，可以直接进行统计。《汉书》的词频统计则利用了Tang and Su(2022)开发的古汉语分词模型，人工检验表明其准确率较高，可以用于统计分析。

本文首先统计了两书词语的平均词长和不同长度词语的比例，如附录A Table 8所示。可以发现，《汉书》相比于《史记》平均词长更大，单音节词的占比更低，而2字、3字、4字或以上的词的占比均更高。之后，本文分别统计了两书中频率最高的单音节词、双音节词、三音节词和四音节词（详细结果见附录A Table 9）。由于时间词不具有实际意义（例如“二年”、“六月”等），因此没有列入此表。

在高频词方面，《史》《汉》二书的高频词总体较为接近，但是又呈现出几处明显的不同。单音节词频率的分布和字频的分布比较相似，《史》《汉》的高频单音节词基本一致，说明《史》《汉》从语言学角度看大致处在同一时代，用字上的总体差异不大。双音节词中，《汉书》与《史记》相比，“诸侯”、“孔子”两个词的排名大幅下降，而“匈奴”、“丞相”、“单

(c) 2022 中国中文信息学会计算语言学专业委员会

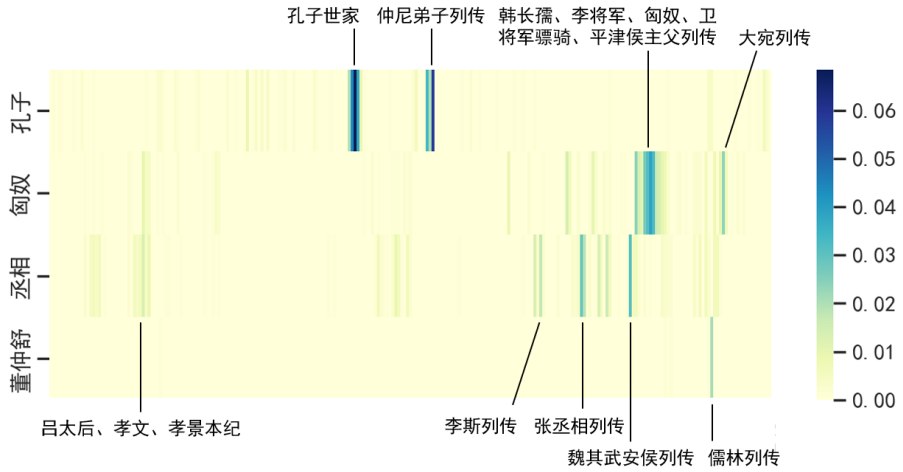


Figure 2: 四个典型词在《史记》全文的分布密度图

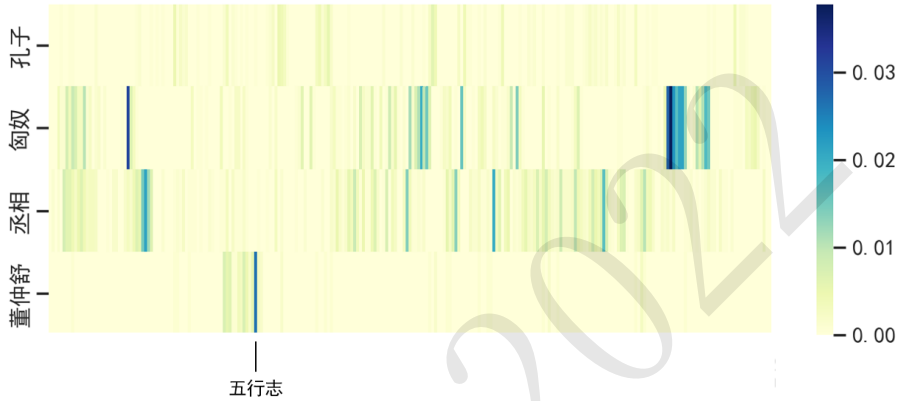


Figure 3: 四个典型词在《汉书》全文的分布密度图

于”三个词的排名大幅上升。结合历史来看，“丞相”和“诸侯”词频的一升一降很大程度上可以归因于汉朝的大一统政治体制和地方分封势力的大幅削弱。长度为4的词中，仅有5个在《史记》和《汉书》中词频均排前十，其余仅在《史记》或《汉书》中进入前十的词均体现出明显的时代特征，反映了先秦和西汉两个时期主要历史人物和官职设置的差异。

以上结果中有几个词比较值得注意，即“孔子”、“匈奴”、“丞相”和“董仲舒”。为了更深入地理解这几个词的词频所体现的文本含义，本文绘制了它们在《史》《汉》两书全文中出现的密度图，如Figure 2和3所示。Figure 2和3将每本书分别按照字数从前到后均分成250份，然后将每个词在每一份中出现的密度按照高低不同赋以不同颜色。此处，记一段文字$\boldsymbol{t} = [a_1, ..., a_N]$，表示$\boldsymbol{t}$由$a_1, \ldots, a_N$这$N$个词组成，则一个词$w$在文字$\boldsymbol{t}$中的密度由以下公式定义：

$$\rho(w, \boldsymbol{t}) = \frac{\boldsymbol{t}.\text{count}(w) \cdot w.\text{length}}{\sum_{i=1}^{N} a_i.\text{length}},$$

其中$\boldsymbol{t}.\text{count}(w)$是$w$在$\boldsymbol{t}$中出现的次数，$w.\text{length}, a_i.\text{length}$ 分别是w和a_i的长度。不同颜色所对应的密度值可见图右侧的图例。

从Figure 2和3可以看出，所关注的四个词在两书中的分布各有特点。具体而言，“孔子”在《史记》中集中分布在《孔子世家》和《仲尼弟子列传》中，在该书的其他篇章中频率极低，而在《汉书》中出现次数寥寥。这表明尽管孔子是中国文化史上最重要的人物之一，但其在《史》《汉》两书中的出现仍然比较局限。在《史记》中，“匈奴”一词主要集中在《韩长孺列传》、《李将军列传》等6篇列传中，而在《汉书》中“匈奴”一词的分布较为分散，在纪、传中均大量出现，存在多个分布高峰。“丞相”一词在《史记》中集中在《吕太后本纪》、《孝文本纪》等6篇中，而在《汉书》中的分布状况与“匈奴”类似，在全书各个部分都有高频出现。在

(c) 2022 中国中文信息学会计算语言学专业委员会

《史记》中，“董仲舒”绝大多数出现在《儒林列传》中且总的出现次数不多，而在《汉书》中则大量集中于《五行志》。这表明董仲舒的“五行”思想在当时已经有了很大的影响。

以上四个词在《史》《汉》二书中的分布具有很大的启示性，因为它们很好地勾勒了两个不同历史时期之间政治、文化、思想领域的重要变化。从政治的角度看，“匈奴”分布的变化很大程度上呼应了匈奴在战国兴起、在西汉强盛并与中原政权展开持久拉锯的事实，其在《汉书》中频繁而分散的出现忠实地体现出匈奴势力在西汉政治舞台上所扮演的重要角色。“丞相”一词在《史记》中的集中出现几乎仅限于记述秦汉两朝历史的三篇本纪和三篇列传中，与丞相这一官职的历史沿革高度吻合：丞相一职初创于战国时期，当时各国虽有与丞相地位等同的“相国”一职，但是名称有所不同，而到了战国中后期，秦国才率先设立“丞相”一职，这一官职也在秦汉两朝得以沿袭(袁祖亮, 1988)。这也就很好地解释了为何“丞相”在《史记》中的出现非常局限，而在《汉书》中多有分布。从文化的角度看，“孔子”和“董仲舒”均是历史上重要的儒学家，而二者在《史》《汉》二书中分布频次的巨大差异，无疑暗示了二者在不同的时代所具备的影响力大不相同。

3.3 《史》《汉》命名实体频率分布对比

命名实体识别（Named Entity Recognition, NER）是一项重要的NLP任务。本文使用基于四库BERT的NER模型来对《史记》《汉书》中的命名实体进行识别。该模型采用BERT+BiLSTM+CRF（Conditional Random Field，条件随机场）架构，其中BERT参数通过《四库全书》语料预训练初始化，之后利用《资治通鉴》NER数据集对模型进行微调。该模型在《资治通鉴》测试集上的F1约98%，在《史记》上测试的F1值约为90%，表现出较高的精度和可用性。该模型的超参数和《四库全书》语料的详细信息见附录C。

在数字人文研究中，针对中国古代典籍的特点，我们一般将古汉语文本中的命名实体分为人物、地点、时间、书籍、官职等类别。命名实体区别于文本中其他字词和短语的最重要特点之一是它们都对应于现实世界中的一个真实存在的实体。因而，对语料中命名实体出现的频率和分布特点进行分析，能够在很大程度上反映出语料所述历史时期中的人、地、时等实体的特征，有助于我们透过表面洞察文字背后的历史脉络。因此，对于史料中的命名实体分布进行分析，对于我们更深入地发掘和理解史料具有重要的意义。

利用前文所述的命名实体识别模型，本文对《史》《汉》二书中每个实体类型频次最高的10个实体进行了统计，结果列在附录A Table 10中。该结果有丰富的解读空间。从人物来看，《史记》和《汉书》的高频人物实体存在一些差异，例如前者包含有“项羽（项王）”、“赵王”等仅在汉朝建立以前存在的人物，而后者则包含有“莽（王莽）”、“光”等主要活动于汉武帝以后的人物。以上差异并不令人惊奇，然而《汉书》排名前十的高频人物中还包含了“禹”、“汤”、“武”等先秦时期的人物，其生活时间与《汉书》所叙时间毫无重合。经过查找原始语料，本文发现，《汉书》中的“禹”、“汤”、“武”指的不全是三代的三位君王，实际上有时指的是汉朝的同名人物。客观上说，《汉书》中“禹”、“汤”、“武”三个人物实体的频繁出现，是《汉书》反复引用夏禹、商汤、周武三位君王的事迹和其他重名人物的存在这两个因素共同作用的结果。

从地点来看，《史记》中出现频次最高的十个地点实体中，除了“汉”以外，其余均为春秋和战国时期力量最强大的诸侯国的国名，而在《汉书》中则出现了“匈奴”、“长安”、“河”等与西汉政治关联更紧密的地点实体。

两部书频次最高的十个书籍实体差异总体不明显，都包含了《春秋》、《诗》、《书》、《易》等先秦时期的经典著作，且在两部书中都排名前四，表明了这些典籍在西汉建立前后均保持着重要的地位。但是，在《史记》中排名靠前的《老子》在《汉书》中并未进入前十，而《汉书》中《五经》和《左氏传》均出现在前十，这一结果与司马迁和班固二位作者本身的思想倾向不无关系，同时也某种程度上说明了西汉建立前后文人思想观念的微妙变化。

排名靠前的时间和官职实体均未显示太多区别。《史记》和《汉书》频次前十的官职实体有9个是相同的，除了这9个外，《史记》有“夫人”，而《汉书》有“太守”。这种现象的原因也容易解释：根据唐朝杜佑《通典》记载，“太守”原名“郡守”，秦朝行郡县制，在郡一级设立郡守一职，汉景帝年间更名为“太守”。因而，“太守”这一官职名实际上在西汉时期才广泛被使用，从而在《汉书》中出现的频次相比《史记》大幅上升。

总的来说，《史记》和《汉书》中出现的命名实体有同有异，呈现出了各自时代的独有特点，并在某种程度上反映了从秦之前到西汉时期历史发展的轨迹。我们使用模型以可观的准确

(c) 2022 中国中文信息学会计算语言学专业委员会

率对古代史籍中的绝大多数命名实体进行识别，用一种精确和客观的方式来描述《史》《汉》中的最主要人物、地点、官职等等历史主体，为探究人文问题提供了一种新的手段。这些结果有助于我们从一个新的角度来认识和理解这段历史。

4 计算语言学视角的《史》《汉》异文研究

4.1 《史》《汉》文本袭用的自动发现

4.1.1 方法

本文使用一种基于TF-IDF和命名实体的文本相似度算法来实现《史》《汉》二书异文的自动发现。对于两个字符串$\boldsymbol{p}_i = a_{i,1}...a_{i,N_i}, \boldsymbol{p}_j = a_{j,1}...a_{j,N_j}$，该方法可以输出二者的相似度$u(\boldsymbol{p}_i, \boldsymbol{p}_j) \in \mathbb{R}$。本方法描述如下：首先，利用3.3节介绍的NER模型得到每个字符串中包含的所有实体提及$\boldsymbol{e}_i$和$\boldsymbol{e}_j$，其中

$$\boldsymbol{e}_i = \{(a_{i,b}...a_{i,c}, h_{i,b,c}) \mid a_{i,b}...a_{i,c}\text{是一个类型为}h_{i,b,c}\text{的命名实体}\},$$

$h_{i,b,c} \in \boldsymbol{H}$，$\boldsymbol{H}$是所有实体类型标签的集合。然后，选取一个字符串集合$\mathcal{D}$，记$\mathcal{D}$中出现的所有n-gram为$\boldsymbol{g} = [g_1, ..., g_M]$，定义$g_i$的逆文档频率（Inverse Document Frequency, IDF）为$\text{idf}(g_i) = \log \frac{|\mathcal{D}|}{|\{\mathbf{d}|g_i\text{在}\mathbf{d}\text{中出现},\ \mathbf{d}\in\mathcal{D}\}|}$，其中$|\mathcal{D}|$是集合$\mathcal{D}$中元素的数量。记$n$-gram g_j在字符串$\boldsymbol{p}_i$中出现的频次为$\text{tf}(\boldsymbol{p}_i, g_j)$，则字符串$\boldsymbol{p}_i$的TF-IDF向量为

$$\mathbf{v}_i = [\text{tf}(\boldsymbol{p}_i, g_1) \cdot \text{idf}(g_1), ..., \text{tf}(\boldsymbol{p}_i, g_M) \cdot \text{idf}(g_M)]^{\mathrm{T}},$$

因此，字符串$\boldsymbol{p}_i, \boldsymbol{p}_j$的相似度可定义为

$$u(\boldsymbol{p}_i, \boldsymbol{p}_j) = \lambda \frac{\mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} + (1-\lambda)\frac{|\boldsymbol{e}_i \cap \boldsymbol{e}_j|}{|\boldsymbol{e}_i \cup \boldsymbol{e}_j|}, \tag{1}$$

其中λ是人为设定的常数且$\lambda \in [0, 1]$。

从公式(1)容易看出，这一相似度定义着重强调了两段文本中命名实体的重合度。当两段文字中所提到的人物、地点、官职等命名实体高度重合时，两段文字很有可能在讲述同样的历史事件，语义相似的可能性也更高。因而，与不考虑命名实体的方法相比，本方法相比更加适用于人、地、时等对象较为密集的史部文献。

本文将此方法与古籍异文发现的现有最好方法(梁媛等, 2021)的性能进行了对比。该论文提出使用预训练语言模型的有监督方法进行异文识别，具体而言是将每组待检测文本对同时输入BERT编码器，对输出隐向量进行分类以确定该文本对是否为同事异文（下称基线方法）。为了对比基线方法和本文方法的性能，本文遵循梁媛等(2021)的描述在《史记》《汉书》语料上复现了该方法。由于本文方法是无监督的，因此为了进行更好的对比，训练基线模型时正例和负例均只提供了1或3个训练样本（1-shot/3-shot）。模型的其余设定与原论文完全相同。测试使用的是人工标注的《史记》《汉书》异文段落数据集，总数约100条。测试结果如Table 3。可以看到，在少样本条件下，本文方法的P、R、F1明显高于基线方法。

此外，基线方法的时间复杂度也值得注意。异文发现场景要求在大量语料中寻找相似文本。若待发现的文本条数为n，则基线方法需要使用BERT进行$n(n-1)/2$次推理，其时间成本相当高昂。相比之下，本文的方法只需要少量的向量内积操作和n次BERT推理，时间成本低，在文本数量较大时仍能保持较高的可用性。

方法		P	R	F1
BERT	1-shot	57.01(± 18.45)	90.61(± 17.93)	66.29(± 9.03)
	3-shot	83.73(± 15.39)	91.47(± 8.71)	86.51(± 9.04)
本文方法		84.21	96.97	90.14

Table 3: 本文方法与现有方法的性能比较（括号内为10次重复实验的标准差）

(c) 2022 中国中文信息学会计算语言学专业委员会

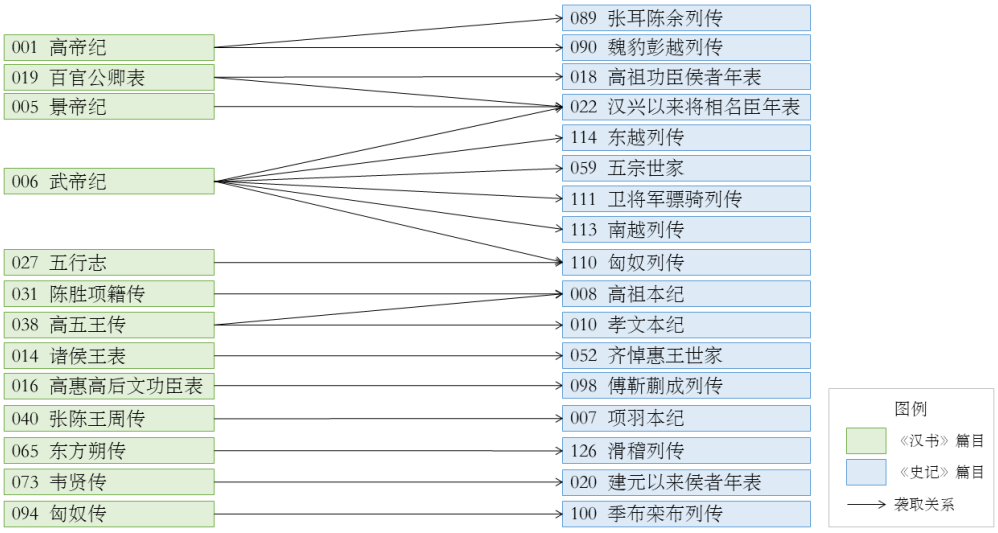


Figure 4: 本文新发现的《汉书》各篇承袭《史记》篇目（仅展示朴未记载的，其余省略）

4.1.2 结果

利用本文方法，将《史记》和《汉书》按照中华书局标点本的分段切割成段落，将得到的所有段落作为字符串集合$\mathcal{D}$，计算其中所有n-gram的IDF。对于《汉书》中的每一段文本$\boldsymbol{p}$，利用公式(1)计算其与《史记》中所有段落的相似度，并选出其中相似度最高的一段（记为$\boldsymbol{q}$），记二者的相似度为$u(\boldsymbol{p},\boldsymbol{q})$。设定阈值$\mu$，当$u(\boldsymbol{p},\boldsymbol{q}) > \mu$时，认为$\boldsymbol{p}$和$\boldsymbol{q}$是相似的。在本研究中，公式(1)中的加权参数取$\lambda = 0.7$，阈值$\mu = 0.25$。通过这一方法，本文发现了《汉书》共2082段文本与《史记》存在高度相似（长度不超过10个字的段落被排除了）。经过逐一人工比对确认后，本文对于《汉书》对《史记》可能的袭用情况进行了全面梳理。为了将结果与之前学者的研究作比较，本文以朴宰雨(1994)在《〈史记〉〈汉书〉比较研究》第三节第三部分“《汉书》各篇承袭《史记》之情况概述”为对照对象。经对照发现，本文的方法不但能够识别出朴宰雨书中所记录的所有袭用篇目，还发现了多达21对本书中没有记载的袭用篇章（见Figure 4）。

可以看到，Figure 4所呈现出的《汉书》袭用《史记》的情况较为复杂，下面本文对其中列出的几组典型的袭用案例进行详细阐释。

案例1：《高帝纪》袭用《张耳陈余列传》、《魏豹彭越列传》

朴宰雨在其书中指出，《高帝纪》“增补与改写之处甚多……为汉书中最大规模者”，并指出班固将《项羽本纪》、《留侯世家》、《韩信卢绾世家》等篇中有关刘邦之事移入了《高帝纪》。但是，朴的列举仍然不够全面，有所遗漏。事实上，《高帝纪》还袭取了《魏豹彭越列传》中五年冬十月张良为汉王献计使韩信、彭越引兵会师之事以及《张耳陈余列传》中高祖逃脱贯高谋杀阴谋之事。Table 4列出了《高帝纪》、《高祖本纪》、《张耳陈余列传》对逃脱谋杀之事的不同记载。虽然《高祖本纪》也记录了同一件事，但是并未记载高祖询问县名的对话。综合三段文本判断，《高帝纪》的这一段文字更可能是从《张耳陈余列传》袭取而来的。

篇名	内容
《汉书·高帝纪》	八年冬，上东击韩信馀寇于东垣。还过赵，赵相贯高等耻上不礼其王，阴谋欲弒上。上欲宿，心动，**问“县名何？”曰：“柏人。”上曰：“柏人者，迫于人也。”**去弗宿。
《史记·高祖本纪》	高祖之东垣，过柏人，赵相贯高等谋弒高祖，高祖心动，因不留。
《史记·张耳陈余列传》	汉八年，上从东垣还，过赵，贯高等乃壁人柏人，要之置厕。上过欲宿，心动，问**曰：“县名为何？”曰：“柏人。”“柏人者，迫于人也！”**不宿而去。

Table 4: 《史记》和《汉书》对于高祖逃脱贯高谋杀阴谋之事的不同说法

(c) 2022 中国中文信息学会计算语言学专业委员会

案例2：《武帝纪》袭用《汉兴以来将相名臣年表》、《五宗世家》、《匈奴列传》、《卫将军骠骑列传》、《南越列传》、《东越列传》

朴宰雨在书中指出，由于《史记》有录无书，且所补者“言辞鄙陋，非迁本意”，因此《武帝纪》并未袭用而是进行了重新创作。本文的结果表明，这一说法不够准确。事实上，本文发现《武帝纪》中有若干段落与《汉兴以来将相名臣年表》、《五宗世家》、《匈奴列传》等共6卷存在高度相似，经分析极有可能为班固袭用。为了更清楚地说明袭用情况，Table 5中列出了《武帝纪》和《匈奴列传》的四个相似段落中的一段。可以看到，《武帝纪》和《匈奴列传》的这些段落高度雷同，基本可以确信二者之间存在袭取关系。

《汉书·武帝纪》内容	《史记·匈奴列传》内容
夏五月，贰师将军三万骑出酒泉，与右贤王战于天山，斩首虏万馀级。又遣因将军出西河，骑都尉李陵将步兵五千人出居延北，与单于战，斩首虏万馀级。陵兵败，降匈奴。	其明年，汉使贰师将军广利以三万骑出酒泉，击右贤王于天山，得胡首虏万馀级而还。匈奴大围贰师将军，几不脱。汉兵物故什六七。汉复使因将军敖出西河，与强弩都尉会涿涂山，毋所得。又使骑都尉李陵将步骑五千人，出居延北千馀里，与单于会，合战，陵所杀伤万馀人，兵及食尽，欲解归，匈奴围陵，陵降匈奴，其兵遂没，得还者四百人。

Table 5: 《汉书·武帝纪》和《史记·匈奴列传》部分雷同段落

案例3：《东方朔传》袭用《滑稽列传》

朴宰雨在其书中认为，《东方朔传》为班固“根据另外的资料有意新创”，因而未将该传列入其袭用情况概述之中。本文认为，《东方朔传》并非由班固完全新创，而是有所承袭《滑稽列传》。现学界普遍认为，今本《史记》的《滑稽列传》只有一部分是司马迁原作，另一部分是由褚少孙补写的，而司马迁原稿中并未写东方朔，东方朔的事迹是褚少孙增补的(李林晓, 2020)。然而值得注意的是，褚少孙是西汉人，而班固是东汉人，因此在班固写作《汉书》时，他显然能够接触到褚少孙增补后的《史记》版本，从而在写作时加以袭用在逻辑上是完全可能的。事实上，本文发现《东方朔传》和《滑稽列传》存在如Table 6所示的雷同段落。综合以上事实，我们有理由认为《汉书·东方朔传》袭用了《史记·滑稽列传》。

《汉书·东方朔传》内容	《史记·滑稽列传》内容
客难东方朔曰：“苏秦、张仪一当万乘之主，而都卿相之位，……同胞之徒无所容居，其故何也？”	时会聚宫下博士诸先生与论议，共难之曰：“苏秦、张仪一当万乘之主，而都卿相之位，……官不过侍郎，位不过执戟，意者尚有遗行邪？其故何也？”
东方先生喟然长息，仰而应之曰：“是固非子之所能备也。彼一时也，此一时也，岂可同哉？……使苏秦、张仪与仆并生于今之世，曾不得掌故，安敢望常侍郎乎！故曰时异事异。	东方生曰：“是固非子所能备也。彼一时也，此一时也，岂可同哉！……使张仪、苏秦与仆并生于今之世，曾不能得掌故，安敢望常侍侍郎乎！传曰：‘天下无害灾，虽有圣人，无所施其才；上下和同，虽有贤者，无所立功。’故曰时异则事异。

Table 6: 《汉书·东方朔传》和《史记·滑稽列传》雷同段落

总的来说，过去的学者在研究《汉书》对《史记》的承袭情况时常常只关注明显的大段袭用而忽略较为零散琐碎的小段袭用，并且限于各种主、客观原因，遗漏了部分袭用关系。而本文利用算法自动地发现了若干过往学者未发现的袭用篇章，有效深化了我们对史汉关系的认知，使我们对于班固创作过程中对《史记》材料的剪裁、重排、整理工作有了更全面的认识。

4.2 基于最长公共子序列的《史》《汉》异文对比分析

最长公共子序列（Longest common subsequence, LCS）算法被广泛地用于解决文本比对问题。本文通过计算异文的LCS来分析《史记》《汉书》异文段落的差异，从宏观统计和微观案例两个角度进行分析阐释。

(c) 2022 中国中文信息学会计算语言学专业委员会

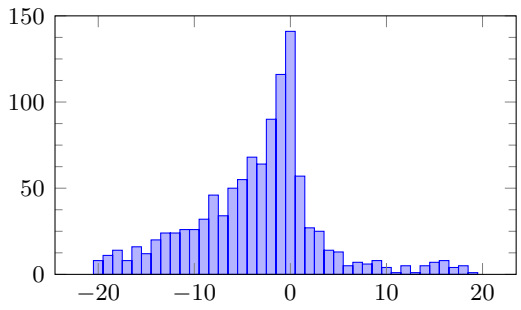


Figure 5: 《汉书》《史记》异文净变化字数直方图（绝对值大于20的部分省略；横轴为净变化字数，纵轴为频数）

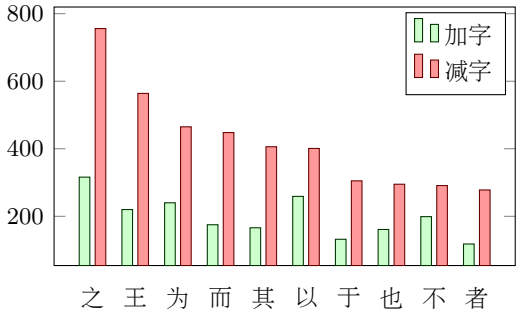


Figure 6: 《汉书》相比《史记》异文删除次数最高的十个字的加字与减字数量（纵轴为加字或减字的频数）

4.2.1 宏观统计分析

此处本文继续利用4.1节提出的文本相似度算法来筛选可以进行比对的文本。在4.1节中，算法的阈值 $\mu = 0.25$，这导致了一些实际并不相关的文本被算法筛选出来。因此，本节将算法的阈值 μ 提高到0.5，从而得到了1364对异文。本文分别计算出这些异文的LCS，将两段文本分别与LCS对齐之后，统计了《汉书》文段在《史记》的基础上增加和删除了哪些字。本文将增加的字数与删除的字数之差称为这对异文的“净变化字数”。Figure 5是算法得出的所有异文的净变化字数的直方图，其中均值为−8.37，中位值为−3。为了从统计上证明《汉书》相比《史记》存在明显的“删字”现象，本文对于异文的净变化字数进行了单尾单样本T检验。检验的原假设为“异文净变化字数的均值大于等于0”。经计算，T统计量的值为−8.950，$p < 0.0001$，故原假设不成立，表明在统计意义上，《汉书》对于《史记》有显著的“减字”现象。

Figure 6中画出了《汉书》异文相比于《史记》删除的字中频率最高的十个字的加字和减字的数量（异体字已去除）。虽然这些字多为无意义的虚词，但是总体上仍能看出删除字数比增加字数更多的趋势，这也与前人(朴宰雨, 1994; 沙志利, 2005)所认为《汉》比《史》异文的文字更加简短凝练的研究结论高度相符。

4.2.2 微观典型例子分析

本文从《史》《汉》异文中选取了两个典型例子，对LCS对齐结果进行对比分析。限于篇幅限制，具体的例子及分析列入附录B。从例子可以看出，LCS可以有效帮助我们在微观层面上对两书异文进行剖析，与宏观分析相配合，对于文本进行更立体的解读。

总的来说，由以上分析可见，本节所使用的LCS算法能够有效地对《史记》《汉书》中的异文进行比对。通过利用LCS对于《史》《汉》异文进行宏观和微观对比，我们可以更好地把握异文在字词增删、用字用词、历史事实、叙事顺序、人物塑造上存在的差异。

5 结论

本文借助基于BERT的古汉语分词模型和命名实体识别模型，对于《史记》《汉书》的字、词、命名实体进行了全量统计对比，对其基本统计数据进行了比较，分析了高频字、高频词、高频命名实体的异同，对于典型词语在全书的分布密度进行了对比，从中挖掘了有关西汉及西汉之前我国历史、政治、文化、思想等方面的沿革，不但为我们理解这段历史提供了新的手段和视角，也为我们用大数据方法处理其他历史文献提供了借鉴。

本文利用一种以命名实体作为外部特征的基于TF-IDF的文本相似度算法对于《史记》《汉书》中的异文进行了自动发现，成功发现了过往学者所未能发现的异文片段，大大拓宽了我们对于《汉书》承袭《史记》规模与程度的认识。进一步，基于这些结果，本文利用最长公共子序列算法对特定异文对进行了对齐，从宏观统计和微观案例两个视角分析了异文加、减、改字的特点，利用更加科学严谨的手段对《史》《汉》写作风格和语言特点的异同作了新的阐释。

本研究站在数字人文的视域下，利用先进的计算方法对于传世千年的中国古代经典进行了再审视、再发现，其方法对于今人研究古籍有一定的借鉴价值，对于数字人文学科的发展做出了独特的贡献。

参考文献

Agirre E, Cer D, Diab M, et al. 2013. *SEM 2013 shared task: Semantic Textual Similarity. *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, 32–43.

李林晓. 2020. 《史记·滑稽列传》选录标准及司马迁不载东方朔之原因. 太原学院学报*(*社会科学版*)*, 21(04):51–59.

李越. 2014. 《左传》《史记》同事异文自动发现及分析. 南京师范大学, 硕士学位论文.

梁媛, 王东波, 黄水清. 2021. 古籍同事异文的自动发掘研究. 图书情报工作, 65(09): 97-104.

朴宰雨. 1994. 《史记》《汉书》比较研究. 中国文学出版社, 北京.

Ramos J. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 29-48.

沙志利. 2005. 《史》《汉》比较研究. 北京大学, 博士学位论文.

Tang X and Su Q. 2022. That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7830–7840, Dublin, Ireland. Association for Computational Linguistics.

王海平. 2003. 《史记》《汉书》异文研究. 暨南大学, 硕士学位论文.

吴福助. 1975. 史汉关系. 文史哲出版社, 台北.

夏德靠. 2019. 从“帝王传记”到“帝王大事记”——《史记》《汉书》“本纪”叙事异同简论. 四川师范大学学报*(*社会科学版*)*, 46(06):115-126.

肖磊, 陈小荷. 2010. 古籍版本异文的自动发现. 中文信息学报, 24(05):50-55.

袁祖亮. 1988. 战国秦汉魏晋南北朝时期的相国与丞相. 郑州大学学报*(*哲学社会科学版*)*, 1988(06):58-65.

曾小霞. 2009. 明清《史记》《汉书》比较研究综述. 苏州大学学报*(*哲学社会科学版*)*, 30(02):71-73.

曾小霞. 2012. 从《史记》和《汉书》看汉代文学之演变. 山西师大学报*(*社会科学版*)*, 39(03):58-61.

张明月. 2021. 《史记》《汉书》重合篇章比较研究. 辽宁师范大学, 硕士学位论文.

张添雅. 2021. 《汉书》八表研究. 哈尔滨师范大学, 硕士学位论文.

朱一玄, 刘毓忱. 2012. 《三国演义》资料汇编. 南开大学出版社, 天津.

诸雨辰. 2016. 历史文本的独断读法——章学诚的《史记》《汉书》解读. 求索, 2016(10):142-148.

CCL 2022

附录A. 统计结果

排序	《史记》		《汉书》	
	字	频率（‰）	字	频率（‰）
1	之	24.71	之	20.93
2	王	16.36	为	14.95
3	为	14.80	以	14.83
4	以	14.65	不	13.23
5	不	14.53	王	10.08
6	子	12.96	年	8.94
7	而	11.79	其	8.54
8	年	11.39	而	7.90
9	曰	10.99	十	7.87
10	其	10.12	子	7.83
11	公	9.62	人	7.66
12	于	9.50	大	7.62
13	人	8.90	侯	7.61
14	侯	8.61	于	7.48
15	也	8.26	曰	7.30

Table 7: 《史记》《汉书》高频字

统计量		《史记》	《汉书》
平均词长		1.247	1.304
不同长度词语占比(%)	1	78.56	74.14
	2	18.80	22.32
	3	2.10	2.53
	4及以上	0.54	1.00

Table 8: 《史记》《汉书》词长分布对比

附录B. 异文LCS微观分析举例

本附录列出两个使用LCS对《史》《汉》异文进行对比的案例，以帮助我们理解LCS方法对于异文研究的辅助作用。引文中，《汉书》相比于《史记》删除的字标红并缩小，增加的字标绿加下划线，替换视作先删除后增加。

例1：《西域传》与《大宛列传》（节选）

大月氏在大宛西可二三千里，居妫水北。其南则大夏，西则安息，北则康居。本行国也，随畜移徙，与匈奴同俗。控弦者可一二十余万。故时，轻匈奴。本居敦煌、祁连间，及至冒顿立，单于攻破月氏，至匈奴而老上单于，杀月氏王，以其头为饮器。始月氏居敦煌、祁连间，及为匈奴所败，乃远去，过大宛，西击大夏而臣之，遂都妫水北，为王庭。其余小众不能去者，保南山羌，号小月氏。

本段主要讲大月氏和小月氏的渊源，差异主要体现在历史事实的明确程度和叙事顺序上。在谈论“控弦者”（士兵）时，《史记》言“可一二十万”，而《汉书》言“十余万”，明确地说明大月氏的士兵少于二十万，比《史记》精准。另外“过大宛”一句《汉书》相比于《史记》明确指出是大宛而非小宛，颇为严谨。在叙述大月氏最开始居敦煌、祁连时，《史记》采用倒叙手法，而《汉书》将这一句提前，按照时间顺序叙述。

例2：《杜周传》与《酷吏列传》（节选）

杜周者，南阳杜衍人也。义纵为南阳太守，以周为爪牙，举为廷尉史。事荐之张汤，汤数言其无害，至御为廷尉史。使案边失亡，所论杀甚众多。奏事中上意，任用，与减宣相编，更为中丞者十余岁。其治与宣相放，然周少言重迟，外宽，而内深次骨。

(c) 2022 中国中文信息学会计算语言学专业委员会

词长	排序	《史记》		《汉书》	
		词	频率（‰）	词	频率（‰）
2	1	天下	3.01	天下	2.53
	2	诸侯	2.32	匈奴	1.63
	3	于是	2.06	以为	1.54
	4	太子	1.52	丞相	1.44
	5	天子	1.50	天子	1.22
	6	匈奴	1.03	诸侯	1.17
	7	孔子	0.97	陛下	1.07
	8	以为	0.96	于是	1.05
	9	丞相	0.88	太子	1.01
	10	将军	0.88	单于	0.98
3	1	太史公	0.39	大将军	0.62
	2	大将军	0.37	大司马	0.51
	3	平原君	0.29	二千石	0.45
	4	孟尝君	0.25	关内侯	0.25
	5	淮南王	0.20	京兆尹	0.22
	6	二千石	0.20	皇太后	0.20
	7	齐桓公	0.18	左将军	0.20
	8	孝文帝	0.17	光禄勋	0.19
	9	秦昭王	0.15	董仲舒	0.19
	10	春申君	0.13	大司农	0.18
4	1	御史大夫	0.25	御史大夫	0.66
	2	骠骑将军	0.10	光禄大夫	0.31
	3	越王勾践	0.07	车骑将军	0.22
	4	孝文皇帝	0.06	太皇太后	0.14
	5	车骑将军	0.05	票骑将军	0.12
	6	吴王夫差	0.04	太中大夫	0.11
	7	齐悼惠王	0.04	水衡都尉	0.09
	8	太中大夫	0.04	孝文皇帝	0.07
	9	公子弃疾	0.03	司隶校尉	0.06
	10	二师将军	0.03	孝武皇帝	0.06

Table 9: 《史记》《汉书》高频词

本段中，《汉书》对于《史记》做了两处重要的改动。第一是杜周任廷尉史一事，《史记》记载杜周是先当了廷尉史再为张汤做事，而《汉书》则记载杜周是先被推荐给张汤才当上了廷尉史。这一差异说明《汉书》对其先后顺序进行了考订，肯定了张汤对酷吏杜周任职的关键作用。第二，《汉书》相比《史记》突出描写了杜周“少言”之特点，更显得杜周为人冷酷，不近人情，其人物形象顿时丰满。可以说，尽管《汉书》这一段对《史记》的改动十分微小，但是“微言大义”，把杜周的人物形象刻画地更加准确、立体、生动。

附录C. NER模型超参数与《四库全书》数据集信息

本文使用的NER模型采用BERT+Bi-LSTM+CRF架构，其主要超参数如Table 11。

《四库全书》是中国古代现存规模最大的丛书，包含了从先秦到清朝的古籍超过三千种。预训练使用的数据集是现代人对大多数现存《四库全书》进行电子化的成果。该数据集质量高、覆盖广，适合于构建预训练模型。语料全部为繁体字，没有标点符号，总字数在6亿左右。

(c) 2022 中国中文信息学会计算语言学专业委员会

实体类型	排序	《史记》		《汉书》	
		提及	频次	提及	频次
人物	1	孔子	400	莽	734
	2	汉王	338	光	370
	3	项羽	273	汉王	337
	4	高祖	265	王莽	286
	5	秦王	246	汤	281
	6	沛公	242	武	269
	7	赵王	227	信	242
	8	齐王	216	高祖	234
	9	汤	208	禹	232
	10	项王	203	羽	210
地点	1	秦	2601	汉	1545
	2	楚	1681	匈奴	968
	3	齐	1622	秦	771
	4	汉	1041	楚	668
	5	赵	1030	齐	555
	6	魏	805	长安	407
	7	晋	746	赵	359
	8	燕	580	周	309
	9	韩	567	河	291
	10	吴	511	吴	258
书籍	1	春秋	81	诗	243
	2	诗	77	春秋	220
	3	书	46	易	210
	4	易	31	书	133
	5	尚书	23	易传	73
	6	老子	19	尚书	55
	7	颂	16	礼	43
	8	礼	13	五经	35
	9	武	11	论语	33
	10	雅	9	左氏传	32
官职	1	太子	693	丞相	859
	2	丞相	497	将军	682
	3	将军	446	陛下	639
	4	大夫	284	太子	605
	5	陛下	269	上	539
	6	太后	257	太守	499
	7	御史大夫	195	太后	496
	8	大将军	189	御史大夫	395
	9	上	160	大夫	380
	10	夫人	139	大将军	372

Table 10: 《史记》《汉书》高频命名实体

第二十一届中国计算语言学大会论文集，第656页-第670页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

超参数	值
编码器骨架	BERT-base
BERT编码器层数	12
BERT隐向量维度	768
BERT自注意力头数	12
Bi-LSTM层数	2
Bi-LSTM隐向量维度	100

Table 11: NER模型超参数

CCL 2022

第二十一届中国计算语言学大会论文集，第656页-第670页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

生成模型在层次结构极限多标签文本分类中的应用

陈林卿*，何大望，肖燕思，刘依林，陆剑平，王为磊
（智慧芽信息科技有限公司，江苏 苏州 215000）
{chenlinqing,hedawang,xiaoyansi,liuyilin,lujianping,wangweilei}@patsnap.com

摘要

层次结构极限多标签文本分类是自然语言处理研究领域中一个重要而又具有挑战性的课题。该任务类别标签数量巨大且自成体系，标签与标签之间还具有不同层级间的依赖关系或同层次间的相关性，这些特性进一步增加了任务难度。该文提出将层次结构极限多标签文本分类任务视为序列转换问题，将输出标签视为序列，从而可以直接从数十万标签中生成与文本相关的类别标签。通过软约束机制和词表复合映射在解码过程中利用标签之间的层次结构与相关信息。实验结果表明，该文提出的方法与基线模型相比取得了有意义的性能提升。进一步分析表明，该方法不仅可以捕获利用不同层级标签之间的上下位关系，还对极限多标签体系自身携带的噪声具有一定容错能力。

关键词： 极限多标签文本分类 ； 层次结构极限多标签 ； 生成模型

Generation Model for Hierarchical Extreme Multi-label Text Classification

CHEN Linqing, HE Dawang, XIAO Yansi, LIU Yilin, LU Jianping, WANG Weilei
(PatSnap Co., LTD. Suzhou, Jiangsu 215000)

Abstract

Hierarchical extreme multi-label text classification task is an important yet challenging task in Natural Language Processing. This task is complex due to the enormous number of labels and corresponding hierarchy relationships in the label system. We propose to view the hierarchical extreme multi-label text classification task as a generation problem and present a novel soft-constrained method for label decoding, which views the output labels as a sequence, rather than as a single label. Rigorous experiments demonstrated that our model is effective at picking out relevant labels directly from thousands of hierarchical labels. Experiments also show that the proposed methods have achieved significant improvements across several datasets. With further analysis, our methods not only capture and utilize the hierarchical structure information between labels at different levels, but also represent the relationships of the labels within the same level.

Keywords: extreme multi-label text classification , hierarchical labels , generation model

*Corresponding author
©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

多标签文本分类利用自然语言处理高效归纳海量文本信息：给定输入文本，从标签集合中返回与输入文本最相关的标签子集。可以将多标签文本分类问题看作学习评分函数的过程，该函数将(实例，标签)对(x, y)映射到分数$f(x, y)$。函数f在模型训练过程中不断优化，使高度相关的(x, y)对获得高分，而不相关的配对得分较低。具有数千甚至更多类别标签的文本分类任务被称为极限多标签文本分类(Liu et al., 2017)。许多现实世界的应用场景都可以看作这种形式。例如，在开放领域的问答中，x代表一个问题，y代表一个包含答案的文章 (Chang et al., 2020; Lee et al., 2019)。相应的，在层次结构极限多标签文本分类（HXMC, Hierarchical EXtreme Multi-label text Classification ）任务中，x表示文本，y表示具备层级结构的标签体系中的一个或多个类别标签。现实应用场景中数据产生速度快，体量大，具有明显的多样性和复杂性。与之对应的标签数量可能高达数万甚至数十万。极限多标签文本分类在档案文献管理，文本资料分类检索等场景具有广泛的应用前景。

极限多标签文本分类任务类内类间样本关系复杂，导致标签语义存在部分重叠并非完全正交。微软发布的学术图谱数据集MAG（Microsoft Academic Graph）(Shen et al., 2018)就是层级结构极限多标签文本分类数据集中的典型。现有研究方法主要有两大类，其中一类利用标签向量压缩 (Liu et al., 2017; Bhatia et al., 2015)等方法减少标签向量维度实现十万级别标签的分类任务，该类方法丢失部分标签语义信息，忽略标签之间的相关性及其层次结构。另一类考虑到标签结构信息的研究工作则多使用硬性约束，通过多次分类，聚类 (Chang et al., 2020)等方法间接或分步骤实现极限多标签分类。忽略了同一文本可以属于多个交叉领域的事实，也没有充分考虑极限多标签体系由于信息量巨大，需要一定的容错能力，若分类/聚类中间步骤产生错误，会一直传播到后续分类结果，形成由错误传递导致的系统性偏见。

本文受到序列到序列(Seq2Seq)模型在机器翻译 (Bahdanau et al., 2014; Luong et al., 2015; Sun et al., 2017)，摘要总结 (Rush et al., 2015; Lin et al., 2018)，风格转移 (Shen et al., 2017; Xu et al., 2018)等一系列序列转换任务上广泛应用的启发，提出利用基于并行多头注意力机制的生成模型来解决层次结构极限多标签文本分类任务。该序列生成模型由带有注意机制的编码器和解码器组成,显著缓解了之前研究工作中CNN感知域太小，LSTM长文本编码能力弱并且编码解码速度慢的缺点，同时多头注意力机制可以分别关注文本的不同部分，保留了CNN多通道输出的优点。解码器基于具备软约束机制的注意力和柱状搜索算法，在之前预测的标签基础上预测下一个标签，挖掘并利用标签序列内部依赖信息。此外，本文提出的标签词表复合映射机制在保留标签体系结构信息的前提下大幅减少词表维度，避免将极限多标签任务拆分为多个分类模型，并确保模型最终输出标签一定存在于标签体系中。

本文主要贡献如下：

- 提出将层次结构极限多标签文本分类任务视为序列转换问题，利用编码器-解码器结构学习类别标签的层级结构关系。
- 提出软约束解码及词表复合映射，在保留标签语义信息的同时为文本从数十万分类中选出高相关标签。而不是将极限分类任务拆分，组合多个分类模型的输出结果作为输出标签。
- 实验表明本文提出的方法与基线模型相比取得有意义的性能提升。进一步的分析表明该方法在学习标签体系层次结构信息方面的有效性。

2 方法

本文利用神经网络自主习得层次结构极限多标签之间的从属，依赖关系，从而通过生成模型为待分类文本匹配多个具有层级依赖关系的类别标签。为了实现这个目标，本文方法在训练过程中将源端编码器自注意力层输出的词级隐藏状态作为文本编码结果。使用解码器对含有软约束层次结构信息的标签序列进行解码。在预测阶段通过柱状搜索预测标签序列，并通过恢复子词化标签及词表映射得到最终输出标签。该章节将详细介绍本文提出的模型的必要细节，为了方便理解还将对一些概念及问题做出定义和解释。

2.1 层次结构极限多标签文本分类

在层次结构极限多标签文本分类任务中。给定具有L个标签的标签空间$\mathcal{L} = \{l_1, l_2, \cdots, l_L\}$，一个包含$m$个字符的文本序列$\mathcal{X} = \{x_1, x_2, \cdots, x_m\}$，目标是将一个与文本高

(c) 2022 中国中文信息学会计算语言学专业委员会

输入

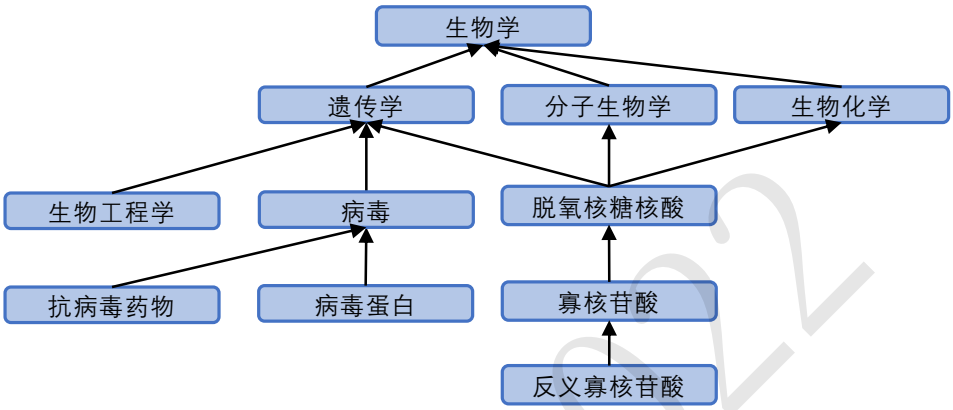
标题：抗冠状病毒的反义寡聚核苷酸及其制药用途

摘要：本发明公开了抗冠状病毒的反义寡聚核苷酸及其制药用途，涉及生物工程领域，解决现有抑制病毒复制的研究大多靶向冠状病毒入侵细胞后被宿主细胞已经翻译生成的RdRp蛋白本身，而不是抑制RdRp的翻译。本发明公开的反义寡聚核苷酸及其联合应用，可特异性地结合冠状病毒5'UTR中IRES序列上的关键茎环结构，寡聚核苷酸序列选自A21，B21，E21和F21等。本发明针对冠状病毒基因组和亚基因组5'UTR进行抗病毒药物设计，是从干扰病毒侵染后多肽段非结构蛋白ORF1ab基因和病毒蛋白翻译的角度考虑，而不是等病毒蛋白大量翻译后再以病毒蛋白作为药物靶点，本发明是以最小的抗冠状病毒成本投入获得最大抗冠状病毒效益产出。

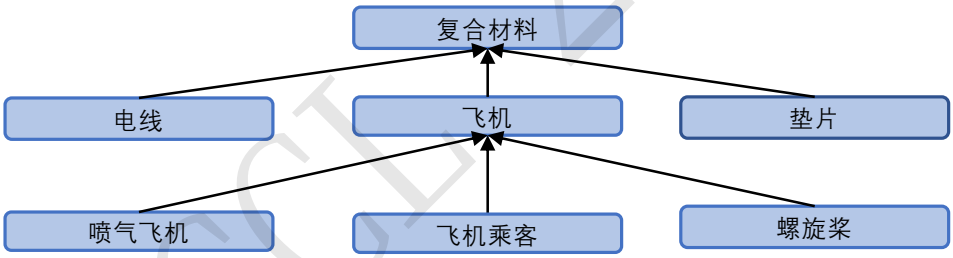
输出

药物；抗病毒药物；病毒蛋白；生物工程学；反义寡核苷酸；基因组；冠状病毒；

(a)：抗病毒生物医药相关专利



(b)：样例专利部分标签的关系示意



(c)：MAG标签体系噪声示例

Figure 1: (a): 样例文本及模型输出的部分标签; (b): 相关标签的层间及层内关系;(c): MAG标签的噪声示例

度相关的$\mathcal{L}$的子集$\mathcal{Y}$分配给文本。与传统的单标签分类只分配给每个文本一个标签不同，每个样例可以有多个标签，且标签之间有一定的从属，依赖关系。标签空间$\mathcal{L}$与文本$\mathcal{X}$之间可以映射n个l。从数学角度来看，HXMC任务可建模为寻找最优标签序列$\mathcal{Y}$的最大化条件概率$p(\mathcal{Y}|\mathcal{X})$问题，其计算公式如下:

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{P}(\mathbf{y_i}|\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_{i-1}}, \mathcal{X}), \tag{1}$$

其中，训练集以待分类文本及标签子集对$(\mathcal{X}, \{y_i\}_{i=1}^{n})$的形式给出。$x_i \in \mathbb{R}^D$，$y_i \in \mathbb{R}^L$。$D$表示文本$\mathcal{X}$中字符$x_i$的特征向量维度,$L$表示标签$y_i$特征向量维度。

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

MAG[0]是由微软构建的开源异构知识图。如图 1所示，作为层次结构极限多标签体系的典型代表，MAG标签体系具备以下特点：

- 标签数量多，MAG标签体系有约70万个分布在6个不同层次上的类别标签，如图 1(a)所示，一条文本可以有多个高度相关的标签。
- 标签关系复杂，如图 1(b)所示，MAG标签间有层次关系，且一个标签可以从属于多个父节点标签。同层标签之间有一定关联。
- 存在错误信息，由于MAG标签数量高达数十万，其上下位关系甚至标签本身可能存在噪声。图 1(c)展示了其中一个样例，在“复合材料”父节点标签下存在“飞机乘客”子标签，这使得其他研究工作中将极限多标签分类任务按类别拆分成多个分类任务的硬约束方法可能造成错误传递并最终导致系统偏见。

2.2 文本编码与标签生成

该小节介绍G-HXMC模型（Generation mode for Hierarchical EXtreme Multi-label text Classification）的必要细节。本文编码-解码结构基于自然语言处理任务中广泛应用的Transformer(Vaswani et al., 2017)，这里主要介绍涉及本文模型编码，解码的核心部分，其他通用组件如前馈神经网络，残差连接等与经典Transformer相同，因篇幅所限不再详细展开。

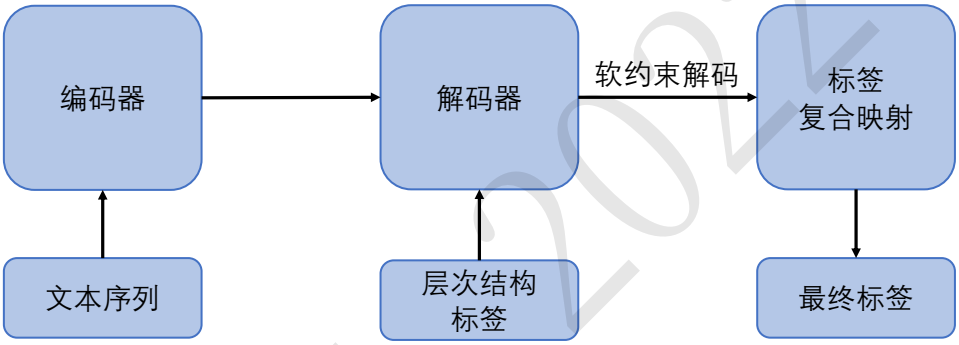


Figure 2: 层次结构极限多标签分类

如图 2 所示，本文提出模型的主要结构由基于多头并行注意机制的编码器和包含软约束及词表复合映射机制的解码器组成。由于模型的生成特性，训练阶段的标签解码与预测阶段的标签生成过程不完全相同。

文本编码 编码器输入由文本标题及摘要拼接成的长序列。与其他序列转换任务类似 (Press et al., 2016)，本文通过词嵌入层将输入序列和输出序列转换为维度D的向量，模型两端的词嵌入层共享同一个线性变换权重矩阵。由于本文长度较长，为了感知输入文本的顺序增加与向量相同维度$\mathbb{R}^{model}$的绝对位置编码(Vaswani et al., 2017)作为位置信息。同时，为了使模型可以感知标题与摘要的区别，本文利用“分段位置编码”区分文本的不同部分。输入序列公式表达如下：

$$I = Concat(PE_1 + emb(T), PE_2 + emb(A)), \tag{2}$$

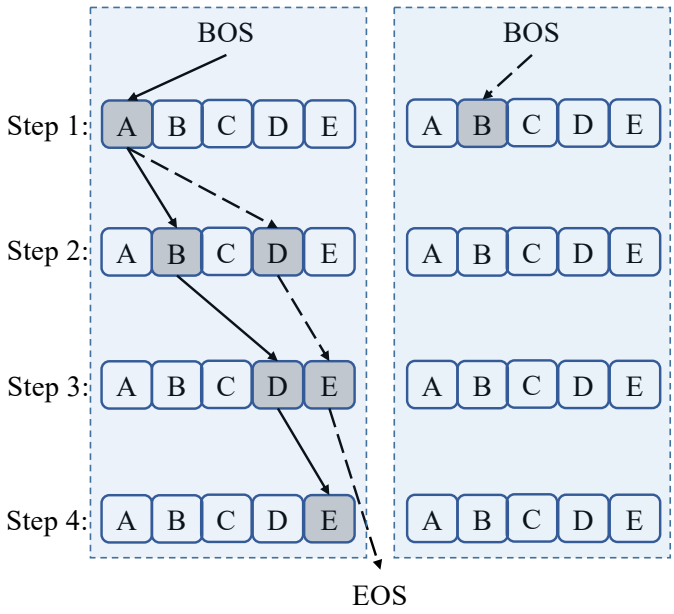
其中，I表示输入序列，T和A分别表示待分类文本的标题与摘要，emb表示词嵌入，PE表示位置编码。

输入文本的不同组成部分并非同等重要，本文使用多头注意力机制捕获不同位置字符之间的关系，对文本进行编码。其公式表达如下：

$$I^{(k)} = MultiHead\left(I^{(k-1)}, I^{(k-1)}, I^{(k-1)}\right), \tag{3}$$

其中，$I \in \mathbb{R}^{model}$表示经过词嵌入的输入序列，K代表编码器层数。

[0]https://www.microsoft.com/en-us/research/project/microsoft-academic-graph



A:生物学;B:遗传学;C:分子生物学;D:生物化学;E:生物工程学

Figure 3: 层次结构标签的生成

标签解码 G-HXMC通过自回归函数:$score(\mathcal{L}|\mathcal{X}) = p_\theta(\mathcal{Y}|\mathcal{X}) = \prod_{i=1}^{N} p_\theta(y_i|y_{i<i}, \mathcal{X})$。其中$\mathcal{Y}$是属于$\mathcal{L}$的具有$N$个子词化字符的标签集合，$\theta$为模型的参数。利用Techer Forcing (Sutskever et al., 2014)最大化输出序列似然度（likelihood），并用dropout(Srivastava et al., 2014)和标签平滑 (Szegedy et al., 2016)进行归一化。简洁地说，本文训练目标即神经机器翻译中的常用目标，通过模型参数θ最大化$logp_\theta(\mathcal{Y}|\mathcal{X})$。

解码器主要组件为两个多头注意力层，一个对标签序列进行编码另外一个利用编码器输出的文本编码结果对标签序列进行解码。解码过程的多头注意力机制公式表达如下:

$$O^{(k)} = MultiHead\left(I^{(k)}, L^{(k)}, L^{(k)}\right), \tag{4}$$

其中，L表示编码后的标签序列，I表示编码后的文本序列，K表示解码器的层数。

标签序列进行编码的方式与公式 3 表示的编码过程类似，但编码内容是目标端标签序列。其他常见组件如前馈神经网络等因篇幅所限亦不再展开。

Cao (2021)等人在具有层次结构的实体召回研究中提出对标签输出过程进行限制。即每一步都进行检查，使得输出结果必然属于前一标签的下位词。强制约束可能存在一些弊端：每个词基于上个输出可能造成错误传递；每一步都进行检查增加时间开销；最重要的是，该约束过程在模型训练过程中并不存在，模型无法学习这种硬约束模式。通过观察数据我们发现，高层级标签天然的具有出现频次高的特性。基于此，本文提出同时在模型的输出及训练过程中对标签进行软约束。根据训练集标签的出现频率对目标端标签进行排序，高层标签排在低层标签前面，同一层标签中高频标签放置在序列前端。

标签生成 极限多标签分类研究工作大多为每条输入文本计算一个维度与标签数一致的输出向量，通过贪心搜索直接为待分类文本选择概率最高的top$N$标签。然而当标签数量很大时，其计算的时间开销和经济成本都十分可观，例如本文使用的MAG数据集有数十万不同标签。同时贪心搜索追求单个位置的最大概率，容错能力较弱，而本文目标是整个序列的概率最大而不是单个标签概率最大。

本文利用介于贪心搜索和广度搜索之间的柱状搜索 (Sutskever et al., 2014)解码策略。如图 3 所示，通过保留一定容错空间缓解贪心搜索可能发生错误传递的缺点。使用柱状搜索的时间开销不取决于词表的大小，只取决于解码过程中保持柱的数量（K）以及解码长度。图中示例K为2，即同时保持2条总体概率最高的候选链路。图 3 中分别以虚线和实现代表两条候选序列实际路径，“Setp2”解码第2个标签时，所有以B标签为第一个标签的候选序列“B-X”概率都小

(c) 2022 中国中文信息学会计算语言学专业委员会

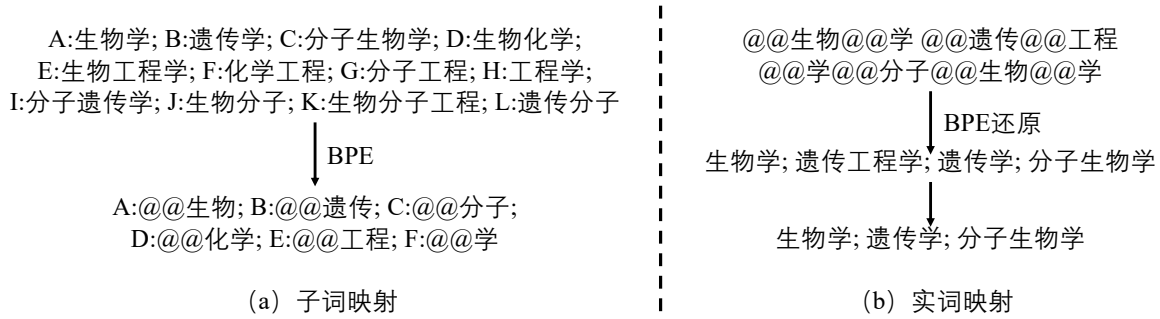


Figure 4: 模型词表与标签的二级映射

数据集	论文			专利		
	训练集	开发集	测试集	训练集	开发集	测试集
样本数	1.97M	5,000	5,000	3.73M	5,000	5,000
标签数	16.90M	42,966	43,206	22.98M	30,797	30,623
平均字符	188.55	190.14	190.34	129.81	131.38	130.91
平均标签	8.54	8.59	8.64	6.15	6.16	6.12

Table 1: 训练集，开发集及测试集的统计信息

于以A标签为第一标签的序列“A-B”和“A-D”，所以右侧柱实际凋零。不同生成路径的长短也有可能不同，虚线路径在生成3个标签后先触发了序列终止符“EOS”。

本文提出的标签词表复合映射主要包括预处理阶段的子词化映射和输出结果阶段的实词映射。受神经机器翻译领域研究工作的启发，本文通过BPE (Sennrich et al., 2016)在保留标签语义信息的前提下缩小解码器端词表。如图 4(a) 所示，子词化后的类别标签数量从约70万减少到不高于3万。如图 4(b) 所示，实词化映射通过筛除不属于MAG体系的标签确保模型输出的最终结果是体系内有实际意义的标签。

3 实验

3.1 数据集

微软学术图(MAG)是一个包含论文和专利等科学出版物，出版物间引用关系，以及作者、机构、期刊、会议和研究领域等信息的开源异构图。基于该知识图谱的数据被用于改善Bing, Cortana, Word和Microsoft Academic的体验。本文实验未使用MAG全量数据，表 1 列举了本文实验使用的专利及论文数据集的部分统计信息，‘M’表示百万，采样方式为随机采样。

3.2 实验设置

本文利用THUMT[1](Tan et al., 2020) 实现基于序列转换的多分类模型，通过拓展词表映射和软约束解码进一步实现层次结构极限多标签文本分类模型。本文在3.5 节列出的实验中，将模型隐藏状态向量的维度设为512，每个编码器解码器的层数都设置为6，多头注意力机制中注意力头的个数都设置为8，柱状搜索的大小设置为5，dropout比例设置为0.1。本文在模型训练过程中将批大小设置为40280个字符并使用$\beta_1 = 0.1$的Adam优化器对模型进行优化 (Kingma and Ba, 2015)。

3.3 评价标准

本文参考之前的研究工作(Wang et al., 2021)，采用微F_1(Micro-F_1)作为主要评价指标。同时报告微准确率（Micro-P）和微召回率（Micro-R）作为辅助参考。Micro-F_1(Manning, 2008)可以理解为Micro-Precision和Micro-Recall的调和平均值。

[1] https://github.com/THUNLP-MT/THUMT

(c) 2022 中国中文信息学会计算语言学专业委员会

模型	MAG-Paper					
	0-1			0-5		
	m-F	m-P	m-R	m-F	m-P	m-R
#1 G-HXMC	76.38	80.65	72.54	60.64	64.39	57.30
#2 + 软约束解码	80.48	**84.68**	76.68	**63.34**	**67.03**	**59.91**
#3 SGM(Yang et al., 2018)	69.89	69.98	69.80	50.81	63.40	42.39
#4 + Global Emb	70.64	70.69	70.59	50.90	63.78	42.35
#5 + 软约束解码	70.98	71.06	70.90	52.90	65.15	44.53
#6 BERT(Devlin et al., 2018)	78.25	79.38	77.15	——	——	——
#7 + 软约束信息	**81.03**	82.09	**80.00**	——	——	——

Table 2: 本文模型在MAG-Paper任务上的MicroF$_1$性能(%)

模型	MAG-Patent					
	0-1			0-5		
	m-F	m-P	m-R	m-F	m-P	m-R
#1 G-HXMC	64.30	68.34	60.71	57.73	66.75	50.85
#2 + 软约束解码	**71.80**	**76.06**	**67.99**	**64.03**	**68.48**	**60.12**
#3 SGM(Yang et al., 2018)	59.81	64.23	55.96	53.01	39.82	79.25
#4 + Global Emb	60.62	65.03	56.77	53.50	58.01	49.64
#5 + 软约束解码	62.50	67.05	58.53	56.84	61.37	52.93
#6 BERT(Devlin et al., 2018)	63.02	67.14	59.38	——	——	——
#7 + 软约束信息	70.69	75.19	66.70	——	——	——

Table 3: 本文模型在MAG-Patent任务上的MicroF$_1$性能(%)

3.4 基线模型

- BERT(Devlin et al., 2018)，利用基于Transformer的双向编码器进行预训练。预训练后的BERT作为语言模型与可以额外业务层结合广泛应用于下游任务，如问答和语言推断。
- SGM(Yang et al., 2018)，使用基于LSTM的序列转换模型解决极限多标签文本分类。该模型记忆之前时间步的输出并加以利用，用于缓解LSTM的遗忘效应。
- HSG(Wang et al., 2021)，提出一种利用层级标签语义信息引导的模型提升策略,在训练和预测过程中给予模型弱监督语义指导信息,从而规约对应的多标签语义边界。

3.5 实验结果

G-HXMC表示本文基于生成范式的层次结构极限多标签文本分类方法。**软约束解码**表示前文提到的对解码端标签进行排序并结合词表复合映射的方法。**SGM**的实验结果中，**Global Embedding**表示相关论文中使用之前时间步解码结果缓解LSTM遗忘信息的方法。本文作者对该模型进行拓展，使其可以应用本文提出的软约束解码，并报告了相关实验结果。**BERT**不具备解码器结构，无法直接从数十万标签中选出与文本高相关的结果，未报告其0-5层标签体系上的分类实验结果，仅在0-1层约200个标签的范围内进行了实验对比。**软约束信息**指仅对目标端标签序列进行软约束处理，用以验证标签的依赖，从属关系是否会给BERT带来精度增益。**HSG**是一种训练策略而非独立模型，表 5 中对比了G-HXMC及该文方法在**Wiki10-31**数据集上的最佳性能。

表 2 中列出本文提出模型及方法在MAG-Paper数据集上的性能结果。其中左侧为0-1层分类结果，右侧为0-5层分类结果。#2，#5，#7中的实验数据表明，本文提出的方法与基线模型相比取得了有意义的性能提升。其中，本文方法在0-1层和0-5层的分类任务上比SGM提高了约10个点，在0-1层分类任务上与使用大规模预料预训练过的BERT相比使用小的多的数据集和训练开销取得了相近性能。#1与#2的对比表明，本文提出的软约束机制给模型性能带来了显著提升。#3与#5，#6与#7的对比表明本文提出的软约束机制不仅可以给本文提出的模型带来性能增益，也可以帮助基准模型提高性能。

表 3 中列出本文提出方法在MAG-Patent数据集上的实验结果。本文提出的模型在所有层级分类任务中都达到了最佳性能。#2，#5，#7中的实验数据表明，本文提出的方法与基线模型相比取得了有意义的性能提升。其中，本文方法在0-1层和0-5层的分类任务上与SGM和BERT的性能比较趋势与MAG-Paper上的实验结果类似，皆表明本文提出的软约束机制不仅可以给本文提出的模型带来性能增益，也可以帮助基准模型提高性能。

(c) 2022 中国中文信息学会计算语言学专业委员会

模型	材料学	计算机	化学	工程学	生物学	物理学	医药	平均
G-HXMC	73.10	59.64	66.08	56.21	61.00	57.74	63.00	62.41
+ 软约束解码	**75.07**	**63.53**	**69.13**	**59.20**	**65.47**	**61.97**	**65.33**	**65.66**
SGM(Yang et al., 2018)	62.98	43.10	53.18	47.77	46.74	50.15	49.10	50.43
+ 软约束解码	65.10	45.80	55.40	49.90	49.70	52.00	53.00	52.99

Table 4: MAG-Paper Level 0 标签分类任务MicroF$_1$性能(%).

表 4 列出只根据部分0层标签对MAG-Paper数据集进行分类的实验结果，并将本文模型与基准模型的分类性能进行对比。观察可以发现，本文提出的模型及方法比基准模型具备更强的层次结构信息利用能力，在使用软约束机制之后获得的增益更高。本文方法在使用软约束机制利用标签层级信息后Micro-F$_1$平均增加了3.15，基准模型在使用软约束机制利用标签层级信息后Micro-F$_1$评测标准上平均提高了2.56。

4 分析讨论

为了观察本文提出的基于软约束机制和复合词表映射的生成模型是如何提高层次结构极限多标签文本分类质量的，本文在该章对模型输出标签个数，输出标签质量及一些实验设置进行进一步的实验与分析。实验结果表明，本文模型不仅可以直接从数十万标签中选出与文本相关的标签，还可以自主习得标签体系内部的层次结构信息。

模型	m-F	m-P	m-R
G-HXMC	**56.62**	**61.87**	**52.19**
SGM(2018)	49.89	49.42	50.37
HSG(2021)	47.63	45.26	50.26

Table 5: Wiki10-31上MicroF$_1$性能(%).

模型	参数（百万）	时间（小时）
G-HXMC	66.86	36.5
BERT(2018)	109.64	**30.1**
SGM(2018)	**30.99**	50.3

Table 6: 训练时间和模型参数.

4.1 训练时间和模型参数

表 6 中列举了本文模型和基线模型的参数量及训练时间。其中，SGM的模型参数最少，但由于循环神经网络的特性，其训练时间最长。BERT模型的参数达到1.09亿，由于没有解码器，在不考虑预训练时间的前提下训练时间最短。同时由于BERT没有解码器结构，不能直接从数十万标签中选出与输入文本高度相关的标签，经过改造增加编码器可以实现这一目标，这一方法本质上也是生成模型的范畴，但会使得模型十分庞大，性能也并未取得有意义的增益，不在本文讨论范围内。本文提出的模型及方法在模型规模，训练时间开销等方面的表现较均衡。

4.2 层次结构信息利用效果

模型	标签（个）
G-HXMC	7.07
+ 上下位关系	**7.75**
SGM(Yang et al., 2018)	5.20
+ 上下位关系	5.51

Table 7: 模型输出标签序列长度对比.

本文在表 7 中通过观察不同模型增加软约束层次结构信息前后输出标签个数评估模型利用层次结构信息的能力。该分析基于MAG-Paper数据集。G-HXMC在增加标签上下位信息后标签数量平均增加了0.7个，而SGM则增加了0.3个。由于SGM在增加标签层次结构信息之前的输出标签个数也较少，本文作者推测LSTM不善于编码长序列的缺点限制了SGM对标签层次信息和依赖关系信息的利用。如表 1 所示，数据集中每条文本的平均标签数量不少于8个，且标签具备6个层级，SGM模型的输出标签数量意味着平均每个层级只有不到一个标签，输出标签较少无疑对SGM模型的分类性能带来了较大负面影响。

4.3 层次结构信息利用方式

本文在图 5 中对比了MAG-Paper数据集上标签层次结构信息不同利用方式对分类性能

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

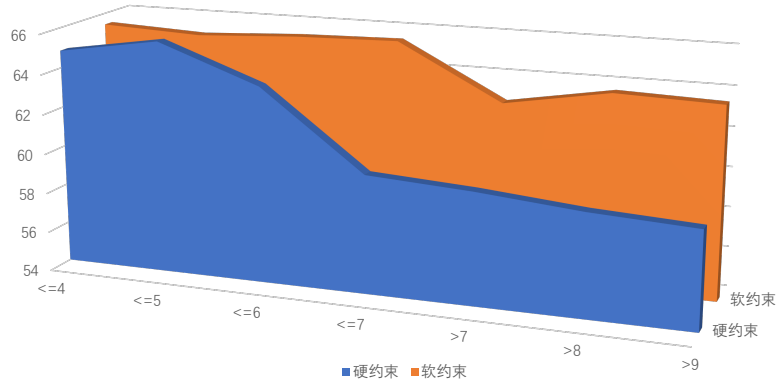


Figure 5: 标签层次结构信息不同利用方式对分类结果的影响MicroF$_1$性能(%).

带来的影响。其中，横坐标表示模型输出标签个数，纵坐标表示对不同长度的结果分别测试MicroF$_1$性能(%)。硬约束表示根据标签层次结构信息对标签从属关系做硬性规定，要求生成的标签必须是前一个标签的下位从属标签。软约束则表述本文所使用的方法。两种利用方式都基于本文提出的G-HXMC模型。通过观察可以看出，硬约束方式输出结果在标签个数较多的时候出现性能大幅下降的现象，而软约束输出结果的性能则较均衡。

4.4 消融实验

标签层级	F1
Level 0-1	71.80
Level 2-5	59.10
Level 0-5	64.03

Table 8: 生成标签层级分布性能对比.

解码长度	F1
5	35.21
10	64.03
15	60.21

Table 9: 不同解码长度对分类性能影响.

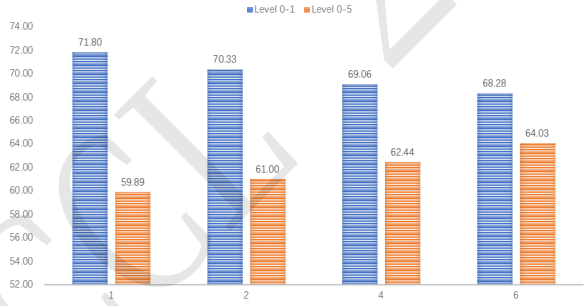


Figure 6: 不同层级分类任务中编码器层数的影响

本文在表 8 中列出了MAG-Patent数据是否利用上位标签信息对分类结果影响的对比。其中，横坐标表示编码器层数，“Level 0-1”指训练数据只有0层和1层标签。“Level 2-5”表示训练数据只有2至5层标签。“Level 0-5”指训练数据包含0-5层的所有标签，但只评估2-5层的分类性能。观察表中结果可知，由于0-1层标签数量较少，标签层次结构清晰无噪声，分类结果总体好于多层次分类结果。另外，对0-5层标签一起学习的分类结果好于只学习2-5层标签的分类结果，表明来自0-1层的优质层次结构信息和依赖关系信息可以给2-5层标签分类带来有意义的增益，验证了本文提出的利用标签体系层次结构信息方法的意义及其有效性。

本文在表 9 中列出了MAG-Paper数据集上，调整解码长度，即输出标签个数对分类结果的影响。表中数据表明使模型输出标签个数与训练数据样本平均标签个数相近可以取得最佳分类效果。这一现象与本文作者的直觉相符。解码长度惩罚系数对模型输出标签个数造成的影响所带来的分类性能变化也体现出类似趋势，由于篇幅所限，本文不再展开介绍。

本文在图 6 中给出编码器层数设置对MAG-Patent数据集不同层级分类任务的影响。其中，“Level 0-1”指只对0层和1层标签进行分类。“Level 0-5”则表示利用标签体系中的所有标签进行分类。观察图中直方图可以发现，0-5层标签分类任务中编码器层数越多性能越好，这一现

(c) 2022 中国中文信息学会计算语言学专业委员会

象和只对0-1层标签进行分类的实验结果相反。该试验结果可能表明输出标签之间的依赖关系越复杂，生成序列越长，需要的编码器层数也越多。

4.5 样例分析

标题	处理和再生废油产品的方法
摘要	从润滑油或工业油中回收再生矿物油和合成油时..... 然后将搅拌的混合物进行滗析操作，.....
G-HXMC	..... 润滑剂| 石油
+ 硬约束	..... 润滑剂| 油脂	汽车
+ 软约束	..... 润滑剂| 矿物油| 滗析

Table 10: 层次结构信息利用方式样例分析

表 10 中对比了同一专利在不利用标签层级信息及软约束，硬约束两种不同方法利用标签层次结构信息时的输出标签结果。观察样例可以发现，不利用任何标签层级信息时，模型输出标签数量较少。硬性约束的方式利用层次结构信息后虽然输出标签数量增加，但出现了明显不合适的标签“油脂”。软约束利用标签层次结构信息时则给出了专业领域技术标签“滗析”。

标题	具有优良耐磨性和低摩擦系数的钛-石墨烧结复合材料
摘要	.....耐磨性和低摩擦特性的烧结钛-石墨的方法。生产具有可控孔隙率的三相结构的..... .....由于其生物相容性.....该复合材料可用于生物医学工程和其他工程领域.....
硬约束	复合材料| 飞机乘客	石墨| 耐磨
G-HXMC	复合材料| 孔隙率	钛| 烧结| 石墨| 耐磨| 生物相容性

Table 11: 模型输出容错能力样例分析

表 11 中样例对比了不同层次结构信息利用方式对MAG噪声的容错能力。观察样例可以清晰发现，软约束方式不但可以生成更多更贴近专利的标签：“生物相容性”，“孔隙率”等，还绕过了从属于“复合材料”的子标签“飞机乘客”。

5 相关工作

极限多标签文本分类早期工作聚焦基于启发式方法改进传统机器学习方法以适应任务。如Liu等人 (2005)尝试将极限多标签分类任务转化为多个基于支持向量机的二分类问题，这种简洁的策略至今仍被广泛应用；Cai等人 (2004)提出了一种基于支持向量机的层次分类方法来解决极限多标签文本分类任务；SLEEC (Bhatia et al., 2015)通过压缩标签向量维度等辅助手段缓解极限多分类标签文本分类任务中类别标签的“长尾分布”问题。这些研究工作多基于机器学习方法，利用词袋模型对文本语义进行建模。词袋模型忽略词出现的顺序及其之间的语义联系，无法利用上下文，限制了模型理解、分类文本的能力。

神经网络近年来在自然语言处理领域取得了一系列引人注目的成果。一些科研工作者开始在极限多标签任务中应用神经网络模型并取得重要进展。例如,Zhang和Zhou (2006)使用具有成对排序损失函数的全联接神经网络处理功能基因和文本的分类任务。Kurata等人 (2016)提出使用卷积神经网络(CNN)进行多标签分类。Chen等人 (2017)使用卷积神经网络（CNN）和递归神经网络(RNN)来捕捉标签及文本的语义信息。Liu等人 (2017)利用多标签文本分类中经典的深度学习算法如Text Convolutional Neural Networks(TextCNN)、Text Recurrent Neural Networks（TextRNN）、FastText等对文本进行特征抽取从而提升文本语义表示性能。这些研究工作取得了长足进展，但仍未突破卷积神经网络感知域的桎梏以及循环神经网络不善于编码长文本的瑕疵。

在上述工作的基础上，研究者们进一步进行各种优化和改进。Chang等人 (2020)开始尝试利用预训练模型帮助分类任务，但需要利用聚类等机器学习方法对文本进行粗分类；Liu等人 (2017)和Bahatia (2015)则通过压缩标签向量维度降低极限多标签分类任务难度。然而，这些方法大多忽略标签之间的关联，也没有让模型自主习得层次结构标签体系内标签的上下位关系及错综复杂的联系；SGM (Yang et al., 2018)提出利用LSTM对文本进行编码和解码，并通过利用全局信息缓解LSTM不利于长文本编码的弊端。该方法仅在小规模多标签文本分类任务上得到应用，并不适用于具备层次结构的极限多标签文本分类任务；Wang等人 (2021)提出通过标签语义信息加强来改善极限多文本分类任务，将极限多标签分类任务划分为多个分类模型，输出

(c) 2022 中国中文信息学会计算语言学专业委员会

组合后的分类结果，忽略了错误传递问题，应用于大型深层网络后增益不明显；BERT (Devlin et al., 2018)是最近几年来得到广泛应用的多任务预训练模型。然而该模型不具备解码结构，无法应用于极限多标签分类任务，更无法习得标签之间的联系。Amigo等人 (2022)则拓展了层次结构极限多标签分类任务的评测维度。

6 总结

本文提出将层次结构极限多标签文本分类任务归入生成范式下的序列转换任务。该模型利用编码-解码结构将文本编码后输出一系列标签。为了让模型更好的自主习得类别标签体系的层级结构和依赖关系，本文提出利用软约束解码和复合词表映射帮助模型生成具有层次结构关系的标签序列。在不同类型文本数据集上与多种基线模型的对比实验表明本文提出的一系列方法取得了有意义的性能提升。进一步的消融实验表明，本文提出的软约束机制和复合词表映射方法相比之前的研究工作可以更好的学习利用极限多标签体系的层级结构信息及依赖关系。

如何通过生成模型为文本从数十万甚至上百万具备错综复杂关系的类别标签中选出相关标签，并尽可能多的利用标签间的从属，依赖关系，是一个值得探索并且具备广泛应用前景的问题。我们将在后续工作中继续从不同角度进行有意义的探索并即将分享最新进展，包括但不限于预训练，多模态信息利用，篇章信息利用等。

参考文献

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu and Yiming Yang. 2017. *Deep Learning for Extreme Multi-label Text Classification.* Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 115–124.

Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen and Wei-Ying Ma. 2005. *Support Vector Machines Classification with a Very Large-scale Taxonomy.* ACM SIGIR Explorations Newsletter, 7(1):36-43.

Lijuan Cai and Thomas Hofmann. 2004. *Hierarchical Document Categorization with Support Vector Machines.* Proceedings of the 13th ACM International Conference on Information and Knowledge Management, 78–87.

Kush Bhatia，Himanshu Jain，Purushottam Kar，Manik Varma and Prateek Jain. 2015. *Sparse local embeddings for extreme multi-label classification.* Proceedings of the Neural Information Processing Systems, 29:730-738.

Diederik P. Kingma and Jimmy Ba. 2015. *Adam: A method for stochastic optimization.* Proceedings of ICLR.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu and Houfeng Wang. 2018. *SGM: Sequence Generation Model for Multi-label Classification.* Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 115–124.

Rico Sennrich, Barry Haddow and Alexandra Birch. 2016. *Neural Machine Translation of Rare Words with Subword Units.* Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of NAACL, 4171–4186.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. *Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising.* WWW.

Yashoteja Prabhu and Manik Varma. 2014. *Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning.* KDD.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. *Pre-training Tasks for Embedding-based Large-scale Retrieval.* International Conference on Learning Representations.

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. *Latent retrieval for weakly supervised open domain question answering.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. *Neural machine translation by jointly learning to align and translate.* CoRR,abs/1409.0473.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. *Effective approaches to attention-based neural machine translation.* CoRR,abs/1508.04025.

Xu Sun, Bingzhen Wei, Xuancheng Ren, and Shuming Ma. 2017. *Label embedding network: Learning label representation for soft training of deep networks.* CoRR,abs/1710.10393.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. *A neural attention model for abstractive sentence summarization.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 379–389.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. *Global encoding for abstractive summarization.* ACL.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. *Style transfer from non-parallel text by cross-alignment.* CoRR,abs/1705.09655.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. *Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach.* ACL.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. *Multilabel neural networks with applications to functional genomics and text categorization.* IEEE Transactions on Knowledge and Data Engineering, 1338–1351.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016. *Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence.* The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 521–526.

Guibin Chen, Deheng Ye, Zhenchang Xing, Jieshan Chen, and Erik Cambria. 2017. *Ensemble application of convolutional and recurrent neural networks for multi-label text categorization.* 2017 International Joint Conference on Neural Networks, 2377–2383.

WANG Yuan, XU Tao, WANG Shilong, ZHOU Yubo and SHI Yancu. 2021. *An Extreme Multi-label Text Classification Strategy via Hierarchical Label Semantic Guidance.* Journal of Chine Information Processing, 35(10):110–118.

Wei-Cheng Chang, Hsiang-Fu Yu and Kai Zhong. 2020. *Taming pre-trained transformers for extreme multi-label text classification.* Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and DataMining, 3163–3171.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2020. *Sequence to sequence learning with neural networks.* Advances in neural information processing systems, 3104–3112.

Ofir Press and Lior Wolf. 2016. *Using the output embedding to improve language models.* preprint arXiv, 1608.05859.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez and Łukasz Kaiser. 2017. *Attention is all you need.* NIPS.

Nicola De Cao, Gautier Izacard, Sebastian Riedel and Fabio Petroni. 2021. *AUTOREGRESSIVE ENTITY RETRIEVAL.* ICLR 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. *Dropout: A simple way to prevent neural networks from overfitting.* Journal of Machine Learning Research, 15(56):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. *Rethinking the inception architecture for computer vision.* Proceedings of the IEEE conference on computer vision and pattern recognition, 2818–2826.

(c) 2022 中国中文信息学会计算语言学专业委员会

Zhixing Tan, Jiacheng Zhang, Xuancheng Huang, Gang Chen, Shuo Wang, Maosong Sun, Huanbo Luan and Yang Liu. 2020. *THUMT: An Open Source Toolkit for Neural Machine Translation.* AMTA 2020.

Christopher D Manning, Prabhakar Raghavan and Hinrich Schutze. 2008. *Introduction to information retrieval.* volume 1. Cambridge university press Cambridge.

Zhihong Shen, Hao Ma and Kuansan Wang. 2018. *A Web-scale system for scientific knowledge exploration.* ACL 2018.

Enrique Amigo and Augstin D. Delgado. 2022. *Evaluating Extreme Hierarchical Multi-label Classification.* ACL 2022,5809-5819.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK.

CCL 2022

基于多源知识融合的领域情感词典表示学习研究

祁瑞华[1,2] **魏佳**[1] **邵震**[1] **郭旭**[1,2] **陈恒**[1,2]†
1.大连外国语大学软件学院/ 辽宁省大连市
2.大连外国语大学语言智能研究中心/ 辽宁省大连市
rhqi@dlufl.edu.cn WeiJ0417@163.com JKL4131@126.com
guoxu@dlufl.edu.cn chenheng@dlufl.edu.cn

摘要

本文旨在解决领域情感词典构建任务中标注数据资源相对匮乏以及情感语义表示不充分问题，通过多源数据领域差异计算联合权重，融合先验情感知识和Fasttext词向量表示学习，将情感语义知识映射到新的词向量空间，从无标注数据中自动构建适应大数据多领域和多语言环境的领域情感词典。在中英文多领域公开数据集上的对比实验表明，与情感词典方法和预训练词向量方法相比，本文提出的多源知识融合的领域情感词典表示学习方法在实验数据集上的分类正确率均有明显提升，并在多种算法、多语言、多领域和多数据集上具有较好的鲁棒性。本文还通过消融实验验证了所提出模型的各个模块在提升情感分类效果中的作用。

关键词： 知识融合 ； 领域情感词典 ； 表示学习

Domain Sentiment Lexicon Representation Learning Based on Multi-source Knowledge Fusion

Ruihua Qi[1,2] **Jia Wei**[1] **Zhen Shao**[1] **Xu Guo**[1,2] **Heng Chen**[1,2]
1.School of Software Engineering of Dalian University of Foreign Languages / Dalian, Liaoning
2.Research Center for Language Intelligence of Dalian University of Foreign Languages / Dalian, Liaoning
rhqi@dlufl.edu.cn WeiJ0417@163.com JKL4131@126.com
guoxu@dlufl.edu.cn chenheng@dlufl.edu.cn

Abstract

This paper is aiming at the problems of lack of annotated data and inadequate sentiment semantic representation in existing domain sentiment lexicon construction methods. In this paper, the joint weight is calculated by multi-source data. Combining prior emotional knowledge and Fasttext word vector representation learning, the sentiment semantic knowledge is mapped to a new word vector space, and the domain sentiment dictionary is automatically constructed from unlabeled data to adapt to the multi-domain and multi-language environment. The comparative experiments on Chinese and English multi-domain public data sets show that, compared with sentiment dictionary and pretrained language model, the proposed multi-source knowledge fusion method of domain sentiment dictionary representation learning has significantly improved the classification accuracy on public data sets, and has good robustness on various algorithms, multi-language, multi-domain and multi-data sets. This paper also verifies the role of each module of the proposed model in improving the effect of sentiment classification through ablation experiments.

Keywords: knowledge fusion , domain sentiment lexicon , representation learning

第二十一届中国计算语言学大会论文集，第684页-第693页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

情感词汇是文本情感表达的主要途径，由情感词汇构成的情感词典能够明显提升情感分析效果的同时具有很好的可解释性，是社交网络情感分析、商品评论观点挖掘等系统中的重要技术手段，已经成为是无监督情感分析的主要依据[1]。当前大数据多语言环境下，情感分析任务主要面临两个挑战：一是网络文本情感词汇语义内涵变化快、表达方式微妙，难以准确捕捉情感倾向；二是情感分析方法具有领域依赖性，在面向特定领域情感分析任务中，通用情感词典起到一定作用，但通用情感词典无法准确判断新词和领域特有情感词，覆盖率和极性判断准确率也难以满足领域变化各异的情感分析需求，通用情感词典或某个领域的情感词典应用于另一个领域时情感分析性能往往下降明显，新兴领域虽然有海量数据但缺乏先验情感知识的指导，因此迫切需求领域情感词典的自动构建方法。

除了情感词典，目前情感知识的来源主要包括领域内规模有限的有标注数据和无标注数据，此外，大量的领域外数据也隐含着对情感知识的有益的情感信息。为充分利用领域内及领域外的有标注和无标注数据中的情感知识，本文提出基于多源知识融合的领域情感词典表示学习方法，融合多源数据语义信息和情感信息弥补先验知识的不足，从无标注数据中抽取情感信息，结合领域情感知识对比方法自动构建适应大数据多领域、多语言环境的领域情感词典。

2 情感词典研究现状

2.1 语义扩展法

语义扩展法基于专家标注的情感知识库，首先人工选定少量的种子词，在情感知识库中查找每个种子词的同义词、反义词等词间关系进行扩展，经过多轮迭代生成新的情感词典。如Westgate等[2]从Thesaurus.com和WordNet语义知识库递归获取单词同义词构成词的同情感极性图，然后对优化路径中词汇的极性值加权平均决定目标词的极性。SAĞLAM等[3]基于同义词反义词数据集构建的词汇图扩展了土耳其语情感词典。Shaukat等[4]利用Vadar和Senticnet情感词典构成领域情感词典，但每个领域只有5至30个情感词汇。语义扩展方法依赖于人工标注的情感词典，一般规模较小，难以适应词义变化和网络新词的出现，通常作为辅助方法。

2.2 词频共现法

词频共现法包括词频法和词共现法，词频法计算词汇频率筛选情感词，如贺飞艳等[5]结合TF-ID和方差统计提出面向微博短文本的情感特征抽取的计算方法。词频共现法假设共现频率越高的词其语义关联越紧密，如Turney等[6]通过点互信息PMI计算候选词与情感种子词的距离，识别候选词的情感倾向。Mullen 等[7] 过PMI计算形容词的情感倾向值，Liu等[8]针对中文情感词典覆盖率低的问题，通过CHI卡方检验与改进的SO-PMI算法关联计算发现新的情感词。词频共现法单纯地依赖词共现统计信息无法有效表示自然语言的复杂语义，人工选择种子词也增加了不确定性。词频共现法的局限在于构建的词表规模太大导致效率低，同时没有充分利用文本的语义信息。

2.3 启发规则法

启发规则法主要通过观察总结自然语言的语法规则和语言学模式建立情感词典，语法规则如连词规则、否定词规则、双向传播规则以及人工定义的其他规则。如Qiu等[9]提出双向传播算法，定义了四类句法依存关系规则通过迭代路径抽取情感词和目标词，Wu [10]加入一致性连词和否定连词等语法规则改进了双向传播算法情感词极性检测。Hutto 等[11]提出基于简单规则的情感分析模型，使用群智方法人工打分选出情感特征集。启发规则法的局限在于需要专家参与人工定义规则，无法概括日新月异的语言现象，通常与其它方法结合应用。

2.4 词向量表示学习

词向量表示学习方面，Li等[12]面向旅游评论领域通过Word2Vec计算候选词与种子词的语义相似度，并用Interior Point Algorithm 内点算法计算候选词的情感值。杨小平等[13]基于Word2Vec算法提出转换约束集多维情感词典构建方法和基于词分布密度的情感类别及强度计

©2022 中国计算语言学大会 根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：大连外国语大学研究创新团队“计算语言学与人工智能创新团队”(2016CXTD06)；大连外国语大学科研基金项目(2021XJYB16)

算和消歧方法。张璞等[14]选择与种子词具有连词关系的词语作为候选情感词，基于种子词和候选情感词之间的Word2Vec词向量相似度构建语义关联图，使用标签传播算法计算情感词的极性构建情感词典，局限在于种子集基于人工选择，增加了成本和不确定性，例如真正的情感词未必与种子词通过连词连接。方法集成方面，Li等[12]面向旅游领域利用集合互信息AMI发现领域新词，结合人工情感评分值、Wordvec词向量与种子词的语义相似度和PMI相似度构建领域情感词典，改善了情感词典构建，但过程中需要人工参与情感词评分过程。蒋翠清等[15]面向社交媒体中的汽车评论，分别利用PMI和Word2Vec 算法识别新词情感极性，根据集成规则对二者识别结果综合判定构建领域情感词典。现有词向量情感词典将语义信息看作情感信息，存在着局限。

3 基于多源知识融合的领域情感词典表示学习

本文提出基于无标注数据的多源知识融合领域情感词典表示学习方法（Multi-source knowledge Fusion based Domain Sentiment Lexicon representation Learning，MFDSL），表示学习框架如图1所示，主要分为四个模块：多源数据融合领域差异联合权重计算模块、情感知识融合模块、Fasttext表示学习模块和情感词典表示学习模块。

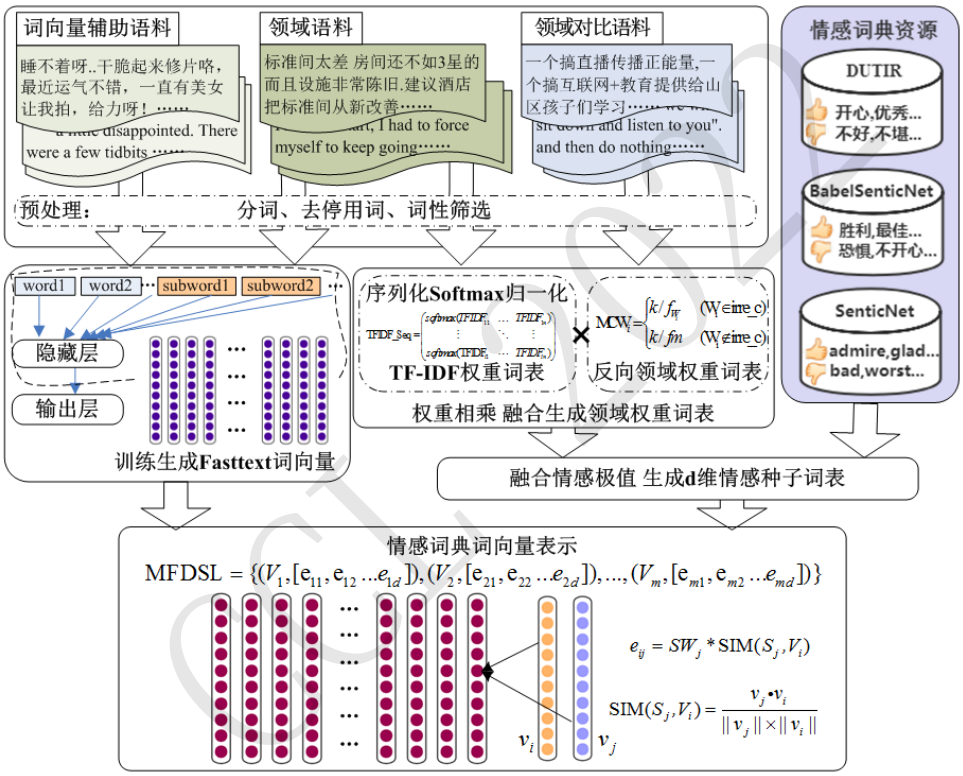


Figure 1: 基于多源知识融合的领域情感词典表示学习框架

3.1 基于多源数据和领域差异权重的情感种子词选取

3.1.1 领域差异联合权重计算

领域情感词提取的首要问题是词汇权重的计算，现有研究大多根据领域语料资源与已有情感知识库结合产生情感词表，计算权重也依赖于领域语料，本研究提出领域差异联合权重计算，引入非领域信息增强领域专有情感词的权重。首先，改进TF-IDF算法作为基础权重计算方法，赋予在少量领域样本中频率高的但总词频不高的词汇比较高的权重。TF权重采用Log标准化，IDF权重采用如(1)所示的逆向文档频率平滑计算方法，其中d为每一条样本，是词t在样本d中出现的频率，N为所有样本数，nt是出现词t的样本数。为避免出现TF-IDF数值过大的情况，对每个样本的TF-IDF值进行Softmax归一化。

(c) 2022 中国中文信息学会计算语言学专业委员会

$$W_{tf-idf} = log(1 + f_{t,d}) * log(\frac{N}{1 + n_t}) + 1 \quad (1)$$

情感词在不同领域语料的分布差异是发现低频领域情感词的重要线索之一。为找出领域中特有的情感词, 本研究基于统计学计算领域差异联合权重，引入外部的领域对比语料强化领域相关词的权重，降低常用的领域无关词的权重。例如在餐厅评论中，“美味可口”的领域优先级应当高于“好”。利用领域对比语料的联合权重计算思路是：计算所有本领域语料词汇在领域对比语料中的词频，以平均词频为基准设置词频和权重的反比关系。设代表每个词汇的词频，k为所有词汇在领域对比语料中的均值，为防止函数分母为0，赋予词汇一个足够小的词频值为该词汇未在领域对比语料中出现时的缺省值，如公式2所示：

$$MCW_i = \begin{cases} k/f_{w_i} & (W_i \in irre\_c) \\ k/f_m & (W_i \notin irre\_c) \end{cases} \quad (2)$$

3.1.2 融合情感知识选取种子词

对于3.1.1节生成的候选词汇联合权重，融合多个情感知识库中情感极性加权求和，再与种子词权重值相乘赋予权重情感极性值，然后通过阈值筛选情感种子词，生成情感种子词表及其对应的权重。中文情感常识知识库采用大连理工大学信息检索研究室的情感词汇本体和Babel SenticNet，英文采用Senticnet 6，多语种情感常识知识库采用Babel SenticNet。大连理工大学信息检索研究室的情感词汇本体包括7大类20小类共27466条中文情感词汇，词典中每个情感词都标注了正向、负向、中性情感极性和情感强度。SenticNet由美国麻省理工学院媒体实验室、斯特灵大学和Sitekit Solutions公司合作构建，目前由Sentic项目组和来自于新加坡南洋理工大学等多家研究机构多领域的专家学者维护，其中SenticNet 6提供了语义和情感关联的20万个英文概念级常识库，标注了情感极性和四个情感维度的情感值，Babel SenticNet是基于SenticNet借助统计翻译方法建立的40种语言的常识知识库。本研究选取上述情感知识库中正向和负向的情感词汇。

3.2 基于Fasttext的词向量表示

词向量保留的语义信息不等同于情感信息, 不同情感极性的词语在语义上可能高度相似, 例如“不错”与“不差”在词向量中有较大相似度，但这两个词的情感极性完全相反，导致词语的情感极性分类不准确。因此本文结合情感知识库和词向量的情感知识和语义信息, 结合多源数据和情感知识库自动构建情感词典，借助深度学习进行词嵌入表示,通过表示学习将Fasttext词向量和情感权重映射到新的情感语义空间，更准确地表示情感语义。

本文采用适应大数据多语言环境的Fasttext词向量进行词汇的表示学习，Fasttext词向量表示学习原理核心思想为[16]：引入子词信息丰富词汇形态学表征信息，将整篇文档的词及n-gram向量叠加平均得到文档向量，使得生僻复杂的单词也能从结构相似的其他单词获得较好的词向量表示。Fasttext突破了土耳其语、芬兰语等形态丰富语种的预训练瓶颈，支持157种语言。Fasttext预训练模型如图1所示，输入层特征向量包括词序列中的所有词、子词和n-gram，并对各个词向量进行加和平均线性变换映射到隐藏层，然后在输出层通过层次softmax函数遍历分类树的叶节点寻找最大概率的分类标签，从而提高了词向量训练速度更适合大规模数据训练。

3.3 基于多源知识融合的领域情感词典学习算法

输入：领域语料re_c、词向量辅助训练语料ft_c、非领域语料irre_c、融合情感词典sl。**输出**：情感词典表示词向量MFDSL。

步骤1：训练词向量，将领域语料与re_c与词向量辅助训练语料ft_c合并，进行分词以及去停用词处理，使用Fasttext预训练得到语义词向量；

步骤2：对领域语料re_c进行分词以及停用词和词性筛选处理，去除助词、标点符号、非语素字、介词、量词、数词、名词、动词，并进行序列化处理。

步骤3：计算各词的TF-IDF值，并对每个句子序列进行softmax处理。得到TFIDF矩阵TFIDF_Seq。然后根据TFIDF_Seq求出每个词的TF-IDF权重，得到TF-IDF值词表，其中m为语料中的词形(Type)数量。

(c) 2022 中国中文信息学会计算语言学专业委员会

步骤4：对非领域语料irre_c进行分词以及去停用词操作，利用公式计算得到每个词的权重MCW，其中代表的词频，参数k为公式的权重值可进行调整，默认情况下取所有情感词在非领域语料中的均值，为当不存在于irre_c中时词频的缺省值，默认情况取值为0.5。最终生成多语料权重词表。

步骤5：将TFIDF_L与MCW_L中相同词的权重相乘，构成情感候选词的融合权重词表，其中权重weight的计算公式为。

步骤6：根据sl中的情感极性，与Weight_L相结合生成维度为d的情感种子词表Seeds。首先求得各情感词的情感权重值，其中p代表sl中的情感词极值，计算方法为各情感词典的极值加权之后求和。然后将词表按照倒序排序，选取正权重值前d/2个词，负权重绝对值前d/2个词，最终得到种子词表;

步骤7：生成情感词典词向量表示，其中每个情感候选词的维度为d，表达式为，其中参数e的计算方法为，SIM函数的计算方法为：利用训练好的fasttext词向量分别得到情感词与种子词的向量表示与，然后利用与计算情感词与种子词之间的相似度。

4 实验结果及分析

情感词典是无监督情感分类任务的主要依据，因此可以通过情感词典在情感分类任务中的效果来间接评估情感词典的有效性[1]，本实验的对照实验包括：表示学习维度对照实验、中文领域情感词向量对照实验和英文领域情感词向量对照实验，同时采用不同领域语料测试本文方法对多语种和多领域的适应性。

4.1 实验数据

本文实验中的中文领域语料来源于谭松波的酒店评论公开数据集[17]，其中正向与负向情感领域语料各2000条，实验选取正负向各1000条数据作为训练数据，正负向各500条作为验证集，剩余的正负向各500条作为测试数据。中文词向量辅助语料采用NLPIR微博内容语料库中新浪微博和腾讯微博评论23万条[18]，以及谭松波整理的1万条酒店评论语料，合计24万条。对这24万条数据进行分词以及去除停用词处理，作为Fasttext词向量的训练语料。中文领域对比语料来源于SMP-EWCT2020的评测数据[19]，包含微博评论共46421条。英文领域语料采用Amazon公开评论数据集[20]，覆盖图书、DVD、电子产品、厨房用品和影像五个领域，每个领域选取标注语料6000条，实验数据随机选取各领域正负向各1000条数据，选取其中50%作为训练集。英文词向量训练辅助语料采用Blitzer收集整理Amazon评论中的无标注数据共80821条[20]，英文领域对比语料来自于纽约时报新闻评论的公开数据共49868条[21]。

Table 1: 情感词典词向量构建采用的语料

语料名称	样本总数	正向样本	负向样本	无标注样本
中文领域语料ChnSentiCorpHtlba4000	4000	2000	2000	0
中文词向量辅助语料ChnSentiCorpHtluba10000	10000	0	0	10000
中文词向量辅助语料	230000	0	0	230000
中文领域对比语料	46421	0	0	46421
英文领域语料Amazon reviews(books)	6000	3000	3000	0
英文领域语料Amazon reviews(dvd)	6000	3000	3000	0
英文领域语料Amazon reviews(electronics)	6000	3000	3000	0
英文领域语料Amazon reviews(kitchen)	6000	3000	3000	0
英文领域语料Amazon reviews(video)	6000	3000	3000	0
英文词向量辅助语料Amazon reviews	80821	0	0	80821
英文领域对比语料	49868	0	0	49868

4.2 实验参数

预处理模块中，中文语料采用结巴分词的paddle模式处理，去除助词、标点符号、非语素字、介词、量词、数词和叹词，采用哈工大的中文停用词表。英文语料通过Spacy筛选词

(c) 2022 中国中文信息学会计算语言学专业委员会

性，采用Spacy模块中的英文停用词表。中英文情感词向量对照实验中，选取五种对照方法与本文方法对比，分别为：(1)情感本体方法，情感词典由相应的情感知识库采用One-hot编码构成，句子向量的维度为情感本体中所有词语的个数，编码值为情感强度；(2)Word2Vec方法，词向量仅由Word2Vec预训练算法生成，词向量维度为100，迭代次数为30，词的最小出现次数为2，句向量的计算采用词向量求和平均生成，句向量的维度与词向量维度相同；(3)Fasttext方法，词向量只由Fasttext预训练算法生成，实验参数与Word2Vec相同；(4)TFIDFSenti2vec，文献[22]中的基于词向量的情感词典方法；(5)TFIDF方法，通过本文的情感种子词生成模块产生情感词典作为输入，未结合Fasttext词向量；(6) MFDSLSVM，本文提出的多源知识融合领域情感词典表示学习方法，情感分类算法采用SVM，实验平台为Sklearn，采用线性核函数Liner和概率估计；(7) BertBiLSTM，采用预训练语言模型Bert和深度学习分类算法BiLSTM；(8) MFDSLBertBiLSTM，采用本文提出的多源知识融合领域情感词典表示学习方法，结合预训练语言模型Bert和深度学习分类算法BiLSTM。

评价指标选取正负向语料的精度、召回率、F1值和总体准确率检验情感词典对文本情感分类任务的有效性。

4.3 实验结果与分析

4.3.1 表示学习维度对照实验

当采用词向量表示文本时，基本原理上是词向量的维度越大效果越好，但完成具体任务时需要达到运算速度和情感分析效果的平衡，因此进行表示学习维度对照实验，选择能达到较好情感分析效果的情感词典表示维度。对照实验中的TFIDF方法和本文的MFDSL算法在情感词典表示维度分别为20维、50维、100维、120维、150维、200维和300维时，在中文实验语料上的情感分类十折交叉验证实验的精度、召回率、F1值和准确率如表2和图2所示，当情感词典表示维度从20维增长到100维，情感分析准确率和各项指标提升明显，而增长到100维之后，准确率提升就比较少并趋于平稳，因此本文选取情感词向量的表示维度为100维。

Table 2: 表示学习维度对照实验

维度	生成方法	macro precision	macro recall	macro f1	accuracy
20	TF-IDF	81.16%	81.07%	81.07%	81.08%
	MFDSL	81.94%	81.90%	81.90%	81.91%
50	TF-IDF	82.68%	82.60%	82.60%	82.61%
	MFDSL	83.95%	83.90%	83.90%	83.91%
100	TF-IDF	83.78%	83.73%	83.72%	83.74%
	MFDSL	84.10%	84.05%	84.05%	84.06%
120	TF-IDF	83.86%	83.80%	83.80%	83.82%
	MFDSL	84.11%	84.06%	84.06%	84.07%
150	TF-IDF	84.09%	84.03%	84.04%	84.05%
	MFDSL	84.18%	84.13%	84.13%	84.14%
200	TF-IDF	84.13%	84.07%	84.08%	84.09%
	MFDSL	84.20%	84.15%	84.15%	84.16%
300	TF-IDF	84.12%	84.07%	84.07%	84.08%
	MFDSL	84.22%	84.17%	84.17%	84.18%

4.3.2 中文领域情感词典对照实验

为验证本文提出的情感词典构建方法，将生成的情感词向量MFDSL与4.2中的情感本体方法、Word2Vec方法、Fasttext方法、TF-IDF方法以及文献[22]中的TF-IDF-Senti2vec方法在中文酒店领域评论上做情感分类实验，设定领域情感词向量MFDSL维度为100维，十折交叉验证结果如表3所示：

从表3可以看出，本文提出的多源知识融合领域情感词典表示学习方法MFDSL在中文领

(c) 2022 中国中文信息学会计算语言学专业委员会

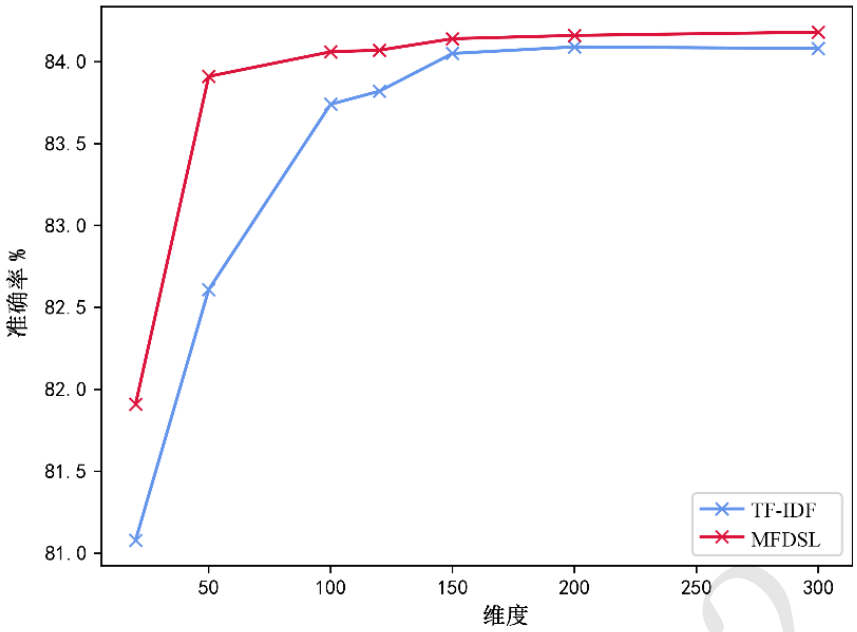


Figure 2: 表示学习维度-准确率变化关系图

Table 3: 中文情感分类对照实验结果

对照方法	macro precision	macro recall	macro f1	accuracy
情感本体-SVM	76.46%	75.21%	74.94%	75.25%
Word2vec-SVM	83.24%	83.21%	83.22%	83.24%
Fasttext-SVM	83.57%	83.55%	83.54%	83.55%
TF-IDF-Senti2vec-SVM	–	–	–	62.97%
TF-IDF-SVM	83.18%	83.10%	83.09%	83.10%
MFDSL-SVM	84.05%	83.99%	83.98%	83.99%
Bert-BiLSTM	89.15%	88.23%	87.96%	88.07%
MFDSL-Bert-BiLSTM	90.94%	90.84%	90.80%	90.82%

域情感词典应用于情感分类的对照实验中取得了比情感本体、词向量方法和TFIDF类方法更高的准确率、召回率、F1值和正确率。其中，单纯采用词向量的方法正确率高于情感本体方法，证明了词向量在语义表示上的优势。文献[22]通过TFIDF和Word2vec结合表示情感词典，效果并不理想，采用SVM分类算法时本文模型正确率比文献[22]方法提高了20.13%，证明了领域对比方法的有效性。采用SVM分类算法时，本文MFDSL结合领域对比方法和Fasttext表示学习后，比较Fasttext方法正确率提高了0.44%，比Word2vec词向量提高了0.75%，比TF-IDF方法提高了0.89%，比情感本体提高了8.74%。在与预训练语言模型Bert的对比中，增加了本文提出的MFDSL模型之后，正确率提高了2.75%，进一步证明了多源知识融合表示学习方法的有效性。

为探究领域对比方法的有效性，实验进一步比较了引入领域对比语料前后的情感词表，如表4所示，引入领域对比语料能够抽取低词频但具有领域代表性的情感词汇，正向情感词汇增加了62个，负向情感词汇增加69个，分别占引入领域对比领域前的13%和16%，更好地实现了领域情感词典构建的目标。

4.3.3 英文领域情感词典对照实验

为验证本文方法在多领域、多语言环境中的有效性，本节选取Blitzer收集整理Amazon图

Table 4: 引入中文领域对比语料前后的情感词表

	极性	数量	示例
引入领域对比语料前	正向	479	...整洁，豪华，宽敞，价廉物美，诚挚...
	负向	427	...最差，不足之处，简陋，大失所望，美中不足...
引入领域对比语料后	正向	541	...宾至如归，方便，便利，便宜，没得说...
	负向	496	...轰鸣声，坑坑洼洼，偏僻，形同虚设，置若罔闻...

书、DVD、电子产品、厨房用品和影像五个领域的英文评论语料，将本文MFDSL方法与情感本体、Word2Vec、Fasttext、TF-IDF和Bert预训练语言模型情感分类对照实验，设定领域情感词向量MFDSL维度为100维，十折交叉验证的平均正确率如表5所示，在五个领域的英文语料上，本文方法均取得了最高的正确率，比Bert预训练语言模型在五个领域分别提升了1.21%,2.2%,0.78%,1.94%和5.51%，验证了其在多语言、多领域环境中的鲁棒性。此外，实验结果表明，对比本文方法与Fasttext、Word2vec词向量在英文数据集上的正确率提升，比在中文数据集上的提升更为明显，原因是中文词向量辅助语料规模更大，并从微博评论中获得了更通用的语义知识。值得注意的是，本文提出的MFDSL领域情感词向量表示方法，需要的标注数据规模小，不仅适用于深度学习模型BiLSTM，还适用于时间复杂度较低的SVM算法，在不同的算法上也具有较好的鲁棒性。

Table 5: 英文情感分类对照实验结果

对照方法accuracy	图书	DVD	电子产品	厨房用品	影像
情感本体-SVM	70.07%	71.00%	69.80%	71.60%	75.60%
Word2vec-SVM	74.71%	76.80%	75.26%	76.70%	76.90%
Fasttext-SVM	74.81%	75.53%	74.19%	76.25%	77.26%
TF-IDF-SVM	73.42%	74.88%	74.28%	75.21%	77.12%
MFDSL-SVM	75.22%	77.57%	77.80%	78.29%	77.82%
Bert-BiLSTM	76.23%	75.74%	77.45%	79.84%	76.23%
MFDSL-Bert-BiLSTM	77.44%	77.94%	78.23%	81.78%	81.74%

为进一步探究领域对比方法的在英文语料上有效性，选取英文图书评论领域比较引入领域对比语料前后的情感词表，如表6所示。可以看出，引入领域对比语料后能够抽取诸如“gastronomic”、“machiavellian”的低频英文情感词，正向情感词汇增加了75个，负向情感词汇增加80个，并有效改进了情感分类效果。

Table 6: 引入英文图书领域对比语料前后的情感词表

	极性	数量	示例
引入领域对比语料前	正向	1804	...great，excellent，wonderful，new，easy...
	负向	1842	...bad，boring，worst，disappointing，confusing...
引入领域对比语料后	正向	1879	...gastronomic，decorative，suggestive，impressively，unforgettably...
	负向	1922	...machiavellian，sissy，unitarian，regretful，musty...

4.3.4 消融实验

本文基于多源知识融合的领域情感词典表示学习模型在词向量的基础上，主要包括领域对比模块、Tfidf模块和情感本体模块。为验证本文模型各个模块的作用，分别进行了中英文语料上的消融实验，如表7和表8所示：

从表7和表8可以看出，总体上在中英文各领域数据集上，仅采用领域对比模块、Tfidf模块和情感本体模块都比整体MFDSL模型的正确率有所下降，当三个模块都移除，仅采用Bert预训

(c) 2022 中国中文信息学会计算语言学专业委员会

Table 7: 中文情感分类消融实验

对照方法	macro precision	macro recall	macro f1	accuracy
MFDSL-Bert-BiLSTM	90.94%	90.84%	90.80%	90.82%
领域对比-Bert-BiLSTM	90.49%	90.33%	90.21%	90.22%
tfidf-BERT-BiLSTM	90.44%	90.41%	90.33%	90.34%
情感本体-BERT-BiLSTM	90.28%	89.93%	89.79%	89.82%
Bert-BiLSTM	89.15%	88.23%	87.96%	88.07%

Table 8: 英文情感分类消融实验结果

对照方法accuracy	图书	DVD	电子产品	厨房用品	影像
MFDSL-Bert-BiLSTM	77.44%	77.94%	78.23%	81.78%	81.74%
领域对比-Bert-BiLSTM	76.64%	75.82%	77.26%	79.03%	80.63%
tfidf-BERT-BiLSTM	75.91%	75.49%	77.23%	79.56%	79.54%
情感本体-BERT-BiLSTM	75.71%	75.20%	76.69%	79.79%	80.62%
Bert-BiLSTM	76.23%	75.74%	77.45%	79.84%	76.23%

练语言模型和BiLSTM分类算法时，正确率降到最低，证实了本文模型中三个模块的有效性。但从实验结果没有明显趋势表示具体哪一个模块在整体性能中贡献最大，正确率主要取决于各个模块之间的交互效果。

5 总结与展望

本文提出基于多源知识融合的领域情感词典表示学习方法自动从无标注数据中构建适应大数据多领域和多语言环境的领域情感词典，引入外部领域对比语料强化领域相关词的权重，融合多源数据语义信息和情感信息弥补先验知识的不足，通过表示学习将词向量和情感权重映射到新的情感语义空间更准确地表示情感语义。在中英文六个领域公开数据集上的对照实验结果表明该模型有效提高了情感词典在情感分类中的有效性，进一步的分析验证了本文方法能够有效抽取其他方法难以自动提取的低频领域情感词。在未来的工作中，将在多领域大规模语料中进一步检验模型的泛化性，进一步探究隐性情感词汇的自动抽取方法和检验标准。

参考文献

王科,夏睿.情感词典自动构建方法综述[J].自动化学报,2016,42(4): 495-511.

Westgate A, Valova I. A Graph Based Approach to Sentiment Lexicon Expansion[C]. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2018: 530-541.

SAĞLAM F, GENÇ B, SEVER H. Extending a sentiment lexicon with synonym–antonym datasets: SWNetTR++[J]. Turkish Journal of Electrical Engineering & Computer Sciences, 2019, 27(3): 1806-1820.

Shaukat K, Hameed I A, Luo S, et al. Domain Specific Lexicon Generation through Sentiment Analysis[J]. International Journal of Emerging Technologies in Learning, 2020, 15(9).

贺飞艳,何炎祥,刘楠,刘健博,彭敏.面向微博短文本的细粒度情感特征抽取方法[J].北京大学学报(自然科学版),2014,50(01):48-54.

Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4): 315-346.

Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources[C].Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 412-418.

(c) 2022 中国中文信息学会计算语言学专业委员会

Liu J, Yan M, Luo J. Research on the construction of sentiment lexicon based on Chinese microblog[C]. In: International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, 2016:56-59.

Qiu G, Liu B, Bu J, et al. Expanding domain sentiment lexicon through double propagation[C].IJCAI. 2009, 9: 1199-1204.

Wu S, Wu F, Chang Y, et al. Automatic construction of target-specific sentiment lexicon[J]. Expert Systems with Applications, 2019, 116: 285-298.

Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text[C].Proceedings of the International AAAI Conference on Web and Social Media. 2014, 8(1).

Li W, Guo K, Shi Y, et al. DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain[J]. Knowledge-Based Systems, 2018, 146: 203-214.

杨小平,张中夏,王良,张永俊,马奇凤,吴佳楠,张悦.基于Word2Vec的情感词典自动构建与优化[J].计算机科学,2017,44(01):42-47+74.

张璞,王俊霞,王英豪.基于标签传播的情感词典构建方法[J].计算机工程,2018,44(05):168-173.

蒋翠清,郭轶博,刘尧.基于中文社交媒体文本的领域情感词典构建方法研究[J].数据分析与知识发现,2019,3(02):98-107.

Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

谭松波. 数据集:谭松波–酒店评论语料[DB/OL].[2020-02-08].https://blog.csdn.net/LiuKingJia/article/details/104228617.

张华平. NLPIR微博内容语料库-23万条[DB/OL].[2017-12-03].http://www.nlpir.org/wordpress/download/weibo_content_

SMP2020-EWECT.SMP2020微博情绪分类评测[DB/OL].[2020-06-19].https://smp2020ewect.github.io/.

Amazon Review Data (2018).Jianmo NiDB/OL].[2022-02-17]. https://nijianmo.github.io/amazon/index.html#complete-data.

Aashita Kesarwani. New York Times Comments.[DB/OL]. [2018]. https://www.kaggle.com/aashita/nyt-comments.

林江豪,周咏梅,阳爱民,陈锦.基于词向量的领域情感词典构建[J].山东大学学报(工学版),2018,48(03):40-47.

CCL 2022

(c) 2022 中国中文信息学会计算语言学专业委员会

俄语网络仇恨言论语料库研究与构建

温昕
西安电子科技大学外国语学院
中国，西安 710071
wenxin@xidian.edu.cn

郑敏娇
国防科技大学信息通信学院
中国，武汉 430010
mjzzheng@126.com

摘要

近年来，网络科技的飞速发展在为整个社会带来极大便利的同时，也加剧了仇恨言论的传播。仇恨言论可能会构成网络暴力，诱发仇恨性的犯罪行为，对社会公共文明和网络空间秩序造成极大的威胁。因此，对网络仇恨言论进行主动的监管和制约具有重大意义。而当前学术界针对俄语的网络仇恨言论研究不足，尤其缺乏俄语网络仇恨言论语料库，这极大地限制了相关技术和应用的发展。2022年俄乌冲突爆发以后，对于俄语网络仇恨言论语料库的研究与构建显得更加迫切。在本文中，作者提出了一种细粒度的俄语网络仇恨言论语料库构建及标注方案，并基于该方案首次创建了包含20476条文本数据，具有针对性、话题统一的俄语仇恨性言论语料库。

关键词： 俄语 ； 语料库 ； 仇恨言论

An Russian Internet Corpus for Hate Speech Detection

Xin Wen
School of Foreign Languages,
Xidian University,
Xi'an, China 700071
wenxin@xidian.edu.cn

Minjiao Zheng
School of Information and Communication,
National University of Defense Technology,
Wuhan, China 430010
mjzzheng@126.com

Abstract

With the rapid development of network society, the spread of hate speech has become increasingly serious, which attracts more and more people's attention. Hateful speech may induce cyber violence, hate crimes, and cause great harm to society. Therefore, it is necessary to supervise online hate speech, and thereby reducing the potential harm. However, there is a lack of relevant research on Russian online hate speech, especially the research on Russian hate speech corpus, which greatly limits the development of the relevant researches. To solve this problem, this paper creates a Russian online hate speech corpus that has a unified topic, and achieves multi-dimensional labeling on the corpus.

Keywords: Russian , Corpus , Hate Speech

1 引言

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十一届中国计算语言学大会论文集，第694页-第703页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

信息爆炸、数据井喷是21世纪社会现状的显著特征。在高速发展的信息流时代，社交网络开始全方位地影响我们的日常生活。全球约有39.6亿人使用社交媒体，占全球77.7亿总人口的50.64%；另有数据表明，有83.36%的互联网用户都在使用社交媒体，网络媒体已逐渐成为人们交流观点、表达情感的最重要途径。看似自由的网络空间实则具有虚拟性、欺骗性、随意性、复杂性，难以监管，网络社交媒体常常成为仇恨性言论的爆发地。

所谓仇恨言论（Hate Speech），是指在仇恨心态的引导下，基于民族、种族、国籍、宗教、性别、年龄和身体状况等身份特征，公开发表，来表达对他人或某些群体的仇恨、欺凌、贬低、侮辱、歧视、威胁或煽动暴力的言论(Erjavec and Kovačič, 2012)，其往往针对于弱势群体进行攻击，涉及到许多社会存在的焦点问题，如种族歧视、性别对立、宗教信仰、性取向等。例如，2014年的乌克兰反政府示威运动（Евромайдан），以及2022年的俄乌冲突爆发以来，俄语网络社交媒体中就产生了大量的仇恨言论。

虽然对于仇恨言论的定义和定义方式各不相同，但多数专家和学者认为，仇恨言论具有极强的负面影响力，不属于言论自由所应当保护的言论，必须受到限制(Василенко, 2019)。仇恨性言论在网络上的传播可能会引发恶劣的结果，诱导人们（尤其是青少年）实施仇恨性犯罪，影响社会安定，危害公共安全，甚至间接导致国际冲突升级。为了制约网络仇恨性言论的产生与传播，各大社交网络平台也制定了相应的规则。例如Twitter、微博等平台都设有专门的内容审查员手动检测和审核包含此类语言现象的评论或帖子。为了提升检测效率，研究自动化、低成本的网络仇恨言论检测方法变得十分必要。而语料库作为语言学研究的基础资源，是研究仇恨言论检测方法的必要条件。当前学术界针对俄语的网络仇恨言论缺乏相关的研究，尤其缺乏俄语网络仇恨言论语料库。因此，研究并建立具有针对性的、话题统一的俄语网络仇恨言论语料库具有重要的理论和实践意义。

针对此问题，本文以乌克兰民族和政治问题为主题，构建了一个俄语仇恨言论语料库，并探讨了语料库的标注方法。本文选取了YouTube这一社交网络平台中的话题：“乌克兰反政府示威运动（Евромайдан）”，对相关视频下方的评论文本进行抓取和收集，以此构建语料库。同时，我们提出了多维度、细粒度的标注方法，从是否含有仇恨言论，有无攻击性、冒犯性、讽刺性、刻板印象、辱骂性等维度及仇恨的强弱程度对评论内容进行标注。此种多维度的标注方法能够使文本内容中的细微差别得到更全面的展现，为仇恨言论的识别和界定提供更精细的参考标准，能够较好地平衡标注的正确性和主观性。最后，对标注结果和方法进行分析与讨论。我们发现，在本语料库所选取的乌克兰民族与政治的主题中，仇恨言论在攻击性和冒犯性两个维度上表现得最为明显，但依然需要考虑刻板印象和讽刺性等隐式煽动的仇恨表达方式。

我们创建的俄语网络仇恨言论语料库包含20476条评论数据，该语料库可作为俄语网络仇恨言论检测技术的测试集和训练集，是相关研究的基础。

2 相关工作

当前在俄语研究范畴内，对于仇恨言论的研究主要是将其作为一种语言现象进行研究，探索仇恨言论在传播理论、多语言环境等语用方面的问题(Василенко, 2019; Хроменков, 2016; Шарнин et al., 2018)，但并未给出关于仇恨言论的定性或定量的描述方法，也没能形成俄语网络仇恨言论相关的权威语料库。在计算语言学和人工智能飞速发展的背景下，语料库缺失这一基础问题极大地限制了俄语相关研究的发展。

目前，对于仇恨言论的语料库研究大多基于英语。Kennedy等(2018)组织语言专家根据编码类型学标记社交网络中的帖子，形成了仇恨语料库GHC（Gab Hate Corpus）；Assimakopoulos等(2020)提出了针对马耳他地中海移民危机、性小众群体（LGBTIQ+）等问题的多层语料库标注方案；Klubička等(2018)强调了现有的仇恨言论检测方式存在的问题，并在语料库的基础上对仇恨言论检测之外的问题进行探索；Huang等(2020)在语料库中标注了文本发布人的年龄、国家、性别和种族/民族等因素，用以分析言论偏差和评论的人口可预测性。

除英语以外，阿拉伯语、意大利语等小语种中也有部分涉及仇恨言论语料库的研究。Alakrot等(2018)基于YouTube数据构建了一个阿拉伯语评论语料库，并将其用于辱骂性语言的检测；Alhuzali与Abdul-Mageed(2018)在Twitter上获取了阿拉伯语的评论内容，构建了语料库和仇恨词典；在意大利语中，针对仇恨言论语料库和仇恨言论检测方法相关的研究有Del Vigna等(2017)；Maisto等(2017)；Bosco等(2018)。

第二十一届中国计算语言学大会论文集，第694页-第703页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

从语料库标注的角度来看，Kwok和Wang(2013)的研究对Twitter文本的攻击性进行了分类，等级为1到5；Ross等(2017)为语料库的标注设计了两个标签，即仇恨言论（是/否）和进攻性，等级从1到6；Del Vigna等(2017)的研究使用了包括无仇恨/弱仇恨/强仇恨三个等级的标签。以上工作表明，简单的二元标签无法满足分析仇恨言论这一复杂的需求，并且可能会影响后续研究和进一步分析的准确性。因此，本文同样采用了非二元标签语料库标注方法，并在此基础上引入了多个注释类别，建立了多维度的语料库标注方案。

从语料库收集的角度看，现有相关研究大多通过仇恨类别或一组典型的仇恨词语来收集语料数据(Del Vigna12 et al., 2017; Sanguinetti et al., 2018)，而Waseem和Hovy(2016)在收集语料时将仇恨词语和经常与仇恨言论一同出现的中性词语结合在一起，可以识别更宽泛的仇恨言论的表达方式。也就是说，通过典型的仇恨词来收集语料库的方法存在片面性的问题。因此，为了解决该问题，本文采用了基于主题的语料数据收集方法，以仇恨言论的高发话题——乌克兰民族和政治问题中的相关网络评论作为语料来源。

总而言之，与上述相关工作相比，本文致力于提出一种更新颖、更细粒度的语料库构建及标注方案。该方案能够用以表示仇恨言论现象的多个方面，因此具有更强的挑战性。

3 语料库的创建

3.1 为什么选择民族政治问题作为语料库的语料来源

由于民族和种族问题是因历史、政治等复杂因素造成的问题，由来已久并且十分尖锐，容易引起言语冲突和言语对抗。敏感的政治问题和暴力冲突事件更容易引发民族和不同派别之间的矛盾。在网络社会中，有关此类事件的评价与讨论大多都是负面的，因此我们在收集语料时将针对乌克兰民族和政治问题的网络评论作为本语料库的语料来源。

我们选取了2013年至2014年发生在乌克兰的“反政府示威运动（Евромайдан）”这一事件为主题，是因为此次运动造成了极大的影响。据统计，为期93天示威造成至少125人死亡，1890多人受伤，65人失踪。由该运动引起的话题和相关视频在网上掀起了轩然大波，引起亲欧派和亲俄派、乌克兰族和俄罗斯族的激烈争论。因此，我们收集了YouTube社交网络平台中与此次运动相关的视频下方的评论文本。

3.2 为什么提出多维度的标注方案

我们将上文提到的评论文本进行整理，获得了20476条评论数据，从中随机选择了1220条评论作为标注子集。我们首先对标注子集进行机器翻译和人工校对。带有翻译内容的数据集分别由语言学专家和以俄语为母语的标注志愿者对其进行标注。标注的具体方案在研究中都将进行详细介绍。我们将对标注的结果进行比照分析，思考标注原则的制定是否合理，分析出现标注偏差的原因，提出日后改进的方向。

我们的标注内容包括是否含有仇恨言论、有无攻击性、冒犯性、讽刺性、刻板印象、辱骂性等维度和仇恨的强弱程度。这样的设计是为了使专家、标注者和我们能够深刻理解和体会仇恨言论的定义和范畴，通过各个含义和角度的交叉验证，帮助我们对文本的仇恨内容进行更加理性的判断。另外，我们建立的语料库话题统一，在内容上具有很强的针对性，能够为仇恨言论自动检测系统的训练提供更加契合的基础支持，并为其他相关研究奠定语料库基础。

4 标注方案

仇恨言论的鉴别是一项非常具有挑战性的任务，不可避免地会受到判断者个体认知和主观思维的影响(Waseem and Hovy, 2016; Ross et al., 2017)。Weber(2009)借助欧洲人权法院的运作方式来说明鉴别仇恨言论的困难之处，并且强调，在划定言论的合法性、非法性的界限时，并没有单一的标准，而是需要借助一组综合性的考量标准，并且进行逐案分析。因此，我们尝试不仅仅鉴别言论中是否存在仇恨性，还参考其他的辅助参数来标注每一条评论以减少标注者的犹豫和不确定，使其能够对每一条数据做出更理性、更准确的判断。

我们制定了一套标注指南，尝试将所有标注项包含在一个统一的框架之中。我们设计的标注项目除了仇恨言论（HS，是/否）以外，还包括攻击性（aggressiveness，强/弱/无）、冒犯性（offensiveness，强/弱/无）、讽刺性（irony，有/无）、刻板印象（stereotype，有/无）和辱骂性（abusive，有/无）。

(c) 2022 中国中文信息学会计算语言学专业委员会

考虑到此注释任务需要面对标注主体（人类）本身的复杂性，以及该任务固有的复杂性，我们分别邀请了两组人员对同一数据进行了标注。我们将由专家组成的A组标注结果和由俄语母语人组成的B组标注结果进行对比，发现两组标注结果也存在一些分歧，根据这些结果，我们深入讨论了发生分歧的可能的原因。

4.1 标注项目

以下是我们对标注的项目所给出的定义和示例：

仇恨言论（HS，是/否）——对于评论是否包含仇恨言论的判定，主要从两个元素上辨别：

- **目标**——仇恨言论的目标需是一定的群体（俄罗斯人/乌克兰人/少数民族/亲俄派/亲欧派）或者是作为这些群体成员的个人。
- **言语意图**——话语的言外之意(Searle and Searle, 1969)。我们需要了解言语的隐含意图，捕捉言语中对目标的渗透、煽动、刺激以及仇恨或暴力的迹象，或是言语中使目标群体失去人性化，丧失合法性，对目标进行伤害或恐吓。

这两个元素在评论数据中同时存在对于确定评论中是否包含仇恨言论至关重要，例如：отстреливайте хохлов да поскорее да боевыми да наглухо валите их（快发动军队向乌克兰人开枪，狠狠地击垮他们）。如果在评论中不包括任何一个元素，那么就认为这条评论中不包含仇恨言论。下面我们对其余的标注类别分别作简要说明：

- **攻击性（aggressiveness，强/弱/无）**——关注用户的攻击、侵略、伤害等直接行为，甚至是煽动对既定目标实施各种形式的暴力行为。如果存在这种意图，还可以对其程度进行强/弱的划分。例如，言论中暗示歧视或将歧视合法化的态度会被视为是攻击性较弱的言论：по русски пожалуйста（请说俄语），这里指对乌克兰人说俄语规范性的一种轻视和瞧不起；而如果提及暴力行为，无论是明示还是暗示，都被认为具有强烈的攻击性：хороший мусор мертвый мусор вырезать их семьи отрезать головы и в футбол играть（好垃圾死垃圾砍了他们的家人砍掉他们的头踢足球）。
- **冒犯性（offensiveness，强/弱/无）**——与攻击性相反，冒犯性关注的是评论内容对给定目标的间接伤害和潜在的影响，对冒犯性的程度也可以进行强/弱的划分。例如，认为给定目标具有典型的人的缺陷时，则评论具有弱冒犯性：наталия вы глупы половина украинцев говорит на русском но при этом русскими не становятся（娜塔莉亚您真是愚蠢，半数乌克兰人都说俄语，但他们也不是俄罗斯人）；而如果目标面临粗暴或侮辱性的表达时，则该评论内容将被标记为具有强冒犯性：ну ты и дура как я посмотрю（好吧，在我看来你就是个傻瓜）。
- **讽刺性（irony，有/无）**——该术语作为通用术语，涵盖了讽刺、幽默、反讽等讽刺意味的细微差别（Bosco等，2013：4159）。在语料库中，对讽刺性的衡量是二元的，即有/无。在数据标注中引入讽刺性源于对数据的初步观察。我们在评论数据中看到了讽刺，尤其是反讽对于仇恨性的削弱或间接表达仇恨性内容的现象非常普遍，尤其是俄罗斯人的语言风格和言语习惯，他们会经常引入一些幽默风趣的表达，所以我们引入讽刺性作为一个仇恨性内容判断的参考项。例如：федорову нужно писать сценарии для голливуда（费多罗夫需要为好莱坞写剧本）；по мне так всё логично эти мирные демонстранты сожгли много зданий взяли в заложники людей захватывают административные здания и не идут на мир думаю многие заслужили пулю от снайпера в лоб（对我来说，一切都如此合乎逻辑，这些和平示威者烧毁了许多建筑物，劫持了许多人质，攻占了行政大楼，就是不要和平，我觉得他们中的很多人都该被枪毙）。
- **刻板印象（stereotype，有/无）**——判断评论内容中是否含有对特定目标明显的或隐含的一种概括或固定的看法，即认为整个目标群体都具有该特征，从而忽视个体差异。在本研究中，刻板印象主要表现为对目标群体的偏见。我们通过对数据的观察发现，对少数群体的仇恨也常常以偏见为特征。在我们的语料库中，对刻板印象的衡量也是以二元的有/无

(c) 2022 中国中文信息学会计算语言学专业委员会

作为判断标准的。例如：прикол в том что украине не предлагают шенген там своих нахлебников хватает（搞笑的是乌克兰不给申根签证，因为那里的寄生虫已经够多了）。

- **辱骂性（abusive，有/无）**——主要指语言中的脏话，禁忌语。由于网络环境相对自由，所以在网络言论中辱骂性非常普遍。仇恨言论在多数情况下表现为辱骂性言论，例如：пшел от сюда дичь（滚吧，畜生）。

4.2 标注程度

在对标注后的数据结果进行更详细的观察时，我们发现被标注为仇恨言论的数据在强度和危害程度上都有着较大的差异，所以仅仅区分评论内容中是否含有仇恨言论是不能精确地反映其仇恨性的。因此，我们引入了对仇恨程度的考量，即用"煽动强度"这个概念来解释不同类型的仇恨言论，甚至是言语中的暴力行为。我们发现，这种程度上的界定对于我们正确理解和深度挖掘评论的含义有着非常大的帮助。我们将仇恨言论的煽动性定义为5个等级，其值分别为0-4级，如果不含有仇恨言论，则值为0。

- **0 级**：完全没有煽动性，文本的内容有些矛盾，尽管评论文本可能被注释为具有攻击性、冒犯性或其他形式，但并不包含仇恨言论：насилие это классно вы че сука（暴力很酷你这个婊子）。

- **1 级**：没有明确的煽动性，但说话人将某些不良特征或品质归因于目标群体。有时，他们认为这些负面特征可能会对说话人和读者构成威胁，这些评论更类似于基于刻板印象的侮辱或判断：они только скакать жрать срать спать могут и им этого вполне хватает для счастья （他们只会跳，吃，拉，睡，有了这些他们就足够幸福了）。

- **2 级**：没有明确的煽动，但是文本中表达的行为旨在使目标群体失去人性化，丧失合法性，或者声称给予他们的基本权利是不公正的特权，亦或是声称目标群体的这些权利会伤害说话人和读者，因此不应再被授予。这些言语行为不是呼吁暴力，但会引起对目标群体的厌恶或仇恨：какая мразь не дай бог придут к власти в россии опять реки крови потеря курил мы и так теряем наших братьев на украине каждый день каждый час заявляю официально хакамада враг снг кто ее поддерживает враг эта гадость не умеющая одеваца откровенная дура и демогог（什么渣滓，上帝保佑，千万别再让俄罗斯掌权了，我们经历着血流成河，在乌克兰我们每天、每个小时都在失去我们的兄弟，我正式宣布哈卡马达（Ирина Муцуовна Хакамада）是独联体的敌人，支持她的人是彻头彻尾的傻瓜、煽动者）。

- **3 级**：明确煽动暴力或歧视性行为，但说话人拒绝承担这些行为的责任，只为这些行为辩护或表达希望发生这种行为的意愿：понаражали долбаёбов майдановцев срочняков которые свою страну защищают просто убивают твари вы позорные надеюсь сдохните скоро（Maidan示威者，这些蠢兵，他们靠杀人保卫自己的国家，可耻畜生，我希望你们快点死）。

- **4 级**：明显煽动暴力或歧视行为；评论者公开提议或呼吁采取这些行动，并宣称自己已准备好执行这些行为，或参与实现这些行动：отстреливайте хохлов да поскорее да боевыми да наглухо валите их（快发动军队向乌克兰人开枪，狠狠地击垮他们）；я конечно не провокатор но я бы на месте беркута стрелял боевыми дабы пресечь этот беспредел（我不是挑衅者，但为了阻止这场混乱，我如果是金雕，我就会开枪）。

综上所述，完整的标注方案由以下类别和标签组成：

1. **hate speech（仇恨言论）：no – yes（无-有）**
2. **aggressiveness（攻击性）：no - weak – strong（无-弱-强）**
3. **offensiveness（冒犯性）:no - weak - strong（无-弱-强）**
4. **irony（讽刺性）：no - yes（无-有）**
5. **stereotype（刻板印象）：no - yes（无-有）**
6. **intensity（程度）：0 - 1 - 2 - 3 – 4**

第二十一届中国计算语言学大会论文集，第694页-第703页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

评论内容	仇恨言论	攻击性	冒犯性	讽刺性	刻板印象	辱骂性	仇恨程度
(1) а нах мне твоя россия россия всегда претендует на землю украины и в добавок уничтожает украинцев так было всегда так что хватит украину приписывать к россии у украины свой путь и нелезте к нам 我才不需要你们俄罗斯呢，俄罗斯觊觎乌克兰的土地，此外还杀乌克兰人，一直如此，不要再往妄想把乌克兰划入俄罗斯了，乌克兰有自己的发展道路，不要把手再伸向我们	有	无	弱	无	有	无	2
(2) пропаганда украинских наци 乌克兰纳粹的宣传	有	无	强	无	无	无	1
(3) молодцы украинцы приятно видеть то как они боролись за свои права и свободы и били поганых мусоров слава украине 干得好，乌克兰人很高兴看到他们如何为自己的权利和自由而战，打击糟糕的垃圾，光荣属于乌克兰	有	强	强	无	无	有	3
(4) давно пора эту мразоту майдановскую боевыми отстреливать 早该用武力打击这个Maidan败类了	有	弱	强	无	无	有	4
(5) менты не люди люди не менты 警察不是人，人们不是警察	有	无	强	有	无	有	1
(6) хочу что бы россия поступила с украиной как с грузией закрыть границу намертво и посмотреть как евродемократия будет процветать 我希望俄罗斯像对待格鲁吉亚一样对待乌克兰，严格关闭边界，看看欧洲民主将如何繁荣	无	弱	无	有	无	无	0
(7)я бы в вас кинул уебков если ..... 如果.....我会向你们这些混蛋扔	有	强	强	无	无	有	4
(8)надо было чтоб они тебе голову проломили чтоб другие такие как ты не рождались 他们应该打断你的脑袋，以免再生出像你这样人	有	强	无	无	无	无	3

Table 1: 标注示例

4.3 标注示例

表1中的示例显示了我们在语料库中如何应用4.1和4.2提及的类别和程度进行标注。如上所述，在我们的方案中，评论文本的仇恨性程度与我们标注的强度直接相关：如果文本中不存在仇恨言论，则其强度等于0，否则其强度的范围为1级到4级。除了强度这个标注项以外，所有其他的标注项都是相互独立的，我们给出的所有标注项既可以单独存在于句子中，也可以同其他标注项一起出现在句子中。因此，一条评论文本中可能包含仇恨言论，但同时几乎不包含或只含有一种其他的现象和含义（见表1中的评论2）；仇恨言论可能伴随着许多其他现象一起存在（评论1、3、4、5）；一条评论中也可能不包含仇恨言论，但是含有其他的现象（评论6）。

示例（1）表达了说话人对俄罗斯一贯行为的印象，认为俄罗斯总是掠夺土地、伤害人民，同时表达了对俄罗斯此种行为的厌恶，期望俄罗斯停止这种行为。此评论中还有对俄罗斯行为的刻板印象，同时话语中隐含俄罗斯的行为会对说话人的群体造成威胁和伤害，会引起对目标群体（俄罗斯人）的厌恶或仇恨，因此，选择将其注释为仇恨言论的强度为2级。在评论（2）中，说话人认为某种行为是乌克兰的纳粹行为，是对群体的一种言语冒犯，“纳粹”这一词反映了对目标群体侮辱性的判断，认为目标群体隐含的这种负面特征可能会对说话人及其群体构成威胁，无明显的煽动性言论，因此仇恨程度为1级。第（3）条评论中“打击糟糕的垃圾”这一言论内容中包含了攻击的行为（“打击”），但这是对句子中主语（“乌克兰人”）暴力行为的一种赞扬，而说话者本身并未表现出参与暴力的意愿。文本中同时包含了言语的冒犯（“垃圾”）和辱骂性词汇（“垃圾”），表达了对暴力的赞扬，因此标注仇恨性的等级为3。第（4）条评论中“败类”是对反政府示威人员的侮辱性的形容，具有辱骂性和冒犯性。本条评论和第（3）条评论中都有“打击”一词，但在本句中，这一行为是说话人对暴力的直接呼吁，并且强调“用武力打击”，可见呼吁暴力的程度非常高，因此仇恨性标记为4级。第（5）条评论带有明显的讽刺意味，该评论运用回文的修辞方式表达了对“警察”的强烈不满，并且辱骂他们“不是人”，具有冒犯性、辱骂性，不具有攻击性，因此仇恨性等级标注为1级。在评论（6）中，不存在仇恨性的言论，说话人“希望”俄罗斯对乌克兰实施“严格关闭边界”政策，这属于一种对目标群体实施某种政策的期许，带有较弱的攻击性，但并没有呼吁暴力或者伤害、侮辱等行为，所以不属于仇恨言论，“看看欧洲民主将如何繁荣”属于一种带有讽刺意味的表达，因此含有讽刺性。

以上描述的示例主要是为了能使读者清楚我们在语料库标注时所采用的判断标准。同时，这些标注的选择也突出了我们研究的关键点——首先要明确各个维度标注项目的定义，然后正确选择标注项，最后实现维度和程度上的精准标注。

(c) 2022 中国中文信息学会计算语言学专业委员会

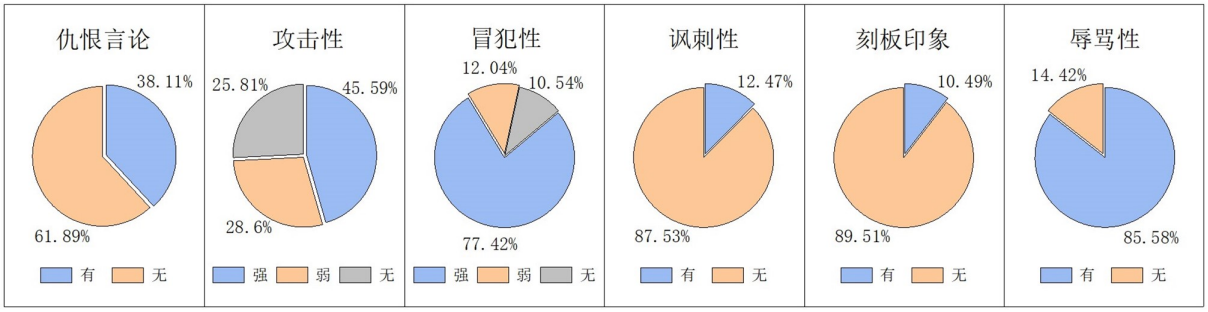


Figure 1: 各标注项在语料库中的占比

5 结果与讨论

5.1 标注结果分析

在本文中，我们对获取到的20476条评论内容中的1220条随机评论文本进行释义和标注，经过统计得到了仇恨言论在总体数据上的分布情况，以及各个维度的标注项在仇恨言论中的分布情况。考虑到我们的标注语料库是为仇恨言论自动检测研究所服务的，因此，在这里我们以仇恨言论为中心，主要研究与仇恨言论相关的其他维度的标注项在评论文本中的表现，并对这些标注项与仇恨言论的关系进行观察和分析。通过对数据的观察，我们发现在以民族与政治问题，特别是本文中选取的反政府示威运动事件为主题的语料库中，仇恨言论多以冒犯性（占评论数据的88%）、辱骂性（86%）和攻击性（74%）作为其主要的表达方式，而讽刺性（12%）和刻板印象（10%）的占比较低（如图1所示）。正如我们之前探讨过的，网络空间具有虚拟性、欺骗性、随意性和复杂性，这使得网络社交平台常常成为仇恨性言论的爆发地。人们在网络上即使是发表了极端的、情绪化的言论，明确地煽动实施暴力或歧视性的行为，似乎也不需要付出什么代价和成本。特别是我们所选取的民族和政治这类话题更容易引起极端情绪的宣泄，甚至是直接的言语对抗和辱骂。因此，在仇恨言论中，带有攻击性和冒犯性的言论占多数，成为人们表达仇恨的主要方式。

随后，我们研究了仇恨言论的强弱程度与攻击性、冒犯性、辱骂性，以及讽刺性和刻板印象之间可能存在的联系。图2中的结果显示了这些标注项在1至4级仇恨强度上的分布。

关于攻击性，在图2-a中我们可以观察到，在较低强度的仇恨等级（1级和2级）中，文本中不存在攻击性或仅存在弱攻击性的频率非常高，但强攻击性的文本出现在低等级仇恨言论中的概率就非常低了。这表明，评论文本在较高的仇恨等级（3级和4级）中，大多数都具有强攻击性。因此，在对仇恨言论及其煽动性进行定义时，需要着重考虑攻击性。整体上来看，无论哪种仇恨程度，攻击性言论的出现频率都很高，并且攻击性的强弱程度和出现的频率与仇恨性的强弱程度基本一致。这些发现与我们上述对仇恨言论的表达方式的理解完全一致。

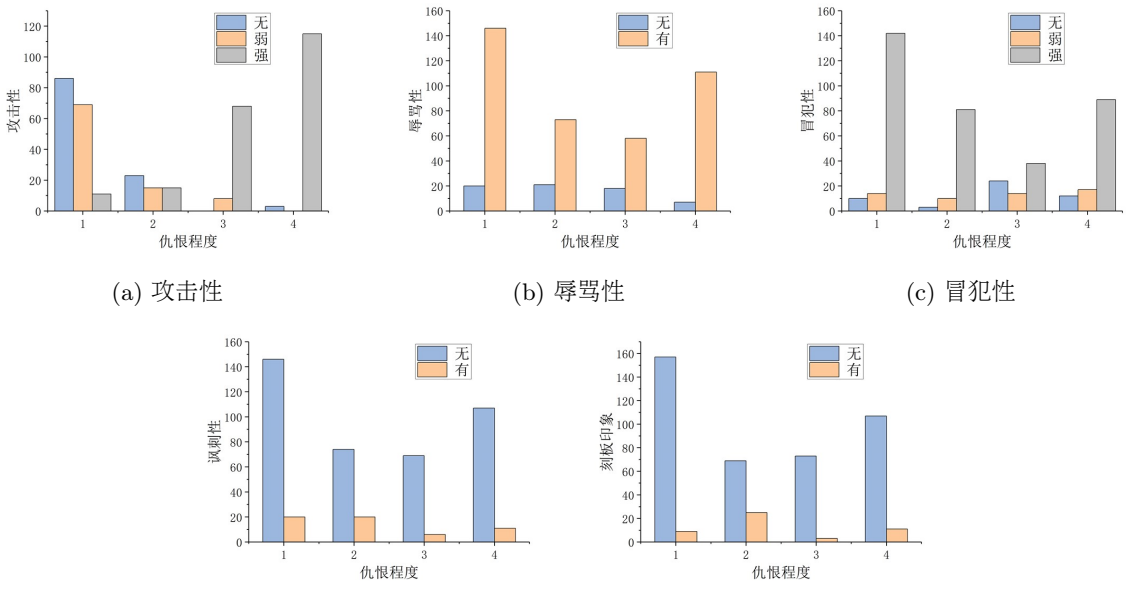
在图2-b中，我们能够看到，与仇恨言论关系同样十分密切的辱骂性言论，其分布几乎涵盖仇恨程度的所有级别，并且所占比例非常高，不含有辱骂性的仇恨言论占比非常低，我们几乎可以得出结论，认为辱骂性和仇恨性具有极高的吻合度和一致性。

从图2-c中我们能够观察到，冒犯性在所有仇恨的等级中均有分布，冒犯性几乎跨越了所有仇恨的强弱等级。然而也有一些规律可循：冒犯性在1级和4级上所占比例更多。这是因为，在标注规则中我们规定仇恨等级为1的是没有明确的煽动性，评论者将不良特征或品质归因于目标群体。这些评论更类似于基于刻板印象的侮辱或判断，也就是对目标群体的人格进行侮辱，甚至是谩骂。但这些言论行为没有煽动对目标群体实施暴力等过激行为，因此很多冒犯性的言论都被归为仇恨程度较弱的1级。在4级仇恨中，冒犯性出现的频率也比较高，这可能是由于在4级仇恨中，出现频率最高的冒犯性和攻击性往往同时出现。

通过对比图2-b和图2-c我们发现，辱骂性和冒犯性，特别是强冒犯性在4个仇恨等级中的分布形势非常一致，这是因为绝大多数强冒犯性的言论都具有辱骂性（当然，不排除少数冒犯性言论不具有辱骂性），所以二者的分布态势基本一致。

在对图2-d和图2-e的观察中我们发现，虽然讽刺性和刻板印象在仇恨言论中的占比不高，但其二者依然对仇恨言论的表达具有很重要的影响，不容忽视。情感文本中的情感分为显式情感和隐式情感。显式情感是指在一个情感文本中包含明显的情感词语，如高兴、漂亮、很棒这

(c) 2022 中国中文信息学会计算语言学专业委员会



(a) 攻击性　(b) 辱骂性　(c) 冒犯性

(d) 讽刺性　(e) 刻板印象

Figure 2: 仇恨程度1-4级在各标注项中的分布

些词语。而隐式情感是指在一个情感文本中并不包含情感词语，比如："您真能说，把黑的说成白的。"带有讽刺性和刻板印象的仇恨言论通常含有一种隐式煽动，这是一种隐式情感的表达。隐式煽动常以更巧妙的方式表达对特定目标群体的偏见、歧视和仇恨，有时还会减轻表达的程度。在民族和政治事件这一主题的背景下，大多数用户在表达仇恨时会很明确地煽动实施暴力或歧视性的行为，甚至是直接对目标群体进行辱骂。少数较为理性的用户在表达仇恨时会在其评论中隐含一定的煽动性，利用讽刺性的语言和基于刻板印象的言论尽可能地降低自身言论可能带来的风险和舆论的谴责。

此外，从计算语言学和自然语言处理的角度来看，当我们利用语料库进行仇恨言论自动检测研究时，所有的这些标注项都将具有非常重要的参考价值。

5.2 标注方案讨论

由于在前文中我们描述的标注方案具有较高的复杂性，并且仇恨言论这一主题会涉及主观性，在专家标注工作（A组）结束后，我们还随机邀请了以俄语为母语的留学生对已标注的文本（1220条）进行二次标注（B组），通过对比实验和数据分析来应对标注工作的复杂性和主观性。

两组（A组和B组）的标注结果对照情况如表2所示：

	仇恨言论	**攻击性**	**冒犯性**	**讽刺性**	**刻板印象**	**辱骂性**
专家	0.38	0.75	0.88	0.13	0.11	0.86
留学生	0.32	0.68	0.79	0.09	0.08	0.85

Table 2: 两个标注子集中每个标注类别的结果比例，即由专家标注的（表中的第一行）和由本土标注者（以俄语为母语的留学生）标注的结果（第二行）

从表格中我们可以看出，两组标注结果虽略有不同，但总体具有较高的一致性，这证明我们提出的标注方案和规则质量较高，具有一定的可理解性和客观性。我们同时对两组结果的偏差进行了深入分析，认为有以下三个方面：

- **个体差异**。虽然两次标注的方案和规则都是相同的，但由于本土标注者（以俄语为母语的留学生）的母语背景、文化程度和受教育程度不同，导致他们和专家对文本内容、标注规

(c) 2022 中国中文信息学会计算语言学专业委员会

则的理解不尽相同。尽管标注一致性偏差并不明显，但这些本土标注者的判断并不能与专家的标注结果完全一致。

- **完成质量**。除了标注者的个体差异，标注时的工作态度也会对结果产生影响。我们收到的标注者的反馈表明，他们有的时候可能没有仔细阅读和充分考虑我们给出的标注规则和示例，往往由于自己的疏忽产生了一些不准确的标注结果。

- **规则制定**。两组数据中存在的不一致性表明标注规范中存在一些不足，这些不足会给标注者造成模棱两可的印象。当标注者产生疑惑时，现有规范和示例可能并不能为其提供有效的参考和帮助。此外，四个仇恨程度的级别（1级到4级）之间的区别通常基于语用而非语义特征，这将导致标注者更加重视评论者的态度，而不是其评论文本的实际内容。

此外，通过对比以上两组数据我们发现，专家标注者之间存在的最大分歧在于仇恨的强度。例如，根据我们的标注方案和规则，表1中的评论文本（7）在仇恨程度上应当被认为比文本（8）更为强烈、更加危险，仅仅是因为前者的评论者使用的是第一人称的语句结构，会涉及到个人的责任，而后者使则用了更独立的形式，没有涉及到评论者本人。但另一种观点认为，文本（8）的仇恨程度可能高于文本（7），由此产生了分歧。相反地，本土标注者（以俄语为母语的留学生）在仇恨程度方面的标注具有高度的一致性。本土标注者对于仇恨言论的标注一致性最高，对仇恨程度的标注一致性次之。这说明本土标注者之间具有较高的标注一致性和可靠性。当专家和本土母语标注者出现分歧时，可以优先考虑本土标注者的标注结果。

鉴于以上对于标注结果的观察，我们发现仇恨程度这个系数在我们的方案中是最有争议的一项。我们的研究结果表明，并非所有仇恨言论的表达方式都是相同的，而且仇恨的强弱程度和表达色彩也不尽相同。在有效定义仇恨强弱程度和表达色彩之前，仍有许多工作要做。

我们制定的标注方案和规则仍然存在一些不足，不能够解决在标注时遇到的所有问题。因此，日后我们对标注方案的思考和制定还有很大的提升空间，尤其是在仇恨强度的概念厘定以及对其进行标注的方式上仍需要改进，需要制定出一个更简单、更具有普适性的方案，例如在标注仇恨言论的程度时只标注为“强”或“弱”(Del Vigna12 et al., 2017)，或是对攻击性和冒犯性在表达色彩方面提出更加清晰和细致的规定等等。

6 结论

在本文中，我们构建了针对乌克兰民族和政治问题的俄语网络仇恨言论语料库，并提出了一种全新的、多维度的语料标注方案，以更深入地研究仇恨言论这一颇具复杂性的问题。具体而言，我们对俄语社交网络评论文本进行了基于话题的选取和收集。除了对评论文本中是否含有仇恨言论进行判断，我们还标注了它的强弱程度（1级到4级），以及攻击性、冒犯性、讽刺性、刻板印象和辱骂性的存在与否和强弱程度。

我们对标注的结果进行了初步分析，总结了仇恨言论的具体特征，同时发现这样层次丰富且细粒度的标注方案并非没有缺陷，这些问题都已在本文示例中予以指出和讨论。该研究一方面为俄语网络仇恨言论的研究开辟了新的前景，尤其是为俄语网络仇恨言论自动检测方法提供了数据基础；另一方面，由于语料库自身的复杂性，我们所构建的语料库虽为基于话题的网络仇恨言论语料库，但我们认为，它不仅可以为仇恨言论本身的研究提供基础，还可能更详细、更系统地用于对其他语言现象的分析和研究。

致谢

感谢所有匿名评审人对本文的审阅，感谢首都师范大学隋然教授、王宗琥教授、北京外国语大学武瑗华教授、北京大学王辛夷教授对本文的建议与启发，谢谢！

参考文献

Azalden Alakrot, Liam Murray, and Nikola S Nikolov. 2018. Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181.

Ali Alshehri, Hassan Alhuzali El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2018. Think before your click: Data and models for adult content in arabic twitter. In *TA-COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety*, volume 15.

(c) 2022 中国中文信息学会计算语言学专业委员会

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The maneco corpus and some input from critical discourse analysis. *arXiv preprint arXiv:2008.06222*.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Karmen Erjavec and Melita Poler Kovačič. 2012. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Filip Klubička and Raquel Fernandez. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. *arXiv preprint arXiv:1805.04661*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Alessandro Maisto, Serena Pelosi, Simonetta Vietri, Pierluigi Vitale, and Via Giovanni Paolo II. 2017. Mining offensive language on social media. *CLiC-it 2017 11-12 December 2017, Rome*, page 252.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Max Weber. 2009. *The theory of social and economic organization*. Simon and Schuster.

Екатерина Николаевна Василенко. 2019. Язык вражды в заголовках новостных интернет-статей (на материале белорусского сегмента всемирной сети). *Вестник Курганского государственного университета*, (1 (52)):80–84.

Павел Николаевич Хроменков. 2016. Лексика вражды в публичной политической риторике периода холодной войны (на материалах инаугурационных речей президентов США середины xx в.-80-х гг. xx в.). *Вестник Московского государственного областного университета. Серия: Лингвистика*, (3):107–117.

ММ Шарнин, НС Ищенко, and НЮ Пахмутова. 2018. Использование методов тематического моделирования многоязычных коллекций для прогноза тревожных событий. In *Шестнадцатая Национальная конференция по искусственному интеллекту с международным участием КИИ-2018*, pages 297–304.

(c) 2022 中国中文信息学会计算语言学专业委员会

基于强化学习的古今汉语句子对齐研究

喻快　　邵艳秋　　李炜*
北京语言大学/信息科学学院，
国家语言资源监测与研究平面媒体中心，
北京市海淀区学院路15号，100083
yukuai_get@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

摘要

基于深度学习的有监督机器翻译取得了良好的效果，但训练过程中需要大量质量较高的对齐语料。对于中文古今翻译场景，高质量的平行语料并不多，而粗对齐的篇章、段语料比较容易获得，因此语料对齐很有研究价值和研究必要。在传统双语平行语料的句子对齐研究中，传统方法根据双语文本中的长度、词汇、共现文字等语法信息，建立一个综合评判标准来衡量两个句对之间相似度。此类方法虽然在单句对齐上取得了较好的效果，但是对于句子语义匹配的能力有限，并且在一些多对多的对齐模式上的性能表现不佳。在本文中我们提出尝试利用现在发展迅速且具有强大语义表示能力的预训练语言模型来考虑双语的语义信息，但是单独使用预训练语言模型只能考虑相对局部的信息，因此我们提出采用基于动态规划算法的强化学习训练目标来整合段落全局信息，并且进行无监督训练。实验结果证明我们提出的方法训练得到的模型性能优于此前获得最好表现的基线模型，尤其相较于传统模型难以处理的多对多对齐模式下，性能提升较大。

关键词： 双语对齐；预训练语言模型；强化学习；动态规划

Research on Sentence Alignment of Ancient and Modern Chinese based on Reinforcement Learning

Kuai Yu　　Yanqiu Shao　　Wei Li*
Information Science School, Beijing Language and Culture University,
Language Resources Monitoring and Research Center,
15 Xueyuan Road, HaiDian District, Beijing, 100083
yukuai_get@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

Abstract

Supervised machine translation based on deep learning has achieved good results, but high-quality aligned corpora are needed in training process. There hasn't been a lot of parallel corpora of high quality for ancient and modern Chinese translation scenes, while coarsely aligned discourse and paragraph corpora are relatively easy to obtain. Therefore corpus alignment is quite valuable and necessary for research. In the study of sentence alignment in traditional bilingual parallel corpus, a comprehensive evaluation criterion is established to measure the similarity between two sentence pairs according to the grammatical information of bilingual text, such as length, vocabulary and context. Although it has achieved good results in single sentence alignment, it has limited ability in sentence matching and poor performance in some many-to-many alignment patterns. We attempt to consider bilingual semantic information by using

* 通讯作者　Corresponding Author

第二十一届中国计算语言学大会论文集，第704页-第715页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

the rapidly developing pre-trained language model with strong semantic representation capabilities. However, the pre-trained language model itself can only cover relatively local information, so we propose reinforcement learning training objectives based on dynamic programming algorithm to integrate global information of paragraphs, and then carry out unsupervised training. Experimental results show that the performance of the model trained by our proposed method is better than that of previous baseline model with best performance, especially compared with many-to-many alignment model, which is difficult to deal with by traditional models.

Keywords: Bilingual Alignment , Pre-trained Language Model , Reinforcement Learning , Dynamic Programming

1 引言

当前基于深度学习的神经机器翻译(Bahdanau et al., 2014; Zoph et al., 2016; Vaswani et al., 2017; Yong et al., 2018)获得了出色的性能表现，基于端到端神经机器翻译相较于传统的统计机器翻译而言有巨大的优势，然而这些有监督的模型训练需要大量双语平行语料。如针对英德、英法翻译等其他多语言翻译采用了数百万条平行语料(Ouyang et al., 2020)，模型的翻译能力很大程度建立在超大规模的平行数据上，而对于一些专门领域的翻译来说，用人工的方式去获取翻译语料，覆盖面有限而且代价高昂。因此，从大规模真实双语语料中自动挖掘双语语料已经成为获取翻译知识的非常重要的途径，其中一项关键技术就是双语对齐技术。

从20世纪80年代开始有众多学者对双语句子对齐展开了研究，句子对齐的方法主要有基于长度和基于词汇对齐的方法。Church和Gale(Gale and Church, 1993)提出以文本长度为特征进行文本对齐，这种方法需要的信息简单，在法英双语对齐中取得了很好的效果，但是这个方法有个天然的缺陷，如果考虑以长度作为特征，计算机容易把两个1-1模式的正确句子给自动对齐成2-2模式，并且忽略了文本中的语义信息。Utsuro等(Utsuro et al., 1997)提出使用统计词典信息在英语和日语两种语言之间进行对齐，规避了仅仅使用长度信息所带来的限制，提高了句子对齐的准确率，但是准确率仍然容易受到翻译风格等方面的影响。Lin等(Lin and Wang, 2007)在古文与现代文对齐任务中结合长度信息与共现汉字信息作为特征，使得对齐性能有了进一步地提高。但是考虑共现汉字这一特征，往往会产生多对多的模式对齐，因为考虑的句子个数越多，共现的汉字个数也多，所以容易产生多个句子对齐多个句子的情况，这实际上是与语料中存在绝大多数1-1的对齐模式(刘颖and 王楠, 2013)情况是互相矛盾的。

前人的研究主要是应用语法规则等特征进行对齐，但是在古代与现代文翻译的语料中，因为翻译风格的差异，有些翻译用词会有意译等情况，单纯的使用语法、句法、共现信息是不够的，应该尝试考虑语义信息。并且实际语料中会存在一些多对多的对齐语料，这类匹配模式恰好是语法规则难以覆盖全面的部分。由于不同时代、领域的古文会存在差异并且对应复杂多样的对齐模式，人工收集各类语料会带来高昂的成本，因此尝试无监督的方法是非常有必要的。

在深度学习和大数据的支撑下，自然语言处理技术迅猛发展。ElMo(Peters et al., 2018)、BERT(Devlin et al., 2018)、GPT-2(Radford et al., 2019)等预训练语言模型在下游任务取得了较好的效果，而基于语言模型预训练和在下游任务精调的框架把自然语言处理带入了一个新的阶段。预训练语言模型的领域适应能力很强，并且在语义建模方面相较于传统语法规则方法更具有优势，因此相比语法规则来说，预训练语言模型更适用于匹配双语句对。

直接使用预训练语言模型进行语料对齐会存在只能建模局部对齐的问题，模型无法综合的考虑全局文本信息而得出最优的对齐结果；此外，如果单纯的使用最长公共子序列（Longest Common Sub-sequence, LCS）算法来构建平行语料进行远程监督训练，不仅会引入大量噪声，影响模型的匹配效果，并且由于多对多对齐模式的出现比例较低，模型很难通过远程监督信号提升此类对齐的能力。

针对这些问题，并基于双语语料中绝大多数句子存在对应翻译的假设，我们提出以覆盖尽量多平行句对为目标的强化学习方法，来整合段落的全局信息，预训练语言模型可以以此从段

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

对齐模式	古文	现代文
1-1对齐	寻此县自不出银，又俚民皆巢居鸟语，不闲货易之宜，每至买银，为损已甚。	考虑到这县不出银矿，加上俚民都住在洞中，说话像鸟语一般，不熟悉交易的方法，每到买银子的时候，损失又很大。
1-2对齐	顷之，世祖命江州众军悉同大举，僧辩乃表皇帝凶问，告于江陵。	不久，世祖命江州各路人马一同大举进攻。//僧辩于是公布皇帝去世的噩耗，把凶讯报告在江陵的世祖。
2-1对齐	城内同时鼓噪，矢石雨下，杀贼既多，贼乃引退。//世祖又命平北将军胡僧率兵下援僧辩。	城内守军同时击鼓呼叫，箭矢码石齐下如雨，杀伤很多叛军，叛军才退兵，世祖又命平北将军曲僧枯率兵沿江而下增援王僧辩。
2-2对齐	又称两受入，易生奸巧，山俚愚怯，不辨自申，官所课甚轻，民以所输为剧。//今若听计丁课米，公私兼利。	另外称两而收，容易导致奸伪狡诈山民。//山民愚蠢而胆怯，无法申明，公家所收的税很轻，俚民却认为很重，如今如果允许按丁收米税，于公私都有好处。

Table 1: 对齐模式示例，//为句子之间的分隔标识

落的全局信息状态来求得最优解，在强化学习过程中，模型对于不同的对齐模式所做出的判断给予相应的奖励信号，模型在决策序列中以最大化双语句对对齐覆盖度这个目标奖励作为探索原则，提升模型在多对多对齐的匹配能力，这样也解决了预训练语言模型在训练中缺乏监督信号的问题，同时模型在大量的无标注的数据中不断地学习，模型对于多对多的对齐模式会有更好的理解。本文的主要贡献有：

- 相较于传统的双语文本对齐方式，本文首次在古文现代文语料对齐任务中引入预训练语言模型来更好的解决古文与现代文的语义匹配问题。
- 本文提出以覆盖尽量多平行句对为目标的全局强化学习训练方法来优化模型，使得模型可以从全局的角度整合段落整体信息，进一步提升了模型在多对多模式的处理能力。
- 我们对于中文古文和现代文对齐的任务构建了大量语料，并且进行了大量的实验和分析，结果表明我们提出的方法在多种对齐模式上的性能均有提升，尤其是在多对多模式的对齐，效果提升尤为明显。

2 相关工作

近些年，在古代和现代文对齐任务中，一种简单结合最长公共子序列与动态规划的算法(Zhang et al., 2018)获得不错的效果。Liu等(Liu et al., 2019)在动态规划的算法基础上结合古文与现代文句子的长度信息、词汇信息、编辑距离的方法，更加全面的考虑到文本中的特征。将依存句法分析融入BiLSTM+CRF(韦希林, 2019)的任务感知模型中，对长句子进行序列标注，使得模型更好的找到长句的分割点，提升句子对齐的准确率。

Gregoire和Langlais(Grégoire and Langlais, 2017)最先提出使用深度学习的方法来解决句子对齐问题，将两个独立的双向循环神经网络对句子进行编码并得到向量化的表示，并将句子向量化表示接入一个全连接层映射成二分类的结果向量，在不依赖手工特征的情况下取得了较好的效果。本文在前人的基础上，引入了更为先进、语义表征能力更好的预训练语言模型BERT对句子进行编码，得到更好的语义匹配能力。

前人的句子对齐工作普遍取得较好的效果是由于数据集中天然存在着大量1-1的对齐模式(刘颖and 王楠, 2013)，该模式对于模型的分类是比较简单的，但是对于数据中存在的部分m-n($0 \leq m$, $n \leq 2$)匹配模式，如表1所示的2-2、1-2、2-1对齐模式。传统利用语法规则特征的方法主要通过提升1-1的对齐模式来提升模型的性能，却忽略了多对多的对齐模式的准确率。如果仅使用预训练语言模型进行改进，必须根据多种匹配模式进行数据标注，这种从段落对齐的语料库中获取数据的过程是非常繁杂的。并且由于语料中各种匹配模式的比例是不均衡的，要获取数量充足且均衡的样本进行有监督学习是不现实的。本文使用强化学习使得模型尽可能注意到多对多的对齐模式的匹配，以覆盖尽可能多的平行句对为目标，进而避免这一问题。

第二十一届中国计算语言学大会论文集，第704页-第715页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

3 任务与数据

3.1 数据选择和预处理

本文的数据集来源于《二十四史》，我们在互联网[0]上爬取并清洗了300篇古代汉语与相对应的现代文文章，由于段落之间的边界比较容易判断，我们人工的对齐了约9000个段落对齐对，并从中抽取了200对段落对齐语料，进行人工句子对齐，得到1550句对并且标记对应的对齐模式作为本次实验的测试集，测试集的统计数据如表2所示，我们将剩余的人工对齐段落对作为本次实验的训练语料。

在古汉语和现代汉语的数据中，存在多种不同的对齐模式，1-1的对齐模式远远高于其他类别的对齐模式，根据(刘颖and 王楠, 2013)对古文与现代文数据集的统计，1-1的对齐模式占比达到了90.12%，2-2以内的对齐模式的占比达到了98.98%，为了兼顾模型的效率，我们只考虑1-0、0-1、1-1、1-2、2-1、2-2这六种对齐模式，其他出现频率不高的模式没有在考虑的范围之内，并且由于1-0、0-1对齐模式对于模型无法判断，在后续实验中我们将二者归并到2-1、1-2对齐模式里面讨论。

	类别	频率	概率	
主要类别	1-0、0-1	6	0.39%	99.48%
	1-1	1346	86.83%	
	1-2	64	4.12%	
	2-1	106	6.85%	
	2-2	20	1.29%	
其他类别	2-3	2	0.13%	0.52%
	3-1	2	0.13%	
	1-3	4	0.26%	
总计		1550	100%	100%

Table 2: 对齐模式统计

3.2 获取粗对齐平行句对

预训练模型需要构建一定的平行句对语料来对模型微调，使得模型获得初步的粗标能力。将现代汉语翻译成古代汉语存在一个特点，在古代汉语的每一个词都倾向于在现代汉语文本中出现，通常有很强的的顺序性，正确的对齐的语料对通常具有最长的公共子序列的最大长度和，根据这一特点，我们将已有的手工对齐的古文与现代文的段落对作为输入，使用LCS算法来进行段落内部的句子的对齐，构建粗粒度的句对齐的平行语料，并且将其应用到后续的实验。

在微调预训练语言模型之前，模型需要从非平行句对中识别出平行句对，因此必须生成负例，我们使用负采样方法构建训练语料(Conneau et al., 2018)。我们获取包含n个平行句对的平行语料库作为正例，为了构建负例，对于每一个平行句对，在语料库里面进行随机采样，构建新的句对生成m个负例，并用这些数据对BERT模型进行微调，并且记录微调BERT模型在测试集上的结果。

4 我们的方法

本文提出的方法希望通过使用预训练语言模型强大的语义匹配能力，提升模型在古文与现代文句子对齐任务的性能。由于段落中出现的句子的匹配需要结合上下文句子的匹配情况进行考量，本文提出结合动态规划的算法，使得模型能够出于段落的整体角度来考虑全局信息进行匹配，在对句子进行匹配的过程中，模型基于先前的决策做出当前状态的最优解，由于难以直接获得句子级对齐的标签，我们提出采用基于动态规划目标的强化学习方法，其中每一次决策过程就是通过预训练语言模型对当前需要判断的一或多对句子判别是否匹配，而总的目标是使得模型得到的匹配能够覆盖尽量全的句子对。

我们将按照以下两个部分介绍我们的方法：基于预训练语言模型的语义匹配模块、基于动态规划的强化学习序列决策模块。

[0]http://m.lishishuwu.com/index.php/list/shi24.html

第二十一届中国计算语言学大会论文集，第704页-第715页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

4.1 基于预训练语言模型的语义匹配

受到BERT(Devlin et al., 2018)在NLI等句对建模任务上成功应用的启发，本文提出使用类似于BERT中下一个句子判断的形式来建模双语句子级别的匹配预测，以构建两段文本之间的关系。文本之间关系的判断是一个二分类任务，在输入层，我们将古文、现代文的文本分别记为$x^{(1)}$，$x^{(2)}$，如式(1)所示：

$$\begin{aligned} x^{(1)} &= x_1^{(1)} + x_2^{(1)} + \cdots + x_n^{(1)} \\ x^{(2)} &= x_1^{(2)} + x_2^{(2)} + \cdots + x_m^{(2)} \end{aligned} \tag{1}$$

我们通过在输入加入特殊[CLS]来指示文本整体的表示，使用[SEP]来分隔来自双语的文本片段。需要注意的是，根据后文中将要介绍的强化学习中具体需要做出的决策的不同，这里x_1和x_2可以是一句或者多句单源文本的拼接。经过如下处理得到输入的表示X。

$$X = [CLS]x_1^{(1)}x_2^{(1)}\cdots x_n^{(1)}[SEP]x_1^{(2)}x_2^{(2)}\cdots x_m^{(2)}[SEP] \tag{2}$$

[CLS]表示文本序列开始的特殊标记；[SEP]表示文本序列之间的分隔标记。

对于给定的输入X，经过式(3)得到BERT的输入表示v

$$v = InputRepresentation(X) \tag{3}$$

在BERT编码层中，输入表示v经过编码，借助自注意力机制充分学习文本中每个词之间的语义关联，最终得到输入文本的上下文语义表示h。

$$h = BERT(v) \tag{4}$$

NSP任务只需要判断输入文本$x^{(2)}$是否是$x^{(1)}$的下一个句子，因此在NSP任务中，BERT使用了[CLS]位的隐含层表示分类预测。具体地，[CLS]位的隐含层表示由上文语义表示h的首个分量h_0构成，因为[CLS]是输入序列中的第一个元素。在得到[CLS]位的隐含层表示h_0后，通过一个全连接层预测输入文本的分类概率P：

$$P = Softmax(h_0W^p + b^0) \tag{5}$$

W^p表示全连接层的权重；b^0表示全连接层的偏置，最后在得到分类概率P后，与真实分类标签y计算交叉熵损失，学习模型的参数。由于没有直接的监督信号，单纯将NSP目标移植到本任务上难以进行有效的训练，所以我们后续将其结合强化学习的方法来构建训练目标，弥补监督信号缺失的问题。

4.2 基于动态规划的强化学习序列决策方法

为了弥补监督信号缺失的问题，我们采用动态规划结合强化学习的方法进行优化。强化学习(Francois-Lavet et al., 2018)通过与环境的交互，在复杂的状态空间序列中不断的决策获得奖励信号，最大化策略奖励。在我们的任务中，输入是古文与现代文的段落对齐文本，环境是动态规划过程每一步根据需要考虑的匹配模式所构建的句对、状态空间是动态规划过程中所有时间步判断的序列、决策是模型对每一步输入的对齐模式句对判断是否匹配。

通过强化学习来学习一个策略网络进行分类，即4.1节我们提出的基于预训练语言模型的语义匹配，分类过程可以被看成是一个决策过程，由于决策过程是离散的，因此我们使用策略梯度算法(Sutton et al., 2000)来更新参数，网络模型的参数θ根据决策过程的奖励(reward)进行批量更新。

4.2.1 动态规划目标与设定

我们使用动态规划算法，从段落到段落之间来寻求全局最优解，动态规划的目标是使得模型预测的匹配能够覆盖尽量全的句子对（即所有句子都能找到对应的翻译），D(i, j)表示第1句到第i句古文与第1句到第j句现代文对齐的分数，每一步我们需要考虑多种对齐模式，由

第二十一届中国计算语言学大会论文集，第704页-第715页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

于2-2以内的对齐模式占比达到了极大部分，故我们只考虑2-2以内的对齐模式，具体的状态转移方程如下所示：

$$D(i,j) = Max \begin{cases} D(i-1,j-1) + BERT(s_i, t_j) \\ D(i-1,j-2) + BERT(s_i, t_j \oplus t_{j-1}) \\ D(i-2,j-1) + BERT(s_i \oplus s_{i-1}, t_j) \\ D(i-2,j-2) + BERT(s_i \oplus s_{i-1}, t_j \oplus t_{j-1}) \end{cases} \tag{6}$$

其中$s_i \oplus s_{i-1}$表示将句子s_i 与s_{i-1}拼接起来，对于1-0与0-1对齐模式，我们将其归并到2-1、1-2模式，我们仅考虑1-1、1-2、2-1、2-2这四种对齐情况。由此可见，时间复杂度为$O(MN)$，其中M是某一古文段落中的句子数，N是对应的现代文段落中的句子数，随着语料库大小的增长，算法将更为耗时，本文提出的方法更适合段落对齐的语料，而不是篇章对齐语料。

如图1所示，段落对齐的语料中，古文段落被切分为4句，现代文段落被切分为3句，在对齐任务中，每一步的环境取决于动态规划过程中，基于先前的匹配所构建的句对，负责分类的策略网络是我们采用的预训练语言模型-BERT。在动态规划求解的过程中，每一步都会根据需要判断的四种对齐模式构建对齐句对作为输入，BERT判断句对是否匹配，如图中标记的节点所在时刻，动态规划基于先前语料中已经对齐的1-1状态，根据设定需要考虑的四种对齐模式来构建句对输入BERT进行判断，模型在此刻判断2-1对齐模式是匹配的，记录当前的匹配的对齐模式以及当前状态的总reward，以此类推，直到动态规划算法求解完毕，即获取到动态规划的最后一步所累积的reward，算法详情参考Algorithm 1。

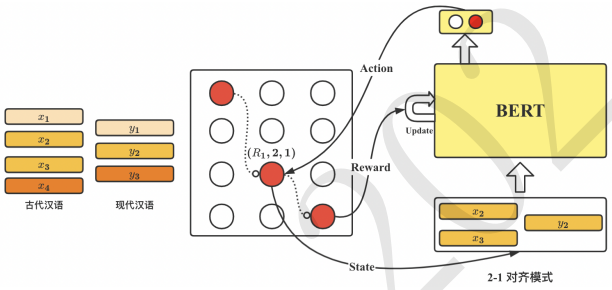


Figure 1: 结合强化学习和预训练语言模型的策略网络模型

4.2.2 Reward目标设定

reward设计总的目标就是使得模型匹配能够涵盖平行句对对齐的覆盖度，BERT模型的决策是一个二分类问题，我们定义为，当$a_t = 1$时，表示当前s_t对应的输入句对判断是匹配的；当$a_t = 0$时，表示当前s_t对应的输入句对判断是不匹配。

为了使得模型能尽可能的覆盖到多对多对齐模式上，我们对reward函数进行了如下设计，在动态规划的决策过程中的每一个时间步，我们根据不同匹配模式给予不同的reward，具体来说，当$a_t = 0$时，不匹配的情况下reward值为0，当$a_t = 1$时，我们对1-1对齐模式reward值设为2，1-2、2-1、2-2对齐模式reward值分别设为3、3、3，对于后三种对齐模式给予更高的reward是因为我们鼓励预训练语言模型尽可能的在对齐过程中察觉到非1-1的对齐模式，而且模型也会凭此获得更多的多对多对齐模式的监督信号，在训练得到更好的语义匹配能力。

4.2.3 策略梯度训练

策略网络BERT接收到当前对齐模式状态s_1，输出动作a_1，接着环境把a_1当成输入，构成新的状态s_2作为输入，进而输出决策a_2。BERT会不断地接收到动态规划过程中所采集到的状态，并且根据模型内部的参数θ而进行决策(action)，进而得到一系列的观测序列τ，动态规划完毕后即为完成一次完整的观测序列，每一次完整的观测可以视为一个轨迹$\tau = (s_1, a_1, s_2, a_2, \cdots, s_t, a_t)$。

轨迹就是当前的输入的对齐模式的句对、采取的策略，根据s_1执行a_1的概率$p_\theta(a_1|s_1)$，$p_\theta(a_1|s_1)$由BERT模型的参数θ决定，该输出是一个分布，BERT会根据这个分布进行采样，决定实际要采取的动作。接下来环境根据动作a_1与s_1产生s_2计算，以此类推，某个完整轨迹τ发生的概率为：

(c) 2022 中国中文信息学会计算语言学专业委员会

Algorithm 1 古文与现代文对齐→ 基于动态规划的强化学习序列决策方法

Require: 策略网络BERT模型的参数θ
初始化分数$D[i,j] \leftarrow 0$
for $i = 1$ to M **do**
  for $j = 1$ to N **do**
    for (d_i, d_j) to $Align - Pattern$ **do**
      if $i - d_i \geq 0$ and $j - d_j \geq 0$ **then**
        获取当前策略网络BERT根据输入所输出的reward, action, prob
        if $D[i,j] \leq (D[i,j] + reward)$ **then**
          $D[i,j] = D[i,j] + reward$
          更新当前的reward, action, prob
        end if
      end if
    end for
  end for
end for
return $R(\tau) = D[i,j]$
// 更新策略网络
Compute the gradient $\nabla \bar{R}_\theta$ base on Equation(11)
Update θ base on Equation(10)

$$p_\theta(\tau) = p(s_1) \prod_{t=1}^{T} p_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t) \tag{7}$$

把环境输出的状态s与BERT输出的动作a组合起来，并且将动态规划过程中每一步的reward累加起来作为当前回合的总reward，就得到了$R(\tau)$，我们的目标就是调整BERT模型参数θ，最大化$R(\tau)$的期望值$\bar{R}_\theta$。在决策过程，需要尽可能穷举每一个轨迹τ，从分布$p_\theta(\tau)$采样一个轨迹τ，计算$R(\tau)$的期望值，并且计算期望$\bar{R}_\theta$的梯度，

$$\nabla \bar{R}_\theta = \sum_{\tau} R(\tau) \nabla p_\theta(\tau) \tag{8}$$

计算某一个状态下某一个动作的对数概率$logp_\theta(a_t^n|s_t^n)$，对这个概率取梯度，在梯度前面乘一个权重，权重就是当前采样轨迹的奖励，计算出梯度后，就可以更新模型，梯度计算如公式(9)所示

$$\nabla \bar{R}_\theta = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} R(\tau^n) \nabla logp_\theta(a_t^n|s_t^n) \tag{9}$$

借鉴(Zinkevich, 2003)的工作，我们使用梯度上升(Gradient Ascent)的方式来更新参数θ，来最大化期望奖励，把θ加上梯度$\nabla \bar{R}_\theta$，学习率设置为η，学习率的调整可以使用Adam(Kingma and Ba, 2014)、AdaGrad(Duchi et al., 2011)等优化器来调整。

$$\theta \leftarrow \theta + \eta \nabla \bar{R}_\theta \tag{10}$$

4.2.4 策略网络初始化

我们从先前手工段对齐语料中抽取300段语料，使用LCS算法构建伪平行语料对BERT-base进行微调后得到的模型，作为本次实验的**基线模型**，并在此基础上进行强化学习训练。

在强化学习的训练过程中，采样轨迹τ得到$R(\tau)$一直是正的，由于只采样了部分的动作序列，某时刻的动作可能从来没被采样到，该动作在策略网络的执行概率会降低，但是并不意味着未被采样到的动作不是好的决策，为了让采样得到的$R(\tau)$有正负之分，可以添加基线b，使得总奖励为$R(\tau) - b$，如果$R(\tau) > b$，就让当前状态的动作发生的概率升高，反之亦然。通过使用这种带基线的策略梯度算法，提升了训练过程的稳定性，梯度计算如公式(11)所示：

$$\nabla \bar{R}_\theta = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} (R(\tau^n) - b) \nabla logp_\theta(a_t^n|s_t^n) \tag{11}$$

(c) 2022 中国中文信息学会计算语言学专业委员会

Model	P	R	F
Length (Gale and Church, 1993)	84.7	75.6	79.9
Statistical (Lin and Wang, 2007)	86.1	85.4	85.7
LCS (Zhang et al., 2018)	84.2	**87.6**	85.9
Proposed Method	**88.9**	86.1	**87.5**

Table 3: 本工作的模型和前人的工作在测试集上的性能表现

在强化学习训练过程中，b设置为基线模型根据输入的段落对齐文本在经过动态规划判断完毕所获得的总reward，reward函数同上4.2.2Reward目标的设定所述。

5 实验部分

5.1 实验设置

对于本实验的基线模型，我们使用LCS与动态规划算法构建的古文与现代文平行句对其进行远程监督训练，本次实验的选用Adam优化器，学习率设置2e-4，权重衰减为1e-4，并且采用Cosin Warmup(He et al., 2019)策略，学习率按照cosin函数进行衰减，为了提升强化学习的训练效率，我们提出的基于动态规划的强化学习的策略分类模型是在基线模型的基础上进行实验，实验的所有结果均为多次实验求平均值。

5.2 评价指标

本次实验的评价指标采用的是精确率(Precision)、召回率(Recall)、F_1-score。假设GB是人工标注的数据集，PB是模型产生的集合，召回率和精确率的计算公式为：

$$Precision = \frac{|GB \cap PB|}{|PB|} \qquad Recall = \frac{|GB \cap PB|}{|GB|} \tag{12}$$

召回率和精确率越高，说明句子的对齐效果越好，将召回率和精确率综合起来考虑的F_1值，计算公式为：

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{13}$$

5.3 实验结果与对比分析

为了验证本文提出的方法的有效性，我们将本文提出的方法与以下基线模型在测试集上进行比较，包括：

- 基于长度的对齐方法（Length）(Gale and Church, 1993)：句子对齐与对齐模式和句子长度相关，以句子的长度特征与对齐模式作为特征进行句子对齐。
- 基于综合特征的对齐方法（Statistical）(Lin and Wang, 2007)：在动态规划的框架下，综合运用长度、对齐模式和汉字信息作为特征，进行句子对齐。
- 基于LCS的对齐方法（LCS）(Zhang et al., 2018)：将句对之间的最长公共子序列作为句对之间相似度的衡量标准。

实验结果如表3所示，可以看出本文提出的方法相较于其他三种方法在性能上均有较大的提升，其中P值和F值为最优，本文提出的将预训练语言模型与动态规划结合强化学习的方法从两方面提升了古文与现代文对齐的性能。一方面，通过BERT的强大语义表征能力提升了匹配的准确率，更好的兼顾了手工特征难以照顾到的方面。另一方面，使用强化学习构建训练目标的方法使得模型能学习更丰富的匹配模式，提升模型的匹配能力。

5.4 分析

5.4.1 消融实验

为了进一步探究我们提出的方法各个模块对于对齐性能提升的作用，我们进行了消融实验。本次实验包括三个模型，如表4所示，基线模型是使用LCS算法构建的粗对齐的

(c) 2022 中国中文信息学会计算语言学专业委员会

Model	P	R	F
Baseline (LCS + BERT)	82.0	85.1	83.5
+DP	83.2	**86.4**	84.8
+DP & RL	**88.9**	86.1	**87.5**

Table 4: 消融实验，Baseline采用4.2.4节策略网络初始化的基线模型

	1-1(1346)	1-2(67)	2-1(109)	2-2(20)
Length (Gale and Church, 1993)	78.5(1056)	56.7(38)	55.0(60)	55.0(11)
Stastiacal (Lin and Wang, 2007)	91.1(1227)	53.7(36)	42.2(46)	40.0(8)
LCS (Zhang et al., 2018)	**93.0(1253)**	58.2(39)	45.9(50)	40.0(8)
Proposed Method	87.5(1178)	**82.0(55)**	**75.2(82)**	**60.0(12)**

Table 5: 不同模型在多对多对齐模式的准确率

平行句对语料对BERT进行微调，然后再使用微调后的模型对齐得到的结果。+DP是表示在Baseline基础上，在对齐过程以动态规划的全局角度对齐得到的结果。+DP & RL是在使用动态规划的基础上，使用强化学习构建训练目标得到的结果，也就是本文所提出的方法。+DP相较于Baseline而言，P值、R值分别提升了1.2%、1.3%，说明结合动态规划全局信息具有一定的效果。并且可以看到我们提出方法的P值、R值、F_1值相较于Baseline分别提高了6.9%、1.0%、4.0%；相较于+DP而言，我们的方法的P值提升了5.7%，召回率R下降了0.3%；召回率的下降是因为模型进行强化学习后，模型将小部分1-1的对齐模式错误地进行多对多的对齐，导致召回率偏低，但模型整体的性能得到了较大幅度的提升，也验证了我们提出的应用动态规划与强化学习的方法的有效性。

5.4.2 多对多对齐

为了验证本文提出的方法在多对多对齐模式情况下的效果，我们对不同对齐模式下的句子对齐的数量进行了单独的统计，从表5可以看出，相较于前人的工作来说，我们提出的方法在1-2、2-1、2-2等多对多对齐模式下准确率均为最优，并且有较大幅度的提升，我们的方法在1-1对齐模式上准确率相较于基于综合特征的对齐方法以及基于LCS的对齐方法分别低了3.6%、5.5%，这是因为我们使用强化学习构建目标过程中，使得模型尽可能匹配到非1-1的对齐模式，我们提出的方法在多对多对齐方面的表现更好了，这是符合直觉的，也证明我们在强化学习构建训练目标对于模型的整体性能提升是有价值的。

我们将人工对齐，+DP以及+DP & RL三种方法以图2中样例来进行对比，如图3所示，人工对齐的路径是(0, 0) → (2, 2) → (3, 3) → (4, 5) → (5, 6)，在Baseline的基础上，使用动态规划从全局的角度进行对齐，+ DP对齐的路径是(0, 0) → (1, 1) → (3, 3) → (4, 5) → (5, 6) 可以看出，因为使用远程监督信号含有大量的噪声，对齐的第一步，+ DP就错误地将2-2对齐判断为1-1对齐，直接导致下一步的对齐也错了，尽管后续(3, 3) → (4, 5)的1-2模式匹配正确，但是整体对齐的准确率降低了，而结合我们提出的结合强化学习构建训练方法，+DP & RL方法，在第一步就正确的匹配了2-2模式的句对，后续的对齐路径也是和人工对齐的路径完全重合，这也证明我们提出的方法在多对多对齐上确实是有一定效果的。

5.5 样例分析

尽管本文提出的方法取得了一定的提升效果，但是我们也发现了我们的方法难以解决的语义匹配的例子，如以下样例所示：

- 样例1
  - ◇ 古文：斩循及父嘏，并循二子，亲属录事参军阮静，中兵参军罗农夫、李脱等，传首京邑。
  - ◇ 现代文：杜军将卢循和他的父亲卢嘏，以及卢的二个儿子，卢的亲属录事参军阮静，中兵参军罗农夫、李脱等人斩首。他们的脑袋被送到京城。

(c) 2022 中国中文信息学会计算语言学专业委员会

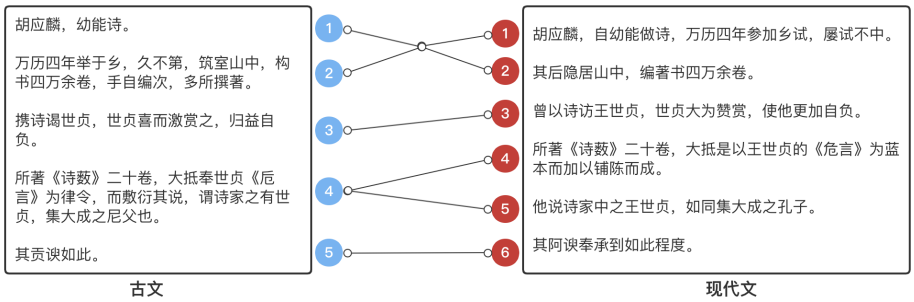


Figure 2: 人工对齐示例；古文第1、2句与现代文第1、2句为2-2对齐模式；古文第4句与现代文第4、5句对1-2对齐模式。

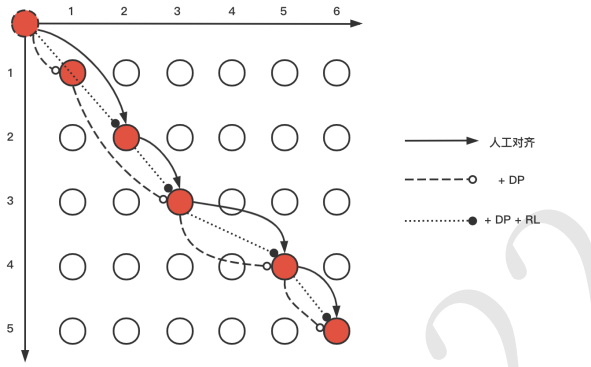


Figure 3: 不同方法的对齐路径对比

- 样例2
  - ◇ 古文：魏建国四年，慕容子晃伐之，入自南陕，战于木底，大破钊军。追至丸都。
  - ◇ 现代文：魏建国四年，慕容的儿子慕容晃攻打他，从南陕进军，大战于木底，大败钊的军队，追到丸都。

如样例1，该样例的正确的对齐模式是1-2，使用LCS算法可以很好的判断出是1-2对齐模式，但是使用我们提出的方法后，BERT模型却将其判断成1-1对齐模式，因为模型认为现代文中的“**他们的脑袋被送到京城。**”这一句在语义上与对应的古文相关度较低，并且该句的长度相较于前一句较短，构成类似于长句+短句的情况，模型在语义编码的时候，该句占比权重较低，该句的语义编码在整体中容易被忽略，造成了模型的漏匹配，故判断成了1-1对齐模式。

同样的问题在样例2中更为明显，BERT模型同样的将古文中的“**追至丸都。**”这一句给漏匹配，样例2的现代文更为明显的构成了长句+短句的情况，使得模型出现了漏匹配的情况。在先前的阐述中，我们将0-1、1-0的对齐模式归纳到1-2、2-1进行讨论，这样会直接的将一句无关的文本加入到相关的1-1对齐的句对中，给训练样本中注入了小部分的噪声，也间接性的造成了模型对于长句+短句出现漏匹配的情况。

6 结论

针对古代汉语和现代汉语对齐任务，我们引入具有强大语义建模能力的预训练语言模型BERT来更好地进行语义匹配，在此基础上，我们采用动态规划的方法使得模型从全局的角度来匹配文本，将动态规划目标和强化学习方法结合使得模型能够尽可能提高双语对齐的覆盖度。我们构建了古文与现代文对齐测试集数据，并验证了我们提出的方法的有效性，相较于传统方式难以解决的多对多对齐模式，有着明显的提升。

但是，本文模型还存在着不足，虽然预训练语言模型有着强大的性能，在语义匹配能力上还是存在优化的空间，对于多对多的对齐模式，长句+短句的情况下，容易出现漏匹配的情况，在未来的研究中会更加关注如何更好的提升模型的语义匹配能力。除此之外，也会对如何更加有效的获取和利用句对齐的粗粒度的平行语料进行进一步的研究。

第二十一届中国计算语言学大会论文集，第704页-第715页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

致谢

本成果受国家自然科学基金项目（61872402）,教育部人文社科规划基金项目(17YJAZH068)，北京语言大学校级项目（中央高校基本科研业务费专项资金）（21YBB19，18ZDJ03），模式识别国家重点实验室开放课题基金资助。

参考文献

D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computer Science*.

A. Conneau, G. Lample, R. Rinott, A. Williams, Samuel R Bowman, H. Schwenk, and V. Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *Conference on Empirical Methods in Natural Language Processing*.

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

Duchi, John, Hazan, Elad, Singer, and Yoram. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*.

Vincent Francois-Lavet, Peter Henderson, Riashat Islam, MarcG Bellemare, and Joelle Pineau. 2018. *An Introduction to Deep Reinforcement Learning*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

F Grégoire and P. Langlais. 2017. A deep neural network approach to parallel sentence extraction.

T. He, Z. Zhang, H. Zhang, Z. Zhang, and M. Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Z. Lin and X. Wang. 2007. Chinese ancient-modern sentence alignment. In *Computational Science - ICCS 2007, 7th International Conference, Beijing, China, May 27 - 30, 2007, Proceedings, Part II*.

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)*.

X. Ouyang, S. Wang, C. Pang, Y. Sun, and H. Wang. 2020. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora.

Matthew Peters, M. Neumann, M. Iyyer, M. Gardner, and L. Zettlemoyer. 2018. Deep contextualized word representations.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Richard S Sutton, David A McAllestern, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation.

T. Utsuro, M. Yamane, Y. Matsumoto, and M. Nagao. 1997. Bilingual text matching using bilingual dictionary and statistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv*.

C. Yong, Z. Tu, F. Meng, J. Zhai, and Yang Liu†. 2018. Towards robust neural machine translation.

Z. Zhang, W. Li, and Qi Su. 2018. Automatic transferring between ancient chinese and contemporary chinese.

M. Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. *icml*.

(c) 2022 中国中文信息学会计算语言学专业委员会

B. Zoph, D. Yuret, J. May, and K. Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

刘颖and 王楠. 2013. 古汉语与现代汉语句子对齐研究. 计算机应用于软件, 30(11):4.

韦希林. 2019. 基于深度学习的双语长句分割方法研究. Ph.D. thesis, 北京交通大学.

CCL 2022

基于情感增强非参数模型的社交媒体观点聚类

刘勘
中南财经政法大学
信息与安全工程学院
/ 武汉430073
liukan@zuel.edu.cn

陈昱
中南财经政法大学
信息与安全工程学院
/ 武汉430073
chen997@stu.zuel.edu.cn

何佳瑞
中南财经政法大学
信息与安全工程学院
/ 武汉430073
hejiarui@stu.zuel.edu.cn

摘要

本文旨在使用文本聚类技术，将社交媒体文本根据用户主张的观点汇总，直观呈现网民群体所持有的不同立场。针对社交媒体文本模式复杂与情感丰富等特点，本文提出使用情感分布增强方法改进现有的非参数短文本聚类算法，以高斯分布建模文本情感，捕获文本情感特征的同时能够自动确定聚类簇数量并实现观点聚类。在公开数据集上的实验显示，该方法在多项聚类指标上取得了超越现有模型的聚类表现，并在主观性较强的数据集中具有更显著的优势。

关键词： 观点分析；短文本流聚类；非参数模型；社交媒体

A Sentiment Enhanced Nonparametric Model for Social Media Opinion Clustering

LIU Kan
School of Information
and Safety Engineering,
Zhongnan University of
Economics and Law
/ Wuhan 430073
liukan@zuel.edu.cn

CHEN Yu
School of Information
and Safety Engineering,
Zhongnan University of
Economics and Law
/ Wuhan 430073
chen997@stu.zuel.edu.cn

HE Jiarui
School of Information
and Safety Engineering,
Zhongnan University of
Economics and Law
/ Wuhan 430073
hejiarui@stu.zuel.edu.cn

Abstract

Based on text clustering techniques, this paper aims to aggregate social media texts according to the different opinions claimed by users. To address the characteristics of short length, large number and complex patterns of social media texts, this paper proposed Sentiment Distribution Enhanced (SDE) method to improve the existing nonparametric-based clustering algorithm. We model text sentiment with a Gaussian distribution, the proposed method automatically determines the number of clusters and achieves opinion clustering while capturing sentiment features. Experiments on public datasets show that the method achieves clustering performance that surpasses existing models on multiple clustering metrics, and has more significant advantages in datasets with strong subjectivity.

Keywords: opinion analysis , short text stream clustering , nonparametric model , social media

1 引言

©2022 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家自然科学基金(72174156)

社交媒体作为自由互联的平台，有着大量的观点从中产生及传播，吸引到了各领域的众多关注。微观层面上，这些观点表现于用户的各类互动行为中，表达了其态度和立场，从宏观的角度看，某事件中的观点集合在一定程度上反映出了舆情。充分挖掘、分析社交媒体平台中丰富的观点信息对网络舆情的引导和治理有着积极的意义。观点分析的研究主要围绕观点倾向性展开，即要求算法能够自动地判别用户发表言论是正面、负面还是中立的观点(Chen et al., 2020)。然而，用户对于事件或话题的观点往往具有更丰富的内容，仅分析观点的倾向不足以全面了解用户的态度和立场。因此，从用户主观言论中总结并提取核心观点成为观点分析亟待解决的问题之一。尤其在社交媒体平台的热点事件或话题中，存在着表达了不同主张的海量用户发言，将其按照观点类别正确划分有助于厘清事件态势，并能从相应的角度直观分析民众诉求和关注点。但是，在归类用户观点的过程中面临着以下困难：（1）用户发言文本长度短，处理长文档的文本挖掘技术难以应用；（2）文本数量动态增长，需要采用能够无限增量处理数据的模型；（3）用户观点模式复杂且无法预先估计，获取实际数据标签成本较大，传统的监督学习方法不再适用。因此，观点挖掘多作为一种聚类问题进行分析。但是普通的文本聚类方法，如Zhao等(2011)基于隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）提出的Twitter-LDA等需要事先设定聚类数量的方法并不符合社交媒体真实情况。

面对上述难点，诸多研究者试图从短文本数据流聚类的方向上解决社交媒体观点聚类问题。其中较经典的研究有Shou等(2013)针对社交媒体的特性设计了Sumblr系统不断监听主题相关推特数据流并进行聚类和摘要，为用户提供主题相关的公众观点总结。还有Yin等(2016)基于非参数贝叶斯模型（Bayesian nonparametric model）提出了GSDPMM，从词的共现特征（co-occurrence）出发，通过概率生成过程灵活地推断聚类数量并完成聚类。

现有的短文本流聚类研究大多仅通过短文本的词特征实现聚类，而社交媒体数据具有更多的特征，比如用户情感就是其中需要重点关注的特征之一。有许多观点分析的研究者假设用户所持看法、立场与其发表文本内容所蕴含的情感具有较强的联系，甚至通过直接或间接的方式将文本情感与观点二者联系起来(O'Connor et al., 2010; 高俊峰and 黄微, 2019)。因此，为了更好地实现社交媒体用户观点聚类，本文使用情感分析的方式挖掘文本观点特征，提出情感分布增强（Sentiment Distribution Enhanced, SDE）方法改进狄利克雷过程混合模型（Dirichlet Process Mixture Model, DPMM）聚类算法。该算法基于非参数模型，使用高斯分布建模文本情感以捕获社交媒体用户言论中的主观信息，融合使用词-情感生成分布实现聚类。

总结本文的贡献有：（1）基于非参数概率模型，提出使用高斯分布建模情感信息以增强聚类算法，解决社交媒体观点聚类难点。（2）在原有的坍缩吉布斯采样（collapsed Gibbs sampling）算法基础上，给出了情感高斯分布的后验参数更新方式以及预测分布（Predictive Distribution）计算。（3）实验证明，SDE方法相比于现有的前沿算法在聚类表现上有所提升，并在主观性较强的社交媒体数据集上取得了更显著的表现。

2 相关研究

观点分析是社交媒体研究领域的经典问题，已有诸多学者对该问题从不同方面进行了探索，本节除了介绍常用的观点分析模型，还重点从短文本数据流的角度介绍社交媒体的用户观点聚类研究。

2.1 观点分析

观点分析，又称观点挖掘或文本意见挖掘，对于社交媒体这一关键的网络舆论平台有重要的现实意义，因而受到了众多研究者的关注。现有研究大多旨在判别用户所发表言论的观点极性(Benkhelifa and Laallam, 2018; Wu et al., 2020)。对观点进行极性分析，其优点是可以转化为分类问题从而采用有监督学习模型，准确率较高，但缺点在于仅从态度倾向或情感标签的角度分析过于简单，难以呈现观点的丰富内涵。由此，有学者使用文本聚类的方法挖掘同类观点，分析民众对于热点事件或话题的各种态度(Ni et al., 2018)。李秀霞等人(2016)针对此问题采取"密度-距离"的快速搜索聚类算法进行共词聚类，该算法可以使数据自动确定聚类中心和数目而不需要人工设定。然而，当新数据到来时此算法也必须重新在完整数据集上运行，而且聚类过程受事先设定的密度阈值影响较大，无法很好地适应社交媒体数据持续增长、分布复杂的实际情况。

2.2 短文本流聚类

面对社交媒体用户观点聚类中的挑战，短文本数据流聚类方法显示出更强的现实场景适应性(Aggarwal, 2013; Nguyen et al., 2015)，其中的工作可分为基于相似度和基于模型的方法。

2.2.1 基于相似度的方法

基于相似度的短文本数据流聚类原理是将文本根据特征映射到向量空间中，并设计向量的相似性度量，在扫描文本数据时计算向量相似度，当对应的相似度大于设置阈值时聚为一类。这种自聚合的聚类机制无需人工设定聚类数量，如Geng等(2020)选择将社交媒体短文本按照其词频、长度等统计信息表示为向量，再计算相应的文本与聚类之间相似度，最终实现聚类。Rakib 等人(2021)改进了此类聚类算法的计算过程，在聚类时通过动态采样一部分聚类计算相似度而无须遍历所有聚类，一定程度上减少了计算成本。

这类基于向量和阈值的方法虽运算较快，但需要在不同数据集上搜索得出最佳阈值，且难以适应数据分布变化较大的数据集。因而该类方法在时间跨度较大的动态数据集中表现不佳。

2.2.2 基于模型的方法

基于模型的短文本数据流聚类方法假设短文本是由概率模型所生成的，文本组成的词汇由聚类相应的词分布抽样得来。其中具有代表性的是Yin等(2014)首先提出的基于狄利克雷过程的混合模型GSDPMM，该模型利用短文本的单词共现信息完成聚类，假设文本之间包含相同的词越多则越可能属于同一类。

基于狄利克雷过程混合模型的短文本聚类方法进一步地克服了固定阈值这一缺陷，能够根据数据分布情况灵活推断聚类数量并完成聚类。很多学者都是在狄利克雷过程混合模型的框架上不断创新，如Yin等(2018)改进了推断算法，使其能以“一遍扫描”（one-pass）的方式处理数据，无需多次迭代使聚类算法收敛，相较于需要迭代求解的GSDPMM更符合社交媒体的观点聚类。Kumar等(2020)深入挖掘了文本特征，在词共现关系之外寻得词的重要性特征，提升了聚类算法的表现。Li等(2016)利用词嵌入（word embedding）表征文本之间的深层相似关系。

短文本流聚类主要以文本与词的统计信息来衡量文本与聚类的相似性，而实际在社交媒体的观点表达中还包含了用户丰富的情感信息。因此本文提出了情感分布增强方法，为社交媒体短文本情感建立概率分布并将其加入生成模型中，与聚类-词的多项式分布共同作为文本的联合生成分布，在推断混合模型参数的同时得到用户观点聚类。

3 问题描述

3.1 概念定义

本文的研究旨在设计一种短文本数据流聚类算法将社交媒体中表达相似观点的用户言论聚合为一类，以直观地分析总结网民看待事件的不同角度。从形式上定义短文本数据流，其中$D=\{d_1,d_2,\ldots,d_\infty\}$是无限长度的数据序列，正如社交媒体平台上不断增加的海量用户发言。其中d_i是长度为$|d_i|$的用户发表文本，由其所包含的词$W^{(i)}=\left\{w_1^{(i)},w_2^{(i)},\ldots,w_{|d_i|}^{(i)}\right\}$组成。同时，每条言论$d_i$都有唯一对应的聚类（即观点）编号$c_i$。本文的最终目标是不断地将社交媒体用户言论分配到对应的聚类$z_k=\{d_{k1},d_{k2},\ldots,d_{k\infty}\}$，使得持同一种观点的文本汇总到一起，即$z_k$内文本聚类编号满足$c_{k1}=c_{k2}=\cdots=k$。此时观点聚类数量是无限的，这是因为随着相关事件或话题发展，新的观点有可能不断出现，但通常观点数量远小于评论数量，有$|Z|\ll|D|$。同时，假设用户评论为短文本，一条文本仅属于一个聚类簇，当$i\neq j$时有$z_i\cap z_j=\emptyset$，若是实际评论长度较长，可以通过拆分等方式处理。

3.2 狄利克雷过程混合模型文本聚类

狄利克雷过程（Dirichlet Process, DP）是一种随机过程，它每次抽样的结果都是一个概率分布。狄利克雷过程在非参数贝叶斯模型中被广泛运用，常作为混合模型的先验，形成狄利克雷过程混合模型(Li et al., 2019)。将混合模型应用于聚类过程中就可以自动地从数据推断聚类簇的数量，无需人工指定类别。

假设观测数据$X=x_1,x_2,\ldots,x_n$相互独立，来自一个具有K个未知形式成分的混合分布，记为$\mathrm{F}(\Phi)$，其中$\Phi=\phi_1,\phi_2,\ldots,\phi_k$是各成分的参数集合，因此有$p\left(X\mid\phi_1,\cdots,\phi_K;\pi_1,\cdots,\pi_K\right)=\sum_{i=1}^{K}\pi_i\,\mathrm{F}\left(X\mid\phi_i\right)$，$\pi_i$代表第i个分布在混合模型中的权重，

满足$\sum_{i=1}^{K}\pi_i=1$。然而在社交媒体等许多现实情况中，只存在用户言论可以作为观测数据，却无从预测可能的分布数量K并推断分布参数和权重。为此，需要建模成分数量无限的混合模型，引入狄利克雷过程作为参数Φ 的先验，定义为$G\sim \mathrm{DP}(\alpha,H)$ 。其中G是由集中参数为α，基分布为H的狄利克雷过程抽样得到的分布。

通常将狄利克雷过程的常见构造方式比喻为“折棍构造法”(stick-breaking construction) (Zhou et al., 2011)。对长度为1 的棍子按比例ξ_1 折断，保留比例为$1-\xi_1$ 的部分并记被切割的长度为π_1，而后对剩余的棍子切除比例为ξ_2 的部分，记其长度为$\pi_2\ldots$ 如此重复以获得一系列长度为π_i 的木棍。利用贝塔分布(Beta Distribution)的性质，以$\xi_i\sim \mathrm{Beta}(1,\alpha)$ 的方式抽样切割比例并保证$0<\xi_i<1$ 。这个过程就是折棍构造，即$\pi_i\sim \mathrm{GEM}(\alpha)$ (GEM 代表Griffiths, Engen和McCloskey)，可以得到$\pi_i=\xi_i\prod_{i=0}^{i-1}(1-\xi_i)$，满足$\sum_{i=1}^{\infty}\pi_i=1$。通过折棍构造有式(1)，其中，当$x=0$ 时有$\delta(x)=1$，其他情况下$\delta(x)=0$ 。

$$G(\phi)=\sum_{k=1}^{\infty}\pi_k\delta\left(\phi-\phi_k\right),\phi_k\sim H \tag{1}$$

根据折棍构造，从连续的基分布H中抽样得到相同的参数是可行的，这代表着狄利克雷过程可以生成无限成分的混合分布，且不同观测数据对应的分布参数可能相同，因此其具备聚类的能力，并能随着数据增长创建新的聚类。

同时，聚类需要考虑到文本间的相似度，因而引入文本建模中常见的狄利克雷-多项式共轭（conjugate）关系。使用多项式分布建模文本数据D，此时F(Φ)就被确定为多项式分布，假设文本由聚类对应的多项式词分布独立生成，即$p\left(d\mid\theta_c\right)=\prod_{w\in d}\mathrm{Mult}\left(w\mid\theta_c\right)$，所以这里的混合分布参数集合$\phi$仅包含多项式分布参数$\theta$ 。这样, 词共现关系较强的文本更可能被聚为一类。综上，用于短文本聚类的狄利克雷过程混合模型生成过程如下式所示(Yin et al., 2018) 。

$$\pi\mid\alpha\sim\mathrm{GEM}(1,\alpha) \tag{2}$$

$$c_i\mid\pi\sim\mathrm{Mult}(\pi)\quad i=1,\ldots,\infty \tag{3}$$

$$\theta_k\mid\beta\sim\mathrm{Dir}(\beta)\quad k=1,\ldots,\infty \tag{4}$$

$$d_i\mid c_i,\{\theta_k\}_{k=1}^{\infty}\sim p\left(d_i\mid\theta_{c_i}\right)=\prod_{w\in d}\mathrm{Mult}\left(w\mid\theta_{c_i}\right) \tag{5}$$

其中，α和β均为超参数，Mult($\cdot$)和Dir($\cdot$)分别代表了多项式分布和狄利克雷分布（Dirichlet distribution）。狄利克雷过程混合模型文本聚类算法概率图如图1所示。

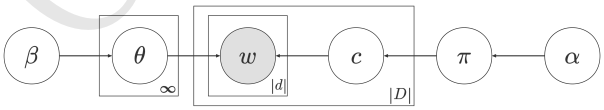


图 1: 狄利克雷过程混合模型概率图

4 情感增强的狄利克雷过程混合模型

4.1 模型结构

情感分布增强方法的目标是在前文所述的狄利克雷过程混合模型中以概率分布形式整合文本情感与词特征，联合学习词-情感分布参数，从而更好地表征用户观点并实现聚类。其中的核心问题为情感值计算、模型结构设计和参数推断过程。

社交媒体情感分析的研究认为，评论等带有情感色彩的主观性文本表现了用户在发出该言论时的情绪状态，其中可能蕴含了用户的个人态度。为了挖掘与表示文本的情感信息，学者们采用了不同的技术将自然语言量化为情感。常见的情感表示法有两类：（1）将情感按照极性分

类为积极、消极和中性等，用标签代表文本对应的情感(Liu, 2012)。该方法将情感看作离散的随机变量，可以视为服从多项式分布。（2）使用实数值表示文本情感的强度，实数的正负分别代表着情感积极或消极的倾向(Zadeh, 2015)。此处情感作为连续的随机变量，依据中心极限定理使用高斯分布建模是合理的。同时，高斯分布的概率密度集中于均值附近，该性质也符合聚类用户言论中相似情感的需求。考虑到连续的情感强度值可以利用在观点情感时序分析等下游任务中，本文选择使用情感值作为文本的情感特征，并使用高斯分布建模。

基于概率生成模型的聚类方法通常假设文本由分布所生成，通过计算文本数据从分布产生的概率进行聚类。情感增强的狄利克雷过程混合模型使用词分布和情感分布联合作用，同时衡量文本的词共现关系和情感相似性完成聚类，其工作过程如图2所示。

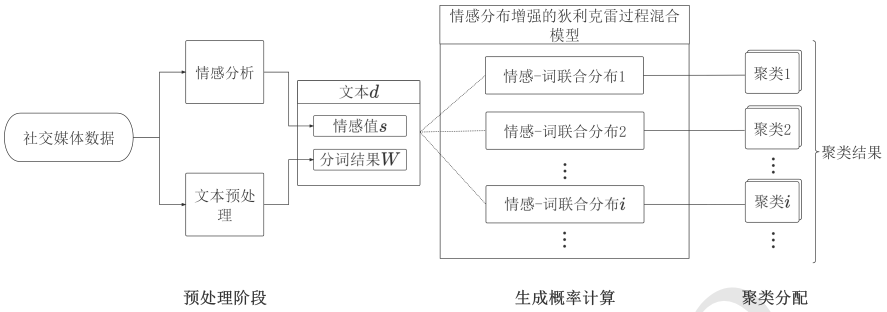


图 2: 基于情感增强非参数模型的观点聚类流程

设每一条短文本d_i 都有唯一对应的情感强度值s_i，其中实数值s_i 越接近正/负无穷则表示文本在积极/消极的情感上拥有越高的强度。因此定义文本情感s_i 由一个高斯分布生成。结合狄利克雷过程混合模型文本聚类框架和情感分布，得到文本生成的似然函数如式(6)。

$$\mathrm{F}\left(d_i \mid \phi_k\right)=\operatorname{Mult}\left(W^{(i)} \mid \theta_k\right) \mathcal{N}\left(s_i \mid \mu_k^s, \sigma_k^s\right) \tag{6}$$

在该似然函数中，此时$\phi_k=\left(\theta_k, \mu_k^s, \sigma_k^s\right)$ 为第k 个聚类(观点) 对应概率分布的参数集合。$\operatorname{Mult}\left(W^{(i)} \mid \theta_k\right)=)=\prod_{w \in d} \operatorname{Mult}\left(w \mid \theta_k\right)$表示在观点聚类$z_k$ 中，每条用户言论中所包含的词都是从聚类相应的多项式分布中依据词袋假设(bag-of-word) 独立生成的。多项式分布的参数θ 其实是一个长度为$|V|$ 的向量，其中V 代表的是算法已处理的所有词汇集合。高斯分布$\mathcal{N}\left(s_i \mid \mu_k^s, \sigma_k^s\right)$ 则拟合了聚类中的文本情感信息，μ_k^s 和σ_k^s 分别是高斯分布的均值和标准差。

接下来需要为上述模型参数设置先验。根据贝叶斯理论，若构造的先验与似然是共轭分布，则它们对应的后验分布与先验将是同一类型的分布。此性质能在狄利克雷过程混合模型参数推断阶段时简化积分式的计算。现已知多项式分布的共轭先验是狄利克雷分布$\mathrm{Dir}(\cdot)$，而高斯分布有多种的共轭先验。在分布参数μ 与σ 都未知的情况下，单维高斯分布的共轭先验可以为高斯逆卡方分布(Normal-inverse-chi-squared Distribution)，记作$\mathrm{Ni}\,\chi^2(\cdot)$ 。本文采用二者的结合设置似然函数的共轭先验，将其定义为式(7)。

$$G_0\left(\Phi_k\right)=\operatorname{Dir}\left(\theta_k \mid \beta_0\right) \mathrm{Ni}\,\chi^2\left(\mu_k^s, \sigma_k^s \mid \Psi_0\right) \tag{7}$$

其中β_0 和$\Psi_0=\left(m_0, \lambda_0, v_0, \epsilon_0^2\right)$ 是先验组成部分中狄利克雷分布与高斯逆卡方分布的参数。在先验分布G_0 中，这些参数均为超参数。基于以上过程图3展示了情感分布增强狄利克雷过程混合模型的概率图，被虚线框起的部分即为文本情感由分布生成的过程。

4.2 参数推断

本文参考经由Yin 等(2018)改进的坍缩吉布斯采样算法，进行混合模型的参数推断。实际上，混合模型的参数推断过程复杂且计算量大，不过在聚类过程中仅需要关注文本d_i所属类别编号c_i，因而可以简化计算，无需求解所有参数。

文本聚类编号的分配由后验预测分布$p\left(c_i=k \mid c_{-i}, d, \Phi\right)$ 确定，它可以根据贝叶斯法则被表示为先验分布与似然函数的乘积，如式(8)。其中，将d展开为d_i和d_{-i}以适应预测分布的形

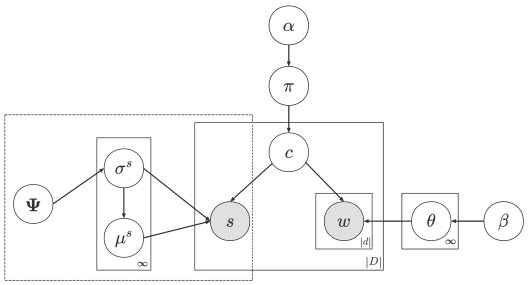


图 3: 情感分布增强狄利克雷过程混合模型概率图

式，并根据条件概率公式从原式变换为第二项，利用D分离（D-Separation）的性质得到最后的结果(Bishop and Nasrabadi, 2006)。

$$p\left(c_i \mid c_{-i}, d, \Phi\right) \propto p\left(c_i \mid c_{-i}, d_{-i}, \Phi\right) p\left(d_i \mid c, d_{-i}, \Phi\right) \propto \left(c_i \mid c_{-i}, \Phi\right) p\left(d_i \mid d_{-i}, c, \Phi\right) \tag{8}$$

经典的中餐馆过程（Chinese Restaurant Process，CRP）(Ferguson, 1973)直观地描述了聚类的过程，该过程说明了假若已知其他评论文本的类别，那么文本i所属观点类别的条件概率$p\left(c_i \mid c_{-i}, \Phi\right)$如式(9)，此条件概率即为式（8）中第一项的展开。

$$p\left(c_i \mid c_{-i}, \Phi\right) = p\left(c_i \mid c_{-i}, \alpha\right) \quad = \begin{cases} \frac{\left|z_{k_i - i}\right|}{|d| + \alpha|d| - 1} \text{(属于已有类的概率)} \\ \frac{\alpha|d|}{|d| + \alpha|d| - 1} \text{(属于新类的概率)} \end{cases} \tag{9}$$

其中$|d|$ 是包含第i 条数据在内，已输入算法的所有文本数量。而$|z_{k,-i}|$ 是第k 个聚类中除去第i 条数据的成员数量。与静态的文本聚类不同，动态的数据流聚类需要配合变量$|d|$，使用$\alpha|d|$ 代替式(2)中的α 作为混合模型成分权重的生成参数。

在计算文本由各聚类生成的概率时，为了便于推断需要假设变量之间的条件独立性，进而$p\left(d_i \mid d_{-i}, c, \Phi\right)$ 可以被分解为式(10) 。

$$p\left(d_i \mid d_{-i}, c, \Phi\right) \quad \propto p\left(W^{(i)} \mid d_{-i}, c, \Phi\right) p\left(s_i \mid d_{-i}, c, \Phi\right) \tag{10}$$

因此式(10)可以由混合模型中的词分布与情感分布分别计算。对于$p\left(s_i \mid d_{-i}, c, \Phi\right)$，它作为建模文本情感强度的后验预测分布，被高斯分布和其共轭先验所确定。根据Murphy (2007)的推导，该共轭关系在积分式中可以简化，计算方式详见式(11) 。

$$p\left(s_i \mid d_{-i}, c, \Phi\right) = \iint \mathcal{N}\left(s_i \mid \mu_k^s, \sigma_k^s\right) Ni\chi_2\left(\mu_k^s, \sigma_k^s \mid \Psi_{k,-i}\right) d\mu_k^s d\sigma_k^s = \mathrm{t}_{v_n}\left(m_n, \frac{\left(1+\lambda_n\right)\epsilon_n^2}{\lambda_n}\right) \tag{11}$$

式(11)中$\mathrm{t}_{v_n}(\cdot)$ 是自由度为v_n 的t 分布概率密度函数，根据该函数就可以计算得到文本情感由各聚类生成的概率，达到为文本分配聚类的目的。在推导过程中，出现的变量$\Psi_{k,-i} = \left(m_n, \lambda_n, v_n, \epsilon_n^2\right)$ 其实是对应于聚类z_k 中逆卡方分布的后验分布参数，在每次成功为文本d_i 指定聚类编号$c_i = k$ 后都需要使用该文本的情感强度值更新先验，即不断通过数据所确定的后验以修正聚类先验。文本情感分布的先验逆卡方分布参数更新方式(Murphy, 2012) 如式(12)，其中n代表用于更新后验的数据量，$\bar{s}$ 是当前聚类内情感均值。

$$\begin{aligned} \lambda_n &= \lambda_0 + n, \quad m_n = \frac{\lambda_0 m_0 + n\bar{s}}{\lambda_n} \\ v_n &= v_0 + n, \quad v_n \epsilon_n^2 = \left(v_0 \epsilon_0^2 + \sum_i \left(s_i - \bar{s}\right) + \frac{n\lambda_0}{\lambda_0 + n}\left(m_0 - \bar{s}\right)^2\right) \end{aligned} \tag{12}$$

有关聚类概率表达式$p\left(d_i \mid d_{-i}, c, \Phi\right)$ 的另一部分，即词分布的后验预测分布$p\left(W^{(i)} \mid d_{-i}, c, \Phi\right)$，同样利用多项式分布-狄利克雷分布共轭的性质推导得到文本词汇由各聚类生成

的概率(Yin and Wang, 2016; Xu et al., 2021), 得到式(13) 。

$$p\left(W^{(i)} \mid d_{-i}, c, \Phi\right) = \frac{\prod_{w \in d_i} \prod_{j=1}^{(w)} \left(f_{z_k,-d_i}^{(w)} + \beta + j - 1\right)}{\prod_{l=1}^{d_i|} \left(|v|_{z_k,-d_i} + |V|\beta + l - 1\right)} \tag{13}$$

$f_{d_i}^{(w)}$ 和$f_{z_k,-d_i}^{(w)}$ 分别指的是词w 在文本d_i 中的词频和在聚类z_k 里出现的次数，$|v|_{z_k,-d_i}$ 是聚类内现有词的数量。综上，包括模型参数推断在内的完整算法过程如表1 所示。本文所使用的自定义数学符号及其意义在表2中列出。

算法1: SDE算法

输入：逆卡方分布参数$m_0, \lambda_0, v_0, \epsilon_0$;文本$D = \{d_1, d_2, \ldots, d_\infty\}$

输出：文本对应的聚类编号$C = \{c_1, c_2, \ldots, c_\infty\}$

For d_i in D do

$s_i = get - sentiment - of(d_i)$//使用情感分析获取文本情感值

计算$p\left(s_i \mid d_{-i}, c, \Phi\right)$见式(11) //情感生成概率

计算$p\left(W^{(i)} \mid d_{-i}, c, \Phi\right)$见式(13) //词生成概率

$p\left(d_i \mid d_{-i}, c, \Phi\right) = p\left(W^{(i)} \mid d_{-i}, c, \Phi\right) p\left(s_i \mid d_{-i}, c, \Phi\right)$见式(10)

$p_k = \frac{|z_{k_i - i}|}{|d| + \alpha|d| - 1} \cdot p\left(d_i \mid d_{-i}, c, \Phi\right)$ //文本由已有类生成的概率

$p_{k+1} = \frac{\alpha|d|}{|d| + \alpha|d| - 1} \cdot p\left(d_i \mid d_{-i}, c, \Phi\right)$ //文本属于新产生的类

$c_i = argmax(p_k)$ //为文本分配生成概率最大的聚类

更新d_i所属聚类z_{c_i}的情感后验分布参数见式(12)

表 1: SDE算法过程

数学符号	含义	数学符号	含义
下标i	向量中的第i个元素	π	混合模型成分权重
$\lvert\cdot\rvert$	向量中的元素数量	α	狄利克雷过程集中参数
$-i$	向量中除第i个元素外所有元素	β	狄利克雷分布参数
$\bar{\cdot}$	向量均值	ξ	贝塔分布抽样结果
D, d	社交媒体文本流、单条文本	θ	词多项式分布参数
$w^{(i)}$	第i条文本的组成词	$\mu^{(s)}, \sigma^{(s)}$	情感分布均值、标准差
c	算法已分配的所有聚类编号	V, v	输入算法的词汇集、V的任意子集
s	文本情感值	$f^{(w)}$	词w 的词频
z	聚类（观点）	n	样本数据数量
Φ, ϕ	混合模型所有成分的参数集合、混合模型成分参数	$\Psi = (m, \lambda, v, \epsilon^2)$	高斯逆卡方分布参数

表 2: 本文所用数学符号及其意义对照

5 实验

5.1 数据集及评价指标

5.1.1 数据集介绍

为了便于对比，本文采用在短文本数据流聚类研究领域中广泛应用的Tweets数据集和Google-News数据集作为实验数据来源。

（1）Tweets数据集由2011至2015年TREC（Text Retrieval Conference）会议提供的推特数据构成[0] ，这些来自社交媒体平台Tweets的文本共30322条，并依据其所讨论的内容被标注

[0]http://trec.nist.gov/data/microblog

为269个不同的主题。该数据集被使用于诸多经典的文本聚类研究(Yin and Wang, 2016; Yin et al., 2018; Kumar et al., 2020; Xu et al., 2021; Chen et al., 2019)。Tweets数据集的平均文本长度为7.97个单词，较为符合社交媒体用户观点聚类的场景。

（2）Google-News数据集收集了11109篇新闻文章，合并同一事件的相关报道，共整理得到152个聚类，最终通过提取新闻标题建立数据集(Yin and Wang, 2014)。Google-News数据集的平均文本长度是6.23单词。相较于Tweets数据集，Google-News数据集作为新闻标题，其文本情感特征并不显著，因此对本文提出的情感分布增强模型更具挑战性。

经由预处理后两个数据集的统计信息如表3所示。之后本文将已集成于自然语言处理工具包[1]（Natural Language Toolkit）中的vader情感分析工具(Hutto and Gilbert, 2014)应用于预处理后的原始数据集上，以还原文本中的情感信息。

数据集	文本数量	聚类数量	词汇数量	平均长度
Tweets	30322	269	12301	7.97
Google-News	11109	152	8110	6.23

表 3: 实验数据集统计信息

5.1.2 评价指标介绍

（1）标准化互信息(Strehl and Ghosh, 2002)（Normalized Mutual Information, NMI）是评估聚类算法的最常见指标之一，若是模型结果越接近真实的聚类情况则NMI越近于1，否则NMI越近于0。

（2）聚类准确度（Accuracy，Acc）更直接地比较算法得到的文本类别标签与真实标签。

（3）聚类同质性（Homogeneity，Ho.）和聚类完整度（Completeness，Cp.）是两个不同的聚类目标(Rosenberg and Hirschberg, 2007)。同质性希望每个聚类中只包含该聚类的成员，而完整度指同一个类别的数据应当归属于同样的聚类簇。这两个指标有助于细致地衡量各算法在不同角度的表现。

（4）FMI（Fowlkes-Mallows Index）是聚类精度和召回的几何平均(Fowlkes and Mallows, 1983)。

实验结果将重点关注NMI、Accuracy与FMI，将三者作为考察不同算法表现的主要指标。

5.2 实验设计

5.2.1 对比方法

对比实验将选取短文本数据流聚类研究领域中最前沿的算法，选择依据以下三个条件：第一，聚类方式基于狄利克雷过程混合模型，才能由此比较加入情感分布后的模型表现。第二，能够以“一遍扫描”的方式处理流式文本数据，符合本文设想的社交媒体场景。第三，算法表现优于其他同类算法，且通过实验能复现原论文中的效果。最终，本文选取了下列两个模型：

（1）OSDM(Kumar et al., 2020)：OSDM是一个基于狄利克雷过程混合模型的短文本流聚类算法，仅支持通过“一遍扫描”的方式处理文本数据，它在Yin等人(2018)工作的基础上增强了词共现关系，并加入了词重要性特征，利用语义信息提升了模型表现。在真实数据集上的实验结果表明，OSDM在各个聚类评价指标中的表现都比较优秀。

（2）DP-BMM(Chen et al., 2020)：DP-BMM不同于其他研究，使用了文本中词对（Biterm）的共现关系来代替原有的单词共现特征。相比于单词，词对代表着更丰富的信息，进一步提升了聚类的表现。作为短文本聚类算法，它不仅能够以一遍扫描的方式实现聚类，也可以通过基于批处理（batch-based）的方式多次迭代处理数据。

5.2.2 模型超参数设置

对于模型OSDM与DP-BMM，本文使用原文献中提供的超参数设定以求重现算法的最优结果。对于OSDM，在两个数据集上采用相同的超参数：$\alpha = 0.002$，$\beta = 0.0004$。对于DP-BMM，在Tweets数据集上令$\alpha = 0.3$，$\beta = 0.02$，$batchsize = 1$，而在Google-News数据集上令$\alpha = 0.6$，$\beta = 0.02$，$batchsize = 1$。本文提出的文本情感分布超参数

[1]https://www.nltk.org/

为$\Psi_0 = (m_0, \lambda_0, v_0, \epsilon_0^2)$，即先验逆卡方分布的参数。通过实验，将OSDM-SDE的情感分布超参数设置为：$m_0 = 0$，$\lambda_0 = 1$，$v_0 = 1$，$\epsilon_0 = 0.01$，在Tweets数据集上将DP-BMM-SDE的超参数设置为：$m_0 = 0$，$\lambda_0 = 0.8$，$v_0 = 1$，$\epsilon_0 = 0.03$，而在Google-News数据集上令$m_0 = 0$，$\epsilon_0 = 0.1$，$v_0 = 1$，$\epsilon_0 = 0.001$。

5.3 实验结果及分析

5.3.1 对比实验结果

实验结果包含了各模型在NMI、Accuracy、Homogeneity、Completeness和FMI等五个指标下的聚类表现，总体的实验结果如表4所示，其中加粗字符表示在该指标上更优的结果。

从表4中可以看出，在整合了文本情感分布后，原有基于狄利克雷过程混合模型的短文本聚类算法在不同程度上均有所提高。情感分布改进后的模型聚类效果在NMI、Accuracy这两个需要关注的指标上都超过了原模型，通过FMI指标也可以看出本文提出的算法在类别分配结果上相较于原模型也更合理。同质性和完整度互相具备冲突的性质，本文提出的算法偏向聚类的完整性，不过二者的调和平均在数值上实际与NMI接近，因此可以看出总体上本文提出的算法更能兼顾二者。

模型	数据集									
	Tweets					Google-News				
	NMI	Acc	FMI	Ho.	Cp.	NMI	Acc	FMI	Ho.	Cp.
OSDM	0.847	0.613	0.583	**0.905**	0.796	0.812	0.617	0.535	**0.829**	0.796
OSDM-SDE	**0.852**	**0.632**	**0.637**	0.902	**0.807**	**0.815**	**0.618**	**0.558**	0.822	**0.807**
DP-BMM	0.799	0.614	0.569	0.773	**0.827**	0.838	0.684	0.635	0.825	0.851
DP-BMM-SDE	**0.801**	**0.629**	**0.578**	**0.778**	0.826	**0.840**	**0.700**	**0.643**	**0.829**	**0.853**

表 4: 对比试验结果

如表5所示，通过对比模型在两个数据集中NMI、Accuracy、FMI评价指标的表现，还可以看出经由文本情感分布增强的聚类算法在社交媒体数据集（Tweets）上的表现提升比在新闻标题数据集（Google-News）上的提升要更加显著。以OSDM为例，情感分布增强后的模型在Tweets数据集上NMI、准确度、FMI分别提升了0.59%、3.1%、9.26%，但是在News数据集上各自只提升了0.37%、0.16%、4.3%。这主要是由于社交媒体平台用户倾向于自由地发表个人言论，因此文本中所蕴含的主观情感色彩较为浓郁。而在新闻标题中则恰恰相反，报道者旨在客观提炼事件信息，用词较为中立。这验证了本文提出的文本情感分布增强方法能够更好地处理情感特征，并在社交媒体用户观点聚类的场景下提升聚类算法的表现。

模型	NMI提升		准确度提升		FMI提升	
	Tweets	Google-News	Tweets	Google-News	Tweets	Google-News
OSDM-SDE	**0.59%**	0.37%	**3.10%**	0.16%	**9.26%**	4.30%
DP-BMM-SDE	**0.25%**	0.23%	**2.44%**	2.34%	**1.58%**	1.26%

表 5: SDE改进模型在不同数据集上的提升对比

5.3.2 数据规模对算法的影响

将Tweets数据集按照3000的步长逐点输出聚类结果，观察数据规模对情感分布增强方法的影响。以OSDM为例，实验结果如表6所示。以NMI、聚类准确度、FMI指标作为评判标准，使用情感分布增强的聚类算法在各数据测试点上取得了超越原模型的表现。观察数据量较少时模型的聚类结果，可以看出SDE方法在小数据量时也一定程度上提升了聚类准确度、NMI与FMI。这项实验结果说明在结合了文本的词特征与情感特征后，能够改进模型在数据较稀疏时的聚类能力，有效利用了社交媒体文本潜藏的用户主观信息。

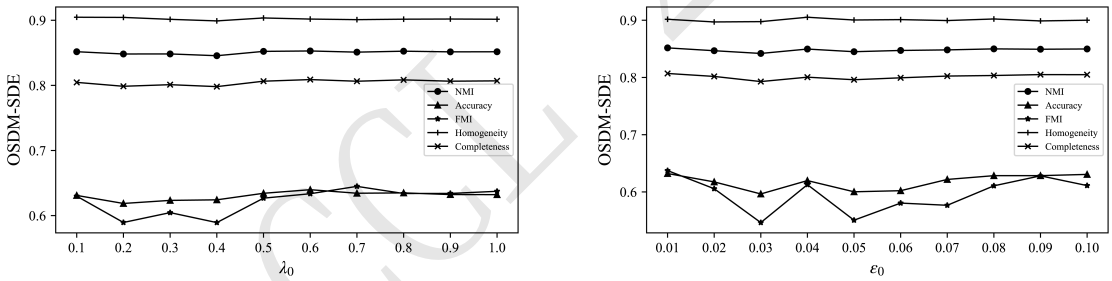
指标	模型	3000	6000	9000	12000	15000	18000	21000	24000	27000	30000
Acc	OSDM	0.601	0.59	0.594	0.621	0.629	0.635	0.637	0.647	0.63	0.614
	+SDE	**0.631**	**0.597**	**0.602**	**0.634**	**0.643**	**0.649**	**0.654**	**0.659**	**0.647**	**0.641**
NMI	OSDM	0.814	0.803	0.807	0.832	0.841	0.848	0.853	0.857	0.852	0.848
	+SDE	**0.818**	**0.803**	**0.81**	**0.834**	**0.844**	**0.851**	**0.857**	**0.86**	**0.856**	**0.853**
FMI	OSDM	0.629	0.585	0.578	0.598	0.629	0.639	0.644	0.655	0.625	0.586
	+SDE	**0.641**	**0.589**	**0.591**	**0.621**	**0.654**	**0.655**	**0.661**	**0.67**	**0.653**	**0.633**

表 6: 不同数据规模的准确度、NMI、FMI(OSDM-SDE)

5.3.3 参数敏感性分析

在使用逆卡方分布Ni$\chi^2(\cdot)$ 作为情感分布$\mathcal{N}(s \mid \mu^s, \sigma^s)$ 的先验时，需要输入其先验逆卡方分布的参数$\Psi_0 = \left(m_0, \lambda_0, v_0, \epsilon_0^2\right)$ 。根据分布定义Ni$\chi^2\left(m_0, \lambda_0, v_0, \epsilon_0^2\right) = \mathcal{N}\left(\mu^s \mid m_0, \sigma^s/\ \lambda_0\right) \times \chi^{-2}\left(\sigma^s \mid v_0, \epsilon_0^2\right)$，可得超参数$\Psi_0$ 所分别对应的解释含义为：m_0 是高斯分布参数μ^s 的先验均值，ϵ_0^2 是参数σ^s 先验分布的缩放参数, λ_0 和v_0 表示了对先验的信任程度。实际应用中通常采取弱信息先验假设，将先验的信任程度设定为一个较小的值(Chipman et al., 2001), 因此本文令λ_0 不大于1 。而v_0 则常被设置为与数据变量维度相同，赋值$v_0 = 1$ 。此外，在建模高斯变量时的一个常见选择是将其均值设为0 ，即$(m_0 = 0)$ 。综上，输入参数中需要人为设定的仅有λ_0 与ϵ_0 两个变量。本小节以OSDM为例，在Tweets数据集下进行敏感性分析。

固定其他参数，分别使λ_0 和ϵ_0在对应取值范围变化，参数敏感性分析实验结果如图4所示。随着λ_0 的变动可以观察到，表示聚类算法表现的NMI、准确度、同质性和完整度指标都相对稳定，仅能在FMI 指标曲线上能观测到一些波动。这主要是因为在弱信息先验假设下，λ_0 的值在较小的区间内变动，对聚类过程的作用并不明显。与λ_0 类似, ϵ_0 的变动也没有对本方法的最终表现造成较大波动，NMI、聚类准确度等指标依然稳定。



(a) (b)

图 4: λ_0(a)，ϵ_0(b)在Tweets上的敏感性分析(OSDM-SDE)

6 结论

本文提出了使用文本情感分布增强（SDE）的方法改进短文本数据流聚类模型，旨在解决社交媒体用户观点聚类中文本长度短、数据量未知，以及缺乏观点数量先验知识等难题，利用文本数据中含有的主观情感信息更有效地实现观点聚类。本文将情感量化为数值，并将其建模为高斯分布，加入狄利克雷过程混合模型中作为文本生成的联合分布，用坍缩吉布斯采样算法推断混合模型参数，同时推导了情感分布的参数更新方式。在真实数据集上的实验结果表明，本文提出的SDE聚类方法在NMI、FMI、聚类准确度等方面均超越了现有模型，验证了SDE方法的合理性和有效性。通过对比社交媒体数据集与新闻数据集上的聚类结果，可以发现使用文本情感分布增强不仅提升了现有模型的聚类效果，还能显著地增进模型在具有较强情感色彩数据集上的表现，符合本文假设的社交媒体用户观点聚类场景。未来的研究方向包括利用社交媒体流式数据的时间顺序信息提升模型的表现以及在聚类结果的基础上文本摘要的自动抽取等。

参考文献

李秀霞and 邵作运. 2016. “密度-距离” 快速搜索聚类算法及其在共词聚类中的应用. 情报学报, 35(4):380–388.

高俊峰and 黄微. 2019. 网络舆情场中观点簇丛的情感极化度测算. 图书情报工作, 63(10):106.

Charu C Aggarwal. 2013. A survey of stream clustering algorithms.

Randa Benkhelifa and Fatima Zohra Laallam. 2018. Opinion extraction and classification of real-time youtube cooking recipes comments. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 395–404. Springer.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Junyang Chen, Zhiguo Gong, and Weiwen Liu. 2019. A nonparametric model for online topic discovery with word embeddings. *Information Sciences*, 504:32–47.

Junyang Chen, Zhiguo Gong, and Weiwen Liu. 2020. A dirichlet process biterm-based mixture model for short text stream clustering. *Applied Intelligence*, 50(5):1609–1619.

Hugh Chipman, Edward I George, Robert E McCulloch, Merlise Clyde, Dean P Foster, and Robert A Stine. 2001. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134.

Thomas S Ferguson. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Edward B Fowlkes and Colin L Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Fei Geng, Qilie Liu, and Ping Zhang. 2020. A time-aware query-focused summarization of an evolving microblogging stream via sentence extraction. *Digital Communications and Networks*, 6(3):389–397.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Jay Kumar, Junming Shao, Salah Uddin, and Wazir Ali. 2020. An online semantic-enhanced dirichlet model for short text stream clustering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 766–776.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.

Yuelin Li, Elizabeth Schofield, and Mithat Gönen. 2019. A tutorial on dirichlet process mixture modeling. *Journal of mathematical psychology*, 91:128–144.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Kevin P Murphy. 2007. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Hai-Long Nguyen, Yew-Kwong Woon, and Wee-Keong Ng. 2015. A survey on data stream clustering and classification. *Knowledge and information systems*, 45(3):535–569.

Ningning Ni, Caili Guo, and Zhimin Zeng. 2018. Public opinion clustering for hot event based on br-lda model. In *International Conference on Intelligent Information Processing*, pages 3–11. Springer.

Brendan O’Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media*.

Md Rashadul Hasan Rakib, Norbert Zeh, and Evangelos Milios. 2021. Efficient clustering of short text streams using online-offline clustering. In *Proceedings of the 21st ACM Symposium on Document Engineering*, pages 1–10.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 533–542.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Peng Wu, Xiaotong Li, Si Shen, and Daqing He. 2020. Social media opinion summarization using emotion cognition and convolutional neural networks. *International Journal of Information Management*, 51:101978.

Wanyin Xu, Yun Li, and Jipeng Qiang. 2021. Dynamic clustering for short text stream based on dirichlet process. *Applied Intelligence*, pages 1–12.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.

Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 625–636. IEEE.

Jianhua Yin, Daren Chao, Zhongkun Liu, Wei Zhang, Xiaohui Yu, and Jianyong Wang. 2018. Model-based clustering of short text streams. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2634–2642.

Amir Zadeh. 2015. Micro-opinion sentiment intensity analysis and summarization in online videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 587–591.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer.

Jian Ying Zhou, Fei Yue Wang, and Da Jun Zeng. 2011. Hierarchical dirichlet processes and their applications: a survey. *Zidonghua Xuebao/Acta Automatica Sinica*, 37(4):389–407.

CCL 2022

Discourse Markers as the Classificatory Factors of Speech Acts

Da Qi[1], **Chenliang Zhou**[1], **Haitao Liu**[*,2,✉]

[1]Department of Linguistics, Zhejiang University, China
[2]Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, China
{da.qi,cl.zhou}@zju.edu.cn, lhtzju@yeah.net

Abstract

Since the debut of the speech act theory, the classification standards of speech acts have been in dispute. Traditional abstract taxonomies seem insufficient to meet the needs of artificial intelligence for identifying and even understanding speech acts. To facilitate the automatic identification of the communicative intentions in human dialogs, scholars have tried some data-driven methods based on speech-act annotated corpora. However, few studies have objectively evaluated those classification schemes. In this regard, the current study applied the frequencies of the eleven discourse markers (*oh, well, and, but, or, so, because, now, then, I mean*, and *you know*) proposed by Schiffrin (1987) to investigate whether they can be effective indicators of speech act variations. The results showed that the five speech acts of *Agreement* can be well classified in terms of their functions by the frequencies of discourse markers. Moreover, it was found that the discourse markers *well* and *oh* are rather efficacious in differentiating distinct speech acts. This paper indicates that quantitative indexes can reflect the characteristics of human speech acts, and more objective and data-based classification schemes might be achieved based on these metrics.

1 Discourse Markers and the (Dis)agreement Continuum

A discourse marker (DM) is a word or phrase that people often use in the process of communication, and its main function is to coordinate and organize discourse to ensure the smooth flow of conversation. In addition, as a carrier of pragmatic information, it usually reflects speakers' mental states and communicative intentions, thus facilitating pragmatic inference (Furkó, 2020). In this regard, Fraser (1996, p.68) defined DMs as "linguistically encoded clues which signal the speaker's potential communicative intentions." Although scholars have never reached a consensus on the definition of DMs, no one would doubt their diverse discursive functions and the capability to transmit communicative intentions.

When analyzing the functions of DMs, scholars also differ considerably in terms of their frameworks and research paradigms. Ariel (1998) distinguished DMs from a semantic perspective: a DM either possesses a semantic meaning, which is interpreted in a particular context with some connection to its form (e.g., *and* and *I mean*); or it does not contain any semantic information (e.g., *well* and *oh*). However, Matei (2010) pointed out that although some DMs contain rich semantic information, there are particular contexts in which the communicative intention it conveys is not related to the semantic information it carries. For example, in some cases, the DM *and* can be used as a discourse continuative, filler word, and buffer term, etc.

Some scholars analyzed the range of functions through the functional-cognitive approach, which shows that DMs have a specific rather than a completely arbitrary range of functions (Redeker, 2006; Fischer, 2006). Ariel (1998, pp.242-243) also expressed support for the non-arbitrary nature of DM functions. She explicated this view in terms of the correspondence between form and function and argued that there are two probabilistically similar possibilities for the form-function correspondence, one in which a form corresponds to multiple functions and the other in which a function corresponds to

✉ Corresponding Author
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 728-737, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

multiple forms. She further claimed that these two possible relationships do not indicate the syntactic arbitrariness but are characterized by unpredictability since the same form may evolve to express many innovative meanings. In this sense, functionalists argue that the universality of DM forms (as opposed to the uniformity of forms) is functionally driven.

The above investigations of DM functions have helped us to gain a deep and broad understanding of DMs' nature and their functional orientations in various contexts. However, as Matei (2010) mentioned, there is a great deal of uncertainty in DMs' functions, and even those with a relatively fixed semantic meaning may produce new and rare uses in some contexts. In addition, the *one form – many functions* and *one function – many forms* nature of DMs, as well as the innovative nature of their functions, also make their functions perform in a variety of ways. Thus, it is difficult to assess all the functions of DMs through an in-depth analysis of the discourse material one by one (the workload is too large). If we want to characterize all aspects of certain DMs and explore the patterns of these linguistic units full of uncertainties and probabilities, it is better to apply an approach that is suitable for approximating all the features possessed by the DMs.

Another consideration in employing this approach is that human communicative intentions are themselves fraught with probabilities and uncertainties. As pointed out by the Speech Act Theory, there is not always a clear correspondence between the words people express and their functions, and speech acts are also characteristic of *one form – many functions* and *one function – many forms* as mentioned by Ariel (1998) (Holtgraves, 2005). A more extreme example, such as *Kennst du das Land wo die Zitronen blühen?* (Knowest thou the land where the lemon trees bloom?), can even express the communicative intention "I am a German soldier" (Searle, 1969). By the same token, the various DMs proposed by previous authors, such as the eleven DMs by Schiffrin (1987) (*oh, well, and, but, or, so, because, now, then, I mean*, and *you know*), may occur in various speech acts depending on the specific speech context.

To address the function of DMs and the probability and uncertainty of human communicative intentions, the present paper tries to introduce some basic probabilistic and statistical methods, such as the hierarchical cluster analysis (HCA), to quantitatively analyze the DMs contained in specific communicative intentions. Our aim is to examine whether certain indicators of DM (e.g., their percentage of frequency of occurrence in different speech acts) can effectively distinguish the communicative functions embodied in differing speech acts to propose a new research methodology for DM-related studies.

In the current study, the frequency of different DMs in differing speech acts was investigated as a possible defining feature for the distinction of communicative intentions. The reason for doing so is that DMs carry diverse pragmatic and contextual information (Redeker, 2006). In this regard, the frequency of DMs may reflect the pragmatic characteristics of different speech acts, which may help us better explore the patterns of human communicative intentions.

Since we want to examine whether the frequency of DMs can effectively distinguish different speech acts, these DMs should first be able to reflect the differences between speech acts that differ significantly, e.g., agreement and disagreement, thanks and apology, etc. Next, we may examine whether it can reflect the slight differences between similar speech acts. Therefore, in the current paper, we applied the continuum of (dis)agreement (*accept, partially accept, hedge, partially reject,* and *reject*) as the object of study to explore whether the frequency distribution of DMs can accurately capture the nuanced differences in speech acts.

When analyzing the agreement-disagreement continuum, scholars have mostly focused on the perception of agreement- and disagreement-like speech acts by people in specific types of discourses. For example, Mulkay (1985) found that strong disagreement is easier to declare in writing than face to face after examining the written letters by biochemists. When investigating the arguments of mentally disabled people, Hewitt et al.'s (1993) study showed that regarding conflict resolution as the primary goal of arguments detracts from the true nature of verbal conflicts – they reflect a social continuum of agreement and disagreement (Jacobs and Jackson, 1981). Trimboli and Walker (1984) compared dyadic discussions following initial agreement and disagreement and found that disagreement was more competitive, characterized by high rates of verbalization, increased numbers of turns, more frequent interruptions, and reduced back channels.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 728-737, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

From the studies above, it can be seen that agreement and disagreement are complex and influenced by various socio-cultural factors, but the specific mechanisms of their intricacies have been seldom studied, and a more systematic and comprehensive understanding has yet to be developed. In the current study, we attempted to employ the frequency of DMs as well as probabilistic and statistical methods to examine the speech acts of agreement and disagreement, complementing the existent findings in discourse analysis.

The research questions of this paper are as follows.

1. How is the frequency distribution of different DMs under the differing speech acts in the agreement-disagreement continuum?

2. Can the frequency of DMs effectively reflect the similarity and peculiarity of the different speech acts?

2 Methods and Materials

2.1 The Hierarchical Cluster Analysis (HCA)

The HCA is an algorithm for clustering the given data. It regards all the data input as a single cluster and then recursively divides each cluster into two subclasses. It enjoys a relatively long history in the study of communicative intentions, including the Speech Act Theory. In the 1960s, scholars had already proposed that human communicative behavior could be structured hierarchically (Scheflen, 1965; Scheflen, 1967). Some researchers then innovatively employed hierarchical organizations for speech acts to analyze specific types of discourse, e.g., therapeutic discourse (Labov and Fanshel, 1977) or interpersonal behavior (D'Andrade and Wish, 1985).

Furthermore, some pragmaticians in recent years started to analyze the data in their experiments with the HCA, especially when they probed into the relationship between existing classificatory schemes and people's perception of a given set of speech acts (Holtgraves, 2005; Liu, 2011). Though word frequency and other textual indices were not applied in their studies, it can be revealed that the HCA may be effective in speech act-related research.

As DMs indicate contextual information and pragmatic relationship, their frequency of use in utterances could be seen as an indicator of speech act. It is then plausible to examine whether these objective indices can be hierarchically clustered in a way that demonstrates the functional similarities and variations between different speech acts.

2.2 The Switchboard Dialog Act Corpus (SwDA)

The SwDA consists of 1,155 five-minute conversations, including around 205,000 utterances and 1.4 million words from the Switchboard corpus of telephone conversations (Jurafsky et al., 1997; Potts, 2022). The dialogs in this corpus all happened between two individuals of different ages, genders, and education levels, and the speech acts of speakers were annotated according to how participants might expect one sort of conversational units to be responded to by another. One of the SwDA's merits is that there can be more than one speech act within each utterance. This annotation scheme perfectly corresponds with the ideas of Labov and Fanshel, who criticized the one-utterance-to-one-speech act method of identifying speech acts in dialogs (Labov and Fanshel, 1977). In this regard, the results obtained through the SwDA may be an accurate reflection of the speech act patterns in human beings' daily dialogs.

According to Jurafsky et al. (1997), there are four sets of speech act hyper-categories that have enough data and meaningful sub-categories – *Agreement, Understanding, Answer,* and *Information Request*. With the 27 kinds of speech acts and the 11 DMs in the four hyper-categories, statistical tests can be conducted to get reliable results. Moreover, traditional speech act classifications such as Searle's (1976), though important, may have some defects, e.g., their abstractness and the overemphasis on speakers. Thus, the SwDA can serve as an ideal research material by virtue of the following attributes.

First, the corpus makes a more detailed and clear distinction between the speech acts of agreement and disagreement. According to Jurafsky et al.'s (1997) classification criteria, speech acts expressing speakers' attitudes are distinguished into a continuum containing five subcategories – direct approval (Agree/Accept), partial approval (Maybe/Accept-part), hold before positive answers (Hedge), partial

Proceedings of the 21st China National Conference on Computational Linguistics, pages 728-737, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

negation (Dispreferred Answers/Reject-part), and direct negation (Reject). All of them were annotated based on Allen and Core's (1997) decision tree (see Figure 1), which helped control the subjectivity and the disagreements of the annotators.

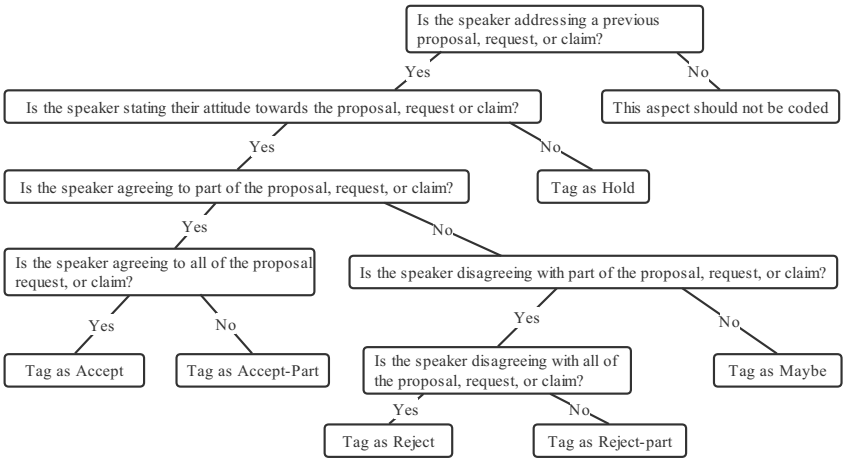


Figure 1: The decision tree for annotating speech acts in the agreement-disagreement continuum.

Second, the SwDA was annotated based on a shallow discourse tag set, which can reduce the abstractness of the speech acts owing to the more direct description of human communicative intentions. In addition to that, eight labelers involved in the project spent about half an hour on labeling each conversation (the conversations lasted five minutes on average). The labeling accuracy and the impact of labelers' subjectivity was evaluated by the *Kappa statistic* (Carletta, 1996; Carletta et al., 1997; McHugh, 2012), and the average pair-wise *Kappa* was .80, which indicated that the annotating results were acceptable (Jurafsky et al., 1997).

We could thus explore not only whether the frequencies of DMs can effectively distinguish speakers' affective attitudes through the HCA, but also whether they can distinguish properties such as the degree of indirectness in communicative acts.

3 Results and Discussions

In this paper, the DM system (*oh, well, and, but, or, so, because, now, then, I mean,* and *you know*) proposed by Schiffrin (1987) was employed to explore whether the frequencies of DMs can reflect the affinity relationship between different speech acts. This system, containing commonly occurring DMs and widely accepted by the academic community, can capture how the diverse DMs with distinctive functions demonstrate the similarities and peculiarities among different communicative intentions. As mentioned in Section 2, the current study analyzed the five speech acts in the SwDA that express agreement or disagreement because they present a typical continuum, which facilitates a more detailed examination of the results of data analysis.

The original text files of the five speech acts were firstly compiled. The five speech acts of *Agreement* contain altogether 24,816 words, among which Agree/Accept has 19,942 words, Maybe/Accept-part 528 words, Hedge 2,703 words, Dispreferred Answers 1,772 words, and Reject 942 words). Next, we applied Antconc 4.0.5 (Anthony, 2021) to automatically get the total word count in the five speech acts, in which the frequency data of the eleven selected words/phrases proposed by Schiffrin (1987) were extracted. Since the automatic process could not distinguish between the 11 words/phrases as DMs and other cases, the authors manually checked the automatically collected data to obtain the exact DM frequencies for follow-up analyses.

It should be noted here that the raw number of DM frequencies may affect the results of the statistical analysis due to the large variation in the total number of words in each speech act. In this regard, this paper calculated the percentages of each DM's frequency relative to the total word number to standardize the data. When converting the numbers to percentages, the authors distinguished two different kinds of

Proceedings of the 21st China National Conference on Computational Linguistics, pages 728-737, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

DMs: one for single words (unigram), such as *oh, well*, etc., and the other for two consecutive words (bigram), *I mean* and *you know*. For the former types of DMs, we counted the percentages of DM frequencies relative to those of all unigrams in each speech act; for the latter, we calculated those of all bigrams, with the aim of making the standard uniform.

After collation and calculation, the frequency data of the eleven DMs proposed by Shiffrin (1987) under each speech act were obtained, as shown in Table 1.

Speech act	oh	well	and	but	or	so	because	now	then	I mean	you know
Agree/Accept	0.0318	0.0117	0.0036	0.0015	0.0003	0.0011	0.0004	0.0004	0.0000	0.0031	0.0005
Maybe/Accept-part	0.0019	0.0455	0.0019	0.0019	0.0038	0.0000	0.0000	0.0000	0.0000	0.0000	0.0019
Hedge	0.0137	0.0396	0.0133	0.0074	0.0007	0.0067	0.0011	0.0037	0.0007	0.0033	0.0004
Dispreferred Answers	0.0045	0.0796	0.0034	0.0028	0.0000	0.0017	0.0000	0.0017	0.0006	0.0051	0.0006
Reject	0.0149	0.0658	0.0042	0.0138	0.0032	0.0011	0.0011	0.0021	0.0011	0.0053	0.0053

Table 1: The proportion of each DM to the total unigrams or bigrams under each speech act.

As can be seen from Table 1, there are significant differences in the proportion of DMs under each speech act, especially the difference between the proportion of *well* in the speech act of agreement and that of disagreement, in which the frequency of *well* is significantly higher than that in the speech act of agreement. In addition, the frequency of *well* in indirect speech acts is higher than that in the direct ones (Dispreferred Answers > Reject > Maybe/Accept-part > Hedge > Agree/Accept). This pattern may indicate a face-saving strategy at work in the politeness principle.

The following excerpts from the SwDA further illustrate the differences between agreement and disagreement as well as those between direct and indirect speech acts.

A. *Dispreferred Answers*
1) Well, I, I think, uh, my background is probably what absolutely turned me off with sixty minutes.
2) Well, I heard tonight on the news that he is willing to come down.
3) Well, I, I, I come from kind of a biased opinion because I'm a, a therapist and a drug and alcohol.
4) Well, that was, you know, with a, with a circular saw.

From the utterances containing *well* in the speech act of Dispreferred Answers, we can see that *well* mainly serves to provide a buffer for the subsequent words. In addition, since the speaker wants to express opposition to the words spoken by the hearer without completely opposing them, he or she tends to use the strategy of repetition (e.g., the repetition of *I* in A. 1) and A. 3)) or continue to apply other DMs as filler words to further moderate the illocutionary force of the speech act of opposition (e.g., *you know* in A. 4)). This phenomenon shows that people would frequently resort to the buffer DM *well* along with other means to minimize the force of opposition they are expressing.

B. *Reject*
1) Well, I don't think you can mail thing, guns through the mail.
2) Well, I doubt that.
3) Well, yes.

When expressing direct opposition to another speaker's opinions, the frequency of *well* is also higher due to the principle of politeness and the consideration of face-saving strategy. Although Reject and Agree/Accept are both direct speech acts, the use of buffer words like *well* in direct disagreement is still significantly higher than that in direct agreement (Reject: 0.0658 > Agree/Accept: 0.0117).

C. *Maybe/Accept-part*
1) Well, even if it's not technical. If it's, uh, some social thing or whatever. It doesn't matter.

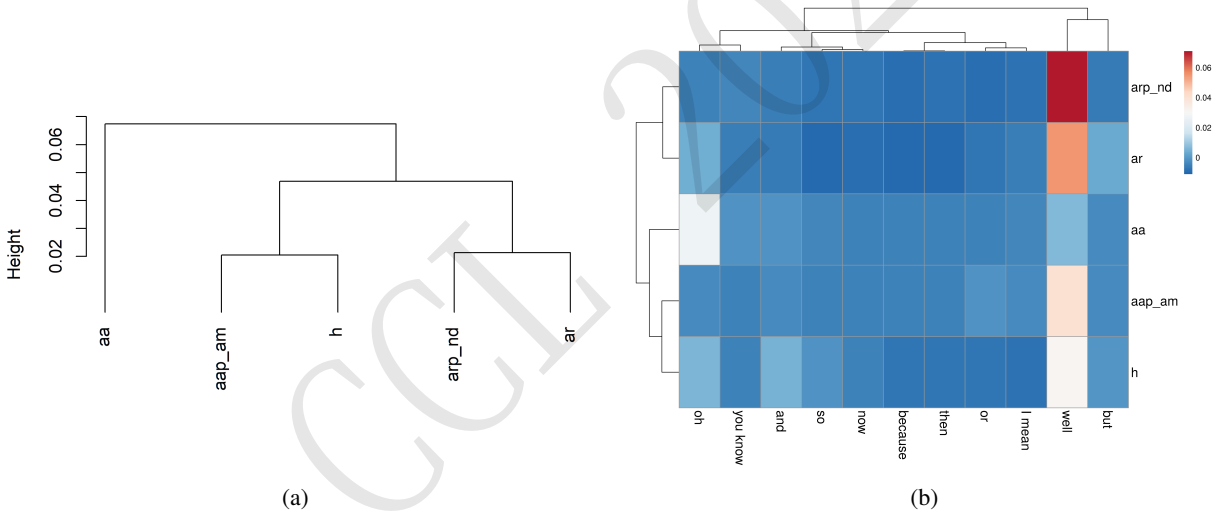
D. *Hedge*
1) Well, uh, it's funny, when I tried, to make the call the other days,

Proceedings of the 21st China National Conference on Computational Linguistics, pages 728-737, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

E. *Agree/Accept*
1) Oh, well yeah.
2) Well, that's true.

Among the three speech acts concerning agreement (Agree/Accept, Maybe/Accept-part, and Hedge), the use of *well* is more convergent, serving as a simple tone buffer, and does not involve a strategy of face protection for the other interlocutor. According to previous studies on *well*, it is often employed as a delay device and a pragmatic marker of insufficiency, indicating the problems with the content of the current or the previous utterances, or as a face-threat mitigator, showing the conflicts in the interpersonal level (Jucker, 1993). Although *well* has a relatively fixed spectrum of discourse functions, its frequency of occurrence varies across discourses expressing different communicative intentions, depending on the specific context and the nature of probability within speakers' language use. Therefore, to accurately capture how the frequency of *well* in different speech acts reflects their affinities, it is best to apply a more suitable method to study these probabilistic linguistic units.

Moreover, from the above analysis, *well* is a DM that can effectively distinguish between agreement and disagreement; however, people cannot merely use *well* when expressing these communicative functions; DMs such as *you know* and *and* also frequently occur in these speech acts. In order to comprehensively and systematically grasp how the frequency of DMs reflects the differences of each speech act, we included in the present study a more comprehensive DM system (that proposed by Schiffrin). Meanwhile, to avoid the overwhelming workload caused by manual qualitative analysis, we adopted established statistical methods to grasp the characteristics embodied in DMs accurately. The *factoextra* and *cluster* packages in *R* were applied to perform an HCA on the data in Table 1. The results are shown in Figure 2.



(a) (b)

Figure 2: (a) The HCA results of the five speech acts in *Agreement*. (b) The HCA results of speech acts (in rows) using the Manhattan distance and the Ward.D2 method. *aa* is referred to as Agree/Accept, *aap_am* is Maybe/Accept-part, *h* is Hedge, *arp_nd* is Dispreferred Answers, and *ar* is Reject.

Figure 2a demonstrates that the clustering results based on the frequency of the eleven DMs neatly reflect the functions of the five speech acts under the *Agreement* hyper-category. The results show a tripartite classification, with Dispreferred Answers and Reject clustered together, Maybe/Accept-part and Hedge in the same cluster, and Agree/Accept in a separate cluster out of the above four speech acts. Hence, we can roughly get a "reject" cluster and an "accept" one in *Agreement*. Nevertheless, this result still has some imperfections: Agree/Accept is clustered out of the other four speech acts, while its function is similar to the "accept" category. After trying different method-distance combinations of the HCA, it was found that the aforementioned classification enjoys the highest probability of occurrence.

We then further altered the combination of clustering methods and distances and found that using the Manhattan distance together with the Ward.D2 method produced a clustering result consistent with the

Proceedings of the 21st China National Conference on Computational Linguistics, pages 728–737, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

functional division of the speech acts in *Agreement* (see Figure 2b)[0]. Moreover, the top panel in Figure 2b displays the clustering result of each DM based on their frequency of use. The cluster of *well* and *but* further corroborates our previous analysis of *well*'s frequent appearance in the speech acts concerning disagreement.

After obtaining the above clustering results, we employed the *cluster* package in *R* to get the proportion of each DM in each cluster for a more detailed analysis. The distribution of DMs' frequency proportions when there are two clusters (henceforth Type A clustering) and three ones (henceforth Type B clustering) are shown in Table 2 and Table 3, respectively.

Cluster	oh	well	and	but	or	so	because	now	then	you know	I mean
A	0.0158	0.0322	0.0063	0.0036	0.0016	0.0026	0.0005	0.0014	0.0002	0.0021	0.0009
B	0.0097	0.0727	0.0038	0.0083	0.0016	0.0014	0.0005	0.0019	0.0008	0.0052	0.0029

Note: Cluster A is the cluster of Agree/Accept, Maybe/Accept-part, and Hedge, and Cluster B is Dispreferred Answers and Reject.

Table 2: The percentages of DM frequencies in different clusters (Type A).

From the data in Table 2 and Table 3, the reason Agree/Accept is separated as an individual cluster in Figure 2a is probably because *oh* appears significantly more frequently in it than in other speech acts. After analyzing the original corpus data, it was found that *oh* usually appears in expressions such as "Oh yes" or "Oh yeah", which constitute a typical feature of Agree/Accept compared with other speech acts. Also, the high frequency of *oh* indicates that most clustering methods are influenced by individual salient values, which lead to the changes in specific cluster branches. In addition, the results in Table 2 and Table 3 show that the frequency of *well* is significantly higher when it expresses negative views than when it expresses positive ones. Since the other DMs accounted for lower frequencies and contributed less to the clustering results, the results obtained in this study may indicate that the two DMs, *well* and *oh*, are more effective in distinguishing between the speech acts of agreement and disagreement. This result also further complements the previous studies on the principle of politeness and the face theory, providing new perspectives for future systematic research on pragmatic principles with large-scale corpus data.

Cluster	oh	well	and	but	or	so	because	now	then	you know	I mean
C	0.0318	0.0117	0.0036	0.0015	0.0003	0.0011	0.0004	0.0004	0.0000	0.0031	0.0005
D	0.0078	0.0425	0.0076	0.0046	0.0023	0.0033	0.0006	0.0018	0.0004	0.0017	0.0011
E	0.0097	0.0727	0.0038	0.0083	0.0016	0.0014	0.0005	0.0019	0.0008	0.0052	0.0029

Note: Cluster C is the cluster of Agree/Accept, Cluster D is Maybe/Accept-part and Hedge, and Cluster E is Dispreferred Answers and Reject.

Table 3: The percentages of DM frequencies in different clusters (Type B).

In summary, from the above analysis, it can be concluded that by employing a method that can accurately grasp the statistical patterns of linguistic units, we may be able to better capture the tendency of each speech act in using DMs and establish the connection between the two important constructions (*speech acts* and *DMs*) with the support from real data. This approach can complement well-developed qualitative analyses of DMs, provide more comprehensive and theoretically supported results (e.g., the DM classification system proposed by Schiffrin), and introduce the advantages of quantitative analysis (big data, objectivity, and accuracy) into the research related to pragmatics and discourse analysis.

4 Conclusions and Implications

In this study, we adopted a quantitative approach to analyze whether DMs, the discourse units that possess discursive and pragmatic information, can effectively distinguish the speech acts of different communicative functions. After calculating the frequencies of the 11 DMs proposed in Schiffrin (1987), we

[0]For all the clustering results using different methods and distance metrics, see Appendix A.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 728–737, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

conducted an HCA using *R* for the examination of such effects. The results showed that the frequencies of DMs were efficacious in differentiating the speech acts of agreement and disagreement. Moreover, the frequencies of DMs also well reflect the intricacies within the indirectness of the five speech acts in the agreement-disagreement continuum, corroborating that DMs are rather precise indicators of speech acts' differences.

The results also indicated that the frequencies of *well* and *oh* might be the key indicators to distinguish between the speech acts of agreement and disagreement, especially *well*, the frequencies of which echo the previous findings in the principle of politeness and the face theory. In this regard, the application of quantitative measures for testing and generalizing the existent theoretical framework may help the research related to pragmatics and discourse analysis develop in a scientific and precise direction. The deficiencies of traditional qualitative research in terms of data size can thus be supplemented by conducting research on the authentic data from large-scale corpus.

In addition, since the current study only examined the five speech acts under the continuum of agreement and disagreement, the patterns found may not fully reflect the patterns in all types of speech acts. Therefore, subsequent studies can further collect the natural corpus data of human conversations and examine more types of speech acts to further explore the effectiveness of DM frequency in reflecting human conversational behaviors. In this way, we may establish a more comprehensive framework for quantitative research in pragmatics and discourse analysis.

Acknowledgements

The authors are grateful to the three anonymous reviewers for providing helpful feedback on this paper.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. http://www.fb10.uni-bremen.de/anglistik/ling/ss07/discourse-materials/DAMSL97.pdf.

Laurence Anthony. 2021. Antconc 4.0.5. Waseda University.

Mira Ariel. 1998. Discourse markers and form-function correlations. In Andreas H. Jucker and Yael Ziv, editors, *Discourse Markers: Descriptions and Theory*, pages 223–260. John Benjamins, Philadelphia.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Roy Goodwin D'Andrade and Myron Wish. 1985. Speech act theory in quantitative research on interpersonal behavior. *Discourse Processes*, 8(2):229–259.

Kerstin Fischer. 2006. Frames, constructions, and invariant meanings: The functional polysemy of discourse particles. In Kerstin Fischer, editor, *Approaches to Discourse Particles*, pages 427–447. Elsevier, Oxford.

Bruce Fraser. 1996. Pragmatic markers. *Pragmatics*, 6(2):167–190.

Péter B. Furkó. 2020. *Discourse Markers and Beyond: Descriptive and Critical Perspectives on Discourse-Pragmatic Devices across Genres and Languages*. Springer International Publishing, Cham.

Lynne E. Hewitt, Judith F. Duchan, and Erwin M. Segal. 1993. Structure and function of verbal conflicts among adults with mental retardation. *Discourse Processes*, 16(4):525–543.

Thomas Holtgraves. 2005. The production and perception of implicit performatives. *Journal of Pragmatics*, 37(12):2024–2043.

Scott Jacobs and Sally Jackson. 1981. Argument as a natural category: The routine grounds for arguing in conversation. *Western Journal of Speech Communication*, 45(2):118–132.

Andreas H. Jucker. 1993. The discourse marker well: A relevance-theoretical account. *Journal of Pragmatics*, 19(5):435–452.

CCL-2022

Daniel Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13. https://web.stanford.edu/˜jurafsky/ws97/manual.august1.html.

William Labov and David Fanshel. 1977. *Therapeutic Discourse: Psychotherapy As Conversation*. Academic Press, New York.

Si Liu. 2011. An experimental study of the classification and recognition of chinese speech acts. *Journal of Pragmatics*, 43(6):1801–1817.

Mădălina Matei. 2010. Discourse markers as functional elements. *Bulletin of the Transilvania University of Braşov*, 3(52):119–126.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

Michael Mulkay. 1985. Agreement and disagreement in conversations and letters. *Text - Interdisciplinary Journal for the Study of Discourse*, 5(3):201–228.

Christopher Potts. 2022. Switchboard dialog act corpus with penn treebank links. https://github.com/cgpotts/swda.

Gisela Redeker. 2006. Discourse markers as attentional cues at discourse transitions running head: Discourse transitions. In Kerstin Fischer, editor, *Approaches to Discourse Particles*, pages 339–358. Brill, Leiden, The Netherlands.

Albert E. Scheflen. 1965. *Stream and Structure of Communicational Behavior: Context Analysis of a Psychotherapy Session*. Eastern Pennsylvania Psychiatric Institute.

Albert E. Scheflen. 1967. On the structuring of human communication. *American Behavioral Scientist*, 10(8):8–12.

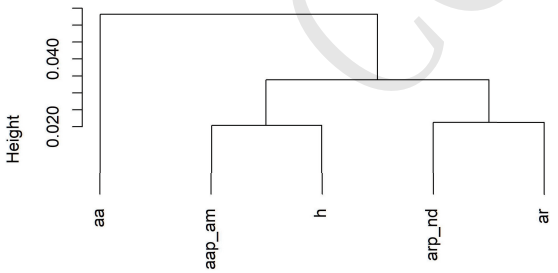
Deborah Schiffrin. 1987. *Discourse Markers*. Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.

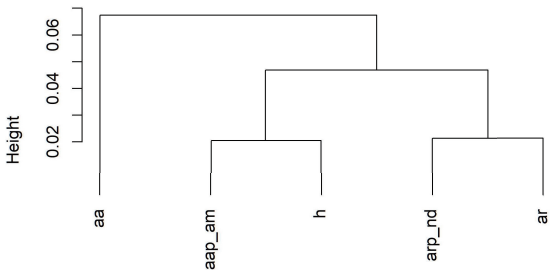
John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5(1):1–23.

Carmelina Trimboli and Michael B Walker. 1984. Switching pauses in cooperative and competitive conversations. *Journal of Experimental Social Psychology*, 20(4):297–311.

Appendix A. The Clustering Results of the Frequencies of Discourse Markers in the Speech Acts of Agreement.



(a) ward.D-euclidean



(b) ward.D2-euclidean

CCL 2022

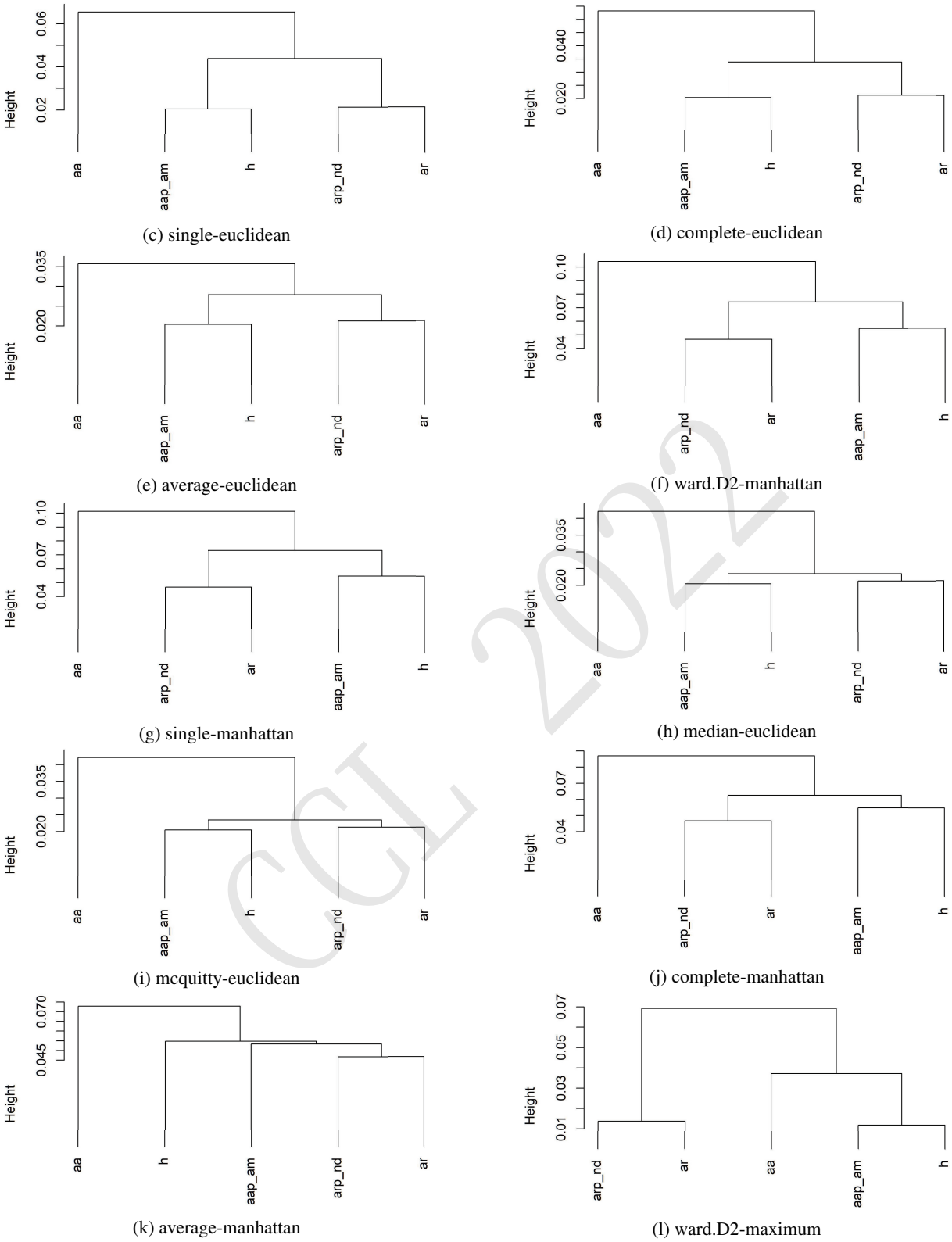


Figure 3: The clustering results of the frequencies of DMs in the speech acts of *Agreement*.

CCL 2022

DIFM:An effective deep interaction and fusion model for sentence matching

Kexin Jiang
Department of Computer Science and Technology, Yanbian University.
2020010075@ybu.edu.cn

Yahui Zhao *
Department of Computer Science and Technology, Yanbian University.
yhzhao@ybu.edu.cn

Rongyi Cui
Department of Computer Science and Technology, Yanbian University.
cuirongyi@ybu.edu.cn

Abstract

Natural language sentence matching is the task of comparing two sentences and identifying the relationship between them. It has a wide range of applications in natural language processing tasks such as reading comprehension, question and answer systems. The main approach is to compute the interaction between text representations and sentence pairs through an attention mechanism, which can extract the semantic information between sentence pairs well. However, this kind of methods fail to capture deep semantic information and effectively fuse the semantic information of the sentence. To solve this problem, we propose a sentence matching method based on deep interaction and fusion. We first use pre-trained word vectors Glove and character-level word vectors to obtain word embedding representations of the two sentences. In the encoding layer, we use bidirectional LSTM to encode the sentence pairs. In the interaction layer, we initially fuse the information of the sentence pairs to obtain low-level semantic information; at the same time, we use the bi-directional attention in the machine reading comprehension model and self-attention to obtain the high-level semantic information. We use a heuristic fusion function to fuse the low-level semantic information and the high-level semantic information to obtain the final semantic information, and finally we use the convolutional neural network to predict the answer. We evaluate our model on two tasks: text implication recognition and paraphrase recognition. We conducted experiments on the SNLI datasets for the recognizing textual entailment task, the Quora dataset for the paraphrase recognition task. The experimental results show that the proposed algorithm can effectively fuse different semantic information that verify the effectiveness of the algorithm on sentence matching tasks.

1 Introduction

Natural language sentence matching is the task of comparing two sentences and identifying the relationship between them. It is a fundamental technique for a variety of tasks. For example, in the paraphrase recognition task, it is used to determine whether two sentences are paraphrased. In the text implication recognition task, it is possible to determine whether a hypothetical sentence can be inferred from a predicate sentence.

Recognizing Textual Entailment (RTE), proposed by Dagan(Dagan and Glickman, 2004), is a study of the relationship between premises and assumptions. It mainly includes entailment, contradiction, and neutrality. The main methods for recognizing textual entailment include the following: similarity-based methods(Ren et al., 2015), rule-based methods(Hu et al., 2020), alignment feature-based machine learning methods(Sultan et al., 2015), etc. However, These methods can't perform well in recognition because they didn't extract the semantic information of the sentences well. In recent years, deep learning-based methods have been effective in semantic modeling, achieving good results in many tasks in NLP(Jin et al., 2021)(Li et al., 2021)(Yang et al., 2020). Therefore, on the task of recognizing textual entailment, deep learning-based methods have outperformed earlier approaches and become the dominant recognizing textual entailment method. For example,Bowman *et al*. used recurrent neural networks to model

*Corresponding author.
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 738-747, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

premises and hypotheses, which have the advantage of making full use of syntactic information(Bowman et al., 2015a).After that, he first applied LSTM sentence models to the RTE domain by encoding premises and hypotheses through LSTM to obtain sentence vectors(Bowman et al., 2015b). WANG *et al*.proposed mLSTM model on this basis, which focuses on splicing attention weights in the hidden states of the LSTM, focusing on the part of the semantic match between the premise and the hypothesis.The experimental results showed that the method achieved good results on the SNLI dataset(Wang and Jiang, 2016).

Paraphrase recognition is also called paraphrase detection.The task of paraphrase recognition is to determine whether two texts hold the same meaning.If they have the same meaning,they are called paraphrase pairs.Traditional paraphrase recognition methods focus on text features.However,there are problems such as low accuracy rate.Therefore,deep learning-based paraphrase recognition methods have become a hot research topic.Deep learning-based paraphrase recognition methods are mainly divided into two types; 1) calculated word vectors by neural networks,and then calculated word vector distances to determine whether they were paraphrase pairs.For example, Huang *et al*.used an improved EMD method to calculate the semantic distance between vectors and obtain the interpretation relationship(Dong-hong, 2017). 2) Directly determining whether a text pair is a paraphrased pair by a neural network model,which is essentially a binary classification algorithm.Wang *et al*.proposed the BIMPM model, which first encodes sentence pairs by a bidirectional LSTM and then matches the encoding results from multiple perspectives in both directions(Wang et al., 2017).Chen *et al*.proposed an ESIM model that uses a two-layer bidirectional LSTM and a self-attention mechanism for encoding, then it extracts features through the average pooling layer and the maximum pooling layer, and finally performs classification(Chen et al., 2017).

These models mentioned above have achieved good results on specific tasks, but most of these models have difficulty extracting deep semantic information and effectively fusing the extracted semantic information,in this paper, we propose a sentence matching model based on deep interaction and fusion .We use the bi-directional attention and self-attention to obtain the high-level semantic information.Then,we use a heuristic fusion function to fuse the low-level semantic information and the high-level semantic information to obtain the final semantic information.We conducted experiments on the SNLI datasets for the recognizing textual entailment task, the Quora dataset for the paraphrase recognition task.,The results showed that the accuracy of the proposed algorithm on the SNLI test set is 87.1%, and the accuracy of the Quora test set is 86.8%. Our contributions can be summarized as follows:

- We propose a sentence matching model based on deep interaction and fusion. It introduces bidirectional attention mechanism into sentence matching task for the first time.
- We propose a heuristic fusion function. It can learn the weights of fusion by neural network to achieve deep fusion.
- We evaluate our model on two different tasks and Validate the effectiveness of the model.

2 BIDAF model based on bi-directional attention flow

In the task of extractive machine reading comprehension, Seo *et al*.first proposed a bi-directional attention flow model BIDAF (Bi-Directional Attention Flow) for question-to-article and article-to-question(Seo et al., 2016). Its structure is shown in Figure 1.

The model mainly consists of an embed layer, a contextual encoder layer, an attention flow layer, a modeling layer, and an output layer. After the character-level word embedding and the pre-trained word vector Glove word embedding, the contextual representations X and Y of the article and the question are obtained by a bidirectional LSTM, respectively.The bi-directional attention flow between them is computed, and it proceeds as follows:

a) The similarity matrix between the question and the article is calculated.The calculation formula is shown in Eq.1.

$$K_{tj} = W^T [X_{:t}; Y_{:j}; X_{:t} \odot Y_{:j}] \tag{1}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 738-747, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

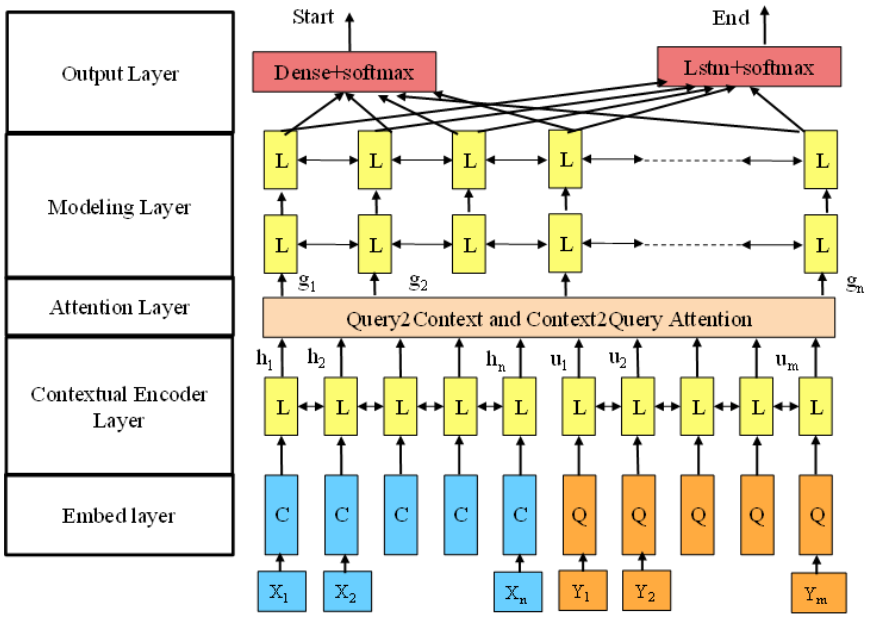


Figure 1: Bi-Directional Attention Flow Model

where K_{tj} is the similarity of the t-th article word to the j-th question word, $X_{:t}$ is the t-th column vector of X, $Y_{:j}$ is the j-th column vector of Y, and W is a trainable weight vector.

b) Calculating the article-to-question attention. Firstly, the normalization operation is performed on the above similarity matrix, and then the weighted sum of the problem vector is calculated to obtain the article-to-problem attention, which is calculated as shown in Eq.2.

$$\begin{aligned} x_t &= soft\max(K) \\ \hat{Y}_{:t} &= \sum_j x_{tj} Y_{:j} \end{aligned} \tag{2}$$

c) Query-to-context (Q2C) attention signifies which context words have the closest similarity to one of the query words and are hence critical for answering the query. We obtain the attention weights on the context words by $y = softmax(max_{col}(K)) \in R^T$, where the maximum function $\max_{col}$ is performed across the column. Then the attended context vector is $\hat{x} = \sum_t y_t X_{:t}$.This vector indicates the weighted sum of the most important words in the context with respect to the query. $\hat{x}$ is tiled T times across the column, thus giving $\hat{X} \in R^{2d*T}$.

d) Fusion of bidirectional attention streams. The bidirectional attention streams obtained above are stitched together to obtain the new representation, which is calculated as shown in Eq.3.

$$L_{:t} = \left[X_{:t}; \hat{Y}_{:t}; X_{:t} \odot \hat{Y}_{:t}; X_{:t} \odot \hat{X}_{:t}\right] \tag{3}$$

We builds on this work by looking at sentence pairs in a natural language sentence matching task as articles and problems for reading comprehension.We use the bi-directional attention and self-attention to obtain the high-level semantic information.Then,we use a heuristic fusion function to fuse the low-level semantic information and the high-level semantic information to obtain the final semantic information.

3 Method

In this section, we describe our model in detail. As shown in Figure 2, our model mainly consists of an embedding layer, a contextual encoder layer, an interaction layer, a fusion layer, and an output layer.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 738-747, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

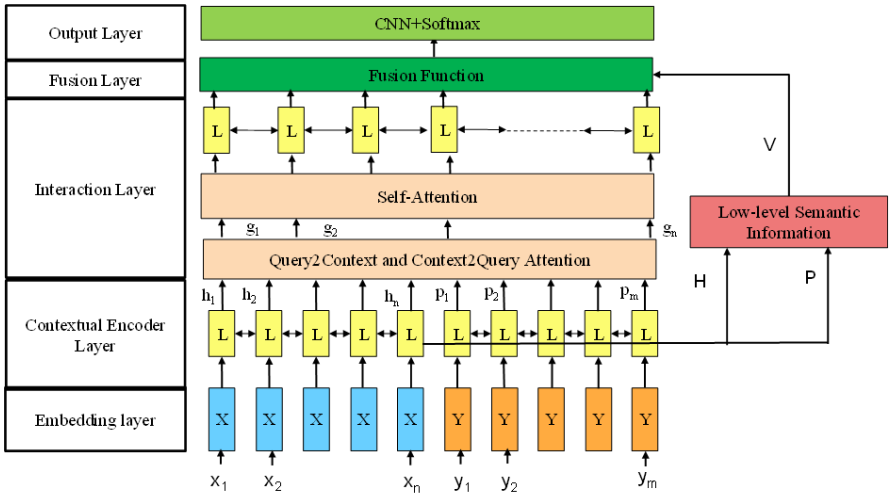


Figure 2: Overview of the architecture of our proposed DIFM model.It consists of an embedding layer, a contextual encoder layer, an interaction layer, a fusion layer, and an output layer.

3.1 Embedding Layer

The purpose of the embedding layer is to map the input sentence A and sentence B into word vectors.The traditional mapping method is one-hot encoding.However,it is spatially expensive and inefficient, so we use pre-trained word vectors for word embedding.These word vectors are constant during training.

Since the text contains unregistered words, we also use character-level word vector embedding.Each word can be seen as a concatenation of characters and characters, and then we use LSTM to get character-level word vectors. It can effectively handle unregistered words.

We assume that the pre-trained word vector for word h is h_w,and character-level word vector is h_c,we splice the two vectors and use a two-tier highway network(Zilly et al., 2017) to get the word vector representation of word h:$h = [h_1; h_2] \in R^{d_1+d_2}$,where d_1 is the dimension of Glove word embedding and d_2 is the dimension of character-level word embedding.Finally,we obtain the word embedding matrix $X \in R^{n*(d_1+d_2)}$ for sentence A and the word embedding matrix $Y \in R^{m*(d_1+d_2)}$ for sentence B, where n,m represent the number of words in sentence A and sentence B.

3.2 Contextual Encoder Layer

The purpose of the contextual encoder layer is to fully exploit the contextual relationship features of the sentences. We use bidirectional LSTM for encoding which can mine the contextual relationship features of the sentences.Then,we can obtain its representation $H \in R^{2d*n}$ and $P \in R^{2d*m}$, where d is the hidden layer dimension.

3.3 Interaction Layer

The purpose of the interaction layer is to extract the effective features between sentences.In this module,we can obtain low-level semantic information and high-level semantic information.

3.3.1 low-level semantic information

The purpose of this module initially fuses two sentences to get the low-level semantic information.We first calculate the similarity matrix S of the context-encoded information H and P,which is shown in Eq.4.

$$S_{ij} = {W_s}^T[h; p; h \odot p] \tag{4}$$

where S_{ij} denotes the similarity between the i-th word of H and the j-th word of P, W_s is weight matrices, h is the i-th column of H , and p is the j-th column of P. Then, we calculate the low-level semantic information V of A and B,which is shown in Eq.5.

$$V = P \cdot softmax(S^T) \tag{5}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 738-747, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.3.2 high-level semantic information

The purpose of this module is mine the deep semantics of the text, and to generate high-level semantic information.In this module,we frist calculate the bidirectional attention of H and P that is the attention of $H \rightarrow P$ and $P \rightarrow H$. It is calculated as follows.

$H \rightarrow P$: The attention describes which words in the sentence P are most relevant to H.The calculation process is as follows; firstly, each row of the similarity matrix is normalized to get the attention weight, and then the new text representation $Q \in R^{2d*n}$ is obtained by weighted summation with each column of P,which is calculated as shown in Eq.6.

$$\begin{aligned} \alpha_t &= softmax(S_{t:}) \in R^m \\ q_{:t} &= \sum_j \alpha_{tj} P_{:j} \end{aligned} \tag{6}$$

where $q_{:t}$ is the t-th column of Q.

$P \rightarrow H$: The attention indicates which words in H are most similar to P.The calculation process is as follows: firstly,the column with the largest value in the similarity matrix $\boldsymbol{S}$ is taken to obtain the attention weight, then the weighted sum of H is expanded by n time steps to obtain $C \in R^{2d*n}$,which is calculated as shown in Eq.7.

$$\begin{aligned} b &= softmax(\max_{col}(S)) \in R^n \\ c &= \sum_t b_t H_{t:} \in R^{2d} \end{aligned} \tag{7}$$

After obtaining the attention matrix Q of $H \rightarrow P$ and the attention matrix C of $P \rightarrow H$,we splice the attention in these two directions by a multilayer perceptron.Finally,we get the spliced contextual representation G, which is calculated as shown in Eq.8.

$$\begin{aligned} G_{:t} &= \beta(C_{:t}, H_{:t}, Q_{:t}) \\ \beta(c, h, q) &= [h; q; h \odot q; h \odot c] \in R^{8d} \end{aligned} \tag{8}$$

Then,we calculate its self-attention(Vaswani et al., 2017) ,which is calculated as shown in Eq.9.

$$\begin{aligned} E &= G^T G \\ Z &= G \cdot softmax(E) \end{aligned} \tag{9}$$

Finally, we pass the above semantic information Z through a bi-directional LSTM to obtain high-level semantic information U.

3.4 Fusion Layer

The purpose of the fusion layer is to fuse the low-level semantic information V and the high-level semantic information U. We innovatively propose a heuristic fusion function, it can learn the weights of fusion by neural network to achieve deep fusion.We fuse V and U to obtain the text representation $L = fusion(U, V) \in R^{n*2d}$, where the fusion function is defined as shown in Eq.10:

$$\begin{aligned} \widetilde{x} &= \tanh(W_1[x; y; x \odot y; x - y]) \\ g &= sigmoid(W_2[x; y; x \odot y; x - y]) \\ z &= g \odot \widetilde{x} + (1 - g) \odot x \end{aligned} \tag{10}$$

Where W_1 and W_2 are weight matrices, and g is a gating mechanism to control the weight of the intermediate vectors in the output vector. In this paper, x refers to U and y refers to V.

3.5 Output Layer

The purpose of the output layer is to output the results.In this paper, we use a linear layer to get the results of sentence matching. The process is shown in Eq.11.

$$y = softmax(\tanh(ZW + b)) \tag{11}$$

where both W and b are trainable parameters. Z is the vector after splicing its first and last vectors.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 738-747, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4 Experimental results and analysis

In this section,we validate our model on two datasets from two tasks.We first present some details of the model implementation, and secondly, we show the experimental results on the dataset. Finally, we analyze the experimental results.

4.1 Experimental details

4.1.1 Loss function

In this paper, the cross-entropy loss function can be chosen as shown in Eq.12.

$$loss = -\sum_{i=1}^{N}\sum_{k=1}^{K} y^{(i,k)} \log \hat{y}^{(i,k)} \tag{12}$$

where N is the number of samples, K is the total number of categories and $\hat{y}^{(i,k)}$ is the true label of the i-th sample.

4.1.2 Dataset

In this paper, we use the natural language inference datasets SNLI, and the paraphrase recognition dataset Quora to validate our model. The SNLI dataset contains 570K manually labeled and categorically balanced sentence pairs. The Quora question pair dataset contains over 400k pairs of data that each with binary annotations, with 1 being a duplicate and 0 being a non-duplicate.The statistical descriptions of SNLI and Quora data are shown in Table 1.

Table 1: The statistical descriptions of SNLI and Quora

dataset	train	validation	test
SNLI	550152	10000	10000
Quora	384290	10000	10000

Table 2: Values of Hyper Parameters

Hyper Parameters	Values
Glove dimension	300
Character embedding dimension	100
Hidden dimension	200
learning rate	0.0005
Optimizer	Adam
Dropout	0.2
activation function	ReLU
Epoch	30
Batch size	128

4.1.3 parameter settings

This experiment is conducted in a hardware environment with a graphics card RTX5000 and 16G of video memory.The system is Ubuntu 20.04, the development language is Python 3.7, and the deep learning framework is Pytorch 1.8.

In the model training process, a 300-dimensional Glove word vector are used for word embedding, and the maximum length of text sentences is set to 300 and 50 words on the SNLI and Quora datasets, respectively. The specific hyperparameter settings are shown in Table 2.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 738-747, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4.2 Experimental results and analysis

We compare the experimental results of the sentence matching model based on deep interaction and fusion on the SNLI dataset with other published models.The evaluation metric we use is the accuracy rate.The results are shown in Table 3. As can be seen from Table 3, our model achieves an accuracy rate of 0.871 on the SNLI dataset, which achieves better results in the listed models.Compared with the LSTM, it is improved by 0.065. Compared with Star-Transformer model,it is improved by 0.004. Compared with some other models, it is observed that our model is better than the others model.

Table 3: The accuracy(%) of the model on the SNLI test set.Results marked with [a] are reported by Bowman et al.(Bowman et al., 2016),[b] are reported by Han et al.(Han et al., 2019), [c] are reported by Shen et al.(Shen et al., 2018), [d] are reported by Borges et al.(Borges et al., 2019), [e] are reported by Guo et al.(Guo et al., 2019), [f] are reported by Mu et al.(Mu et al., 2018).

Model	Acc
300D LSTM encoders[a]	80.6
DELTA[b]	80.7
SWEM-max[c]	83.8
Stacked Bi-LSTMs[d]	84.8
Bi-LSTM sentence encoder[d]	84.5
Star-Transformer[e]	86.0
CBS-1+ESIM[f]	86.7
DIFM	**87.1**

We conduct experiments on the Quora dataset, and the evaluation metric is accuracy. The experimental results on the Quora dataset are shown in Table 4. As can be seen from Table 4,the accuracy of our method on the test set is 0.868.The experimental results improve the accuracy by 0.054 compared to the traditional LSTM model.Compared with the enhanced sequential inference model ESIM,it is improved by 0.004. The experimental results achieved good results compared to some current popular deep learning methods.Our model achieve relatively good results in both tasks, which illustrates the effectiveness of our model.

Table 4: The accuracy(%) of the model on the Quora test set.Results marked with [g] are reported by Yang et al.(Yang et al., 2021), [h] are reported by He et al.(He and Lin, 2016) , [i] are reported by Zhao et al.(Zhao et al., 2021), [j] are reported by Chen et al.(Chen et al., 2017).

Model	Acc
LSTM	81.4
RCNN[g]	83.6
PWIM[h]	83.4
Capsule-BiGRU[i]	86.1
ESIM[j]	85.4
DIFM	**86.8**

4.3 Ablation experiments

To explore the role played by each module, we conduct an ablation experiment on the SNLI dataset .Without using the fusion function, which means that the low-level semantic information are directly spliced with the high-level semantic information. The experimental results are shown in Table 5.

We first verify the effectiveness of character embedding. Specifically,we remove the character embedding for the experiment, and its accuracy drops by 1.5 percentage points, proving that character embedding plays an important role in improving the performance of the model.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 738-747, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 5: Ablation study on the SNLI validation dataset

Model	Acc(%)
DIFM	87.1
w/o character embedding	85.6 (↓1.5)
w/o low-level semantic information	85.9 (↓1.2)
w/o high-level semantic information	79.5 (↓**7.6**)
w/o fusion	86.1(↓1.0)
w/o self-attention	58.8(↓1.3)
w/o $P \rightarrow H$	84.6(↓2.5)
w/o $H \rightarrow P$	86.2(↓0.9)

In addition,we verify the effectiveness of the semantic information and fusion modules. We removed low-level semantic information and high-level semantic information from the original model, and its accuracy dropped by 1.2 percentage points and 7.6 percentage points. At the same time, we remove the fusion function, and its accuracy drops by about 1.0 percentage points. It shows that the different semantic information and the fusion function are beneficial to improve the accuracy of the model, with the high-level semantic information being more significant for the model.

Finally,we verify the effectiveness of each attention on the model.We remove the attention from P to H, the attention from H to P, and the self-attention module respectively. Their accuracy rates decreased by 2.5 percentage points, 0.9 percentage points, and 1.3 percentage points. It shows that all the various attention mechanisms improve the performance of the model, with the P to H attention being more significant for the model.

The ablation experiments show that each component of our model plays an important role, especially the high-level semantic information module and the P to H attention module, which have a greater impact on the performance of the model. Meanwhile, the character embedding and fusion function also play an important role in our model.

5 Conclusion

we investigate natural language sentence matching methods and propose an effective deep interaction and fusion model for sentence matching. Our model first uses the bi-directional attention in the machine reading comprehension model and self-attention to obtain the high-level semantic information.Then,we use a heuristic fusion function to fuse the semantic information that we get.Finally, we use a linear layer to get the results of sentence matching .We conducted experiments on SNLI and Quora datasets. The experimental results show that the model proposed in this paper can achieve good results in two tasks.In this work,we find that our proposed interaction module and fusion module occupie the dominant position and have a great impact on our model.However,Our model is not as powerful as the pre-trained model in terms of feature extraction and lacks external knowledge.The next research work plan will focus on the following two points: 1) we use more powerful feature extractors, such as BERT pre-trained model as text feature extractors; 2) the introduction of external knowledge will be considered. For example, WordNet, an external knowledge base, contains many sets of synonyms, and for each input word, its synonyms are retrieved from WordNet and embedded in the word vector representation of the word to further improve the performance of the model.

Acknowledgements

This work is supported by National Natural Science Foundation of China [grant numbers 62162062]. State Language Commission of China under Grant No. YB135-76, scientific research project for building world top discipline of Foreign Languages and Literatures of Yanbian University under Grant No. 18YLPY13. Doctor Starting Grants of Yanbian University [2020-16], the school-enterprise cooperation project of Yanbian University [2020-15].

References

Luís Borges, Bruno Martins, and Pável Calado. 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–26.

Samuel Bowman, Christopher Potts, and Christopher D Manning. 2015a. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 12–21.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015b. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. Association for Computational Linguistics (ACL).

Samuel R Bowman, Raghav Gupta, Jon Gauthier, Christopher D Manning, Abhinav Rastogi, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 1466–1477. Association for Computational Linguistics (ACL).

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, 2004:26–29.

HJPJ Dong-hong. 2017. Convolutional network-based semantic similarity model of sentences. *Journal of South China University of Technology (Natural Science)*, 45(3):68–75.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. 2019. Star-transformer. 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 1315–1325.

Kun Han, Junwen Chen, Hui Zhang, Haiyang Xu, Yiping Peng, Yun Wang, Ning Ding, Hui Deng, Yonghu Gao, Tingwei Guo, et al. 2019. Delta: A deep learning based language technology platform. *arXiv preprint arXiv:1908.01853*.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pages 937–948.

Chaowen Hu, Changxing Wu, and Yalian Yang. 2020. Extended s-lstm based textual entailment recognition. *Journal of Computer Research and Development*, 57(7):1481–1489.

Jing Jin, Yahui Zhao, and Rongyi Cui. 2021. Research on multi-granularity ensemble learning based on korean. In *The 2nd International Conference on Computing and Data Science*, pages 1–6.

Feiyu Li, Yahui Zhao, Feiyang Yang, and Rongyi Cui. 2021. Incorporating translation quality estimation into chinese-korean neural machine translation. In *China National Conference on Chinese Computational Linguistics*, pages 45–57. Springer.

Norman Mu, Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. 2018. Parameter re-initialization through cyclical batch size schedules. *arXiv preprint arXiv:1812.01216*.

Han Ren, Yaqi Sheng, and Wenhe Feng. 2015. Recognizing textualentailmentbasedonknowledgetopicmodels. *Journal of Chinese Information Processing*, 29(6):119–127.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 440–450.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. Feature-rich two-stage logistic regression for monolingual alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 949–959.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with lstm. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1442–1451.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Feiyang Yang, Yahui Zhao, and Rongyi Cui. 2020. Recognition method of important words in korean text based on reinforcement learning. In *China National Conference on Chinese Computational Linguistics*, pages 261–272. Springer.

Dezhi Yang, Xianxin Ke, and Qichao Yu. 2021. A question similarity calculation method based on rcnn. *Journal of Computer Engineering and Science*, 43(6):1076–1080.

Qi Zhao, Yanhui Du, and Tianliang Lu. 2021. Algorithm of text similarity analysis based on capsule-bigru. *Journal of Computer Engineering and Applications*, 57(15):171–177.

Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutnık, and Jürgen Schmidhuber. 2017. Recurrent highway networks. In *International conference on machine learning*, pages 4189–4198. PMLR.

CCL 2022

ConIsI: A Contrastive Framework with Inter-sentence Interaction for Self-supervised Sentence Representation

Meng Sun
Dalian University of Technology
sunmeng20@mail.dlut.edu.cn

Degen Huang*
Dalian University of Technology
huangdg@dlut.edu.cn

Abstract

Learning sentence representation is a fundamental task in natural language processing and has been studied extensively. Recently, many works have obtained high-quality sentence representation based on contrastive learning from pre-trained models. However, these works suffer the inconsistency of input forms between the pre-training and fine-tuning stages. Also, they typically encode a sentence independently and lack feature interaction between sentences. To conquer these issues, we propose a novel **Con**trastive framework with **I**nter-sentence **I**nteraction (ConIsI), which introduces a sentence-level objective to improve sentence representation based on contrastive learning by fine-grained interaction between sentences. The sentence-level objective guides the model to focus on fine-grained semantic information by feature interaction between sentences, and we design three different sentence construction strategies to explore its effect. We conduct experiments on seven Semantic Textual Similarity (STS) tasks. The experimental results show that our ConIsI models based on $\mathrm{BERT}_{\mathrm{base}}$ and $\mathrm{RoBERTa}_{\mathrm{base}}$ achieve state-of-the-art performance, substantially outperforming previous best models SimCSE-$\mathrm{BERT}_{\mathrm{base}}$ and SimCSE-$\mathrm{RoBERTa}_{\mathrm{base}}$ by 2.05% and 0.77% respectively.

1 Introduction

Learning good universal sentence representation is a fundamental task and benefits a wide range of natural language processing tasks such as text classification and machine translation, especially for large-scale semantic similarity computation and information retrieval. With the rise of pre-trained language models (Devlin et al., 2019; Liu et al., 2019), many downstream tasks have achieved remarkable improvements. However, the native sentence representation derived from pre-trained language models without additional supervision are usually low-quality and can not be used directly (Reimers et al., 2019). Recently, contrastive learning has become a popular approach to improve the quality of sentence representation in a self-supervised way.

Contrastive learning is an approach of learning effective feature representation by positive pairs and negative pairs. It generally takes different views as positive or negative pairs for each sentence using various data augmentation ways. And it works by pulling semantically close positive instances together and pushing negative instances away. However, current approaches based on contrastive learning mainly suffer two problems: *train-tuned bias* and *fine-grained interaction deficiency*. Firstly, previous approaches typically input a single sentence to the encoder at a time, which is inconsistent with the pre-training stage of the language models. Most language models concatenate multiple sentences as the input form at the pre-training stage. We argue that the inconsistency of input forms between the pre-training and fine-tuning stages may harm the performance. Secondly, each sentence in a minibatch is encoded independently while training, which lacks fine-grained interaction information between sentences. According to previous works in text matching (Li et al., 2021; Wang et al., 2021; Lu et al., 2022) , modeling a proper interaction between input sentences can improve the performance of semantic feature embedding

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
* : Corresponding Author

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

for representation-based models, but existing works on sentence representation ignore the importance of this interaction.

Therefore, to conquer these drawbacks of current contrastive learning based methods, we propose ConIsI, a **Con**trastive framework with **I**nter-**s**entences **I**nteraction for self-supervised sentence representation. Firstly, we present to construct a sentence pair as positive instance for each sentence to alleviate the train-tuned bias. By referring to an original sentence and a sentence pair as a positive pair, the model can not only obtain effective representation of a single sentence, but also mitigate the train-tuned bias between the pre-training and fine-tuning stages. Further, to solve the problem of lacking interaction between sentences, we propose a sentence-level objective to perform the inter-sentence interaction during encoding. We pass a pair of sentences as a text sequence into the encoder and the target semantic category of the two sentences is predicted. The sentence pair is sufficiently interacted through the internal interaction mechanism in Transformer-based block (Vaswani et al., 2017) during encoding. Through the inter-sentence interaction, the model can encode fine-grained semantic information and achieve further improvement. Moreover, for a minibatch of n sentences, there are $n \cdot (n-1)/2$ interactive computations. In order to ensure the training efficiency, we do not perform an interactive operation on all data due to too many possible combinations. Instead, we artificially construct a sentence for each original sentence to adjust the difficulty of the interactive objective, which only requires n interactive computations. We propose several models based on three sentence construction strategies, named ConIsI-o1, ConIsI-o2, and ConIsI-s, respectively. The overall model of our proposed ConIsI can be seen in Figure 1.

Our contributions can be summarized as follows:

- We propose to construct each positive pair with an original sentence and a sentence pair based on contrastive learning, which not only learns effective representation by pulling semantically close samples together but also mitigates the train-tuned bias between pre-training and fine-tuning phases.
- We propose a simple but effective sentence-level training objective based on inter-sentence interaction. It alleviates the problem of interaction deficiency among sentences and enriches the semantic information of sentence representation. We also present three sentence construction strategies for interactive sentence pairs and analyze their effects.
- We conduct extensive experiments on seven standard Semantic Textual Similarity (STS) datasets. The results show that our proposed ConIsI-s-$\text{BERT}_{\text{base}}$ and ConIsI-s-$\text{RoBERTa}_{\text{base}}$ achieve 78.30% and 77.34% averaged Spearman's correlation, a 2.05% and 0.77% improvement over SimCSE-$\text{BERT}_{\text{base}}$ and SimCSE-$\text{RoBERTa}_{\text{base}}$ respectively, which substantially outperforms the previous state-of-the-art models.

2 Related Work

Sentence representation built upon the distributional hypothesis has been widely studied and improved considerably. Early works (Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018) inspired by word2vec (Mikolov et al., 2013) lead to strong results by predicting surrounding information of a given sentence. The emergence of pre-trained models such as BERT (Devlin et al., 2019) shows much great potential for sentence representation. Recently, many works have explored how to learn better sentence embeddings from the pre-trained models.

Supervised Methods A common supervised step of learning a model is fine-tuning with labeled data in downstream training sets. Several works build upon the success of using annotated natural language inference (NLI) datasets (including Stanford NLI (Bowman et al., 2015) and Multi-Genre NLI (Williams et al., 2018)) for sentence representation, which projects it as a 3-way classification task (entailment, neutral, and contradiction) to get better sentence embeddings. Conneau et al. (2017) use a BiLSTM-based model as encoder, and they train it on both Stanford NLI and Multi-Genre NLI datasets. Universal Sentence Encoder (Cer et al., 2018) uses the Stanford NLI dataset to enhance the unsupervised training by adopting a Transformer-based model. Sentence-BERT (Reimers et al., 2019) that adopts a Siamese network (Chopra et al., 2005) with a shared BERT encoder is also trained on Stanford NLI and Multi-Genre NLI datasets.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Unsupervised Methods Some works focus on using the regularization method to improve the quality of raw sentence representation generated by original BERT. Bert-flow (Li et al., 2020) puts forward a flow-based approach to solving the problem that native embeddings of BERT occupy a narrow cone in the vector space. Similarly, Bert-whitening (Su et al., 2021) maps BERT's embeddings to a standard Gaussian latent space by whitening the native embeddings. They all try to alleviate the representation degeneration of pre-trained models and yield substantial improvement.

Self-supervised Methods The sentence-level training objective in language models like BERT inspires a line of work over self-supervised sentence representation learning. BERT includes the next sentence prediction (NSP) task, which predicts whether two sentences are neighboring or not. However, Liu et al. (2019) prove that NSP has minimal effect on the final performance and even does harm to the training model. Therefore, many works have proposed various self-supervised objectives for pre-training sentence encoders. Cross-Thought (Wang et al., 2020) and CMLM (Yang et al., 2021) are two similar approaches that present to predict surrounding tokens of given contextual sentences. And Lee et al. (2020) propose to learn an objective that predicts the correct sentence ordering provided the input of shuffled sentences.

As a self-supervised learning method, contrastive learning with no need for scarce labeled data attracts much attention, and many excellent works have been proposed. Inspired by SimCLR (Chen et al., 2020) which applies data augmentation techniques on the same anchor such as image rotating, scaling, and random cropping to learn image representation in the computer vision community, some works pay attention to getting effective positive pairs by using similar approaches. In the natural language process community, many works apply textual augmentation techniques on the same sentence to obtain different views as positive pairs based on the SimCLR framework. Zhang et al. (2020) extract global feature of a sentence as positive pairs, Wu et al. (2020) and Yan et al. (2021) take some token-level transformation ways such as word or subword deletion or replacement, and Gao et al. (2021) apply dropout mask of Transformer-based encoder to get positive pairs. And Zhang et al. (2021) adopt BYOL (Grill et al., 2020) framework using back-translation data.

3 Methodology

In this section, we present ConIsI, a contrastive framework with inter-sentence interaction for self-supervised sentence representation, which contains two parts: (1) the ConIsI model of joint contrastive learning objective and inter-sentence interactive objective (Section 3.1), and (2) the strategies of sentence construction in the inter-sentence interactive objective (Section 3.2).

3.1 Model

The ConIsI model joints contrastive learning and inter-sentence interactive objectives. The inter-sentence interactive objective is a binary classification task that performs fine-grained interaction between sentences and predicts whether two sentences are in the same semantic category. The overall architecture is shown in Figure 1.

3.1.1 Data Augmentation

To alleviate the train-tuned bias caused by different input forms, we perform sentence-level repetition operation to construct positive instances. For each sentence, our approach proposes to take a sentence pair as positive instance. Specifically, given a tokenized sentence $x = \{t_1, t_2, ..., t_l\}$ (l is the max sequence length), we define the sentence pair as $Y = \{t_1, t_2, ..., t_l, t_1, t_2, ..., t_l\}$, which is the concatenation of two original sentences. For each minibatch of sentences $\mathcal{B} = \{x_i\}_{i=1}^{N}$ (N is the batch size), we perform data augmentation operation on each sentence and then get the positive instances $\mathcal{B}_{\text{Aug}} = \{Y_i\}_{i=1}^{N}$.

3.1.2 Sentence Pair Composition

To perform fine-grained interaction between sentences, we take a pair of sentences as a textual sequence to input into the encoder. The input two sentences can get fine-grained interaction with each other through Transformer-based block. Also, considering the training efficiency, we do not perform interaction on all sentences as there are too many combinations of sentence pairs. Instead, we construct the composed

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

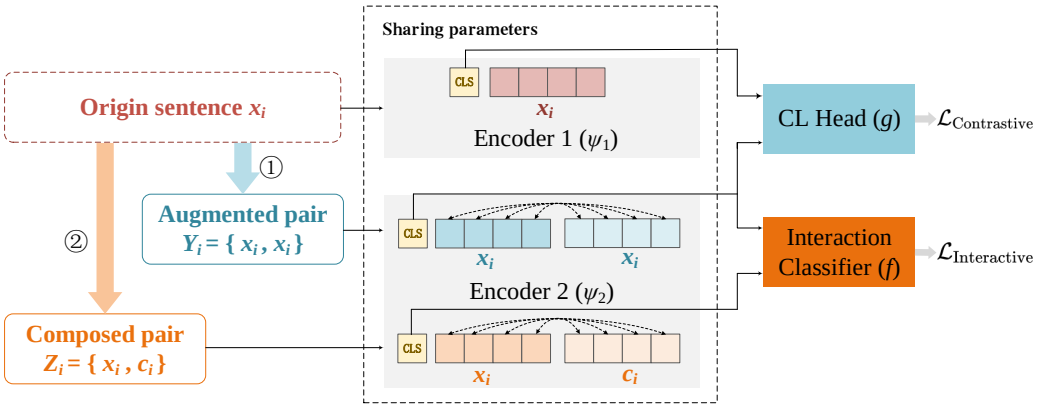


Figure 1: The overall structure of the ConIsI model. It mainly consists of five components: the data augmentation operation (①), the text composition part (②), the encoder $\psi(\cdot)$ mapping the input data to the sentence representation space, the CL Head $g(\cdot)$ and the Interaction Classifier $f(\cdot)$ applying for the contrastive loss and the interactive loss respectively.

sentence pair $Z_i = \{x_i, c_i\}$ for each sentence x_i in $\mathcal{B}$. Specifically, we try to obtain a sentence c_i which belongs to a different semantic category from x_i. Then we concatenate the sentence x_i and the sentence c_i as a composed sentence pair Z_i. We perform the sentence pair composition operation on each sentence in minibatch $\mathcal{B} = \{x_i\}_{i=1}^N$ and then get the composed pairs $\mathcal{B}_{\text{Com}} = \{Z_i\}_{i=1}^N$. We explore three different sentence construction strategies to obtain c_i in section 3.2.

3.1.3 Encoding

We take pre-trained checkpoints of BERT or RoBERTa as the encoder model to obtain sentence representation. For BERT, there are two input forms to fine-tune downstream tasks: one is the single sentence input, and the other is the sentence pair input. Previous works based on contrastive learning input a single sentence to the pre-trained model to learn sentence embeddings, which is inconsistent with the pre-training stage and suffers the train-tuned bias. To alleviate this problem and maintain the model's ability of encoding a single sentence meanwhile, we propose to adopt both two forms. The original sentence x_i is taken as a single sentence and input to the encoder 1. The augmented sentence pair Y_i and the composed sentence pair Z_i are taken as sentence pairs and input to the encoder 2. And to ensure that the augmented sentence pair has the same meaning as the original sentence, the max length of the tokenizer for the former is set double for the latter. The encoder 1 and the encoder 2 share the same parameters.

For RoBERTa whose input forms are a single sentence or several concatenated sentences separated by "</s>" token, we input the original sentence into the encoder 1. And The augmented sequence pair and the composed sentence pair are taken as two concatenated sentences and input to the encoder 2. Similarly, the max length of the tokenizer for encoder 2 is set double for that of encoder 1, and the two encoders share the same parameters.

3.1.4 Contrastive Learning

Contrastive learning aims to learn effective representation by pulling semantically close objects and pushing ones that are dissimilar away. We follow the SimCRL (Chen et al., 2020) contrastive framework and take a cross-entropy objective (Chen et al., 2017) in our approach.

For each minibatch $\mathcal{B} = \{x_i\}_{i=1}^N$, the contrastive loss is defined on $\mathcal{B}$ and the augmented instances $\mathcal{B}_{\text{Aug}} = \{Y_i\}_{i=1}^N$. Let $i \in \{1, .., N\}$ denote the index of an arbitrary instance in augmented set $\mathcal{B}_{\text{Aug}}$, and let $j \in \{1, .., N\}$ be the index of the other instance in $\mathcal{B}_{\text{Aug}}$. We refer to (x_i, Y_i) as a positive pair, while treating the other $N-1$ examples $Y_j (j \neq i)$ in $\mathcal{B}_{\text{Aug}}$ as negative instances for this positive pair. After the positive pair is encoded, we obtain the last hidden state of the special "[CLS]" token as the contextual

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

representation of the corresponding sample, denoted as $h_{[\mathrm{CLS}]}$.

$$\begin{aligned} h^x_{[\mathrm{CLS}]}, h^x_1, ..., h^x_l, h^x_{[\mathrm{SEP}]} &= \psi_1(x) \\ h^Y_{[\mathrm{CLS}]}, h^Y_1, ..., h^Y_l, h^Y_{[\mathrm{SEP}]}, h^{Y'}_1, ..., h^{Y'}_l, h^{Y'}_{[\mathrm{SEP}]} &= \psi_2(Y) \end{aligned} \tag{1}$$

Then we add a predictor layer $g(\cdot)$ to map $h_{[\mathrm{CLS}]}$ to the contrastive embedding space and obtain h , which is given as follows:

$$h = Elu(BN_1(W_1 \cdot h_{[\mathrm{CLS}]} + b_1)) \tag{2}$$

where $W_1 \in R^{d\times d}$ is the weight matrix, $b_1 \in R^{d\times 1}$ is the bias vector, and d is the number of features in hidden layers. Both W_1 and b_1 are trainable parameters. BN_1 is the $\mathrm{BatchNorm1d}$ layer and Elu is the activate function.

Let h^x_i, h^Y_i and h^Y_j be the corresponding outputs of the head $g(\cdot)$. Then for x_i, we try to separate Y_i apart from all negative instances by minimizing the following,

$$\ell^I_i = -log\frac{e^{sim(h^x_i, h^Y_i)/\tau}}{\sum_{j=1}^{N} e^{sim(h^x_i, h^Y_j)/\tau}} \tag{3}$$

where τ denotes the temperature parameter we set as 0.05. We choose cosine similarity $sim(\cdot)$ as the similarity calculation function between a pair of normalized outputs, $sim(h_1, h_2) = \frac{h_1^T h_2}{||h_1||\cdot||h_2||}$.

The contrastive loss is then averaged over all pairs,

$$\mathcal{L}_{\mathrm{Contrastive}} = \sum_{i=1}^{N} \ell^I_i / N \tag{4}$$

3.1.5 Interactive Classification

When applying a training objective after getting sentence embeddings in previous work, each sentence is encoded independently and can not see other sentences while encoding. Therefore, the semantic information contained in each sentence embeddings is insufficient. In contrast, modeling sentence pairs can effectively alleviate this problem. While encoding a sentence pair through the model, the two sentences can obtain fine-grained interaction information from each other. We propose to model an inter-sentence interaction objective between input sentences to enrich semantic information for sentence embeddings.

We encode the sentence pairs into the semantic category space for self-supervised classification. Different from contrastive learning objective, the interactive objective learns fine-grained semantic information through the interaction between sentences. The interactive loss is implemented on the augmented instance Y_i in B_{Aug} and the corresponding composed instance Z_i in $\mathcal{B}_{\mathrm{Com}}$. We refer to the two sentences $\{x_i, x_i\}$ in augmented pair Y_i as being in the same category, and the sentences $\{x_i, c_i\}$ in composed pair Z_i as being in different category. Our model passes Y_i and Z_i to the encoder 2 and obtains the last hidden state of the special "[CLS]" token as their sentence pair embeddings, respectively.

$$\begin{aligned} h^Y_{[\mathrm{CLS}]}, h^Y_1, ..., h^Y_l, h^Y_{[\mathrm{SEP}]}, h^{Y'}_1, ..., h^{Y'}_l, h^{Y'}_{[\mathrm{SEP}]} &= \psi_2(Y) \\ h^Z_{[\mathrm{CLS}]}, h^Z_1, ..., h^Z_l, h^Z_{[\mathrm{SEP}]}, h^{Z'}_1, ..., h^{Z'}_l, h^{Z'}_{[\mathrm{SEP}]} &= \psi_2(Z) \end{aligned} \tag{5}$$

We use a predictor and linear layers to encode $h_{[\mathrm{CLS}]}$ into the semantic category space to obtain r. $r \in R^d$ is the semantic category representation. The formulas are as follows:

$$h = Elu(BN_2(W_2 \cdot h_{[\mathrm{CLS}]} + b_2)) \tag{6}$$

$$r = W_3 \cdot h + b_3 \tag{7}$$

where $W_2, W_3 \in R^{d\times d}$ are the weight matrixs, $b_2, b_3 \in R^{d\times 1}$ are the bias vectors, and d is the number of features in the hidden layers. W_2, W_3 and b_2, b_3 are all learnable parameters, and W_2, b_2 share the same parameters with W_1 and b_1 in $g(\cdot)$ respectively. BN_2 share the same parameters with BN_1 and Elu is the activate function.

Let r^Y_i and r^Z_i denote the corresponding outputs of the head $f(\cdot)$. Then we predict whether each pair

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

is in the same category by optimizing the following objective,

$$\ell_i^{II} = -log\frac{e^{r_i^Y}}{e^{r_i^Y} + e^{r_i^Z}} \tag{8}$$

Then the interactive loss for a mini-batch with N sentence pairs is as follows:

$$\mathcal{L}_{\text{Interactive}} = \sum_{i=1}^{N} \ell_i^{II}/N \tag{9}$$

3.1.6 Overall objective

Finally, our overall objective is,

$$\begin{aligned}\mathcal{L} &= (1-\lambda)\cdot\mathcal{L}_{\text{Contrastive}} + \lambda\cdot\mathcal{L}_{\text{Interactive}} \\ &= (1-\lambda)\cdot\sum_{i=1}^{N}\ell_i^{I}/N + \lambda\cdot\sum_{i=1}^{N}\ell_i^{II}/N\end{aligned} \tag{10}$$

where ℓ_i^I, ℓ_i^{II} are defined in Eq(3) and Eq(8), respectively. λ is the balanced parameter between the contrastive loss and the interactive loss. During training, we jointly optimize a contrastive learning objective and an inter-sentence interactive objective over the original sentences, the augmented sentence pairs and composed sentence pairs. Then we fine-tune all the parameters using the joint objective.

3.2 Sentence Construction Techniques

Intuitively, two semantically opposite sentences are easier for the model to distinguish than two semantically closer sentences. As a self-supervised classification task, the difficulty of the interactive objective can significantly affect the performance of the model. Thus we propose different sentence construction techniques to control the complexity of the inter-sentence interactive objective. We try to construct a sentence c_i that is not in the same semantic category as the original sentence x_i in section 3.1.2. We explore three sentence construction methods, two of which are constructing from the original sentence as shown in section 3.2.1, and one is sampling from other sentences in section 3.2.2.

3.2.1 From Original Sentence

Since the bidirectional language models encode a word based on contextual information, sentences with high textual similarity usually are in high semantic similarity in representation. However, the sentences with high textual similarity may not actually be semantically similar. For example, "this is not a problem." and "this is a big problem." are two sentences with high textual similarity because of similar wording, but they are not semantically similar because of opposite meanings. The models usually fail to distinguish textual similarity and semantic similarity, which has been discussed deeply in the vision field (Robinson et al., 2021; Chen et al., 2021). As a result, a model may overestimate the semantic similarity of any pairs with similar wording regardless of the actual semantic difference between them. Therefore, we propose to construct sentences that are semantically different but are textually similar to the original sentence to improve the fine-grained semantic discrimination ability of the model.

Subword Replacement The subword replacement mechanism randomly substitutes some sub-words in a sentence. Specifically, given a tokenized sub-word sequence $x = \{t_1, t_2, ..., t_l\}$ (l is the max sequence length) after processing by a sub-word tokenizer. Firstly, We mask a certain proportion of the tokenized sequence x at random. If the i-th token is chosen, then we replace the masked token with a random token 80% of the time, leaving the masked token unchanged 20% of the time.

Word Replacement The word replacement mechanism works on full words in a sentence. Different from subword replacement, the word replacement mechanism randomly substitutes some full words with antonyms. If a word is chosen, then we replace the word with its antonym. We use the WordNet (Miller, 1993) to obtain the antonym of a word.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.2.2 From Other Sentences

Different from constructing a new sentence from the original sentence, this method selects one other sentence from the training data at random. Specifically, for a given sentence x_i within the minibatch $\mathcal{B} = \{x_i\}_{i=1}^{N}$, we randomly select sentence x_k $(k \in [1, N], k \neq i)$ as c_i for composed pair.

We apply the three sentence construction strategies to our ConIsI model, named ConIsI-o1, ConIsI-o2, and ConIsI-s. Among them, ConIsI-o1 and ConIsI-o2 represent the joint contrastive objective and interactive objective under the subword replacement and word replacement, respectively. ConIsI-s represents the jointing of contrastive learning and the interactive objective under the sampling from other sentences.

4 Experiments

4.1 Data

We train our model on the same one million sentences randomly sampled from English Wikipedia that are provided by SimCSE[0]. All our experiments are fully self-supervised and note that no STS sets are used for training.

We evaluate our approach on multiple Semantic Textual Similarity (STS) datasets: STS12-16 (STS12 - STS16) (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015; Agirre et al., 2016), STS Benchmark (STS-B) (Cer et al., 2017) and SICK-Relatedness (SICK-R) (Marelli et al., 2014), which are seven standard STS benchmark datasets and are extensively used to measure the sentence embeddings and the semantic similarity of sentence pairs. These datasets are composed of pairs of sentences and one golden score between 0 and 5, where a higher score indicates a higher similarity between two sentences in Table 1. The statistics is shown in Table 2.

Sentence1	Sentence2	Golden Score
a plane is taking off .	an air plane is taking off .	5.000
a cat is playing a piano .	a man is playing a guitar .	0.600
a man is playing a guitar .	a man is playing a trumpet .	1.714

Table 1: The sentence samples of STS datasets.

	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Total
Number of train samples	0	0	0	0	0	5479	4500	-
Number of valid samples	0	0	0	0	0	1500	500	-
Number of test samples	3108	1500	3750	3000	1186	1379	4927	-
Number of Unlabeled Texts	6216	3000	7500	17000	18366	17256	19854	89192

Table 2: The statistics of STS datasets.

4.2 Evaluation Setup

Following previous work, we evaluate our method on STS tasks using the SentEval toolkit (Conneau and Kiela, 2018). We take the "[CLS]" embedding generated by the last hidden layer of the encoder 1 in Figure 1 as the sentence representation. To evaluate the sentence representation for a fair comparison, we follow the settings of Sentence-BERT (Reimers et al., 2019) and SimCSE (Gao et al., 2021): (1) we directly take cosine similarities for all STS tasks without training extra linear regressor on top of frozen sentence embeddings for STS-B and SICK-R; (2) we report Spearman's rank correlation coefficients rather than Pearson's; (3) and we take the "all" setting for STS12-STS16 which fuses data from different topics together to make the evaluation closer to real-world scenarios.

[0]https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse/resolve/main/wiki1m for simcse.txt

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4.3 Training Details

We implement our ConIsI model with Huggingface's transformers package[1] 4.2.1 based on Python 3.8.12 and Pytorch 1.8.0 and run the model on Nvidia 3090 GPU. We start our experiments from pre-trained checkpoints of BERT or RoBERTa. All experiments use the Adam optimizer and the random seed is set as 42. The temperature parameter τ is set as 0.05, and the dropout rate is set as 0.1. Furthermore, the hyper-parameter settings of the models are shown in Table 3. Besides, We train our models for one epoch and evaluate the model every 125 training steps.

Model	Batch size	Max sequence length	Learning rate	Hidden size	λ
ConIsI-s-$\text{BERT}_{\text{base}}$	64	32	3e-5	768	0.8
ConIsI-s-$\text{RoBERTa}_{\text{base}}$	64	32	3e-5	768	0.1
ConIsI-s-$\text{BERT}_{\text{large}}$	64	28	3e-5	1024	0.1

Table 3: Hyper-parameters settings for ConIsI-s models.

4.4 Baselines

We compare our model with previous strong baseline models on STS tasks, including:
(1) Recent state-of-the-art self-supervised models using a contrastive objective: SimCSE (Gao et al., 2021), IS-BERT (Zhang et al., 2020), ConSERT (Yan et al., 2021), Mirror-BERT (Liu et al., 2021), DeCLUTR (Giorgi et al., 2021), CT-BERT (Carlsson et al., 2020), BSL (Zhang et al., 2021), SG-OPT (Kim et al., 2021);
(2) Post-processing methods like BERT-flow (Li et al., 2020) and BERT-whitening (Su et al., 2021);
(3) And naive baselines like averaged GloVe embeddings (Pennington et al., 2014); averaged first and last layer BERT embeddings.

4.5 Main Results

Table 4 shows the evaluation results on seven STS tasks. ConIsI-s-$\text{BERT}_{\text{base}}$ can significantly outperform SimCSE-$\text{BERT}_{\text{base}}$ and raise the averaged Spearman's correlation from 76.25% to 78.30%, which brings a 2.05% average improvement over the SimCSE-$\text{BERT}_{\text{base}}$ model on seven tasks. For the RoBERTa model, ConIsI-s-$\text{RoBERTa}_{\text{base}}$ can also improve upon SimCSE-$\text{RoBERTa}_{\text{base}}$ from 76.57% to 77.34%, a 0.77% increase. And for the ConIsI-s-$\text{BERT}_{\text{large}}$ model, we also achieve better performance, from 78.41% to 79.55%, a 1.14% increase. In general, our method achieves substantial improvement on the seven STS datasets over baseline models.

4.6 Ablation Study

In this section, we discuss the effects of different components. In our model, both the contrastive learning objective and the inter-sentence interactive objective are crucial because they are committed to obtaining the ability of normal semantic encoding and fine-grained semantic information, respectively. If we remove the inter-sentence interactive objective, the model becomes a SimCSE-like model with a different positive instance construction way, causing a drop of 1.30%. If we remove the contrastive learning objective, the performance of **Avg.** drops significantly by more than 10% (see Table 5). This results show that it is important to have common and fine-grained attributes that exist together in the sentence representation space. When compared with SimCSE-$\text{BERT}_{\text{base}}$, our proposed method of taking a sentence pair as positive instance brings an improvement of 0.75%. The result shows that the problem of train-tuned bias is alleviated by the input form of augmented sentence pair.

4.7 Analysis

In this section, we conduct a series of experiments to validate our model better. We use $\text{BERT}_{\text{base}}$ or $\text{RoBERTa}_{\text{base}}$ model and all reported results are evaluated on the seven STS tasks.

[1] https://github.com/huggingface/transformers

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe-embeddings(avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$(first-last avg.)◇	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-flow◇	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening◇	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$§	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
BSL-$\text{BERT}_{\text{base}}$†	67.83	71.40	66.88	79.97	73.97	73.74	70.40	72.03
CT-$\text{BERT}_{\text{base}}$◇	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
ConSERT-$\text{BERT}_{\text{base}}$‡	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SG-OPT-$\text{BERT}_{\text{base}}$♭	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
Mirror-$\text{BERT}_{\text{base}}$♮	69.10	81.10	73.00	81.90	75.70	78.00	69.10	75.40
SimCSE-$\text{BERT}_{\text{base}}$◇	68.40	82.41	74.38	80.91	78.56	76.85	**72.23**	76.25
∗ConIsI-s-$\text{BERT}_{\text{base}}$	**70.92**	**84.35**	**76.67**	**83.53**	**78.94**	**82.15**	71.55	**78.30**
$\text{RoBERTa}_{\text{base}}$(first-last avg.)◇	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\text{base}}$whitening◇	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-$\text{RoBERTa}_{\text{base}}$◇	52.41	75.19	65.52	77.12	78.63	72.41	68.62	69.99
SimCSE-$\text{RoBERTa}_{\text{base}}$◇	70.16	81.77	73.24	**81.36**	80.65	80.22	68.56	76.57
∗ ConIsI-s-$\text{RoBERTa}_{\text{base}}$	**71.21**	**83.31**	**75.11**	81.13	**80.73**	**80.50**	**69.39**	**77.34**
SimCSE-$\text{BERT}_{\text{large}}$◇	70.88	84.16	76.43	84.50	**79.76**	79.26	73.88	78.41
∗ ConIsI-s-$\text{BERT}_{\text{large}}$	**72.33**	**86.14**	**77.42**	**84.83**	79.60	**81.76**	**74.78**	**79.55**

Table 4: Sentence embedding performance on STS tasks in terms of Spearman's correlation and "all" setting. ♣: results from (Reimers et al., 2019); §: results from (Zhang et al., 2020); †: results from (Zhang et al., 2021); ‡: results from (Yan et al., 2021); ♭: results from (Kim et al., 2021); ♮: results from (Liu et al., 2021); ◇: results from (Gao et al., 2021); ∗: results from ours.

Model	Avg.
SimCSE-$\text{BERT}_{\text{base}}$	76.25
ConIsI-s-$\text{BERT}_{\text{base}}$	78.30
w/o fine-grained classification loss	77.00 (-1.30)(+0.75)
w/o contrastive loss	67.68 (-10.62)

Table 5: **Avg.** results of seven STS tasks for ConIsI-s-$\text{BERT}_{\text{base}}$ model variants.

4.7.1 Validation of Sentence Construction Strategies

We compare the three models ConIsI-o1, ConIsI-o2, and ConIsI-s to verify the effects of our proposed sentence construction strategies for the inter-sentence interactive objective.

Table 6 shows that our proposed sentence construction techniques for the inter-sentence interactive objective improve the performance of self-supervised sentence representation. Compared with SimCSE-$\text{BERT}_{\text{base}}$ and SimCSE-$\text{RoBERTa}_{\text{base}}$, the Spearman's correlation of ConIsI-o1-$\text{BERT}_{\text{base}}$ and ConIsI-o1-$\text{RoBERTa}_{\text{base}}$ on seven STS tasks have improved by 0.89% and 1.78% respectively, a 1.34% increase on average. The results of ConIsI-o2-$\text{BERT}_{\text{base}}$ and ConIsI-o2-$\text{RoBERTa}_{\text{base}}$ on seven STS tasks have improved by 1.10% and 1.56% respectively, a 1.33% increase on average. The results of ConIsI-s-$\text{BERT}_{\text{base}}$ and ConIsI-s-$\text{RoBERTa}_{\text{base}}$ have improved by 2.05% and 0.77% respectively, a 1.41% increase on average.

As the Table 6 shown, the ConIsI-o1-$\text{RoBERTa}_{\text{base}}$ and ConIsI-o2-$\text{RoBERTa}_{\text{base}}$ implemented by the strategies of "from original sentence" bring more remarkable improvement to the SimCSE-RoBERTa model, exceeding 1.5%. And the ConIsI-s models implemented by the strategy of "from other sentences" gets a lower boost to the SimCSE-RoBERTa model, but a greater improvement to the SimCSE-BERT model. That is, RoBERTa is more capable of encoding fine-grained features and distinguishing textual similarity and semantic similarity than BERT. In contrast, BERT focuses

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Avg.	Model	Avg.
SimCSE-BERT$_{\text{base}}$	76.25	SimCSE-RoBERTa$_{\text{base}}$	76.57
∗ConIsI-o1-BERT$_{\text{base}}$	77.14	∗ConIsI-o1-RoBERTa$_{\text{base}}$	78.35
∗ConIsI-o2-BERT$_{\text{base}}$	77.35	∗ConIsI-o2-RoBERTa$_{\text{base}}$	78.13
∗ConIsI-s-BERT$_{\text{base}}$	78.30	∗ConIsI-s-RoBERTa$_{\text{base}}$	77.34

Table 6: Validation results of sentence construction strategies.

more on encoding common features in the sentence representation space. We argue that the pre-trained RoBERTa model pays more attention to fine-grained features because of the more refined optimization techniques than BERT in the pre-training phase. So ConIsI-o1-RoBERTa$_{\text{base}}$ and ConIsI-o2-RoBERTa$_{\text{base}}$ achieve better performance than ConIsI-s-RoBERTa$_{\text{base}}$. While ConIsI-s-BERT$_{\text{base}}$ achieves better performance than ConIsI-o1-BERT$_{\text{base}}$ and ConIsI-o2-BERT$_{\text{base}}$.

Overall, our proposed contrastive framework with inter-sentence interaction have improved performance compared with the previous best model SimCSE. The experimental results show that the three sentence construction strategies are effective for the ConIsI model. We take the ConIsI-s model's results as our final ConIsI model's performance.

4.7.2 Effect of Coefficient λ

λ is the weighted hyperparameter for contrastive loss and inter-sentence interactive loss involved in the final joint objective function Eq(10). A smaller λ means a larger contrastive loss weight, indicating that the model pays more attention to common features. And a larger λ means a larger interactive loss weight, indicating that the model focuses more on fine-grained features. Our experiments find that λ plays an essential role in the joint objective, and the experimental results are shown in Table 7. When $\lambda = 0$, the model becomes a SimCSE-like model, and the result shows that our proposed method to take a sentence pair as the positive instance is effective, which brings an improvement over SimCSE-BERT$_{\text{base}}$(Gao et al., 2021) by 0.75%. The results prove that the interactive objective is helpful to enhance the performance of the model under different λ. And when $\lambda = 0.8$, it achieves the best performance on the STS datasets and gets substantial improvement over that when $\lambda = 0$.

λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Avg.	77.00	77.97	77.42	77.78	77.81	78.03	77.76	77.93	**78.30**	77.58

Table 7: **Avg.** results of seven STS tasks under different λ for ConIsI-s-BERT$_{\text{base}}$ model.

5 Conclusion

In this paper, we propose the ConIsI model, which joints contrastive learning and inter-sentence interactive training objective for optimization. We propose to perform a sentence repetition operation on each sentence and then take the augmented pair as a positive instance based on contrastive learning, which alleviates the train-tuned bias of language models. We also propose the inter-sentence interactive objective, which guides the model to focus on fine-grained semantic information by feature interaction between sentences. Moreover, we design three sentence construction strategies in the inter-sentence interactive objective. Experimental results show our proposed ConIsI achieves substantial improvement over the previous state-of-the-art models. In the future, we will further explore more effective inter-sentence interactive way to enrich semantic information in sentence representation, and we hope to apply our approach to other downstream tasks such as machine translation.

References

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–*

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

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.

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.

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.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 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.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Samuel 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.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ting Chen, Calvin Luo, and Lala Li. 2021. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

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)*.

A Conneau, D Kiela, H Schwenk, L Barrault, and A 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. 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 *NAACL-HLT (1)*.

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.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. 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 879–895.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of NAACL-HLT*, pages 1367–1377.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. 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 2528–2540.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Haejun Lee, Drew A Hudson, Kangwook Lee, and Christopher D Manning. 2020. Slm: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Dan Li, Yang Yang, Hongyin Tang, Jingang Wang, Tong Xu, Wei Wu, and Enhong Chen. 2021. Virt: Improving representation-based models for text matching through virtual interaction. *arXiv preprint arXiv:2112.04195*.

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*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, et al. 2022. Ernie-search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. *arXiv preprint arXiv:2205.09153*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Lrec*, pages 216–223. Reykjavik.

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.

George A Miller. 1993. Wordnet: A lexical database for english. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

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.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 671–688. Association for Computational Linguistics.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in Neural Information Processing Systems*, 34.

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

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.

Shuohang Wang, Yuwei Fang, Siqi Sun, Zhe Gan, Yu Cheng, Jingjing Liu, and Jing Jiang. 2020. Cross-thought for sentence encoder pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–421.

Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. 2021. Distilled dual-encoder model for vision-language understanding. *arXiv preprint arXiv:2112.08723*.

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.

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.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6216–6228.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped unsupervised sentence representation learning. 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 5168–5180.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 748-760, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Data Synthesis and Iterative Refinement for Neural Semantic Parsing without Annotated Logical Forms

Shan Wu[1,3], **Bo Chen**[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
{wushan2018,chenbo,xianpei,sunle}@iscas.ac.cn

Abstract

Semantic parsing aims to convert natural language utterances to logical forms. A critical challenge for constructing semantic parsers is the lack of labeled data. In this paper, we propose a data synthesis and iterative refinement framework for neural semantic parsing, which can build semantic parsers without annotated logical forms. We first generate a naive corpus by sampling logic forms from knowledge bases and synthesizing their canonical utterances. Then, we further propose a bootstrapping algorithm to iteratively refine data and model, via a denoising language model and knowledge-constrained decoding. Experimental results show that our approach achieves competitive performance on GEO, ATIS and OVERNIGHT datasets in both unsupervised and semi-supervised data settings.

1 Introduction

Semantic parsing is the task of translating natural language (NL) utterances to their formal meaning representations (MRs), such as lambda calculus (Zettlemoyer and Collins, 2005; Wong and Mooney, 2007), FunQL (Kate et al., 2005; Lu et al., 2008), and SQL queries (Guo et al., 2019; Bogin et al., 2019; Chang et al., 2020). Currently, most neural semantic parsers (Dong and Lapata, 2016; Dong and Lapata, 2018) model semantic parsing as a sequence translation task via a encoder-decoder framework. For instance, given an utterance "*What is the length of river traverses state0*", a SEQ2SEQ parsing model obtains its FunQL representation by sequentially generating its tokens `answer(length(river(traverse_2( state0))))`.

One of the key challenges in building a semantic parser is the scarcity of annotated data. Since annotating utterances with MRs is time consuming and requires specialized expert knowledge. Witnessed the data bottleneck problem, there are many learning algorithms have been proposed, such as denotation-based weak supervised learning (Pasupat and Liang, 2016; Misra et al., 2018), dual learning (Cao et al., 2019) , transfer learning (Su and Yan, 2017; Herzig and Berant, 2018). There are also many studies focus on the quick construction of training data, such as OVERNIGHT (Wang et al., 2015). However, these works still require some degree of human efforts.

In this paper, we propose a data synthesis and iterative refinement framework, which can build semantic parsers without labeled data. Inspired by the idea that, a simple and noise corpus can be synthesized by a grammar-lexicon method, like the one used in OVERNIGHT, and can be refined by leveraging external knowledges, like language models and knowledge base constraints. So, we first obtain a naive corpus based on synchronous context-free grammars and a seed lexicon. Then we improve the corpus with the knowledge of language models and knowledge base constraints by iteratively refining data and model to obtain mature corpus. Finally, we use the refined corpus to train the semantic parser. Figure 1 shows the overview of our method.

Specifically, to get the naive corpus, we sample logical forms from knowledge bases, and then synthesize their corresponding canonical utterances using a grammar-based synthesizing algorithm. For example, like in Overnight, we can synthesize an unnatural utterance "*what is length river traverse state0*" from `answer(length(river(traverse_2( state0))))`. Although the synthesized utterance

*Corresponding Author

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

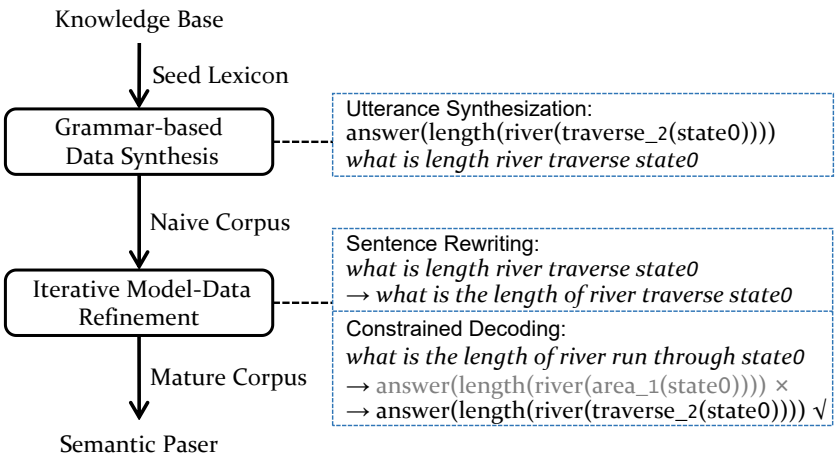


Figure 1: The overview of our approach.

"*what is length river traverse state0*" is different from the real-world utterance "*what is the length of river traverse state0*", the naive corpus can provide a start for unsupervised learning, and can be used to pretrain a base semantic parser.

Then, to improve the synthesized naive corpus, we iteratively refine the model and the data via a bootstrapping process, using the knowledge of language models and knowledge base constraints. Due to the limitation of grammars and seed lexicon, the synthesized training instances in naive corpus are often noisy, differing from real-world utterances, and with limited diversity, which hinder the model from generalizing to natural data. To address these issues, we propose to iteratively refine the model and the synthesized data via a denoising language model and knowledge-constrained decoding. Firstly, we view synthesized canonical utterances as an artificial version of utterances which are often not as fluent as natural utterances, then leverage a denoising language model to rewrite the canonical utterances to be closer to natural utterances. Secondly, to address the noise problem, a knowledge-constrained decoding algorithm is employed to exploit constraints from knowledge bases, therefore meaning representations can be more accurately predicted even when semantic parser is not strong enough. Finally, the *data synthesization* and *semantic parsing* are iteratively refined to bootstrap both the corpus and the semantic parser: the refined corpus is used to train a better semantic parser, and the better semantic parser in turn is used to refine training instances.

The main contributions of this paper are:

- We propose a data synthesis and iterative refinement framework to build neural semantic parsers without labeled logical forms, in which we generate naive corpus from scratch and improve them with the knowledge of language models and knowledge base constraints via an iterative data-model refinement.
- Experimental results on GEO, ATIS and OVERNIGHT datasets show that our approach achieves competitive performance without using annotated data.

2 Background

2.1 Base Semantic Parsing Model

We employ the SEQ2SEQ semantic parser as our base model (Dong and Lapata, 2016), which has shown its simplicity and effectiveness. Notice that our method is not specialized to SEQ2SEQ model and it can be used for any neural semantic parsers.

Encoder. Given a sentence $\mathbf{x} = w_1, w_2, ..., w_n$, the SEQ2SEQ model encodes $\mathbf{x}$ using a bidirectional RNN. Each word w_i is mapped to a fixed-dimensional vector by a word embedding function $\phi(\cdot)$ and then

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

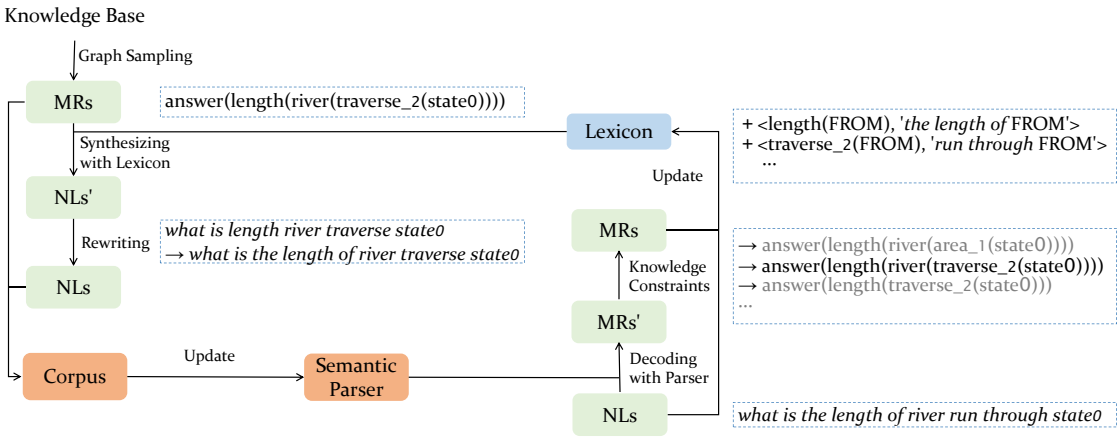


Figure 2: The illustration of our approach. MRs denotes meaning representations, NLs denotes natural language sentences. The naive corpus is synthesized by seed lexicon. In each bootstrapping iteration, the corpus is refined via denoising language model and knowledge-constrained decoding. The data and the models are improved iteratively.

fed into a bidirectional LSTM (Hochreiter and Schmidhuber, 1997). The hidden states in two directions are concatenated $\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$, and the encoding of the whole sentence is: $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n$.

Attention-based Decoder. Given the sentence representation, the SEQ2SEQ model sequentially generates the tokens of its logical form. Specifically, the decoder is first initialized with the hidden states of encoder $\mathbf{s}_0 = [\overrightarrow{\mathbf{h}}_n; \overleftarrow{\mathbf{h}}_1]$. Then at each step t, let $\phi(y_{t-1})$ be the vector of the previous predicted logical form token, the current hidden state $\mathbf{s}_t$ is obtained from $\phi(y_{t-1})$ and $\mathbf{s}_{t-1}$. Then we calculate the attention weights for the current step t, with the i-th hidden state in the encoder:

$$\alpha_t^i = \frac{\exp(\mathbf{s}_t \cdot \mathbf{h}_i)}{\sum_{i=1}^{n} \exp(\mathbf{s}_t \cdot \mathbf{h}_i)} \tag{1}$$

and the next token is generalized from the vocabulary distribution:

$$\begin{aligned} \mathbf{c}_t &= \sum_{i=1}^{n} \alpha_t^i \mathbf{h}_i \\ P(y_t | y_{<t}, \mathbf{x}) &= \text{softmax}(\mathbf{W}_o[\mathbf{s}_t; \mathbf{c}_t] + \mathbf{b}_o) \end{aligned} \tag{2}$$

where $\mathbf{W}_o \in \mathbb{R}^{|V_y| \times 3n}$, $\mathbf{b}_o \in \mathbb{R}^{|V_y|}$ and $|V_y|$ is the output vocabulary size.

Learning. Given a training corpus consisting of <utterance, logical form> pairs, the SEQ2SEQ model is trained by optimizing the objective function:

$$J = -\sum_{(\mathbf{x},\mathbf{y}) \in \mathbf{D}} \sum_{t=1}^{m} \log p(y_t | y_{<t}, \mathbf{x}) \tag{3}$$

where $\mathbf{D}$ is the corpus, $\mathbf{x}$ is the utterance, $\mathbf{y}$ is its logical form label.

2.2 SCFG for Data synthesization

Wang, Berant, and Liang (2015) use a synchronous context-free grammar(SCFG) to generate logical forms paired with canonical utterances, and use crowdsourcing to paraphrase these canonical utterances into natural utterances. The SCFG consists of a set of production rules (lexicon): $N \rightarrow \langle \alpha, \beta \rangle$, where N is a non-terminal, and α and β are sequence of terminal and non-terminal symbols. Any non-terminal symbol in α is aligned to the same non-terminal symbol in β, and vice versa. Therefore, SCFGs define a set of joint derivations of aligned pairs of strings. The seed lexicon in OVERNIGHT is specified by the

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

builder containing types, entities, and properties in databases. Type checking is also performed to rule out some uninterpretable canonical utterances.

3 Approach

This section describes our data synthesis and iterative refinement method for semantic parsing. Firstly, we generate a naive training corpus by sampling meaning representations from knowledge bases and synthesizing their utterances using a grammar-based algorithm. Then, to reduce the noise and eliminate the gap with real corpus, we propose to iteratively refine the data and the model by rewriting synthesized utterances via a denoising language model and generating meaning representations via knowledge-constraint decoding. Figure 2 shows the overview of our approach and we describe all components in detail as follows.

3.1 Data Synthesis

In OVERNIGHT (Wang et al., 2015) and PARASEMPRE (Berant and Liang, 2014), they use simple grammars to generate logical forms paired with canonical utterances. To generate corpus from scratch, we also synthesize data via a grammar-based algorithm.

Specifically, we first sample MRs from knowledge bases via a graph sampling algorithm, then we synthesize their utterances by mapping predicates to words from a seed lexicon and composing these words using context free grammars. Different from the corpus generation method in OVERNIGHT, our method starts from not only grammar but also the knowledge base schema, and can be easier to extended to other datasets like GEO and ATIS.

Generating MRs via Graph Sampling

The graph sampling algorithm aims to sample meaning representations from knowledge bases. Given a knowledge base, Graph Sampling regards MRs as subgraphs of the knowledge base. To ensure the truthfulness and integrality of generated meaning representations, we sample subgraph-based MRs according to both the structure of MRs and the schemas of knowledge bases.

Specifically, to generate MRs, we start from the nonterminal token `root` and then recursively expand all nonterminal tokens in current MRs. For general/functional nonterminal tokens such as `root`, `argmax` and `count`, because they are domain-independent, we expand them using hand-crafted general production rules. For nonterminal tokens about entities and relations such as `river`, `state` and `city` for GEO, because they are domain dependent, we expand them by production rules sampled from knowledge base schemas.

To utilize the schema to produce MRs, we extend the original schema by adding the attribute value as

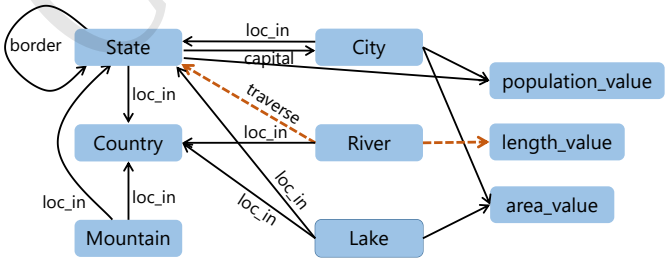


Figure 3: The extended schema of GEO (partial). To sample the subgraph from the dotted edges , the root nonterminal token `root` is recursively extended by the production rules:

`root` → `answer(length_value)`
`length_value` → `length(river_set)`
`river_set` → `river(river_attri)`
`river_attri` → `traverse_2(state_set)`
`state_set` → `state0`

, generating the MR: `answer(length(river(traverse_2 (state0))))`

value type nodes and the aggregation operations as self-loop edges. We provide the extended schema and sampling examples in the Fig 3.

Based on the schema graph, the meaning representations can be effectively sampled by utilizing context-free grammar (i.e., the production rules) for grammatical correctness and knowledge base schemas for semantic correctness.

Synthesizing Utterances via SCFG-based Algorithm

Based on canonical compositionality assumption in Wang, Berant, and Liang (2015), we also use SCFG to generate utterances. We extend the context-free grammar in Graph Sampling to synchronous context-free grammar. For example in Fig 2, based on the SCFG rules, we can synthesize the utterance "what is length river traverse state0" from the sampled MR:

$$\texttt{root} \rightarrow \langle \texttt{answer}\,(\textsc{Form})\,, \text{what is } \textsc{Form}\rangle$$
$$\textsc{Form} \rightarrow \langle \texttt{length}\,(\textsc{Form})\,, \text{length } \textsc{Form}\rangle$$
$$\textsc{Form} \rightarrow \langle \texttt{river}\,(\textsc{Form})\,, \text{river } \textsc{Form}\rangle$$
$$\textsc{Form} \rightarrow \langle \texttt{traverse\_2}\,(\textsc{Form})\,, \text{traverse } \textsc{Form}\rangle$$
$$\textsc{Form} \rightarrow \langle \texttt{state0}, \text{state0}\rangle$$

Seed Lexicon Construction To synthesize utterances from sampled semantic representations, a lexicon is further needed for SCFG, which maps logical tokens to their natural language words. For OVERNIGHT, we simply use its original seed lexicon. For other datasets, we use the following simple way to build an initial lexicon:

For domain-general logical tokens we manually write their natural language templates. The number of domain-general rules is usually very small. Some examples of our domain-general rules are in Table 1.

Category	Domain-general Rules	NL Templates
Query	answer (FORM)	what is FORM
Count	count (FORM)	the number of FORM
Exclusion	exclude (FORM_1, FORM_2)	FORM_1 do not FORM_2
Superlative(max)	largest_one (VALUE (FORM))	FORM with largest VALUE
Filter(type)	λtλs: ($t $s)	$t $s
Filter(property)	λpλvλs: ($p $v $s)	$s whose $p is $v
Comparative(<)	λpλvλs: (< ($p $v) $s)	$s whose $p is smaller than $v
Superlative(max)	λpλs: argmax $s ($p $s)	$s with largest $p

Table 1: Examples of our domain-general rules on GEO (above) and ATIS (below). We write seed lexicon of domain-general grammar manually, the number of which is usually very small (only 5 needed in GEO and 12 in ATIS and 23 in OVERNIGHT).

For domain-dependent entity tokens and relation tokens, we simply use the words in their logical tokens, with a simple preprocessing which removes numbers and underlines. For example, the `area_1` denotes the words "area" and `departure_time` denotes the words "departure time".

Using the above SCFG with seed lexicon, an initial training corpus can be synthesized. Although, this seed lexicon is obviously with limited coverage and lack of diversity. This naive corpus can still provide a helpful start for semantic parsing. Next, we describe how to iterative refine the parsing mode and data.

3.2 Iterative Data-Model Refining

Due to the limitation of grammar and lexicon, the synthesized training instances in naive corpus are often noisy, differing from real-world sentences, and with limited diversity. To address these issues, we refine the corpus with the knowledge of language models and knowledge base constraints through a bootstrapping process: 1) we rewrite synthesized utterances via a denoising language model, so the utterances will be more fluent and closer to natural utterances; 2) we propose to exploit knowledge during decoding, so that meaning representations can be more accurately predicted even when the model is not strong enough; 3) we iteratively refine the data and the model via a bootstrapping process. After several iterations of refinement, we obtain the mature corpus and the final semantic parser.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Utterance Rewriting via Denoising Language Model

The synthesized utterances are often not fluent, differing from real-world sentences. For example, the synthesized utterance in Fig 2: "*what is length river traverse state*" is very different to its natural expression "*what is the length of river traverses state0*". And this discrepancy misleads models to learn incorrect patterns.

Thanks to the current powerful language models, we can use a denoising language model to rewrite synthesized utterances to more natural sentences. Specifically, we regard the synthesized utterances as a noisy version of natural expressions, and then denoise them via neural language model-based language denoising techniques (Lample et al., 2018).

Specifically, we train a language model based on GPT2.0 (Radford et al., 2019), which is then used to denoise by minimizing:

$$\mathcal{L}^{lm} = \mathbb{E}_{x \sim \mathbf{X}}[-\log P(x|C(x))] \tag{4}$$

where C is a noise model with some words dropped and swapped as in Lample et al. (2018).

Generating High-quality Lexicon via Knowledge-Constrained Decoding

To obtain high-quality lexicon, which can be used to synthesize better ⟨MR, canonical utterance⟩ pairs, we use the current parser to generate parallel data. Without manually annotated corpus, the initial semantic parser is often not strong enough, therefore it is difficult to find high-quality meaning representations. So we also apply knowledge-constrained decoding.

Like previous work (Xiao et al., 2016; Krishnamurthy et al., 2017; Yin and Neubig, 2017), we decode the meaning representations under the grammar we mentioned in Graph Sampling. Only the grammatical logical forms are generated during the decoding. Additionally, we leverage knowledge base schemas to effectively filter out illegal logical forms. Given a semantic parser, we first obtain the top K meaning representations for each sentence. Then if there exists an executing program or search engine for logical forms, we will only keep the executable logical forms. Otherwise, we verify whether the logical form is well-typed under the knowledge base schema constraints, and only preserve the eligible logical forms.

After obtaining the higher quality parallel data, following Wong and Mooney (2006), we apply the GIZA++ on the parallel data to get the alignments between words and grammar rules and induce a new SCFG lexicon.

Iterative Learning

It is obviously that the model promotion and the data refining can reinforce each other: better parsers can generate data of higher quality, and higher quality data can be used to train stronger models. Based on this intuition, we propose to iteratively refine model and data by leveraging the duality between them.

Specifically, in each data-model refining iteration, we: 1) first synthesize the utterances $\mathbf{X}'$ of the sampled MRs $\mathbf{Y}'$ using the current lexicon and the denoising model; 2) train a new semantic parser using the synthesized data; 3) parse the unlabeled utterances via knowledge-constrained decoding; 4) induce a new lexicon using both the highly confident automatically labeled data and the synthesized data.

We gradually increase the proportion of parsing data at each iteration. In the k-th iteration, we select the top $\delta \times (k+1)$ confident parsing pairs for lexicon learning. The confidence scores are calculated as the normalized likelihood:

$$Score(x, y) = \frac{1}{N_y} \log P(y|x) \tag{5}$$

4 Experiments

4.1 Experimental Settings

Datasets We conduct experiments on three standard datasets: GEO, and ATIS, OVERNIGHT, which use different meaning representations and contain different domains.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

		Bas.	Blo.	Cal.	Hou.	Pub.	Rec.	Res.	Soc.	Avg.
Supervised										
SEQ2SEQ		84.3	57.9	78.1	69.9	76.2	80.7	78.0	80.5	75.7
RECOMBINATION (Jia and Liang, 2016)		85.2	58.1	78.0	71.4	76.4	79.6	76.2	81.4	75.8
CROSSDOMAIN (Su and Yan, 2017)		86.2	60.2	79.8	71.4	78.9	84.7	81.6	82.9	78.2
SEQ2ACTION (Chen et al., 2018)		88.2	61.4	81.5	74.1	80.7	82.9	80.7	82.1	79.0
DUAL (Cao et al., 2019)		87.5	63.7	79.8	73.0	81.4	81.5	81.6	83.0	78.9
Unsupervised (with nonparallel data)										
Two-stage (Cao et al., 2020)		64.7	53.4	58.3	59.3	60.3	68.1	73.2	48.4	60.7
WmdSamples (Cao et al., 2020)		31.9	29.0	36.1	47.9	34.2	41.0	53.8	35.8	38.7
Mature Corpus + Samples		58.5	55.3	62.4	65.1	66.7	62.2	72.3	47.1	**61.2**
Unsupervised										
Cross-domain Zero Shot* (Herzig and Berant, 2018)		-	28.3	53.6	52.4	55.3	60.2	61.7	-	-
GENOVERNIGHT (Wang et al., 2015)		15.6	27.7	17.3	45.9	46.7	26.3	61.3	9.7	31.3
Naive Corpus	EMBED BERT	15.9	24.6	18.6	44.1	46.9	27.0	62.2	9.7	31.1
	Glove	16.2	23.6	16.2	30.3	36.9	27.0	43.2	9.2	25.3
	Rand	13.8	21.1	15.6	28.2	21.9	27.0	31.1	8.2	20.9
Mature Corpus	EMBED BERT	**45.9**	**52.5**	**52.7**	**58.5**	**61.9**	**52.1**	**69.8**	**33.6**	**53.4**
	Glove	44.1	51.5	48.5	56.4	58.8	50.2	68.9	32.0	51.3
	Rand	35.1	43.2	36.5	44.7	46.9	46.5	65.0	25.6	42.9
	w/o Denoising	32.8	45.0	40.1	46.8	52.5	45.6	63.1	26.6	44.1
	w/o Constraint	29.0	39.7	35.3	37.8	41.9	42.8	64.7	23.4	39.3

Table 2: Accuracies on OVERNIGHT. The previous methods with superscript $*$ means they use different unsupervised settings.

GEO This is a semantic parsing benchmark about U.S. geography (Zelle and Mooney, 1996). The variable-free semantic representation FunQL (Kate et al., 2005) is used in this dataset. We follow the standard 600/280 train/test instance splits.

ATIS This is a large dataset, which contains 5,410 queries to a flight booking system. Each question is annotated with a lambda calculus query. Following Zettlemoyer and Collins (2007), we use the standard 4,473/448 train/test instance splits in our experiments.

OVERNIGHT OVERNIGHT contains natural language paraphrases paired with lambda DCS logical forms across eight domains. We evaluate on the standard train/test splits as Jia and Liang (2015).

In all our experiments, we only use the unlabeled sentences in each dataset. The standard accuracy is used to evaluate different systems, which is obtained as the same as Jia and Liang (2016).

Synthesized Training Corpus We generate training instances proportional to the original dataset sizes (1500 for GEO, 5000 for ATIS, and 1500 for each domain in OVERNIGHT). For OVERNIGHT, we use its original defined grammar and lexicon.

Denoising Language Model We train an individual denoising language model for each dataset (each domain for OVERNIGHT). For each utterance in unlabeled queries, we sample 5 noisy sentences to construct the training pairs by dropping words randomly or slightly shuffling the utterance as Lample et al. (2018). The pretrained language model GPT2.0 is adapted on paraphrase generation dataset, then fine-tuned on denoising sentences with 15 epochs and the learning rate of 1e-5.

System Settings We train all our models with 5 data-model refining iterations. In each iteration, the neural semantic parser is trained 15 epochs, with the initial learning rate of 0.001. We use Adam algorithm (Kingma and Ba, 2015) to update parameters, with batch size is 20. Our model uses 200-dimensional hidden units and 200-dimensional word vectors for sentence encoding. We initialize all parameters by uniformly sampling within [-0.1, 0.1]. $\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2019) is uesd to get word representations. The beam size K during decoding is 5. The hyper-parameter δ is 0.1. Following Dong and Lapata (2016), we handle entities with a Replacing mechanism, which replaces identified entities with their types and IDs.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	GEO	ATIS
Supervised		
SEQ2SEQ	88.2	84.2
Dong and Lapata (2016)	87.1	84.6
Jia and Liang (2016)	89.3	83.3
Susanto and Lu (2017)	90.0	-
Xu et al. (2018)	88.1	85.9
Chen, Sun, and Han (2018)	88.9	85.5
Jie and Lu (2018)	89.3	-
Guo et al. (2020)	87.1	83.1
Unsupervised		
Confidence-driven*	66.4	-
Two-stage*	63.7	-
Naive Corpus	29.3	25.0
Mature Corpus		
EMBED BERT	**58.2**	**52.9**
GloVe	55.0	52.5
Rand	44.6	43.3
w/o Denoising	45.0	39.5
w/o Constraint	38.9	37.1

Table 3: Accuracies on GEO and ATIS. The previous methods with superscript $*$ means they use different unsupervised settings. Confidence-driven and Two-stage both use the nonparallel data.

4.2 Experimental Results

Overall Results

We compare our model with different settings:

1) **Naive Corpus** – the semantic parser is trained from the naive corpus, which is generated by meaning representation sampling and utterance synthesizing;

2) **Mature Corpus** – the corpus is improved by iterative data-model refining;

3) **Supervised** – the model is trained using the original training corpus with the same settings.

For Overnight, we further compare with the Cross-domain Zero Shot (Herzig and Berant, 2018) which is trained on other source domains and then generalized to new domains and GENOVERNIGHT (Wang et al., 2015) in which all the canonical utterances are also generated without manual annotation. With the nonparallel data: Two-stage (Cao et al., 2020) employs the cycle learning framework. WmdSamples (Cao et al., 2020) labels each input sentences with the most possible outputs in the unparallel corpus and deals with these faked samples in a supervised way. Our Mature Corpus + Samples method follows WmdSamples, using the parser built on the refined data to label each input.

The results are shown in Table 2 and Table 3. We can see that:

1) **Our learning framework is promising for resolving the training data bottleneck problem of semantic parsing.** In all datasets, our method outperforms other baselines in the same unsupervised settings. On OVERNIGHT, our method also surpasses the previous approaches in unsupervised data settings. These results verify that data synthesis and iterative data-model refinement is a promising method for semantic parsing without annotated logical forms.

2) **The iterative data-model refining is effective to bootstrap semantic parsers.** Compared with Naive Corpus, after corpus refinement our Mature Corpus gains 27.9 accuracy improvement in ATIS. This verifies the effectiveness of the data-model refining. We believe it results from: i) denoising language model can improve the quality of generated utterances and knowledge-constrained decoding can filter out invalid meaning representations; ii) the bootstrapping can leverage the duality between data and model for iterative refining.

Detailed Analysis

Effects of Utterance Denoising and Constrained Decoding. Table 2 and 3 show the accuracies by removing denoising language model (–Denoising) and by removing knowledge constraints during decoding (–Constraint). We can see that: 1) Both utterance denoising and constrained decoding are effective. In

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

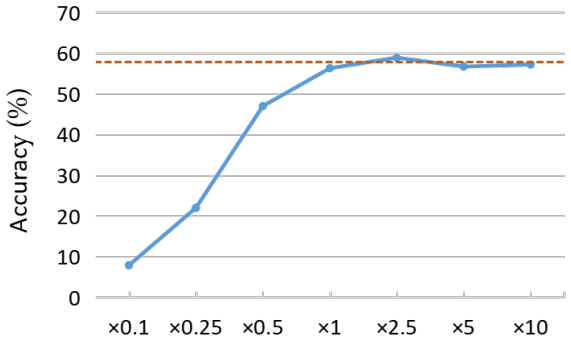


Figure 4: Test accuracies on GEO with different size of synthesized data. The number of sampled meaning representations has increased from 0.1 times the amount of original data to 10 times. The dash line shows the accuracy of Golden MRs

	GEO	ATIS
Iterative Updating		
Iter.1	41.4	37.7
Iter.2	49.3	44.6
Iter.3	57.1	48.0
Iter.4	58.9	52.5
Iter.5	58.2	52.9

Table 4: Evaluation Accuracies on GEO and ATIS with the increase of iterations.

average on all three datasets, removing denoising results in 12.0 accuracy drop and removing constrained decoding results in 16.4 accuracy drop. 2) Constrained decoding is more helpful than denoising. We believe this is because the grammar and the knowledge-base can effectively improve the quality of automatically generated parallel data, from which a new lexicon is built and is further used to synthesize new parallel data.

Effects of Word Embeddings. To analyze the effects of word embeddings settings, we compare our method with different settings of word embeddings: BERT – word representations are from the pretrained $\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2019); GloVe – word embeddings are initialized by GloVe (Pennington et al., 2014); Rand – the word embeddings are initialized by uniformly sampling within the interval [-0.2, 0.2], and the unseen words are all presented as UNK token. We can see that the pretrained word embeddings can effectively improve the model. We believe this is because it empowers the model with better representation and helps the model generalize to similar words.

Effect of Data Synthesis. To analyze the effectiveness of synthesized data, we: 1) compare our models with Golden MRs – in which all utterances are synthesized from the manually labeled meaning representations in original corpus; 2) increase the amount of sampled meaning representations from $\times 0.1$ to $\times 10$ size of the original labeled data. The results on GEO are shown on Figure 4.

We can see that: 1) the graph sampling algorithm can effectively sample meaning representations – compared with Golden-MRs, our method can achieve nearly the same performance with $\times 1$ dataset. 2) The data synthesis is useful, when the size of data increases from $\times 0.1$ to $\times 1$, the performance gradually increases. We also noticed that when the data size exceeds the original data, the performance of the model does not improve much. We believe that this is because too much data generated with a certain amount of noise can no longer provide useful supervision information.

Effect of Iterative Bootstrapping. Table 4 shows the accuracies by increasing the number of iterations. We can see that: 1) the iterative data-model refining is effective: when we conduct more refining iterations, the performance gradually increases and stabilizes at a reasonable level – from 41.4 accuracy in Iter 1 to

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

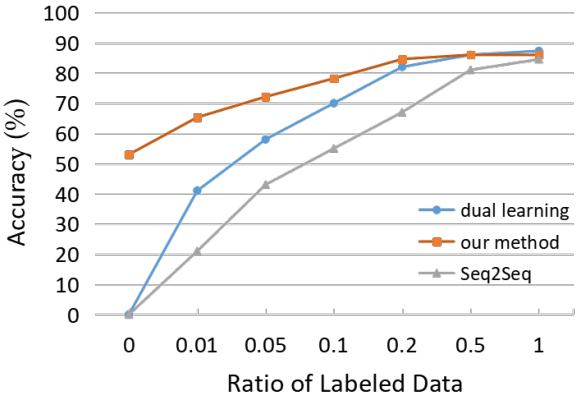


Figure 5: Test accuracies on ATIS with different amounts of labeled data.

58.9 in Iter 4 in GEO; 2) The bootstrapping process can reach its equilibrium within few iterations: for GEO in 5 iterations and for ATIS in 4 iterations.

Semi-supervised learning. To investigate the effectiveness of our method given some additional labeled instances, we vary the amount of labeled data from 0 to all labeled data. Our model can use the labeled data to train semantic parser and induce lexicon in each iteration. Seq2Seq can only use the labeled data. Dual learning (Cao et al., 2019) forms a closed loop to learn unlabeled data in reinforcement learning. In Figure 5, We can see that our model enhances semantic parsing over most settings. Especially, our model has obvious advantages when there is a small amount of labeled data.

5 Related Work

Neural semantic parsers In recent years, neural semantic parsers have achieved significant progress. Neural parsers model semantic parsing as a sentence to logical form translation task (Xiao et al., 2016; Jia and Liang, 2016; Iyyer et al., 2017; Jie and Lu, 2018), And many constrained decoding algorithms are also proposed (Krishnamurthy et al., 2017; Liang et al., 2017; Iyyer et al., 2017; Chen et al., 2018);

Data scarcity in semantic parsing Witnessed the labeled data bottleneck problem, many techniques have been proposed to reduce the demand for labeled logical forms. Many weakly supervised learning are proposed (Artzi and Zettlemoyer, 2013; Berant et al., 2013; Reddy et al., 2014; Agrawal et al., 2019), such as denotation-base learning (Pasupat and Liang, 2016; Goldman et al., 2018), iterative searching (Dasigi et al., 2019). Semi-supervised semantic parsing is also proposed, such as variational auto-encoding (Yin et al., 2018), dual learning (Cao et al., 2019), dual information maximization (Ye et al., 2019), and back-translation (Sun et al., 2019). Constrained language models are also proposed to resolve few-shot semantic parsing (Wu et al., 2021; Shin et al., 2021).

Unsupervised semantic parsers There are also some unsupervised semantic parsers, such as USP (Poon and Domingos, 2009) proposes the first unsupervised semantic parse, and GUSP (Poon, 2013) builds semantic parser by annotating the dependency-tree nodes and edges. Wang et al. (2011) select high confidence pairs for unsupervised learning. Two-stage (Cao et al., 2020) train unsupervised paraphrasing model with non-parallel data for semantic parsing.

6 Conclusions

We propose a data synthesis and iterative data-model refining algorithm for neural semantic parsing, which can build semantic parsers without labeled data. In our method, the naive corpus is generated from scratch by grammar-based method and knowledge base schemas, and the corpus is improved on bootstrapping to refine model and data with the knowledge of language models and knowledge bases constraints. Experimental results show our approach can achieve promising performance in unsupervised settings.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

References

Priyanka Agrawal, Ayushi Dalmia, Parag Jain, Abhishek Bansal, Ashish R. Mittal, and Karthik Sankaranarayanan. 2019. Unified semantic parsing with weak supervision. In *ACL*.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *ACL*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Ben Bogin, Jonathan Berant, and Matt Gardner. 2019. Representing schema structure with graph neural networks for text-to-sql parsing. In *ACL*.

Ruisheng Cao, Su Zhu, Chen Liu, Jieyu Li, and Kai Yu. 2019. Semantic parsing with dual learning. In *ACL*.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *ACL*.

Shuaichen Chang, Pengfei Liu, Yun Tang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Zero-shot text-to-sql learning with auxiliary task. In *AAAI*.

Bo Chen, Le Sun, and Xianpei Han. 2018. Sequence-to-action: End-to-end semantic graph generation for semantic parsing. In *ACL*.

Pradeep Dasigi, Matt Gardner, Shikhar Murty, Luke Zettlemoyer, and Eduard H. Hovy. 2019. Iterative search for weakly supervised semantic parsing. In *NAACL-HLT*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *ACL*.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *ACL*.

Omer Goldman, Veronica Latcinnik, Ehud Nave, Amir Globerson, and Jonathan Berant. 2018. Weakly supervised semantic parsing with abstract examples. In *ACL*.

Dan Goldwasser, Roi Reichart, James Clarke, and Dan Roth. 2011. Confidence driven unsupervised semantic parsing. In *ACL*.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. In *ACL*.

Jiaqi Guo, Qian Liu, Jian-Guang Lou, Zhenwen Li, Xueqing Liu, Tao Xie, and Ting Liu. 2020. Benchmarking meaning representations in neural semantic parsing. In *EMNLP*.

Jonathan Herzig and Jonathan Berant. 2018. Decoupling structure and lexicon for zero-shot semantic parsing. In *EMNLP*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *ACL*.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *ACL*.

Zhanming Jie and Wei Lu. 2018. Dependency-based hybrid trees for semantic parsing. In *EMNLP*.

Rohit J. Kate, Yuk Wah Wong, and Raymond J. Mooney. 2005. Learning to transform natural to formal languages. In *IAAI,*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *EMNLP*.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *EMNLP*.

Chen Liang, Jonathan Berant, Quoc Le, Kenneth D. Forbus, and Ni Lao. 2017. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. In *ACL*.

Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S. Zettlemoyer. 2008. A generative model for parsing natural language to meaning representations. In *EMNLP*.

Dipendra Misra, Ming-Wei Chang, Xiaodong He, and Wen-tau Yih. 2018. Policy shaping and generalized update equations for semantic parsing from denotations. In *EMNLP*.

Panupong Pasupat and Percy Liang. 2016. Inferring logical forms from denotations. In *ACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Hoifung Poon and Pedro M. Domingos. 2009. Unsupervised semantic parsing. In *EMNLP*.

Hoifung Poon. 2013. Grounded unsupervised semantic parsing. In *ACL*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics*.

Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. 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 7699–7715. Association for Computational Linguistics.

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *EMNLP*.

Yibo Sun, Duyu Tang, Nan Duan, Yeyun Gong, Xiaocheng Feng, Bing Qin, and Daxin Jiang. 2019. Neural semantic parsing in low-resource settings with back-translation and meta-learning. *CoRR*.

Raymond Hendy Susanto and Wei Lu. 2017. Semantic parsing with neural hybrid trees. In *AAAI*.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *ACL*.

Yuk Wah Wong and Raymond J. Mooney. 2006. Learning for semantic parsing with statistical machine translation. In *NACL*.

Yuk Wah Wong and Raymond J. Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *ACL*.

Shan Wu, Bo Chen, Chunlei Xin, Xianpei Han, Le Sun, Weipeng Zhang, Jiansong Chen, Fan Yang, and Xunliang Cai. 2021. From paraphrasing to semantic parsing: Unsupervised semantic parsing via synchronous semantic decoding. 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 5110–5121. Association for Computational Linguistics.

Chunyang Xiao, Marc Dymetman, and Claire Gardent. 2016. Sequence-based structured prediction for semantic parsing. In *ACL*.

Kun Xu, Lingfei Wu, Zhiguo Wang, Mo Yu, Liwei Chen, and Vadim Sheinin. 2018. Exploiting rich syntactic information for semantic parsing with graph-to-sequence model. In *EMNLP*.

Hai Ye, Wenjie Li, and Lu Wang. 2019. Jointly learning semantic parser and natural language generator via dual information maximization. In *ACL*.

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. In *ACL*.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. 2018. Structvae: Tree-structured latent variable models for semi-supervised semantic parsing. In *ACL*.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *AAAI*.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI*.

Luke S. Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *EMNLP-CoNLL*.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 761-773, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

EventBERT: Incorporating Event-based Semantics for Natural Language Understanding

Anni Zou[1,2,3], Zhuosheng Zhang[1,2,3], Hai Zhao[1,2,3,*]
[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{annie0103,zhangzs}@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn

Abstract

Natural language understanding tasks require a comprehensive understanding of natural language and further reasoning about it, on the basis of holistic information at different levels to gain comprehensive knowledge. In recent years, pre-trained language models (PrLMs) have shown impressive performance in natural language understanding. However, they rely mainly on extracting context-sensitive statistical patterns without explicitly modeling linguistic information, such as semantic relationships entailed in natural language. In this work, we propose EventBERT, an event-based semantic representation model that takes BERT as the backbone and refines with event-based structural semantics in terms of graph convolution networks. EventBERT benefits simultaneously from rich event-based structures embodied in the graph and contextual semantics learned in pre-trained model BERT. Experimental results on the GLUE benchmark show that the proposed model consistently outperforms the baseline model.

1 Introduction

Recent years have witnessed deep pre-trained language models (PrLM) such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and ERNIE (Sun et al., 2020) significantly prospering the performance of a wide range of natural language understanding (NLU) tasks. The remarkable advancements brought by PrLM have shown the effectiveness of leveraging contextualized representation. However, they mainly rest on extracting context-sensitive statistical patterns without explicitly modeling linguistic information such as semantic relationships in natural language.

It is clear that natural language itself abounds with ample, multi-level linguistic information. Although PrLMs like BERT implicitly represent linguistic knowledge more or less (Rogers et al., 2020), studies disclose that linguistic knowledge is far from fully absorbed (Ettinger, 2020; Rogers et al., 2020). Therefore, there emerges a series of derivatives of PrLM intending to fuse explicit linguistic knowledge so as to acquire better language representation, including syntactic (Bai et al., 2021; Xu et al., 2021; Zhang et al., 2020b) and semantic information (Zhang et al., 2020a; Guo et al., 2020b; Guan et al., 2021).

In cognition practice, human needs to distill semantics of different levels to gain a comprehensive understanding, whereas neural language models learn semantic representation to deal with downstream tasks (Geeraerts and Cuyckens, 2007). Thus, effective learning of semantic knowledge plays a crucial role in NLU tasks and has gained growing attention recently. For instance, Zhang et al. (2020a) proposed SemBERT, which directly connects multiple predicate-argument structures acquired by semantic role labeler (SRL) to get the joint representation.

The essence of SRL (Shi and Lin, 2019) lies in that every sentence possesses multiple predicate-specific structures which can represent different frames of events, while semantic roles express the abstract role that arguments of a predicate can take in the event. Besides, the events inside a sentence have interactions with each other that serve together to present the overall semantic knowledge. As shown

*Corresponding author. This work was supported in part by the Key Projects of National Natural Science Foundation of China under Grants U1836222 and 61733011.

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

in Figure 1, SRL parses every sentence with multiple predicate-specific structures which can serve as events inferring *who did what to whom*, *when and why*. Each event has an inner structure centered on the predicate to which several arguments are associated such as *Hoy[ARG0]*, *the woman's age[ARG1]* and *Tuesday[ARGM-TMP]* connected to *confirmed[V]*. Meanwhile, the multiple events work together to give a comprehensive meaning of a sentence, like the events centered on *said*, *confirmed* and *left*. With regard to delving into the inner interactions between the events and effectively capturing multiple objects, we are motivated to build a graph to reveal the intrinsic structures between and inside the events.

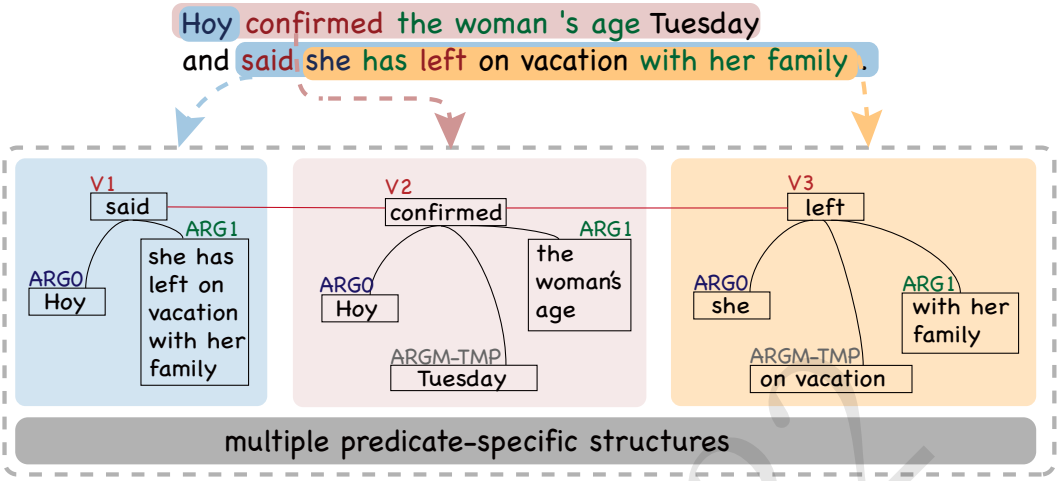


Figure 1: An example showing how SRL parses sentences and the intuition of constructing event-based graph.

Inspired by the above ideas, we propose EventBERT: an event-based semantic representation model which takes BERT as the backbone and refines with event-based structural semantics. Our EventBERT benefits simultaneously from rich event-based structures embodied in the graph and contextual semantics learned in the pre-trained BERT.

Our proposed model works in three steps: it first applies an off-the-shelf SRL toolkit to parse every sentence with semantic role labels; then it constructs event-based graphs and employs Graph Convolutional Networks (GCNs) (Schlichtkrull et al., 2018) to propagate and aggregate information from neighboring nodes on the graph; at last, it combines the contextualized representation acquired by BERT encoder together with the graph-level representation to obtain an event-based contextualized representation.

The key contributions of our work are summarized as follows:

1) We extract event-based semantic knowledge from SRL to enrich language representation.

2) We employ GCNs to construct sentence-level graphs which better reveal interactions inside and between the events in a sentence.

2 Related Work

2.1 Semantics in Language Representation

Recent studies show that current prominent pre-trained language models have already incorporated semantic information to some extent (Clark et al., 2019), yet such implicit semantic information is far from enough for comprehensive natural language understanding (Ettinger, 2020). Thus there emerges a research line that focuses on fusing semantic information into contextualized language representation. ERNIE2.0 (Sun et al., 2020) adopts three-stage masking in which entity-level masking helps to obtain a word representation containing richer semantic information. SemBERT (Zhang et al., 2020a) makes use of PropBank (Palmer et al., 2005) to fuse semantic role tags into language representation. FMSR (Guo et al., 2021) utilizes FrameNet (Baker et al., 1998) to extract multi-level semantic information within sentences. SS-MRC (Guo et al., 2020a) takes advantage of syntax and frame semantics in an attempt to carve out information from two complementary perspectives to obtain richer language representation.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Besides simply employing semantic knowledge, other recent works shift the focus to exploring deeper structural semantics. Guan et al. (2021) leverage frame semantics and graph neural networks to model sentences from both intra-sentence level and inter-sentence level. Wu et al. (2021) introduce SIFT to inject predicate-argument semantic dependencies into pre-trained language models via R-GCNs. Xie et al. (2022) introduce structured knowledge through multi-tasking to get a unified model, which inspires the potential of leveraging structural information. Unlike previous works that attempt to capture shallow semantic structures by semantic tags, our model digs deeper into semantics itself and aims to find the structured event-based information behind semantics, thus unveiling richer structural-semantic information inside the sentence.

2.2 Graph Modeling for Language Understanding

As natural language itself abounds with dependencies and intricate relations between different levels of language units, graph neural networks (GNNs), which model the units as nodes in the graph and learn the weight via the message passing between nodes of the graph (Scarselli et al., 2008; Kipf and Welling, 2016; Velickovic et al., 2017), stand out by explicitly and intuitively capturing the relations. Besides, a number of extensions to the original graph neural networks have been developed, the most notable of which include graph convolutional networks (GCNs) (Kipf and Welling, 2016), graph attention networks (GANs) (Velickovic et al., 2017) and the models from Li et al. (2015) and Pham et al. (2017) utilizing gating mechanisms to facilitate optimization.

In response to the outstanding performance of GCNs, several efforts have been made in recent years to improve performance on natural language understanding using GCNs, including GraphRel (Fu et al., 2019) which considers the interaction between named entities and relations via relation-weighted GCNs to better extract relations, NumNet (Ran et al., 2019) which utilizes a numerically-aware graph to perform numerical reasoning, DFGN (Qiu et al., 2019) which dynamically builds the entity graph by adding the edges with co-occurrence relations, HGN (Fang et al., 2019) which creates a hierarchical graph by constructing nodes on different levels of granularity and social information reasoning (Li and Goldwasser, 2019) which uses GCNs to capture the documents' social context.

Moreover, R-GCNs (Schlichtkrull et al., 2018) have shown effectiveness in relational graph modeling. For example, Entity-GCN (De Cao et al., 2019) employs R-GCNs to link mentions of candidate answers for multi-document question answering. DFGN (Qiu et al., 2019) dynamically builds the entity graph by adding the edges with co-occurrence relations and softly masking out irrelevant entities. DGM (Ouyang et al., 2021) constructs two discourse graphs and uses R-GCNs to fully capture interactions among the elements. Ma et al. (2022) employs R-GCNs to enhance reference dependencies for dialogue disentanglement. In contrast with previous works, our work proposes a sentence-level graph that is finely designed to mine the relationships between multiple elements in a sentence, extract rich structural semantics and facilitate information flow over the graph as well.

3 Model

Figure 2 gives an overview of our proposed EventBERT, which consists of two major components:

1. Context Encoder which acquires deep and contextualized representations for raw input sequences by following BERT architecture;

2. Event-based Encoder which obtains richer structural-semantic representation by modeling event-based intra-sentence graphs.

We omit the details of BERT which is widely used and ubiquitous and leave readers to resort to Devlin et al. (2019) for more information.

3.1 Context Encoder

The raw input sentence $X = \{x_1, \ldots, x_n\}$ is a sequence of words in length n. It is first tokenized to a sequence of sub-words with `[SEP]` inserted at the end as the end marker and `[CLS]` inserted at the beginning to get a sentence-level representation: $X' = \{token_1, \ldots, token_m\}$. Then we pass it

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

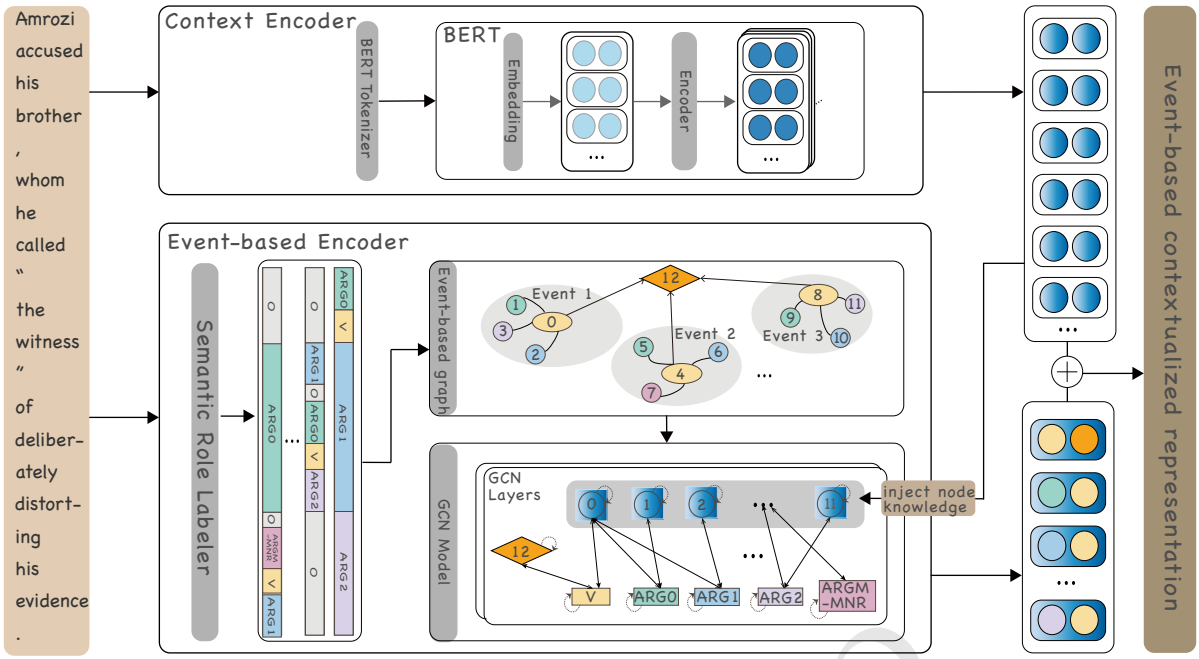


Figure 2: The overall structure of EventBERT.

through the embedding block and encoder block of BERT to produce a context-informed representation $C = \{c_1, \ldots, c_m\} \in \mathbb{R}^{m \times d_{hs}}$ using the equation below:

$$C = BERT(X'), \tag{1}$$

where m denotes the length of sentence on sub-word level and d_{hs} stands for the dimension of hidden states.

3.2 Event-based Encoder

3.2.1 Semantic Role Labeler

The raw input sentence is simultaneously fed into Semantic Role Labeler (Shi and Lin, 2019) to fetch multiple predicate-specific structures tagged by PropBank semantic roles:

$$T = \{t_1, \ldots, t_d\}, \tag{2}$$

where d is the number of semantic structures for one sentence. Notably, t_i can be represented under the format $\{tag_1^i, tag_2^i, ..., tag_n^i\}$ and every tag span in t_i is recorded with its corresponding index in the context for further alignment.

3.2.2 Graph Construction

Figure 3 shows the process of graph construction: the predicates in the original input text are firstly extracted and an event subgraph is constructed with each predicate as the center; then a super event node (SEN) is applied to link all the predicates to collect the integral event information within the aggregated sentence; the Levi graph is finally constructed with reference to the method of Levi (1942), which is used to prepare the next stage of further computational operations on the graph.

For each sentence with the argument-predicate roles, we construct an event-based graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ with span-level nodes $v_i \in \mathcal{V}$ and labeled edges $(v_i, r, v_j) \in \mathcal{E}$, where $r \in \mathcal{R}$ a relation type. Since every sentence has several semantic structures, here we take one structure as example and show the modeling method. Given $Seq_{tag} = \{tag_1, tag_2, ..., tag_n\}$ a word-level tag sequence,

1. We first transform it to a span-level sequence $Seq'_{tag} = \{tag'_1, tag'_2, ..., tag'_l\}$ by aggregating the same neighboring tags with $l \leq n$ representing the length of tags on span-level;

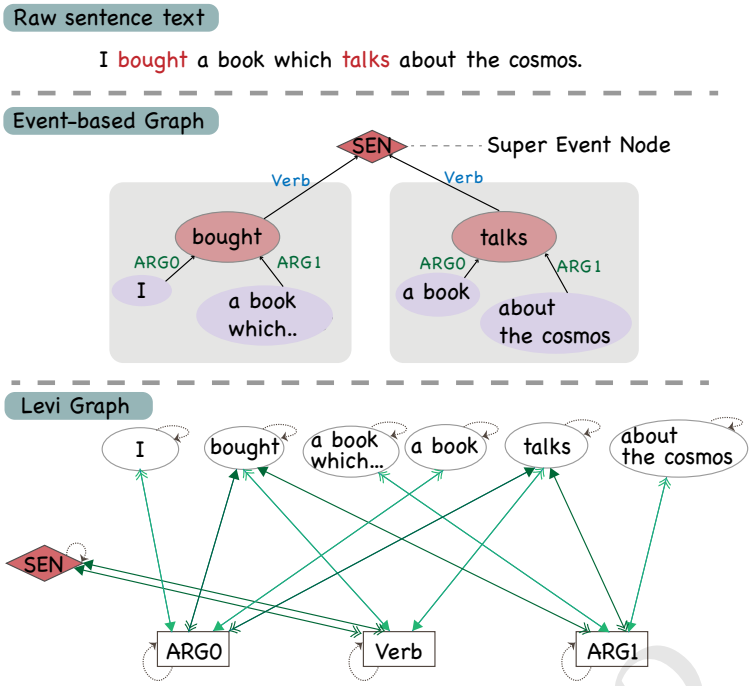


Figure 3: The process of graph construction: from raw sentence text to event-based graph and corresponding Levi graph.

2. Then, we add a Super Event Node ($v = SEN$) to seize global graph information;

3. After that, we add other nodes and edges to G based on the following process:
   (a) we first find tag'_p which corresponds to predicate (*Verb* in e'),
   (b) we add a node $v = n_p$ and a directed edge $e = (n_p, Verb, SEN)$ with $r = Verb$,
   (c) for the rest tags referring to arguments of the predicate, tag'_q for example, we add a node $v = n_q$ and a directed edge linking to the predicate $e = (n_q, tag'_q, n_p)$ with relation $r = tag_q$;

4. Finally, the corresponding Levi graph (Levi, 1942) is extended from G to $G_L = (\mathcal{V}_L, \mathcal{E}_L, \mathcal{R}_L)$. For nodes $\mathcal{V}_L$, we add the nodes representing relations to the original: $\mathcal{V}_L = \mathcal{V} \cup \mathcal{R}$. For edges $\mathcal{E}_L$, we transform each edge $e = (n_q, tag'_q, n_p)$ in G into two corresponding edges: $e_1 = (n_q, tag'_q)$ and $e_2 = (tag'_q, n_p)$ in G_L. For $\mathcal{R}_L$, we follow the setting of Ouyang et al. (2021) and refine it to five types: *default-in*, *default-out*, *reverse-in*, *reverse-out*, *self* according to the direction of edges towards the relation vertices, as is shown in Table 1.

Table 1: Relation types in our extended Levi graph

$\mathcal{R}_L$ in Levi graph	Illustration
default-in	the propagation path pointing to the node as the end point
default-out	the propagation path pointing to the node as the starting point
reverse-in	the propagation path in the opposite direction of *default-in*
reverse-out	the propagation path in the opposite direction of *default-out*
self	the propagation paths pointing to the node itself

3.2.3 Event-based Contextualized Representation

We adopt Relational Graph Convolutional Networks (R-GCNs) (Schlichtkrull et al., 2018) to implement explicit event graphs since traditional Graph Convolutional Networks (GCNs) cannot handle graphs con-

CCL 2022

taining edge features with multiple relations. For predicate and argument nodes, we inject the corresponding span-level encoding results obtained from Context Encoder in Section 3.1. For relation nodes, we regard the relations as embeddings and use a lookup table to get the initial representation. Given that the initial representation of each node v_i is h_i^0, the propagation process can be written as:

$$h_i^{(l+1)} = \text{ReLU}\left(\sum_{r\in\mathcal{R}_L}\sum_{v_j\in\mathcal{N}_r(v_i)}\frac{1}{c_{i,r}}w_r^{(l)}h_j^{(l)}\right), \tag{3}$$

where $h_i^{(l)} \in \mathbb{R}^{d^{(l)}}$ is the hidden state of node v_i in layer l with $d^{(l)}$ being the dimensionality of this layer's representations, $\mathcal{N}_r(v_i)$ denotes the set of neighbor indices of node v_i under the relation r, $c_{i,r}$ is a problem-specific normalization constant equal to $|\mathcal{N}_i^r|$, $w_r^{(l)}$ is the learnable parameters of layer l.

Since the importance of these relations cannot be treated the same, for example, the relation *Verb* is much more important than the relation *ARG2*, we introduce the gating mechanism (Marcheggiani and Titov, 2017). The basic idea is to compute a value between 0 and 1 for message passing control as is shown in Equation 4. Finally, the propagation process of R-GCNs under the gating mechanism is as follows:

$$g_j^{(l)} = \text{Sigmoid}\left(h_j^{(l)}W_{r,g}^{(l)}\right) \tag{4}$$

$$h_i^{(l+1)} = \text{ReLU}\left(\sum_{r\in\mathcal{R}_L}\sum_{v_j\in\mathcal{N}_r(v_i)}g_j^{(l)}\frac{1}{c_{i,r}}w_r^{(l)}h_j^{(l)}\right), \tag{5}$$

where $W_{r,g}^{(l)}$ is the learnable parameter under the l-th level relation type r.

With R-GCNs model, we obtain a graph-level semantic representation:

$$R = \{r_1, \ldots, r_f\} \in \mathbb{R}^{f\times d_{hs}} \tag{6}$$

where f is the number of nodes in the graph and d_{hs} is the same dimension as the representation C in Equation 1 obtained from the context encoder.

At last, we concatenate R with the contextual sub-word-level representation C provided by Context Encoder and generate an event-based contextualized representation taking the mean value of both sub-word-level and graph-level information, which is then used as the new sequence representation for downstream tasks following the same way of Devlin et al. (2019).

4 Experiments

4.1 Setup

4.1.1 Datasets

We build EventBERT on the BERT backbone and fine-tune the model on GLUE (General Language Understanding Evaluation) benchmark (Wang et al., 2018) to evaluate the performance, which includes two single-sentence tasks CoLA (Warstadt et al., 2018), SST-2 (Socher et al., 2013)), three similarity and paraphrase tasks MRPC (Dolan and Brockett, 2005), STS-B (Cer et al., 2017), QQP (Chen et al., 2018) , three inference tasks MNLI (Nangia et al., 2017), QNLI (Rajpurkar et al., 2016), RTE (Bentivogli et al., 2009). We exclude the controversial and problematic dataset WNLI (Levesque et al., 2012).

4.1.2 Evaluation Metrics

According to Wang et al. (2018), different datasets in GLUE correspond to different evaluation metrics, which include accuracy (acc), Matthew's correlation (mc) and Pearson correlation (pc). Among the eight datasets, STS-B is reported by Pearson correlation, CoLA is reported by Matthew's correlation, and other tasks are reported by accuracy.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	CoLA (mc)	SST-2 (acc)	MNLI (acc)	QNLI (acc)	RTE (acc)	MRPC (acc)	QQP (acc)	STS-B (pc)	Avg -
Base-size									
$\text{BERT}_{\text{BASE}}$	58.4	92.8	83.2	88.6	68.5	86.0	86.5	87.8	81.5
$\text{EventBERT}_{\text{BASE}}$	59.6	93.3	83.9	91.8	69.7	89.7	89.8	88.9	83.3(↑1.8)
Large-size									
$\text{BERT}_{\text{LARGE}}$	60.3	93.1	85.2	91.5	70.3	88.5	90.2	89.3	83.6
$\text{EventBERT}_{\text{LARGE}}$	63.1	94.0	85.3	92.6	71.4	89.5	90.6	89.5	84.5(↑0.9)

Table 2: Comparisons between our models and baseline models on GLUE dev set. STS-B is reported by Pearson correlation, CoLA is reported by Matthew's correlation, and other tasks are reported by accuracy.

4.1.3 Implementation Details

For the experiments, we use an initial learning rate in {1e-5, 2e-5, 3e-5} with warm-up rate of 0.1 and L2 weight decay of 0.01. The batch size is selected in {16, 32}. The maximum number of epochs is set in $[2, 5]$ depending on tasks. Texts are tokenized with maximum length of 256 for the tasks. We use 2 layers of R-GCNs in our model.

4.2 Results

Table 2 presents the results on the GLUE benchmark, which show that EventBERT achieves consistent gains over all the subtasks under both base and large models.

The results indicate that our model performs better on longer sentences as shown in Section 5.3. Furthermore, our analysis shows that EventBERT can effectively benefit from the fine-grained graph-like event-based structures, as illustrated in case studies in Section 5.4. The results also disclose that modeling intrinsic structures between and inside events is crucial for language understanding.

In addition, the experimental results show that EventBERT has a significant performance gain on small datasets such as CoLA and MRPC, which indicates that semantic information involving event modeling is more advantageous and competitive in smaller datasets. In practice or industry, large-scale annotated data is rare and scarce due to the high cost and required expensive human resources, so language models that dominate in small-scale datasets are more valuable and important for most NLP tasks.

5 Analysis

5.1 Ablation Study

We conduct the ablation study to investigate the effects of the gating mechanism and the addition of global nodes in the event-based encoder module. Results in Table 3 show that both the gating mechanism and global nodes are non-trivial.

5.2 Methods of Aggregation

During the period of concatenating and aggregating the graph level semantic representation R and the contextual representation C, we further analyze the influence of different methods of aggregation such as max-pooling and mean-pooling by comparing the models with the same hyper-parameters on three datasets CoLA, MRPC and RTE respectively. Results in Table 3 demonstrate that employing mean-pooling presents better performance.

Model	CoLA (mc)	MRPC (acc)	RTE (acc)
Ablation study			
EventBERT_{base}	59.6	89.7	69.7
w/o gating	58.6	86.8	69.0
w/o global node	58.4	87.0	67.9
Aggregation methods			
BERT_{base}	58.4	86.0	68.5
w/ max-pooling	59.1	86.8	68.2
w/ mean-pooling	59.6	89.7	69.7

Table 3: Ablation study and comparison of aggregation methods on three datasets.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

5.3 Effectiveness of semantic structures

In order to dig deeper into the rationale behind the effectiveness of the model, we select two datasets QNLI and MRPC, representing large-scale and small-scale datasets respectively. We statistically calculate the accuracy of the corresponding models on different word-level sequence length intervals for EventBERT and baseline. Figure 4 shows that our model outperforms the baseline especially when the sequence is relatively long and our model performs better on longer sentences compared with shorter ones, which implies that modeling intrinsic semantic structures is potential to guide the model to learn richer structural semantics more than contextualized information. Thus, the analysis of word sequence lengths shows that EventBERT performs better on data with longer sequence lengths, which indicates that event-level modeling is promising and competitive for understanding long texts. Under many practical situations where available data are long texts, the idea of extracting event-level structural-semantic information is promising in many NLP tasks.

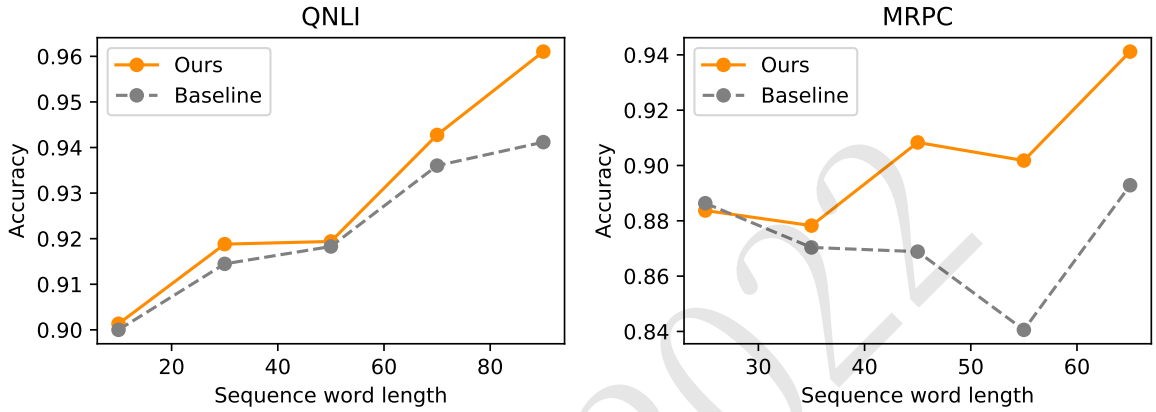


Figure 4: Accuracy of different sequence word lengths on QNLI and MRPC.

5.4 Interpretability: Case Study

We select three cases in Classification, Sentence Similarity and Language Inference from SST-2, MRPC and QNLI respectively which are shown in Figure 5, aiming to further explore the mechanism. It can be seen that our model can perceive explicit structural meaning to better understand the language. We will analyze each of the three cases in detail so as to analyze the advantages of EventBERT more intuitively.

5.4.1 Classification

In the case from SST-2, our model succeeds in capturing and understanding the event *Friel and william's exceptional performances[ARG0] anchored[V] the film's power[ARG1]*, whereas the baseline does not manage to capture this meaning, thus leading to the failure.

5.4.2 Sentence Similarity

The case from MRPC demonstrates that our model grabs the distinct semantic structures centered on *is* and *has* and thus gives the right answer *not equivalent*. The event centered on the predicate *donate* belongs to the same structure, which contains the arguments *ARG0*, *ARG1* and *ARGM-TMP* having the same contents (i.e., *the woman donated blood*). Nevertheless, the remaining events which center on the predicate *is* and the predicate *has* in the sentence pair are semantically different as one structure includes the arguments *ARG1* and *ARG2* while the other contains only *ARG0* and *ARG1*.

In Sentence Similarity tasks, two sentences in a sentence pair are likely to have one or several events in common, such as the event centered *donate* in this case. However, a subtle difference in a key element in the semantic structure of the sentence may also lead to a very different semantics of the whole sentence, such as the events centered on *is* and *has*. Our proposed model EventBERT precisely appreciates the

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

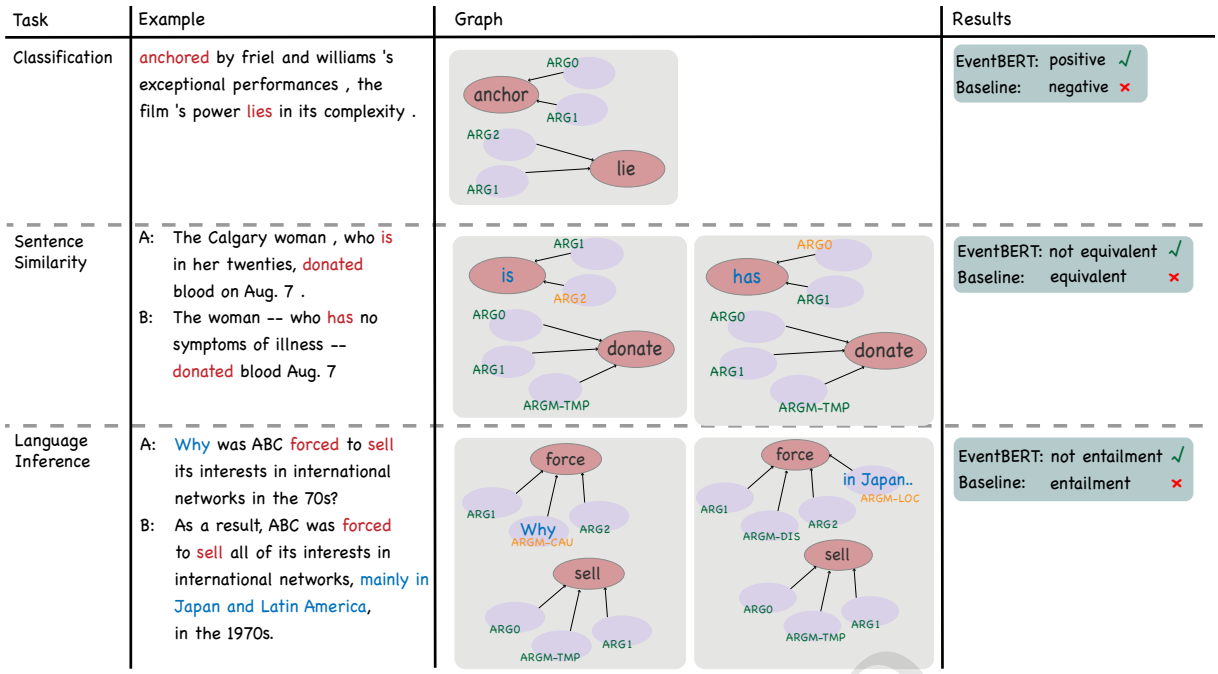


Figure 5: Examples selected from the dev set of SST-2, MRPC and QNLI where baseline fails but our model succeeds.

value of abstracting structural semantics, benefiting from capturing event-based semantic knowledge to perceive the differences between sentences and thus make more accurate judgments.

5.4.3 Language Inference

Referring to the case from QNLI, as can be seen from Figure 5, the question and paragraph texts are broadly similar in terms of *sell*-centered structure, both containing the arguments labeled *ARG0*, *ARG1*, and *ARGM-TMP*. However, by means of graph modeling, it can be clearly and explicitly observed that the structures centered on *force* are distinct, with the structure in the interrogative sentence containing the argument *ARGM-CAU* and the corresponding structure in the paragraph texts containing the argument *ARGM-LOC* instead. It is worth noting that one of the most crucial steps in determining whether a paragraph entails the correct answer to a question is whether the corresponding semantic structure in paragraph texts has the span labeled with the semantic role referring to the interrogative in the question. For example, in this case, the interrogative *Why* is exactly the *ARGM-CAU* of the predicate *force*; whereas the structure centered on *force* in the paragraph lacks the corresponding argument content and is replaced by *ARGM-LOC* instead. Therefore, it can be easily inferred that the paragraph focuses on the location (i.e., *in Japan and Latin America*) while the question concentrates on the cause (i.e., *Why*), which exactly reflects that there is no answer span for the interrogative of the question.

It is known that interrogative in the question and corresponding answer span should belong to the same semantic role. EventBERT takes full advantage of extracting abstracted semantics based on predicates, thus conducting language inference tasks more efficiently.

5.5 Error Analysis

We select bad cases of the baseline model and further investigate the ones of which our EventBERT also fails to predict the correct answers. We study two cases respectively from MRPC and QNLI as is shown in Table 4. The first error is caused by EventBERT's identification of the argument *in a written statement* of the predicate *said* in the first sentence, which is not entailed in the second sentence. However, the lack of this argument does not affect the main semantic information. The second error is due to argument reference confusion for the special predicate *is*. For instance, the interrogative *What* is labeled as *ARG2* whereas the correct answer *Hypersensitivity* is labeled as *ARG1*. From the above error cases, it may

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

suggest that our model needs to have a more accurate perception of semantic relationships, which is left for future studies.

Example	EventBERT	Golden Answer
This decision is clearly incorrect , " FTC Chairman Timothy Muris *said in a written statement*. The decision is " clearly incorrect , " FTC Chairman Tim Muris *said*.	*Not equivalent*	Equivalent
What is the name for a response of the immune system that damages the body's native tissues? *Hypersensitivity is* an immune response that damages the body's own tissues.	*Not entailment*	Entailment

Table 4: Errors in predictions for cases in MRPC and QNLI dev set. The words in magenta indicate the key predicate. The words in blue indicate the key arguments referred to the predicate.

6 Conclusion

In this work, we propose EventBERT, an event-based semantic representation model that builds on BERT architecture and incorporates event-based structural semantics in terms of graph network modeling for fine-grained language representation. Experiments on a wide range of NLU tasks show the effectiveness of our model by consistently surpassing the baseline. While most existing works focus on fusing accurate semantic signals to enhance semantic information, we open up a novel perspective to model intrinsic structural semantics for deeper comprehension and inference in an intuitive and explicit way.

References

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-bert: Improving pre-trained transformers with syntax trees. *arXiv preprint arXiv:2103.04350*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *ACL-PASCAL*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. 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 2306–2317, Minneapolis, Minnesota, 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.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *IWP2005*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2019. Hierarchical graph network for multi-hop question answering. *arXiv preprint arXiv:1911.03631*.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418.

Dirk Geeraerts and Hubert Cuyckens. 2007. Introducing cognitive linguistics. In *The Oxford handbook of cognitive linguistics*.

Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hongye Tan. 2021. Frame semantic-enhanced sentence modeling for sentence-level extractive text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4045–4052, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2020a. Incorporating syntax and frame semantics in neural network for machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2635–2641, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020b. A frame-based sentence representation for machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 891–896, Online, July. Association for Computational Linguistics.

Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2021. Frame-based multi-level semantics representation for text matching. *Knowledge-Based Systems*, 232:107454.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Friedrich Wilhelm Levi. 1942. *Finite geometrical systems: six public lectues delivered in February, 1940, at the University of Calcutta*. University of Calcutta.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks forpolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2022. Structural characterization for dialogue disentanglement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–297, Dublin, Ireland, May. Association for Computational Linguistics.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *RepEval*.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Dialogue graph modeling for conversational machine reading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3158–3169, Online, August. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2017. Column networks for collective classification. In *Thirty-first AAAI conference on artificial intelligence*.

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, July. 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*.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. Numnet: Machine reading comprehension with numerical reasoning. *arXiv preprint arXiv:1910.06701*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

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.

Peng Shi and Jimmy J. Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

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 *EMNLP*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *stat*, 1050:20.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Zhaofeng Wu, Hao Peng, and Noah A. Smith. 2021. Infusing finetuning with semantic dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. Syntax-enhanced pre-trained model. 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 5412–5422, Online, August. Association for Computational Linguistics.

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.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020a. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020b. Sg-net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9636–9643.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 774-785, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

An Exploration of Prompt-Based Zero-Shot Relation Extraction Method

Jun Zhao[1*], Yuan Hu[1*], Nuo Xu[1], Tao Gui[1†], Qi Zhang[1†], Yunwen Chen[2], Xiang Gao[2]
[1] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[2] DataGrand Information Technology (Shanghai) Co., Ltd., Shanghai, China
{zhaoj19,yuanhu20,tgui,qz}@fudan.edu.cn
xun22@m.fudan.edu.cn
{chenyunwen,gaoxiang}@datagrand.com

Abstract

Zero-shot relation extraction is an important method for dealing with the newly emerging relations in the real world which lacks labeled data. However, the mainstream two-tower zero-shot methods usually rely on large-scale and in-domain labeled data of predefined relations. In this work, we view zero-shot relation extraction as a semantic matching task optimized by prompt-tuning, which still maintains superior generalization performance when the labeled data of predefined relations are extremely scarce. To maximize the efficiency of data exploitation, instead of directly fine-tuning, we introduce a prompt-tuning technique to elicit the existing relational knowledge in pre-trained language model (PLMs). In addition, very few relation descriptions are exposed to the model during training, which we argue is the performance bottleneck of two-tower methods. To break through the bottleneck, we model the semantic interaction between relational instances and their descriptions directly during encoding. Experiment results on two academic datasets show that (1) our method outperforms the previous state-of-the-art method by a large margin with different samples of predefined relations; (2) this advantage will be further amplified in the low-resource scenario.

1 Introduction

Relation extraction (RE) aims to extract the relation between entity pairs from unstructured text. The extracted relation facts can benefit various downstream applications such as knowledge graph completion (Wang et al., 2014), web search (Xiong et al., 2017) and dialog systems (Madotto et al., 2018). However, many effective RE methods (Wu and He, 2019; Du et al., 2018) work within predefined relation sets. They failed to deal with a real-world environment where new relations will emerge after the training phase. These fast-growing new relations make it impossible for us to gather labeled training data for all of them. To recognize the newly emerging relations lacking labeled data, zero-shot RE is of the utmost practical interest.

Despite the great potential of zero-shot RE in real-world applications, there have been relatively few studies focusing on this challenging task. To enable models to predict unseen relations, previous works usually model zero-shot relation extraction as a well-designed task form. Levy et al. (2017) consider relation extraction as a machine reading comprehension. They first associate a few question templates for each relation and then determine which relation satisfies the given sentence and question by model prediction. However, a reasonable and effective question template usually needs careful design, which cannot meet the extraction needs of rapidly growing new relations (Chen and Li, 2021). Therefore, instead of manually constructing question templates, subsequent works (Obamuyide and Vlachos, 2018; Chen and Li, 2021) take advantage of the readily available textual description to represent the new relations, and formulate zero-shot RE as a semantic matching task achieving superior results.

However, current methods usually require a large number of in-domain labeled data of predefined relations to train the model parameters. The learned relational knowledge is mainly from labeled data

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
* Equal contribution.
† Corresponding authors.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

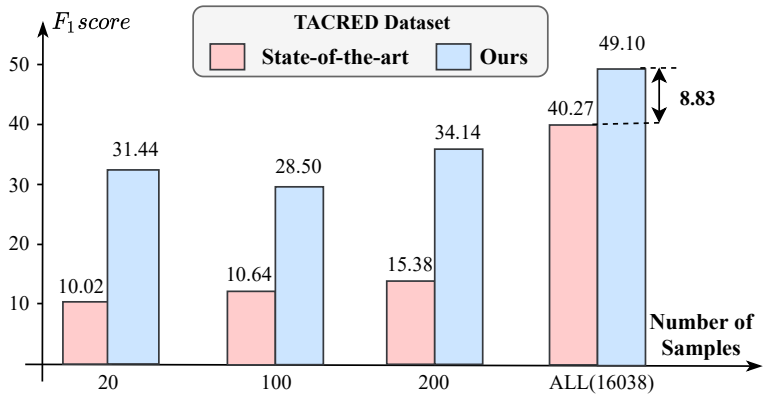


Figure 1: When shifting to some special domains (e.g. medicine, finance) where large-scale labeled data are not available, the performance of these methods on new relations decreases significantly. By inducing the knowledge in the pre-trained language model, our method can approach the results of previous state-of-the-art method ZS-BERT (Chen and Li, 2021) using only 200 labeled data. When using all data, our method improves the F1 score by 8.83%.

itself. As a result, when shifting to some special domains where large-scale labeled data are not available, the performance of these methods on new relations decreases significantly. An experimental illustration is shown in Figure 1. Fortunately, pre-trained language models (PLMs) such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), can learn a wealth of linguistic (Peters et al., 2018), local syntactic (Hewitt and Manning, 2019) and long-range semantic (Jawahar et al., 2019) from large-scale corpora by self-supervised learning. An interesting question is whether we can reduce the dependence on labeled data of predefined relations with the help of knowledge in PLMs?

To answer this question, in this work, we propose a prompt-based zero-shot RE method. Different from previous methods, in which the learned relational knowledge mainly comes from the labeled data of predefined relations, we leverage prompt to stimulate the rich knowledge distributed in PLMs to reduce dependence on these labeled data. Specifically, we model zero-shot RE as a semantic matching task between relational instance and description. In order to induce the knowledge in PLMs, we fuse the original input with the prompt template to formulate a cloze-style task. Then, we count the probability distribution of the model output and take the words with significant differences between classes as label words. In addition, each predefined relation corresponds to **many** instances and **one** description. The significant quantity gap makes the two-tower methods unable to effectively model the semantics of relation description. Therefore, we directly model the semantic interaction between instances and descriptions during training. Based on the reformulated input and these selected label words, we optimize a semantic matching model, which predicts whether the relation and the textual description match. Experimental results show that our method has very significant advantages when the large-scale labeled data of predefined relations are not available.

To summarize, the main contributions of our work are as follows: (1) We propose a prompt-based zero-shot relation extraction method, which maintains high generalization ability when using even one labeled data per predefined relations. (2) We design comprehensive experiments to analyze the impact of predefined relations and prompt composition on the generalization performance of the model in the low-resource scenario, which may enlighten the following work. (3) Experiment results on two academic datasets show that our method outperforms the previous state-of-the-art method by a large margin and this advantage will be further amplified in low resource scenarios.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2 Related Work

2.1 Knowledge in Pretrained Language Model

Contextual word representations derived from pre-trained language models have recently been shown to provide significant improvements to the state of the art for a wide range of NLP tasks, motivating a growing body of research investigating what aspects of linguistic knowledge they are able to learn from unlabeled data. Peters et al. (2018) showed that different neural architectures (e.g., LSTM, CNN, and Transformers) can hierarchically structure linguistic information that varies with network depth. (Jawahar et al., 2019; Clark et al., 2019; Goldberg, 2019) show that such hierarchy exists as well for BERT models that are not trained using the standard language modeling objective. More recently, many studies (Tenney et al., 2019; Hewitt and Manning, 2019) probe the knowledge within PLMs from various perspectives and find that the existing models trained on language modeling and translation produce strong representations for syntactic phenomena. Together, these results suggest that pre-trained language models entail comprehensive linguistic knowledge, which accounts for its great performance on downstream tasks and proves its potential to represent the samples of zero-shot relation extraction tasks, which has limited training data.

2.2 Prompt-Based Optimization

Since the advent of prompt tuning, it has soon become the prevailing paradigm of natural language processing. Prompt tuning is based on language models that estimate the probability of text. It modifies the original input of downstream tasks to a prompt with unfilled positions, and predicts the output based on the slot-filling result by language models (Liu et al., 2021). This method has been proven to be helpful on various NLP tasks, including text classification (Han et al., 2021), entity typing (Ding et al., 2021), text generation (Li and Liang, 2021), and also multi-modal tasks (Tsimpoukelli et al., 2021). Current studies have made some attempts to derive knowledge from PLMs with prompts. Jiang et al. (2020) proposed mining-based and paraphrasing-based methods to automatically generate high-quality prompts, which boosted the performance of knowledge-driven tasks. Zhong et al. (2021) conducted a set of control experiments to disentangle the efforts of training data and pre-trained knowledge. Inspired by these works, compared with direct fine-tuning, using the limited labeled data to derive the existing relational knowledge in the pretrained model is a better choice.

3 Method

We reformulate the task of zero-shot relation extraction as a semantic matching task optimized by prompt-tuning. In this section, we will introduce our proposed method in detail. We start by defining the problem we will tackle. Then we introduce how we reformulate zero-shot relation extraction, our prompt design and the selection of label tokens. Finally, we introduce the strategy of making predictions with our model.

3.1 Problem Definition

For the zero-shot relation extraction task, we expect the model $\mathcal{M}$ to predict the right relation of two annotated entities within the text, where the candidate relations are unseen during training.

Formally, let $R_s = \{r_s^1, \ldots, r_s^n\}$ denotes the set of predefined relations. Each relation in R_s has a corresponding textual description, composing the set of relation descriptions $D_s = \{d_s^1, \ldots, d_s^n\}$. In the train set $S_s = \{S_s^1, \ldots, S_s^N\}$, each sample $S_s^i = (x^i, r_s^i)$ consists of a relational instance x^i and its relation label $r_s^i \in R_s$, in which the relational instance x^i is a piece of text s^i with annotated entities e_1^i and e_2^i, namely $x^i = \langle s^i, e_1^i, e_2^i \rangle$. Similarly, the set of unseen relations for testing is denoted as $R_u = \{r_u^1, \ldots, r_u^m\}$, together with the corresponding description set $D_u = \{d_u^1, \ldots, d_u^m\}$. Note that all relations in R_u are unseen during training, i.e. $R_s \cap R_u = \varnothing$. The test set is denoted as $S_u = \{S_s^1, \ldots, S_u^M\}$, in which each test sample $S_u^j = (x^j, r_u^j)$.

3.2 Task Reformulation

In our work, we model zero-shot relation extraction as a semantic matching task where we need to recognize the semantic equivalence relations between relational instances and the description of their

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Input		Label
Original Sample	**Prompt**	
Cloud Nothings was formed in Cleveland .	[CLS] [CT] Premise : *input text* [SEP] [CT] Hypothesis : *relation description* . Answer : [MASK] [SEP]	Place of Foundation
Reformulated Samples		
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : location where a group or organization was formed . Answer : [MASK] [SEP]		match
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : musical instrument that a person plays . Answer : [MASK][SEP]		not_match
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : league in which team or player plays or has played in . Answer : [MASK] [SEP]		not_match
[CLS] [CT] Premise : Cloud Nothings was formed in Cleveland [SEP] [CT] Hypothesis : heritage designation of a historical site . Answer : [MASK] [SEP]		not_match

Table 1: An example of the reformulation of zero-shot relation extraction task. Each original sample is paired with various descriptions to form new samples.

corresponding relation labels. Specifically, we pair each test sample with the description of every candidate relation, and label them with `match`/`not_match` to form semantic matching samples. And we set it to have half the probability of pairing the training sample with the non-corresponding relation description and half the probability of pairing it with the corresponding relation description. Therefore, the number of positive and negative semantic examples in the training set is roughly equal. As shown in Table 1, the pair is labeled as `match` only when the description matches the corresponding relation label of the relational instance.

Formally, taking the training sample $S_s^i = (x^i, r_s^i)$ for example, we can derive a semantic matching sample $\{(x^i, d_s^k, y^k)\}$ from it, where

$$y^k = \begin{cases} \texttt{match} & r_s^i = r_s^k \\ \texttt{not\_match} & otherwise, \end{cases} \tag{1}$$

We denote the newly derived train set for semantic matching as $S_s' = \{(x^i, d_s^k, y^{ik})\}_{i=1...N}$. Note that from each test sample we will derive m semantic matching samples. The test set is denoted as $S_u' = \{(x^j, d_u^l, y^{jl})\}_{j=1...M, l=1...m}$. In summary, the above efforts convert the original problem to a semantic matching task, which is basically a 2-classification task that we could handle.

Is the two-tower architecture suitable for this task? The state-of-the-art zero-shot methods (Obamuyide and Vlachos, 2018; Chen and Li, 2021) adopt a two-tower architecture to implement the above semantic matching model. However, encoding instances and descriptions in isolation is not a good choice. Assuming that we use 10 relations and 100 instances of each relation to train a two-tower model, there are 1000 different inputs for instance encoder and only 10 inputs for the description encoder. This significant gap makes it difficult for description encoder to learn semantics effectively. Different from the two-tower architecture, the proposed method directly models the semantic interaction between instances and description during encoding. We will show the significant improvement brought by this change in the experiments.

3.3 Model with Prompt Tuning

To model the semantic matching between relational instances and descriptions, we take advantage of pre-trained language models together with prompt tuning. Noticeably, for zero-shot relation extraction, the most critical issue during training is that very few relation descriptions are exposed to the model. Furthermore, all of the descriptions in the test set are unseen in training. Thus, the rich linguistic knowledge of PLM is necessary to ensure that the model understands the descriptions with limited training. Additionally, to tackle the discrepancy of PLM between the pre-training and fine-tuning stage, prompt tuning is necessary to reformulate downstream tasks as cloze-style tasks that BERT is good at. We believe that prompt tuning provides an effective way to fully export knowledge from pre-trained language models and also enables few-shot learning of the task. Due to the discussions, we build our model based on BERT, which learns the objectives by prompt tuning.

Prompt Design For each reformulated sample (x, d, y), we fill the original text of relational instance and the description into a prompt. We define the prompt x' for relational instance x and relation description d

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

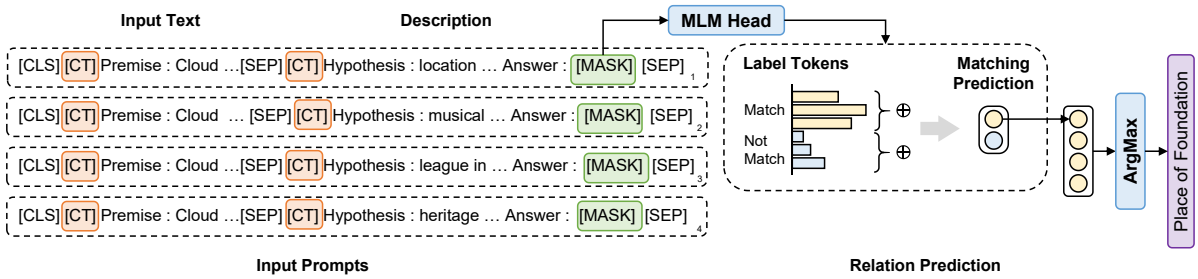


Figure 2: An overview of the process of relation prediction. The [MASK]ed positions within input prompts are firstly filled by language model, then the logits of label tokens are collected to predict the matching probability of input text and description. Lastly, we collect the matching probabilities for each pair and estimate the distribution of relational labels based on them.

as

$$\begin{aligned} x' = & \texttt{[CLS][CT]}\ s'\ \texttt{[SEP][CT]}\ d' \\ & \texttt{[MASK][SEP]} \\ \text{where}\ \ s' = & \text{Premise: } s \\ d' = & \text{Hypothesis: } d \text{ Answer:}, \end{aligned} \tag{2}$$

where s is the input text, which is the original text of x; s' and d' denotes the prompt-formulated input text and description respectively; [CT] denotes T different continuous tokens that make up the template. Examples of input prompts could be seen in Table 1. The design of prompt aims to fully utilize the ability of BERT as a rich knowledge base, and the introduction of continuous tokens in template aims to enhance the representation ability of the prompt, since these tokens could be optimized in the whole embedding space.

Label Token Selection Following the common settings of prompt tuning on classification task, we also determine label tokens for each category (namely match or not_match) for consequential prompt tuning. Basically, for the two categories, the probability distributions of masked language modeling should be different and distinguishable. Thus, retrieving label tokens is the process of capturing features that indicate the distribution associated with a certain category. We solve the problem by estimating the distributions and retrieving tokens that have the most significant difference of probability among distributions.

Formally, we partition the reformulated train set S'_s by category of label y. The matched and unmatched samples are denoted as $S_{sm} = \{(x, d, y) \in S'_s | y = \texttt{match}\}$ and $S_{sn} = \{(x, d, y) \in S'_s | y = \texttt{not\_match}\}$, respectively. The prompts of samples are then fed to BERT. For sample (x, d, y), the estimated distribution of the [MASK] token is calculated as

$$P(w|x, d) = \text{softmax}(W(\text{MLM}(x')) + b), \tag{3}$$

where w denotes every token in vocabulary, $P(w|x, d)$ indicates the estimated MLM distribution of the sample, MLM denotes the output embedding of [MASK] token, W and b denote trainable weights of linear projection.

The MLM distribution of categories is estimated by averaging the predicted distributions among samples in the category:

$$P_m(w) = \frac{1}{|S_{sm}|} \sum_{(x,d,y)}^{S_{sm}} P(w|x, d), \tag{4}$$

$$P_n(w) = \frac{1}{|S_{sn}|} \sum_{(x,d,y)}^{S_{sn}} P(w|x, d), \tag{5}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

where $P_m(w)$ and $P_n(w)$ indicate the estimated MLM distribution of the category `match` and `not_match`, $|S_{sm}|$ and $|S_{sn}|$ denote the number of matched and unmatched samples, respectively.

Finally, for each category, the tokens with top-K possibility difference between the MLM distribution within and without the category are selected as the label tokens. The possibility difference of each word is divided by their estimated occurrence possibility to ensure fair comparison.

$$\{w_m^1, \ldots, w_m^K\} = \underset{w}{\text{topK}} \frac{P_m(w) - P_n(w)}{P_m(w) + P_n(w)}, \tag{6}$$

$$\{w_n^1, \ldots, w_n^K\} = \underset{w}{\text{topK}} \frac{P_n(w) - P_m(w)}{P_m(w) + P_n(w)}. \tag{7}$$

In Eq.6 and 7, K is the number of tokens selected for each category, $\{w_m^1, \ldots, w_m^K\}$ and $\{w_n^1, \ldots, w_n^K\}$ denote the selected label tokens of the category of {`match` and `not_match`} respectively.

3.4 Training and Inference

In this part, we introduce our strategy to derive relation predictions from the semantic matching model, along with the training objectives.

Similar to other prompt-based methods, the output possibilities of label tokens are collected to perform a 2-classification on label y. The possibility of categories is proportional to the production possibility of label words. As shown in Eq.8, in implementation, we achieve this by adding the output logits of label tokens and applying softmax on them:

$$P(y = c|x, d) = \text{softmax}\left(\sum_{k=1}^{K} \log P(w_c^k|x, d)\right), \tag{8}$$

where $c \in \{\texttt{match}, \texttt{not\_match}\}$.

The prediction of relation label for relational instance x_i is done by collecting the possibilities of match between x_i and the descriptions of every candidate relation $R^k \in R$. As in Eq. 9, the matching possibilities of x_i and all candidate relations are collected as logits and are put to a softmax function to predict the distribution of the relation label.

$$p^{ik} = P(y^{ik} = \texttt{match}|x^i, D^k), \tag{9}$$

$$P(r^k|x^i) = \frac{\exp(p^{ik})}{\sum\limits_{r^k \in R} \exp(p_{ik})}. \tag{10}$$

Lastly, the model is trained on cross-entropy loss L_{CE} to maximize the log-likelihood of all training samples.

$$L_{CE} = \sum_{i=1}^{N} \text{CrossEntropy}(r_s^i, \{p_s^{ik}\}_{k=1}^n). \tag{11}$$

As for making prediction on unseen samples, i. e. evaluating model on test sets, for each test sample S_u^j, the predicted relation distribution of relational instance x_j is illustrated in Eq. 12 and 13. We pick the relation with the highest possibility as the predicted result.

$$p^{jl} = P(y^{jl} = \texttt{match}|x^j, d_u^l), \tag{12}$$

$$P(R_u^l|x^j) = \frac{\exp(p^{jl})}{\sum\limits_{r_u^j \in R_u} \exp(p_{ik})}, \tag{13}$$

$$\hat{r}_u^j = \underset{l}{\text{argmax}} P(R_u^l|x^j). \tag{14}$$

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Dataset	# Inst.	# relations	% N/A
FewRel	56000	80	-
TACRED	106264	42	79.5%

Table 2: Original statistics of datasets FewRel and TACRED. %N/A is the proption of label "no_relation" and "-" represents there is no N/A instances.

4 Experimental Setup

In this section, we describe the datasets for training and evaluating the proposed method. We also detail the baseline models for comparison. Finally, we clarify the implementation details and hyperparameter configuration of our method.

4.1 Datasets

Our main experiments are conducted on two relation extraction datasets: FewRel and TACRED. The original statistics of the two datasets are listed in Table 2.

FewRel (Han et al., 2018a). There are 80 relations included in FewRel, a high-quality RE dataset with 56,000 instances from Wikipedia. To be consistent with the previous state-of-the-art method, we rearrange the dataset. To be specific, we choose 65 relations as labeled set with predefined relation and select 15 relations as the unlabeled set with unseen relations.

TACRED (Zhang et al., 2017). TACRED is a human-annotated relation extraction dataset that contains 106,264 examples with 42 kinds of relations(including “no_relation”). The instances of special class "no_relation" is removed, and we use the remaining 21,773 instances for training and evaluation.

We also add a low-resource setting, which means the size of training data is small. Under the setting, the development set is provided, with about 5 examples per relation. As shown in Table 3 and Table 4, the three different values of n represents the number of data used for training are only 20, 100 and 200 respectively. For the setting, We randomly sample training data from each relation category roughly evenly. Note that when sampling 20 training data, the number of relation categories in the training set of both datasets is also reduced to 20. For both of the two datasets, we use the Macro-F1 score as the main metric to evaluate the model’s performance.

4.2 Compared Methods

To verify the effectiveness of our proposed method, we select the following models for comparison. The state-of-the-art method ZS-BERT (Chen and Li, 2021) adopted the two-tower architecture, this method encodes sentences and relation descriptions separately and uses nearest neighbor search as the matching function to obtain the prediction of unseen relations. When comparing with R-BERT (Wu and He, 2019) and Attentional Bi-LSTM (Zhou et al., 2016), two supervised relation extraction (SRE) models, we take the same way as ZS-BERT (Chen and Li, 2021) so that SRE models can carry out zero-shot prediction. Specifically, we change the last layer to a fully-connected layer with tanh activation function. Based on the input instance embedding and relation description’s embedding, the nearest neighbor search will be applied to generate the zero-shot prediction. We also compare our method with ESIM (Chen et al., 2017), a semantic matching model. To have a fair comparison, the strategy to generate relation predictions from the semantic matching model is the same as ours. Finally, we introduce BERT(CLS) (Devlin et al., 2019) to intuitively show the performance improvement brought by modeling the semantic interaction between instances and descriptions during encoding.

4.3 Implementation Details

We adopt BERT-base-cased as the encoder and all experiments are conducted using a NVIDIA GeForce RTX 3090 with 24GB memory. The number of continuous tokens is $t = 4$. We use AdamW for optimization, in which the initial learning rate is 3e-5. Taking into account the randomness of network initialization and random selection of n training instances, we run our experiment 5 times and the results

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	FewRel(m=15)											
	n=20			n=100			n=200			n=all		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Att Bi-LSTM (Zhou et al., 2016)	14.19	13.88	14.03	15.75	19.8	17.55	20.83	26.00	23.13	38.13	32.05	34.82
ESIM (Chen et al., 2017)	0.60	5.45	1.08	0.90	6.56	1.58	7.66	7.38	7.52	36.97	32.51	34.60
R-BERT (Wu and He, 2019)	8.40	8.38	8.39	13.61	15.90	14.67	16.05	18.58	17.22	32.25	25.58	28.53
ZS-BERT (Chen and Li, 2021)	6.04	6.36	6.20	6.34	7.93	7.05	8.35	9.59	8.93	35.54	38.19	36.82
BERT(CLS) (Devlin et al., 2019)	**44.95**	33.65	38.49	49.99	47.20	48.55	**53.14**	52.13	52.62	**67.62**	59.12	63.09
Ours	44.94	**45.72**	**45.33**	**50.21**	**51.72**	**50.96**	52.49	**53.98**	**53.23**	64.48	**62.45**	**63.45**

Table 3: Main results on FewRel. The best results are bold. n is the number of provided training data and m represents unseen relations' number.

Method	TACRED(m=11)											
	n=20			n=100			n=200			n=all		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Att Bi-LSTM (Zhou et al., 2016)	14.33	11.38	12.68	13.73	10.64	11.99	15.68	21.70	18.20	25.20	20.17	22.41
ESIM (Chen et al., 2017)	9.09	0.15	0.29	8.54	9.41	8.96	1.52	9.15	2.61	26.99	18.38	21.87
R-BERT (Wu and He, 2019)	14.59	7.27	9.70	18.93	12.12	14.78	23.62	19.67	21.47	44.66	45.86	45.25
ZS-BERT (Chen and Li, 2021)	10.79	9.35	10.02	12.53	9.25	10.64	14.98	15.79	15.38	38.08	42.72	40.27
BERT(CLS) (Devlin et al., 2019)	25.53	19.78	22.29	9.34	10.55	9.91	**37.97**	**34.43**	**36.11**	**51.90**	44.71	48.03
Ours	**32.40**	**30.54**	**31.44**	**38.12**	**22.75**	**28.50**	34.56	33.73	34.14	51.85	**46.63**	**49.10**

Table 4: Main results on TACRED. The best results are bold. n is the number of provided training data and m represents unseen relations' number.

we report are the average results. Other results of compared methods are gotten when the parameters remain the same as its own published source code. We follow Soares et al. (2019) to augment each instance with four reserved word pieces to mark the begin and end of each entity. The relation descriptions of FewRel are obtained from (Han et al., 2018b) and TACRED's are obtained from the TAC-KBP relation ontology guidelines[2].

5 Results and Discussion

5.1 Main Results

The main results of our experiments on FewRel and TACRED are listed in Table 3 and Table 4. **First**, as can be seen, the method we propose steadily outperforms compared methods, and even the previous state-of-the-art method (Chen and Li, 2021) performs much worse than our method when targeting at different number of training instances. The reason is that the two-tower model which the previous state-of-the-art method (Chen and Li, 2021) encodes the input instances and candidate relations with large quantitative differences separately, and we argue that this modeling choice is insufficiently expressive for modeling the semantic matching between instances and relation descriptions. What's more, the simple matching function (ZS-BERT uses nearest neighbor search) is incapable of capturing the complicated interactions between input sentences and relation descriptions. Our proposed method yields rich interactions between the input instance and candidate relation description, as they are jointly encoded to obtain a final representation. At the layers of transformer, every word in the candidate relation description can attend to every word in the input instance, and vice-versa, so our proposed method can produce a candidate-sensitive input representation, which the ZS-BERT cannot. **Second**, it can be apparently found that the baseline's performance decreases significantly when the number of labeled data decreases, which indicates that large number of in-domain labeled data of predefined relations is a prerequisite for their good performance. While our method manage to derive the original knowledge in PLMs with prompt so that our method still performs well when the labeled data is scarce. For FewRel, our MACRO-F1 score reaches 45.33% training with 20 instances, which is better than the result of previous state-of-the-art using the complete

[2]https://tac.nist.gov/2015/KBP/ColdStart/guidelines/ TAC_KBP_2015_Slot_Descriptions_V1.0.pdf

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	FewRel_TACRED n=all			TACRED_FewRel n=all		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Att Bi-LSTM (Zhou et al., 2016)	21.86	27.72	24.44	31.27	39.26	34.82
ESIM (Chen et al., 2017)	22.67	18.91	20.62	19.38	11.93	14.77
R-BERT (Wu and He, 2019)	23.10	28.49	23.98	15.31	14.70	15.00
ZS-BERT (Chen and Li, 2021)	35.90	29.78	32.55	17.69	11.81	14.16
Ours	**41.26**	**37.62**	**39.36**	**60.01**	**50.74**	**54.99**

Table 5: Results on two constructed cross-domain tasks.

Prompt	FewRel	TACRED
`[PRE]` Question : `[HYP]` . true or false ? Answer : `[MASK]`	63.11	47.19
`[PRE]` Question : `[HYP]` ? `[MASK]`	61.09	48.79
`[PRE]` Is `[HYP]` true ? Answer : `[MASK]`	**63.58**	47.72
Does `[HYP]` agree with `[PRE]` ? `[MASK]`	62.44	45.73
Ours	63.45	**49.10**

Table 6: Results on different prompts.

train dataset. Such results verify the strong ability of low-resource learning for our proposed method.

5.2 Cross Domain Analysis

Through the analysis of main results, we have concluded that large-scale labeled data of predefined relations is a prerequisite for the existing model to achieve good generalization performance on unseen relations. An ensuing question is: when we deal with the problem of a field that lacks labeled data, can we solve this problem by using labeled data with existing relations in common fields? To answer this question, we conducted experiments on two constructed cross-domain zero-shot relation extraction tasks.i.e.,: FewRel to TACRED and TACRED to FewRel. Specifically, pre-defined relations and their labeled instances come from the source domain training dataset, and we evaluate performance on the target domain testing dataset.

Table 5 shows the results. By comparing with the in-domain experimental results in the main experiment, we can find: the change of domain does increase the semantic gap between the pre-defined and unseen relations. As a result of that: For FewRel to TACRED, the experimental result of our method is reduced from 49.10% to 39.36%, and for TACRED to FewRel, the result is reduced from 63.45% to 54.99%. But our performance still outperforms compared methods, which shows the proposed method's generalization on unseen relations.

5.3 Influence of Pre-defined Relation Number

In this subsection, we study the effect of the number of seen predefined relations in the train dataset. And we conduct the experiment on FewRel. For FewRel, the original number of predefined relations is 65, we sample 33,17,9,5 classes from the original train dataset in turn, which correspond to 50%, 25%, 12.5%, 6.25% of the original classes represented by the scale on the horizontal axis in the figure. The results of Figure 3 prove that the number of pre-defined relations does matter. As the number decreases, the knowledge learned from the training set also decreases, which can weaken the model's generalization of unseen relations. So the performance of our proposed method also gets worse. Nevertheless, our method can still be said to perform well. For FewRel, When we reduce the number of predefined relation types to 5, our performance still outperforms the previous state-of-the-art, which can validate the effectiveness of our proposed method.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

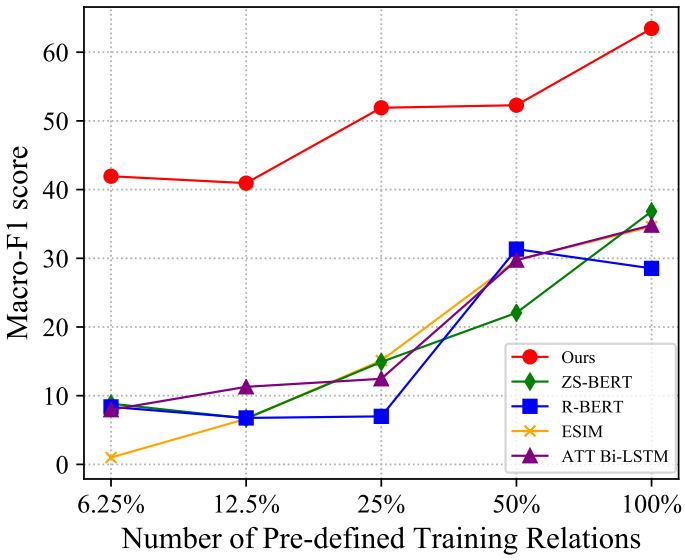


Figure 3: Model results with different number of predifined training relations on FewRel.

5.4 Analysis on Different Prompt Forms

To explore the impact of different forms of prompt on the performance of the proposed method, we conducted experiments on two datasets based on different prompts. Because the continuous tokens' position relative to `[HYP]` and `[PRE]` doesn't change, it is omitted from the table. As is shown in Table 6, inappropriate forms may lead to worse results, but on the other hand, a suitable prompt form can also improve model performance since it can help elicit the existing knowledge in PLMs. Among all the prompt forms, the form we have chosen is relatively well-behaved. Moreover, the prompt's performance is not necessarily the same as our intuition, in other words, the prompt we think good is not necessarily good for PLMs and we think the automatic generation of prompts is a promising research direction.

6 Conclusions

In this work, we introduce a prompt-based zero-shot relation extraction method, which still maintains superior generalization performance under low-resource settings. We clarify the limitations of the two-tower architecture in previous state-of-the-art methods, and directly model the interaction between instances and descriptions during encoding, which breaks the performance bottleneck of the previous model. The introduce of prompt-tuning effectively elicit the knowledge in PLMs and significantly reduces the dependence on predefined relations. We believe that these are the reasons why our method achieves excellent results. Experiment results on two academic datasets show that our method outperforms the previous state-of-the-art method by a large margin and this advantage will be further amplified in low resource scenarios.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

References

Chih-Yao Chen and Cheng-Te Li. 2021. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. *arXiv preprint arXiv:2104.04697*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341.

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*.

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.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2021. Prompt-learning for fine-grained entity typing. *arXiv preprint arXiv:2108.10604*.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. *arXiv preprint arXiv:1809.00699*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018a. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018b. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, October-November. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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. *arXiv preprint arXiv:2107.13586*.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia, July. Association for Computational Linguistics.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium, November. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*.

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.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1271–1279, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online, June. Association for Computational Linguistics.

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 21st China National Conference on Computational Linguistics, pages 786-797, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Abstains from Prediction: Towards Robust Relation Extraction in Real World

Jun Zhao[1*], **Yongxin Zhang**[1*], **Nuo Xu**[1], **Tao Gui**[1†], **Qi Zhang**[1†], **Yunwen Chen**[2], **Xiang Gao**[2]
[1] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[2] DataGrand Information Technology (Shanghai) Co., Ltd., Shanghai, China
{zhaoj19,yongxinzhang20,tgui,qz}@fudan.edu.cn
xun22@m.fudan.edu.cn
{chenyunwen,gaoxiang}@datagrand.com

Abstract

Supervised learning is a classic paradigm of relation extraction (RE). However, a well-performing model can still confidently make arbitrarily wrong predictions when exposed to samples of unseen relations. In this work, we propose a relation extraction method with rejection option to improve robustness to unseen relations. To enable the classifier to reject unseen relations, we introduce contrastive learning techniques and carefully design a set of class-preserving transformations to improve the discriminability between known and unseen relations. Based on the learned representation, inputs of unseen relations are assigned a low confidence score and rejected. Off-the-shelf open relation extraction (OpenRE) methods can be adopted to discover the potential relations in these rejected inputs. In addition, we find that the rejection can be further improved via readily available distantly supervised data. Experiments on two public datasets prove the effectiveness of our method capturing discriminative representations for unseen relation rejection.

1 Introduction

Relation extraction aims to predict the relation between entities based on their context. The extracted relational facts play a vital role in various natural language processing applications, such as knowledge base enrichment (Distiawan et al., 2019), web search (Xiong et al., 2017), and question answering (Honovich et al., 2021).

To improve the quality of extracted relational facts and benefit downstream tasks, many efforts have been devoted to this task. *Supervised relation extraction* is a representative paradigm built upon the closed world assumption (Gallaire and Minker, 1978). Benefiting from artfully designed network architectures (Miwa and Bansal, 2016; Huang and Wang, 2017; Zhang et al., 2018) and valuable knowledge in pretrained language model (Du et al., 2018; Verga et al., 2018; Wu and He, 2019; Baldini Soares et al., 2019), models effectively capture semantic-rich representations and achieves superior results. However, conventional supervised relation extraction suffer from the lack of large-scale labeled data. To tackle this issue, *distantly supervised relation extraction* has attracted much attention. The existing works mainly focus on how to alleviate the noise generated in the automatic annotation. Common approaches include selecting informative instances (Lin et al., 2016), incorporating extra information (Zhang et al., 2019), and designing sophisticated training (Ma et al., 2021).

Although a supervised relation classifier achieves excellent performance on known relations, real-world inputs are often mixed with samples of unseen relations. A well-performing model can still confidently make arbitrarily wrong predictions when dealing with these unseen relations (Nguyen et al., 2014; Recht et al., 2019). The unrobustness is rooted in the *Shortcut* feature (Geirhos et al., 2020) of neural networks. Models optimized by a supervised objective does not actively learn features beyond the bare minimum necessary to discriminate between known relations. As shown in Figure 1, if there is only president relation in the training data between Obama and the United States, the model tends

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
* Equal contribution.
† Corresponding authors.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

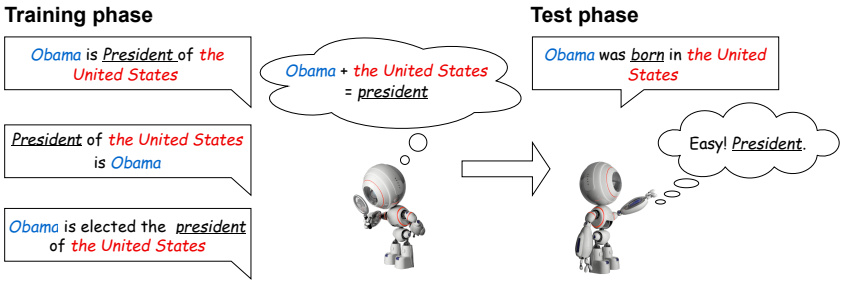


Figure 1: Neural models tend to use the simplest way to meet the supervised objective (*Shortcut* phenomenon (Geirhos et al., 2020)), which would lead to negative predictions on unseen relations. Hence, for the unseen relations, we hope neural models can reject prediction through embracing sufficient features.

Model/ Dataset	SpanBERT	Roberta	CP
Ori(F_1-score)	0.919	0.928	0.936
Mix(ΔF$_1$-score)	0.317↓	0.310↓	0.310↓

Table 1: Supervised RE models' performance when encountering new relations. These models are from previous papers(Joshi et al., 2019; Liu et al., 2019; Peng et al., 2020). Ori: all relations in the test set are present in the training set. Mix: 50% of the relations in the test set do not appear in the training set.

to predict the president relation when it encounters them again. However, entities are not equivalent to relation definitions. Models severely biased to the extraction of overly simplistic features can easily fail to generalize to discriminate between known and unseen relations. As shown in Table 1, when the unseen relations appears in the test set, the supervised RE models' F_1-score drops by at least 30 points.

In this work, we propose a robust relation extraction method in real world settings. By integrating rejection option, the classifier can effectively detect whether inputs express unseen relations instead of making arbitrary bad predictions. Specifically, we introduce contrastive training techniques to achieve this goal. A set of carefully designed class-preserving transformations are used to learn sufficient features, which can enhance the discriminability between known and unknown relation representations. The classifier built on the learned representation is confidence-calibrated. Thereby samples of unseen relations are assigned a low confidence score and rejected. Off-the-shelf OpenRE methods can be used to discover potential relations in these samples. In addition, we find the rejection can be further improved via the readily available distantly-supervised data. Experimental results show the effectiveness of our method capturing discriminative representations for unseen relation rejection.

To summarize, the main contributions of our work are as follows: (1) We propose a relation extraction method with rejection option, which is still robust when exposed to unseen relations. (2) We design a set of class-preserving transformations to learn sufficient features to discriminate known and novel relations. In addition, we propose to use readily available distantly-supervised data to enhance the discriminability. (3) Extensive experiments on two academic datasets prove the effectiveness of our method capturing discriminative representations for unseen relation rejection.

2 Related Work

2.1 Relation Extraction

Relation extraction has advanced for more than a couple of decades. Supervised/Distantly supervised relation extraction is oriented at predefined relational types. Researchers have explored different network architectures (Zhang et al., 2018), training strategies (Ma et al., 2021) and external information (Zhang et al., 2019). Superior results have been achieved. Open relation extraction is oriented at emerging

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

unknown relation. Well-designed extraction forms (e.g. sequence labelling (Fader et al., 2011), clustering (Zhao et al., 2021)) are used to deal with relations without pre-specified schemas. Different from them, we consider a more general scenario, in which known and unknown relations are mixed in the input. We effectively separate them by a rejection option, which enables us to use the optimal paradigm to deal with the corresponding relations.

2.2 Classification with Rejection Option

Most existing classification methods are based on the closed world assumption. However, inputs are often mixed with samples of unknown classes in real-world applications. The approaches used to handle it roughly fall into one of two groups. The first group calculates the confidence score based on the classifier output. The score can be used to measure whether an input belongs to unknown classes. Maximum softmax probability (MSP) (Hendrycks and Gimpel, 2017) is a represetative method and Liang et al. (2018) further improve MSP by introducing temperature scaling. Furthermore, Shu et al. (2017) build a multi-class classifier with a 1-vs-rest final layer of sigmoids to reduce the open space risk. The second group considers classification with rejection option as an outlier detection problem. Off-the-shelf outlier detection algorithms (Breunig et al., 2000; Schölkopf et al., 2001; Liu et al., 2008) are leveraged. Different optimization objectives such as large margin loss (Lin and Xu, 2019), gaussian mixture loss (Yan et al., 2020) are adopted to learn more discriminative representations to facilitate anomaly detection. Recently, Zhang et al. (2021) propose to learn the adaptive decision boundary (ADB) that serves as the basis for judging outliers.

3 Approach

In this paper, we propose a robust relation extraction method in real world settings. By integrating rejection option, the classifier can effectively detect whether inputs express unseen relations instead of making arbitrary bad predictions. Off-the-shell OpenRE methods can be used to discover potential relations in these rejected samples.

The problem setting in this work is formally stated as follows. Let $\mathcal{K} = \{\mathcal{R}_1, ..., \mathcal{R}_k\}$ be a set of known relations and $\mathcal{U} = \{\mathcal{R}_{k+1}, ..., \mathcal{R}_n\}$ be a set of unseen relations where $\mathcal{K} \cap \mathcal{U} = \emptyset$. Let $\mathcal{X}$ be an input space. Given the training data $\mathcal{D}^\ell = \{(x_i^\ell, y_i^\ell)\}_{i=1,...,N}$ where $x_i^\ell \in \mathcal{X}$, $y_i^\ell \in \mathcal{K}$, we target constructing a mapping rule $f : \mathcal{X} \rightarrow \{\mathcal{R}_1, ..., \mathcal{R}_k, \mathcal{R}^*\}$ where $\mathcal{R}^*$ denotes rejection option. Let $\mathcal{D}^u = \{(x_i^u, y_i^u)\}_{i=1,...,M}$ be the testing dataset where $y_i^u \in \mathcal{K} \cup \mathcal{U}$. An desirable mapping rule f should meet the following objective as much as possible:

$$f(x) = \begin{cases} y_i^u & y_i^u \in \mathcal{K} \\ \mathcal{R}^* & y_i^u \in \mathcal{U}. \end{cases}$$

3.1 Method Overview

We approach the problem by introducing contrastive learning techniques. As illustrated in Figure 2, the proposed method comprises four major components: relation representation encoder $\boldsymbol{g}(\cdot)$, confidence-calibrated classifier $\boldsymbol{\eta}(\cdot)$, class-preserving transformations $\mathcal{T}$, and the OpenRE module.

Our overview starts from the first two components. There is no doubt that an encoder and classifier are the basic components of a supervised relation extractor. However, the supervised training objective does not encourage the model to learn features beyond the bare minimum necessary to discriminate between known relations. Consequently, the classifier can misclassify unseen relations to known relations with high confidence.

In order to calibrate the confidence of the classifier, we introduce contrastive learning techniques. Given training batch $\mathcal{B}$, an augmented batch $\widetilde{\mathcal{B}}$ is obtained by applying random transformation $t \in \mathcal{T}$ to mask partial features. Then the supervised contrastive learning objective max/minimize the representation agreement according to whether their relations are the same. By doing this, the model is forced to find more features to discriminate between relations and the classifier can be calibrated. Based on the confidence-calibrated classifier, unknown relations are rejected if the maximum softmax probability of the classifier does not exceed a preset threshold θ.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

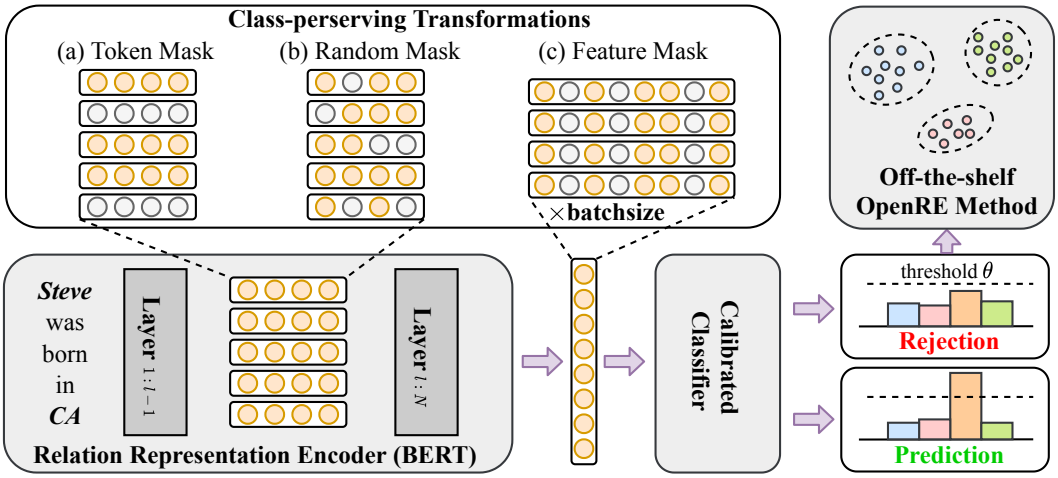


Figure 2: An overview of the proposed method. Three steps are included: (1) Contrastive training techniques and a set of class-preserving transformations are utilized to learn sufficient features. (2) The classifier extract known relations and rejects samples of unseen relations according to these features. (3) Off-the-shelf OpenRE method (SelfORE) is incorporated to discovery unseen relations in these rejected samples.

In order to discriminate unknown relations rather than just detect their existence, we further integrate the off-the-shelf OpenRE method into our framework. The samples rejected by the classifier are sent to the OpenRE module to detect potential unknown relations.

3.2 Relation Representation Encoder

Given a relation instance $x_i^\ell = (\boldsymbol{w}_i, h_i, t_i) \in \mathcal{D}^\ell$ where $\boldsymbol{w}_i = \{w_1, w_2, ..., w_n\}$ is the input sentence and $h_i = (s^h, e^h)$, $t_i = (s^t, e^t)$ mark the position of head and tail entities, relation representation encoder $\boldsymbol{g}(\cdot)$ aims to encode contextual relational information to a fixed-length representation $\boldsymbol{r}_i = \boldsymbol{g}(x_i) \in \mathbb{R}^d$. We opt for simplicity and adopt the commonly used BERT (Devlin et al., 2018) to obtain $\boldsymbol{r_i}$ while various other choices of the network architecture are also allowed without any constraints. Formally, the process of obtaining $\boldsymbol{r_i}$ is:

$$\boldsymbol{h}_1, ..., \boldsymbol{h}_n = \text{BERT}(w_1, ..., w_n) \tag{1}$$

$$\boldsymbol{h}_{ent} = \text{MAXPOOL}(\boldsymbol{h}_s, ..., \boldsymbol{h}_e) \tag{2}$$

$$\boldsymbol{r}_i = \langle \boldsymbol{h}_{head} | \boldsymbol{h}_{tail} \rangle , \tag{3}$$

where $\boldsymbol{h}_1, ..., \boldsymbol{h}_n$ is the result of the input sentence after BERT encoding, subscript s and e represent the start and end positions of the entity, $\boldsymbol{h}_{ent}$ represents the result of the maximum pooling of the entity, $\boldsymbol{h}_{ent}$ can be divided into head entity $\boldsymbol{h}_{head}$ and tail entity $\boldsymbol{h}_{tail}$, and $\langle \cdot | \cdot \rangle$ is the concatenation operator.

3.3 Confidence-calibrated Classifier

In order to alleviate overconfidence to unseen relations, we introduce contrastive learning techniques to calibrate classifier. A well-calibrated classifier should not only accurately classify known relations, but also give low confidence to unseen relations, that is, $\max_y p(y|x)$.

Given a training batch $\mathcal{B} = (x_i^\ell, y_i^\ell)_{i=1}^B$, we obtain a augmented batch $\widetilde{\mathcal{B}} = (\widetilde{x}_i^\ell, y_i^\ell)_{i=1}^B$ by applying random transformation $t \in \mathcal{T}$ on $\mathcal{B}$. For brevity, the superscript ℓ is omitted in the subsequent elaboration of this section. For each labeled sample $(\widetilde{x}_i, y_i)$, $\widetilde{\mathcal{B}}$ can be divided into two subsets $\widetilde{\mathcal{B}}_{y_i}$ and $\widetilde{\mathcal{B}}_{-y_i}$. $\widetilde{\mathcal{B}}_{y_i}$ denotes a set that contains samples of relation y_i and $\widetilde{\mathcal{B}}_{-y_i}$ contains the rest. The supervised contrastive

CCL 2022

learning objective is defined as follows:

$$\mathcal{L}_{cts}^{sup}(\mathcal{B},\mathcal{T}) = \frac{1}{2B}\sum_{j=1}^{2B}\mathcal{L}_{cts}(\widetilde{x}_i, \widetilde{\mathcal{B}}_{y_i}\backslash\{\widetilde{x}_i\}, \widetilde{\mathcal{B}}_{-y_i}) \tag{4}$$

$$\mathcal{L}_{cts}(x,\mathcal{D}^+,\mathcal{D}^-) = -\frac{1}{|\mathcal{D}^+|}log\frac{\sum_{x'\in\mathcal{D}^+}q(x,x')}{\sum_{x'\in\mathcal{D}^+\cup\mathcal{D}^+}q(x,x')} \tag{5}$$

$$q(x,x') = \exp(sim(\boldsymbol{z}(x),\boldsymbol{z}(x'))/\tau), \tag{6}$$

where $|\mathcal{D}|$ denotes the number of samples in $\mathcal{D}$, $sim(x,x')$ denotes the cosine similarity between x and x' and τ denotes a temperature coefficient. Following Chen et al. (2020), we use a additional projection layer $\boldsymbol{t}$ to obtain the contrastive feature $\boldsymbol{z}(x) = \boldsymbol{t}(\boldsymbol{g}(x))$.

Benifiting from contrastive training, the encoder $\boldsymbol{g}(\cdot)$ learns rich features to discriminate between known and novel relations. Accordingly, we train a confidence-calibrated classifier $\boldsymbol{\eta}(\cdot)$ upon $\boldsymbol{g}(\cdot)$ as follows:

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}^\ell}[\mathcal{L}_{ce}(\boldsymbol{\eta}(\boldsymbol{g}(x_i)),y)], \tag{7}$$

where $\mathcal{L}_{ce}$ is the cross entropy loss. In addition, we can easily obtain a large number of training data $\mathcal{D}^{dist}$ through distant supervision. None of the y_i^{dist} in $\mathcal{D}^{dist}$ are known relation, that is, $\{y_i^{dist}\}\cap\{y_j^\ell\} = \emptyset$. These data are only used as negative examples, so the noise in the data will not be a problem. We force the classifier output distribution of negative examples to approximate the uniform distribution by optimizing the cross-entropy between them. Using $\mathcal{D}^{dist}$, we optimize model by following objective instead of equation 7.

$$\mathcal{L}^{dist} = \mathcal{L} + \lambda\mathbb{E}_{x\sim\mathcal{D}^{dist}}[\mathcal{L}_{ce}(\boldsymbol{\eta}(\boldsymbol{g}(x)),y_{uni})], \tag{8}$$

where $\mathcal{L}$ refers to the optimization objective of equation 7. λ is the hyperparamters that balances the known relation data and distantly supervised data. We can achieve good results simply by setting λ to 1 without adjustment. y_{uni} represents a uniform distribution.

Based on the confidence-calibrated classifier, we specify the rejection rule $f(\cdot)$ as follows:

$$f(x_i) = \begin{cases} y & max_y p(y|x_i) > \theta \\ \mathcal{R}^* & Otherwise, \end{cases} \tag{9}$$

where θ is a threshold hyperparameters, the posterior probability $p(y|x_i)$ is the output of classifier $\boldsymbol{\eta}$ and $\mathcal{R}^*$ denotes the rejection option.

3.4 Class-preserving Transformations

Transformations is the core component of contrastive learning. Our intuition in designing transformation is that feature masks at different views force the model to find more features to discriminate between known relations. These new features can play a vital role in recognizing unseen relations. Why do the above methods work? As shown in Figure 1, due to the *shortcut* phenomenon, the model is more inclined to remember the relations between entities and it would make mistakes when predicting new relations between the same entity pair. Intuitively through the mask mechanism, the model could mask out some features that belong to Obama and the United States, and then it will have to find more other features to distinguish *the president of* from other relations. Therefore it will not learn the *Shortcut* bias of *Obama + the United States = the president of*. In this work, we design three class-preserving transformations to mask partial features as follows.

Token Mask. Token mask works in the process of sentence encoding. In this transformation, we randomly mask a certain proportion of tokens to generate a new view of relation representation.

Random Mask. Random mask also works in the process of sentence encoding. Instead of completely masking representation of selected tokens, each dimension of the representation of each word is considered independently in this transformation.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Algorithm 1: Robust Relation Extraction

Input: known relation dataset $\mathcal{D}^{\ell}$, distantly supervised dataset $\mathcal{D}^{dist}$ (optional), testing dataset $\mathcal{D}^{u}$, transformation set $\mathcal{T}$, model parameters Θ, Φ for encoder and classifier, OpenRE module $\mathcal{O}$ and learning rate α.

Training Phase
repeat
  sample a training batch $\mathcal{B}$ from $\mathcal{D}^{\ell}$;
  obtain transformed batch $\widetilde{\mathcal{B}} = t(\mathcal{B}), t \sim \mathcal{T}$;
  enrich representation by contrastive training (equ 4): $\Theta = \Theta - \alpha \nabla_{\Theta} \mathcal{L}_{cts}^{sup}$;
  sample a distant batch $\mathcal{B}^{dist}$ from $\mathcal{D}^{dist}$;
  optimize classifier by supervised training (equ 7 or 8):
  $\{\Theta, \Phi\} = \{\Theta, \Phi\} - \alpha \nabla_{\{\Theta,\Phi\}} \mathcal{L}^{dist}$;
until *convergence*;
Testing Phase
Filter the unseen relations subset $\mathcal{D}^{rej}$ from $\mathcal{D}^{u}$ by the rejection rule f (equ 9);
Output predictions $\{y_i^u\}$ for the rest samples of known relations;
Run the OpenRE module $\mathcal{O}$ to obtain potential relations in $\mathcal{D}^{rej}$;

Feature Mask. Feature mask works after sentence encoding. Given a relation instance $x_i^{\ell} \in \mathcal{D}^{\ell}$, we first obtain its relation representation $\boldsymbol{r}_i = \boldsymbol{g}(x_i)$. Then we randomly mask a certain proportion of feature dimensions of $\boldsymbol{r}_i$ to generate a new view.

It is certain that a more complicated and diverse transformations will bring additional improvement. This will be one of our future work.

3.5 OpenRE Module

We introduce the OpenRE module for the integrity of the framework, although it is not our main concerns. Based on the rejection rules f described in section 3.3, we can classify samples of known relations while rejecting unseen relations. In this section, we take a step forward. By integrating the off-the-shelf OpenRE method, we try to discover the potential unseen relations in the rejected samples instead of only detecting their existence. We adopt SelfORE (Hu et al., 2020), a clustering-based OpenRE method, as the building block of our OpenRE module. Various other methods can also be used as the alternative to SelfORE without any constraints. More details about OpenRE methods can be found in the related papers. Overall, the method proposed in this paper is detailed in algorithm 1.

4 Experimental Setup

In this section, we describe the datasets for training and evaluating the proposed method. We also detail the baseline models for comparison. Finally, we clarify the implementation details.

4.1 Datasets

We conduct our experiments on two well-known relation extraction datasets. In addition, a distantly supervised dataset are used in a auxiliary way.

FewRel. Few-Shot Relation Classification Dataset (Han et al., 2018). FewRel is a human-annotated dataset containing 80 types of relations, each with 700 instances. We use the top 40 relations as known and the middle 20 relations as unseen. Since the relations of FewRel dataset is exactly the same as that of FewRel-Distance, we hold out the last 20 relations for the use of distant supervision. The training set contains 25600 randomly selected samples of known relations. In order to evaluate the rejection performance to the unseen relations, the test/validation set contains 3200/1600 samples composed of known and unseen relations.

TACRED. The TAC Relation Extraction Dataset (Zhang et al., 2017). TACRED is a human-annotated large-scale relation extraction dataset that covers 41 relation types. Similar to the setting of FewRel, we

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

use the top 31 relations as known and the rest 10 relations as unseen. The training set consists of 18113 randomly selected samples of known relations. The size of validation set and test set are 900 and 1800 respectively, including known and unseen relations. It should be noted that 50% of the unseen relation samples in the validation set and test is `no_relation`.

FewRel-distant. FewRel-distant contains the distantly-supervised data obtained by the authors of FewRel before human annotation. We use this dataset as the distantly supervised data in our experiments.

4.2 Baselines and Evaluation Metrics

MSP (Hendrycks and Gimpel, 2017). MSP assumes that correctly classified examples tend to have greater maximum softmax probabilities than examples of unseen classes. Thereby the maximum softmax probabilities are used as confidence score for unseen classes detection.

MSP-TC (Liang et al., 2018). MSP-TC uses maximum softmax probabilities with temperature scaling and small perturbations to enhance the separability between known and unseen classes, allowing for more effective detection.

DOC (Shu et al., 2017). DOC builds n 1-vs-rest sigmoid classifiers for n known classes respectively. The maximum probability of these binary classifiers is considered as the confidence score for unseen classes detection.

LMCL (Lin and Xu, 2019). Large margin cosine loss (LMCL) aims to learn a discriminative deep representations. It forces the model to not only classify correctly but also maximize inter-class variance and minimize intra-class variance. Based on the learned representations, local outlier factor (LOF) is used to detect unseen classes.

ADB (Zhang et al., 2021). Labeled known classes samples are first used for representation learning. Then the learned representations are utilized to learn the adaptive spherical decision boundaries for each known classes. Samples outside the hypersphere will be rejected for recognition.

Evaluation Metrics. We follow previous work (Zhang et al., 2021; Lin and Xu, 2019) and take all the unseen relations as one rejected class. The accuracy and macro F1 metrics are used as the scoring function to evaluate the unseen relation detection.

4.3 Implementation Details

We use the Adam (Kingma and Ba, 2015) as the optimizer, with a learning rate of $1e-4$ and batch size of 100 for all datasets. If the results don't improve on the validation set for 10 epochs, we stop the training to avoid overfitting. All experiments are conducted using a NVIDIA GeForce RTX 3090 with 24GB memory.

5 Results and Analysis

In this section, we present the experimental results of our method on FewRel and TACRED datasets to demonstrate the effectiveness of our method.

5.1 Main Results

Our experiments in this section focus on the following three related questions.

Can the proposed method effectively detect unseen relations? To answer this question, we consider all the known relations as one predicted class and the rest unseen relations as one rejected class. Table 2 reports model performances on FewRel, TACRED datasets, which shows that the proposed method achieves state-of-the-art results on unseen relation detection. Benefiting from the contrastive training objectives and the carefully designed transformations, the *Shortcut* phenomenon is effectively alleviated, and the model learns sufficient features to discriminate between known and unseen relations. Therefore, the proposed method consistently outperforms the compared baselines by a large margin in different mixing-ratio settings.

Does the detection of unseen relations impair the extraction of known relations? Integrating the rejection option can make the classifier more robust in real applications. However, we do not want the unseen relations detection impair known relations classification, which is the basic function of the

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Dataset	Method	25%		50%		75%	
		Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
FewRel	MSP (Hendrycks and Gimpel, 2017)	0.805	0.781	0.786	0.786	0.797	0.774
	MSP-TC (Liang et al., 2018)	0.802	0.772	0.769	0.769	0.786	0.768
	DOC (Shu et al., 2017)	0.794	0.768	0.781	0.781	0.784	0.761
	LMCL (Lin and Xu, 2019)	0.810	0.785	0.740	0.740	0.835	0.777
	ADB (Zhang et al., 2021)	0.801	0.800	0.837	0.799	0.837	0.784
	Ours	**0.888**	**0.852**	**0.844**	**0.824**	**0.838**	**0.827**
TACRED	MSP (Hendrycks and Gimpel, 2017)	0.758	0.691	0.698	0.688	0.734	0.650
	MSP-TC (Liang et al., 2018)	0.789	0.687	0.674	0.670	0.765	0.671
	DOC (Shu et al., 2017)	0.793	0.687	0.707	0.678	0.775	0.681
	LMCL (Lin and Xu, 2019)	0.737	0.705	0.667	0.684	0.785	0.654
	ADB (Zhang et al., 2021)	0.772	0.714	0.711	0.710	0.767	0.699
	Ours	**0.827**	**0.758**	**0.723**	**0.742**	**0.788**	**0.715**

Table 2: Main results of unseen relation detection with different known class proportions (25%, 50% and 75%) on two relation extraction datasets. Compared with the best results of all baselines, our method improves F_1-score by an average of 2.6%, 3.5% on FewRel and TACRED dataset, respectively.

Dataset	Method	25%	50%	75%
FewRel	MSP	0.730	0.769	0.814
	MSP-TC	0.675	0.771	0.764
	DOC	0.737	0.780	0.805
	LMCL	0.765	0.767	0.809
	ADB	0.778	0.770	0.810
	Ours	**0.827**	**0.793**	**0.828**
TACRED	MSP	0.610	0.619	0.668
	MSP-TC	0.378	0.438	0.639
	DOC	0.628	0.627	0.686
	LMCL	0.616	0.615	0.687
	ADB	0.625	**0.640**	0.665
	Ours	**0.637**	0.633	**0.688**

Table 3: Macro F_1-score of known relation classification with different proportion of known relations.

classifier. From table 3 we can observe that the proposed model not only effectively detect unseen relations, but also accurately classify known relations. This demonstrate that the designed transformation will not affect the original relational semantics, so the rich features obtained by comparative learning remain discriminability for the known relations.

Can the model achieve superior performance under different threshold settings? We show the receiver operating characteristic (ROC) curve in Figure 3. The area under ROC curve (AUROC) summarize the performance of a classifier detecting unseen relations across different thresholds. From Figure 3 we can observe that the AUROC of the proposed method is the largest. Therefore, the proposed method has certain advantages under different threshold settings.

5.2 Ablation Study

To understand the effects of each component of the proposed model, we conduct an ablation study on it and report the results (Macro-F_1) on the two dataset in Table 4. The results show that the detection of unseen relations is degraded if any transformation is removed. It indicates that (1) These transformations force model learn sufficient features through mask mechanism from different views. The learned features are beneficial for the detection of unseen relations. (2) Since the transformations are from different views, they can be superimposed and further enhance the detection of unseen relations. In addition, we find that distantly supervised data can significantly improve the detection of unseen relations. Because there are a large number of diverse relations in the external knowledge base, we can easily construct a large number of negative samples. So this improvement can be seen as a free lunch.

5.3 Relation Representation Visualization

To intuitively show the influence of the rich features learned through contrastive training, we visualize the relational representation with t-SNE (van der Maaten and Hinton, 2008). We select five semantically

CL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

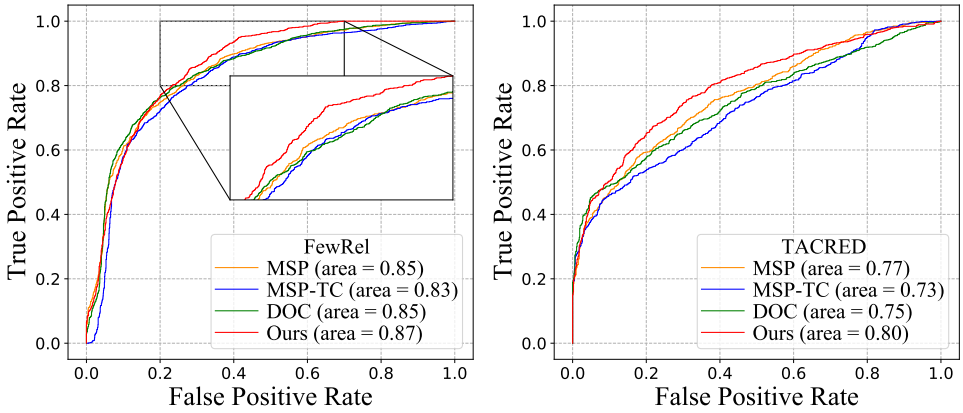


Figure 3: ROC curves on two datasets.

Dataset	Method	25%	50%	75%
FewRel	w/o Feature Mask	0.845	0.807	0.816
	w/o Random Mask	0.846	0.814	0.809
	w/o Token Mask	0.833	0.810	0.803
	w/o Distant	0.810	0.805	0.815
	Ours	**0.852**	**0.824**	**0.827**
TACRED	w/o Feature Mask	0.753	0.728	0.703
	w/o Random Mask	0.740	0.735	0.706
	w/o Token Mask	0.750	0.738	0.706
	w/o Distant	0.716	0.700	0.684
	Ours	**0.758**	**0.742**	**0.715**

Table 4: Abalation study of our method.

similar known relations from FewRel dataset, and randomly select 40 samples for each of them. 100 hard samples of unseen relations misclassified by MSP method are selected to show the superiority of our method. From the visualization results in Figure 4, we can observe that, before training (upper left), the relation representations are scattered in the semantic space. After supervised training (upper right), samples can be roughly divided by relation, but different relations are still close to each other. This is consistent with the *Shortcut* feature in neural network. We note that samples of unseen relations are mixed with known relation samples. After contrastive training (down left), model learns sufficient features to discriminate unseen relations. Therefore, samples of unseen relations are effectively separated. Finally, a best relation representation are obtained by applying both supervised and contrastive optimization (down right).

5.4 A Case Study on OpenRE

For the samples rejected by the classifier, the off-the-shelf OpenRE method can be used to discovery potential unseen relations. In this section, we provide a brief case study to show the discovered unseen relations by SelfORE (Hu et al., 2020). OpenRE module outputs the cluster assignment of these

Extracted surface-form	Golden surface-form
university	schools_attended
was found	founded
charges with	charges
died in	country_of_death
was born in	date_of_birth

Table 5: Extracted and golden surface-form relation names on TACRED.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

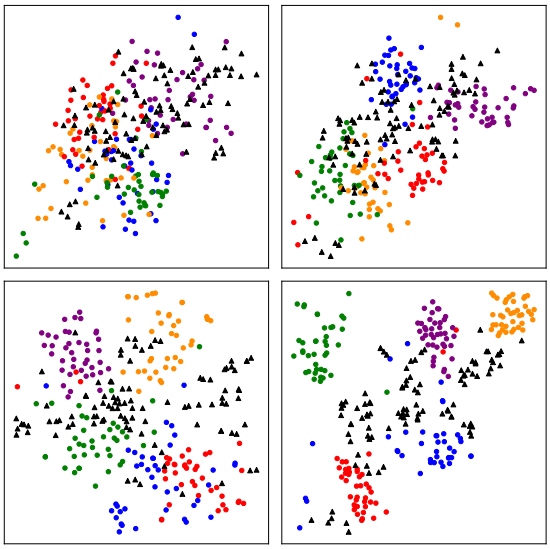


Figure 4: Visualization of the relation representation after t-SNE dimension reduction. The representations are colored with their ground-truth relation labels. Black triangles indicate unknown relations. These four from top left to bottom right sequentially illustrate the relation representation of initial state, after supervised optimization, after contrastive optimization, after both of them.

rejected samples. We extract the relation names using the frequent n-gram in each cluster and the extraction results are shown in table 5. By integrating the OpenRE module, our method complete (1) the classification of known relations, (2) the rejection of unseen relations, (3) discovery of unseen relations. Based on the above process, robust relation extraction in real applications is realized.

6 Conclusions

In this work, we introduce a relation extraction method with rejection option to improve the robustness in real-world applications. The proposed method employs contrastive training techniques and a set of carefully designed transformations to learn sufficient features. The classification of known relations and rejection of unseen relations can be done with these features. Unseen relations in the rejected samples can be discovered by incorporating off-the-shelf OpenRE methods. Experimental results show that our method outperforms SOTA methods for unseen relation rejection.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

References

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.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

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.

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2216–2225, Brussels, Belgium, October-November. Association for Computational Linguistics.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Hervé Gallaire and Jack Minker, editors, 1978. *On Closed World Data Bases*, pages 55–76. Springer US, Boston, MA.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium, October-November. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *CoRR*, abs/2104.08202.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip S. Yu. 2020. Selfore: Self-supervised relational feature learning for open relation extraction. *CoRR*, abs/2004.02438.

Yi Yao Huang and William Yang Wang. 2017. Deep residual learning for weakly-supervised relation extraction. *CoRR*, abs/1707.08866.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *arXiv e-prints*, page arXiv:1907.10529, July.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496, Florence, Italy, July. Association for Computational Linguistics.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

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.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

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 e-prints*, page arXiv:1907.11692, July.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Yaqian Zhou, and Xuanjing Huang. 2021. SENT: sentence-level distant relation extraction via negative training. *CoRR*, abs/2106.11566.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August. Association for Computational Linguistics.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2014. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. *arXiv e-prints*, page arXiv:2010.01923, October.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471, 07.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916, Copenhagen, Denmark, September. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. 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 872–884, New Orleans, Louisiana, June. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam. 2020. Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060, Online, July. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium, October-November. Association for Computational Linguistics.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. 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 3016–3025, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14374–14382.

Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. 2021. A relation-oriented clustering method for open relation extraction.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Using Extracted Emotion Cause to Improve Content-Relevance for Empathetic Conversation Generation

Minghui Zou[*], **Rui Pan**[*], **Sai Zhang**[†], **Xiaowang Zhang**
College of Intelligence and Computing, Tianjin University, Tianjin, China
zhang_sai@tju.edu.cn

Abstract

Empathetic conversation generation intends to endow the open-domain conversation model with the capability for understanding, interpreting, and expressing emotion. Humans express not only their emotional state but also the stimulus that caused the emotion, i.e., emotion cause, during a conversation. Most existing approaches focus on emotion modeling, emotion recognition and prediction, and emotion fusion generation, ignoring the critical aspect of the emotion cause, which results in generating responses with irrelevant content. Emotion cause can help the model understand the user's emotion and make the generated responses more content-relevant. However, using the emotion cause to enhance empathetic conversation generation is challenging. Firstly, the model needs to accurately identify the emotion cause without large-scale labeled data. Second, the model needs to effectively integrate the emotion cause into the generation process. To this end, we present an emotion cause extractor using a semi-supervised training method and an empathetic conversation generator using a biased self-attention mechanism to overcome these two issues. Experimental results indicate that our proposed emotion cause extractor improves recall scores markedly compared to the baselines, and the proposed empathetic conversation generator has superior performance and improves the content-relevance of generated responses.

1 Introduction

Open-domain conversation generation has made remarkable progress over recent years, relying on deep learning and neural networks (Serban et al., 2016; Wolf et al., 2019; Zhou et al., 2020; Huang et al., 2020). However, previous works primarily centre around improving the linguistic quality of the generated responses, such as grammatical correctness, content variety, and topic relevance, neglecting the important factor of emotion (Zhou et al., 2018). The information conveyed by humans during communication contains not only syntactic and semantic information but also emotional information. Emotion is one of the essential cognitive behaviors in humans, and artificial intelligence has the objective of enabling machines to mimic human intelligent behaviors. As an important research branch of AI, one of the long-term goals of dialogue systems is to enable machines to perceive, comprehend and express emotions. In addition, studies (Martinovski and Traum, 2003; Prendinger and Ishizuka, 2005; Prendinger et al., 2005) have shown that introducing emotional information into conversation systems can improve user engagement and satisfaction, make human-computer conversation more natural, and reduce the number of conversation terminations.

As a new research hotspot for the NLP community, most existing approaches on empathetic conversation generation focus on identifying the emotion category of the input sequence and generating a response based on the predicted emotion label. Song et al. (2019) introduce an external emotion lexicon into the generation module to achieve explicit and implicit emotion expression. Li et al. (2020) create a coarse-grained and fine-grained emotion representation for the input sequence by using an emotion classifier to identify the input sequence's emotion category and an external emotion lexicon to identify

[*]These authors contributed equally to this work and should be considered co-first authors.
[†]Corresponding author.

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the emotion terms. Majumder et al. (2020) improve the empathetic expression in the generated response by mimicking the input sequence's emotion while taking into account their emotional polarity. Firdaus et al. (2021) incorporate multiple emotions to varying degrees in the generation of responses to make the model more anthropomorphic.

Existing works mainly focus on emotion-related issues, focusing less on content-relevance. However, a case study of the responses generated by the existing models shows that the existing models do not guarantee the content-relevance of the generated responses very well. As shown in Table 1, in case 1, EmpDG (Li et al., 2020) generates an emotionally irrational and irrelevant response, whereas GPT2 (Radford et al., 2019) can express empathy for the user's emotion, but the generated response deviates from the conversation's topic (from diet to age); in case 2, EmpDG and GPT2 both focus too much on the user's lonely emotion but fail to develop the conversation around the user's specific situation, resulting in the generation of a safe response, which is also irrelevant to the situation.

Input 1	My son really needs to cut down on the amount of sugary snacks he has!
Emotion	Disappointed
EmpDG	That is so disgusting! I bet he is a very disgusting situation.
GPT2	I know how you feel. I have a son who is 6 and a daughter who is 5.
Input 2	I feel so lonely sometimes because all my friends live in a different country .
Emotion	Lonely
EmpDG	I am sorry you had to go through that.
GPT2	Oh no! I hope you are ok!

Table 1: Examples of empathetic conversation.

We suggest there are two main reasons: Firstly, as Gao et al. (2021) pointed out, existing works deal with emotions on a surface level and do not consider the underlying causes of the emotion, making it difficult to comprehend user's complicated emotions and badly affecting the subsequent links of emotion prediction and empathetic conversation generation. Secondly, emotion category is a strong supervisory signal, and overemphasizing its importance in the process of generating responses can easily lead to the generation of safe responses for specific emotion categories. For examples in Table 1, if the model can accurately capture the emotion cause in the input sequence (as highlighted in yellow) and incorporate them into the process of generating responses, the model will have the ability to understand the user's emotion better and generate responses with more relevant content by developing topics around the facts conveyed by the user during the generation process.

To this end, we propose an empathetic conversation generation model enhanced by emotion cause to improve the content-relevance of generated responses. Specifically, our model involves two components, an emotion cause extractor and an empathetic conversation generator. In order to accurately identify emotion cause in the absence of large-scale labeled data, we present a semi-supervised training method to optimize the emotion cause extractor. To integrate the extracted emotion cause into the empathetic conversation generator and minimize the damage to the general language knowledge already learned by the pre-trained language model, we introduce a biased self-attention mechanism to enhance the model's attention to the emotion cause when generating responses.

The contributions of our work are summarized as follows:

- To compensate for the scarcity of large-scale word-level emotion-cause labeled datasets, a semi-supervised training method using labeled and unlabeled data for joint training is proposed.
- To integrate the extracted emotion cause into the generation process, a biased self-attention mechanism that does not introduce new additional parameters is proposed.
- Experimental results indicate that our proposed model performs superior to the baselines and improves the content-relevance of the generated responses.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2 Related Work

Empathetic conversation generation has made great progress in recent years. Several works (Song et al., 2019; Shen and Feng, 2020; Welivita and Pu, 2020; Zheng et al., 2021; Sabour et al., 2022; Shen et al., 2021) attempt to make dialogue models more empathetic and have achieved promising results. Song et al. (2019) introduce an external emotion lexicon into the generation module to achieve explicit and implicit emotion expression. Shen et al. (2020) present a novel framework that extends the emotional conversation generation through a dual task and alternatively generates the responses and queries. Welivita et al. (2020) combine dialogue intent modeling and neural response generation to obtain more controllable and empathetic responses. Zheng et al. (2021) propose a multi-factor hierarchical framework to model communication mechanism, dialog act and emotion in a hierarchical way. Sabour et al. (2022) introduce external commonsense information to absorb additional information about the situation and help the model better understand the user's emotion.

Emotion cause extraction is intended to discover the stimulus reasons behind the user's emotion (Lee et al., 2010; Chen et al., 2010). Although there has been a lot of excellent works in this research direction (Xia and Ding, 2019; Bao et al., 2022; Turcan et al., 2021), most of the existing datasets are at the sentence/sub-sentence level (Kim et al., 2021). There is still a lack of a large-scale word-level emotion-cause labeled dataset up till now.

Most existing approaches on empathetic conversation generation only consider superficial emotional information in the dialogue context but ignore deeper emotional causes. Recently, some researches (Gao et al., 2021; Kim et al., 2021) have attempt to investigate emotion cause in empathetic conversation generation, resulting in more relevant and empathetic responses. Since there is no large-scale word-level emotion-cause labeled dataset, Gao et al. (2021) train an emotion cause extractor using a sentence-level labeled dataset and then automatically construct a word-level labeled dataset. Kim et al. (2021) use a Bayesian conditional probability formula based on the emotion category of the dialogue context to train an emotion cause extractor in a weakly supervised way. In order to incorporate emotion cause into the process of generating responses, Gao et al. (2021) introduce a soft gating mechanism and a hard gating mechanism to make model boost the attention on emotion cause; while Kim et al. (2021) introduce the RSA framework, which is essentially a Bayesian conditional probability-based response rewriting module based on the original decoder.

3 Task Formulation

Emotion cause extraction. Given an input sequence $X_e = (x_1, x_2, ..., x_k)$, the goal is to predict the emotion cause probability $C = (c_1, c_2, ..., c_k)$ that indicates whether the token is an emotion cause. Specifically, we add special tokens [CLS] and [SEP] at the beginning and end of the sequence, respectively (as shown in Figure 1).

Empathetic conversation generation. Given an input sequence $X_g = (x_1, x_2, ..., x_n)$, the goal is to generate a response $Y = (y_1, y_2, ..., y_m)$ that is empathetic and relevant to the conversation. Specifically, follow the previous works (Lin et al., 2019; Shin et al., 2020; Gao et al., 2021), we concatenate all utterances in the dialogue context together as input and separate utterances by [SEP] tokens (as shown in Figure 1).

4 Approach

Our proposed emotion-cause-enhanced empathetic conversation generation model consists of two main modules: Emotion Cause Extractor and Empathetic Conversation Generator. The overview is shown in Figure 1. Since there is no large-scale word-level emotion cause dataset available, we present a semi-supervised training method to obtain the emotion cause extractor using small-scale labeled data jointly trained with large-scale unlabeled data. To involve the emotion cause in the generation process, we introduce multiplicative signals to implement the biased self-attention mechanism. The multiplicative signal enhances the model's attention to the emotion cause in the generation process and improves the content-relevance of the generated responses.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

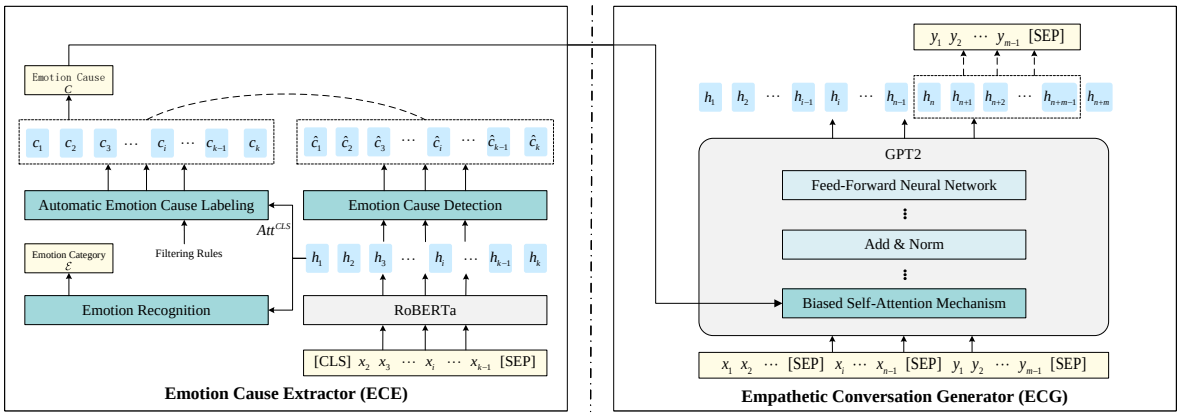


Figure 1: The overview of our proposed ECE and ECG.

4.1 Emotion Cause Extractor

The RoBERTa model (Liu et al., 2019) created by stacking the Transformer encoder (Vaswani et al., 2017) can better model contextual information in both directions. We construct the Emotion Cause Extractor (ECE for short) based on the RoBERTa to identify the emotion categories of the input sequence and its emotion causes. Thus the tasks of the ECE can be divided into emotion recognition and emotion cause detection.

4.1.1 Emotion Recognition

Emotion recognition is a classification problem aiming to predict the emotion category of the input sequence. Given a input sequence X_e, the forward propagation process of the model can be defined as:

$$H_h^E = \text{RoBERTa}(X_e) \tag{1}$$

$$P = \text{softmax}\big(W_e H_{h,1}^E + b_e\big) \tag{2}$$

where H_h^E denotes the output of the last hidden layer, and $H_{h,1}^E$ denotes the output of the first token (i.e., [CLS]) in the last hidden layer. W_e and b_e denote the parameters of the feed-forward neural network.

After obtaining the probability distribution P of emotion category, the emotion category of the X_e can be defined as $\mathcal{E} = \text{argmax}(P)$.

We employ the following loss function to optimize the parameters:

$$\mathcal{L}_{emo}(P) = -\sum_{i \in labels} t(i) \log p_i \tag{3}$$

where $labels \in \{1, 2, ..., s\}$ denotes emotion categories, and $t(i)$ denotes the ground truth distribution corresponding to the input sequence.

It is noted that the input representation of the RoBERTa contains both word embedding and positional embedding:

$$H_0^E = X_e W_e^W + X_e^P W_e^P \tag{4}$$

where W_e^W denotes the word embedding matrix, X_e^P denotes the absolute position of tokens in X_e, and W_e^P denotes the positional embedding matrix.

4.1.2 Emotion Cause Detection

Emotion cause detection is a sequence labeling problem that aims to predict whether each token in the input sequence is the emotion cause, i.e., a word-level $\{0, 1\}$ labeling problem. Since no large-scale word-level emotion cause dataset is available, this section proposes a semi-supervised training method using small-scale labeled data jointly with large-scale unlabeled data.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

For the labeled data, given an input sequence X_e, the context-aware word representation is obtained by encoding using the RoBERTa. Then, a layer of the feed-forward neural network is used for $\{0, 1\}$ sequence labeling:

$$H_h^E = \mathrm{RoBERTa}(X_e) \tag{5}$$

$$\widehat{C} = \mathrm{softmax}\big(W_c H_h^E + b_c\big) \tag{6}$$

where $\widehat{C}$ represents the emotion cause probability of each token, W_c and b_c denote the parameters of the feed-forward neural network.

The loss function applied for parameter learning is as follows:

$$\mathcal{L}_{cau}\left(\widehat{C}\right) = -\sum_{i=1}^{k} \log \mathrm{P}\left(\widehat{C}_i\right) \tag{7}$$

where k indicates the length of the input sequence, and $\mathrm{P}(\cdot)$ denotes obtaining the probability corresponding to the ground truth label of each token.

For the unlabeled data, we observe that the model needs to pay attention to the emotion cause when predicting the emotion category of the input sequence. Thus the attention weight distribution of the model in predicting emotion categories can be used to predict whether each token is an emotion cause or not. Given an input sequence X_e, emotion recognition is performed using the RoBERTa to obtain the attention weight distribution Att^{CLS} of the first [CLS] token in the last hidden layer. Then, simple filtering based on the rules (including removing punctuation, special words, stop words, etc.) is applied, and the tokens with top-k weights are selected as the emotion cause of the input sequence. In this way, emotion cause labels can be automatically constructed for unlabeled data, and the rest of the processing is similar to labeled data.

However, the above method of automatic emotion cause labeling requires converting each token from vector to text at the realization and then performing rule-based filtering. This leads to the fact that the computational graph of automatic emotion cause labeling module is not fully linked with that of emotion cause detection module, i.e., the loss function $\mathcal{L}_{cau}$ of emotion cause detection is not derivable for Att^{CLS}, and cannot be directly involved in the optimization of Att^{CLS}. Thus we propose an additional auxiliary loss function to link the computational graph and introduce the regularization constraint by computing the vector inner product of Att^{CLS} and $\widehat{C^1}$:

$$\mathcal{L}_{aux}\left(Att^{CLS}, \widehat{C}\right) = Att^{CLS} \cdot \widehat{C^1} \tag{8}$$

where $\widehat{C^1} = \widehat{C}\,[1, :]$ denotes the probability that each token is the emotion cause.

In summary, we employ the following loss function to optimize the emotion cause extractor:

$$\mathcal{L}^{ECE} = \lambda_1 \mathcal{L}_{emo} + \lambda_2 \mathcal{L}_{cau} + \lambda_3 \mathcal{L}_{aux} \tag{9}$$

where λ_i indicates the weight of each loss function (we set $\lambda_1 = 1/3$, $\lambda_2 = \lambda_3 = 1$).

4.2 Empathetic Conversation Generator

4.2.1 Conversation Generation

Given a input sequence X_g and the corresponding probability of emotion cause C, the goal of the Empathetic Conversation Generator (ECG for short) is to maximize the probability $P\left(Y|X_g, C\right)$. The empathetic conversation generator proposed in this section is implemented based on the GPT2 (Radford et al., 2019). Forward propagation process of the GPT2 in conversation generation task can be defined as:

$$H_h^G = \mathrm{GPT2}(X_g) \tag{10}$$

$$\widehat{Y} = \mathrm{softmax}\big(W_g H_h^G + b_g\big) \tag{11}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

where W_g and b_g denote the parameters of the feed-forward neural network.

The loss function is as follows:

$$\mathcal{L}^{ECG}\left(\widehat{Y}\right) = -\sum_{i=1}^{m} \log \mathrm{P}\left(\widehat{Y}_i\right) \tag{12}$$

where m denotes the length of the sequence, and $\mathrm{P}(\cdot)$ denotes obtaining the probability corresponding to the ground truth.

It is noted that the input representation of the GPT2 contains three parts: word embedding, positional embedding and role embedding:

$$H_0^G = X_g W_g^W + X_g^P W_g^P + X_g^R W_g^R \tag{13}$$

where X_g^R denotes the role identifier of each token in the input sequence X_g (used to distinguish different speakers), and W_g^R denotes the role embedding matrix.

4.2.2 Biased Self-Attention Mechanism

In order to integrate the emotion cause into the generation progress of the GPT2, it is typical to introduce a new attention mechanism layer. However, considering that the GPT2 has large-scale, trained parameters, if a new attention mechanism layer is introduced in the fine-tuning phase, it may greatly impact the original parameters and destroy the general knowledge already learned by the GPT2. Therefore we chose to introduce multiplicative signals based on emotion cause on top of the original self-attention mechanism of the GPT2 to enhance the model's attention to emotion cause during generation. Meanwhile, the above possible problems are avoided since no additional parameters are introduced.

Moreover, considering that deep neural networks are biased toward modelling syntactic information at the bottom level and semantic information at the top level, the first few layers of the GPT2 network do not require special attention for the emotion cause. We use the layer number information to scale the above multiplicative signals. As the number of layers increases, the multiplicative signals based on the emotion cause gradually strengthen.

The original self-attention mechanism of the GPT2 is defined as:

$$\mathrm{MaskedAttention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M - \lambda\left(I - M\right)\right) V \tag{14}$$

where $\odot$ denotes the multiplication of the corresponding elements of the matrix, λ denotes an infinite scalar (generally taken as $\lambda = 10000$). M denotes the lower triangular matrix with all non-zero elements being 1, I denotes the matrix where all elements are 1.

Our proposed biased self-attention mechanism based on the emotion cause can be defined as:

$$\mathrm{MaskedScore}(Q, K) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M - \lambda\left(I - M\right)\right) \tag{15}$$

$$\mathrm{BiasedScore}(Q, K) = \mathrm{Normalize}\left(\mathrm{MaskedScore}(Q, K) \odot \left(I + \frac{h_i}{h} C\right)\right) \tag{16}$$

$$\mathrm{Normalize}(X) = \frac{x_{i,j}}{\sum_i x_{i,j}} \tag{17}$$

$$\mathrm{BiasedAttention}(Q, K, V) = \mathrm{BiasedScore}(Q, K)\, V \tag{18}$$

where C represents the probability of each token being an emotion cause, $h_i \in \{1, 2, ..., h\}$ denotes the serial number of the self-attention layer, $\mathrm{Normalize}(\cdot)$ denotes the function for normalization by row.

4.3 Training Strategy

Our proposed model is trained using a two-stage training strategy.

In the first stage, the ECE is trained using a semi-supervised training method, as shown in Algorithm 1.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Algorithm 1: The training process of ECE

```
Input: ECE, EmoCause-1 dataset and EmpDialog dataset
Loading the RoBERTa and randomly initializing other parameters;
for training iteration do
    for data ∈ EmoCause-1 do
        Train ECE in a supervised method;
    end
    for data ∈ EmpDialog do
        Construct emotion cause labels automatically;
        Train ECE in a supervised method based on the emotion cause labels;
    end
end
Output: ECE
```

In the second stage, the ECG is trained based on the emotion cause extracted by the ECE, and the parameters of the ECE are frozen in this stage. The training process is shown in Algorithm 2.

Algorithm 2: The training process of ECG

```
Input: ECG, ECE and EmpDialog dataset
Loading the ECE;
Loading the GPT2 and randomly initializing other parameters;
for training iteration do
    for data ∈ EmpDialog do
        Extract the emotion cause of the input sequence using ECE;
        Integrate the extracted emotion cause into ECG using biased self-attention mechanism;
        Update the parameters of the ECG;
    end
end
Output: ECG
```

5 Experiments

5.1 Datasets

We use the following two datasets to conduct experiments.

Label	Hopeful	
Situation	I have been making goals each week for earning money. I'm hoping to save enough to start renovations on my house.	
Conversation	Speaker:	I have big renovation plans for my house. I've made a money plan and have kept to it so far.
	Listener:	Well at least you have a plan. Are you planning to start the renovation soon?
	Speaker:	Yes, hopefully it will all go as planned. So far so good.
	Listener:	Awesome. I'm sure it's going to turn out great.

Table 2: An example of the EmpDialog dataset.

EmpatheticDialogues (EmpDialog for short) is a dataset for empathetic conversation generation created by Rashkin et al. (2019). The dataset, which contains 19,533 conversations in the training set, 2770 conversations in the validation set and 2547 conversations in the test set, is collected and created by the Amazon Mechanical Turk platform. EmpDialog defines 32 emotion categories, and each conversation

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

is created based on an emotional category and a situation description. An example of the EmpDialog dataset is shown in Table 2.

Label	Hopeful
Situation	I have been making goals each week for earning money. I'm hoping to save enough to start renovations on my house.
Cause	goals, earning, money

Table 3: An example of the EmoCause dataset.

EmoCause is a word-level emotion cause dataset created by Kim et al. (2021) based on the validation and test sets of EmpDialog. The dataset is also collected and created by the Amazon Mechanical Turk platform. The workers are asked to vote for each token in a given $situation$ to determine whether it is the emotion cause. EmoCause have 2770 validation data and 2547 test data. An example of the EmoCause dataset is shown in Table 3.

As described in subsection 4.3 our proposed model is trained in two stages and the experimental data used in different stages are different.

Experimental Data for ECE: The experimental data used by ECE are obtained from EmpDialog and EmoCause. First, the validation set of EmoCause is randomly divided into two equal parts (denoted as EmoCause-1 and EmoCause-2). Then, the training set (unlabeled) of EmpDialog is combined with EmoCause-1 (labeled) to form the training set used in the experiments, EmoCause-2 is used as the validation set for experiments, and the test set of EmoCause is used as the test set for experiments.

Experimental Data for ECG: The experimental data used in ECG are derived from EmpDialog, and the division of the training set, validation set and test set is the same as the original dataset.

5.2 Comparison Methods

For ECE, we chose the following three models as baselines: (1) **EmpDG** (Li et al., 2020): a Transformer-based model that creates the coarse and fine-grained emotion representation by emotion classification and external emotion lexicon. In addition, it uses two discriminators to interact with user feedback. Here, we select the coarse-grained tokens as the emotion cause. (2) **RoBERTa_Att:** a RoBERTa-based (Liu et al., 2019) model that is trained on the emotion recognition task, we obtain emotion cause by the attention weight distribution of the first special token [CLS]. (3) **GEE** (Kim et al., 2021): a BART-based (Lewis et al., 2020) model that uses a Bayesian conditional probability formula based on the emotion category labels of context to predict emotion cause.

For ECG, we chose the following three models as baselines: (1) **EmpDG** (Li et al., 2020): the same as mentioned above. (2) **RecEC** (Gao et al., 2021): a Transformer-based model that incorporates emotion cause into response generation with gating mechanisms. It constructs emotion cause labels using a pre-trained sentence-level emotion cause extractor. (3) **GPT2** (Radford et al., 2019): a GPT2-based model that is fine-tuned on the conversation generation task.

5.3 Evaluation Metrics

For ECE, we conducted the automatic evaluation to evaluate with the following metrics: emotion classification accuracy (Accuracy for short) and emotion cause recall rate (Recall for short).

For ECG, we used automatic evaluation and manual evaluation to verify the effectiveness. The metrics used for the automatic evaluation included Perplexity, Distinct-1, Distinct-2, and emotion classification accuracy (Accuracy for short), well-known metrics commonly used to evaluate conversation generation. Additionally, we introduced BERTscore (Zhang et al., 2020) to measure the cosine similarity between the generated response and the gold response. BERTscore contains three more specific metrics, namely recall rate ($\mathrm{R_{BERT}}$), precision rate ($\mathrm{P_{BERT}}$) and F1 score ($\mathrm{F_{BERT}}$).

The manual evaluation included both quantitative and qualitative components. The quantitative component required scorers to score on three dimensions of Empathy, Relevance, and Fluency, with each dimension being scored in an increasing value domain from 1 to 5. The qualitative component required

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

scorers to rank the response generated by different models in order of preference. The manual evaluation randomly selected 100 test data and disrupted the responses generated by different models. Afterwards, these responses are distributed to 3 scorers for scoring, and the final results are averaged. The above approach fully ensures the fairness of the manual evaluation.

5.4 Parameter Settings

ECE is constructed based on RoBERTa-base, and ECG is constructed based on GPT2-base. Table 4 is drawn to show the parameter settings in detail.

	ECE	ECG
Initial learning rate	0.00002	0.00002
Gradient reduce strategy	ReduceLROnPlateau	ReduceLROnPlateau
Gradient clip threshold	1	1
Gradient accumulation threshold	1	2
Batch size	64	8
Early stopping strategy	Top-5 Recall	Perplexity
Early stopping threshold	5	5

Table 4: Parameter setting of ECE and ECG.

5.5 Experimental Results and Analysis

Model	Accuracy	Top-1 Recall	Top-3 Recall	Top-5 Recall
EmpDG	0.31	0.134	0.362	0.493
Roberta_Att	0.58	0.148	0.399	0.596
GEE	0.40	0.173	0.481	0.684
ECE (Ours)	**0.58**	**0.227**	**0.565**	**0.727**

Table 5: Results on comparative experiments of the different Emotion Cause Extractors.

Table 5 shows the experimental results of different emotion cause extractors. Our ECE performs optimally in all metrics compared to the comparison methods. Compared with the Roberta_Att, ECE maintains its original strong competitiveness in emotion classification accuracy while achieving remarkable improvement in emotion cause recall rate. These achievements demonstrate that our proposed semi-supervised training method can effectively narrow the gap between emotion recognition and emotion cause detection and significantly improve the emotion cause detection ability of the model.

Training Dataset	Accuracy	Top-1 Recall	Top-3 Recall	Top-5 Recall	Training Method
train	0.56	0.147	0.410	0.607	unsupervised
valid	0.56	0.246	0.514	0.556	supervised
merge (ours)	**0.58**	**0.227**	**0.565**	**0.727**	semi-supervised
merge w/o $\mathcal{L}_{aux}$	0.58	0.208	0.523	0.709	semi-supervised

Table 6: Results on ablation study of the ECE.

We design the ablation study to further analyze the effectiveness of our proposed semi-supervised training method. In Table 6, the "train" (or "valid") in Training Dataset represents that ECE uses only the training (or validation) set of EmoCause for unsupervised (or supervised) training. Similarly, "merge" represents that ECE uses the training set of EmpDialog with EmoCause-1 for semi-supervised training. Note that in the "valid" set of experiment, the test set of EmoCause is used as the validation set, which is actually not a regular practice and is only required here to meet the need of the ablation experiments because we do not have more labeled data.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

The experimental results in Tabel 6 show that the supervised training method is outstanding on Top-1 Recall and Top-3 Recall compared with the unsupervised training method. Still, the supervised training method is significantly weaker than the unsupervised training method on Top-5 Recall. This phenomenon declares that the supervised training method is superior to the unsupervised training method in performance, but it can easily cause overfitting and lead to instability. In contrast, the semi-supervised training method has the advantage of combining the two. On the one hand, supervised training can be used to provide a clear, task-appropriate optimization goal for emotion cause detection. On the other hand, the labeled data can guide the processing of automatic emotion cause labeling and the unlabeled data can avoid overfitting that may result from using only labeled data. In addition, an ablation study on $\mathcal{L}_{aux}$ under the semi-supervised training method also validates the effectiveness of our proposed auxiliary loss function.

Model	Perplexity	Distinct-1	Distinct-2	P_{BERT}	R_{BERT}	F_{BERT}	Accuracy
EmpDG	34.311	0.018	0.069	0.252	0.213	0.232	0.314
RecEC	177.825	0.019	0.090	0.225	0.177	0.201	0.412
GPT2	14.132	**0.027**	**0.112**	0.304	0.238	0.271	/
ECG (Ours)	**14.063**	0.025	0.109	**0.307**	**0.240**	**0.273**	**0.598**

Table 7: Results on Automatic Evaluation of the ECG. It should be noted that the particularly large Perplexity of RecEC is because the model is trained with F_{BERT} as the optimization target for the early stop strategy.

Table 7 demonstrates the automatic evaluation results of different empathetic conversation generation models. Our ECG achieves remarkable improvements in all metrics compared with EmpDG and RecEC, which are Transformer-based models. ECG also makes a small improvement in all metrics except Distinct compared with the pre-trained language model GPT2. The above phenomenon suggests that our ECG can improve the quality of the generated responses by introducing attention to emotion cause on the basis of pre-trained language models. Regarding the poor performance of ECG on Distinct, it may be due to the limitations caused by the emotion cause in the generation process.

Model	Empathy	Relevance	Fluency
EmpDG	2.927	2.763	4.497
RecEC	2.893	2.790	4.677
GPT2	3.213	3.257	4.753
ECG (Ours)	**3.383**	**3.553**	**4.763**

Table 8: Results on Manual Evaluation of the ECG.

Pref. (%)	EmpDG	RecEC	GPT2	**ECG (Ours)**
EmpDG	/	47.1	26.7	29.9
RecEC	52.9	/	38.2	31.0
GPT2	73.3	61.8	/	42.5
ECG (Ours)	**70.1**	**69.0**	**57.5**	/

Table 9: Preference test (%) between any two method.

Table 8 shows the manual evaluation results of different empathetic conversation generation models. The improvement in Empathy and Relevance of the responses generated by ECG is remarkable, which indicates that introducing attention to the emotion cause in the generation process can promote the model's understanding of user emotion and generate more content-relevant emotional responses. Table 9 presents the preferences of scorers for different models. The scorers' preference for our ECG is greater than the other models, which verifies the validity of the ECG.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

5.6 Case Study

Input	I recently went to eat a restaurant that was not very clean. The whole place had an odd odor and made me uncomfortable.
Emotion	Disgusted
EmpDG	I would have cried right now.
RecEC	Oh no! I hate when people do that.
GPT2	Oh no! Did you clean it up?
ECG (Ours)	That's not good. Did you complain to the manager?
Input	My son was just born, My first boy after 6 girls !
Emotion	Joyful
EmpDG	Wow, That is amazing. How old is your son?
RecEC	Oh that's great. Is he a big boy?
GPT2	Wow! That's amazing! I bet you are proud of him!
ECG (Ours)	Congratulations ! That is a very exciting time for you and your family .

Table 10: Two cases of responses generated by different models.

To further illustrate that focusing on the emotion cause helps improve the content-relevance of the generated responses, we show two cases in Table 10. In the first case, ECE identifies the emotion cause in user input (as highlighted in yellow) and understands the stimulus behind the user's disgusted emotion is the poor environment of the restaurant, which prompts ECG to generate an empathetic response expressing sympathy and concerning for subsequent development (as highlighted in pink). In the second case, ECE recognizes the emotion cause in user input (as highlighted in yellow) and understands the stimulus behind the user's joyful emotion is the long-awaited birth of a son, prompting ECG to generate an empathetic response that congratulates to the user and fits the user's family situation (as highlighted in pink).

Comparing the responses generated by different models in the above two cases, it can be seen that our proposed model can accurately capture the emotion cause in user input and effectively incorporate it into the generation process, showing stronger content-relevance compared to other baselines, which further illustrates the important role of the emotion cause in the content-relevance of generated responses.

6 Conclusion

In this paper, we present an empathetic conversation generation model enhanced by the emotion cause to make the generated responses more content-relevant. Our proposed model comprises an emotion cause extractor and an empathetic conversation generator. To compensate for the scarcity of large-scale word-level emotion-cause labeled datasets, we suggest a semi-supervised training method that simultaneously uses labeled and unlabeled data for training. To integrate the extracted emotion cause into the generation process, we propose a biased self-attention mechanism that does not introduce new additional parameters. Experimental results indicate that our proposed model performs superior to the baselines and the generated responses of our model are more empathetic and content-relevant.

Acknowledgements

This work was supported by the Joint Project of Tianjin University-Bohai Bank Joint Laboratory for Artificial Intelligence Technology Innovation and Bayescom.

References

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Multi-granularity semantic aware graph model for reducing position bias in emotion cause pair extraction. In *Findings of the Association for*

Computational Linguistics: the 60th Conference of the Association for Computational Linguistics (ACL), pages 1203–1213. Association for Computational Linguistics, Dublin, Ireland.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 179–187. Tsinghua University Press, Beijing, China.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2021. More the merrier: Towards multi-emotion and intensity controllable response generation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 12821–12829. AAAI Press, Virtual Event.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 807–819. Association for Computational Linguistics, Punta Cana, Dominican Republic (Virtual Event).

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems*, 38(3):1–32.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2227–2240. Association for Computational Linguistics, Punta Cana, Dominican Republic (Virtual Event).

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53. Association for Computational Linguistics, Los Angeles, USA.

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 (ACL)*, pages 7871–7880. Association for Computational Linguistics, Online.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4454–4466. International Committee on Computational Linguistics, Barcelona, Spain (Online).

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132. Association for Computational Linguistics, Hong Kong, China.

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*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: mimicking emotions for empathetic response generation. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979. Association for Computational Linguistics, Online.

Bilyana Martinovski and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the International Speech Communication Association Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 11–16. ISCA Archive, Château-d'Oex, Vaud, Switzerland.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users'affective states. *Applied artificial intelligence*, 19(3-4):267–285.

Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies*, 62(2):231–245.

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.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pages 5370–5381. Association for Computational Linguistics, Florence, Italy.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: Commonsense-aware empathetic response generation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 11229–11237. AAAI Press, Virtual Event.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Lei Shen and Yang Feng. 2020. CDL: curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 556–566. Association for Computational Linguistics, Online.

Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. 2021. Constructing emotional consensus and utilizing unpaired data for empathetic dialogue generation. In *Findings of the Association for Computational Linguistics: the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3124–3134. Association for Computational Linguistics, Punta Cana, Dominican Republic (Virtual Event).

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *Proceedings of the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7989–7993. IEEE, Barcelona, Spain.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pages 3685–3695. Association for Computational Linguistics, Florence, Italy.

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: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 3975–3989. Association for Computational Linguistics, Online Event.

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 the 30th Conference on Annual Conference Neural Information Processing Systems (NeurIPS)*, pages 5998–6008. MIT Press, Long Beach, USA.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4886–4899. International Committee on Computational Linguistics, Barcelona, Spain (Online).

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pages 1003–1012. Association for Computational Linguistics, Florence, Italy.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. OpenReview.net, Addis Ababa, Ethiopia.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 813–824. Association for Computational Linguistics, Online.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 730–739. AAAI Press, New Orleans, USA.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

To Adapt or to Fine-tune: A Case Study on Abstractive Summarization

Zheng Zhao* **Pinzhen Chen**
School of Informatics, University of Edinburgh
{zheng.zhao, pinzhen.chen}@ed.ac.uk

Abstract

Recent advances in the field of abstractive summarization leverage pre-trained language models rather than train a model from scratch. However, such models are sluggish to train and accompanied by a massive overhead. Researchers have proposed a few lightweight alternatives such as smaller adapters to mitigate the drawbacks. Nonetheless, it remains uncertain whether using adapters benefits the task of summarization, in terms of improved efficiency without an unpleasant sacrifice in performance. In this work, we carry out multifaceted investigations on fine-tuning and adapters for summarization tasks with varying complexity: language, domain, and task transfer. In our experiments, fine-tuning a pre-trained language model generally attains a better performance than using adapters; the performance gap positively correlates with the amount of training data used. Notably, adapters exceed fine-tuning under extremely low-resource conditions. We further provide insights on multilinguality, model convergence, and robustness, hoping to shed light on the pragmatic choice of fine-tuning or adapters in abstractive summarization.

1 Introduction

In the current era of research, using large pre-trained language models (PLM) and fine-tuning these models on a downstream task yields dominating results in many tasks (Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Brown et al., 2020). The scope of our work is on abstractive summarization, which is the task of generating a concise and relevant summary given a long document. Recent works have demonstrated the success of fine-tuning PLMs on summarization (Liu and Lapata, 2019; Zhang et al., 2020; Rothe et al., 2020). Nonetheless, such a paradigm becomes increasingly expensive with the ever-growing sizes of PLMs, since both the training time and space requirement increase along with the number of parameters. The issue becomes more severe when multiple languages or domains are introduced, as separate models need to be trained and saved depending on the setup.

Houlsby et al. (2019) proposed lightweight adapters as an alleviation of the large overhead of fine-tuning PLM on a downstream task. While many researchers have followed and adopted their idea, experiments are rarely done on summarization; from both quantitative and qualitative perspectives, it remains a myth of which direction one should pick in practice. In this work, we perform a thorough exploration of using adapters with a PLM on the task of abstractive summarization by examining different scenarios.

Our experiments are designed along three dimensions: 1) languages involved: monolingual, cross-lingual, and multilingual; 2) data availability: high, medium, low, and scarce; 3) knowledge being transferred: languages, domains as well as tasks. Through comprehensive experimental results, we demonstrate that with a realistic availability of resources, fine-tuning a PLM is superior to using adapters for the purpose of obtaining the best text quality. However, the game changes under low-resource settings: adapters have shown better, if not, on par performances compared to fine-tuning, especially in domain adaption.

*Corresponding author
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2 Related Work

Fine-tuning a PLM with downstream task-specific objectives is a useful paradigm. It not only speeds up training, but also transfers the knowledge from abundant pre-training data to lower-resourced tasks. Whilst it has been proven successful in the field of summarization (Ladhak et al., 2020; Zhang et al., 2020; Zou et al., 2020; Rothe et al., 2020), this strategy requires optimizing and updating all parameters in the fine-tuned model, and is particularly expensive when a number of (sub-)tasks need to be approached.

To mitigate these problems, Houlsby et al. (2019) proposed to insert small neural modules named "adapters" to each layer of the PLM sequentially, and only update the adapters during fine-tuning while freezing most of the PLM parameters. When dealing with different sub-tasks – languages, domains, etc. – it is especially storage-efficient as only adapter weights need to be saved instead of the whole fine-tuned model. Several adapter architectures have been designed since then. Pfeiffer et al. (2020b) suggested simply placing adapters after the feed-forward block in each layer of the PLM, instead of adding adapters after both the multi-head attention and feed-forward block as proposed in the original work. Apart from adding adapters sequentially, He et al. (2022) designed an adapter that is parallel to the PLM.

Recent research that had utilized adapters in the task of summarization, argued that the low availability of opinion summarization datasets often leads to the standard fine-tuning method overfitting on tiny datasets (Brazinskas et al., 2022). Thus, they presented an efficient few-shot fine-tuning method based on adapters for opinion summarization. They added adapters to pre-trained models, trained the adapters on a large unlabelled customer reviews dataset, then fine-tuned them on the human-annotated corpus. Their method outperformed standard fine-tuning methods on various datasets. In addition, they showed that the proposed method can generate better-organized summaries with improved coherence and fewer redundancies in the case of summary personalization. Chen and Shuai (2021) created a meta-transfer learning framework for low-resource abstractive summarization, aiming to leverage pre-trained knowledge to improve the performance of the target corpus with limited examples. They inserted adapter modules into their model to perform meta-learning and leverage pre-trained knowledge simultaneously. Their methods are particularly effective under manually constructed low-resource settings on various summarization datasets with diverse writing styles and forms.

In comparison, our work investigates fine-tuning and using adapters in summarization, by comparing the performance of models using the fine-tuning strategy with models using adapters in the case of language adaptability, data availability, and knowledge transfer. For language adaptability, we examine the case of monolingual, cross-lingual, and multilingual summarization. For data availability, we study models trained under low, medium, and high resource scenarios. Lastly, for knowledge transfer, we investigate several factors: languages, domains, and tasks. To the best of our knowledge, adapters have not been tested in these scenarios.

3 Methodology

3.1 Method overview

Our aim is to study two fine-tuning variants for summarization under several settings using a PLM: the *fine-tuning* paradigm, and the *adapter* strategy. Fine-tuning initializes a PLM from a pre-trained checkpoint, then trains and updates the whole model on a summarization dataset. On the other hand, the adapter strategy also initializes a PLM from a pre-trained checkpoint, with adapter modules then inserted into the model. During training, we only update the adapter, the layer normalization parameters, and the final output layer.

We use mBART (Liu et al., 2020) as our backbone PLM for settings involving non-English languages. It is a sequence-to-sequence model pre-trained on large-scale monolingual corpora in 25 languages, with a denoising autoencoding objective. The model is designed to do multilingual machine translation tasks. After training it on a summarization dataset, the model is capable of doing monolingual, cross-lingual, and multilingual summarization. For English-only settings, we use BART (Lewis et al., 2020)

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

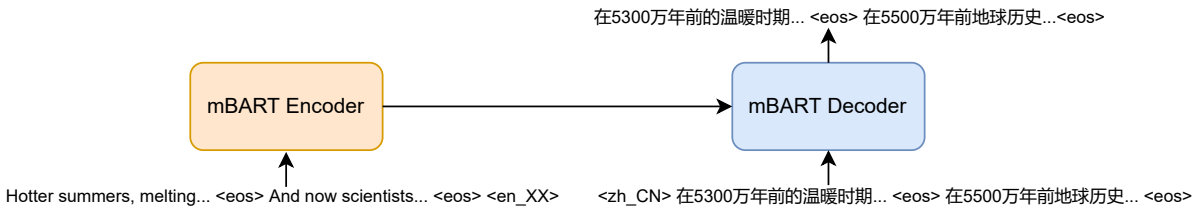


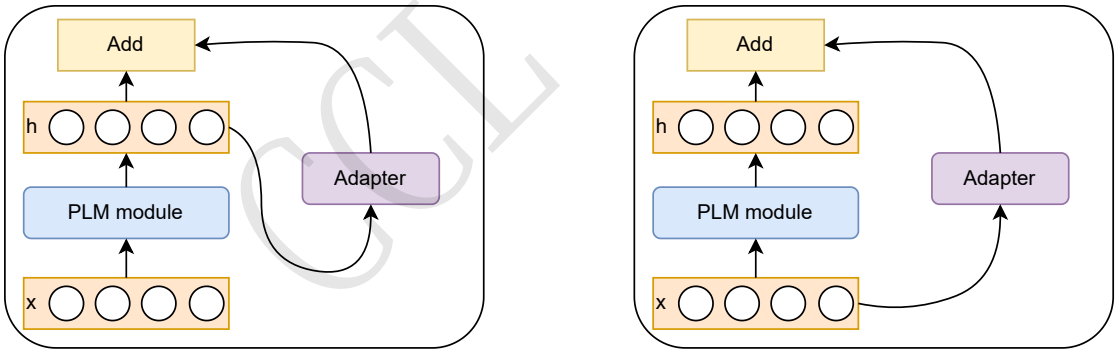
Figure 1: An illustration of our mBART based model for cross-lingual summarization from English to Chinese.

as the PLM. Similar to mBART, BART is also a sequence-to-sequence model pre-trained on large-scale corpora with denoising autoencoder architecture.

We have two kinds of models: mBART-FT which employs the fine-tune strategy, and mBART-Adapt which uses the adapt strategy. In order to recognise the source and target languages, following Liu et al. (2020), our models take a special separator token between each sentence, a language code token at the end of the source document, and at the beginning of the target summary. We provide a cross-lingual demonstration for our model in Figure 1. In addition, we propose BART-FT and BART-Adapt which use the fine-tune strategy and the adapt strategy, respectively.

3.2 Adapter variants

As mentioned earlier, there are various adapter variants. We experiment with two variants: one with sequential connections (Houlsby et al., 2019), and one with parallel connections (He et al., 2022). We display an illustration of these variants in Figure 2. After trying out different learning rates and reduction factors (the ratio between PLM's hidden dimension and adapter's bottleneck dimension), we discover that sequential adapters always outperform the parallel ones in our tasks. Thus we use Houlsby et al. (2019)'s sequential adapter for all of our mBART/BART-Adapt models.



(a) A sequential adapter

(b) A parallel adapter

Figure 2: An illustration of adapter variants, adapted from He et al. (2022). "PLM module" represents a certain sub-layer of the PLM (e.g. attention or feed-forward layer) that is frozen.

3.3 Evaluation

The evaluation metrics are F1 scores of ROUGE-1/2/L (Lin, 2004). Since we deal with multiple languages, we use the multilingual ROUGE implemented in a previous paper.[1] We stick to the toolkit's default settings, e.g., sentence segmentation and word stemming.

[1] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Dataset	Language	train / valid / test	Source
NCLS	zh→en	1.7m / 3.0k / 3.0k	Sina Weibo
	en→zh	365k / 3.0k / 3.0k	CNN/Daily Mail
Wiki-Lingua	en→ar	20.4k / 2.9k / 5.8k	wikiHow
	en→vi	13.7k / 2.0k / 3.9k	
	en↔ja	8.9k / 1.3k / 2.5k	
XL-Sum	gu	9.1k / 1.1k / 1.1k	BBC
	fr	8.7k / 1.1k / 1.1k	
	ne	5.8k / 0.7k / 0.7k	
	ko	4.4k / 0.6k / 0.6k	
	si	3.2k / 0.5k / 0.5k	

Table 1: Statistics of datasets and languages for the language adaption experiment.

4 Language Experiments

4.1 Experimental setup

We test our proposed paradigm on NCLS[2], WikiLingua[3], and XL-Sum[4] datasets particularly designed for cross-lingual and multilingual summarization (Zhu et al., 2019; Ladhak et al., 2020; Hasan et al., 2021). These datasets are either machine-translated or crawled from the web.

NLCS is built by machine-translating an existing English (en) dataset (CNN/Daily Mail, by Nallapati et al. (2016)) to Chinese (zh), and vice versa (Sina Weibo, by Hu et al. (2015)). A translated document is only kept if its round-trip translation reaches a certain threshold score. Plain translations and human-corrected translations are supplied as separate test sets; we use the human-corrected set in this work. WikiLingua is constructed by extracting and aligning article-summary pairs from wikiHow. We experiment with three languages that resemble medium and low-resource scenarios: Arabic (ar), Vietnamese (vi), and Japanese (ja).

Different from the cross-lingual datasets, XL-Sum is monolingual. It consists of professionally annotated article-summary pairs from BBC in many languages. The datasets come in various sizes for a number of languages, as shown in Table 1. This dataset allows for multilingual experiments since the data come from the same domain and are not centred on English. We experiment on five low-resource languages: Gujarati (gu), French (fr), Nepali (ne), Korean (ko), and Sinhala (si). For the monolingual scenario, we directly use the monolingual summarization data to train the model. For the cross-lingual setting, since machine translation is a cross-lingual task, we also directly train the model using the cross-lingual summarization data. Lastly, in a multilingual configuration, we simply mix summarization data in different languages, and train the model using the mixed data.

Our experiments are based on a public mBART checkpoint[5]. We use the adapter from Houlsby et al. (2019). Fine-tuning an mBART model updates around 610M parameters in total; the addition of adapters introduces 50M parameters, yet only this 8% are being optimized during training. We use the Adam optimizer for training (Kingma and Ba, 2015), with a learning rate of 1e-5 for mBART, and 1e-4 for mBART with adapters. We set the adapter reduction factor to 2, which means that the bottleneck dimension in an adapter is half of the hidden dimension in mBART. We perform hyperparameter searches on the following: learning rate and reduction factor, and monitor ROUGE scores on the validation set to select the best value. We provide further details of the grid search in Appendix A.

All models are trained on 4 NVIDIA A100 GPUs with a batch size of 12 on NCLS, and 4 on WikiLingua and XL-Sum. The model convergence time is from 1 to 30 hours depending on the dataset used. We use PyTorch (Paszke et al., 2019) for our model implementation. We use the Huggingface library (Wolf et al., 2020) and AdapterHub (Pfeiffer et al., 2020a) for mBART and adapter implementation.

[2] https://github.com/znlp/ncls-corpora
[3] https://github.com/esdurmus/wikilingua
[4] https://github.com/csebuetnlp/xl-sum
[5] https://huggingface.co/facebook/mbart-large-cc25

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Lang.	mBART-FT			mBART-Adapt		
	R1	R2	RL	R1	R2	RL
zh→en	**46.46**	**30.18**	**42.26**	41.41	22.73	36.56
en→zh	**45.22**	**22.49**	**34.38**	40.74	16.83	29.27

(a) High-resource, NCLS.

Lang.	mBART-FT			mBART-Adapt		
	R1	R2	RL	R1	R2	RL
en→ar	**25.85**	**7.35**	**21.01**	24.68	7.26	20.40
en→vi	**33.63**	**15.17**	**26.65**	30.98	13.94	24.59
en→ja	**35.70**	**12.34**	**28.34**	34.06	11.43	27.08
ja→en	**35.24**	**12.38**	**28.09**	33.14	11.54	26.46

(b) Medium and low-resource, WikiLingua.

Table 2: Results for cross-lingual summarization.

Lang.	Multilingual						Monolingual					
	mBART-FT			mBART-Adapt			mBART-FT			mBART-Adapt		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
gu	**20.18**	**6.96**	**18.09**	20.12	6.82	17.99	**20.23**	**6.43**	**17.67**	19.20	5.95	16.96
fr	**33.53**	**14.37**	**26.11**	33.44	14.01	25.63	**33.29**	**13.68**	**25.13**	32.37	13.02	24.73
ne	**24.70**	**9.52**	**22.23**	23.26	8.55	20.94	**24.06**	**9.05**	**21.62**	23.31	8.36	21.01
ko	17.73	**8.76**	16.27	**18.82**	8.12	**17.23**	**19.73**	9.12	**18.07**	19.05	**9.24**	17.73
si	**26.95**	**13.51**	**22.36**	25.68	12.69	21.80	**25.59**	12.25	**21.92**	24.99	**12.30**	21.44

Table 3: Results for low-resource multilingual and monolingual summarization on XL-Sum.

4.2 Results

We first provide results on high-recourse cross-lingual summarization on NCLS in Table 2a. We can see that mBART-FT achieves significantly higher ROUGE scores than mBART-Adapt in both Chinese-to-English as well as English-to-Chinese settings. We then list result numbers on medium and low-recourse cross-lingual summarization on WikiLingua in Table 2b. Similar to the behaviour under the high-resource setting, mBART-FT consistently achieves better ROUGE performance than mBART-Adapt, regardless of the source or target languages. However, we spot that the difference in ROUGE scores is smaller for language pairs with lower resources, which suggests a positive correlation between the gap in performance and training data availability.

In Table 3, we show results of both multilingual (left) and monolingual (right) summarization on XL-Sum. In a multilingual setup, a single model is trained on five languages. whereas in a monolingual setup, five individual models are trained on the five languages separately. We can first see that mBART-FT generally surpasses mBART-Adapt, in both multilingual and monolingual setups. In addition, multilingual models generally outperform monolingual models by a small margin. This behaviour is corroborated by Hasan et al. (2021)'s work that mixing multiple languages altogether during training can result in a positive transfer among them (Conneau et al., 2020).

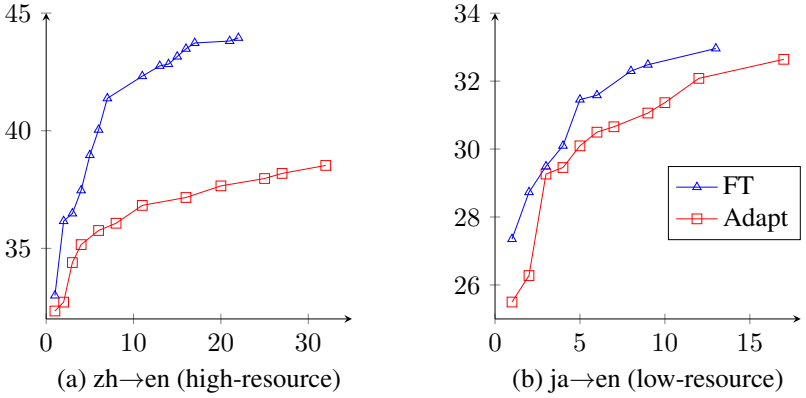
It is straightforward from our work, that, for summarization tasks with high data availability, it is not worth trading performance for efficiency with adapters. For low-resource scenarios, adapters achieve similar results as fine-tuning, and can therefore be a convenient choice for fast training and compact disk storage. When multiple low-resource languages are concerned, especially if they are related languages, it might be beneficial to build a multilingual model instead of individual monolingual models.

4.3 Convergence

To measure the convergence difference between mBART-FT and mBART-Adapt, we plot validation set ROUGE-1 scores against epochs for two previous experiments (high-resource zh→en and low-resource

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

ja→en) in Figure 3. Plotting stops when validation does not improve. We measure convergence in terms of epochs, rather than wall-time. In our experiments, we find that wall-time per epoch for mBART-FT is about merely 1.5 times that for mBART-Adapt, since validation takes a large portion especially when the dataset is small.



(a) zh→en (high-resource) (b) ja→en (low-resource)

Figure 3: Validation ROUGE-1 (y-axis) against epochs (x-axis) for mBART-FT and mBART-Adapt in different data conditions.

As Figure 3(a) shows, with sufficient resources, mBART-FT and mBART-Adapt started with similar ROUGE scores, then the gap quickly increases, suggesting a faster and better convergence rate for fine-tuning. We also observe that mBART-FT converged within fewer epochs. Furthermore, Figure 3(b) suggests that, in a low-resource condition, even though mBART-FT surpasses mBART-Adapt in terms of ROUGE, they both have similar convergence rates with the gap reduced. These trends indicate that in a high-resource scenario fine-tuning is preferred, whereas in a low-resource scenario, adapters can be used to reduce overhead while maintaining performance.

5 Domain Adaptation Experiments

5.1 Experimental setup

In addition to multilinguality, we conduct extra experiments on domain adaptation, which is typically tackled using the same pre-training then fine-tuning paradigm. In our setting, we adapt CNN/Daily Mail to XL-Sum, both in English, with various data sizes. Although both datasets are news articles, they differ hugely in writing styles. We start with a BART model (Lewis et al., 2020) fine-tuned on the CNN/Daily Mail (Nallapati et al., 2016) dataset for summarization; it is available as a public model checkpoint.[6]

To further understand the impact of data availability, we artificially and iteratively make the training data 10 times smaller. This results in five data conditions with sizes ranging from merely 31 to 306.5k. We make sure that larger training splits are supersets of the smaller splits. The validation and test sets remain unchanged at 11.5k as provided in the original dataset. In addition to the XL-Sum dataset, which is in the news domain, we also experimented with adapting CNN/Daily Mail to the BookSum[7] (Kryscinski et al., 2021) dataset, a collection of narratives from the literature domain such as novels, plays, and stories. Their human written summaries have three levels of granularity, and we use the paragraph-level summaries for our experiment. Unlike the CNN/Daily Mail dataset, we only experiment on the full size of the BookSum dataset.

The English BART checkpoint has in total 139M parameters to be fine-tuned, while adapters have 14.2M parameters (10%). As an additional parameter-controlled fine-tuning variant, we choose to freeze the entire BART but the last decoder layer, which has 9.5M parameters. The final decoder layer makes up 7% of the entire model, and has a comparable amount of trainable parameters to an adapter. Similar to the previous setting, we use the Adam optimizer with a learning rate of 1e-5 for BART-FT, and 1e-4

[6] https://huggingface.co/ainize/bart-base-cnn
[7] https://github.com/salesforce/booksum

Domain	Data Size		BART-FT			BART-Adapt			BART-FT-LastLayer		
			R1	R2	RL	R1	R2	RL	R1	R2	RL
XL-Sum	original	306.5k	**34.48**	**14.73**	**28.93**	32.94	13.46	27.60	30.20	11.69	25.17
	medium	30.65k	**30.63**	**11.38**	**25.31**	30.15	11.10	25.05	26.70	8.67	21.94
	small	3065	27.27	**8.91**	**22.27**	**27.32**	8.79	22.20	23.06	6.21	18.76
	tiny	307	24.10	**6.52**	19.38	**24.29**	6.41	**19.50**	19.13	4.12	15.54
	micro	31	19.69	4.26	15.73	**20.74**	**4.65**	**16.45**	16.30	2.20	11.43
BookSum	111.6k		**20.27**	**4.01**	15.50	20.22	3.95	**15.57**	19.33	3.56	14.93

Table 4: Results for domain adaptation from CNN/Daily Mail to XL-Sum on English (top) with artificially constrained data sizes, and to BookSum (bottom) with full data size.

Article: Lewis Williams, 20, died on 11 January from a shotgun wound suffered in Wath Road, Mexborough. South Yorkshire Police said two men aged 20 and 49 were arrested on Friday in connection with his death, bringing the total number of arrests to eight . . . **Gold Summary:** Two more people have been arrested in connection with a fatal shooting. **BART-FT Summary:** Two more people have been arrested in connection with the fatal shooting of a man in South Yorkshire. **BART-Adapt Summary:** *Eight* more people have been arrested in connection with the death of a man in South Yorkshire.
Article: BBC News Officials say the country's Olympic Committee will "oversee participation of women athletes who can qualify". The decision will end recent speculation as to whether the entire Saudi team could have been disqualified on grounds of gender discrimination . . . For the desert kingdom, the decision to allow women to compete in the Olympics is a huge step, overturning deep-rooted opposition from those opposed to any public role for women . . . **Gold Summary:** Saudi Arabia is to allow its women athletes to compete in the Olympics for the first time. **BART-FT Summary:** Saudi Arabia is to allow women to compete in next year's Olympic and Paralympic Games. **BART-Adapt Summary:** Saudi Arabia is to allow women to take part in the *2012 Winter* Olympics, officials say.
Article: The vehicle was seen at about 03:45 BST at the fast food giant's branch in Catterick, North Yorkshire. A 19-year-old man was arrested at the site, a short distance from the local golf club, on suspicion of theft and driving while unfit through drink. Police said it was the "most unusual job" of the night but officers managed to "avoid a high-speed pursuit" . . . **Gold Summary:** A stolen golf buggy was seized after being spotted at a McDonald's drive-thru. **BART-FT Summary:** A suspected stolen car was spotted at a McDonald's drive-thru. **BART-Adapt Summary:** A man has been arrested after a car was seen driving into a McDonald's branch.

Table 5: Examples of gold and generated summaries (from models trained on the full dataset) with their corresponding articles selected from the XL-Sum (English) dataset. Summary phrases italicized and highlighted in red denote hallucinations.

for BART-Adapt. We use a batch size of 4 on XL-Sum, and 8 on BookSum. All other hyperparameter settings are identical to those in the language adaptation experiment.

5.2 Results

We report the experiment results in Table 4. The pattern is that for medium to large CNN/Daily Mail data sizes, BART-FT outperforms BART-Adapt significantly. The two methods tie at around 300-3000 training sizes. BART-Adapt wins notably when there are only a handful of examples. This implies that adapters only stand out when the amount of data is extremely limited. In this case, we doubt the importance of training efficiency in adapters when the data size is so small. Instead, we argue that a potential benefit of using adapters is to reduce overfitting. As for BookSum, we can observe that numbers are very similar for both models with BART-FT slightly outperforming BART-Adapt. We argue adapters can do well in domain adaption despite the domain difference as long as there are sufficient training data. Finally, we notice the performance of fine-tuning only the last decoder layer is nowhere near BART-FT or BART-Adapt; this implies the practicability of adapters in summarization.

5.3 Qualitative analysis

To understand the quality of generated summaries between BART-FT and BART-Adapt, we examined a set of randomly selected model outputs from the XL-Sum dataset. We show some examples in Table 5. We find that summaries generated by the two models are roughly the same in terms of informativeness,

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Task	Data Size	Model	R1	R2	RL
DialogSum	12.5k	BART-FT	**47.40**	**24.66**	**39.03**
		BART-Adapt	47.24	24.57	38.56
SAMSum	14.7k	BART-FT	**49.52**	**24.91**	40.64
		BART-Adapt	49.38	24.69	**40.99**

Table 6: Results for task adaption from CNN/Daily Mail to DialogSum and SAMSum.

Task	Data Size	Model	R1	R2	RL
DialogSum	12.5k	BART-FT*	35.60	16.59	29.69
		BART-Adapt*	**36.35**	**17.03**	**30.25**
SAMSum	14.7k	BART-FT**	**40.91**	**14.82**	**32.32**
		BART-Adapt**	40.42	14.65	32.28

Table 7: Results for robustness analysis of task adaption experiments. Results are directly obtained by using the trained model from the other task without any further training. *denotes the model trained on SAMSum, and **denotes the model trained on DialogSum.

grammaticality, and fluency. Despite summaries being similar in these aspects, we find that BART-Adapt summaries are more prone to hallucinations, which is a well-known problem in abstractive summarization that summaries are not factual with respect to the source or general knowledge.

6 Task Transfer Experiments

6.1 Experimental setup

In previous settings, we conduct experiments with the fine-tuning paradigm on the subject of language and domain adaption. We also conduct experiments on task adaption to further verify our findings. In particular, we experiment with adapting a news summarization model to dialogue summarization. Dialogue summarization is often considered a much different task from monologic texts (e.g. news in our case) summarization due to its unique challenges. Chen et al. (2021) point out that: information flow is reflected in the dialogue discourse structures, summaries are required to be objective, and dialogue is acted at the pragmatic level. For these reasons, we choose to work with the DialogSum (Chen et al., 2021) and the SAMSum (Gliwa et al., 2019) datasets. We follow the previous setting and start with a BART model already fine-tuned on the CNN/Daily Mail dataset, then further train the model on these two datasets separately. We use a batch size of 8 for both DialogSum and SAMSum. All other hyperparameter settings are identical to those in the domain adaptation experiment.

6.2 Results

We report the experiment results in Table 6. We can observe that despite the dataset, BART-FT almost always beats BART-Adapt. However, we can notice that the performance gap is rather small, possibly due to the small dataset sizes. This is consistent with our earlier findings that adapters are on par with fine-tuning when the amount of training data is limited.

6.3 Model robustness

In addition to model performance, we also examine the robustness of models with either fine-tuning or adapters. In particular, we evaluate the model in a zero-shot manner where we directly test the DialogSum model on the SAMSum dataset, and vice versa. We present the results in Table 7. We can first observe that performance drops significantly compared to those in Table 6. Moreover, BART-Adapt has better performance than BART-FT on the DialogSum dataset, and it achieves very similar results on the SAMSum dataset. This suggests that adapters are more robust in a zero-shot setup with fewer data; the reason could be less overfitting introduced by a limited number of parameters in adapters.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

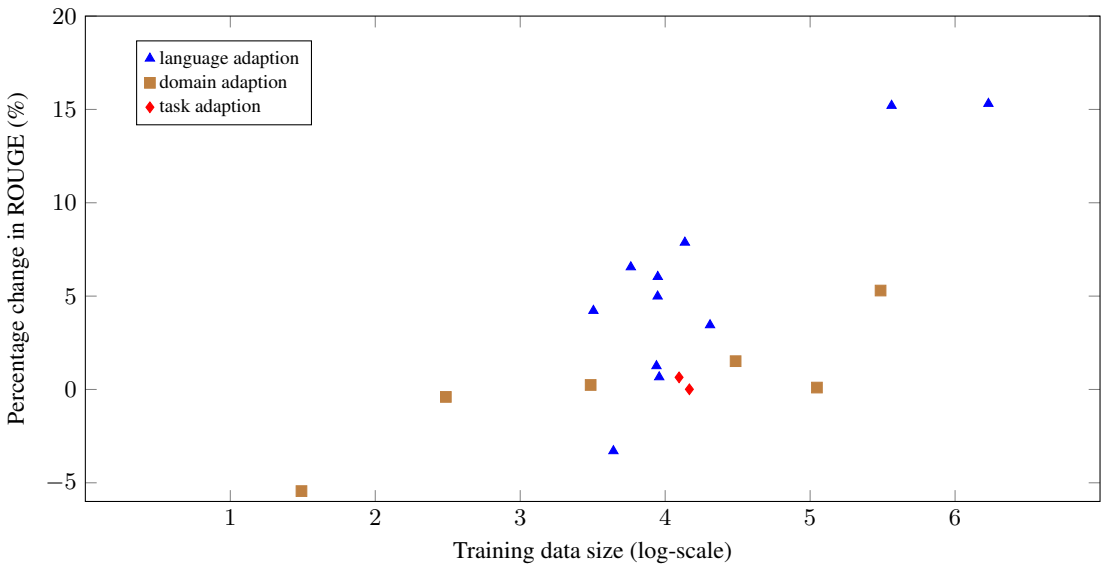


Figure 4: The effect of the training data size on ROUGE difference between the fine-tuning and adapter strategy. We display how much percent FT is better than using adapters. Note that data points from different tasks (with different shapes and colors) are not strictly comparable.

6.4 Effect of data availability on performance

Our results suggest that fine-tuning generally surpasses adapters under all three settings (language, domain, and task adaption). In addition, we observe that the amount of training data affects the performance gap between the two methods. To further validate this observation, we plot the percentage change in ROUGE performance (between those of fine-tuning and those of adapters) against the training size (log-scale) and we provide the visualization in Figure 4. We use the average number of ROUGE-1/2/L to represent the performance. From the plot, we can see that percentage change in ROUGE has an obvious positive relationship with the training data size which means that as the amount of training data increases, the performance gap between BART-FT and BART-Adapt increases as well. Looking at the tasks individually, we can see that for language adaption tasks with relatively small amounts of data, this trend is not very notable. The trend is most salient on domain adaption tasks since we manually controlled the data size for the experiment for adapting CNN/Daily Mail to XL-Sum.

7 Conclusions and Future Work

With large PLMs coming to light, we investigate fine-tuning and adapter strategies for transfer learning in abstractive summarization. We demonstrated that the performance gap between the two strategies is positively correlated with the availability of training resources, despite the languages being tested. Further analysis on domain adaptation and task adaption produces agreeing observations. We conclude that for realistically large summarization datasets, full fine-tuning will guarantee the best output quality. On the other hand, when resources are scarce, the advantages of adapters emerge in the niche market.

Most summarization datasets are web-crawled or machine-translated, resulting in non-optimal data quality. We plan to perform more qualitative analysis on the model outputs such as linguistic interpretation and human evaluation. In addition, we only experimented with fine-tuning and using adapters on mBART and BART for abstractive summarization, so there is room for research on other large PLMs, as well as other NLP tasks in the future.

Acknowledgements

We thank the reviewers of the paper for their feedback. Zheng Zhao is supported by the UKRI Centre for Doctoral Training in Natural Language Processing (grant EP/S022481/1). Pinzhen Chen is supported by a donation to Kenneth Heafield. This work does not necessarily reflect the opinion of the funders.

References

Arthur Brazinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. *Findings of the Association for Computational Linguistics: NAACL 2022*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. *34th Conference on Neural Information Processing Systems*.

Yi-Syuan Chen and Hong-Han Shuai. 2021. Meta-transfer learning for low-resource abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. DialogSum challenge: Summarizing real-life scenario dialogues. *Proceedings of the 14th International Conference on Natural Language Generation*.

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. *Proceedings of the 58th Annual Meeting of the 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. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. *Proceedings of the 2nd Workshop on New Frontiers in Summarization*.

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. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning*.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. *In Findings of the Association for Computational Linguistics: EMNLP 2020*.

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. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

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. *33rd Conference on Neural Information Processing Systems*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

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. *Journal of Machine Learning Research*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*.

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. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *Proceedings of the 37th International Conference on Machine Learning*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Yanyan Zou, Xingxing Zhang, Wei Lu, Furu Wei, and Ming Zhou. 2020. Pre-training for abstractive document summarization by reinstating source text. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

A Model Configurations

We tuned the hyperparameters using the validation set. We list the hyperparameters in Table 8, and highlight the selected ones in bold if multiple values are tried out. Instead of an expensive grid search on all combinations, we searched for the best configurations one by one. We performed a single run for each experiment.

Configuration	**Value**
training toolkit	PyTorch (Paszke et al., 2019)
stopping criterion	validation ROUGE
learning rate	1e-3, 5e-3, **1e-4** (mBART+Adapt), 5e-4, **1e-5** (mBART-FT), 5e-5
optimizer	Adam (Kingma and Ba, 2015)
beta1, beta2	0.9, 0.999
weight decay	1e-6
loss function	cross-entropy
decoding batch size	1
decoding beam size	5
decoding len. penalty	1.0
adapter reduction factor	1, **2**, 8, 16
trainable parameters	mBART-FT: 610M mBART-Adapt: 50M

Table 8: Model and training configurations.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 824-835, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

MRC-based Medical NER with Multi-task Learning and Multi-strategies

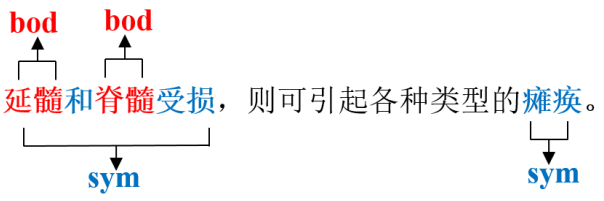
Xiaojing Du, Yuxiang Jia*, and Hongying Zan
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
zzu_dxj@163.com,{ieyxjia,iehyzan}@zzu.edu.cn

Abstract

Medical named entity recognition (NER), a fundamental task of medical information extraction, is crucial for medical knowledge graph construction, medical question answering, and automatic medical record analysis, etc. Compared with named entities (NEs) in general domain, medical named entities are usually more complex and prone to be nested. To cope with both flat NEs and nested NEs, we propose a MRC-based approach with multi-task learning and multi-strategies. NER can be treated as a sequence labeling (SL) task or a span boundary detection (SBD) task. We integrate MRC-CRF model for SL and MRC-Biaffine model for SBD into the multi-task learning architecture, and select the more efficient MRC-CRF as the final decoder. To further improve the model, we employ multi-strategies, including adaptive pre-training, adversarial training, and model stacking with cross validation. Experiments on both nested NER corpus CMeEE and flat NER corpus CCKS2019 show the effectiveness of the MRC-based model with multi-task learning and multi-strategies.

1 Introduction

With the fast development of medical digitalization, more and more medical documents are generated, including electronic medical records, medical reports, etc. Medical information extraction, including medical named entity recognition (NER), becomes increasingly important to applications like knowledge graph construction, question answering system, and automatic electronic medical record analysis. Medical NER is a task to automatically recognize medical named entities, like body (bod), disease, clinical symptom (sym), medical procedure, medical equipment, drug, medical examination item, etc., from medical texts. Medical named entities are usually long, nested and polysemous, which pose great challenges to medical NER. For example, in Fig 1, the two "bod" entities "延髓"(medulla oblongata) and "脊髓"(spinal cord) are nested in the "sym" entity "延髓和脊髓受损"(damage to the medulla oblongata and spinal cord).



Damage to the medulla oblongata and spinal cord can cause various types of paralysis.

Figure 1: An example with nested entity

To tackle both flat and nested NER, like (Li et al., 2020b), we take NER as a machine reading comprehension (MRC) problem. In addition, from different views, NER can be treated as a

*Corresponding author
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

sequence labeling (SL) task or a span boundary detection (SBD) task. We integrate MRC-CRF model for SL and MRC-Biaffine model for SBD into the multi-task learning (MTL) architecture. There is no nested NEs composed of entities of the same type in the datasets, so we select the more efficient MRC-CRF as the final decoder. To further improve the model, we employ multi-strategies (MS), including adaptive pre-training, adversarial training, and model stacking with cross validation. The main contributions of this paper are as follows:

- We improve MRC-CRF for medical NER with Biaffine through a multi-task learning architecture, which is a lighter way than traditional ensemble learning.
- We propose multi-strategies to improve the NER model, including adaptive pre-training, adversarial training, and model stacking with cross validation.
- Experimental results on both the nested NER corpus CMeEE (Zhang et al., 2022) and the flat NER corpus CCKS2019 (Han et al., 2020) show the effectiveness of the proposed model.

2 Related Work

Just like NER in other application domains, medical NER borrows methods from NER in general domain. The methods evolve from rule-based methods, traditional machine learning-based methods, deep learning-based methods, to the present mainstream pre-training-based methods.

Pre-trained models like BERT (Dai et al., 2019; Li et al., 2020a; Qin et al., 2021), ELMo (Li et al., 2019; Li et al., 2020c; Wan et al., 2020), etc., have become a standard module to encode the input texts. To better represent a text, RNN (Chowdhury et al., 2018), LSTM (Dai et al., 2019), GRU (Qin et al., 2021), CNN (Kong et al., 2021) and other neural networks are usually employed after the pre-trained model. Taking the NER as a sequence labeling problem, CRF (Lafferty et al., 2001) is finally used to generate the sequence labels.

For Chinese, characters (Li et al., 2020c), radicals, strokes (Li et al., 2019; Luo et al., 2020) and glyphs (Zhong and Yu, 2021) can provide useful information besides words. Thus such linguistic units are encoded together with words using LatticeLSTM (Zhao et al., 2019), ELMo (Li et al., 2019; Li et al., 2020c; Wan et al., 2020) and other networks. Domain data can be used to improve medical NER. (Liu et al., 2021) pre-train a Med-BERT based on medical texts to boost the performance significantly. (Chen et al., 2020) integrate domain dictionary and rules with Bi-LSTM-CRF.

Multi-task learning is another way to improve the performance. NER model can be enhanced by parameter sharing with models of other tasks. (Chowdhury et al., 2018) take NER and POS tagging as two tasks. (Luo et al., 2020) take NER on two different datasets as two tasks. To tackle nested NER problem and encode knowledge from entity types, NER is formulated as a machine reading comprehension task (Li et al., 2020b), and two binary classifiers are used to detect the span of a named entity. To enhance the information interaction between the head and tail of an entity, (Cao et al., 2021) introduce biaffine to MRC. (Zhu et al., 2021) ensemble sequence labeling and span boundary detection by voting strategies while (Zheng et al., 2021) ensemble CRF and MRC.

3 The MRC-MTL-MS model

MRC model extracts answer fragments from paragraphs by a given question. Suppose X is the input text, for each entity type y, designing a query q_y, extracting a subsequence x of type y from X, and we can get the triple (q_y, x, X), which is exactly the $(question, answer, context)$ a MRC model needs. The model only calculates the loss of context during training, and masks the loss of query and padding.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

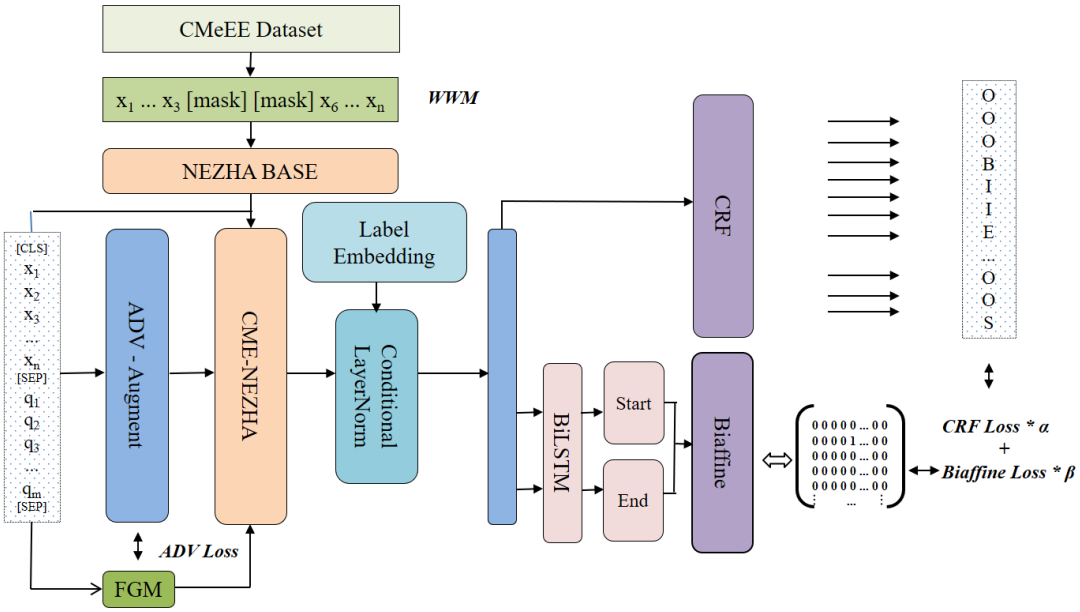


Figure 2: The architecture of the proposed NER model

3.1 Multi-task Learning (MTL)

The overall architecture of the model is shown in Fig 2. The multi-task learning architecture consists of the main task of sequence labeling by CRF and the auxiliary task of span boundary detection by Biaffine. For each entity type y, the input to the model is context X followed by query q_y, which is proved experimentally better than reversed concatenating way. The input is encoded by an adaptive pre-trained model CME-NEZHA, then goes through a Conditional LayerNorm guided by entity label embedding to further untilize entity type knowledge, and finally is decoded by CRF and Biaffine respectively.

3.1.1 Sequence Labeling with CRF

Suppose $h = (h_1, h_2, ..., h_N)$ is the encoded hidden layer sequence after Conditional LayerNorm, and $y = (y_1, y_2, ..., y_N)$ is the tag sequence, as shown in Fig 2. The score of sequence y is computed as follows,

$$s(h, y) = \sum_{n=1}^{N} \mathrm{W}_{n,y_n} + \sum_{n=2}^{N} T_{y_{n-1},y_n} \tag{1}$$

where W is the score matrix of each tag at each time step and T is the transition matrix between tags.

The probability of sequence y is calculated by softmax function, where $Y(h)$ represents all possible tag sequences.

$$p(y \mid h) = \frac{e^{s(h,y)}}{\sum_{\tilde{y} \in Y(h)} e^{s(h,\tilde{y})}} \tag{2}$$

The maximum likelihood loss function is used for training.

$$L_{\mathrm{CRF}} = \log(p(y \mid h)) \tag{3}$$

During inference, the predicted tag sequence with the maximum score is obtained with Viterbi algorithm.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$$y^* = \arg \max_{\widetilde{y} \in Y(h)} s(h, \widetilde{y}) \tag{4}$$

3.1.2 Span Boundary Detection with Biaffine

As shown in Fig 2, the hidden layer sequence after Conditional LayerNorm goes through a bidirectional LSTM and two seperate nonlinear layers to learn the representation of start and end of the span. Finally, the score of a span i is calculated by a Biaffine classifier as follows,

$$h_i^s = MLP_{\text{start}}\ (h_i) \tag{5}$$

$$h_i^e = MLP_{\text{end}}\ (h_i) \tag{6}$$

$$r(i) = h_i^{s^T} U h_i^e + W\left(h_i^s \oplus h_i^e\right) + b \tag{7}$$

Where U is a $N * C * N$ tensor, W is a $2N * C$ matrix, b is the bias, N is the length of the sentence, C is the number of entity categories $+1$(non-entity).

We assign span i a NER category y_i:

$$y_i = \arg\max r(i) \tag{8}$$

The learning objective of our named entity recognizer is to assign a correct category to each valid span. Hence it is a multi-class classification problem and we optimise the model with softmax cross-entropy:

$$p\left(i_c\right) = \frac{\exp\left(r\left(i_c\right)\right)}{\sum_{\hat{c}=1}^{C} \exp\left(r\left(i_{\hat{c}}\right)\right)} \tag{9}$$

$$L_{\text{Biaffine}} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log p\left(i_c\right) \tag{10}$$

3.1.3 The Combined Loss

The final loss function of the model is weighted by the loss function of CRF and the loss function of Biaffine, as shown below:

$$L = \alpha^* L_{\text{CRF}} + \beta^* L_{\text{Biaffine}} \tag{11}$$

Where α and β are positive real number and their sum equals 1. They can be learned and updated iteratively with the training and we initialize both of them as 0.5.

3.2 Multi-strategies (MS)

Three strategies are adopted to enhance the performance, including Adaptive Pre-training (AP), Adversarial Training (AT) and model stacking with Cross Validation (CV). In order to reduce distribution differences between the task data and data used by the pre-trained model, we use CMeEE data for task-adaptive pre-training (Gururangan et al., 2020) based on the pre-trained model NEZHA (Wei et al., 2019) with Whole Word Masking (WWM) strategy to get a new domain adaptive pre-trained model CME-NEZHA. In order to improve the robustness of the model, we employ adversarial training (Kurakin et al., 2016) with Fast Gradient Method (FGM) strategy. Lastly, 5-fold cross validation is adopted to prevent model overfitting and exploit advantages of multi-models. Five models are trained and contribute equally to the final decision.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Dataset	Training set	Validation set	Test set	Average sentence length
CMeEE	15000	5000	3000	>50 characters
CCKS2019	800	200	379	>390 characters

Table 1: Statistics of datasets

Entity Type	Entity number	Percent	Average entity length
bod	26589	28%	3.37
dis	24077	26%	5.35
sym	18579	20%	6.70
pro	9610	10%	5.30
dru	6331	7%	4.74
ite	4091	4%	4.37
mic	3019	3%	4.31
equ	1392	1%	4.30
dep	494	1%	2.86
Total	94182	100%	4.91
Anatomy	11520	49%	2.48
Disease	5535	23%	6.98
Drug	2307	10%	3.71
Laboratory	1785	8%	4.00
Image	1317	5%	3.79
Operation	1191	5%	12.85
Total	23655	100%	4.36

Table 2: Entity statistics of CMeEE and CCKS2019

4 Datasets

Two public datasets are used for experiments, CMeEE for nested NER and CCKS2019 for flat NER. Statistics of the two datasets are shown in Table 1, including sizes of the training, validate and test sets. As can be seen, the size of CMeEE is larger while the average text length of CCKS2019 is longer.

The texts of CMeEE are from textbooks of clinical pediatrics, which contain 9 types of entities, including Body (bod), Disease (dis), Symptom (sym), Medical procedure (pro), Medical equipment (equ), Drug (dru), Medical examination item (ite), Department (dep) and microorganism (mic). The texts of CCKS2019 are from electronic medical records, which contain 6 types of entities, including Disease and diagnosis, Image examination, Laboratory examination, Operation, Drug and Anatomy. As show in Table 2, the distributions of entities are imbalanced in both corpora. The top 3 dominant types of entities in CMeEE are bod, dis, and sym, while the top 3 dominant types of entities in CCKS2019 are Anatomy, Disease and Drug. On average, entities of sym and Operation are the longest in the two corpora respectively.

Flat entity	Nested entity	Percent of nested	Percent of nested in sym
84119	10063	10.68%	30.21%

Table 3: Nested entity statistics of CMeEE

As shown in Table 3, 10.68% of all entities in CMeEE are nested entities and 30.21% entities of sym are nested entities. Entities nested in sym entities are shown Table 4. All entity types except dep have entities nested in sym, where bod is the dominant type.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Entity type	Number	Percent	Example of nested entity
bod	4706	84.84%	{无色胶冻样 [痰]bod}sym {Colorless jelly like [sputum]bod}sym
ite	486	8.76%	{[胸片]ite 异常}sym {[Chest radiograph]ite Abnormal}sym
dis	229	4.13%	{逐步发生全身弛缓性 [瘫痪]dis}sym {Progressive generalized flaccid [paralysis]dis}sym
pro	59	1.06%	{[肺部听诊]pro 呼吸音减弱}sym {[Lung auscultation]pro respiratory sound is reduced}sym
dru	28	0.50%	{[维生素 A]dru 摄入不足}sym {[vitamin A]dru Insufficient intake}sym
mic	26	0.47%	{气道分泌物 [细菌]mic 培养阳性}sym {Airway secretion [bacteria]mic culture positive}sym
equ	13	0.23%	{长期 [呼吸机]equ 依赖}sym {Long-term [respirator]equ dependence}sym

Table 4: Entities nested in sym

5 Experiments

5.1 Query Generation

As shown in Table 5, for CMeEE, we put example entities into the query, while for CCKS2019, we take the description of the entity type as the query.

5.2 Experimental Settings

We retrain the pre-trained model NEZHA based on the CMeEE corpus by 100 epochs. Then we fine-tune the model for NER by 4 epochs. We set the batch size to 16, dropout to 0.1, NEZHA learning rate to 2.5e-5, other learning rate to 2.5e-3, and maximum text length to 256. NVIDIA GTX2080Ti is used to run the program. Micro average F1 is chosen as the evaluation metric.

5.3 Comparison with Previous Models

5.3.1 Baselines on CMeEE Corpus

(1) MacBERT-large and Human are from (Zhang et al., 2022). MacBERT is variant of BERT, taking a MLM (Masked Language Model) as correction strategy. Human denotes the annotating result of human. (2) BERT-CRF, BERT-Biaffine and RICON are from (Gu et al., 2022). BERT-CRF solves sequence labeling with CRF, BERT-Biaffine detects span boundary with Biaffine, and RICON learns regularity inside entities. (3) Lattice-LSTM, Lattice-LSTM+Med-BERT, FLAT-Lattice, Medical-NER, and Medical NER+Med-BERT are from (Liu et al., 2021). Lattice-LSTM, Lattice-LSTM+Med-BERT and FLAT-Lattice incorporate lexicon to decide entity boundary. Medical NER and Medical NER+Med-BERT introduce big dictionary and pre-trained domain model.

5.3.2 Baselines on CCKS2019 Corpus

(1)BERT-BiLSTM-CRF is from (Dai et al., 2019), taking CRF for sequence labeling. (2)BBC+Lexicon+Glyph is from (Zhong and Yu, 2021), introducing lexicon and glyph information. (3) WB-Transformer+SA is from (Zhang et al., 2021), taking self-attention for semantic enrichment. (4) ELMo-lattice-LSTM-CRF is from (Li et al., 2020c), fusing ELMo and lexicon to improve sequence labeling performance. (5) ACNN is from (Kong et al., 2021), composed of hierarchical CNN and attention mechanism. (6) FS-TL is from (Li et al., 2019), fusing stroke information with transfer learning.

CCL 2022

Entity type	Query
bod	在文本中找出身体部位，例如细胞、皮肤、抗体 Find body parts in the text, for example, cells, skin and antibodies
dep	在文本中找出科室，例如科、室 Find departments in the text, for example, department and room
dis	在文本中找出疾病，例如癌症、病变、炎症、增生、肿瘤 Find diseases in the text, for example, cancer and pathological changes
dru	在文本中找出药物，例如胶囊、疫苗、剂 Find drugs in the text, for example, capsule, vaccine and agent
equ	在文本中找出医疗设备，例如装置、器、导管 Find medical devices in the text, for example, device and conduit
ite	在文本中找出医学检验项目，例如尿常规、血常规 Find medical test items in the text, for example, urine routine and blood routine
mic	在文本中找出微生物，例如病毒、病原体、抗原、核糖 Find micro organisms in the text, for example, virus and pathogen
pro	在文本中找出医疗程序，例如心电图、病理切片、检测 Find medical procedures in the text, for example, electrocardiogram and pathological section
sym	在文本中找出临床表现，例如疼痛、痉挛、异常 Find clinical manifestations in the text, for example, pain and spasm
Anatomy	找出疾病、症状和体征发生的人体解剖学部位 Find where in the human anatomy the disease, symptoms and signs occur
Disease	找出医学上定义的疾病和医生在临床工作中对病因、病生理、分型分期等所作的判断 Find medically defined diseases and physicians' judgments regarding etiology, pathophysiology, staging, etc., in clinical work-up
Drug	找出用于疾病治疗的具体化学物质 Find specific chemicals for disease treatment
Image	找出影像检查（X 线、CT、MR、PETCT 等）+ 造影 + 超声 + 心电图 Find imaging examinations (X-ray, CT, Mr, PETCT, etc.) + contrast + ultrasound + ECG
Laboratory	找出在实验室进行的物理或化学检查 Find physical or chemical examinations performed in the laboratory
Operation	找出医生在患者身体局部进行的切除、缝合等治疗，是外科的主要治疗方法 Find the main treatment in surgery that doctors perform locally on the patient's body, such as excision, suture, etc.

Table 5: Query for different entity types in CMeEE and CCKS2019

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Precision/%	Recall/%	F1 score/%
MacBERT-large(Zhang et al., 2022)	-	-	62.40
Human(Zhang et al., 2022)	-	-	67.00
BERT-CRF(Gu et al., 2022)	58.34	64.08	61.07
BERT-Biaffine(Gu et al., 2022)	64.17	61.29	62.29
RICON(Gu et al., 2022)	66.25	64.89	65.57
Lattice-LSTM(Liu et al., 2021)	57.10	43.60	49.44
Lattice-LSTM+Med-BERT(Liu et al., 2021)	56.84	47.58	51.80
FLAT-Lattice(Liu et al., 2021)	66.90	70.10	68.46
Medical NER(Liu et al., 2021)	66.41	70.73	68.50
Medical NER+Med-BERT(Liu et al., 2021)	**67.99**	70.81	69.37
MRC-MTL-MS(Ours)	67.21	**71.89**	**69.47**

Table 6: Comparison with previous models on CMeEE

Model	Precision/%	Recall/%	F1 score/%
BERT-BiLSTM-CRF(Dai et al., 2019)	73.84	75.31	74.53
BBC+Lexicon+Glyph(Zhong and Yu, 2021)	85.17	84.13	84.64
WB-Transformer+SA(Zhang et al., 2021)	-	-	84.98
ACNN(Kong et al., 2021)	83.07	**87.29**	85.13
FS-TL(Li et al., 2019)	-	-	85.16
ELMo-lattice-LSTM-CRF(Li et al., 2020c)	84.69	85.35	85.02
MRC-MTL-MS(Ours)	**85.29**	85.32	**85.31**

Table 7: Comparison with previous models on CCKS2019

As shown in Table 6 and 7, our MRC-MTL-MS model outperforms all comparison models on both the nested NER corpus CMeEE and the flat NER corpus CCKS2019.

5.4 Ablation Experiments

The ablation experiments are shown in Table 8. MRC-Base is the same with (Li et al., 2020b), pointer network is used to detect span boundary. MRC-CRF only uses CRF for decoding. MRC-Biaffine only uses Biaffine for decoding. MRC-MTL integrates CRF and Biaffine with multi-task learning and use CRF as the final decoder. We can see that multi-task learning model outperforms single-task models. Adaptive Pre-training (AP), Adversarial Training (AT), and model stacking with Cross Validation (CV) strategies further improve the performance. Among which, CV contributes the most. Compared with MRC-Base, the improvement of F1 score on the nested NER corpus is 2.56%, which is higher than that of 1.63% on the flat NER corpus.

5.5 Experiments on Different Types of NEs

Experimental results of different types of NEs on the two corpora are shown in Table 9 respectively. As can be seen, on CMeEE, the entity type dru has the highest F1 score 81.17%, while the entity type ite has the lowest F1 score. The averagely longest and most nested entity type sym also has low F1 score and needs further study. The overall F1 scores on CCKS2019 are high and the entity type Drug also has the highest F1 score 95.25%, indicating that Drug entities are easier to recognize. For those entity types with low scores, like ite and Laboratory, constructing related lexicons maybe useful for improvement.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	CMeEE/%			CCKS2019/%		
	Precision	Recall	F1 score	Precision	Recall	F1 score
MRC-Base	67.98	65.87	66.91	82.63	84.76	83.68
MRC-CRF	67.17	67.25	67.21	84.40	84.91	84.65
MRC-Biaffine	**70.71**	64.09	67.24	83.22	83.77	83.49
MRC-MTL	64.58	71.76	67.98	84.42	84.97	84.70
+AP	66.28	70.34	68.25	84.23	85.24	84.73
+AP+AT	68.04	69.16	68.59	84.20	**85.39**	84.79
+AP+AT+CV	67.21	**71.89**	**69.47**	**85.29**	85.32	**85.31**

Table 8: Ablation experiments on CMeEE and CCKS2019

CCL 2022

Entity type	Precision/%	Recall/%	F1 score/%
bod	62.92	71.33	66.86
dis	76.78	80.69	78.69
dru	75.38	87.93	81.17
dep	54.24	88.89	67.37
equ	74.48	81.20	77.70
ite	51.06	49.23	50.13
mic	76.64	82.16	79.30
pro	61.91	71.50	66.36
sym	58.49	54.68	56.52
Mac-Avg	65.77	74.18	69.72
Anatomy	85.25	87.07	86.15
Disease	85.63	85.56	85.60
Drug	95.45	95.05	95.25
Image	86.65	87.64	87.14
Laboratory	74.54	67.97	71.10
Operation	85.91	79.01	82.32
Mac-Avg	85.57	83.72	84.63

Table 9: Results of different types of NEs on CMeEE and CCKS2019

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

5.6 Case Study

Table 10 gives two examples from CMeEE. In the first example, the MRC-Base model does not correctly detect the boundary of the entity "郎飞结上的补体被激活" (Complement on Ranvier knot is activated), while the MRC-MTL-MS model correctly recognizes the boundary and the entity type. In the second example, the MRC-Base model correctly detects the boundary of the entity "高血压" (hypertension), but predicts a wrong label. The MRC-MTL-MS model correctly recognizes the polysemous entity, indicating its superiority in disambiguating polysemous entities.

Sentence	AMAN 的一个早期表现就是郎飞结上的补体被激活。 An early manifestation of AMAN is that complement on Ranvier knot is activated.
Entity	郎飞结上的补体被激活 Complement on Ranvier knot is activated.
Golden Labels	B-SYM I-SYM I-SYM I-SYM I-SYM I-SYM I-SYM I-SYM I-SYM E-SYM
MRC	B-BOD I-BOD E-BOD O O O O O O O
MRC-MTL-MS	B-SYM I-SYM I-SYM I-SYM I-SYM I-SYM I-SYM I-SYM I-SYM E-SYM
Sentence	患儿情况好，只 1 例发生慢性排异及高血压。 The condition of the child is good, and only one develops chronic rejection and hypertension.
Entity	高血压 hypertension
Golden Labels	B-SYM I-SYM E-SYM
MRC	B-DIS I-DIS E-DIS
MRC-MTL-MS	B-SYM I-SYM I-SYM

Table 10: Two cases with labels BIES

6 Conclusion

This paper proposes a MRC-based multi-task model for Chinese medical NER, enhancing MRC-CRF with Biaffine to recognize the named entities more accurately. To further improve the model, we introduce multi-strategies, including adaptive pre-training, adversarial training and model stacking with cross validation. Our model can cope with both flat NER and nested NER. Experiments on the nested NER corpus CMeEE and the flat NER corpus CCKS2019 show the effectiveness of our model. In the future, we will incorporate domain knowledge to improve the recognition performance on hard named entities.

7 Acknowledgements

We would like to thank the anonymous reviewers for their insightful and valuable comments. This work was supported in part by Major Program of National Social Science Foundation of China (Grant No.17ZDA318, 18ZDA295), National Natural Science Foundation of China (Grant No.62006211), and China Postdoctoral Science Foundation (Grant No.2019TQ0286, 2020M682349).

References

Jun Cao, Xian Zhou, Wangping Xiong, Ming Yang, Jianqiang Du, Yanyun Yang, and Tianci Li. 2021. Electronic medical record entity recognition via machine reading comprehension and biaffine. *Discrete Dynamics in Nature and Society*, 2021.

Xianglong Chen, Chunping Ouyang, Yongbin Liu, and Yi Bu. 2020. Improving the named entity recognition of chinese electronic medical records by combining domain dictionary and rules. *International Journal of Environmental Research and Public Health*, 17(8):2687.

Shanta Chowdhury, Xishuang Dong, Lijun Qian, Xiangfang Li, Yi Guan, Jinfeng Yang, and Qiubin Yu. 2018. A multitask bi-directional rnn model for named entity recognition on chinese electronic medical records. *BMC bioinformatics*, 19(17):75–84.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)*, pages 1–5.

Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Yi Zheng, Baoxing Huai, and Nicholas Jing Yuan. 2022. Delving deep into regularity: A simple but effective method for chinese named entity recognition. *arXiv preprint arXiv:2204.05544*.

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.

Xianpei Han, Zhichun Wang, Jiangtao Zhang, Qinghua Wen, and Yinnian Lin. 2020. Overview of the ccks 2019 knowledge graph evaluation track: entity, relation, event and qa. *arXiv preprint arXiv:2003.03875*.

Jun Kong, Leixin Zhang, Min Jiang, and Tianshan Liu. 2021. Incorporating multi-level cnn and attention mechanism for chinese clinical named entity recognition. *Journal of Biomedical Informatics*, 116:103737.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Nan Li, Ling Luo, Zeyuan Ding, Yawen Song, Zhihao Yang, and Hongfei Lin. 2019. Dutir at the ccks-2019 task1: improving chinese clinical named entity recognition using stroke elmo and transfer learning. In *Proceedings of the 4th China Conference on Knowledge Graph and Semantic Computing (CCKS 2019)*, pages 24–27.

Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. 2020a. Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of biomedical informatics*, 107:103422.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020b. A unified mrc framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859.

Yongbin Li, Xiaohua Wang, Linhu Hui, Liping Zou, Hongjin Li, Luo Xu, Weihai Liu, et al. 2020c. Chinese clinical named entity recognition in electronic medical records: Development of a lattice long short-term memory model with contextualized character representations. *JMIR Medical Informatics*, 8(9):e19848.

Ning Liu, Qian Hu, Huayun Xu, Xing Xu, and Mengxin Chen. 2021. Med-bert: A pre-training framework for medical records named entity recognition. *IEEE Transactions on Industrial Informatics*.

Ling Luo, Zhihao Yang, Yawen Song, Nan Li, and Hongfei Lin. 2020. Chinese clinical named entity recognition based on stroke elmo and multi-task learning. *Chinese Journal of Computers*, 43(10):1943–1957.

Qiuli Qin, Shuang Zhao, and Chunmei Liu. 2021. A bert-bigru-crf model for entity recognition of chinese electronic medical records. *Complexity*, 2021.

Qian Wan, Jie Liu, Luona Wei, and Bin Ji. 2020. A self-attention based neural architecture for chinese medical named entity recognition. *Mathematical Biosciences and Engineering*, 17(4):3498–3511.

CCL 2022

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*.

Zhichang Zhang, Xiaohui Qin, Yanlong Qiu, and Dan Liu. 2021. Well-behaved transformer for chinese medical ner. In *2021 3rd International Conference on Natural Language Processing (ICNLP)*, pages 162–167.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2022. Cblue: A chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915.

Shan Zhao, Zhiping Cai, Haiwen Chen, Ye Wang, Fang Liu, and Anfeng Liu. 2019. Adversarial training based lattice lstm for chinese clinical named entity recognition. *Journal of biomedical informatics*, 99:103290.

Hengyi Zheng, Bin Qin, and Ming Xu. 2021. Chinese medical named entity recognition using crf-mt-adapt and ner-mrc. In *2021 2nd International Conference on Computing and Data Science (CDS)*, pages 362–365.

Shanhao Zhong and Qingsong Yu. 2021. Improving chinese medical named entity recognition using glyph and lexicon. In *Proceedings of 2021 International Conference on Advanced Education and Information Management (AEIM 2021)*, pages 75–80.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. Hitsz-hlt at semeval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 836-847, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

A Multi-Gate Encoder for Joint Entity and Relation Extraction

Xiong Xiong[1,2], **Yunfei Liu**[1,2], **Anqi Liu**[1], **Shuai Gong**[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,liuyunfei,liuaq,gongshuai19,shyli}@csu.ac.cn

Abstract

Named entity recognition and relation extraction are core sub-tasks of relational triple extraction. Recent studies have used parameter sharing or joint decoding to create interaction between these two tasks. However, ensuring the specificity of task-specific traits while the two tasks interact properly is a huge difficulty. We propose a multi-gate encoder that models bidirectional task interaction while keeping sufficient feature specificity based on gating mechanism in this paper. Precisely, we design two types of independent gates: task gates to generate task-specific features and interaction gates to generate instructive features to guide the opposite task. Our experiments show that our method increases the state-of-the-art (SOTA) relation F1 scores on ACE04, ACE05 and SciERC datasets to 63.8% (+1.3%), 68.2% (+1.4%), 39.4% (+1.0%), respectively, with higher inference speed over previous SOTA model.

1 Introduction

Extracting relational facts from unstructured texts is a fundamental task in information extraction. This task can be decomposed into two sub-tasks: Named Entity Recognition (NER) (Florian et al., 2003), which aims to recognize the boundaries and types of entities; and Relation Extraction (RE) (Zelenko et al., 2002), which aims to extract semantic relations between entities. The extracted relational triples in the form of (subject, relation, object) are basic elements of large-scale knowledge graphs (Lin et al., 2015).

Traditional approaches perform NER and RE in a pipelined fashion (Zhou et al., 2005; Chan and Roth, 2011; Gormley et al., 2015). They first extract all the entities in a given text, and then identify pairwise relations between the extracted entities. However, because the two sub-tasks are modeled independently, pipelined methods are vulnerable to error propagation issue. Since the interaction between NER and RE is neglected, the errors accumulated in the previous NER stage cannot be corrected in the subsequent RE stage. To resolve this issue, some joint models have been proposed to model these two tasks simultaneously. Early feature-based joint models (Yu and Lam, 2010; Miwa and Sasaki, 2014) rely on complicated feature engineering to build interaction between entities and relations. More recently, neural joint models have attracted increasing research interest and have demonstrated promising performance on joint entity and relation extraction.

In existing neural joint models, there are mainly two ways to build the interaction between NER and RE: parameter sharing and joint decoding. In parameter sharing methods (Zeng et al., 2018; Bekoulis et al., 2018a; Dixit and Al-Onaizan, 2019), NER model and RE model are built on top of a shared encoding layer to achieve joint learning. However, approaches based on parameter sharing implicitly rather than explicitly model the inter-task interaction, leading to insufficient excavation of the inherent association between the two tasks. Moreover, these two tasks focus on different contextual information (Zhong and Chen, 2021; Ye et al., 2022), but methods of sharing representations cannot provide task-specific features with enough specificity for the two tasks. In terms of error propagation, parameter sharing methods alleviate the error propagation between tasks, but to a limited extent, because these models still perform pipelined decoding.

*Corresponding author.
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Another family of approaches adopt unified tagging framework in the form of sequences (Zheng et al., 2017), tables (Zhang et al., 2017; Ren et al., 2021), or graphs (Fu et al., 2019; Xue et al., 2021) to integrate the information of entities and relations as a whole and perform joint decoding to extract relational triples. Although these methods enhance the inter-task interaction, the specificity of task features is not well considered since the entities and relations still share contextual representations in essence. Moreover, all these joint decoding methods require complex joint decoding algorithms, and it is challenging to balance the accuracy of joint decoding and the abundance of task-specific features.

Accordingly, the main challenge of joint entity and relation extraction is to construct proper interaction between NER and RE while ensuring the specificity of task-specific features. Wang and Lu (2020) adopt two types of representations to generate task-specific representations, sequence representations for NER and table representations for RE, separately. These two types of representations interact with each other to model inter-task interaction. Yan et al. (2021) perform neuron partition in an autoregressive manner to generate task-specific features jointly in order to build inter-task interaction. They combine the task-specific features and global features as the final input to the task modules. Inspired by Yan et al. (2021)'s work, we adopt the task modules they used that model each relation separately with tables (Miwa and Sasaki, 2014), and we propose a simple but effective feature encoding approach for joint entity and relation extraction, achieving excellent results while being less computationally intensive. We will detail the differences and our advantages in Section 3.5.

In this work, we propose a **M**ulti-**G**ate **E**ncoder (**MGE**) that control the flow of feature information based on gating mechanism, so as to filter out undesired information and retain desired information. MGE has two types of gates: task gates and interaction gates. Task gates are used to generate task-specific features, and interaction gates control how much information flows out to guide the opposite task. The output of interaction gate is combined with the opposite task-specific features to generate the input of corresponding task module, resulting in a bidirectional interaction between NER and RE while maintaining sufficient specificity of task-specific features.

The main contributions of this work are summarized below:

1. A multi-gate encoder for joint entity and relation extraction is proposed, which effectively promotes interaction between NER and RE while ensuring the specificity of task features. Experimental results show that our method establishes the new state-of-the-art on three standard benchmarks, namely ACE04, ACE05, and SciERC.
2. We conduct extensive analyses to investigate the superiority of our model and validate the effectiveness of each component of our model.
3. The effect of relation information on entity recognition is examined. Our additional experiments suggest that relation information contributes to predicting entities, which helps clarify the controversy on the effect of relation signals.

2 Related Work

The task of extracting relational triples from plain text can be decomposed into two sub-task: Named Entity Recognition and Relation Extraction. The two tasks can be performed in a pipelined manner (Chan and Roth, 2011; Gormley et al., 2015; Zhong and Chen, 2021; Ye et al., 2022) or in a joint manner (Miwa and Sasaki, 2014; Zheng et al., 2017; Wang and Lu, 2020; Yan et al., 2021).

Traditional pipelined methods (Zhou et al., 2005; Chan and Roth, 2011; Gormley et al., 2015) firstly train a model to extract entities and then train another model to classify the relation type between subject and object for each entity pair. Recent pipelined approaches (Zhong and Chen, 2021; Ye et al., 2022) still follow this pattern and adopt marker-based span representations to learn different contextual representations between entities and relations, and between entity pairs, which sheds some light on the importance of feature specificity. Although Zhong and Chen (2021) and Ye et al. (2022) achieve better performance than previous pipelined methods and some joint methods, they still run the risk of error propagation and do not adequately account for interactions between tasks. To ease these issues, some joint models that extract entities and relations jointly has been proposed.

Joint entity and relation extraction is a typical multi-task scenario, and how to handle the interaction

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

between tasks is a frequently discussed topic. Early joint models (Yu and Lam, 2010; Miwa and Sasaki, 2014) rely on feature engineering to build task interaction. More recently, many neural joint models have been proposed and show promising performance. Miwa and Bansal (2016) builds a sequence tagging model for NER and a dependency tree model for RE separately on top of a shared LSTM layer and performs joint learning, achieving task interaction through parameter sharing. Zeng et al. (2018) uses sequence-to-sequence learning framework with copy mechanism to jointly extract entities and relations. Bekoulis et al. (2018b) builds a CRF layer for NER and a sigmoid layer for RE on a shared LSTM layer. Eberts and Ulges (2020) proposes a span-based joint model for entity and relation extraction. They performs span classification and span filtering to extract entity spans and then performs relation classification based on the contextual span representations from BERT (Devlin et al., 2019) encoder. All these approaches construct the interaction between NER and RE through parameter sharing. Another class of methods adopts joint decoding to fuse the two tasks together. Li and Ji (2014) uses structured perceptron with beam search to extract entities and relations simultaneously. Wang et al. (2018) proposes a transition system to convert the joint task into a directed graph. Wang et al. (2020b) introduces a novel handshaking tagging scheme to formulate joint extraction as a token pair linking problem. Zhang et al. (2017) and Ren et al. (2021) convert the task into a table-filling problem.

In addition to building interaction between tasks, another important issue is the specificity of task features. As recent studies (Zhong and Chen, 2021; Ye et al., 2022) have shown, generating specific contextual features for different tasks can achieve better results on the overall task than sharing input features. Zhong and Chen (2021) and Ye et al. (2022) both use a pre-trained language model (e.g., BERT) for NER and another for RE to obtain different contextual representations for specific task. However, fine-tuning distinct pre-trained encoders for the two task separately is computationally expensive. In our work, we adopts gating mechanism to balance the flow of feature information, taking into account both the interaction between tasks and the specificity of task features.

3 Method

In this section, we first formally define the problem of joint entity and relation extraction and then detail the structure of our model. Finally, we discuss how our model differs from the approach we follow and explain why our method performs better.

3.1 Problem Definition

The problem of joint entity and relation extraction can be decomposed into two sub-tasks: NER and RE. Let $\mathcal{E}$ denotes the set of predefined entity types and $\mathcal{R}$ denotes the set of predefined relation types. Given a sentence containing N words, $X = \{x_1, x_2, \ldots, x_N\}$, the goal of NER is to extract an entity type $e_{ij} \in \mathcal{E}$ for each span $s_{ij} \in S$ that starts with x_i and ends with x_j, where S is the set of all the possible spans in X. For RE, the goal is to extract a relation type $r_{i_1 i_2} \in \mathcal{R}$ for each span pair whose start words are x_{i_1} and x_{i_2} respectively. Combining the results of NER and RE, we get the final output of this problem $Y_r = \{(e_{i_1 j_1}, r_{i_1 i_2}, e_{i_2 j_2})\}$, where $e_{i_1 j_1}, e_{i_2 j_2} \in \mathcal{E}, r_{i_1 i_2} \in \mathcal{R}$.

3.2 Multi-Gate Encoder

We adopt BERT (Devlin et al., 2019) to encode the contextual information of input sentences. As shown in Figure 1, our proposed MGE employs four gates to control the flow of feature information based on gating mechanism. The two task gates are designed to generate task-specific features for NER and RE, while the two interaction gates aim to generate interaction features that have a positive effect on the opposite task. The task-specific features and interaction features are combined to form the input of task modules, carrying out bidirectional task interaction through feature exchange.

Let $H_b \in \mathbb{R}^{N \times d}$ denotes the contextual feature matrix of sentence X extracted by BERT encoder, where d is the hidden size of BERT layer. In order to preliminarily build the specificity between entity recognition features and relation extraction features, we generate candidate entity features H_e^c and candidate relation

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

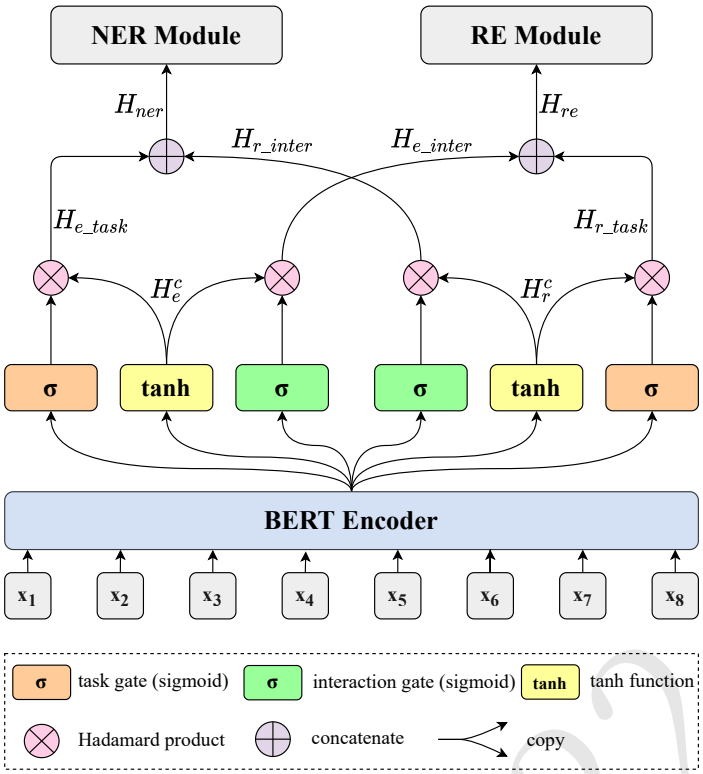


Figure 1: The architecture of our proposed MGE. There are two types of gates in the encoder: task gates and interaction gates. H_e^c and H_r^c denote candidate entity features and candidate relation features respectively. H_{e_task} and H_{r_task} denote task-specific features generated by task gates. H_{e_inter} and H_{r_inter} denote interaction features generated by interaction gates to guide the opposite task. H_{ner} and H_{re} are the final input features to NER module and RE module.

features H_r^c based on BERT output representations as follows:

$$\begin{aligned} H_e^c &= \tanh\left(H_b W_e + b_e\right) \\ H_r^c &= \tanh\left(H_b W_r + b_r\right), \end{aligned} \tag{1}$$

where $W_{(\cdot)} \in \mathbb{R}^{d \times h}$ and $b_{(\cdot)} \in \mathbb{R}^h$ denote trainable weights and bias and h is the hidden size in MGE. $\tanh(\cdot)$ means $tanh$ activation function. The candidate features will be input to the task gates and interaction gates of corresponding task for further feature filtering to generate task-specific features and interaction features.

The task gates decide what information in the candidate features is contributing to the corresponding specific task, which is implemented by a sigmoid layer. The sigmoid layer produces values in the range of zero to one, indicating how much information is to be transmitted. A value of zero means no information is allowed to pass, whereas a value of one means all the information is allowed to pass. We calculate entity task gate G_{e_task} and relation task gate G_{r_task} as below:

$$\begin{aligned} G_{e\_task} &= \sigma\left(H_b W_{e\_task} + b_{e\_task}\right) \\ G_{r\_task} &= \sigma\left(H_b W_{r\_task} + b_{r\_task}\right), \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ represents sigmoid activation function. $W_{(\cdot)} \in \mathbb{R}^{d \times h}$ and $b_{(\cdot)} \in \mathbb{R}^h$ denote weights and bias. The entity task gate G_{e_task} and relation task gate G_{r_task} work independently and are specialized in filtering information useful for specific task in candidate features to obtain task-specific features for entity recognition and relation extraction respectively. We calculate the Hadamard (element-wise) product between task gates and candidate features to generate task-specific features for NER and RE:

$$\begin{aligned} H_{e\_task} &= G_{e\_task} \odot H_e^c \\ H_{r\_task} &= G_{r\_task} \odot H_r^c, \end{aligned} \tag{3}$$

CCL 2022

where $\odot$ denotes Hadamard product operation. H_{e_task} and H_{r_task} represent entity task-specific features and relation task-specific features respectively.

Similarly, the interaction gates decide what information in entity candidate features H_e^c is helpful for guiding relation extraction and what information in H_r^c is helpful for guiding entity recognition. This is also implemented through sigmoid activation function:

$$\begin{aligned} G_{e\_inter} &= \sigma\left(H_b W_{e\_inter} + b_{e\_inter}\right) \\ G_{r\_inter} &= \sigma\left(H_b W_{r\_inter} + b_{r\_inter}\right), \end{aligned} \tag{4}$$

where G_{e_inter} denotes entity interaction gate and G_{r_inter} denotes relation interaction gate. $W_{(\cdot)} \in \mathbb{R}^{d\times h}$ and $b_{(\cdot)} \in \mathbb{R}^h$ denote weights and bias. These two interaction gates are then applied to candidate features to generate interaction features:

$$\begin{aligned} H_{e\_inter} &= G_{e\_inter} \odot H_e^c \\ H_{r\_inter} &= G_{r\_inter} \odot H_r^c, \end{aligned} \tag{5}$$

where H_{e_inter} denotes entity interaction features used to guide RE and H_{r_inter} denotes relation interaction features used to guide NER.

Finally, we perform feature exchange based on the task-specific features and interaction features to achieve bidirectional interaction between NER and RE. Specifically, we concatenate entity task-specific features H_{e_task} and relation interaction features H_{r_inter}, and relation task-specific features H_{r_task} is concatenated with entity interaction features H_{e_inter}:

$$\begin{aligned} H_{ner} &= H_{e\_task} \oplus H_{r\_inter} \\ H_{re} &= H_{r\_task} \oplus H_{e\_inter}, \end{aligned} \tag{6}$$

where $\oplus$ means concatenation operation. $H_{ner} \in \mathbb{R}^{N\times 2h}$ and $H_{re} \in \mathbb{R}^{N\times 2h}$ are the final features to be input to NER and RE task modules respectively. Through exchanging features that are designed to guide the opposite task and combining task-specific features, H_{ner} and H_{re} balance the task interaction and feature specificity of NER and RE.

3.3 Table-filling Modules

Following Yan et al. (2021), we adopt table-filling framework to extract entities and relations, which treats both NER and RE as table filling problems. For NER, the goal is to predict all the entity spans and corresponding entity types. Specifically, we construct a $N \times N$ type-specific table for each entity type $k \in \mathcal{E}$, whose element at row i and column j represents the probability of span $s_{ij} \in S$ belonging to type k. We firstly concatenate the representations of every two tokens based on H_{ner} and connect a fully-connected layer to reduce the hidden size. Then we employ layer normalization (Ba et al., 2016) and ELU activation (Clevert et al., 2015) to obtain table representations of spans. Formally, for span s_{ij} that starts with x_i and ends with x_j, we compute the table representation $T_{ner}^{i,j} \in \mathbb{R}^h$ as follows:

$$T_{ner}^{i,j} = \mathrm{ELU}(\mathrm{LayerNorm}([H_{ner}^i; H_{ner}^j]W_e^h + b_e^h)), \tag{7}$$

where $H_{ner}^i \in \mathbb{R}^{2h}$ and $H_{ner}^j \in \mathbb{R}^{2h}$ denote the vectors corresponding to words x_i and x_j in entity features $H_{ner} \in \mathbb{R}^{N\times 2h}$ that containing both entity task-specific information and relation interaction information. $W_e^h \in \mathbb{R}^{4h\times h}$ and $b_e^h \in \mathbb{R}^h$ are trainable parameters. To predict the probability of span s_{ij} belonging to entity type k, we project the hidden size to $|\mathcal{E}|$ with a fully-connected layer followed by a sigmoid activation function:

$$p(e_{ij} = k) = \sigma(T_{ner}^{i,j} W_e^{tag} + b_e^{tag}), \forall k \in \mathcal{E}, \tag{8}$$

where $W_e^{tag} \in \mathbb{R}^{h\times|\mathcal{E}|}$ and $b_e^{tag} \in \mathbb{R}^{|\mathcal{E}|}$ are trainable parameters and $|\mathcal{E}|$ represents the number of predefined entity types.

The goal of RE table-filling module is to predict the start word of each entity and classify the relations between them. The structure of RE module is formally analogous to the NER module. Similar to NER,

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

we construct a $N \times N$ type-specific table for each relation type $l \in \mathcal{R}$. For the table corresponding to relation l, the element at row i and column j represents the probability that the i-th word x_i and the j-th word x_j in a sentence are respectively the start words of subject entity and object entity of relation type l. For x_i and x_j, we compute the table representations $T_{re}^{i,j} \in \mathbb{R}^h$ as follows:

$$T_{re}^{i,j} = \text{ELU}(\text{LayerNorm}([H_{re}^i; H_{re}^j]W_r^h + b_r^h)), \tag{9}$$

where $H_{re}^i \in \mathbb{R}^{2h}$ and $H_{re}^j \in \mathbb{R}^{2h}$ denote the vectors corresponding to words x_i and x_j in features $H_{re} \in \mathbb{R}^{N \times 2h}$ that containing both relation task-specific information and entity interaction information. $W_r^h \in \mathbb{R}^{4h \times h}$ and $b_r^h \in \mathbb{R}^h$ are trainable parameters. The probability that x_i and x_j are the start words of the subject and object of relation type l is calculated as follows:

$$p(r_{ij} = l) = \sigma(T_{re}^{i,j} W_r^{tag} + b_r^{tag}), \forall l \in \mathcal{R}, \tag{10}$$

where $W_r^{tag} \in \mathbb{R}^{h \times |\mathcal{R}|}$ and $b_r^{tag} \in \mathbb{R}^{|\mathcal{R}|}$ are trainable parameters and $|\mathcal{R}|$ represents the number of predefined relation types. We obtain the prediction results of NER module and RE module under the following conditions:

$$p(e_{i_1 j_1} = k_1) \geq 0.5;\ p(r_{i_1 i_2} = l) \geq 0.5;\ p(e_{i_2 j_2} = k_2) \geq 0.5 \tag{11}$$

where $k_1, k_2 \in \mathcal{E}, l \in \mathcal{R}$. For a fair comparison, the hyper-parameter threshold is set to be 0.5 without further fine-tuning as in previous works.

Combining the prediction results of NER and RE task modules, we can get the final relational triples in a given sentence:

$$Y_r = \{(e_{i_1 j_1}, r_{i_1 i_2}, e_{i_2 j_2})\}, e_{i_1 j_1}, e_{i_2 j_2} \in \mathcal{E}, r_{i_1 i_2} \in \mathcal{R}, \tag{12}$$

where $e_{i_1 j_1}$ and $e_{i_2 j_2}$ are entity spans predicted by NER task module, and $r_{i_1 i_2}$ denotes the relation between head-only entities predicted by RE task module.

3.4 Loss Function

During training, we adopt binary cross entropy loss for both NER and RE task modules. Given a sentence containing N words, we compute the NER loss and RE loss as follows:

$$\begin{aligned}
\mathcal{L}_{\text{NER}} &= -\sum_{i=1}^{N}\sum_{j=i}^{N}\sum_{k \epsilon \mathcal{E}} \hat{p}(e_{ij} = k) \log p(e_{ij} = k) + (1 - \hat{p}(e_{ij} = k)) \log (1 - p(e_{ij} = k)) \\
\mathcal{L}_{\text{RE}} &= -\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{l \epsilon \mathcal{R}} \hat{p}(r_{ij} = l) \log p(r_{ij} = l) + (1 - \hat{p}(r_{ij} = l)) \log (1 - p(r_{ij} = l)),
\end{aligned} \tag{13}$$

where $\hat{p}(e_{ij} = k)$ and $\hat{p}(r_{ij} = l)$ represent ground truth labels. $p(e_{ij} = k)$ and $p(r_{ij} = l)$ are the probability predicted by NER and RE modules. The final training goal is to minimize the sum of these two losses:

$$\mathcal{L} = \mathcal{L}_{\text{NER}} + \mathcal{L}_{\text{RE}}. \tag{14}$$

3.5 Differences from PFN

Our method differs from PFN (Yan et al., 2021) in the following ways: (1) We generate interaction features using distinct interaction gates, which are independent of the process of generating task-specific features. (2) All feature operations in MGE are performed in a non-autoregressive manner, i.e., all tokens in the sentence are processed in a single pass, resulting in increased efficiency. As a result, our method is simpler while still ensuring proper NER-RE interaction. Furthermore, as demonstrated in Section 4, our model outperforms PFN on three public datasets and achieves faster inference speed while employing the same task modules and pre-trained encoders.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4 Experiments

4.1 Dataset

We evaluate our model on three popular English relation extraction datasets: ACE05 (Walker et al., 2006), ACE04 (Doddington et al., 2004) and SciERC (Luan et al., 2018). The ACE05 and ACE04 datasets are collected from various domains, such as news articles and online forums. Following Luan et al. (2019), we split ACE04 into 5 folds and ACE05 into 10051 sentences for training, 2424 sentences for validation, and 2050 sentences for test [0]. And we follow Yan et al. (2021) to construct the development set of ACE04 with 15% of the training set.

Dataset	$\|\mathcal{E}\|$	$\|\mathcal{R}\|$	#Entities	#Relations	#Sentences		
					Train	Dev	Test
ACE05	7	6	38,287	7,070	10,051	2,424	2,050
ACE04	7	6	22,708	4,084	8,683 (5-fold)		
SciERC	6	7	8,094	4,684	1,861	275	551

Table 1: Statistics of datasets. $|\mathcal{E}|$ and $|\mathcal{R}|$ are numbers of entity and relation types.

The SciERC dataset is collected from 500 AI paper abstracts, and includes annotations for scientific entities, their relations, and coreference clusters. It consists six predefined scientific entity types and seven predefined relation types. In our experiments, we only use the annotation information of entities and relations. We download the processed dataset from the project website [1] of Luan et al. (2018), including 1861 sentences for training, 275 sentences for validation and 551 sentences for test. Table 1 shows the statistics of ACE04, ACE05 and SciERC datasets.

4.2 Evaluation

Following standard evaluation protocol, we use micro F1 score as an evaluation for both NER and RE. For NER task, an entity is considered as correct if its boundary and type are both predicted correctly. For RE task, a relational triple is correct only if its relation type and the boundaries and types of entities are correct.

4.3 Implementation Details

For fair comparison, we use *albert-xxlarge-v1* (Lan et al., 2020) as the base encoder for ACE04 and ACE05. And for SciERC, we use *scibert-scivocab-uncased* (Beltagy et al., 2019) as the base encoder. Regarding the use of cross-sentence context (Luan et al., 2019; Luoma and Pyysalo, 2020), that is, to extend each sentence by its context for better contextual representations, we don't adopt this experimental setting considering the fairness of experimental comparisons. Zhong and Chen (2021) extend each sentence to a fixed context window size of 300 words for entity model and 100 words for relation model. Ye et al. (2022) set the context window size to be 512 words for entity model and 256 / 384 words for relation model. Although cross-sentence context may further enhance the performance of entity recognition and relation extraction, if the research focus is not on the cross-sentence context, the different cross-sentence context lengths will greatly affect the experimental results, making it difficult to conduct fair comparisons. All our experiments are carried out in single-sentence setting and we compare with the experimental results of other baselines under the single-sentence setting.

Our model is implemented with PyTorch and we train our models with Adam optimizer of a linear scheduler with a warmup ratio of 0.1. For all the experiments, the learning rate and training epoch are set to be 2e-5 and 100 respectively. We set the batch size to be 4 for SciERC and 16 for ACE04 and ACE05. Following previous work (Yan et al., 2021), the max length of input sentence is set to be 128. All the

[0] We process the datasets with scripts provided by Luan et al. (2019): https://github.com/luanyi/DyGIE/tree/master/preprocessing.

[1] http://nlp.cs.washington.edu/sciIE/

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Encoder	ACE05		ACE04		SciERC	
		NER	RE	NER	RE	NER	RE
SPTree (Miwa and Bansal, 2016)	LSTM	83.4	55.6	81.8	48.4	-	-
Katiyar and Cardie (2017)	LSTM	82.6	53.6	79.6	45.7	-	-
Multi-turn QA (Li et al., 2019)	BERT	84.8	60.2	83.6	49.4	-	-
Table-Sequence (Wang and Lu, 2020)	ALBERT	89.5	64.3	88.6	59.6	-	-
SPE (Wang et al., 2020a)	SciBERT	-	-	-	-	68.0	34.6
PURE (Zhong and Chen, 2021)	ALBERT	**89.7**	65.6	88.8	60.2	-	-
	SciBERT	-	-	-	-	66.6	35.6
PFN (Yan et al., 2021)	ALBERT	89.0	66.8	**89.3**	62.5	-	-
	SciBERT	-	-	-	-	66.8	38.4
MGE (Ours)	ALBERT	**89.7**	**68.2**	**89.3**	**63.8**	-	-
	SciBERT	-	-	-	-	**68.4**	**39.4**

Table 2: Overall F1 scores on the test set of ACE04, ACE05, and SciERC. Results of PURE are reported in single-sentence setting for fair comparison.

models are trained with a single NVIDIA Titan RTX GPU. We select the model with the best average F1 score of NER and RE on the development set, and report the average F1 of 5 runs on the test set.

4.4 Baselines

We compare our model with the following baselines: (1) **BiLSTM** (Miwa and Bansal, 2016; Katiyar and Cardie, 2017): these models perform NER and RE based on shared Bi-directional LSTMs. Miwa and Bansal (2016) treats entity recognition as a sequence tagging task and represents the relations between entities in dependency tree. Katiyar and Cardie (2017) formulates both entity recognition and relation detection as sequence tagging tasks. (2) **Multi-turn QA** (Li et al., 2019): it converts the task into a multi-turn question answering task: each entity type and relation type has its corresponding pre-designed question template, and entities and relations are extracted by answering template questions with standard machine reading comprehension (MRC) (Seo et al., 2018) framework. (3) **Table-Sequence** (Wang and Lu, 2020): this work uses a sequence encoder and a table encoder to learn task-specific representations for NER and RE separately, and models task interaction through combining these two types of representations. (4) **SPE** (Wang et al., 2020a): this method proposes a span encoder and span pair encoder to add intra-span and inter-span information to the pre-trained model for entity and relation extraction task. (5) **PURE** (Zhong and Chen, 2021): this work builds two independent encoders for NER and RE separately and performs entity relation extraction in a pipelined fashion. PURE experimentally validates the importance of learning different contextual representations for entities and relations separately. (6) **PFN** (Yan et al., 2021): this work proposes a partition filter network to generate task-specific features and shared features of the two tasks, and then combining global features to extract entities and relations with table-filling framework.

Among these baselines, the two BiLSTM based methods build task interaction through parameter sharing, Multi-turn QA is a paradigm shift based method, PURE is a pipelined method, and Table-Sequence, SPE and PFN are methods based on multiple representations interaction.

4.5 Main Results

Table 2 reports the results of our approach MGE compared with other baselines on ACE05, ACE04 and SciERC. As is shown, MGE achieves the best results in terms of F1 score against all the comparison baselines. For NER, MGE achieves similar performance to PURE (Zhong and Chen, 2021) on ACE05 but surpasses PURE by an absolute entity F1 of +0.5%, +1.8% on ACE04 and SciERC. And for RE, our method obtains a substantially +2.6%, +3.6%, +3.8% absolute relation F1 improvement over PURE on ACE05, ACE04, and SciERC respectively. This demonstrates the superiority of the bidirectional task

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	SciERC		ACE05	
	RE (F1)	Speed (sent/s)	RE (F1)	Speed (sent/s)
PFN	38.4	342.2	66.8 / 60.8†	34.2 / 387.2†
MGE (Ours)	**39.4**	**479.2**	**68.2 / 62.0**†	**36.0 / 567.6**†

Table 3: We compare our MGE model with PFN model in both relation F1 and inference speed. We use $scibert - scivocab - uncased$ for SciERC and $albert - xxlarge - v1$ / $bert - base - cased$ for ACE05. † marks the inference speed on ACE05 when using $bert - base - cased$ encoder. The speed is measured on a single NVIDIA Titan V GPU with a batch size of 32.

interaction in our model compared to the unidirectional interaction in PURE.

In comparison to the previous state-of-the-art model PFN (Yan et al., 2021), we can see that our method achieves a similar entity F1 to PFN on ACE04, but an absolute relation F1 improvement of +1.3%. This suggests that, given the same NER performance, our method can obtain a better RE performance, implying that the entity knowledge in our method more effectively leads the RE task. Furthermore, on ACE05, MGE surpasses PFN by an absolute F1 improvement of +0.7% and +1.4% in NER and RE, respectively. On SciERC, we get a 1.6% higher entity F1 and a 1.0% higher relation F1 compared to PFN. Note that we use the same pre-trained encoders and task modules as PFN, and these improvements demonstrate the effectiveness of our proposed multi-gate encoder.

4.6 Inference Speed

As described in Section 3.5, our method employs a non-autoregressive way for feature encoding, which is simpler and faster than the autoregressive approach in PFN. In order to experimentally compare the model efficiency, we conduct experiments to evaluate these two models' inference speed on the test set of ACE05 and SciERC datasets. We perform inference experiments on a single NVIDIA Titan V GPU with a batch size of 32.

Table 3 shows the relation F1 scores and the inference speed of PFN and MGE. We use $scibert - scivocab - uncased$ encoder for SciERC and $albert - xxlarge - v1$ / $bert - base - cased$ (Devlin et al., 2019) encoder for ACE05. As is shown, with the same pre-trained model, our method obtains +1.0% improvement in relation F1 score with +40% speedup on the test set of SciERC. On ACE05, our model achieves a relation F1 improvement of +1.4% compared to PFN, but only slightly accelerates the inference speed (34.2 vs 36.0) when using $albert - xxlarge - v1$ pre-trained model. This is because $albert - xxlarge - v1$ contains 223M parameters, which is much larger than the 110M parameters in $scibert - scivocab - uncased$ and $bert - base - cased$, and most of the computational cost of the model is concentrated in the pre-trained model part. As a result, the speedup provided by MGE does not appear to be significant. Therefore, we also evaluate the inference speed on ACE05 using $bert - base - cased$. As Table 3 shows, our model achieves +47% speedup and an absolute relation F1 improvement of +1.2% on ACE05 when using $bert - base - cased$. This clearly demonstrates that our proposed MGE can improve the performance of joint entity and relation extraction while accelerating the model inference speed.

5 Analysis

In this section, we conduct ablation study on ACE05, ACE04 and SciERC to investigate how each component of MGE affects the final performance, where we apply $albert - xxlarge - v1$ encoder for ACE05 and ACE04, $scibert - scivocab - uncased$ encoder for SciERC. Specifically, we ablate the task gate or interaction gate to verify their effectiveness.

5.1 Effect of Task Gates.

We remove task gates from the complete MGE structure to explore whether they can generate effective task-specific features. As shown in Table 4, when we remove the entity task gate, the entity F1 scores on the ACE04 and SciERC datasets decrease by 0.5% and 0.2%, respectively. And when we remove the

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Encoder					ACE05		ACE04		SciERC	
B	G_{e_task}	G_{r_task}	G_{e_inter}	G_{r_inter}	NER	RE	NER	RE	NER	RE
✓	✓	✓	✓	✓	89.7	**68.2**	**89.3**	**63.8**	68.4	**39.4**
✓	-	✓	✓	✓	89.7	67.4	88.8	62.2	68.2	37.5
✓	✓	-	✓	✓	89.9	67.8	88.8	62.6	68.0	39.1
✓	✓	✓	-	✓	89.4	67.4	89.1	63.0	**68.5**	38.9
✓	✓	✓	✓	-	**90.0**	66.6	89.2	63.6	68.2	38.7
✓	✓	✓	-	-	**90.0**	66.1	88.4	62.8	67.9	37.8

Table 4: F1 scores of ablation study on ACE05, ACE04 and SciERC. B denotes BERT encoder. G_{e_task}, G_{r_task}, G_{e_inter} and G_{r_inter} means entity task gate, relation task gate, entity interaction gate and relation interaction gate.

relation task gate, the relation F1 scores on ACE05, ACE04 and SciERC datasets decrease by 0.4%, 1.2% and 0.3%, respectively. This indicates that task gates can effectively generate task-specific features to improve the performance of NER and RE.

5.2 Effect of Interaction Gates.

We also investigate the effect of the MGE entity interaction gate and relation interaction gate on task interaction. As there is no entity interaction gate, it is similar to weakening the guidance of entity information on the relation extraction task when compared to the unaffected MGE model. After deleting the entity interaction gate, the relation F1 scores on the ACE05, ACE04, and SciERC datasets decrease by 0.8%, 0.8%, and 0.5%, respectively, as shown in Table 4. In MGE, this highlights the effectiveness of the entity interaction gate.

Although it is widely accepted that entity information is necessary for relation extraction, previous research on the impact of relation information on entity recognition has been mixed. Zhong and Chen (2021) claims that relation information has no significant improvement on entity model. However, Yan et al. (2021) discover that relation signals have a significant impact on entity prediction. Our research also sheds light on this contentious issue. In MGE, the guidance of relation information on entity recognition is cut off when the relation interaction gate is ablate. The entity F1 scores decrease on ACE04 and SciERC but increase on ACE05 when the relation interaction gate is removed. Our experimental results match the experimental analysis of Yan et al. (2021). They conclude that relation information is helpful for predicting entities that appear in relational triples, but not for entities outside relational triples. According to Yan et al. (2021), there are fewer entities belonging to relational triples in ACE05, compared with ACE04 and SciERC. Consequently, the relation information is comparatively less helpful for entity recognition in ACE05 but has a positive effect on entity recognition in ACE04 and SciERC. To sum up, the relation interaction gate can effectively generate interaction features to facilitate the recognition of entities within triples.

Moreover, when we remove both the entity interaction gate and the relation interaction gate, the relation F1 scores on ACE05, ACE04 and SciERC datasets decrease by 2.1%, 1.0% and 1.6%, respectively. This shows the effectiveness of interaction gates in MGE for task interaction in joint entity relation extraction.

5.3 Bidirectional Interaction Vs Unidirectional Interaction.

From Table 4, we also observe that employing only an entity interaction gate or only a relation interaction gate in the encoder performs worse than adopting these two gates simultaneously. This means that the two tasks of entity recognition and relation extraction are mutually reinforcing, and bidirectional interaction between NER and RE is more effective than unidirectional interaction.

6 Conclusion

In this paper, we propose a multi-gate encoder for joint entity and relation extraction. Our model adopts gate mechanism to build bidirectional task interaction while ensuring the specificity of task features by

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

controlling the flow of feature information. Experimental results on three standard benchmarks show that our model achieves state-of-the-art F1 scores for both NER and RE. We conduct extensive analyses on three datasets to investigate the superiority of our model and validate the effectiveness of each component of our model. Furthermore, our ablation study suggests that relation information contributes to entity recognition, which helps to clarify the controversy on the effect of relation information.

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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018a. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium, October-November. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018b. Joint entity recognition and relation extraction as a multi-head selection problem. *CoRR*, abs/1804.07847.

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 (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, 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.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

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.

Kalpit Dixit and Yaser Al-Onaizan. 2019. Span-level model for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5308–5314, Florence, Italy, July. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 168–171.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy, July. 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.

Arzoo Katiyar and Claire Cardie. 2017. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 917–928, Vancouver, Canada, July. Association for Computational Linguistics.

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. In *iclr*.

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.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy, July. 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*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, October-November. Association for Computational Linguistics.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. 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 3036–3046, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 904–914, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August. Association for Computational Linguistics.

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.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2646–2656, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Bidirectional Attention Flow for Machine Comprehension, June. Number: arXiv:1611.01603 arXiv:1611.01603 [cs].

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

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 *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4461–4467. AAAI Press.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Yijun Wang, Changzhi Sun, Yuanbin Wu, Junchi Yan, Peng Gao, and Guotong Xie. 2020a. Pre-training entity relation encoder with intra-span and inter-span information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1692–1705, Online, November. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020b. 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.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2021. Gdpnet: Refining latent multi-view graph for relation extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 2–9.

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.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Pack together: Entity and relation extraction with levitated marker. In *Proceedings of ACL 2022*.

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.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 71–78. Association for Computational Linguistics, July.

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.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. End-to-end neural relation extraction with global optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1730–1740, Copenhagen, Denmark, September. Association for Computational Linguistics.

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 21st China National Conference on Computational Linguistics, pages 848-860, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Improving Event Temporal Relation Classification via Auxiliary Label-Aware Contrastive Learning

Tiesen Sun
School of Computer Science and Technology, Dalian University of Technology, Dalian, China

Lishuang Li *
School of Computer Science and Technology, Dalian University of Technology, Dalian, China
lils@dlut.edu.cn

Abstract

Event Temporal Relation Classification (ETRC) is crucial to natural language understanding. In recent years, the mainstream ETRC methods may not take advantage of lots of semantic information contained in golden temporal relation labels, which is lost by the discrete one-hot labels. To alleviate the loss of semantic information, we propose learning Temporal semantic information of the golden labels by Auxiliary Contrastive Learning (TempACL). Different from traditional contrastive learning methods, which further train the PreTrained Language Model (PTLM) with unsupervised settings before fine-tuning on target tasks, we design a supervised contrastive learning framework and make three improvements. Firstly, we design a new data augmentation method that generates augmentation data via matching templates established by us with golden labels. Secondly, we propose patient contrastive learning and design three patient strategies. Thirdly we design a label-aware contrastive learning loss function. Extensive experimental results show that our TempACL effectively adapts contrastive learning to supervised learning tasks which remain a challenge in practice. TempACL achieves new state-of-the-art results on TB-Dense and MATRES and outperforms the baseline model with up to 5.37%F_1 on TB-Dense and 1.81%F_1 on MATRES.

1 Introduction

The temporal relations of events are used to describe the occurring sequence of events in an article. Therefore understanding the temporal relations of events in articles is useful for many downstream tasks such as timeline creation (Leeuwenberg and Moens, 2018), generating stories (Goldfarb-Tarrant et al., 2020), forecasting social events (Jin et al., 2021), and reading comprehension (Ning et al., 2020). Hence, the ETRC task is an important and popular natural language understanding research topic among NLP community.

The ETRC task is to determine the occurrence sequence of a given event pair. The context of the event pair is usually given to aid judgment. Ning et al. (2019) first encoded the event pairs into embedded representations and then used fully connected layers as a classifier to generate confidence scores for each category of temporal relations. All related works of the NLP community since then have followed the classification view: classifying the embedded representations. Naturally, we can encode the context and events into a better embedding space in which the different relations are distinguished well, to get better classification results.

Traditionally, all recent works use one-hot vectors to represent golden temporal relation labels in the training stage. However, the one-hot vector reduces the label with practical semantics to the zero-one vector. It makes the embedded representations extracted by the ETRC models waiting for classifying be the similarities of the instances with the same label. But, the similarities are not equal to the label seman-

*Corresponding authors
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
Foundation: The National Natural Science Foundation of China (No. 62076048), The Science and Technology Innovation Foundation of Dalian (2020JJ26GX035).

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

tics, and lead to arbitrary prediction and poor model generalization, especially for confused instances. In brief, the one-hot vectors which represent temporal relation categories lose much semantic information.

To cope with the loss of semantic information in golden labels, we propose to learn the lost semantic information by contrastive learning, which is well confirmed and most competitive method for learning representations under unsupervised settings, so that the ETRC model can obtain better event representations. However, effectively adapting contrastive learning to supervised learning tasks remains a challenge in practice. General methods such as (Fang et al., 2020), which continue to train the PTLM model using unsupervised contrastive learning on the input texts (without labels) from the target task before fine-tuning, apply contrastive learning to supervised representation learning mechanically. They discard the category information in the process of further training. In the supervised ETRC task, we want the event pair representations with the same category to be as close as possible without collapsing. But direct application of the unsupervised contrastive learning loss function would prevent them from getting closer, because it discard the category information. It's an inherent problem of self-supervised contrastive learning. So the standard contrastive learning is not natural for the supervised ETRC task. To solve this problem we designed label-aware contrastive learning loss and design a new contrastive learning framework. Additionally, we argue that we can do contrastive learning in the intermediate layers of the PTLM as same as the last layer simultaneously. In a cascade structure, a change in previous layers affects the subsequent layers and continuous positive changes will make the learning process easier. Hence, we propose patient contrastive learning and design three patient strategies.

Overall, we propose TempACL: Firstly, we manually construct templates based on the semantics of labels and get augmentation sentences by matching the labels of instances. Secondly, we train the encoder of key samples which are necessary for contrastive learning by the augmentation datasets established by the ETRC datasets and the augmentation sentences. Thirdly, we jointly train the ETRC model with cross entropy loss and label-aware contrastive learning loss using a patient contrastive learning strategy.

The main contributions of this paper can be summarized as follows:

- We propose learning the lost semantic information in golden labels by contrastive learning, and then design TempACL, a supervised contrastive learning framework based on a new data augmentation method designed by us. To our knowledge, we are the first to propose using contrastive learning on the ETRC task.
- In order to make our TempACL achieve better performance, we design label-aware contrastive learning loss and patient contrastive learning strategy.
- We demonstrate the effectiveness of our TempACL on TB-Dense and MATRES datasets. Our TempACL outperforms the current best models with up to 2.13%F_1 on TB-Dense and 1.26%F_1 on MATRES and outperforms the baseline model with up to 5.37%F_1 on TB-Dense and 1.81%F_1 on MATRES.

2 Related work

2.1 Event Temporal Relation Classification

Since the birth of pre-trained language models, researchers have mainly used them to encode event representations and design many new methods based on them. Wang et al. (2020) propose a JCL method that makes the classification model learn their designed logical constraints within and across multiple temporal and subevent relations by converting these constraints into differentiable learning objectives. Zhou et al. (2021) propose the CTRL-PG method, which leverages the Probabilistic Soft Logic rules to model the temporal dependencies as a regularization term to jointly learn a relation classification model. Han et al. (2021) propose the ECONET system, which further trains the PTLM with a self-supervised learning strategy with mask prediction and a large-scale temporal relation corpus. Zhang et al. (2021) propose the TGT network that integrates both traditional multi-head self-attention and a new temporal-oriented attention mechanism and utilizes a syntactic graph that can explicitly find the connection between two events. Tan et al. (2021) propose the Poincaré Event Embeddings method which

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

encodes events into hyperbolic spaces. They argue that the embeddings in the hyperbolic space can capture richer asymmetric temporal relations than the embeddings in the Euclidean space. And they also proposed the HGRU method which additionally uses an end-to-end architecture composed of hyperbolic neural units, and introduces common sense knowledge (Ning et al., 2019).

All of the above methods use the one-hot vector and lose the semantic information of the golden label. To take advantage of the missing semantic information, we make the target ETRC model learn from them via contrastive learning.

2.2 Contrastvie Learning

Contrastive learning aims to learn efficient representations by pulling semantically close neighbors together and pushing non-neighbors away (Hadsell et al., 2006). In recent years, self-supervised contrastive learning and supervised contrastive learning have attracted more and more researchers to study them.

Self-Supervised Contrastvie Learning. In computer vision (CV), Wu et al. (2018) propose MemoryBank, which maintain a large number of representations of negative samples during training and update negative sample representations without increasing batch size. He et al. (2020) propose MoCo, which designs the momentum contrast learning with two encoders and employs a queue to save the recently encoded batches as negative samples. Chen et al. (2020) proposed the SimCLR which learns representations for visual inputs by maximizing agreement between differently augmented views of the same sample via a contrastive loss. Grill et al. (2020) propose BYOL, which uses asymmetric two networks and discards negative sampling in self-supervised learning. In Natural Language Processing (NLP), Yan et al. (2021) propose ConSERT, which has a similar model structure to SimCLR, except that ResNet is replaced by Bert and the mapping header is removed. And they also propose multiple data augmentation strategies for contrastive learning, including adversarial attack, token shuffling, cutoff and dropout.

Supervised Contrastvie Learning. Khosla et al. (2020) extend the self-supervised contrastive approach to the fully-supervised setting in the CV domain, and take many positives per anchor in addition to many negatives (as opposed to self-supervised contrastive learning which uses only a single positive). Gunel et al. (2020) extends supervised contrastive learning to the NLP domain with PTLMs.

Different from ConSERT we design a new data augmentation method based on templates in our contrastive learning framework. And different from Khosla's work, we design a new supervised contrastive loss which still uses only a single positive but does not treat the sentence representations with the same label as negative examples.

3 Our Baseline Model

Our baseline model is comprised of an encoder and a classifier. We use RoBERTa (Liu et al., 2019) as our encoder and use two fully connected layers and a tanh activation function between them as our classifier. Recently, most of the related works use RoBERTa as an encoder, because RoBERTa can achieve better results on the ETRC task than BERT in practice.

Each instance is composed of an event temporal triplet t (i.e. ($< e_1 >$, $< e_2 >$, r), where $< e_1 >$ and $< e_2 >$ are event mentions and r is the temporal relation of the event pair.) and the context s of the events which may be a single sentence or two sentences.

We first tokenize the context and get a sequence of tokens $X_{[0,n)}$ with length n. Then we feed the $X_{[0,n)}$ into RoBERTa. One event mention may correspond to multiple tokens, so we send the token embeddings corresponding to these tokens to an average pooling layer to get the final event representation e_i. Next, we combine e_1 and e_2 into a classification vector $e_1 \oplus e_2$, where $\oplus$ is used to denote concatenation. Finally, we feed the classification vector into the classifier followed by a soft-max function to get confidence scores for each category of temporal relations.

4 Self-Supervised Contrastive Learning

Contrastive learning is learning by pulling similar instance pairs closer and pushing dissimilar instance pairs farther. The core of self-supervised contrastive learning is to generate augmented examples of original data examples, create a predictive task where the goal is to predict whether two augmented

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

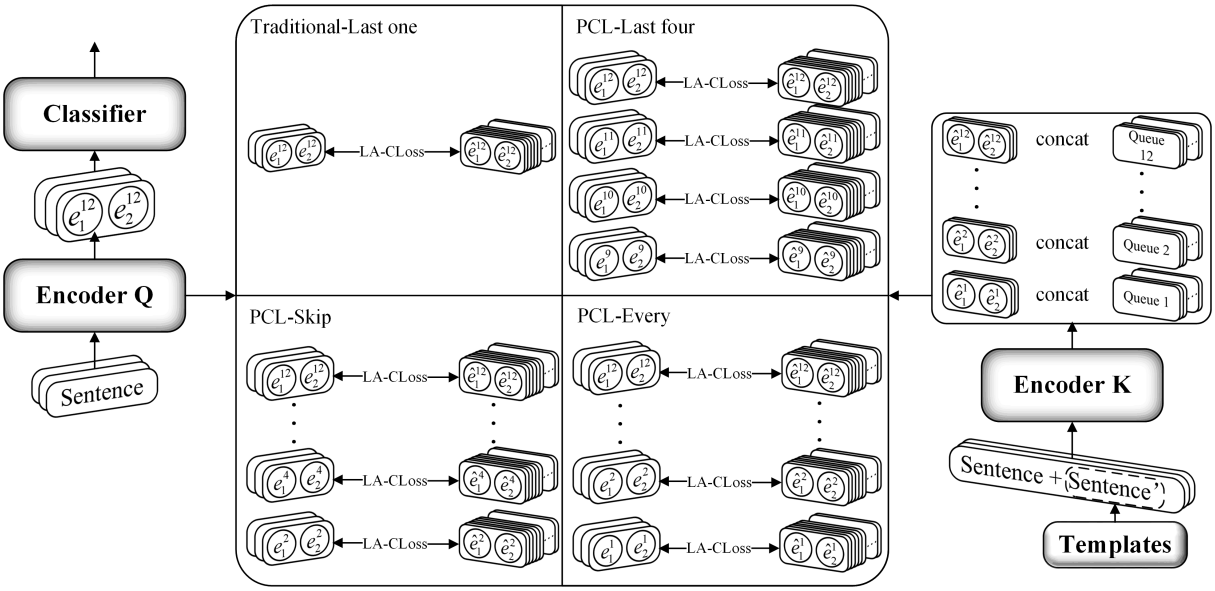


Figure 1: Joint training with patient contrastive learning. We name the PLTM which encodes positive and negative key samples as Encoder K and the PLTM used for ETRC as Encoder Q.

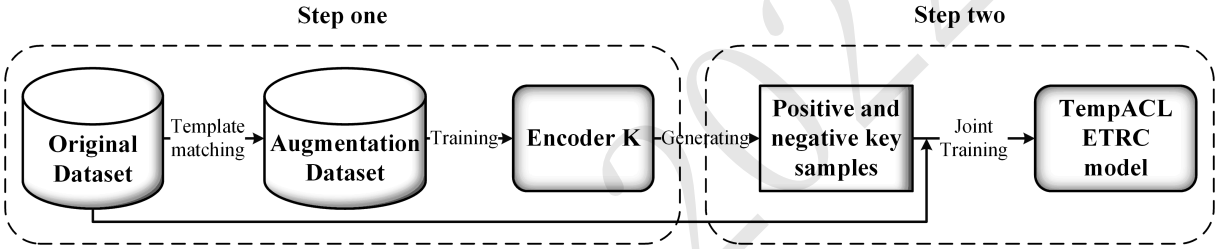


Figure 2: Overall process of TempACL

examples are from the same original data example or not, and learn the representation network by solving this task. He et al. (2020) formulate contrastive learning as a dictionary look-up problem and propose an effective contrastive loss function L_{CL} with similarity measured by dot product:

$$L_{CL} = -\log \frac{\exp\left(q \cdot k^{+}/\tau\right)}{\exp\left(q \cdot k^{+}/\tau\right) + \sum_{\{K^{-}\}} \exp\left(q \cdot k^{-}/\tau\right)} \tag{1}$$

where q is a query representation, k^{+} is a representation of the positive (similar) key sample, k^{-} are representations of the negative (dissimilar) key samples, K^{-} is a negative key samples set, and τ is a temperature hyper-parameter. He et al. (2020) also propose maintaining the dictionary as a queue of data samples. It allows contrastive learning to reuse the previous batch of key samples so that we can increase the number of negative samples without increasing the batch size, thus improving the performance of the model. The dictionary size is a flexible hyper-parameter. The samples in the dictionary are progressively replaced. The current batch is enqueued to the dictionary, and the oldest batch in the queue is removed. In this paper, we follow this part of their work and transfer it to the supervised ETRC task.

5 TempACL Approach

In this section, we introduce our TempACL approach in details and draw the overall process of TempACL in Figure 2. TempACL aims to encoder semantic information of golden temporal relation labels and uses contrastive learning to make the baseline model extract better event representations. Hence, we first train

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Temporal Relation	Templates
AFTER*	the beginning of the event of $< e_1 >$ is after the end of the event of $< e_2 >$.
BEFORE*	the end of the event of $< e_1 >$ is before the beginning of the event of $< e_2 >$.
INCLUDES	the beginning of the event of $< e_1 >$ is before the beginning of the event of $< e_2 >$ and the end of event of $< e_1 >$ is after the end of the event of $< e_2 >$.
IS_INCLUDED	the beginning of the event of $< e_1 >$ is after the beginning of the event of $< e_2 >$ and the end of event of $< e_1 >$ is before the end of the event of $< e_2 >$.
VAGUE*	the temporal relation between the event of $< e_1 >$ and the event of $< e_2 >$ is vague.
SIMULTANEOUS*	the event of $< e_1 >$ and the event of $< e_2 >$ have the same beginning and end time.

Table 1: Templates. All the six temporal relation labels are in TB-Dense and * indicates the temporal relation label also exists in MATRES.

Encoder K used for encoding semantic information of golden temporal relation labels, and then jointly train the baseline model with auxiliary contrastive learning via the label-aware contrastive learning loss function and a patient strategy. Specially, we fix the parameters of the Encoder K in the joint training stage.

5.1 Training Encoder K

First of all, we need to establish templates. In order to make the positive key samples encoded by Encoder K contain as much and as detailed semantic information of golden temporal relation labels as possible, we need to create efficient templates that automatically convert each golden temporal relation label into a temporal information-enriched sentence s' to enrich the semantic information of golden temporal relation labels. We argue that the time span of events (i.e., the duration of the events) guides ETRC. So we use the start and end times of events and the temporal relation between events to describe the temporal relation of the event pair on a subtle level. We show the templates in Table 1.

Subsequently, we build the augmentation dataset. For each record (t, s) in original Dataset, we use r to match the templates and get s' by filling events into the corresponding positions in the template, then concatenate s and s' to get an augmentation sentence $s_{aug} = s + s'$, finally get a new record (t, s_{aug}). We combine all new records into an augmentation dataset.

Finally, we use the augmentation dataset to train the Encoder K with the help of the classifier which we propose in section 3 under supervised setting. Encoder K is a RoBERTa model.

5.2 Joint Training with Patient Label-aware Contrastive Loss

The trained Encoder K has been obtained, we can start joint training in Figure 1. We send s in the original dataset to Encoder Q, and then get event pair representations $\{e_{1j} \oplus e_{2j}\}_{j=1}^{12}$ in different layers of Encoder Q. e_{ij} is the hidden state corresponding to the event i from the j-th RoBERTa Layer. We simultaneously send s_{aug} in the augmentation dataset to Encoder K, and then get event pair representations $\{\hat{e}_{1j} \oplus \hat{e}_{2j}\}_{j=1}^{12}$ in different layers of Encoder K. $\hat{e}_{ij}$ is the hidden state corresponding to the event i from the j-th RoBERTa Layer, and $\hat{}$ is used to denote the hidden state from the Encoder K. We normalized $e_{1j} \oplus e_{2j}$ as the query q and $\hat{e}_{1j} \oplus \hat{e}_{2j}$ as key k with L2 Norm. According to different patient strategies, queries and keys of different layers were selected for comparative learning.

We should not mechanically apply the loss function of self-supervised contrastive learning in equation 1 to the supervised ETRC directly. In the supervised ETRC task, we want the event pair representations with the same category to be as close as possible without collapsing. But L_{CL} treat the key samples in the queue, whose event pair have the same temporal relation with the event pair of the query sample, as negative key samples. Therefore, in the process of minimizing the L_{CL}, the event pair representations

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	TB-Dense		MATRES	
	Documents	Triplets	Documents	Triplets
Train	22	4032	204	10097
Dev	5	629	51	2643
Test	9	1427	20	837

Table 2: Data statistics for TB-Dense and MATRES

with the same category are mutually exclusive, which confuse the ETRC model. So we propose label-aware contrastive loss function L_{LACL} :

$$L_{LACL} = -\sum_{i=1}^{N}\left(\log \frac{\exp\left(q \cdot k^{+}/\tau\right)}{\exp\left(q \cdot k^{+}/\tau\right) + \sum_{\{K'^{-}\}} \exp\left(q \cdot k'^{-}/\tau\right)}\right)_{i} \tag{2}$$

where $\bar{K}^{-}$is negative key samples set which except the key samples with the same label as q, and N is the number of training samples. In practice, we convert $q \cdot k$ where $k \in \{k : k \in K^{-}, k \notin K'^{-}\}$ to -10^{6} by matrix operations.

Inspired by Sun et al. (2019), we argue that using the event pair representations of the intermediate layers of the Encoder Q and the event pair representations of the intermediate layers of the Encoder Q for additional contrastive learning can enhance the learning of semantics of the Encoder Q, and improve the performance of the baseline model. Hence we propose patient label-aware contrastive learning loss L_{PCL} based on equation 2:

$$L_{PCL} = -\sum_{j \in J}\sum_{i=1}^{N}\frac{1}{\|J\|}\left(\log \frac{\exp\left(q \cdot k^{+}/\tau\right)}{\exp\left(q \cdot k^{+}/\tau\right) + \sum_{\{K'^{-}\}} \exp\left(q \cdot k'^{-}/\tau\right)}\right)_{i,j} \tag{3}$$

where J is the set of intermediate layers involved in contrastive learning. Specifically, we propose three patient contrastive learning strategies: (1) PCL-Last four: we contrast the last four layers of the Encoder Q and Encoder K (Figure 1 upper right). (2) PCL-Skip: we contrast every two layers of the Encoder Q and Encoder K (Figure 1 lower left). (3) PCL-Every: we contrast every layers of the Encoder Q and Encoder K (Figure 1 lower right).

Finally, we jointly train ETRC task and auxiliary label-aware contrastive learning task with the final loss function L_{final}:

$$L_{finall} = \alpha L_{CE} + \beta L_{PCL} \tag{4}$$

where L_{CE} is cross-entropy loss function, α and β are hyper-parameters which weight the importances of ETRC task and auxiliary label-aware contrastive learning task.

6 Experiments and Results

In this section, we perform experiments on TB-Dense and MATERS and prove our TempACL performs better than previous state-of-the-art methods. Details on the datasets, experimental setup, and experimental results are provided in the following subsections.

6.1 Dataset

6.1.1 TB-Dense

TB-Dense(Cassidy et al., 2014) is a densely annotated dataset for the ETRC and annotated based on TimeBank. It also annotates the temporal relations of pairs of events across sentences, different from TimeBank which only annotates events in the same sentence. It annotates a total of 6 temporal relations (AFTER, BEFORE, INCLUDE, IS INCLUDED, VAGUE, SIMULTANEOUS). We follow the split strategy of Han et al. (2021) and Zhang et al. (2021) which uses 22 documents as train set, 5 documents as dev set and 9 documents as test set.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

6.1.2 MATERS

MATERS(Ning et al., 2018) is refined from 275 documents in TimeBank and TempEval (containing AQUAINT and Platinum). Ning et al. (2018) design a novel multi-axis (i.e., main, intention, opinion and hypothetical axes) annotation scheme to further annotate the 275 documents. There are only 4 temporal relations (BEFORE, AFTER, EQUAL and VAGUE) different from TB-Dense and the EQUAL is the same as SIMULTANEOUS. We follow the official split strategy that uses TimeBank and AQUAINT for training and Platinum for testing. We also follow the previous works (Ning et al., 2019; Tan et al., 2021) that randomly select 20 percents of the official train documents as dev set.

We briefly summarize the data statistics for TB-Dense and MATRES in Table 2.

6.2 Experimental Setup

In the process of training Encoder K, we add a dropout layer between the Encoder K and the Classifier and set the drop probability to 0.5, in order to make the key samples contain more useful temporal information. We train Encoder K 10 and 20 epochs respectively on TB-Dense and MATRES. We set the batch size to 24, the τ to 0.1, the learning rate of the Classifier to 5e-4 and the learning rate of RoBERTa to 5e-6. We use grid search strategy to select the best $\alpha \in$ [0.7: 1.4] and $\beta \in$ [0.01: 0.001]. As for the dimension of the hidden states between two fully connected layers in the Classifier, we set it to 36. We set the size of the queue to 3840 and 9600 respectively on TB-Dense and MATRES.

6.3 Main Results

As shown in Table 3, we compare our approach with other state-of-the-art methods in recent years on TB-Dense and MATRES. We report the best F_1 value for each method. The compared methods have been introduced in section 2. And the results of compared methods are directly taken from the cited papers except CERT[1]. We reproduce CERT and record the results.

We observe that our baseline model achieves 63.56%F_1 on TB-Dense and 79.95%F_1 on MATRES. It demonstrates that our baseline model can effectively classify temporal relation, and even achieves a competitive performance that is close to the current best 80.5%F_1 on MATRES. Furthermore, our TempACL outperforms previous state-of-the-art methods on ETRC with up to 2.13%F_1 on TB-Dense and 1.26%F_1 on MATRES. Compared with CERT, the traditional self-supervised contrastive learning method, our TempACL achieves 4.01%F_1 and 1.30%F_1 improvement respectively. These experimental results prove the effectiveness of learning semantic information of golden temporal relation labels via patient label-aware contrastive learning. There are three possible reasons for the effectiveness: (1) The difference between the query representation and the key representation comes from the semantic information of the golden temporal relation label, because the input of Encoder Q doesn't have the label information but the input of Encoder K input does. The L_{LACL} forces q closer to K to reduce the difference. So that in the process of minimizing L_{LACL} Encoder Q learns the label semantic information and forces itself to extract more useful information related to golden temporal relation labels from the sentences that do not contain any golden temporal relation label information. (2) The supervised contrastive learning framework and L_{LACL} designed by us is more suitable for the ETRC task than the traditional self-supervised contrastive learning method. (3) The data augmentation method proposed by us not only utilizes the semantic information of labels but also enriches the semantic information of labels.

Different from JCL and HGRU, which use external commonsense knowledge to enrich the information contained in event representations, TempACL enables the model to better mine the information contained in original sentences. Compared to ECONET and TGT, which use a larger pre-trained language model, or TGT and HGRU, which use networks with complex structures followed RoBERTa base or BERT Large, TempACL enables a smaller and simpler model which only contains a RoBERTa base and two fully connected layers to achieve the state-of-the-art performance.

[1]https://github.com/UCSD-AI4H/CERT

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method		TB-Dense	MATRES
JCL(Wang et al., 2020)	RoBERTa base	-	78.8
ECONET(Han et al., 2021)	RoBERTa Large	66.8	79.3
TGT(Zhang et al., 2021)	BERT Large	66.7	80.3
Poincaré Event Embeddings(Tan et al., 2021)	RoBERTa base	-	78.9
HGRU+knowledge(Tan et al., 2021)	RoBERTa base	-	80.5
CERT(Fang et al., 2020)	RoBERTa base	64.92	80.46
Baseline (ours)	RoBERTa base	63.56	79.95
TempACL (ours)	RoBERTa base	**68.93**	**81.76**

Table 3: Comparison of various approaches on ETRC on TB-Dense and MATRES. Bold denotes the best performing model. F_1-score (%)

Method	TB-Dense	MATRES
Traditional-Last one	66.17	80.95
PCL-Last four	68.93	81.76
PCL-Skip	67.73	80.46
PCL-Every	65.23	80.37

Table 4: Results of TempACL with different strategies. F_1-score (%)

6.4 Ablation Study and Qualitative Analysis

We observe that, TempACL make improvements of 5.37%F_1 and 1.81%F_1 on TB-Dense and MATRES respectively compared with the baseline model. In this section, we first qualitatively analyze key samples, and then we do the ablation experiments to further study the effects of patient strategies and label-aware contrastive learning loss. We ensure that all ablation results are optimal by using optimal strategies under the given conditions.

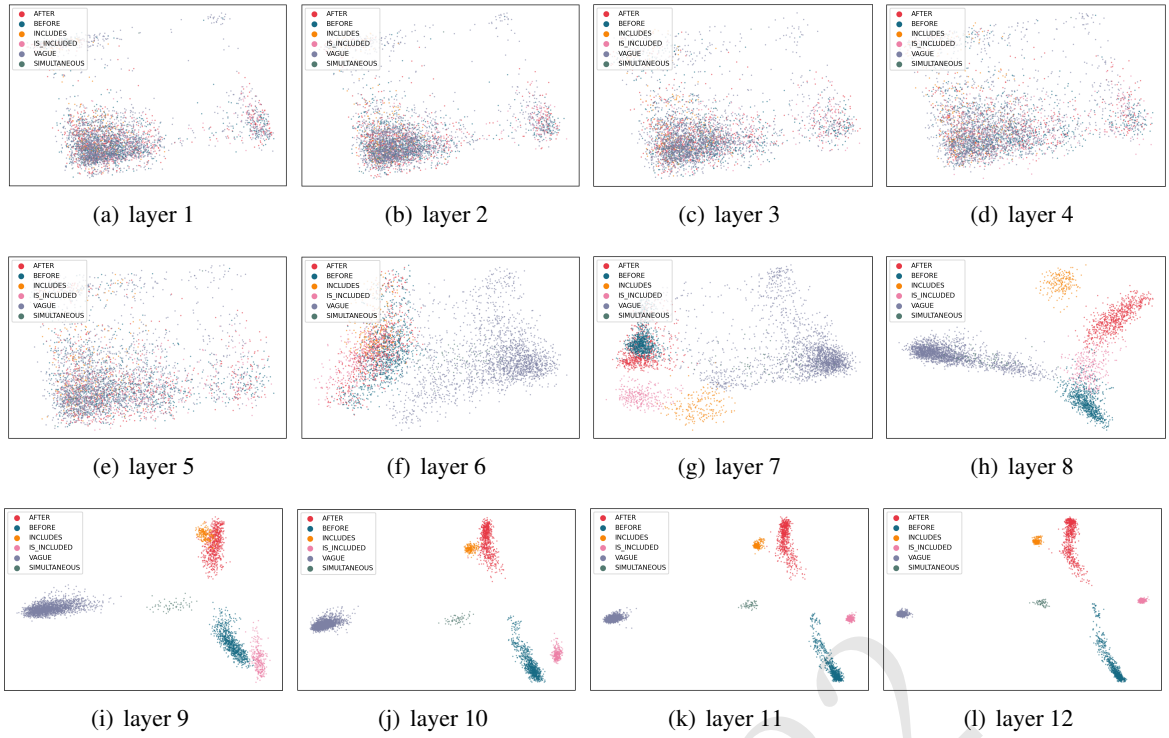
6.4.1 Qualitative analysis.

Wang and Isola (2020) propose to justify the effectiveness of contrastive learning in terms of simultaneously achieving both alignment and uniformity. Hence we reduce the dimension of key samples in each layer through PCA and represent it in Fig.3 on TB-Dense. All four contrastive strategies we used to utilize the key samples of the last layer, so we take Figure 3(l) to analyze the alignment and uniformity of TempACL. On the one hand, we can see that there are 6 clusters of representations that are well-differentiated even in two dimensions. Our method maps key samples with the same category to a relatively dense region. These well demonstrate that our embedded knowledge has a strong alignment. On the other hand, we also can see that the 5 clusters, which represent temporal categories in Figure 3(l) right, are farther from the VAGUE cluster than each other. It means that our embedded knowledge retains as much category information as possible. The farther away different clusters are, the more category information and differences are retained. Moreover, different key samples with the same category distribute evenly within the dense region, which means that our key samples retain as much instance information as possible. Furthermore, the more evenly distributed they are, the more information they retain. These well demonstrate that our embedded knowledge has a strong uniformity. We find that the key samples encoded by the last four layers of the Encoder K have strong alignment and uniformity.

6.4.2 Last one strategy VS Patient strategy

In section 5.2 we propose three patient strategies. In this section, we do experiments to study which strategy is optimal and report the experimental results in Table 4. PCL-Last four achieves the best results on both TB-Dense and MATRES. On the one hand, PCL-Last four provides more positive and negative samples. In Figure 3, the distribution of key samples in the last four layers also indicates that these positive and negative samples have great value in learning. On the other hand, this layer-by-layer approach greatly reduces the difficulty of learning. In the PTLM, different sub-layers are cascade, and the

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China



(a) layer 1 (b) layer 2 (c) layer 3 (d) layer 4

(e) layer 5 (f) layer 6 (g) layer 7 (h) layer 8

(i) layer 9 (j) layer 10 (k) layer 11 (l) layer 12

Figure 3: The distributions of key samples of each RoBERTa layers on TB-Dense.

Method	TB-Dense	MATRES
TempACL-LACL	68.93	81.76
TempACL-TCL	66.03	80.89
baseline	63.56	79.95

Table 5: Results of TempACL with different contrastive learning loss. F_1-score (%)

changes in the output in the front layers influence the latter layers. PCL-every performs poorly and worse than Traditional-Last one, because the first eight layers do not provide good positive and negative key samples, and learning them confuses the model. However PCL-Skip performs better than Traditional-Last one. This is because the number of bad key samples in PCL-Skip is relatively small, which makes the negative impact of these bad key samples much smaller. The layer-by-layer approach reduces the difficulty of learning and the benefits outweigh the negative impact.

6.4.3 Label-aware contrastive loss vs traditional contrastive loss

In order to determine whether our proposed label-aware contrastive loss has a positive effect, we conduct a comparative experiment and record the experimental results in Table 5. We compare the TempACL with label-aware contrastive learning loss (TempACL-LACL) and the TempACL with traditional contrastive learning loss (TempACL-TCL) on TB-Dense and MATRES respectively. We can see that the TempACL-LACL achieves 2.90%F_1 and 0.87%F_1 performance improvement over the TempACL-TCL respectively. It shows the benefit of eliminating key samples with the same label as the query from the negative samples set. The reason is that using key samples, which have the same label as the query, as negative samples prevent instances of the same label from learning similar event representations to some extent, which runs counter to the ETRC's aims. And the label-aware contrastive learning loss can avoid such a situation.

7 Conclusion

In recent years, the mainstream ETRC methods focus on using discrete values to represent temporal relation categories and lose too much semantic information contained in golden labels. So we propose

TempACL, which makes the ETRC model learn the lost semantic information in golden labels via contrastive learning. Extensive experiments prove the contrastive learning framework in TempACL is more suitable for the supervised ETRC task than traditional self-supervised contrastive learning. The patient contrastive learning strategy designed by us provides more useful positive and negative key samples and reduces the difficulty of contrastive learning. The label-aware contrastive learning loss designed by us avoids the negative interactions between different queries and keys in the same category, which is an inherent problem of self-supervised contrastive learning.

References

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *ACL (Volume 2: Short Papers)*, pages 501–506.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv e-prints*, pages arXiv–2005.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *EMNLP*, pages 4319–4338.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

R. Hadsell, S. Chopra, and Y. 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.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. ECONET: Effective continual pretraining of language models for event temporal reasoning. In *EMNLP*, pages 5367–5380.

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. IEEE.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. Forecastqa: A question answering challenge for event forecasting with temporal text data. pages 4636–4650.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *ACL (Volume 1: Long Papers)*, pages 1318–1328.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *EMNLP-IJCNLP*, pages 6203–6209.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *EMNLP*, pages 1158–1172.

CCL 2022

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *EMNLP-IJCNLP*, pages 4323–4332.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *EMNLP*, pages 8065–8077.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *EMNLP*, pages 696–706.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv e-prints*, pages arXiv–2105.

Shuaicheng Zhang, Lifu Huang, and Qiang Ning. 2021. Extracting temporal event relation with syntactic-guided temporal graph transformer. *arXiv preprint arXiv:2104.09570*.

Yichao Zhou, Yu Yan, Rujun Han, J Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *AAAI*, volume 35, pages 14647–14655.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 861-871, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Towards Making the Most of Pre-trained Translation Model for Quality Estimation

Chunyou Li[1], **Hui Di**[2], **Hui Huang**[1], **Kazushige Ouchi**[2]
Yufeng Chen[1], **Jian Liu**[1], **Jinan Xu**[1*]
[1] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China
[2] Toshiba (China) Co., Ltd.
{21120368,chenyf,jianliu,jaxu}@bjtu.edu.cn;
dihui@toshiba.com.cn; huanghui_hit@126.com;
kazushige.ouchi@toshiba.co.jp

Abstract

Machine translation quality estimation (QE) aims to evaluate the quality of machine translation automatically without relying on any reference. One common practice is applying the translation model as a feature extractor. However, there exist several discrepancies between the translation model and the QE model. The translation model is trained in an autoregressive manner, while the QE model is performed in a non-autoregressive manner. Besides, the translation model only learns to model human-crafted parallel data, while the QE model needs to model machine-translated noisy data. In order to bridge these discrepancies, we propose two strategies to post-train the translation model, namely Conditional Masked Language Modeling (CMLM) and Denoising Restoration (DR). Specifically, CMLM learns to predict masked tokens at the target side conditioned on the source sentence. DR firstly introduces noise to the target side of parallel data, and the model is trained to detect and recover the introduced noise. Both strategies can adapt the pre-trained translation model to the QE-style prediction task. Experimental results show that our model achieves impressive results, significantly outperforming the baseline model, verifying the effectiveness of our proposed methods.

1 Introduction

Machine translation has always been the hotspot and focus of research. Compared with traditional methods, neural machine translation (NMT) has achieved great success. However, current translation systems are still not perfect to meet the real-world applications without human post-editing. Therefore, to carry out risk assessment and quality control for machine translation, how to evaluate the quality of machine translation is also an important problem.

Quality Estimation (QE) aims to predict the quality of machine translation automatically without relying on reference. Compared with commonly used machine translation metrics such as BLEU (Papineni et al., 2002) and METEOR (Lavie and Denkowski, 2009), QE can be applicable to the case where reference translations are unavailable. It has a wide range of applications in post-editing and quality control for machine translation. The biggest challenge for QE is data scarcity. Since QE data is often limited in size, it is natural to transfer bilingual knowledge from parallel data to the QE task.

One well-known framework for this knowledge transfer is the predictor-estimator framework, in which the predictor is trained on large parallel data and used to extract features, and the estimator will make quality estimation based on features provided by the predictor. The predictor is usually a machine translation model, which can hopefully capture the alignment or semantic information of the source and the target in a pair. Kim et al. (2016) first proposed to use an RNN-based machine translation model as the feature extractor, to leverage massive parallel data to alleviate the sparsity of annotated QE data. Wang et al. (2020) employed a pre-trained translation model as the predictor and added pseudo-PE information to predict translation quality.

* Corresponding author.
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

However, there are two discrepancies between machine translation and quality prediction, which impedes the NMT model to be directly adopted for feature extraction. i) Translation task is usually a language generation task trained in an autoregressive manner, where each token is only conditioned on previous tokens unidirectionally. But QE is a language understanding task performed in a non-autoregressive manner, therefore each token could attend to the whole context bidirectionally. ii) The predictor is trained on human-crafted parallel data and only learns to model the alignment between correct translation pairs. However, the QE task needs to model machine-translated, imperfect translation pairs. Both discrepancies may hinder the adaptation of the pre-trained NMT model to the downstream QE task, leading a degradation of model performance (Weiss et al., 2016).

In this paper, we propose two strategies to alleviate the discrepancies, named as Conditional Mask Language Modeling (CMLM) and Denoising Restoration (DR). Both strategies are applied to the pre-trained NMT model and can be deemed as a post-training phase. The CMLM is to train the NMT model to recover the masked tokens at the target side in a non-autoregressive manner, where each token can attend to the whole target sequence bidirectionally. Furthermore, the DR first generates erroneous translation by performing conditionally masked language modeling, and then trains the NMT model to detect the introduced noise and recover the target sequence, which is also performed in a non-autoregressive manner. Both methods can adapt the autoregressive NMT model to non-autoregressive QE prediction. Moreover, compared with CMLM, DR removes the introduction of [MASK] token (which may also cause the discrepancy between pre-training and QE prediction). Besides, adversarially using another model with knowledge distillation to generate noise could provide more natural and harder training samples, thereby pushing the translation model better model the semantic alignment between the imperfect translation and source sequence. After the post-training phase, the NMT model is better adapted to the quality prediction task, and can serve as a better feature extractor.

Our contributions can be summarized as follows:

- We propose two strategies for post-training the NMT model to bridge the gaps between machine translation and quality estimation, which can make the NMT model more suitable to act as the feature extractor for the QE task.
- We conduct experiments on the WMT21 QE tasks for En-Zh and En-De directions, and our methods outperform the baseline model by a large margin, proving its effectiveness. We also perform in-depth analysis to dig into the discrepancies between translation and quality prediction.

2 Background

2.1 Task description

Quality Estimation aims to predict the translation quality of an MT system without relying on any reference. In this task, the dataset is expressed in the format of triplet (s, m, q), where s represents the source sentence, m is the translation output from a machine translation system, and q is the quality score of machine translation.

Generally, Quality Estimation task includes both word-level and sentence-level tasks. In word-level task, the prediction is done both on source side (to detect which words caused errors) and target side (to detect mistranslated or missing words). In sentence-level task, it will mark each sentence with a score, which can be calculated based on different standards, consists of Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), Direct Assessment (DA) (Graham et al., 2017), Multidimensional Quality Metrics (MQM) (Lommel et al., 2014), etc. In this work, we mainly focus on sentence level post-editing effort prediction, which is measured by:

$$HTER = (I + D + R)/L, \tag{1}$$

where I, D and R are the number of Insertions, Deletions and Replacement operations required for post-editing, and L is the reference length. However, labeling the data requires post-editing for the machine translations by experts, leading the label of QE data too expensive to obtain, which makes QE highly data-sparse.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

2.2 Previous work

Generally, sentence-level QE is fomulated as a regression task. Early approaches were based on features fed into a traditional machine learning method, such as QuEst++ (Specia et al., 2015) and MARMOT (Logacheva et al., 2016) system. These model usually has two modules: the feature extraction module and the classification module. But they relied on heuristic artificial feature designing, which limits their development and application (Huang et al., 2020). With the increasing popularity of deep learning methods, researchers resort to distributed representations and recurrent networks to encode translation pairs. However, the limited size of training samples impedes the learning of deep networks (Martins et al., 2017). To solve this problem, a lot of research has been done to use additional resource (both bilingual and monolingual) to strengthen the representation (Kim and Lee, 2016) . After the emergence of BERT (Devlin et al., 2018), some work attempts to use the pre-trained language model as a predictor directly and add a simple linear on top of the model to obtain the predictions (Chen et al., 2021; Chowdhury et al., 2021), which has led to significant improvements.

Among all the deep learning-based methods, one commonly used framework for QE is the predictor-estimator framework, where the predictor is used as a feature extractor and the estimator uses the features to make predictions. The predictor is usually a translation model, which can alleviate the problem of data sparsity by transferring bilingual knowledge from parallel data. Kim et al. (2016) firstly proposed the predictor-estimator framework to leverage massive parallel data to improve QE results, they applied an RNN-based machine translation model as the predictor and added a bidirectional RNN as estimator to predict QE scores, which achieved excellent performance especially in sentence-level QE. Fan et al. (2019) used Transformer-based NMT model as the predictor to extract high-quality features, and used 4-dimensional mis-matching features from this model to improve performance. Wang et al. (2019) pretrained left-to-right and right-to-left deep Transformer models as the predictor and introduced a multilayer bidirectional Gated Recurrent Unit (Bi-GRU) as the estimator to make prediction. Wu et al. (2020) reformed Transformer-based predictor-estimator by using multidecoding during the machine translation module, then implemented LSTM-based and Transformer-based estimator with top-K and multi-head attention strategy to enhance the sentence feature representation. Wang et al. (2020) employed a pretrained translation model as the predictor and added pseudo-PE information to predict translation quality, which obtained the best result in the English-German direction of WMT20. However, despite various of improvement has been made on the predictor-estimator framework, the discrepancy problem between machine translation and quality estimation is not systematically investigated.

3 Approach

In this section, we first describe the NMT-based QE architecture, and then describe our proposed post-training strategies.

3.1 QE Architecture

The QE architecture is shown in Figure 1. Our work follows the predictor-estimator framework. The predictor is a translation model trained with the transformer architecture on parallel data, which has learned the feature extraction ability of bilingual inputs after a long-term and large-scale pre-training. Therefore, adding only a linear layer on the top of translation model and fine-tuning with a small amount of QE data can achieve promising results.

As shown in Figure 1, the final hidden vector of the neural machine translation model corresponding to the first input token is fed into a simple linear layer to make quality prediction, which is given by:

$$HTER_{pred} = W_s^T h^{(0)} + b_0, \tag{2}$$

where $h^{(0)} \in \mathbb{R}^H$ is the hidden vector of the first input token, $W_s \in \mathbb{R}^H$ represents a weight matrix, H is the dimension of hidden states, $b_0 \in \mathbb{R}^1$ is the bias. The loss function is the mean squared error between $HTER_{pred}$ and $HTER_{true}$, which can be written as:

$$L_{QE} = MSE(HTER_{pred}, HTER_{true}) \tag{3}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

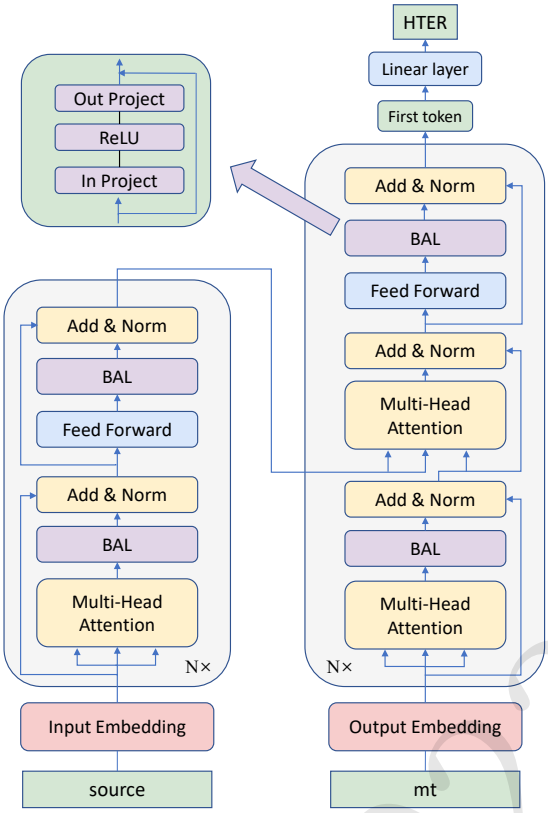


Figure 1: The illustration of the QE model. The *source* and *mt* sentence are fed into encoder and decoder respectively. The BAL is integrated after the self-attention layer and the FFN layer, respectively. In order to better adapt to QE task, the causal mask in decoder is removed.

Since the size of training dataset is relatively small, the model is easy to be over-fitted when all parameters are updated. Incorporating the insights from Wang et al. (2020), the Bottleneck Adapter Layers (BAL) (Houlsby et al., 2019) are integrated into the neural machine translation model, which alleviates the problem of overfitting by freezing the parameters of the original model. The BAL is implemented with two simple fully-connected layers, a non-linear activation and residual connections, where the hidden representations are first expanded two times and then reduced back to the original dimension.

3.2 Conditional Masked Language Modeling

The Conditional Masked Language Modeling is illustrated in Figure 2. Despite using the same architecture as the machine translation model, the CMLM utilizes a mask language modeling objective at the target side (Ghazvininejad et al., 2019). The source sentence is sent to the encoder, while some tokens are corrupted at the target side. Then the CMLM is trained to recover the corrupted target sentence.

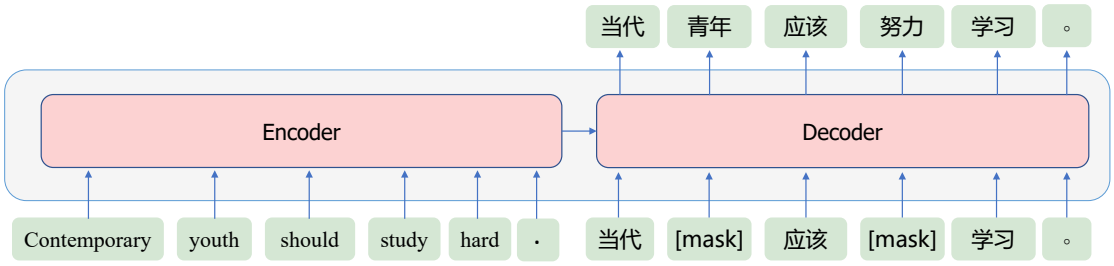


Figure 2: The illustration of the CMLM. At the target side, some tokens are replaced with [mask] symbol or random token. Note that it also needs to remove the casual mask in decoder.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

In terms of implementation, given a parallel sentence pair $< x, y >$, we generate a corrupted sentence y' with a 25% mask ratio. When the i-th token is chosen to be masked, it may be replaced with the [MASK] token 20% of the time or a random token 80% of the time. The training objective for CMLM is to maximize: $P(y_i|x, y')$, where y_i is the i-th token, x and y' represent the source sentence and the corrupted target sentence, respectively. More specifically, we reuse the parameters of the neural machine translation model instead of training the model from scratch, and the model is trained with data in the same domain as the QE data.

Translation model is a natural language generation model trained in an autoregressive manner, where each token can only pay attention to the tokens before it, and the tokens after it are masked out. On the contrary, QE task is a natural language understanding task in which each token needs to be concerned with the whole context. Through this mask-prediction task focusing on bidirectional information, the model can learn the context-based representation of the token at the target side, thereby adapting the unidirectional NMT decoder to the bidirectional prediction task.

3.3 Denoising Restoration

Inspired by Electra (Clark et al., 2020), to further mitigate the discrepancy of data quality, we apply the Denoising Restoration strategy to post-train the neural machine translation model. The model architecture is illustrated in Figure 3, which can be divided into the Noiser and the Restorer. The Noiser is used to create noisy samples, and the restorer is used to recover the noisy samples. After that, only the Restorer would be used as the predictor and the Noiser would be dropped.

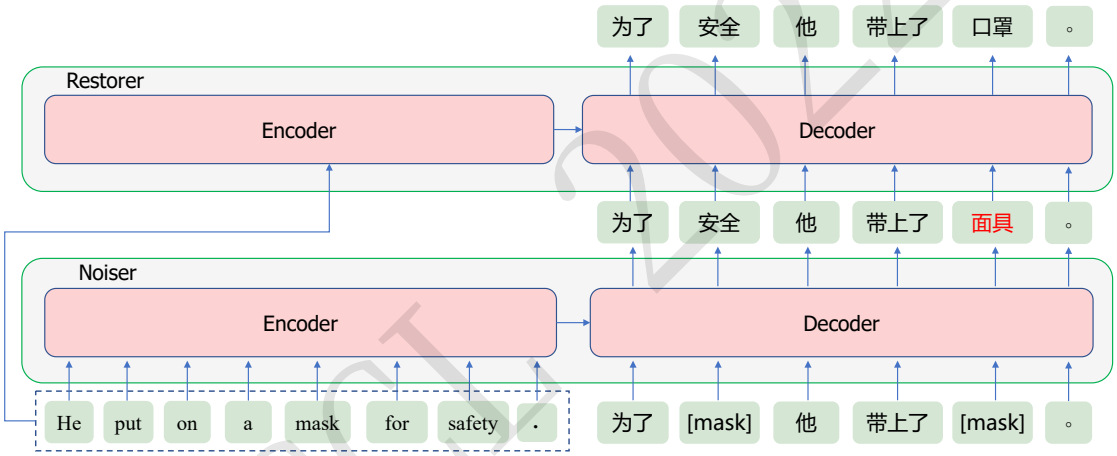


Figure 3: The Noiser-Restorer architecture.

The Noiser is first trained to introduce noise at the target side. It has the same architecture as the CMLM, the difference is that we utilize the decoding results of the to-be-evaluated NMT model as the training objective of the Noiser, where the to-be-evaluate NMT model is used to generate QE data. Specifically, given a parallel sentence pair $< x, y >$, we use the to-be-evaluted NMT model to generate the translation $\tilde{y}$ of x. Then the Noiser is trained with the new parallel sentence pair $< x, \tilde{y} >$. After the training of the Noiser, we put the Noiser and the Restorer together for training with parallel data$< x, y >$. Moreover, it is performed by dynamic mask strategy with the masked positions decided on-the-fly, where the mask ratio is same as that of the CMLM. The loss function is defined as follows:

$$L_{DR} = -\sum_{i=1}^{L} logP(l = l_i|x, \hat{y}), l_i \in \{1, 2, ..., V\}, \tag{4}$$

where L is the length of sentence, $\hat{y}$ is the sentence generated by the Noiser, V is the size of vocabulary.

The reason for introducing Noiser is that in the CMLM strategy, there is a large deviation between the sentences generated by randomly adding noise and real machine translation, which is easily detected and may limit the performance. Limited by the performance of the Noiser, it is certain that not all tokens can

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

be recovered completely and correctly. Therefore, the target sequence generated by the Noiser is noisy compared with reference translation. Meanwhile, since the Noiser utilizes a decoder with language modeling capabilities for generation, the generated sentences are more natural without obvious lexical and syntactic errors. Similarly, real machine translation noise is also natural and does not have significant lexical and syntactic errors, so the noise generated by the Noiser is closer to the real noise distribution than the noise generated by random replacement. A possible example is shown in the Figure 3.

In addition, we utilize knowledge distillation technique (Kim and Rush, 2016) in the Noiser, which is used to transfer specific patterns and knowledge among different sequence generation models. In our scenario, the decoding process of the to-be-evaluated NMT model has a fixed pattern, so the translation results obtained by decoding the source sentences with this NMT model contains the noise distribution of the to-be-evaluated NMT model. When the Noiser learns to recover a corrupted token, both training objectives and context are generated by this NMT model. Hence, the obtained Noiser would have a similar decoding space with the to-be-evaluated NMT model. Note that the Noiser could produce pseudo translations with the same length as the reference translation, which is convenient for later training.

Despite both adopting non-autoregressive training objective, the difference between CMLM and Restorer lies in the source of noise. The noise of CMLM comes from random masking, while the noise of Restorer comes from language model generation. On the one hand, the noise generated by the Noiser is more consistent with the noise distribution of the to-be-evaluated NMT model, so during the training, the Restorer can learn the modeling ability for noise data with specific distribution. On the other hand, since the noise generated by the Noiser is more natural and more difficult to identify, the obtained Restorer would have a better feature extraction ability and can identify trivial translation errors. In cases where QE needs to model machine-translated noisy data, the Restorer is more suitable for QE task.

4 Experiments

4.1 Settings

Dataset. Our experiments focus on the WMT21 QE tasks for English-to-Chinese (En-Zh) and English -to-German (En-De) directions. The QE data in each direction contains a training set of 7000, a validation set of 1000, and a test set of 1000. Besides, we also use the test set of WMT20. To train our own NMT model, we use the En-Zh and En-De parallel data released by the organizers[0], which contains roughly 20M sentence pairs for each direction after cleaning.

For the CMLM and DR, We first trained a BERT-based domain classifier and then screened 200K in-domain data from WikiMatrix for each direction[1]. The validation set we use is the training set of the QE task.

Implementation Details. All our programs are implemented with Fairseq (Ott et al., 2019). For the NMT model, we use Transformer-base architecture. We apply byte-pair-encoding (BPE) (Sennrich et al., 2015) tokenization to reduce the number of unknown tokens and set BPE steps to 32000. The learning rate is set to 5e-4. This setting is adopted in both En-Zh and En-De directions.

For the CMLM, the casual mask is removed and learning rate is set to 5e-5. For the Noiser-Restorer model, the parameters of the Noiser are frozen and the learning rate for the Restorer is 5e-5. For the Noiser, we use the decoding results of the to-be-evaluated NMT model as the training objective. We use inverse-square-root scheduler in above three models. For the QE model, it trained for 30 epochs and the hyperparameter patience is set to 5. The activation function in the BAL is ReLU. We batch sentence pairs with 4096 tokens and use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-8}$. The learning rate is 1e-4 without any scheduler.

The training data for all models is preprocessed by Fairseq based on the vocabulary and BPE vocabulary of the NMT model. For fair comparison, we tune all the hyper-parameters of our model on the validation data, and report the corresponding results for the testing set. The main metric we use is Pearson's Correlation Coefficient. We also calculate Spearman Coefficient, but it is not a ranking reference in the QE task.

[0]https://www.statmt.org/wmt21/quality-estimation-task.html
[1]http://data.statmt.org/wmt21/translation-task/WikiMatrix

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Direction	System	Test21		Test20		Avg
		Pearson↑	Spearman↑	Pearson↑	Spearman↑	
En-Zh	XLM-R(WMT-baseline)	0.282	-	-	-	0.282
	DistilBert	0.257	0.223	0.340	0.334	0.299
	XLM-R	0.265	0.219	0.323	0.318	0.294
	NMT	0.286	0.242	0.322	0.312	0.304
	NMT(finetune)	0.294	0.243	0.322	0.311	0.308
	CMLM	0.334	0.273	0.355	0.345	0.345
	DR	**0.342**	**0.275**	**0.362**	**0.353**	**0.352**
En-De	XLM-R(WMT-baseline)	0.529	-	-	-	0.529
	DistilBert	0.466	0.433	0.432	0.427	0.449
	XLM-R	0.537	0.492	**0.469**	**0.464**	0.503
	NMT	0.528	0.491	0.427	0.424	0.478
	NMT(finetune)	0.532	0.491	0.438	0.430	0.485
	CMLM	0.569	0.518	0.450	0.437	0.509
	DR	**0.577**	**0.521**	0.460	0.424	**0.519**

Table 1: Experiment results on both En-Zh and En-De directions. 'XLM-R' and 'DistilBERT' are implemented by us based on XLM-RoBERTa and DistilBERT. 'Avg' represents the average value of the pearson over two datasets. '-' indicates missing results.

4.2 Main Results

We compare our models with the following methods:

PLM-Baseline: Pre-training language models (PLM) are directly used as the predictor without integrating the BAL layer. In our experiments, DistilBert (Sanh et al., 2019) and XLM-RoBERTa (Conneau et al., 2019) were selected, and the baseline of organisers is also implemented by XLM-RoBERTa.

NMT-Baseline: An NMT model pre-trained on parallel data is used as the predictor, where NMT(finetune) is obtained by continuing to finetune on the in-domain data used by CMLM and DR.

The experimental results in both En-Zh and En-De directions are reported in Table 1. The Test20 is officially corrected, so there are no up-to-date results. As can be seen, the performance of the baseline model is relatively poor. By leveraging MLM training strategies, the CMLM can better focus on contextual information and achieves much better performance than NMT model. Moreover, the denoising restoration strategy further enhances the feature extraction ability of Restorer by introducing noise that is consistent with the distribution of NMT and outperforms the CMLM in two language pairs. This illustrates that our approaches alleviate the discrepancy between the NMT model and the QE model, thereby making the NMT model better adapted to the QE task. Combined with the official ranking, in En-Zh direction, our single model outperforms other systems except the first place (which adapt multiple ensemble techniques and data-augmentation).

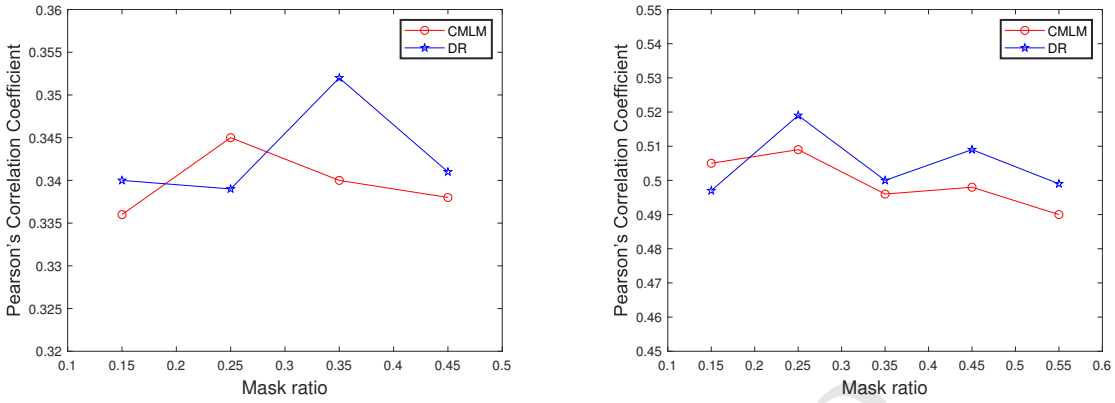
The CMLM and DR also perform better than the fine-tuned NMT model, which indicates the performance gains of them are not due to the introduction of additional datasets. Besides, the NMT-based models are more effective than PLM-Baseline in most of the comparisons, we consider that the NMT model is naturally fit for machine translation related tasks, benefiting from the knowledge of bilingual alignment.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

5 Analysis

5.1 The Impact of Mask Ratio and [MASK] symbol

During the training stage, the number of corrupted tokens may affect the performance of the model, which is related to the mask ratio. We conduct experiments to study the impact of different mask ratio and the results are illustrated in Figure 4.



(a) En-Zh direction (b) En-De direction

Figure 4: The effect of mask ratio on model performance. The values in the diagrams are the average of the pearson over two datasets.

We find that the two diagrams exhibit roughly the same pattern. The QE performance first improves, but when the mask ratio is too high, the results start to decline. This is because as the mask ratio increases, the quality of the pseudo data is gradually approaching the real machine translation, therefore the model can better model semantic alignment between the imperfect translation and source. However, when the mask ratio is too high, most of the input sentence is covered and it is too difficult for the model to restore them, thus the model can barely learn anything useful and the performance is degraded. We also observe that the performance peak of the Noiser-Restorer model in En-Zh direction comes later than that in the En-De direction. One possible reason is that the Noiser in the En-Zh direction performs better than that in the En-De direction, we will explain this in the next subsection.

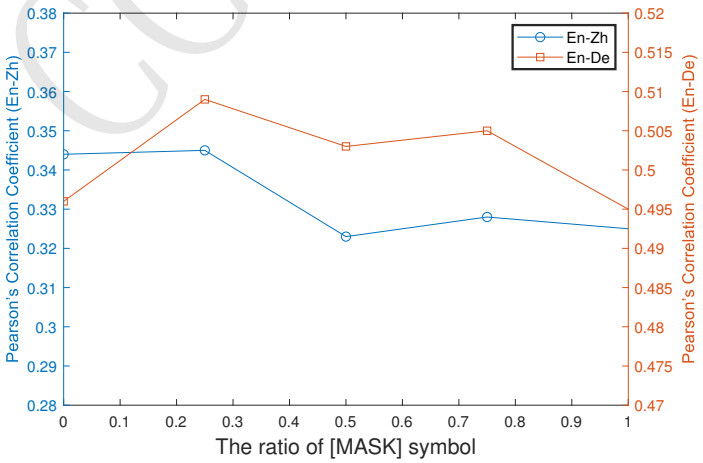


Figure 5: The impact of the [MASK] symbol.

In the CMLM strategy, among the corrupted tokens, some will be replaced with [MASK] symbol, and the others will be replaced with random tokens. We fix the mask ratio and then gradually increase the proportion of corrupted tokens replaced with [MASK] symbol to study the impact of introducing

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

[MASK] symbol. The results are presented in Figure 5. We can observe that performance get worse as the introduced [MASK] symbol increases. It may be caused by the mismatch between pre-training and fine-tuning when too many [MASK] tokens are introduced, as they never appear during the fine-tuning stage. Furthermore, using only random replacement does not give the best results, which proves that the performance improvement brought by DR is not only due to the removal of [MASK] symbol but also benefits from the introduction of natural noise close to the real machine translation.

5.2 The Impact of Knowledge Distillation

In the implementation of the Noiser, we use the decoding results of the to-be-evaluated NMT model as the training objective of the Noiser. Our motivation is to make the Noiser learn the knowledge implied by to-be-evaluated model, so as to generate sentences that is closer to the noise of real machine translation. We conduct experiments to verify the effective of this scheme, and the results are shown in Table 2.

Direction	System	Test21		Test20		Avg
		Pearson↑	MAE↓	Pearson↑	MAE↓	
En-Zh	Noiser-Restorer *w/o* kd	0.328	0.240	0.346	**0.226**	0.337
	Noiser-Restorer *w/* kd	**0.334**	**0.202**	**0.360**	0.233	**0.347**
En-De	Noiser-Restorer *w/o* kd	0.546	**0.125**	**0.449**	0.144	**0.498**
	Noiser-Restorer *w/* kd	**0.549**	0.128	0.436	**0.133**	0.493

Table 2: The comparison results of Noiser-Restorer under two strategies. '*w/* kd' and '*w/o* kd' denote with or without knowledge distillation, respectively. The 'MAE' is the Mean Absolute Error.

For a fair comparison, we extracted another dataset from WikiMatrix instead of the one used to train the Noiser for experiments. According to the experimental results, we find that the scheme plays an obvious role in the En-Zh direction, which shows that the Noiser generates pseudo data consistent with the noise distribution of the to-be-evaluated NMT model, thereby improving the performance. However, the situation is different for the En-De direction, where the results are not improved or even slightly decreased as a whole. We speculate that it may be affected by the performance of the to-be-evaluated neural machine translation model. We studied the QE dataset and came up with the results shown in the Table 3.

Direction	train	valid	test21	test20
En-Zh	0.4412	0.2839	0.2283	0.3329
En-De	0.1784	0.1830	0.1754	0.1667

Table 3: The statistical results of translation quality for QE dataset in En-Zh and En-De directions. The values in the table represent the average value of hter label.

HTER indicates human-targeted translation edit rate, and the higher HTER is, the worse the translation quality is. As can be seen in Table 3, the average value of HTER in the En-Zh direction is generally higher than that in the En-De direction. This shows that the to-be-evaluated NMT model has a better translation effect in the En-De direction, thus the machine translation is not much different from the reference translation. It is difficult for Noiser to learn the pattern contained in the NMT model, so the knowledge distillation does not play a significant role.

5.3 Different Loss Calculation Methods

Base on previous researches, there are two ways to calculate the loss:

i. Following BERT, calculating the loss only on the small subset that was masked out.

ii. Calculating the loss over all input tokens at the target side.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Direction	System	Test21		Test20		Avg
		Pearson↑	MAE↓	Pearson↑	MAE↓	
En-Zh	Only-Corrupted	0.328	0.217	0.348	**0.227**	0.338
	All-Tokens	**0.334**	**0.202**	**0.355**	0.233	**0.345**
En-De	Only-Corrupted	**0.574**	**0.125**	0.445	0.136	**0.510**
	All-Tokens	0.568	0.126	**0.450**	**0.132**	0.509

Table 4: Experimental results of different loss calculation methods in En-Zh and En-De directions. 'Only-Corrupted' and 'All-Tokens' mean the loss is calculated on the corrupted tokens and all input tokens, respectively.

We compare these two methods on the CMLM strategy and the results are shown in Table 4. In the En-Zh direction, the method of calculating the loss on all tokens is better than that only on the corrupted tokens. However, the situation is a little different in the En-De direction. We speculate that English and German belong to the same family of languages, and the prediction is relatively simple, so adding this additional information has little effect. Overall, the performance of the two methods is roughly equivalent.

6 Conclusion

When applying the pre-trained machine translation model to feature extraction for QE, there are two discrepancies between the NMT model and the QE model. One is the difference in data quality, the other is the regressive behavior of the decoder. In this paper, we propose two strategies to adapt the neural machine translation model to QE task, namely Conditional Masked Language Modeling and Denoising Restoration. The CMLM adopts a mask-prediction task at the target side, which allows the model to learn context-based representations. Moreover, the DR employs a Noiser-Restorer architecture, where the Noiser is used to generate sentences with the same noise distribution as machine translation, then the Restorer will detect and recover the introduced noise. Compared with the original NMT model, our methods bridge the gaps between the NMT model and the QE model, making it more suitable for the QE task. The experimental results verify the effectiveness of our methods.

The main work in this paper focuses on sentence-level task. Intuitively, the discrepancy also exists on word-level quality estimation when applying the pre-trained NMT model, and our strategies could function without any adaptation. Besides, enhancing the estimator can also improve QE performance, and we will leave this as our future work.

Acknowledgement

The research work descried in this paper has been supported by the National Key R&D Program of China (2020AAA0108001) and the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Yimeng Chen, Chang Su, Yingtao Zhang, Yuxia Wang, Xiang Geng, Hao Yang, Shimin Tao, Guo Jiaxin, Wang Minghan, Min Zhang, et al. 2021. Hw-tsc's participation at wmt 2021 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 890–896.

Shaika Chowdhury, Naouel Baili, and Brian Vannah. 2021. Ensemble fine-tuned mbert for translation quality estimation. *arXiv preprint arXiv:2109.03914*.

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*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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*.

Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, and Luo Si. 2019. "bilingual expert" can find translation errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6367–6374.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Hui Huang, Hui Di, Jin'an Xu, Kazushige Ouchi, and Yufeng Chen. 2020. Unsupervised machine translation quality estimation in black-box setting. In *China Conference on Machine Translation*, pages 24–36. Springer.

Hyun Kim and Jong-Hyeok Lee. 2016. Recurrent neural network based translation quality estimation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 787–792.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115.

Varvara Logacheva, Chris Hokamp, and Lucia Specia. 2016. Marmot: A toolkit for translation quality estimation at the word level. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3671–3674.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

André FT Martins, Marcin Junczys-Dowmunt, Fabio N Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

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*.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, pages 115–120.

Ziyang Wang, Hui Liu, Hexuan Chen, Kai Feng, Zeyang Wang, Bei Li, Chen Xu, Tong Xiao, and Jingbo Zhu. 2019. Niutrans submission for ccmt19 quality estimation task. In *China Conference on Machine Translation*, pages 82–92. Springer.

CL 2024

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, Yimeng Chen, et al. 2020. Hw-tsc's participation at wmt 2020 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1056–1061.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Haijiang Wu, Zixuan Wang, Qingsong Ma, Xinjie Wen, Ruichen Wang, Xiaoli Wang, Yulin Zhang, Zhipeng Yao, and Siyao Peng. 2020. Tencent submission for wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1062–1067.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 872-883, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Supervised Contrastive Learning for Cross-lingual Transfer Learning

Shuaibo Wang[1], **Hui Di**[2], **Hui Huang**[3], **Siyu Lai**[1]
Kazushige Ouchi[2], **Yufeng Chen**[1]*, **Jinan Xu**[1]
[1] School of Computer and Information Technology, Beijing Jiaotong University,
Beijing 100044, China
[2] Toshiba (China) Co., Ltd. [3] Harbin Institute of Technology
{wangshuaibo,20120374,chenyf,jaxu}@bjtu.edu.cn;
dihui@toshiba.com.cn; huanghui_hit@126.com;
kazushige.ouchi@toshiba.co.jp

Abstract

Multilingual pre-trained representations are not well-aligned by nature, which harms their performance on cross-lingual tasks. Previous methods propose to post-align the multilingual pre-trained representations by multi-view alignment or contrastive learning. However, we argue that both methods are not suitable for the cross-lingual classification objective, and in this paper we propose a simple yet effective method to better align the pre-trained representations. On the basis of cross-lingual data augmentations, we make a minor modification to the canonical contrastive loss, to remove false-negative examples which should not be contrasted. Augmentations with the same class are brought close to the anchor sample, and augmentations with different class are pushed apart. Experiment results on three cross-lingual tasks from XTREME benchmark show our method could improve the transfer performance by a large margin with no additional resource needed. We also provide in-detail analysis and comparison between different post-alignment strategies.

1 Introduction

Cross-lingual transfer learning aims to transfer the learned knowledge from a resource-rich language to a resource-lean language. The main idea of crosss-lingual transfer is to learn a shared language-invariant feature space for both languages, so that a model trained on the source language could be applied to the target language directly. Such generalization ability greatly reduces the required annotation efforts, and has urgent demand in real-world applications.

Recent multilingual pre-trained models, such as XLM-RoBERTa(XLM-R) (Conneau et al., 2020), have been demonstrated surprisingly effective in the cross-lingual scenario. By fine-tuning on labeled data in a source language, such models can generalize to other target languages even without any additional training. This has become a de-facto paradigm for cross-lingual language understanding tasks.

Despite their success in cross-lingual transfer tasks, multilingual pre-training commonly lacks explicit cross-lingual supervision, and the representations for different languages are not inherently aligned. To further improve the transferability of multilingual pre-trained representations, previous works propose different methods for cross-lingual alignment. Zheng et al. (2021) and Lai et al. (2021) propose to augment the training set with different views, and align the pre-trained representations of different languages by dragging two views closer. However, simply bringing different views closer would easily lead to representation collapse and performance degradation (Tao et al., 2021). Meanwhile, Pan et al. (2021) and Wei et al. (2021) propose to incorporate additional parallel data, and align the pre-trained representations by contrasting positive and negative samples. However, monotonously treating all random samples equally negative is inconsistent with the classification objective.

In this work, we propose a simple yet effective method to better post-align the multilingual representations on downstream tasks, which can both avoid representation collapse and meanwhile induce classification bias. With only training data for the source language available, our method performs cross-lingual fine-tuning by two steps. 1) Firstly, the original training data is augmented with different views,

* Corresponding author.
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

including code-switching, full-translation and partial-translation. All views could provide cross-lingual supervision for post-alignment. 2) Given one training sentence as the anchor point, the corresponding augmented view serves as the positive sample, and other augmented views with different labels serve as the negative samples, contrastive learning is performed by pulling positive samples together and pushing apart negative samples. This is called Supervised Contrastive Learning (SCL), and can be deemed as a cross-lingual regularizer to be combined with conventional fine-tuning.

We perform experiments on two cross-lingual classification tasks, namely XNLI (cross-lingual inference) and PAWS-X (cross-lingual paraphrase identification) (Conneau et al., 2018; Yang et al., 2019a). We compare different alignment methods, and our proposed method outperforms previous methods by a large margin, proving its effectiveness. Besides, we also apply our method on the cross-lingual retrieval task of BUCC[0] and tatoeba (Artetxe and Schwenk, 2019). We use the data from PAWS-X as supervision, and fine-tune the pretrained model by contrasting samples with their machine translation. Our proposed method again outperforms other methods by a large margin.

Detailed analysis and discussion are provided to compare different post-alignment methods for pre-trained representations, and to prove the necessity of label-supervision when performing cross-lingual contrastive learning.

2 Background

2.1 Contrastive Learning

Contrastive learning aims at maximizing the similarity between the encoded query q and its matched positive samples k^+ while keeping randomly sampled keys $\{k_0, k_1, k_2, ...\}$ far away from it. With similarity measured by a score function $s(q, k)$, InfoNCE (van den Oord et al., 2018) loss is commonly used to this end:

$$L_{ctl} = \frac{\exp(s(q, k^+))}{\exp(s(q, k^+)) + \sum_{i=1}^{n} \exp(s(q, k_i^-))}$$

Contrastive learning has led to significant improvements in various domains (He et al., 2020; Gao et al., 2021). Recently, Khosla et al. (2020) propose to incorporate label-supervision to the fine-tuning of pre-trained models, and obtain improvement on multiple datasets of the GLUE benchmark, and our work is inspired by them. However, their method is only targeted at monolingual tasks.

2.2 Cross-lingual Transfer

Cross-lingual transfer learning aims to transfer the learned knowledge from a resource-rich language to a resource-lean language. Despite recent success in large-scale language models, how to adapt models trained in high-resource languages (e.g., English) to low-resource ones still remains challenging. Several benchmarks are proposed to facilitate the progress of cross-lingual transfer learning (Hu et al., 2020; Liang et al., 2020), where models are fine-tuned on English training set and directly evaluated on other languages.

Recently, several pre-trained multilingual language models are proposed for cross-lingual transfer, including multilingual BERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), and XLM-R (Conneau et al., 2020). The models work by pre-training multilingual representations using some form of language modeling, and have made outstanding progress in cross-lingual tasks. However, most existing models use only single-language input for language model finetuning, without leveraging the intrinsic cross-lingual alignment. Therefore, several methods have been proposed to post-align the pre-trained representations, by introducing some form of cross-lingual supervision. Cao et al. (2020) and Dou et al. (2021) propose to generate word alignment information from parallel data, and push the aligned words in parallel data to have similar representations. Pan et al. (2021), Wang et al. (2021) and Wei et al. (2021) propose to utilize contrastive learning for post-alignment by contrasting positive and negative samples, where positive samples are parallel to each other while negative samples are randomly picked.

[0]https://comparable.limsi.fr/bucc2017/

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Zheng et al. (2021) and Lai et al. (2021) propose to augment the training set with different views, and align the representations by dragging two views close to each other. In a nutshell, despite all variations of supervision in both sentence or word-level, from both parallel data or automatically crafted data, the alignment must be performed by inter-lingual comparing, either by bringing two representations closer or contrasting a representation with random sampled representations. However, we argue that both methods are in contradiction with the cross-lingual classification objective, for which we will give detailed analysis in Section 3.2.

3 Approach

In this section, we first introduce the three cross-lingual data augmentation methods. Based on that, we propose three paradigms to post-align the multilingual representations, and provide theoretical analysis and comparison for them.

3.1 Cross-lingual Data Augmentation

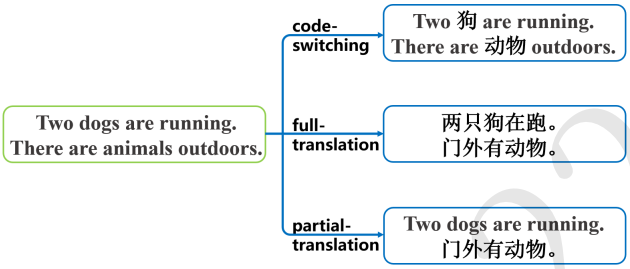


Figure 1: Different cross-lingual data-augmentation methods. Here we use sentence-pair classification as an example, therefore each sample contains two sentences.

In this work, we do not want to incorporate any parallel data (which is inaccessible in a lot of scenarios, especially for a resource-lean language that we want to transfer to). Therefore, to provide cross-lingual supervision for post-alignment, we propose three data augmentation methods:

1. Code-switching: Following Qin et al. (2020), we randomly select words in the original text in the source language and replace them with target language words in the bilingual dictionaries, to generate code-switched data. The intuition is to help the model automatically and implicitly align the replaced word vectors in the source and all target languages by mixing their context information, and the switched words can serve as anchor point for aligning two representation space.

2. Full-translation: Machine translation has been proved to be an effective data augmentation strategy under the cross-lingual scenario. It can provide translations almost in-line-with human performance, and therefore serves as a strong baseline for cross-lingual tasks.

3. Partial-translation: This method simply takes a portion of input and replace it with its translation in another language. According to Singh et al. (2019), partial-translation could provide inter-language information, where the non-translated portion serves as the anchor point. This is somehow akin to code-switching, and can be deemed as code-switching in segment-level.

The three methods can provide cross-lingual supervision in a coarse-to-fine manner (sentence-level, segment-level, word-level). We perform all the three methods to the whole training set. Each training sample could be code-switched multiple times with different results, and each task contains translation into multiple languages, leading to multiple views from a cross-lingual perspective.

3.2 Cross-lingual Alignment: What do we want?

Many experiments (Cao et al., 2020; Kulshreshtha et al., 2020) suggest that to achieve reasonable performance in the cross-lingual setup, the source and the target languages need to share similar representations. However, current multilingual pre-trained models are commonly pre-trained without explicit cross-lingual supervision. Therefore, the cross-lingual transfer performance can be further improved by additional cross-lingual alignment.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Given the training sample in source language and its cross-lingual augmentations, previous methods perform cross-lingual alignment in two different trends: Multi-view Alignment (Zheng et al., 2021; Lai et al., 2021) or Contrastive Learning (Wei et al., 2021; Pan et al., 2021). The multi-view alignment is to bring the sample and the corresponding augmentation together, while the contrastive learning is to bring these two together while pushing apart other random sampled augmentations. Suppose we are working with a batch of training examples of size N, $\{x_i, y_i\}, i = 1, ...N$, x_i denotes the training sample, while y_i is the label, the two different objectives can be denoted as follows:

$$L_{MVA} = -s(\Phi(x_i), \Phi(\hat{x}_i))$$

$$L_{CL} = -\log \frac{s(\Phi(x_i), \Phi(\hat{x}_i))}{s(\Phi(x_i), \Phi(\hat{x}_i)) + \sum_{j=1}^{N} \mathbb{I}_{j \neq i} s(\Phi(x_i), \Phi(\hat{x}_j))}$$

where $\Phi(\cdot) \in R_d$ denotes the $L2$-normalized embedding of the final encoder hidden layer before the softmax projection, and $\hat{x}_i$ denotes the augmented view (code-switching, full-translation, partial-translation, etc.), and s(q, k) denotes the similarity measure (cosine similarity, KL divergence, etc.). MVA is short for multi-view alignment, and CL is short for contrastive learning.

Since in vanilla contrastive learning, the similarity function is normally in the form of exponential, therefore L_{CL} can be detached into two terms:

$$L_{CL} = \underbrace{-s(\Phi(x_i), \Phi(\hat{x}_i))}_{alignment} + \underbrace{\log(e^{s(\Phi(x_i), \Phi(\hat{x}_i))} + e^{\sum_{j=1}^{N} \mathbb{I}_{j \neq i} s(\Phi(x_i), \Phi(\hat{x}_j))})}_{uniformity}$$

where the first term optimize the alignment of representation space, and the second term optimize the uniformity, as discussed in Wang et al. (2020). According to Gao et al. (2021), let W be the sentence embedding matrix corresponding to x_i, i.e., the i-th row of W is $\Phi(x_i)$, optimizing the $uniformity$ term essentially minimizes an upper bound of the summation of all elements in $WW^\top$, and inherently "flatten" the singular spectrum of the embedding space.

However, the $uniformity$ term in L_{CL} is in contradiction with the classification objective. In classification task, we want the representations to be clustered in several bunches, each bunch corresponds to a class. Or else to say, we want the representations to be inductively biased, rather than uniformly distributed.

On the other hand, it is obvious that the multi-view alignment objective L_{MVA} is to solely maximize the alignment. This would easily lead to representation collapse, since simply projecting all representations to one data point could easily reduce the $alignment$ term to zero. Contrast between samples is necessary to avoid collapse, and simply removing the $uniformity$ term is also not what we want.

3.3 Better Alignment with SCL

To better perform cross-lingual alignment, we propose to introduce label information to the vanilla contrastive learning, named as Supervised Contrastive Learning (SCL):

$$L_{SCL} = -\log \frac{s(\Phi(x_i), \Phi(\hat{x}_i))}{s(\Phi(x_i), \Phi(\hat{x}_i)) + \sum_{j=1}^{N} \mathbb{I}_{y_j \neq y_i} s(\Phi(x_i), \Phi(\hat{x}_j))}$$

More concretely, our modification is based on InfoNCE loss (van den Oord et al., 2018), therefore the similarity function is written as:

$$s(\Phi(x_i), \Phi(\hat{x}_i)) = e^{cos(\Phi(x_i), \Phi(\hat{x}_i))/\tau}$$

where $\tau > 0$ is an adjustable scalar temperature parameter that controls the separation of classes. Empirical observations show that both $L2$-normalization of the encoded embedding representations (which is incorporated in the calculation of cosine similarity) and an adjustable scalar temperature parameter τ

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

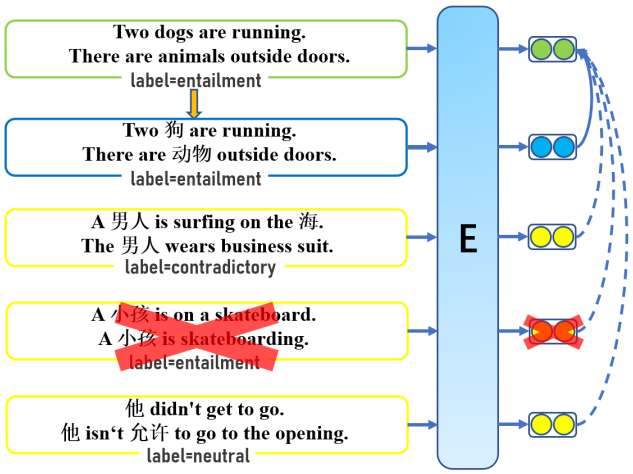


Figure 2: Our proposed supervised contrastive learning. Solid line connects positive pairs while dashed line connects negative pairs. Notice the false negative sample is removed.

improve performance. This can serve as a cross-lingual regularization term and be combined with the canonical classification loss:

$$L_{CE} = y_i \cdot \log(1 - \hat{y}_i) + \hat{y}_i \cdot \log(1 - y_i)$$

$$L_{total} = L_{CE} + \lambda L_{SCL}$$

where λ is a scalar weighting hyperparameter that we tune for each downstream task.

The core idea is simple, just to remove the negative samples which belong to the same class with the anchor point. Therefore, only samples from different classes would be pulled apart. The modified $uniformity$ term is not to unify the representations any more, but to push the multilingual decision clusters apart from each other.

This loss can be applied to a variety of encoders, not just limited to multilingual pre-trained transformer-like models. The loss is meant to capture similarities between examples of the same class and contrast them with examples from other classes. This is in line with the objective of cross-lingual alignment. When we are doing cross-lingual alignment, what we really want to do is to transfer the representation for a certain class to another language, rather than to learn a unified multilingual representation space.

4 Experiments

4.1 Data Preparation

In this work, we mainly focus on sentence-level tasks, for which the aggregated representation is easily accessible. We conduct experiments on two cross-lingual sentence-pair classification tasks: natural language inference and paraphrase identification. The Cross-lingual Natural Language Inference corpus (XNLI) (Conneau et al., 2018) asks whether a premise sentence entails, contradicts, or is neutral toward a hypothesis sentence. The Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X) (Yang et al., 2019a) dataset requires to determine whether two sentences are paraphrases. Both tasks are from XTREME benchmark (Hu et al., 2020). Despite their intrinsic different objective, both tasks can be formalized as sentence-pair classification tasks. For both tasks, the training set is in English, while human annotated development and test sets are available for a bunch of different languages. The model is evaluated on the test data of the task in the target languages.

For cross-lingual data augmentation, we first randomly sample a target language and then adapt the generating method for each data augmentation method. Since XNLI covers more target languages than PAWS-X, we set $t_f = 2, t_p = 2, t_c = 1$ in XNLI, and $t_f = 1, t_p = 1, t_c = 1$ in PAWS-X, where t_f, t_p

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
cross-lingual transfer (Models are fine-tuned on English training data only.)																
InfoXLM	86.4	74.2	79.3	79.3	77.8	79.3	80.3	72.2	77.6	67.5	74.6	75.6	67.3	77.1	77.0	76.5
HITCL	86.3	74.8	80.6	79.5	78.9	81.3	80.5	73.1	79.0	69.9	75.7	75.4	69.7	77.4	77.6	77.3
xTune*	84.7	76.7	81.0	79.9	79.4	81.6	80.5	75.6	77.9	68.4	75.4	77.2	72.2	78.1	77.4	77.7
XLMR-base	84.8	72.7	78.8	77.9	76.5	79.8	78.9	72.2	76.5	66.8	73.9	73.7	68.0	76.8	75.4	75.5
MVA	85.0	75.0	79.1	78.2	78.1	79.7	79.1	72.5	76.8	68.9	75.5	74.5	70.0	76.9	77.4	76.5
CL	84.4	75.5	80.0	79.3	78.7	80.4	79.8	74.1	78.3	71.5	76.1	76.0	71.0	78.2	77.8	77.4
SCL	**86.3**	**77.8**	**81.7**	**81.3**	**80.6**	**82.7**	**81.8**	**76.3**	**80.4**	**73.8**	**78.9**	**78.1**	**73.1**	**80.5**	**80.2**	**79.6**
translate-train (Models are fine-tuned on both English data and its translations.)																
InfoXLM	86.5	78.9	82.4	82.3	81.3	83.0	82.6	77.8	80.6	73.3	78.9	79.5	71.6	81.0	80.7	80.0
HITCL	86.5	78.1	82.2	80.8	81.6	83.2	82.3	76.7	81.3	73.8	78.6	80.5	73.9	80.4	80.7	80.0
xTune*	86.6	79.7	82.7	82.2	81.9	83.1	82.3	78.9	80.9	75.7	78.4	79.8	75.3	80.5	80.0	80.5
XLMR-base	84.3	76.9	80.3	79.8	79.1	81.5	80.3	75.3	78.1	72.9	77.1	77.4	70.8	79.8	79.7	78.2
MVA	85.4	78.5	81.5	81.8	**80.6**	82.3	81.0	77.3	79.9	74.1	**78.8**	78.2	73.5	80.2	80.2	79.6
CL	85.9	77.2	81.6	80.5	80.0	81.7	81.5	76.5	80.3	73.5	77.8	78.2	72.5	79.9	79.9	79.1
SCL	**86.4**	**78.8**	**82.0**	**82.0**	80.5	**82.9**	**82.3**	**77.3**	**80.5**	**74.5**	78.6	**79.7**	**74.2**	**80.9**	**80.3**	**80.1**

Table 1: Experiment results on XNLI. Results with * are reimplemented by us with their released codes. InfoXLM (Chi et al., 2021a) and HITCL (Wei et al., 2021) use contrastive learning while xTune (Zheng et al., 2021) uses multi-view alignment. Notice xTune uses more augmentation data and model ensemble compared to us.

and t_c respectively represent the number of samples generated by full-translation, partial translation and code-switching for each training data. Therefore, each training batch contains $6 \times batch\_size$ sentence pairs in XNLI and $4 \times batch\_size$ sentence pairs in PAWS-X. The code-switching ratio r_c is set as 0.75 in XNLI and 0.5 in PAWS-X. For cross-lingual retrieval tasks mentioned below, each training pair from PAWS-X is detached into two sentences when feeding to the model, and we do not incorporate code-switching as data augmentation.

4.2 Setup

For sentence pair classification tasks of XNLI and PAWS-X, we concatenate the input as the formation defined by XLM-R:

```
[s] input1 [\s] input2 [\s]
```

and we use the final hidden layer corresponding to `[s]` as aggregated representation. For retrieval tasks of BUCC and tatoeba, we perform alignment on the same aggregated representation, but the retrieval is performed on the averaged pooled eighth layer, following the related works (Chi et al., 2021b; Chi et al., 2021a). Adam optimizer is applied with a learning rate of 5e-6. Batch size is set as 24 for XNLI, 36 for PAWS-X and 48 for retrieval.

We evaluate a number of strong baselines and the three post-align strategies discussed in the former section. The baseline is trained with cross-entropy loss with no alignment term serving as cross-lingual regularizer. Then we create cross-lingual augmentations with different methods, and apply different alignment strategies. Three groups of augmentations (full-translation, partial translation, code-switching) are mixed together. The bilingual dictionaries we used for code-switch substitution are from MUSE (Lample et al., 2018). For languages that cannot be found in MUSE, we ignore these languages since other bilingual dictionaries might be of poorer quality. The machine translated training set is taken from the XTREME repository, which is obtained by an in-house translation model from Google.

We mainly compare with models that learn multilingual contextual representations as they have achieved state-of-the-art results on cross-lingual tasks. All cross-lingual alignment strategies are applied to pre-trained XLM-R-base. Following the trend of Hu et al. (2020), we mainly consider the following two scenarios:

Cross-lingual Transfer: the models are fine-tuned on English training data, and directly evaluated on different target languages.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	en	de	es	fr	ja	ko	zh	avg
cross-lingual transfer (Models are fine-tuned on English training data only.)								
InfoXLM*	94.7	89.7	90.1	90.4	78.7	79.0	82.3	86.4
xTune*	93.7	90.2	89.9	90.4	82.6	81.9	84.3	87.6
XLMR-base	94.5	88.4	89.4	89.3	76.0	77.2	82.6	85.3
MVA	95.0	89.1	90.9	90.6	79.5	81.1	83.7	87.1
CL	94.6	89.8	91.3	90.9	78.9	80.0	82.8	86.9
SCL	**95.3**	**91.3**	**91.8**	**91.7**	**83.2**	**84.5**	**85.7**	**89.0**
translate-train (Models are fine-tuned on both English data and its translations.)								
InfoXLM*	94.5	90.5	91.6	91.7	84.4	83.9	85.8	88.9
xTune*	93.9	90.4	90.9	91.7	85.6	86.8	86.6	89.4
XLMR-base	95.0	89.8	91.8	91.6	81.2	84.3	84.4	88.3
MVA	95.3	90.9	92.0	91.8	83.1	83.6	85.3	88.8
CL	95.4	90.2	92.1	91.4	81.7	84.0	85.3	88.6
SCL	**95.5**	**91.4**	**92.3**	**92.3**	**83.2**	**85.0**	**87.2**	**89.5**

Table 2: Experiment results on PAWS-X. Results with * are reimplemented by us with their released codes.

Translate-train: the models are fine-tuned on the concatenation of English training data and its translation to all target languages. Translate-train is normally a strong baseline for cross-lingual transfer tasks. For classification tasks, it is straightforward that the translation should be assigned with the same label.

In both settings, the alignment term is combined with the canonical cross-entropy loss to be back-propagated together. We use KL Divergence as the similarity measure for multi-view alignment. For contrastive learning, we only consider in-batch negative samples, leaving more complicated methods (e.g. to maintain a memory bank for negative samples (He et al., 2020)) to the future.

4.3 Main Results

As shown in Table 1 and Table 2, we can see that our proposed method could improve the cross-lingual transfer results of pre-trained XLM-R by a large margin. Our method is especially effective in zero-shot setting, where the accuracy is improved by 4.1 points on XNLI and 3.7 points on PAWS-X. Our method can also achieve significant improvement in translate-train setting, where the accuracy is improved by 1.9 points on XNLI and 1.2 points on PAWS-X. Results are consistently improved among all languages, despite their relation with English close or not.

The results of multi-view alignment and vanilla contrastive learning, despite using the same augmentation data, underperform our method on both datasets. This proves the pre-trained representations are better aligned according to the label information after SCL. Different representations, despite belonging to different languages, are projected to the same cluster if they belong to the same class.

SCL is a simple yet effective framework to align the pre-trained multilingual representations on downstream tasks. Cross-lingual signals can be obtained by machine translation or bilingual dictionary, therefore no extra human annotation is needed. While previous works also propose other methods to align the pre-trained representations, the results in Table 1 and 2 prove the superiority of our method.

5 Analysis and Discussion

5.1 Different Augmentations

In this section, we want to explore the influence of different cross-lingual augmentations. We apply different groups of augmentations under the zero-shot setting, and compare the results on different tasks.

As shown in Table 3, we can see that the results of full translation and partial translation are better than code-switching. We think it is because the information provided by code-switching is comparably sparse, only a few anchor words covered by the bilingual dictionary. On the other side, well-trained machine translation system can provide fluent and accurate translation, therefore the multilingual representation can be better aligned. We can also tell that the results of our proposed method outperform the counterparts again on both datasets, proving its superiority.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

AugData	Method	XNLI en	XNLI avg	PAWS-X en	PAWS-X avg
None	XLMR	84.9	75.5	94.5	85.3
full-trans	MVA	85.2	76.6	94.9	87.1
	CL	85.0	77.9	94.9	87.2
	SCL	85.6	**79.2**	95.3	**88.7**
partial-trans	MVA	83.7	75.7	95.2	86.5
	CL	84.5	76.9	94.9	86.6
	SCL	85.3	78.4	95.3	88.1
code-switch	MVA	85.3	76.4	94.7	86.1
	CL	84.5	76.1	95.2	86.5
	SCL	84.8	76.2	95.1	87.2

Table 3: Experiment results on XNLI and PAWS-X based on different cross-lingual data augmentations, including full-translation, partial translation, and code-switching. For each group of data, we apply all three post-align methods.

similarity measure	lambda	XNLI en	XNLI avg
KLDiv	1	85.19	76.64
	10	85.05	76.71
Symmetric KLDiv	1	84.67	76.17
	10	83.85	76.20
Cosine Similarity	1	83.03	75.16
	10	84.05	76.38
Mean-Square Error	1	83.95	75.37
	10	84.35	76.58

Table 4: Experiment results of different similarity measures and loss weight λ on XNLI. Here we only use the augmentation of full-translation, and the results is in cross-lingual setting. We do not experiment on PAWS-X due to resource limitation.

5.2 Similarity Measure

The similarity measure in L_{MVA} has many alternatives. Previous studies on multi-view learning propose all kinds of measures (Yang et al., 2019b), such as Cosine-Similarity, Mean-Square Error, Kullback-Leibler Divergence and Symmetric Kullback-Leibler Divergence. Suppose we are dealing with an input x and its augmentation $\hat{x}$, different similarity measures can be denoted as:

$$L_{KLDiv} = \Phi(x) log \frac{\Phi(\hat{x})}{\Phi(x)}$$

$$L_{SymKLdiv} = \Phi(x) log \frac{\Phi(\hat{x})}{\Phi(x)} + \Phi(\hat{x}) log \frac{\Phi(x)}{\Phi(\hat{x})}$$

$$L_{cosine} = \frac{\Phi(x) \cdot \Phi(\hat{x})}{||\Phi(x)|| \, ||\Phi(\hat{x})||}$$

$$L_{MSE} = ||\Phi(x) - \Phi(\hat{x})||^2$$

where $\Phi(\cdot)$ denotes the $L2$-normalized aggregated representation. We experiment different similarity measures on the multi-view alignment objective, in combination with different loss weight λ, and the results are shown in Table 4. Surprisingly, we do not see a clear difference between different measures, and in the end we decide to use cosine similarity with $\lambda = 10$ in all experiments. On the other hand, λ is set as 1 for contrastive learning.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

setting	temp	XNLI en	XNLI avg	PAWS-X en	PAWS-X avg
cross-transfer	1.0	85.6	79.2	95.3	88.7
	0.3	85.2	79.1	94.8	88.7
	0.1	85.8	79.2	95.3	88.2
translate-train	1.0	86.4	79.8	95.4	89.0
	0.3	85.8	79.8	95.4	89.1
	0.1	85.9	79.5	95.3	89.2

Table 5: Experiment results of different contrast temperatures on XNLI and PAWS-X. Here we only use the augmentation of full-translation, and the results are based on supervised contrastive learning.

5.3 Contrast Temperature

Previous empirical observations show that an adjustable scalar temperature parameter τ can improve the performance of contrastive learning (Wang and Isola, 2020; He et al., 2020). Lower temperature increases the influence of examples that are harder to separate, effectively creating harder negatives. However, we do not find such a pattern in our experiments, as shown in Table 5, and finally we decide to set the temperature τ as 1.0 in all experiments.

5.4 SCL for Cross-lingual Retrieval

To further prove the importance of label information in cross-lingual fine-tuning, we also apply the alignment methods on cross-lingual sentence retrieval tasks. We experiment on two datasets, BUCC[1] and tatoeba (Artetxe and Schwenk, 2019). Both datasets aim at extracting parallel sentences from a comparable corpus between English and other languages, with BUCC covering 4 languages and tatoeba covering more than 100 languages. To compare with previous works, we only use a subset of tatoeba (33 languages) in this work.

The pre-trained multilingual models are able to provide language-deterministic representations by nature. Previous works directly calculate the similarity of different sentences by representations from the pre-trained model, to determine whether two sentences are parallel or not (Hu et al., 2020; Chi et al., 2021b; Chi et al., 2021a). In this work, we propose to use the data of paraphrase identification, including the original training sentence pairs and their translations to six languages, to post-align the pre-trained representations.

We compare the previously proposed three strategies to post-align the pre-trained representations. Since we are dealing with retrieval task, the sentence pair from two different languages are encoded separately by the pre-trained XLM-R. We apply the alignment training methods on the aggregated representation. For multi-view alignment, only two translation pairs are pulled closer to each other. For vanilla contrastive learning, we treat all translation pairs as positive while the others as negative. For our proposed SCL, both translation pairs and translation with paraphrasing pairs are deemed as positive, while the others are deemed as negative, as denoted by the following formula:

$$L_{SCL} = -\sum_{j=1}^{N} \mathbb{I}_{y_{ij}=1} \log \frac{s(\Phi(x_i), \Phi(\hat{x}_j))}{s(\Phi(x_i), \Phi(\hat{x}_j)) + \sum_{k=1}^{N} \mathbb{I}_{y_{ik}\neq 1} s(\Phi(x_i), \Phi(\hat{x}_k))}$$

where x_i is a training sample and $\hat{x_i}$ is its translation, and $y_{ij} = 1$ denotes x_i and x_j are a paraphrase pair. After the fine-tuning stage, following previous work, we utilize the average pooled hidden representation of the eighth layer of the pre-trained model as the sentence representation.

As shown in Table 6 and Table 7, paraphrase identification dataset with translated augmentation, despite containing noise generated by the MT model, can provide cross-lingual signal to post-align the multilingual representations. Vanilla contrastive learning can perform alignment space by pulling translation pairs together and pushing translation pairs apart, but paraphrase pairs also possess the same semantics, and should not be contrasted as negative samples. After introducing label information into contrast, the

[1] https://comparable.limsi.fr/bucc2017

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	en-de	en-fr	en-ru	en-zh	avg
mBERT*	62.5	62.6	51.8	50.0	56.7
XLM*	56.3	63.9	60.6	46.6	56.8
XLMR-large*	67.6	66.5	73.5	56.7	66.0
XLMR-base	82.68	74.85	82.08	64.09	75.93
MVA	43.92	26.24	38.71	7.58	29.11
CL	87.22	79.93	86.88	78.83	83.21
SCL	**88.82**	**81.88**	**88.01**	**82.47**	**85.29**

Table 6: Experiment results on BUCC2018 test set. Results with * are released by XTREME(Hu et al., 2020). We apply different post-align strategies on pre-trained XLM-RoBERTa-base model using the training set of PAWS-X with translation augmentation.

Method	en-xx	xx-en
XLMR-base*	55.50	53.40
XLM-E(Chi et al., 2021b)	65.00	62.30
InfoXLM(Chi et al., 2021a)	68.62	67.29
XLMR-base	55.60	53.49
MVA	28.00	27.79
CL	78.80	77.87
SCL	**80.41**	**80.84**

Table 7: Experiment results on tatoeba. Result with $*$ is released by (Chi et al., 2021b). xx denotes the 33 languages as experimented in (Chi et al., 2021a) and (Chi et al., 2021b), and we release the averaged accuracy in both directions.

retrieval accuracy is further improved by 2-3 points. On the contrary, multi-view alignment would lead to representation collapse and cannot converge at all. This is in line with our previous analysis.

6 Conclusion

In this paper, we propose to improve cross-lingual fine-tuning with supervised contrastive learning. Cross-lingual supervision is created by augmenting the training set, and different methods to post-align the multilingual pre-trained representation are compared. We propose to incorporate label-information when performing cross-lingual contrastive fine-tuning, and outperforms previous methods by a large margin on four cross-lingual transfer benchmark datasets.

Canonical cross-entropy has many intrinsic problems, especial when performing transfer learning tasks, and contrastive learning can be a decent supplementary. By alleviating the commonality and differences between different examples, representations are efficiently transferred from one domain or language to another. In the future, we would explore the application of supervised contrastive learning on other transfer learning tasks, including token-level classification, language generation, cross-domain transfer, etc.

Acknowledgement

This research work is supported by the National Key R&D Program of China (2019YFB1405200), the National Nature Science Foundation of China (No. 61976016, 61976015 and 61876198) 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

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proc. of ICLR*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proc. of ACL*.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021b. XLM-E: cross-lingual language model pre-training via ELECTRA. *CoRR*.

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

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 *Proc. of ACL*.

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

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proc. of ACL*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proc. of EMNLP*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. of CVPR*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proc. of NeurIPS*.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Proc. of ACL*.

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 *Proc. of ACL*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proc. of ICLR*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proc. of ACL*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proc. of IJCAI*.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *CoRR*.

Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. 2021. Exploring the equivalence of siamese self-supervised learning via A unified gradient framework. *CoRR*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. of ICML*.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Liang Wang, Wei Zhao, and Jingming Liu. 2021. Aligning cross-lingual sentence representations with dual momentum contrast. In *Proc. of EMNLP*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *Proc. of ICLR*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019b. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proc. of ACL*.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proc. of ACL*.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 884-895, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Interactive Mongolian Question Answer Matching Model Based on Attention Mechanism in the Law Domain

Yutao Peng, Weihua Wang✉, Feilong Bao
College of Computer Science, Inner Mongolia University, China
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, China
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, China
yutao.peng@mail.imu.edu.cn
{wangwh,csfeilong}@imu.edu.cn

Abstract

Mongolian question answer matching task is challenging, since Mongolian is a kind of low-resource language and its complex morphological structures lead to data sparsity. In this work, we propose an Interactive Mongolian Question Answer Matching Model (IMQAMM) based on attention mechanism for Mongolian question answering system. The key parts of the model are interactive information enhancement and max-mean pooling matching. Interactive information enhancement contains sequence enhancement and multi-cast attention. Sequence enhancement aims to provide a subsequent encoder with an enhanced sequence representation, and multi-cast attention is designed to generate scalar features through multiple attention mechanisms. Max-Mean pooling matching is to obtain the matching vectors for aggregation. Moreover, we introduce Mongolian morpheme representation to better learn the semantic feature. The model experimented on the Mongolian corpus, which contains question-answer pairs of various categories in the law domain. Experimental results demonstrate that our proposed Mongolian question answer matching model significantly outperforms baseline models.

1 Introduction

Question answer matching is used to identify the relationship between the question-answer pairs, and it is one of the application scenarios of text matching. Text matching is an important fundamental technology in Natural Language Processing (NLP) and can be applied to a large number of NLP tasks, such as Information Retrieval (IR), Natural Language Inference (NLI), question answering (QA) system, dialogue system, etc. For the tasks of Information Retrieval, text matching is utilized to compute the relevance between queries and documents to select the relevant documents (Huang et al., 2013). For the tasks of Natural Language Inference, text matching is employed to judge whether the premise can infer the hypothesis (Bowman et al., 2015). And for the question answering tasks, text matching is applied to pick the answers that are most relevant to a given question (Tan et al., 2016).

With the development of deep learning, text matching methods with neural network are increasingly emerging. These methods can be divided into two types—representation-based match and interaction-based match. The first type is representation-based match (Huang et al., 2013; Shen et al., 2014; Palangi et al., 2014), which is focused on modeling the representations of the two sentences, so that they are encoded into semantic vectors in the same embedding space. The second type is interaction-based match (Chen et al., 2017; Tay et al., 2017; Wang et al., 2017), which is targeted at interacting with each information between sentence pairs to improve the process of representation learning. Interaction-based match performs better than representation-based match, because representation-based match lacks a comparison of lexical and syntactic information between sentence pairs, while interaction-based match can take advantage of the interactive information across sentence pairs to enhance their own representations. Therefore, interactive matching methods are currently the mainstreaming methods of text matching.

However, the development of Mongolian question answering system is relatively slow, and there are few studies about it. The first reason for the slow development is that Mongolian is a kind of low-resource language. It lacks public labeled corpus. The second reason is the data-sparse problem caused

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896-907, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

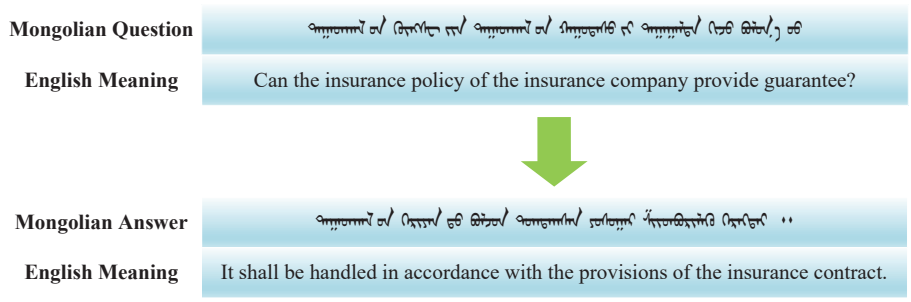


Figure 1: A Mongolian question-answer pair example.

by complex Mongolian morphological structures. Mongolian is an agglutinative language and its root can be followed by different suffixes, which is different from Chinese and English.

In this paper, we construct a Mongolian question answering data set in the law domain. An example is shown in Figure 1, there are two Mongolian sentences, including the question and its corresponding answer. Our task is to judge whether the answer corresponds to the question. In order to solve the problem of insufficient information enhancement of previous interaction-based matching methods, we propose an Interactive Mongolian Question Answer Matching Model (IMQAMM), which combines interactive information enhancement and max-mean pooling matching. Interactive information enhancement concatenates a series of feature vectors to get the enhanced sequence representation and adopts a compression function to reduce feature vectors to scalars based on multiple attention mechanisms for the issue of data-sparse. Max-Mean pooling matching is to compute the maximum and average cosine similarities corresponding to each morpheme representation.

The paper is organized as follows: Section 2 introduces the related work. Section 3 presents the details of our Mongolian question answer matching model. Section 4 shows our experimental setup and results. Section 5 gives the conclusion and the future work.

2 Related Work

Text semantic matching is the core of retrieval question answering system. In recent years, there has been a lot of research on text semantic matching, which has driven the development of question answering systems. Early work on text semantic matching was based on relatively simple models to compare the sentence pairs. Most of these work was based on sentence encoding methods. Huang et al. (2013) proposed Deep Structured Semantic Models (DSSM), which projected queries and documents into a common low-dimensional space to get similarities between them. Since the CNN has the characteristics of local perception and parameter sharing, Shen et al. (2014) improved the DSSM model and proposed the Convolutional Latent Semantic Model (CLSM). To tackle the limitations of RNN in capturing contextual dependencies, Palangi et al. (2014) presented the Long Short Term Memory DSSM (LSTM-DSSM). Mueller and Thyagarajan (2016) proposed the Manhattan LSTM (MaLSTM) model, which utilized the siamese LSTM to compute the Manhattan distance between sentence pairs. Shen et al. (2018) proposed Directional Self-Attention Network (DiSAN) without any RNN or CNN for sentence encoding. Yang and Kao (2020) proposed a Siamese sentence encoder that would not propagate any interactive information between sentence pairs.The matching methods based on attention networks have gradually developed (Tan et al., 2016; Yin et al., 2016; Kim et al., 2019). Tan et al. (2016) used the attentive LSTM to match questions and passage answers with question information. Yin et al. (2016) explored computing the attention matrix before and after convolution to select correct answers for a question. Kim et al. (2019) used densely-connected recurrent neural network and concatenated the co-attentive features in multiple layers for semantic sentence matching.

Compare-Aggregate networks are very popular in different tasks (Wang and Jiang, 2017; Wang et al., 2017; Chen et al., 2017; Tay et al., 2017; Tay et al., 2018). Wang and Jiang (2017) presented a compare-aggregate model that performed word-level matching by element-wise multiplication or subtraction and

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896-907, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

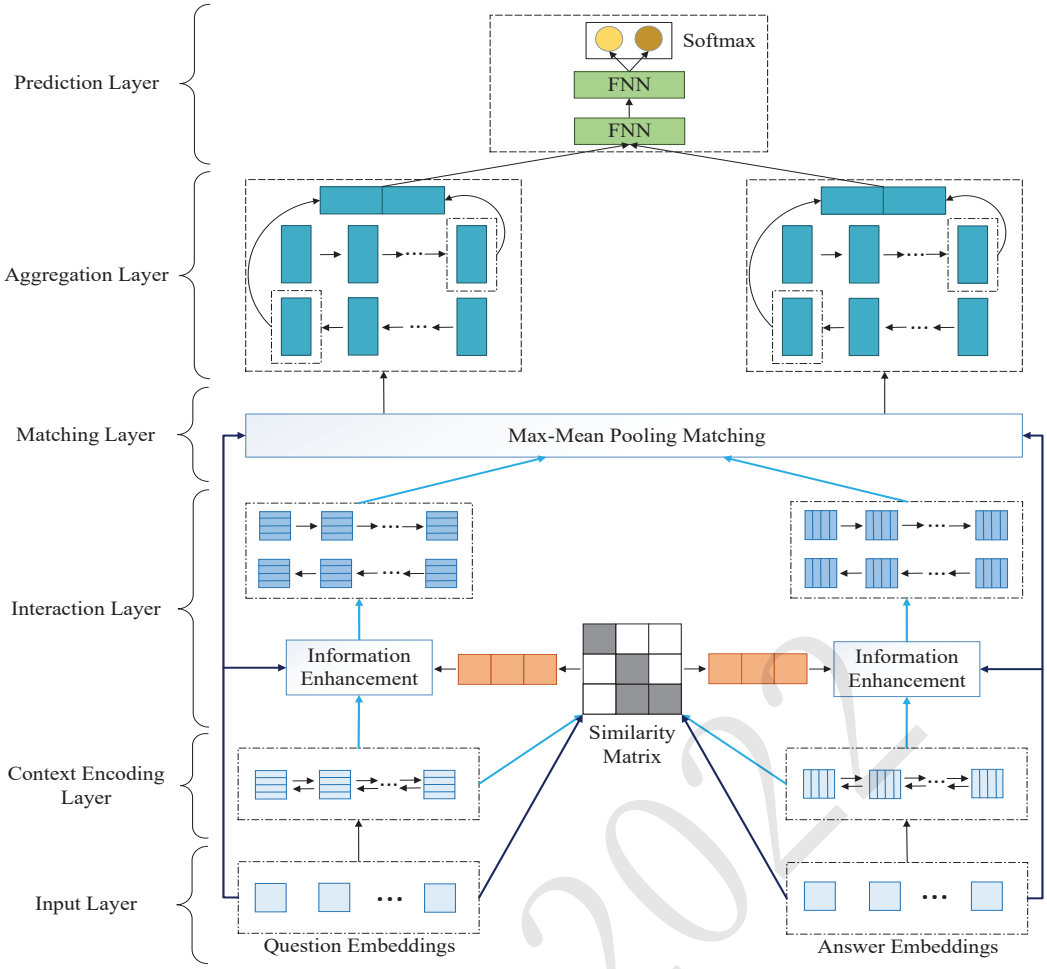


Figure 2: Architecture for Interactive Mongolian Question Answer Matching Model (IMQAMM), where the initial morpheme representations and the contextual representations are respectively applied to compute the similarity matrix for interactive information enhancement.

aggregated by CNN. Wang et al. (2017) proposed a bilateral multi-perspective matching (BiMPM) model, which used multi-perspective cosine matching strategy between encoded sentence pairs. Chen et al. (2017) improved the approach proposed by Parikh et al. (2016) and achieved sequential inference model using chain LSTMs. Tay et al. (2017) presented ComProp Alignment-Factorized Encoders (CAFE) that used factorization machines to compress the alignment vectors into scalar features, which can effectively augment the word representations. Tay et al. (2018) explored using Multi-Cast Attention Networks (MCAN) to improve learning process by adopting several attention variants and performing multiple comparison operators.

These text semantic matching models laid the foundation for later IR models and QA systems. Although these models have achieved state-of-the-art performance on various datasets, they may not be suitable for low-resource agglutinative languages. In this paper, we introduce Mongolian morpheme representation, then use interactive information enhancement to take full advantage of the information across Mongolian question-answer pairs and apply max-mean pooling matching to capture the maximum influence and the overall influence between Mongolian question-answer pairs.

3 Model Architecture

In this section, we will describe our model architecture layer by layer. Figure 2 shows a high-level view of the architecture, and then the details of our model are given as follows.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896–907, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

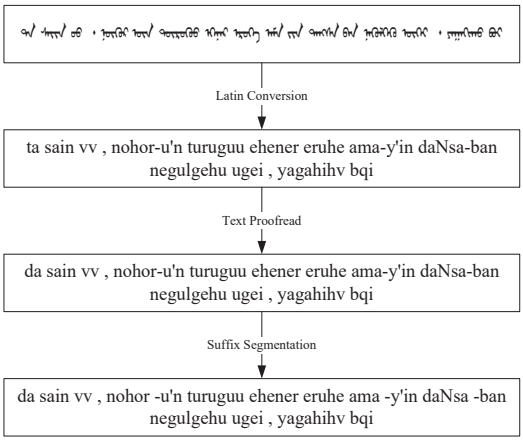


Figure 3: An example of traditional Mongolian transformation steps.

3.1 Input Layer

Mongolian is a kind of agglutinative language with complex morphological structures (Wang et al., 2015). Although there are natural spaces between Mongolian words, morphological segmentation is still needed for us. Mongolian word-formation is achieved by appending different suffixes to the stem, and they can also be concatenated layer by layer, which can lead to data sparsity. In this paper, we use Latin to deal with Mongolian and segment the suffixes to get the morpheme representations (Wang et al., 2019).

Before getting the morpheme representations of Mongolian question-answer pairs, we need to make some transformations to the traditional Mongolian language. As shown in Figure 3, the steps of transformation are divided into three steps. First of all, we convert the traditional Mongolian alphabet to the corresponding Latin alphabet. Next, because a Mongolian glyph can map to different letters, it is necessary to proofread the text (Lu et al., 2019). Finally, the suffixes connect to the stem through a Narrow No-Break Space (NNBS) (U+202F, Latin:"-"), so we can segment the suffixes to get the independent training units.

To obtain the morpheme embeddings of Mongolian question-answer pairs, we adopt Word2Vec (Mikolov et al., 2013), which contains CBOW (Continuous Bag of Word) and Skip-gram. And we choose the Skip-gram model to train the morpheme vectors.

3.2 Context Encoding Layer

LSTM is a variant of RNN, which can capture contextual dependencies effectively. In order to better represent the semantic information, we utilize the bi-directional LSTM (BiLSTM) to extract contextual features from question embeddings q and answer embeddings a.

$$\bar{q_i} = BiLSTM(q,\ i)\ ,\ \forall i\ \in\ [1,\ \ldots,\ m] \tag{1}$$

$$\bar{a_j} = BiLSTM(a,\ j)\ ,\ \forall j\ \in\ [1,\ \ldots,\ n] \tag{2}$$

where m is the length of question sentence, and n is the length of answer sentence.

3.3 Interaction Layer

In this layer, we introduce the interactive information enhancement, which contains sequence enhancement based on LSTMs and multi-cast attention using four variants of attention mechanism.

3.3.1 Sequence Enhancement

Inspired by the ESIM proposed by Chen et al. (2017), we also adopt the non-parameterized comparison strategy for sequence enhancement. Firstly, we calculate the similarity matrix between a question-answer pair encoded by BiLSTM.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896-907, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$$e_{ij} = \bar{q_i}^T \bar{a_j} \tag{3}$$

Then the key of the strategy is soft alignment attention, which can get an attentive vector of a weighted summation of the other hidden states ($\bar{a_j}$ or $\bar{q_i}$). This process is shown in the following formulas:

$$\tilde{q_i} = \sum_{j=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} \bar{a_j} \ , \ \forall i \in [1, \ldots, m] \tag{4}$$

$$\tilde{a_j} = \sum_{i=1}^{m} \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{kj})} \bar{q_i} \ , \ \forall j \in [1, \ldots, n] \tag{5}$$

where $\tilde{q_i}$ is a weighted summation of $\{\bar{a_j}\}_{j=1}^{n}$, $\tilde{a_j}$ is a weighted summation of $\{\bar{q_i}\}_{i=1}^{m}$.

Finally, we use the original hidden states and the attentive vectors to compute the difference and the element-wise product, which are then concatenated with the original hidden states and the attentive vectors.

$$T_i^q = [\bar{q_i}; \tilde{q_i}; \bar{q_i} - \tilde{q_i}; \bar{q_i} \odot \tilde{q_i}] \ , \ \forall i \in [1, \ldots, m] \tag{6}$$

$$T_j^a = [\bar{a_j}; \tilde{a_j}; \bar{a_j} - \tilde{a_j}; \bar{a_j} \odot \tilde{a_j}] \ , \ \forall j \in [1, \ldots, n] \tag{7}$$

3.3.2 Co-Attention

Co-attention is a pair-wise attention mechanism, which has a natural symmetry between sentence pairs or other pairs (Lu et al., 2017). Co-attention is a kind of variant of attention mechanism, and in this work, we decide to adopt four variants of attention mechanism: (1) **max-pooling co-attention**, (2) **mean-pooling co-attention**, (3) **alignment-pooling co-attention**, and (4) **self attention**.

The first step is to connect question and answer by calculating the similarity matrix between the initial morpheme embeddings of question-answer pairs.

$$s_{ij} = q_i^T M a_j \tag{8}$$

where M is a trainable parameter matrix.

Extractive pooling includes max-pooling and mean-pooling. **Max-pooling co-attention** aims to attend each morpheme of the sequence based on the maximum effect on each morpheme of the other sequence, while **mean-pooling co-attention** is focused on the average effect. The formulas are as following:

$$q_1' = Softmax(\max_{col}(s))^\top q \qquad a_1' = Softmax(\max_{row}(s))^\top a \tag{9}$$

$$q_2' = Softmax(\underset{col}{mean}(s))^\top q \qquad a_2' = Softmax(\underset{col}{mean}(s))^\top a \tag{10}$$

where q_1', q_2', a_1' and a_2' are the co-attentive representations of q or a.

Similar to the sequence enhancement mentioned above, **alignment-pooling co-attention** is computed individually to softly align each morpheme to the other sequence. The process is shown in the following formulas:

$$\tilde{q_i}' = \sum_{j=1}^{n} \frac{\exp(s_{ij})}{\sum_{k=1}^{n} \exp(s_{ik})} a_j \ , \ \forall i \in [1, \ldots, m] \tag{11}$$

$$\tilde{a_j}' = \sum_{i=1}^{m} \frac{\exp(s_{ij})}{\sum_{k=1}^{m} \exp(s_{kj})} q_i \ , \ \forall j \in [1, \ldots, n] \tag{12}$$

where $\widetilde{q_i}'$ is a weighted summation of $\{a_j\}_{j=1}^{n}$, $\widetilde{a_j}'$ is a weighted summation of $\{q_i\}_{i=1}^{m}$.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896-907, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

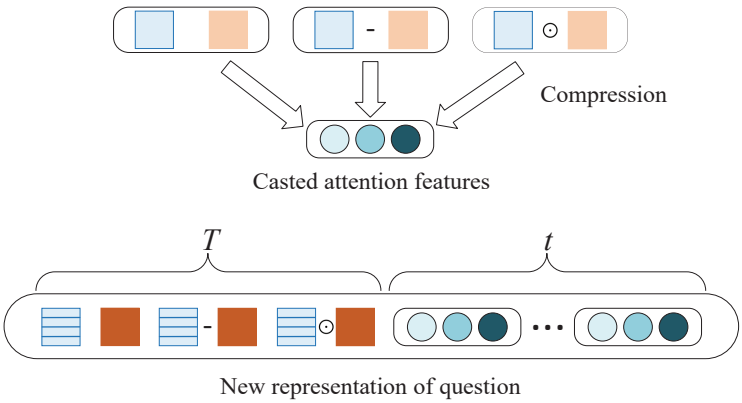


Figure 4: Information enhancement of question.

Self attention is applied to both question and answer independently. The sentence representation is denoted by x instead of q or a. The self attention function is computed as:

$$x_i' = \sum_{j=1}^{l} \frac{\exp(s_{ij})}{\sum_{k=1}^{l} \exp(s_{ik})} x_j \tag{13}$$

where $x_i{}'$ is the self-attentional representation of x_j, l is the length of the sentence.

3.3.3 Multi-Cast Attention

Multi-cast attention can get a multi-casted feature vector from multiple attention mechanisms. Each attention mechanism performs concatenation, subtractive and multiplicative operations respectively, and uses a compression function to get three scalars. The initial morpheme embeddings of a question-answer pair q and a are replaced by x, and $\tilde{x}$ is the attentive vector. The casted attention features for each attention mechanism are shown in the following formulas:

$$f_{con} = F_c([\tilde{x}; x]) \tag{14}$$

$$f_{sub} = F_c(\tilde{x} - x) \tag{15}$$

$$f_{mul} = F_c(\tilde{x} \odot x) \tag{16}$$

where F_c is a compression function, $[.;.]$ is the concatenation operator and $\odot$ is the element-wise product.

Factorization Machines (FM) can make predictions on any real-valued feature vector (Rendle, 2010). Therefore, we adopt FM as a compression function to get casted scalars. The function is as follows:

$$F(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j \tag{17}$$

where $w_0 \in \mathbb{R}$, $w_i \in \mathbb{R}^n$, $v_1, \ldots, v_n \in \mathbb{R}^{n \times k}$, and k is the number of latent factors of the FM model.

For each Mongolian question-answer pair, we apply four variants of attention mechanism mentioned above: (1) Max-pooling co-attention (2) Mean-pooling co-attention (3) Alignment-pooling co-attention and (4) Self-attention. Take the question sentence as an example, as shown in the Figure 4, three scalars are generated from each attention mechanism, so the final multi-casted feature vector is $t \in \mathbb{R}^{12}$. As such, for each morpheme, we concatenate the enhanced sequence representation T and the multi-casted feature vector t to get the new representation O^q. And O^a can be obtained in the same way.

$$O_i^q = [T_i^q; t_i^q] \ , \ \forall i \in [1, \ldots, m] \tag{18}$$

$$O_j^a = [T_j^a; t_j^a] \ , \ \forall j \in [1, \ldots, n] \tag{19}$$

CCL 2022

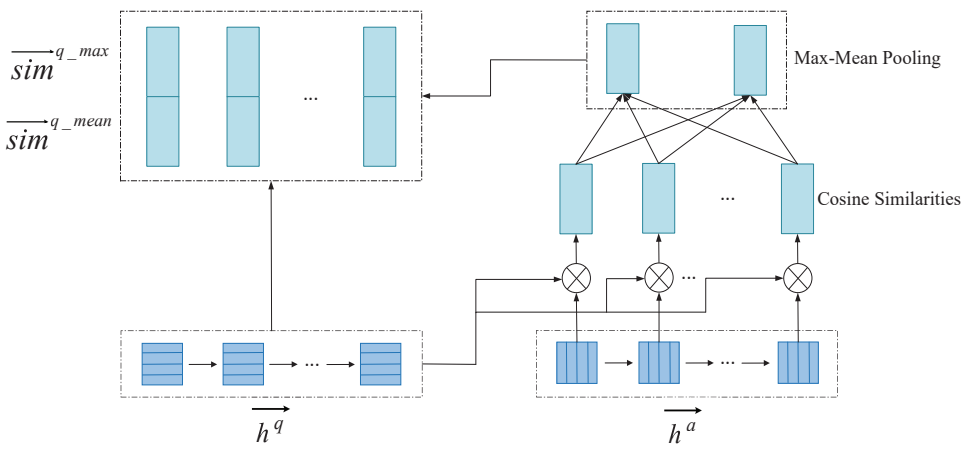


Figure 5: The max-mean pooling matching in forward direction of matching direction $q \rightarrow a$.

We use BiLSTM to encode interaction information at each time-step of O^q and O^a.

$$\overrightarrow{h_i^q} = \overrightarrow{LSTM}(h_{i-1}^q, O_i^q) \quad i = 1, \ldots, m \qquad \overleftarrow{h_i^q} = \overleftarrow{LSTM}(h_{i+1}^q, O_i^q) \quad i = m, \ldots, 1 \tag{20}$$

$$\overrightarrow{h_j^a} = \overrightarrow{LSTM}(h_{j-1}^a, O_j^a) \quad j = 1, \ldots, n \qquad \overleftarrow{h_j^a} = \overleftarrow{LSTM}(h_{j+1}^a, O_j^a) \quad j = n, \ldots, 1 \tag{21}$$

3.4 Matching Layer

To match question-answer pairs, we adopt the max-mean pooling matching strategy. Firstly, the cosine function is defined as follows:

$$sim = f_s(v_1, v_2; W) \tag{22}$$

where v_1 and v_2 are the d-dimensional vectors to be matched, $W \in \mathbb{R}^{l \times d}$ is the trainable parameter matrix, and l is the number of perspectives. For each dimension of the dimension space, it can be assigned different weights. Thus, the matching value from the k-th perspective is calculated by the formula as follows:

$$sim_k = cosine(W_k \circ v_1, W_k \circ v_2) \tag{23}$$

where $\circ$ represents the element-wise product, W_k is the k-th low of W.

Then we compare each time-step of question (or answer) representation against all time-steps of answer (or question) representation. For convenience, we only define the matching direction $q \rightarrow a$.

Morpheme Matching For the initial morpheme embeddings of question-answer pairs, we define the max-mean pooling matching strategy. The formulas are as following:

$$\overrightarrow{sim}_i^{q\_max} = \max_{j \in (1 \ldots n)} f_s(q, a; W^1) \tag{24}$$

$$\overrightarrow{sim}_i^{q\_mean} = \operatorname*{mean}_{j \in (1 \ldots n)} f_s(q, a; W^1) \tag{25}$$

Interaction Matching And for the representations of question-answer pairs after interaction, we also define the max-mean pooling matching strategy in forward direction and backward direction. Figure 5 shows the max-mean pooling matching in forward direction. The formulas are as following:

$$\overrightarrow{sim}_i^{q\_max} = \max_{j \in (1 \ldots n)} f_s\left(\overrightarrow{h_i^q}, \overrightarrow{h_j^a}; W^2\right) \qquad \overleftarrow{sim}_i^{q\_max} = \max_{j \in (1 \ldots n)} f_s\left(\overleftarrow{h_i^q}, \overleftarrow{h_j^a}; W^3\right) \tag{26}$$

$$\overrightarrow{sim}_i^{q\_mean} = \operatorname*{mean}_{j \in (1 \ldots n)} f_s\left(\overrightarrow{h_i^q}, \overrightarrow{h_j^a}; W^2\right) \qquad \overleftarrow{sim}_i^{q\_mean} = \operatorname*{mean}_{j \in (1 \ldots n)} f_s\left(\overleftarrow{h_i^q}, \overleftarrow{h_j^a}; W^3\right) \tag{27}$$

CCL 2022

At last, we concatenate all the results of the max-mean pooling matching.

$$sim_i^q = [\overrightarrow{sim}_i^{q\_max}; \overrightarrow{sim}_i^{q\_mean}; \overrightarrow{sim}_i^{q\_max}; \overrightarrow{sim}_i^{q\_mean}; \overleftarrow{sim}_i^{q\_max}; \overleftarrow{sim}_i^{q\_mean}] \tag{28}$$

where $i \in [1, \ldots, m]$, max is element-wise maximum and $mean$ is element-wise mean. The calculation process of sim_j^a is similar to that of sim_i^q.

3.5 Aggregation Layer

We utilize BiLSTM to aggregate the matching vectors sim_i^q and sim_j^a, which are calculated from two matching directions $q \to a$ and $a \to q$.

$$\overrightarrow{v_i^q} = \overrightarrow{LSTM}(v_{i-1}^q, sim_i^q) \quad i = 1, \ldots, m \qquad \overleftarrow{v_i^q} = \overleftarrow{LSTM}(v_{i+1}^q, sim_i^q) \quad i = m, \ldots, 1 \tag{29}$$

$$\overrightarrow{v_j^a} = \overrightarrow{LSTM}(v_{j-1}^a, sim_j^a) \quad j = 1, \ldots, n \qquad \overleftarrow{v_j^a} = \overleftarrow{LSTM}(v_{j+1}^a, sim_j^a) \quad j = n, \ldots, 1 \tag{30}$$

Then we concatenate the last hidden states of BiLSTM models used in two matching directions.

$$y_{out} = [\overrightarrow{v_m^q}; \overleftarrow{v_1^q}; \overrightarrow{v_n^a}; \overleftarrow{v_1^a}] \tag{31}$$

3.6 Prediction Layer

Mongolian question answer matching in this paper is a binary classification problem. We then pass the output of aggregation y_{out} into a two-layer feed-forward neural network and a softmax layer.

$$y_{pred} = softmax(W_2^F \cdot tanh(W_1^F \cdot y_{out} + b_1^F) + b_2^F) \tag{32}$$

where $W_1^F \in \mathbb{R}^{h_1 \times h_2}$, $b_1^F \in \mathbb{R}^{h_2}$, $W_2^F \in \mathbb{R}^{h_2 \times 2}$, $b_2^F \in \mathbb{R}^2$.

3.7 Model training

To train our model, we minimize the binary cross-entropy loss.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log P_i + (1 - y_i)\log(1 - P_i)] \tag{33}$$

where N is the number of labels, $y_i \in \{0, 1\}$ and P_i is the predicted probability.

4 Experiments

In this section, we describe our experimental setup and give our experimental results.

4.1 Data set and Evaluation Metrics

Our Mongolian question answering data set is translated from the Chinese question answering corpus and crawled from the Mongolian web sites. In order to improve the generalization ability of the model, we extend the original data set and construct negative samples. The ratio of positive and negative samples is 1 : 1. The data set contains 265194 question-answer pairs and each category is randomly divided into train, dev and test with the percent 80%, 10% and 10%, respectively.

We adopt Precision (P), Recall (R), F1-score (F1) and Accuracy (Acc) as the evaluation metrics of our experiments.

4.2 Model Configuration

We implement our model in TensorFlow. The batch size is set to 128, the epoch is set to 20, the max sentence length is set to 50 and the number of perspectives is set to 5. We use pre-trained 300-dimensional Mongolian Word2Vec embeddings. The size of hidden layers of all BiLSTM layers is set to 100. We use dropout with a rate of 0.1, which is applied to every layer. For training, we use the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0005 to update parameters.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896-907, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4.3 Baselines

In this subsection, we compare our model with several matching models on the Mongolian question answering data set. The first two models are based on sentence encoding methods, the next two models are based on attentive networks, while the others are based on compare-aggregate networks.

1) **SINN:** Yang and Kao (2020) proposed the model that applied self-attention based on RNN and CNN for sentence encoding.
2) **DiSAN:** Shen et al. (2018) proposed the model that used directional self-attention for encoding, and compressed features with multi-dimensional self-attention.
3) **ABCNN:** Yin et al. (2016) proposed the model that computed the attention matrix before and after convolution for modeling sentence pairs.
4) **DRCN:** Kim et al. (2019) proposed the model that used stacked RNN and co-attentive features to enhance representation.
5) **MULT:** Wang and Jiang (2017) presented the model that performed word-level matching by element-wise multiplication and aggregated by CNN.
6) **CAFE:** Tay et al. (2017) presented the model that adopted factorization machines to compress the alignment vectors into scalar features for augmenting the word representations.
7) **MCAN:** Tay et al. (2018) presented the model that adopted several attention variants and performed multiple comparison operators.
8) **ESIM:** Chen et al. (2017) presented the sequential inference model using chain LSTMs.

4.4 Results

Table 1 and Table 2 report the overall performance of the different models and the performance comparison of each category.

Model	Acc(%)
SINN	75.21
DiSAN	81.69
ABCNN	73.78
DRCN	75.31
MULT	81.19
CAFE	81.27
MCAN	81.63
ESIM	81.79
IMQAMM	**83.02**

Table 1: Test accuracy on Mongolian question answering data set.

Model	Matched			Mismatched		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
SINN	72.53	81.17	76.60	78.62	69.25	73.64
DiSAN	82.73	80.11	81.40	80.72	83.27	81.98
ABCNN	71.10	80.11	75.34	77.23	67.44	72.00
DRCN	77.53	71.29	74.28	73.43	79.33	76.27
MULT	82.80	78.74	80.72	79.73	83.64	81.64
CAFE	80.97	81.74	81.35	81.57	80.79	81.18
MCAN	80.78	83.01	81.88	82.53	80.25	81.37
ESIM	82.23	81.10	81.66	81.36	82.47	81.91
IMQAMM	83.68	82.04	**82.85**	82.39	84.00	**83.18**

Table 2: Performance comparison of different methods on test set.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896-907, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 1 presents that our Interactive Mongolian Question Answer Matching Model (IMQAMM) achieves an accuracy of 83.02%, which has already outperformed all the baseline models. Notably, IMQAMM has an improvement of about 1.23% compared to the highest ESIM in the baseline models. It shows that the introduction of multi-cast attention is helpful. IMQAMM outperforms MCAN and CAFE by 1.39% and 1.75%, which proves the significance of sequence enhancement. Compared with DRCN and ABCNN, the five models at the bottom of Table 1 have significant improvements, thus compare-aggregate networks can provide more interactive information than attentive networks in this task. And the performance of our model is higher than SINN and DiSAN, which indicates that our interactive model is better than the sentence encoding based methods on Mongolian question answering data set.

Table 2 presents the performance comparison of different methods. The improvements of IMQAMM over the highest ESIM on the matched F1 score and mismatched F1 score are 1.19% and 1.27%. Compared with all the baseline methods, our IMQAMM is competitive in each category.

4.5 Ablation Study

As shown in Table 3, we conduct an ablation study to analyze the influence of each component. We remove three parts from IMQAMM to examine the influence: 1) Multi-Cast Attention. 2) Morpheme Matching. 3) Interaction Matching.

According to the results of ablation experiments in Table 3, we can see the key components of our model. Firstly, when removing Multi-Cast Attention, the accuracy decreases by 0.38%, which proves that Multi-Cast Attention is helpful for our model. Secondly, we find that Morpheme Matching is necessary for our model. When we remove it, the accuracy is reduced by 0.6%. Finally, when removing Interaction Matching, we can observe that the performance of our model drops dramatically. The accuracy drops from 83.02% to 80.52%. This result shows that Interaction Matching is crucial for our model.

Model	Acc(%)
IMQAMM	**83.02**
w/o Multi-Cast Attention	82.64
w/o Morpheme Matching	82.42
w/o Interaction Matching	80.52

Table 3: Ablation study on Mongolian question answering data set.

5 Conclusion

In this paper, we propose an Interactive Mongolian Question Answer Matching Model (IMQAMM), which mainly combines interactive information enhancement and max-mean pooling matching. First of all, we make some transformations to traditional Mongolian language and introduce the morpheme vectors. Second, we enhance the sequence representation by concatenating a series of feature vectors. Third, the multi-cast attention is introduced to alleviate the data-sparse problem caused by complex Mongolian morphological structures. Finally, the max-mean pooling matching strategy is applied to match question-answer pairs in two directions. Experimental results show that our model performed well on the Mongolian question answering data set.

However, there is still a lot of room for improvement. In the future work, we will consider using the pre-trained language model BERT to get a better initialization, which may help improve the performance of our model.

Acknowledgements

This work is supported by National Key R&D Program (Nos. 2018YFE0122900); National Natural Science Foundation of China (Nos. 62066033, 61773224); Inner Mongolia Applied Technology Research and Development Fund Project (Nos. 2019GG372, 2020GG0046, 2021GG0158, 2020PT0002); Inner Mongolia Achievement Transformation Project (Nos. 2019CG028); Inner Mongolia Natural Science Foundation (2020BS06001); Inner Mongolia Autonomous Region Higher Education Science and

Technology Research Project (NJZY20008); Inner Mongolia Autonomous Region Overseas Students Innovation and Entrepreneurship Startup Program; Inner Mongolia Discipline Inspection and Supervision Big Data Laboratory Open Project. We are grateful for the useful suggestions from the anonymous reviewers.

References

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.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1657–1668.

Yichen Gong, Heng Luo and Jian Zhang. 2017. Natural Language Inference over Interaction Space. *arXiv preprint arXiv:1709.04348.*

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Seonhoon Kim, Inho Kang and Nojun Kwak 2019. Semantic Sentence Matching with Densely-Connected Recurrent and Co-Attentive Information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6586–6593.

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

Jiasen Lu, Jianwei Yang, Dhruv Batra and Devi Parikh. 2017. Hierarchical Question-Image Co-Attention for Visual Question Answering. *arXiv preprint arXiv:1606.00061.*

Min Lu, Feilong Bao, Guanglai Gao, Weihua Wang, and Hui Zhang. 2019. An automatic spelling correction method for classical mongolian. In *International Conference on Knowledge Science, Engineering and Management*, pages 201–214.

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.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese Recurrent Architectures for Learning Sentence Similarity. In *Thirtieth AAAI Conference on Artificial Intelligence.*

H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song and R. Ward. 2014. Semantic Modelling with Long-Short-Term Memory for Information Retrieval. *arXiv preprint arXiv:1412.6629.*

Ankur P. Parikh, Oscar Täckström, Dipanjan Das and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. *arXiv preprint arXiv:1606.01933.*

Steffen Rendle. 2010. Factorization machines. In *IEEE International conference on data mining*, pages 995–1000.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng and Grégoire Mesnil. 2014. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 101–110.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan and Chengqi Zhang. 2018. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

Ming Tan, Cicero dos Santos, Bing Xiang and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473.

Yi Tay, Luu Anh Tuan and Siu Cheung Hui. 2017. Compare, Compress and Propagate: Enhancing Neural Architectures with Alignment Factorization for Natural Language Inference. *arXiv preprint arXiv:1801.00102.*

CCL 2022

Yi Tay, Luu Anh Tuan and Siu Cheung Hui. 2018. Multi-Cast Attention Networks for Retrieval-based Question Answering and Response Prediction. *arXiv preprint arXiv:1806.00778.*

Weihua Wang, Feilong Bao and Guanglai Gao. 2015. Mongolian Named Entity Recognition using Suffixes Segmentation. In *2015 International Conference on Asian Language Processing (IALP)*, pages 169–172.

Shuohang Wang and Jing Jiang. 2017. A compare-aggregate model for matching text sequences. In *proceedings of 5th International Conference on Learning Representations, ICLR 2017.*

Zhiguo Wang, Wael Hamza and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. *arXiv preprint arXiv:1702.03814.*

Weihua Wang, Feilong Bao, and Guanglai Gao. 2019. Learning morpheme representation for mongolian named entity recognition. In *Neural Processing Letters*, 50(3): 2647–2664.

Kai-Chou Yang and Hung-Yu Kao. 2020. Generalize Sentence Representation with Self-Inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, paegs: 9394–9401.

Wenpeng Yin, Hinrich Schütze, Bing Xiang and Bowen Zhou. 2016. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. In *Transactions of the Association for Computational Linguistics*, 4: 259–272.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 896-907, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

TCM-SD: A Benchmark for Probing Syndrome Differentiation via Natural Language Processing

Mucheng Ren[1], **Heyan Huang**[1], **Yuxiang Zhou**[1], **Qianwen Cao**[1], **Yuan Bu**[2], **and Yang Gao**[1]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Xuzhou City Hospital of Traditional Chinese Medicine, Xuzhou, China
{renm,hhy63,yxzhou,qwcao,gyang}@bit.edu.cn
buyuantcm@gmail.com

Abstract

Traditional Chinese Medicine (TCM) is a natural, safe, and effective therapy that has spread and been applied worldwide. The unique TCM diagnosis and treatment system requires a comprehensive analysis of a patient's symptoms hidden in the clinical record written in free text. Prior studies have shown that this system can be informationized and intelligentized with the aid of artificial intelligence (AI) technology, such as natural language processing (NLP). However, existing datasets are not of sufficient quality nor quantity to support the further development of data-driven AI technology in TCM. Therefore, in this paper, we focus on the core task of the TCM diagnosis and treatment system—syndrome differentiation (SD)—and we introduce the first public large-scale benchmark for SD, called TCM-SD. Our benchmark contains 54,152 real-world clinical records covering 148 syndromes. Furthermore, we collect a large-scale unlabelled textual corpus in the field of TCM and propose a domain-specific pre-trained language model, called ZY-BERT. We conducted experiments using deep neural networks to establish a strong performance baseline, reveal various challenges in SD, and prove the potential of domain-specific pre-trained language model. Our study and analysis reveal opportunities for incorporating computer science and linguistics knowledge to explore the empirical validity of TCM theories.

1 Introduction

As an essential application domain of natural language processing (NLP), medicine has received remarkable attention in recent years. Many studies have explored the integration of a variety of NLP tasks with medicine, including question answering (Pampari et al., 2018; Tian et al., 2019), machine reading comprehension (Li et al., 2020; Yue et al., 2020), dialogue (Zeng et al., 2020), named entity recognition (Jochim and Deleris, 2017; He et al., 2020), and information retrieval (Liu et al., 2018). Meanwhile, numerous datasets in the medical domain with different task formats have also been proposed (Pampari et al., 2018; Li et al., 2020; Tian et al., 2019). These have greatly promoted the development of the field. Finally, breakthroughs in such tasks have led to advances in various medical-related applications, such as decision support (Feng et al., 2020; Panigutti et al., 2021) and International Classification of Disease (ICD) coding (Cao et al., 2020; Yuan et al., 2022).

However, most existing datasets and previous studies are related to modern medicine, while traditional medicine has rarely been explored. Compared to modern medicine, traditional medicine is often faced with a lack of standards and scientific explanations, making it more challenging. Therefore, it is more urgent to adopt methods of modern science, especially NLP, to explore the principles of traditional medicine, since unstructured texts are ubiquitous in this field.

TCM, as the representative of traditional medicine, is a medical system with a unique and complete theoretical basis formed by long-term medical practice under the influence and guidance of classical Chinese materialism and dialectics. Unlike modern medicine, in which medical professionals assign treatments according to disease type, TCM practitioners conduct in-depth analyses based on evidence collected from four diagnostics methods—inspection, auscultation and olfaction, interrogation, and palpation—to determine which type of **syndrome (zheng,** 证) the patient experiencing. Different treatment methods are then adopted according to the type of syndrome. Therefore, patients with the same

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

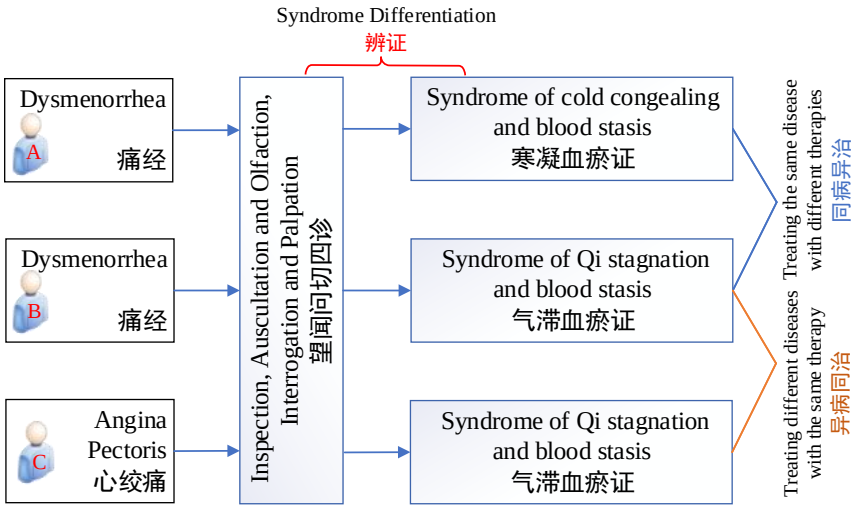


Figure 1: Concept of Traditional Chinese Medicine (TCM) syndrome differentiation.

disease may have different syndromes and thus receive different treatments, while patients with different diseases may have the same syndrome and thus undergo the same treatment. These concepts are called "treating the same disease with different therapies (同病异治)" and "treating different diseases with the same therapy (异病同治)," respectively, which are the core methods upheld by TCM.

For the example shown in Figure 1, patients A and B have the same disease—dysmenorrhea—but one is influenced by cold while the other is driven by Qi stagnation (which is a specific concept in TCM). Thus, different therapies would be assigned. However, patient C suffered from angina pectoris but shared the same syndrome as patient B. Therefore, they would be treated with similar therapies. Thus, the **syndrome**, instead of the disease, can be regarded as the primary operating unit in the TCM medical system, which not only effectively summarizes the patients' symptoms but also determines the subsequent treatment. In this process, known as **syndrome differentiation**, *the inferencing task of deciding which syndrome is associated with a patient based on clinical information*, is a vital pivot of the TCM medical system.

In recent years, with the discovery of artemisinin (Tu, 2016) and the beneficial clinical manifestations of TCM to treat COVID-19 (Yang et al., 2020; Zhang et al., 2020b), TCM has increasingly attracted attention. There have been some studies in which NLP techniques were used to explore SD tasks (Zhang et al., 2019; Zhang et al., 2020a; Wang et al., 2018; Liu et al., 2020; Pang et al., 2020), but the development has been significantly hindered by the lack of large-scale, carefully designed, public datasets.

Therefore, this paper aims to further integrate traditional medicine and artificial intelligence (AI). In particular, we focus on the core task of TCM—syndrome differentiation (SD)—to propose a high-quality, public SD benchmark that includes 54,152 samples from real-world clinical records. To our best knowledge, this is the first time that a textual benchmark has been constructed in the TCM domain. Furthermore, we crawled data from the websites to construct a TCM domain text corpus and used this to pre-train a domain-specific language model called as ZY-BERT (where ZY came from the Chinese initials of TCM). The experiments and analysis of this dataset not only explored the characteristics of SD but also verified the effectiveness of domain-specific language model.

Our contributions are summarized as follows:

1. We have systematically constructed the first public large-scale SD benchmark in a format that conforms to NLP, and established the strong baselines. This can encourage researchers use NLP techniques to explore the principles of TCM that are not sufficiently explained in other fields.

2. We proposed two novel methods, pruning and merging, which could normalize the syndrome type, improve the quality of the dataset, and also provide a reference for the construction of similar TCM datasets in the future.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3. We proposed a domain-specific language model named as ZY-BERT pre-trained with a large-scale unlabeled TCM domain corpus, which produces the best performances so far.

2 Preliminaries

To facilitate the comprehension of this paper and its motivation and significance, we will briefly define several basic concepts in TCM and analyze the differences between TCM and modern medicine.

2.1 Characteristics of Traditional Chinese Medicine (TCM) Diagnosis

The most apparent characteristic of TCM is that it has a unique and complete diagnostic system that differs from modern medicine. In modern medicine, with the assistance of medical instruments, the type of disease can be diagnosed according to the explicit digital indicators, such as blood pressure levels. However, TCM adopts abstract indicators, such as Yin and Yang, Exterior and Interior, Hot and Cold, and Excess and Deficiency.

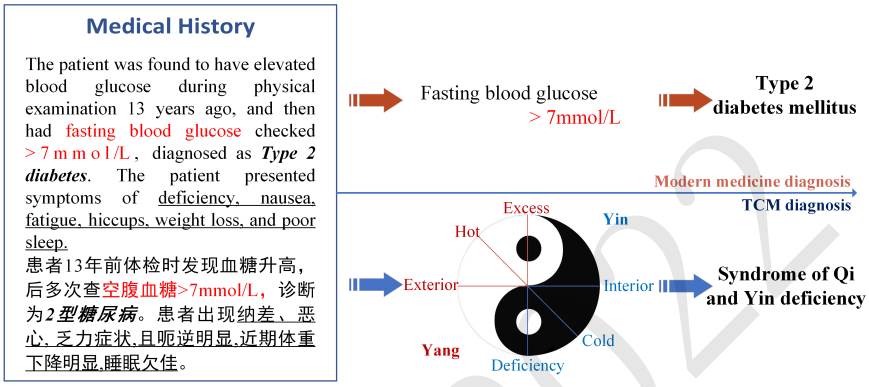


Figure 2: Different diagnostic processes of TCM and modern medicine for the same medical history.

As shown in Figure 2, given a medical history, modern medicine diagnoses the disease based on the level of fasting blood glucose, while TCM would map the various symptoms into a specific space with a unique coordinate system, analyze the latent causes, and combine them to determine a certain syndrome. Compared with the apparent numerical indicators of modern medicine, the concept of TCM is far more abstract and challenging to explain with modern medical theories.

However, TCM's difficult-to-describe nature does not mean that it has no value or rationality. In contrast, TCM has various complete and self-contained SD theories. Therefore, to explore TCM, we should not confine ourselves to the biomedical field. We may adopt NLP to explore TCM, which mainly consists of unstructured text. The linguistic characteristics may offer a scientific way to explain TCM theories. Therefore, in this paper, we present an SD dataset for further development.

2.2 Differences between ICD coding and Syndrome Differentiation

Automatic ICD coding is defined as assigning disease codes to Electronic Medical Records (EMR) , which is similar to TCM syndrome differentiation. Yet the two tasks are worlds apart in difficulty. Generally, the name of a patient's disease is directly recorded in EMR, and the task of the ICD coding is simply to normalize the names of these diseases in the manner of the ICD standard, without requiring a deep understanding of the context. For the example shown in Figure2, **Type 2 diabetes** has already described in the medical history so that ICD coding can be easily completed. While the syndrome differentiation not only requires collecting scattering evidence from the context through deep understanding, but also need to execute reliable and feasible inference, which brings a huge challenge to the model.

3 Related Works

There are three main streams of work related to this manuscript: medical dataset, natural language processing in syndrome differentiation and domain specific pre-trained language model.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908–920, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	Medical system	Domain	# of syndromes	# of samples	Task Type	Is avaliable?	Language
This Work	Traditional Medicine	General	148	54,152	Class.,MRC	Yes	Chinese
Wang (2009)	Traditional Medicine	Liver Cirrhosis	3	406	Class.	No	Chinese
Zhang (2019)	Traditional Medicine	Stroke	3	654	Class.	No	Chinese
Wang (2018)	Traditional Medicine	Diabetes	12	1,180	Class.	No	Chinese
Pang (2020)	Traditional Medicine	AIDS	7	12,000	Class.	No	Chinese
Johnson (2016)	Modern Medicine	Critical Care	-	53,423	-	Yes	English
Stubbs (2015)	Modern Medicine	General	-	1,304	De-ID.	Yes	English
Dougan (2014)	Modern Medicine	General	-	6,892	DNR	Yes	English
Abacha (2019)	Modern Medicine	General	-	405;203;383	NLI;RQE;QA	Yes	English
Tian (2019)	Modern Medicine	General	-	46,731	MRC	Yes	Chinese

Table 1: Comparison of medical datasets in traditional and modern medicine. This table only includes textual data. The abbreviations in the table are defined as follows: classification (Class.), machine reading comprehension (MRC), de-identification (De-ID.), disease name recognition (DNR), natural language inference (NLI), recognizing question entailment (RQE), and question answering (QA).

3.1 Medical Datasets

In recent years, health record systems in hospitals have been moving towards digitalization and electronization, and a large amount of clinical data has been accumulated. To make more effective use of these data and provide better medical services, some studies led by MIMIC-III (Johnson et al., 2016) have shared these valuable data with medical researchers around the world (Stubbs et al., 2015; Doğan et al., 2014). Subsequently, with the development of AI, the domain characteristics of various studies have been combined to design various task-oriented datasets (Pampari et al., 2018; Li et al., 2020; Tian et al., 2019). These datasets have greatly promoted the development of AI in the medical field and have had a profound impact on society in terms of health and well-being.

However, as shown in Table 1, most of these publicly available datasets focus on modern medicine, there are far fewer datasets on traditional medicine. This is because, compared with traditional medicine, modern medicine has a rigorous, scientific, and standardized medical system, which can efficiently collect high-quality data. Furthermore, the standardization of traditional medicine is still in the development stage, which makes the collection and construction of relevant datasets extremely challenging. Thus the scarce TCM SD datasets has hindered the development of AI in this field. To alleviate this issue, we constructed the first large-scale, publicly available dataset for TCM SD.

3.2 Natural Language Processing (NLP) in Syndrome Differentiation

At present, most existing studies have treated SD as a multi-class classification task (i.e., taking the medical records as the input and output the predicted one from numerous candidate syndrome labels). Zhang (2019) used support vector machines to classify three types of syndromes for stroke patients. Zhang (2020a) also introduced an ensemble model consisting of four methods, a back-propagation neural network, the random forest algorithm, a support vector classifier, and the extreme gradient boosting method, to classify common diseases and syndromes simultaneously. Wang (2018) proposed a multi-instance, multi-task convolutional neural network (CNN) framework to classify 12 types of syndromes in 1,915 samples. Pang (2020) proposed a multilayer perceptron (MLP) model with an attention mechanism to predict the syndrome types of acquired immunodeficiency syndrome (AIDS). Similarly, Liu (2020) proposed a text-hierarchical attention network for 1,296 clinical records with 12 kinds of syndromes. However, these approaches only worked well for small-scale datasets. Our work established a series of strong baseline models and conducted comparisons on a larger-scale datasets.

3.3 Domain Specific Pre-trained Language Model

Large-scale neural language models pre-trained on unlabelled text has proved to be a successful approach for various downstream NLP tasks. A representative example is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), which has become a foundation block for building task-specific NLP models. However, most works typically focus on pre-training in the general domain, while domain-specific pre-training has not received much attention. Table 2 summarizes common language

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908–920, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Corpus	Domain	Language	Corpus Size
BERT (Devlin et al., 2018)	Wiki+Books	General	EN	3.3B tokens
RoBERTa-wwm (Cui et al., 2021)	Web Crawl	General	CN	5.4B tokens
MacBERT (Cui et al., 2020)	Web Crawl	General	CN	5.4B tokens
SciBERT (Beltagy et al., 2019)	Web Crawl	Science	EN	3.2B tokens
BioBERT (Lee et al., 2020)	PubMed	Medical	EN	4.5B tokens
ClinicalBERT (Alsentzer et al., 2019)	MIMIC	Medical	EN	0.5B tokens
BlueBERT (Peng et al., 2019)	PubMed+MIMIC	Medical	EN	4.5B tokens
PubMedBERT (Gu et al., 2021)	PubMed	Medical	EN	3.1B tokens
TCM-BERT* (Yao et al., 2019)	Web Crawl	Medical (TCM)	CN	0.02B tokens
ZY-BERT (Ours)	Web Crawl	Medical (TCM)	CN	0.4B tokens

Table 2: Summary of pre-training details for the various BERT models.

models pre-trained in either general domain or specific domain. In general, biomedical and science are mainstream fields of pre-training language model, but in the filed of TCM, there is no much work that has been conducted as for as we know.

The reasons may be two-fold. On the one hand, TCM lacks large-scale public text corpus, like Wikipedia and PubMed. We deal with this issue by presenting a corpus in TCM domain via crawling and collecting related documents from the websites and books. On the other hand, there is also a lack of downstream tasks that can verify the performance of the pre-training language model, thus we propose the syndrome differentiation task to measure its effectiveness.

To be noticed, an existing work already proposed a language model in the filed of TCM, named as TCM-BERT (Yao et al., 2019), but it did not undergo pre-training of large-scale corpus, but was only finetuned on small-scale nonpublic corpus (0.02B tokens). While, our work provide a more completed TCM-domain corpus (over 20 times larger) and verify its effectiveness during pre-training stage.

4 Benchmark and Methods

The TCM-SD benchmark that we collected contains over 65,000 real-world Chinese clinical notes. Table 3 presents an example. Specifically, each clinical note contains the following five components: **Medical history** is the critical information for completing SD. It mainly describes a patient's condition at admission; **Chief complaint** is a concise statement describing the main symptoms that appeared in the medical history; **Four diagnostic methods record (FDMR)** is a template statement consisting of four main TCM diagnostic methods: inspection, auscultation and olfaction, interrogation, and palpation; **ICD-10 index number and name** represents the name and corresponding unique ID of the patient's disease; **Syndrome name** is the syndrome of the current patient. However, the raw data could not be used directly for the SD task due to the lack of quality control. Therefore, a careful normalization was further conducted to preprocess the data.

4.1 Syndrome Normalization

Like ICD, TCM already has national standards for the classification of TCM diseases, named *Classification and Codes of Diseases and Zheng of Traditional Chinese Medicine* (GB/T15657-1995), which stipulates the coding methods of diseases and the zheng of TCM. However, TCM standardization is still in its early phase of development and faces inadequate publicizing and implementation (Wang et al., 2016). Some TCM practitioners still have low awareness and different attitudes toward TCM standardization, resulting in inconsistent naming methods for the same syndrome.

Therefore, based on the above issues, we accomplish syndrome normalization in two stages: merging and pruning.

Merging operation is mainly used in two cases. The first is cases in which the current syndrome has multiple names, and all appear in the dataset. For example, *syndrome of wind and heat* (风热证) and

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Medical History
The patient began to suffer from repeated dizziness more than eight years ago, and the blood pressure measured in a resting-state was higher than normal many times. The highest blood pressure was 180/100 mmHg, and the patient was clearly diagnosed with hypertension. The patient usually took Nifedipine Sustained Release Tablets (20 mg), and the blood pressure was generally controlled, and dizziness occasionally occurred. Four days before the admission, the patient's dizziness worsened after catching a cold, accompanied by asthma, which worsened with activity. Furthermore, the patient coughed yellow and thick sputum. The symptoms were not significantly relieved after taking antihypertensive drugs and antibiotics, and the blood pressure fluctuated wildly. On admission, the patient still experienced dizziness, coughing with yellow mucous phlegm, chills, no fever, no conscious activity disorder, no palpitations, no chest tightness, no chest pain, no sweating, a weak waist and knees, less sleep and more dreams, forgetfulness, dry eyes, vision loss, red hectic cheeks, and dry pharynx, five upset hot, no nausea and vomiting, general eating and sleeping, and normal defecation.
患者8年余前开始反复出现头晕，多次于静息状态下测血压高于正常，最高血压180/100 mmHg，明确诊断为高血压，平素服用硝苯地平缓释片20 mg，血压控制一般，头晕时有发作。此次入院前4天受凉后头晕再发加重，伴憋喘，动则加剧，咳嗽、咳黄浓痰，自服降压药、抗生素症状缓解不明显，血压波动大。入院时：仍有头晕，咳嗽、咳黄粘痰，畏寒，无发热，无意识活动障碍，无心慌、胸闷，无胸痛、汗出，腰酸膝软，少寐多梦，健忘，两目干涩，视力减退，颧红咽干，五心烦热，无恶心呕吐，饮食睡眠一般，二便正常。
Chief Complaint
Repeated dizziness for more than eight years, aggravated with asthma for four days.
反复头晕8年余，加重伴喘憋4天。
Four Diagnostic Methods Record
Mind: clear; spirit: weak; body shape: moderate; speech: clear,..., tongue: red with little coating; pulse: small and wiry.
神志清晰，精神欠佳，形体适中，语言清晰, ... , 舌红少苔，脉弦细。
ICD-10 Name and ID: Vertigo (眩晕病) BNG070
Syndrome Name: Syndrome of Yin deficiency and Yang hyperactivity 阴虚阳亢证

External Knowledge Corpus:
A syndrome with Yin deficiency and Yang hyperactivity is a type of TCM syndrome. It refers to Yin liquid deficiency and Yang loss restriction and hyperactivity. Common symptoms include dizziness, hot flashes, night sweats, tinnitus, irritability, insomnia, red tongue, less saliva, and wiry pulse. It is mainly caused by old age, exposure to exogenous heat for a long period, the presence of a serious disease for a long period, emotional disorders, and unrestrained sexual behavior. Common diseases include insomnia, vertigo, headache, stroke, deafness, tinnitus, premature ejaculation, and other diseases.
阴虚阳亢证，中医病证名。是指阴液亏虚，阳失制约而偏亢，以头晕目眩，潮热盗汗，头晕耳鸣，烦躁失眠，舌红少津，脉细数为常见证的证候，多因年老体衰，外感热邪日久，或大病久病迁延日久，情志失调，房事不节等所致。常见于不寐、眩晕、头痛、中风、耳聋耳鸣、早泄等疾病中。

Table 3: A sample clinical record from the TCM-SD dataset with related external knowledge. An explicit match between the medical history and external knowledge is marked in blue, while the text in orange is an example of an implicit match that required temporal reasoning.

syndrome of wind and heat attacking the external (风热外袭证) belong to the same syndrome, and we would merge them into one unified name. In this case, we used the national standards for screening. Another is that the current syndrome name does not exist in a standardized form. Therefore, we recruited experts to conduct syndrome differentiation according to the specific case clinical records and finally merge the invalid syndromes into standard syndromes. For example, *syndrome of spleen and kidney yang failure* (脾肾阳衰证) would be merged into *syndrome of spleen and kidney yang deficiency* (脾肾阳虚证).

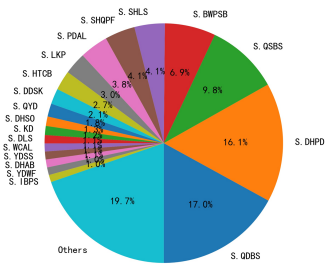
Pruning operation is mainly applied to syndromes with non-standard names that experts fail to differentiate due to vague features. In addition, since syndrome names are hierarchically graded, we pruned out syndromes with higher grades to ensure that the syndromes that appear in the current dataset are the most basic grade, that is the most specific ones that determine the subsequent treatment. For example, *syndrome of wind and cold* (风寒证) is a high-grade syndrome, and its clinical manifestations can be a *syndrome of exterior tightened by wind-cold* (风寒束表证) or *syndrome of wind-cold attacking lung* (风寒袭肺证); each has different symptoms and treatment methods.

4.2 Dataset Statistics

After normalization, the number of syndromes in the dataset was reduced from the original 548 categories to 244. Considering that some syndromes are infrequent, we further filtered out syndrome categories containing fewer than 10 samples when partitioning the dataset. Then, the processed dataset with 148 syndrome categories and 54,152 samples was divided into a training set, a development (Dev) set, and a

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908–920, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	Train	Dev	Test	Total
# Samples	43,180	5,486	5,486	54,152
# Vovabulary size	2,964	2,393	2,377	3,039
# Chief complaint tokens	13.4	13.5	13.5	13.4
# Medical history tokens	249.2	272.3	258.7	252.5
# FDMR tokens	131.9	132.3	132.2	132
# Syndrome tokens		4.9		
# Syndrome categories		148		
# Knowledge		1,027		
# Knowledge tokens		99.1		



Abbreviation	Full Name	证名
S.QDBS	syndrome of qi deficiency with blood stasis	气虚血瘀证
S.DHPD	syndrome of dampness and heat pouring downwards	湿热下注证
S.QSBS	syndrome of qi stagnation and blood stasis	气滞血瘀证
S.BWPSB	syndrome of blockade of wind-phlegm-static blood	风痰瘀阻证
S.SHLS	syndrome of stagnant heat in the liver and stomach	肝胃郁热证
S.SHQPF	syndrome of stagnated healthy qi and pathogenic factors	正虚瘀结证
S.PDAL	syndrome of phlegm and dampness accumulating in the lung	痰湿蕴肺证
S.LKP	syndrome of liver and kidney depletion	肝肾亏虚证
S.HTCB	syndrome of heat toxin congestion and binding	热毒蕴结证
S.DDSK	syndrome of dual deficiency of spleen and kidney	脾肾两虚证
S.QYD	syndrome of qi and yin deficiency	气阴两虚证
S.DHSO	syndrome of dampness and heat stagnation and obstruction	湿热瘀阻证
S.KD	syndrome of kidney deficiency	肾虚证
S.DLS	syndrome of disharmony of liver and stomach	肝胃不和证
S.WCAL	syndrome of wind and cold attacking the lung	风寒袭肺证
S.YDSS	syndrome of yang deficiency in spleen and stomach	脾胃虚寒证
S.DHAB	syndrome of dampness and heat accumulation and binding	湿热蕴结证
S.YDWF	syndrome of yang deficiency with water flooding	阳虚水泛证

Figure 3: The characteristics and syndrome distribution in the dataset.

test set with a ratio of 8:1:1. The dataset characteristics and syndrome distribution shown in Figure 3.

Since the data were collected from real-world scenarios, the distribution of syndromes was inevitably unbalanced, leading to a significant gap between the number of rare syndromes and the number of the common ones. The subsequent experiments demonstrate the challenges brought by long-tail distribution issues, and we show that this issue can be mitigated by introducing external knowledge and domain-specific pre-training.

4.3 External Knowledge

Current clinical records do not contain any relevant knowledge about the target syndromes, which causes models to have to rely on remembering patterns to complete the task. Therefore, we constructed an external unstructured knowledge corpus encompassing 1,027 types of TCM syndromes by web crawling for information on all the TCM syndromes on the online[1]. Specifically, the knowledge of each syndrome consisted of three parts: the cause of the syndrome, the main manifestations, and common related diseases. Table 3 shows an example. We demonstrate the effectiveness of this knowledge in the experimental section.

4.4 ZY-BERT

In general, ZY-BERT differs with TCM-BERT in two main parts: data and pre-training task.

First, the scale and quality of unlabelled text corpus directly affect the performance of pre-trained language models. Previous work TCM-BERT (Yao et al., 2019) directly used clinical records as pre-training corpus, resulting in monotonic data type and limited corpus size, which could not meet the needs of large-scale pre-training language model. To deal with this issue, we collected unlabelled data varies in different types from the TCM related websites, including books, articles from websites and academic papers from China National Knowledge Infrastructure (CNKI), counting over 400 million tokens.

Furthermore, the previous work TCM-BERT adopts char masking (CM) and next sentence prediction (NSP) as the pre-training tasks. However, Chinese words usually consist of multiple characters and masking single character might destroy the meaning of the whole word. For example, the word phrase *Yang Deficiency(*阳虚*)* consists of two characters. Thus, we borrowed the idea of Whole Word Masking from Cui (2021) and replace NSP with it , which could add challenges to the model training process and allow the model to learn more complex linguistic features.

Finally, the pre-trained language model consists of 24 Transformer layers, with input dimensionality of 1024. Each transformer contains 16 attention heads. Then we trained the model 300K steps with a maximum learning rate 5e-5 and a batch size of 256. Other hyperparameters and pre-training details are kept same as the ones used in Liu (2019).

5 Experiments

We selected the multi-class classification task as the primary form of SD to directly compare the performances of the existing models against the TCM-SD dataset, and used the accuracy and Macro-F1 as evaluation metrics. Specifically, the chief complaint and medical history were concatenated as the inputs,

[1] www.dayi.org.cn

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908–920, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	Dev				Test			
	Acc.	Macro-F1	Macro-R	Macro-P	Acc.	Macro-F1	Macro-R	Macro-P
DT	59.42%	20.68%	21.33%	21.52%	59.10%	21.67%	22.38%	22.20%
SVM	77.63%	32.13%	29.56%	43.10%	78.53%	36.37%	32.98%	49.35%
BiLSTM	69.30%	17.53%	15.08%	14.76%	69.65%	15.15%	15.65%	17.08%
BiGRU	73.57%	19.53%	20.12%	21.81%	74.43%	20.93%	21.90%	23.76%
CNN	77.56%	31.79%	30.39%	37.99%	78.58%	32.83%	31.29%	39.19%
BERT	79.44%	34.18%	34.12%	38.00%	80.17%	35.45%	34.99%	42.00%
distilBERT	79.09%	36.07%	36.62%	38.13%	80.46%	40.24%	39.99%	45.84%
ALBERT	79.62%	37.88%	37.65%	41.94%	80.51%	40.50%	39.57%	46.54%
RoBERTa	80.81%	43.18%	42.55%	47.68%	**82.26%**	47.55%	45.72%	54.15%
TCM-BERT	79.48%	37.84%	37.60%	42.00%	80.55%	41.58%	40.91%	48.47%
ZY-BERT(ours)	**81.43%**†	**49.47%**†	**48.89%**†	**54.08%**†	82.19%†	**51.01%**†	**49.42%**†	**57.70%**†

Table 4: Performance for the classification task. The marker † refers to p-value <0.01.

i.e. *[CLS] Chief Complaint [SEP] Medical History [SEP]*, where [CLS] and [SEP] are special tokens used for classification and separation. Then the model predicts the target syndromes from 148 candidate labels based on the representation of [CLS] token.

5.1 Baseline

The baseline methods we used consisted of four types: statistical methods, classical neural-network-based (NN-based) methods, language-model-based (LM-based) methods and domain-specific LM-based methods.

Statistical methods. These methods were the decision tree (DT) and support vector machine (SVM) methods. These two statistical methods have been widely used in previous studies on SD.

Classical NN-based methods. These methods included a Bi-LSTM (Schuster and Paliwal, 1997), a Bi-GRU (Qing et al., 2019), and a two-layer CNN (Kim, 2014). Word embeddings were retrieved from the Chinese version of BERT (Cui et al., 2021).

LM-based methods. These methods included several popular LMs, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), distillBERT (Sanh et al., 2019), and ALBERT (Lan et al., 2019). These models concatenate multiple pieces of text with special tokens as inputs, make classifications based on the hidden states of the first token, or determine the start and end of the answer by training two classifiers.

Domain-specific LM-based methods. These methods are similar with LM-based ones but usually pre-trained on domain-specific corpus rather than general domain corpus. TCM-BERT (Yao et al., 2019) and our proposed ZY-BERT are the two LM used in this manuscripts.

5.2 Main Results

Table 4 presents the performances of all the methods for the classification task. Generally, all the methods had good accuracy, which demonstrated that the models were effective at fitting when enough examples were supplied. However, each syndrome in the TCM-SD dataset should have the same importance. Thus, the Macro-F1 is a more accurate metric to evaluate the performances of the models. The Macro-F1 scores achieved by the models were much lower than the accuracy, which demonstrated the challenges of the imbalanced TMC-SD datasets.

Moreover, the statistical methods achieved better scores than the classical NN-based methods. This is because the structures designed for focusing on contextualized representations, such as the Bi-LSTM and Bi-GRU networks, were not good at capturing features, and the performances were worse. In contrast, the SVM and CNN methods were good at extracting local features and obtained better scores. Nonetheless, the language models still achieved the highest scores, demonstrating the effectiveness of the large-scale corpus pre-training.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	Dev				Test			
	EM	Macro-F1	Macro-R	Macro-P	EM	Macro-F1	Macro-R	Macro-P
Medical History + All Syndromes								
BERT	77.27%	40.71%	41.10%	43.26%	78.20%	45.60%	45.32%	50.15%
RoBERTa	78.71%	45.09%	44.30%	49.38%	80.42%	47.57%	46.42%	51.89%
Medical History + Five Syndromes								
BERT	95.59%	77.12%	76.32%	81.04%	95.83%	82.33%	81.35%	86.34%
RoBERTa	95.75%	79.16%	78.74%	82.79%	95.86%	84.42%	84.92%	86.74%
Medical History + Five Syndromes + Knowledge								
BERT	95.24%	81.21%	81.33%	84.61%	96.06%	85.15%	84.48%	87.92%
RoBERTa	**95.33%**	**81.53%**	**81.76%**	**84.49%**	**96.26%**	**85.88%**	**85.59%**	**89.09%**

Table 5: Performance with the machine reading comprehension (MRC) task.

6 Discussion

6.1 Effect of Domain-specific Pre-training

The last two rows in Table 4 indicates the effects of domain-specific pre-training. To be noticed, our proposed ZY-BERT achieved the astonishing performance improvement and mitigated long-tail distribution issue greatly. On the one hand, Macro-F1 score achieved by ZY-BERT is over 4% larger than that achieved by RoBERTa, demonstrating the effectiveness of large-scale domain-specific corpus for domain-specific tasks. On the other hand, ZY-BERT also achieves over 10% Macro-F1 scores higher than the previous domain-specific model TCM-BERT, which proves the quality and reliability of the TCM domain corpus constructed by us.

6.2 Effect of Knowledge

To testify the effectiveness of the external knowledge corpus, we leveraged knowledge into the model by concatenating the relevant syndrome knowledge with the medical history. However, due to the length limits of the language models, feeding knowledge of all syndromes into the model is infeasible under classification setting. Thus we converted the task from classification to extractive MRC, and designed the following three settings shown in Table 5 to evaluate the significance of the knowledge.

Firstly, we concatenated the original inputs with all syndrome names, and asked the model to extract the target syndrome spans from the context. The competitive results shown between MRC and classification tasks demonstrated that the model had a consistent ability among different task formats without external knowledge. Then we further conducted two groups of experiments. In the first group, instead of concatenating all syndrome names, we only included five syndromes, where one was the target syndrome and the other four were randomly selected. In the second group, we appended the corresponding knowledge for each syndrome selected in the first group. The superior results achieved by the latter group demonstrate the importance of knowledge.

However, the outstanding performance, either with knowledge or without knowledge, was mainly due to the fact that we manually narrowed down the search range to five syndromes. We used the term frequency–inverse document frequency (TFIDF) to search for relevant knowledge from the knowledge corpus based on medical history, and P@5 was only 3.94%. Thus, knowledge is essential, but finding it is difficult.

6.3 Ablation Study

Table 6 shows the results of the ablation study on the TCD-SD dataset. Removing either the medical history or the chief complaint resulted in lower performances, especially if only the chief complaint was taken into account. This was because the chief complaint was typically too short to include sufficient features for classification. However, the chief complaint and medical history complemented each other in a coarse-to-fine fashion.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Method	Dev Acc.	Dev Macro-F1	Dev Macro-R	Dev Macro-P	Test Acc.	Test Macro-F1	Test Macro-R	Test Macro-P
Only Chief Complaint								
BERT	70.56%	23.15%	26.34%	26.34%	71.58%	24.08%	25.38%	24.08%
RoBERTa	71.36%	28.55%	28.85%	33.13%	72.91%	30.78%	34.54%	34.54%
Only Medical History								
BERT	79.40%	33.50%	33.46%	37.90%	79.62%	35.57%	35.13%	42.18%
RoBERTa	79.80%	41.40%	40.12%	45.38%	81.83%	45.19%	43.03%	53.78%
Chief Complaint + Medical History								
BERT	79.44%	34.18%	34.12%	38.00%	80.17%	35.45%	34.99%	42.00%
RoBERTa	**80.81%**	**43.18%**	**42.55%**	**47.68%**	**82.26%**	**47.55%**	**45.72%**	**54.15%**

Table 6: Ablation study on the TCM-SD dataset.

6.4 Error Analysis

By analyzing the error cases, we found that the vast majority of errors occurred in the category with few samples, and fitting only according to the data distribution was still the most significant issue. Except for algorithmic problems, we concluded that there were three main error types:

Complex Reasoning. As shown in Table 3, besides the explicit match marked in blue, there was an implicit match marked in orange that required temporal reasoning. Additionally, the task also included complex reasoning, such as numerical reasoning, spatial reasoning and negative reasoning.

Incomplete Knowledge. The current models do not take into account the concepts that arise from the SD task, such as Yin and Yang. Therefore, the models do not know how to map the symptoms into the special coordinate system of the TCM diagnostics system.

Out-Of-Vocabulary. In the clinical records, there exists not only academic medical-related terms but also various rare traditional characters in TCM, which impeded the understanding of the context.

7 Conclusions

This paper introduced a meaningful task, SD, in TCM and its connection with NLP and presented the first public large-scale benchmark of SD: TCM-SD. Furthermore, a knowledge corpus supporting the model understanding and the large-scale TCM domain corpus for pre-training were constructed. Moreover, one domain-specific pre-training language model named as ZY-BERT was proposed. The experiments on this dataset demonstrated the challenges, the inadequacy of existing models, the importance of knowledge and the effectiveness of domain-specific pre-training. This work can greatly promote the internationalization and modernization of TCM, the proposed benchmark and associated baseline models provide a basis for subsequent research.

Acknowledgements

This work is supported by funds from the National Natural Science Foundation of China (No.U21B2009). The data used in this paper were only routine diagnosis and treatment data of patients, excluding any personal information of the patients (such as name, age, and telephone number). This study did not interfere with normal medical procedures or create an additional burden to medical staff, and no experiments were conducted on patients. **All the data have been desensitized**. Therefore, this paper does not involve ethical issues and waives the requirement of individual patient consent. We public TCM-SD dataset, TCM-domain corpus and ZY-BERT model at https://github.com/Borororo/ZY-BERT. We thank the reviewers for their helpful and constructive comments. And we thank M.D. Yonglan Zhou for her insightful and professional suggestions.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Emily Alsentzer, John Murphy, William Boag, et al. 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.

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 (EMNLP-IJCNLP)*, pages 3615–3620.

Pengfei Cao, Chenwei Yan, Xiangling Fu, et al. 2020. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online, July. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, et al. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online, November. 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/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable clinical decision support from text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1478–1489, Online, November. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, et al. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yun He, Ziwei Zhu, Yin Zhang, et al. 2020. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online, November. Association for Computational Linguistics.

Charles Jochim and Léa Deleris. 2017. Named entity recognition in the medical domain with constrained CRF models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 839–849, Valencia, Spain, April. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, et al. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, et al. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dongfang Li, Baotian Hu, Qingcai Chen, et al. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1427–1438, Online, November. Association for Computational Linguistics.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Ziqing Liu, Enwei Peng, Shixing Yan, et al. 2018. T-know: A knowledge graph-based question answering and infor-mation retrieval system for traditional Chinese medicine. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 15–19, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ziqing Liu, Haiyang He, Shixing Yan, et al. 2020. End-to-end models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: Model development and validation. *JMIR Medical Informatics*, 8(6):e17821.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Others. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium, October-November. Association for Computational Linguistics.

Huaxin Pang, Shikui Wei, Yufeng Zhao, et al. 2020. Effective attention-based network for syndrome differentiation of AIDS. *BMC Medical Informatics and Decision Making*, 20(1):1–10.

Cecilia Panigutti, Alan Perotti, André Panisson, et al. 2021. Fairlens: Auditing black-box clinical decision support systems. *Information Processing & Management*, 58(5):102657.

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.

Li Qing, Weng Linhong, and Ding Xuehai. 2019. A novel neural network-based method for medical text classification. *Future Internet*, 11(12):255.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Yuanhe Tian, Weicheng Ma, Fei Xia, et al. 2019. ChiMed: A Chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260, Florence, Italy, August. Association for Computational Linguistics.

Youyou Tu. 2016. Artemisinin—A gift from traditional Chinese medicine to the world (nobel lecture). *Angewandte Chemie International Edition*, 55(35):10210–10226.

Yan Wang, Lizhuang Ma, and Ping Liu. 2009. Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. *Computer Methods and Programs in Biomedicine*, 95(3):249–257.

Juan Wang, Yi Guo, and Gui Lan Li. 2016. Current status of standardization of traditional Chinese medicine in china. *Evidence-Based Complementary and Alternative Medicine*, 2016.

Zeyuan Wang, Shiding Sun, Josiah Poon, et al. 2018. CNN based multi-instance multi-task learning for syndrome differentiation of diabetic patients. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1905–1911. IEEE.

Yang Yang, Md Sahidul Islam, Jin Wang, et al. 2020. Traditional Chinese medicine in the treatment of patients infected with 2019-new coronavirus (sars-cov-2): A review and perspective. *International journal of biological sciences*, 16(10):1708.

Liang Yao, Zhe Jin, Chengsheng Mao, et al. 2019. Traditional chinese medicine clinical records classification with bert and domain specific corpora. *Journal of the American Medical Informatics Association*, 26(12):1632–1636.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. *arXiv preprint arXiv:2203.01515*.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online, July. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, et al. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, November. Association for Computational Linguistics.

Dongxue Zhang, Zhichao Gan, and Zhihui Huang. 2019. Study on classification model of traditional Chinese medicine syndrome types of stroke patients in convalescent stage based on support vector machine. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, pages 205–209. IEEE.

Hong Zhang, Wandong Ni, Jing Li, et al. 2020a. Artificial intelligence–based traditional Chinese medicine assistive diagnostic system: Validation study. *JMIR medical informatics*, 8(6):e17608.

Leyin Zhang, Jieru Yu, Yiwen Zhou, et al. 2020b. Becoming a faithful defender: Traditional Chinese medicine against coronavirus disease 2019 (covid-19). *The American journal of Chinese medicine*, 48(04):763–777.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 908-920, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

COMPILING: A Benchmark Dataset for Chinese Complexity Controllable Definition Generation

Jiaxin Yuan[12*], Cunliang Kong[123*], Chenhui Xie[123], Liner Yang[123†], Erhong Yang[13]

[1]National Language Resources Monitoring and Research Center Print Media Language Branch, Beijing Language and Culture University

[2]School of Information Science, Beijing Language and Culture University

[3]Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University

jiaxinyuan625@gmail.com

Abstract

The definition generation task aims to generate a word's definition within a specific context automatically. However, owing to the lack of datasets for different complexities, the definitions produced by models tend to keep the same complexity level. This paper proposes a novel task of generating definitions for a word with controllable complexity levels. Correspondingly, we introduce **COMPILING**, a dataset given detailed information about Chinese definitions, and each definition is labeled with its complexity levels. The COMPILING dataset includes 74,303 words and 106,882 definitions. To the best of our knowledge, it is the largest dataset of the Chinese definition generation task. We select various representative generation methods as baselines for this task and conduct evaluations, which illustrates that our dataset plays an outstanding role in assisting models in generating different complexity-level definitions. We believe that the COMPILING dataset will benefit further research in complexity controllable definition generation.

1 Introduction

Definition Generation (DG) is the task of describing the meaning that a word takes in a specific context. This task can help language learners by providing explanations for unfamiliar words. Recent researches (Ishiwatari et al., 2019; Zheng et al., 2021) attempted to apply the task to the field of Intelligent Computer-Assisted Language Learning (ICALL), and have made a significant progress.

Previous studies on DG mainly concentrate on generating different definitions for polysemous words (Gadetsky et al., 2018; Mickus et al., 2019; Reid et al., 2020), or generating definitions with appropriate specificity (Huang et al., 2021a). In these studies, researchers have faced various issues, such as the high complexity problem. High complexity definitions contain words that are more difficult than the defined word, and hence are labored for language learners to read and understand. Nevertheless, there have been few focuses on complexity controllable generation of definitions. A possible reason is that the complexities of definitions are not provided in currently existed datasets, which leads to the difficulty of automatic training and evaluation.

Actually, the problems mentioned above are especially prominent in the language environment of Chinese. Definitions with suitable complexity are in urgent practical needs for Chinese as Foreign Language (CFL) learners. According to the Ministry of Education of China, by the end of 2020, more than 20 million foreign students are learning Chinese. But as Zhang (2011) pointed out, since the difficulty of definitions is not considered, most existing dictionaries cannot meet CFL learner's requirements. Besides, the existing Chinese learner dictionaries contain only a small number of words. For instance, the Contemporary Chinese Learner Dictionary (CCLD) only has about 6,400 words. In contrast, the Modern Chinese Dictionary (MCD), which is designed for native speakers, has about 69,000 words.

Therefore, in this work, we focus on the task of generating definitions for CFL learners with appropriate complexities. At present, there are two datasets used for the Chinese definition generation task, but

*Equal contribution

†Corresponding author: Liner Yang (lineryang@gmail.com)

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

neither of them can meet the needs of this task. The most widely used CWN dataset (Yang et al., 2020; Fan et al., 2020; Kong et al., 2020) was built from the Chinese WordNet (Huang et al., 2010), which is a knowledge base of sense distinction[1]. This dataset is limited in size with 8,221 words. Zheng et al. (2021) constructed a dataset from the 5th edition of MCD. But it only collects disyllabic nouns and verbs, and additional annotation of formation rules is required. Besides, both datasets didn't provide the complexity of definitions, which is essential information in the controllable generation.

To enhance the study of this task, we propose to build a novel benchmark dataset named **COMPILING** (**C**hinese c**OMP**lex**I**ty contro**L**lable def**IN**ition **G**eneration). The dataset is large and of high quality, which contains 127,757 entries in total. Each entry consists of a word, an example, a definition, and two complexity measurements of this definition. More specifically, we build the dataset by using two Chinese dictionaries, namely the CCLD and the 7th edition of MCD. The former collects fewer words, but the definitions are simpler. The latter is the opposite. By combining these two dictionaries, we obtain a large amount of definitions that vary in different complexities.

In order to quantitatively measure the *complexity* of definitions, we refer to the graded vocabularies formulated by HSK (Chinese Proficiency Test). HSK is set up to test the proficiency of non-native speakers. It has nine levels from easy to hard, and each level corresponds to a vocabulary. The COMPILING dataset contains an average level and a maximum level of each definition.

We find that both dictionaries tend to use phrases rather than complete sentences as examples in some cases. For instance, the word "规模" (scale) has two example phrases in MCD (Modern Chinese Dictionary), which are "规模宏大" (large scale) and "初具规模" (begin to take shape). We think that short phrases might be helpful for language learners to understand, but complete sentences can provide more context in the automatic definition generation. Thus, we design an algorithm to expand the phrases into sentences (Section 4.2).

We believe that this dataset can further enhance the research on Chinese complexity controllable definition generation, which could not only benefit the language learners, but also low literacy readers, as well as people with aphasia or dyslexia. We also provide baselines of mainstream generation methods as references (Section 6).

In summary, our contributions are listed below:

- We propose a novel task of generating definitions for a word with appropriate complexity. The task is of great use in helping CFL learners to learn the vocabulary.
- We propose the **COMPILING** dataset that is of large scale and high quality. This dataset could serve as the benchmark of the task we proposed.
- We perform several experiments on the COMPILING dataset and the results demonstrate it could assist models to achieve effective complexity controllable generations.

2 Related Work

2.1 Definition Generation

Noraset et al. (2017) first proposed the definition modeling task and use word embeddings to generate definitions of the corresponding words. Referencing on the problem of word sense disambiguation, Ishiwatari et al. (2019) and Gadetsky et al. (2018) incorporated word contexts into definition modeling and demonstrated its effectiveness of distinguishing different meanings. Recent work (Huang et al., 2021b) reformulates the task as generating descriptions using extracted knowledge. Research on Chinese definition modeling was first proposed by Yang et al. (2020), they adapted a transformer-based model and incorporated sememes into the model to provide more external semantic knowledge. Fan et al. (2020) redefined the Chinese definition modeling as generating the corresponding definition for a target word and its context. Zheng et al. (2021) utilized the characteristics of Chinese by adding formation features to enhance definition modeling. Besides, there are also studies on multilingual definition generation (Kong et al., 2020) and combining extraction and generation for this task (Huang et al., 2021c).

[1] http://lope.linguistics.ntu.edu.tw/cwn2

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Notably, Kong et al. (2022) proposed to generate simple definitions employing a multitasking framework. Since the lack of a definition dataset with different complexities, they managed to generate both complex and simple definitions in an unsupervised way.

Differently, we focus on building the benchmark dataset for different Chinese definition generation tasks and hope it could be beneficial for further research.

2.2 Controllable Generation

Controllable generation is widely adapted in kinds of language modeling tasks. For instance, data augmentation (Amin-Nejad et al., 2020), dialog generation (Firdaus et al., 2020), storytelling (Goldfarb-Tarrant et al., 2020), and so on. And the objects controlled in different studies vary from each other. Specifically, considering the significance of sentiment in poetry definition, Chen et al. (2019) proposed a model to generate poetry with controllable emotions. Gao et al. (2019) first presented a framework to develop questions about specific answers that meet target difficulty levels. To attract more readers, Jin et al. (2020) introduced a headline generation model to produce enticing titles with target three styles. Likewise, in order to explore and release the practical value of definition generation, we propose the complexity controllable definition generation task committed to producing definitions satisfying users of all levels.

Currently, the most controllable generation tasks are achieved based on pre-trained learning models. And Zhang et al. (2022) summarized the common methods as Finetuning, Retrain PLMs, and Post-Process. And we utilize the first method to control the complexity of the definition more efficiently.

2.3 Prompt Learning

In recent years, the pre-trained model with fine-tuning has gradually become the mainstream of natural language processing tasks. Due to the complex training objectives and large hyperparameter groups, large-scale pre-training models can effectively extract features from a large amount of supervised and unsupervised data. By storing the learned knowledge in parameters and fine-tuning the model for specific tasks, the same model can be applied to a series of downstream natural language processing tasks (Han et al., 2021a).

Prompt learning is a method of fully learning knowledge by adding additional text to the model's input. Prompt can be divided into artificial and automatic construction according to the text attached to the input (Han et al., 2021a). Among them, automatically constructed prompts are divided into discrete and continuous ones. A discrete prompt refers to the fact that the constructed prompt is composed of actual text symbols, and applicable tasks include text classification (Han et al., 2021b), text generation (Zheng and Huang, 2021), etc.

Although the combination of pre-training and fine-tuning methods can be adapted to most NLP tasks, when it comes to each specific task, the number of parameters that need to be adjusted for are vast. By adopting prompt learning, the pre-training model can be applied to the required tasks by only modifying the part of the prompt for different downstream tasks. Therefore, the training process will become more efficient.

3 Problem Formulation

In this work, we aim to generate a definition $\boldsymbol{d}^c$ with appropriate complexity c, for a given word and example sentence $(w^*, \boldsymbol{e})$. This task is feasible because the word and it's corresponding definition should be assumed to have the same semantics. A common solution is to predict tokens in the definition one by one, depending on the previous words and the other conditions, which can be formulated as:

$$P(\boldsymbol{d}^c|w^*, \boldsymbol{e}, c) = \prod_{t=1}^{T} P(\boldsymbol{d}_t^c|\boldsymbol{d}_{<t}^c, w^*, \boldsymbol{e}, c), \tag{1}$$

where d_t^c is the t-th token in the definition, and T is the total length of definition. Each probability distribution can be approximated by the following equation:

$$P(\boldsymbol{d}_t^c|\boldsymbol{d}_{<t}^c, w^*, \boldsymbol{e}, c) \propto \exp(Wh_t/\tau), \tag{2}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

where W is a matrix collecting word vectors, h_t is a vector summarizing inputs at current time-step, and τ is a hyper-parameter for temperature, set to 1 in default.

4 Dataset Construction

The source corpora are extracted from the MCD and CCLD, both published by the Commercial Press. For corpus from MCD and CCLD, we process them separately with the same construction methods and finally put them together.

The construction of the COMPILING dataset is divided into three stages: data structured annotation, example sentences expansion, and post-processing. First, we propose a strategy for building structured datasets due to the high complexity and compact construction of automatically extracted data. In this phase, we set up a platform. It not only helps annotators proofread and audit corpus data more efficiently but is also conducive for us to check and collect data. Besides, since the context of a targeted word in the dictionary is always a collocation instead of a complete sentence, we then conduct expanding context to enhance the overall abundance of language for our proposed datasets. Furthermore, to divide definitions into different complexity levels, we calculate the HSK level of each description.

4.1 Data Structured Annotation

In the beginning, we collect initial data and find they are disorganized and complex in structure, which is problematic to conduct automatic processing. Hence, we start up data structured annotation. To better manage and boost the whole process, we build up a platform before the formal annotation and deploy it on two servers, one for corpus from MCD, and the other for corpus from CCLD. This platform could not only serve specifically for this task, but it is also appropriate for the construction of any resource by replacing the data.

Concentrating on tackling the problem of disorganized data, we suggest a series of rules for annotation. For a particular word, its attached contents include its spell, definition, example sentences of the usage of a specific definition, and so on. Hence, we propose to add labels before corresponding contents to distinguish different types of data, which is conducive for computers to extract this information based on their labels automatically. Both dictionaries have instructions illustrating the meta-information, such as the organization of entries, the style of definitions and examples, and basic usages. We invite a student who majors in linguistics to formulate the annotation guidelines based on the instructions, which will be the reference for annotators. By doing so, we hope annotators could restore that language information and the relationships between them to a large extent. Then, we invite 20 students majoring in linguistics to annotate the corpora on our platform regarding the guidelines. This phase lasted for two months.

Algorithm 1 Example Sentences Expansion

Input: phrase p, corpus C
Output: examples E

$D \leftarrow \{\}, E \leftarrow []$
for $sentence$ **in** C **do**
  if p **in** $sentence$ **then**
    $score \leftarrow pplScore(sentence)$ ▷ Compute the PPL score for each sentence.
    $D[sentence] \leftarrow score$
  end if
end for
$sortedExamples \leftarrow descSortByValue(D)$ ▷ Descendant sort by the scores.
for $i = 0 \rightarrow topN$ **do** ▷ $topN$ is set to 5 in practice.
  $E.add(sortedExamples[i])$
end for

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4.2 Example Sentences Expansion

While the information extracted from dictionaries is large and abundant, the context attached to the targeted words given in dictionaries is too short to provide enough knowledge for the model to learn and generate descriptions. In the second stage of construction, considering the significance of sentences, we start up example sentence expansion. For contexts without sufficient length in the original corpus, we tend to find sentences with a longer length and higher quality in the new canon for replacement, and the specific process is as follows. We first screened each example sentence in the annotated texts. We set the length threshold to six, and if the length of the initial context is longer than the threshold, we will retain the sentences; otherwise, we will find longer sentences with more abundant information in the new corpus to cover the original ones. It is worth noting that if a term contains more than one sentence (collocation), for each sentence (collocation), we will replace it with new matching contexts.

We design Algorithm 1 to match and gain new high-quality sentences. Given the ambiguity of most words, we utilize an allocation as the input of Algorithm 1 instead of a phrase to ensure the found sentences contain the corresponding usage of a specific definition. As shown in Algorithm 1, we collect all the sentences that fit the requirements and grade them by utilizing Perplexity (PPL)[2], which is one of the most common metrics for evaluating language fluency. Eventually, the top five sentences in the rankings are designated to replace those original short contexts.

4.3 Post Processing

Difficulty classification The most crucial step of constructing a complexity-controlled dataset is integrating the difficulty level of definition into the dataset. We utilize the HSK metric to represent the complexity degree. HSK[3], called the Chinese Proficiency Test, set to evaluate the Chinese proficiency and application of non-native speakers. It is divided into nine levels, and the difficulty increases progressively from low to high. For convenience, we regard the seventh, eighth, and ninth levels as a whole. Finally, we set seven complexity levels of HSK, and each level corresponds to a vocabulary. For words that are not included in the first seven-level, we classify them as the highest level.

Entry construction Besides, For each definition, we first conduct word segmentation, then calculate the average and highest HSK level, and combine the HSK level into the dataset. Eventually, each entry of the COMPILING dataset consists of a target word, its definition, the average and highest HSK level, and the contexts of the corresponding usage of this description.

5 Dataset Analysis

Table 1: The main statistics of the COMPILING dataset.

Datasets	Count		Average Length	
	Words	Entries	Definition	Context
MCD	67,801	101,314	13.8	27.5
CCLD	6,502	26,443	13.4	20.4

As mentioned before, the smallest unit of the COMPILING dataset consists of five parts. In particular, if a word is polysemous or has numerous contexts, they are regarded as distinct entries. For instance, as shown in Table 2, the word "收拾" (clear up) has four different definitions, and each of them follows an example sentence. Hence there are four entries of "收拾" (clear up) in total.

As shown in Table 1, we analyze statistics of data extracted from MCD and CCLD, respectively. Table 3 shows the basic statistics of the COMPILING dataset and another dataset of Chinese definition modeling. For training, the given definitions of each entry are seen as the ground truth.

[2] https://huggingface.co/docs/transformers/perplexity
[3] http://www.chinesetest.cn

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 2: Example entries of COMPILING dataset.

Word	Definition	Average	Maximum	Sentence	Source
收拾 clear up	使变干净整齐；整理 To make clean and tidy	2	3	东西都收拾好了，可以出门了。 With everything packed up, we're ready to go.	CCLD
收拾 repair	使有毛病的东西功能正常；修理 To make something defective function properly	2	4	我的手机坏了，得找厂家收拾一下。 My phone is out of order so I have to ask manufacturer for help.	CCLD
收拾 settle	整理；整顿 Put in order	4	6	冬储夏衣，夏藏冬衣，收拾屋子，还要照看外孙女。 Store summer clothes in the winter, hide winter clothes in the summer, clean the house, and look after her granddaughter.	MCD
收拾 kill	消灭；杀死 Eliminate	8	10	据点的敌人，全叫我们收拾了。 All the enemies in the stronghold have been eliminated.	MCD

Table 3: Statistics of Chinese definition modeling datasets.

Datasets	Count		Average Length	
	Words	Entries	Definition	Context
CWN	8,221	84,542	9.07	21.57
COMPILING	74,303	127,757	13.60	23.95

To better highlight the complexity degree of the dataset, we set levels 1-3 in HSK as the simple grade, levels 3-7 as the medium grade, and levels 7-9 and 9+ as hard quality. We count the HSK level distribution of definitions of COMPILING, as shown in Figure 1.

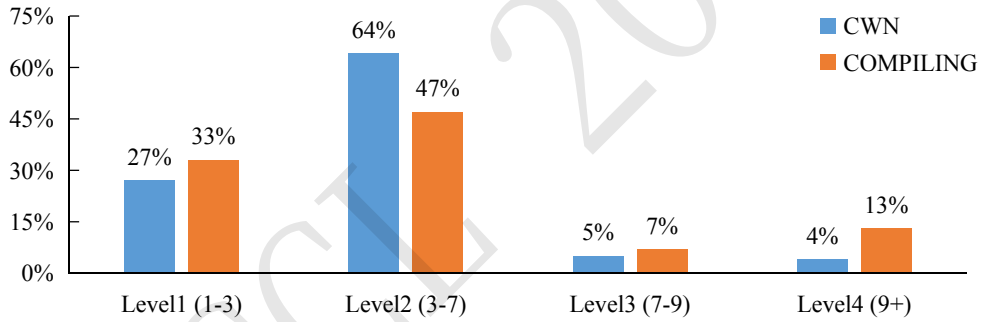


Figure 1: The distribution of average HSK level in CWN and COMPILING.

The distribution of definitions in the COMPILING dataset in the three levels is closer than CWN. Given the particularity of the Complexity Controllable definition generation task, it is necessary to construct a dataset including entries covering all difficulty levels. In this way, the model can learn and distinguish the complexity of descriptions, hence generating a new definition of a word with a target complexity level.

Hence, the COMPILING dataset could be applied to both general definition generation tasks and those which incorporate the complexity of definitions, demonstrating its value in being as a benchmark dataset.

6 Experiments

6.1 Baselines

This section introduces several methods for common generation tasks, which can serve as baselines for our proposed task.

LOG-CaD LOG-CaD (Ishiwatari et al., 2019) is a model for generating descriptions for words and phrases. This model summarizes clues from the static, contextualized, as well as character-level em-

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

beddings of the given word, and then employs an LSTM-decoder for the generation. A gated attention mechanism is employed to capture and filter information from the embeddings during decoding.

Transformser We treat the task as a special type of single language translation and directly use the original transformer model proposed by Vaswani et al. (2017). We concatenate the word and example sentence as the input sequence and train the model to generate the definition. We use the same approach to deal with the input and output in BERT and BART models. All hyper-parameters are set according to the original paper for a fair comparison.

BERT Pretrained language models have been widely used in various NLP tasks in recent years. By obtaining prior knowledge during pretraining, the PLMs can encode the input sentence more effectively. Thus, we use the Chinese-bert-base (Devlin et al., 2019) model to initialize all the parameters in a transformer encoder and employ a transformer decoder for the generation. Note that the decoder is trained from scratch without initialization.

BART Unlike BERT, BART (Lewis et al., 2019) is a pretrained encoder-decoder language model, which is more suitable for generation tasks. Since the monolingual BART only support English, we use the multilingual version of BART and set both source and target language as Chinese for this task.

MT5 T5 is one of the representative pre-training language models. It considers all NLP tasks as a uniform text-to-text paradigm. mT5 (Xue et al., 2021) is a multi-language variant of T5, and its performance on various benchmark tasks is generally outstanding. Therefore, we choose mT5 to perform the prompt learning method.

Table 4: Datasets divided by HSK level.

Complexity	HSK	Entries
Easy	1-3	48,458
Medium	4-7	53,945
Hard	7+	25,354

6.2 Settings

As a benchmark dataset introduced to enhance the Chinese definition generation task, we set up the experiments to verify the effectiveness of the COMPILING dataset.

Regardless of complexity levels We first design the experiment to evaluate the overall performance of the baseline models on our dataset. In this setting, we train the models using the entire training set, despite of the different complexity levels. And the purpose of this setting is to provide a comparison standard for other experiments. We divide the dataset into training, development, and test sets according to 8:1:1. The training data are fine-tuned according to the input formats of different models.

Complexity specific models To evaluate the significant role of the COMPILING dataset in generating definitions across various difficulties, we set up an experiment to train the model on different complexity-level sub-datasets. First of all, we split the dataset into three subsets on basis of the average HSK level. As shown in Table 4, the HSK levels of definitions in Easy Set are between 1 to 3, Medium Set corresponding to level 4-6, and Hard Set corresponding to level 7+. Then we split each subset into training, development, and test sets according to the ratio of 8:1:1. Finally, we fine-tune the BART model utilizing these three training sets, and hence getting three models. Each one could generate definitions with its corresponding complexity level.

Unified model based on prompt learning To assist the model to generate descriptions with different complexity of demand, we adopt the method of prompt learning. It allows the model to learn by adding tokens that represent difficulty information to the inputs, such as <extra_id_1> for level 1 (lowest), <extra_id_2> for level 2, and so on. The training set is formed by prefacing each definition of the

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

COMPILING dataset with the corresponding special tokens. Each entry of the final dataset includes: <extra_id_x>, target word, its corresponding definitions and context. During the training phase, the model encodes both complexity and definition information. In the analysis stage, aiming to verify the effectiveness of this method, we select 10 entries from the test set of the COMPILING dataset. For each entry, only its difficulty token is modified with the other information keep remaining, so as to construct two copies of the entry. It is worth noting that the principle of constructing the new complexity tokens is, that the two new entries and the original one(a group of data) differ by at least 2 levels or more, which means they can represent easy, medium, and hard complexity respectively. For example, if the definition of the source entry is specified with the difficulty as 3, the complexities of the two copies of it need to be constructed as at least 1 and 5. Finally, a total of 30 entries are included in the new test set. Then, we perform the model on this new test set to observe whether the generated definitions are differentiated in line with their specified complexity.

6.3 Evaluation Metrics

In order to better analyze and quantify the experimental results, we select three evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and HSK, which are used to comprehensively evaluate the quality and complexity level of generated definitions.

BLEU BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) was originally proposed for the evaluation of machine translation research. The core of BLEU is to separately calculate the N-gram in the generated and the reference sentence, and then compare them one by one to count the times that can be matched. The higher times illustrate higher accuracy. However, the shorter reference segment always leads to more co-occurrence times, which means the shorter generated definitions tend to get a higher BLEU score.

NIST On the basis of BLEU, NIST (National Institute of Standards and Technology) (Doddington, 2002) adds the calculation of the information weight of N-gram. While the BLEU simply sums up the number of N-grams, the NIST sums up the information weights and then divides it by the number of N-gram segments in the whole sentence. In this way, the weightage of those N-grams which appear less frequently will be heavier.

HSK As mentioned in section 4.3, HSK is a test set to evaluate the Chinese proficiency and application ability of non-native Chinese speakers. Based on the purpose of assisting CFL learners to understand Chinese well, we select HSK to measure the complexity level of definitions. Besides, we set seven difficulty levels (scores) of HSK and each of them corresponds to a vocabulary. The final level of a definition is determined by the average score of its segments.

6.4 Results and Analysis

Regardless of complexity levels We report the experimental results on the entire COMPILING dataset in Table 5. The results show that PLMs outperforms the other two methods in terms of the BLEU and NIST scores apparently. However, the results of BERT and BART models diverged on these two metrics. Since NIST assigns different weights to tokens, we believe it better reflects the model's performance. We confirmed this by reading the generated samples. We also notice that as the model performance im-

Table 5: Experiment results on the COMPILING dataset.

Models	Dev			Test		
	BLEU	NIST	HSK	BLEU	NIST	HSK
LOG-CaD	27.66	25.55	3.74	27.71	27.88	3.85
Transformer	28.61	25.85	3.92	28.58	31.00	3.96
BERT	**32.95**	29.66	4.05	**32.03**	30.56	4.08
BART	29.49	**36.90**	**4.76**	30.63	**42.79**	**4.80**

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

proves, so does the average HSK level of the generated definitions. This phenomenon is because simpler words are used more frequently, and hence are more easily learned by models. As the modeling ability improves, the better-performing models learn to use more complex words. This can be challenging for future complexity controllable definition generation works, i.e., improving the performance and reducing the generation complexity at the same time.

Complexity specific models Table 6 illustrates experiment results on three different subsets. As listed in the table, we not only test on the subset in which the model is trained, but also on other subsets. Generally, all the models perform best on the subset it was trained, and poorly on other subsets. Moreover, the performance decays as the complexity level between the model and data increases. Definitions with different complexity have different lexical and syntax, resulting in poor cross-complexity generalization. Besides, we found that even on different test sets, definitions generated by the same model have similar complexity.

Table 6: Experiment results in terms of complexity controllable generation on three test sets.

Models	Easy Set			Medium Set			Hard Set		
	BLEU	NIST	HSK	BLEU	NIST	HSK	BLEU	NIST	HSK
BART-Easy	**32.44**	**64.40**	2.40	21.56	27.61	2.73	25.89	7.95	2.74
BART-Medium	22.92	24.59	4.70	**27.69**	**40.68**	4.86	29.37	16.09	5.01
BART-Hard	22.49	3.55	**8.46**	23.70	7.04	**8.45**	**46.57**	**18.22**	**8.76**

Unified model based on prompt learning MT5-base (Xue et al., 2021) was selected as the benchmark model in this experiment. The best PPL obtained from the definitions generated on the validation set is 38.44. The BLEU and NIST of the model on the test set are 27.42 and 4.66, respectively. The model generates interpretations based on the new test mentioned in Section 6.2. Table 7 lists two examples where it is fairly obvious that the resulting definitions are differentiated and conform to the expectations for their specified complexity levels. To evaluate the complexity of generating definitions more accurately, we adopt automatic evaluation, ranking the difficulty of each group[4]. The automatic evaluation is based on the Chinese Text Complexity Analysis Platform (CTAP)[5] (Cui et al., 2022). We selected the features of word diversity and word density that reflect the difficulty of paraphrases and calculated the scores of definitions in each group based on the above features. Finally, the scatter distribution diagram is shown in Figure 2. It can be seen that the complexity score of the Hard Group is mainly above 5,

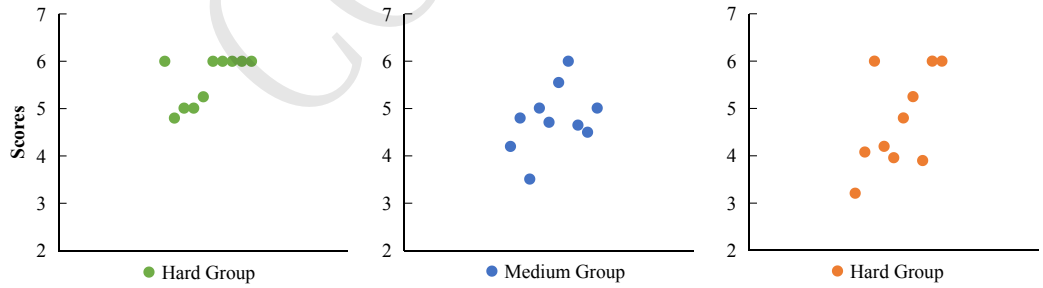


Figure 2: The automatic evaluation results. For example, the scatters of the Hard Group represent those definitions that are specified as the hardest, and the ordinate corresponds to the scores obtained by the automatic rating.

and the number of definitions with the highest score is the largest. The definition in the Easy Group scored the lowest overall score. This means the difficulty level of the model-generated interpretations

[4]Each group of data refers to one original entry and its two copies, their specified complexity of definition is different and other information keep the same.

[5]http://ctap.wenmind.net

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

obtained by automatic evaluation is roughly in line with expectations. The result proves the effectiveness of prompt learning on complexity controllable task, but since the difference in the overall distribution of scattered points in each group in the figure is not particularly obvious, it also reflects that there is room for exploration and improvement of this task in the future.

7 Conclusion

In this work, we propose a novel task of generating Chinese complexity controllable definitions for a given word and example sentence. This task is of great use in helping CFL learners and low literacy readers. Meanwhile, we introduce the COMPILING dataset, which is a benchmark adapting to kinds of definition generation tasks. We also provide several baselines for this task, among which the prompt learning method better assist models in generating definitions with specified complexity. Nevertheless, the experimental results also show that this task is challenging, and the performance needs further improvement.

Acknowledgments

This work was supported by the funds of Research Project of the National Language Commission (No. ZDI135-131, No. ZDI145-24) and Fundamental Research Funds for the Central Universities in BLCU (No. 21PT04). We would like to thank all anonymous reviewers for their valuable comments and suggestions on this work.

References

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of LREC*.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of IJCAI*.

Yue Cui, Junhui Zhu, Liner Yang, Xuezhi Fang, Xiaobin Chen, Yinging Wang, and Erhong Yang. 2022. CTAP for Chinese:A Linguistic Complexity Feature Automatic Calculation Platform. In *Proceedings of LREC*.

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*.

George R. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT*.

Qinan Fan, Cunliang Kong, Liner Yang, and Erhong Yang. 2020. Chinese definition modeling based on bert and beam seach. In *Proceedings of CCL*.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Emosen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry P. Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of ACL*.

Yifan Gao, Lidong Bing, Wang Chen, Jianan Wang, and Irwin King. 2019. Difficulty controllable generation of reading comprehension questions. In *Proceedings of IJCAI*.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph M. Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of EMNLP*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021a. Pre-trained models: Past, present and future. *AI Open*, 2.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021b. Ptr: Prompt tuning with rules for text classification. *ArXiv*, abs/2105.11259.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Chu-Ren Huang, S. Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I. Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet : design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese information processing*, 24.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021a. Definition modelling for appropriate specificity. In *Proceedings of EMNLP*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2021b. Cdm: Combining extraction and generation for definition modeling. *ArXiv*, abs/2111.07267.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2021c. CDM: combining extraction and generation for definition modeling. *CoRR*, abs/2111.07267.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of NAACL*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of ACL*.

Cunliang Kong, Liner Yang, Tianzuo Zhang, Qinan Fan, Zhenghao Liu, Yun Chen, and Erhong Yang. 2020. Toward cross-lingual definition generation for language learners. *CoRR*, abs/2010.05533.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupervised simple definition generation. In *Proceedings of ACL*.

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*.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on DLNLP*.

Thanapon Noraset, Chen Liang, Lawrence Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of AAAI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. VCDM: Leveraging Variational bi-encoding and Deep contextualized Word Representations for Improved Definition Modeling. In *Proceedings of EMNLP*.

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 *Proceedings of NIPS*.

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 NAACL*.

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

Han 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*, abs/2201.05337.

Yihua Zhang. 2011. Discussion on the Definitions in Chinese Learner's Dictionaries: Comparative Study of Domestic and Foreign Learner Dictionaries (Translated from Chinese). *Chinese Teaching in the World*.

Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *ArXiv*, abs/2109.06513.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021. Decompose, fuse and generate: A formation-informed method for chinese definition generation. In *Proceedings of NAACL*.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 921-931, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Can We Really Trust Explanations? Evaluating the Stability of Feature Attribution Explanation Methods via Adversarial Attack

Zhao Yang[1,2], **Yuanzhe Zhang**[1,2], **Zhongtao Jiang**[1,2],
Yiming Ju[1,2], **Jun Zhao**[1,2], **Kang Liu**[1,2,3*]
[1]School of Artificial Intelligence, University of
Chinese Academy of Sciences / Beijing, 100049, China
[2]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences / Beijing, 100190, China
[3]Beijing Academy of Artificial Intelligence / Beijing, 100084, China
{zhao.yang,yzzhang,zhongtao.jiang}@nlpr.ia.ac.cn
{yiming.ju,jzhao,kliu}@nlpr.ia.ac.cn

Abstract

Explanations can increase the transparency of neural networks and make them more trustworthy. However, can we really trust explanations generated by the existing explanation methods? If the explanation methods are not stable enough, the credibility of the explanation will be greatly reduced. Previous studies seldom considered such an important issue. To this end, this paper proposes a new evaluation frame to evaluate the *stability* of current typical feature attribution explanation methods via textual adversarial attack. Our frame could generate adversarial examples with similar textual semantics. Such adversarial examples will make the original models have the same outputs, but make most current explanation methods deduce completely different explanations. Under this frame, we test five classical explanation methods and show their performance on several stability-related metrics. Experimental results show our evaluation is effective and could reveal the *stability* performance of existing explanation methods.

1 Introduction

Fueled by recent rapid development in deep learning, NLP systems have obtained promising results in several fields, such as medical, law and commerce (Rudin, 2019; Bommasani et al., 2021). However, besides the predicted results, users concern more on how these results are generated (Lipton, 2018). To this end, lots of emphases have been set upon the explanation methods for neural networks (Ribeiro et al., 2016; Li et al., 2016; Simonyan et al., 2013; Bastings et al., 2019).

Although the current explanation methods have increased the transparency of the neural networks and provided explanations as supports for predicted results, most of them ignored important questions: *are these methods reliable and the generated explanations really trustful?* Besides the widely used focused properties of explanation methods, such as faithfulness, plausibility (Adebayo et al., 2018; Jacovi and Goldberg, 2020; Atanasova et al., 2020), readableness (Bastings et al., 2019) and compactness (Miller, 2019; Jiang et al., 2021), we believe *stability* is an important but often overlooked property (Robnik-Šikonja and Bohanec, 2018). When we put a small perturbation on the input, which would not change the input semantic and the output of the original model, we believe that the explanation method is not stable enough when we obtain the same outputs with quite different explanations. For example, Figure 1 shows all results of major explanation methods would change when we just replace `fine` by `refined`, including LIME (Ribeiro et al., 2016), Leave-one-out (Li et al., 2016), Vanilla Gradient (Simonyan et al., 2013), Smooth Gradient (Smilkov et al., 2017), Integrated Gradient (Sundararajan et al., 2017).

To fulfill the *stability* testing, we intuitively consider existing word-substitution based textual adversarial attack methods[0](Ren et al., 2019; Zang et al., 2020), since it is under the black-box[1] settings and

*Corresponding author.
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

[0]Feature attribution based explanation methods show the importance of each token to the prediction. Therefore, paraphrase-based attack methods do not fit because they would modify too many parts of inputs at once.

[1]Black-box refers to we can only utilize the outputs of the model during the attack. However, some explanation methods are

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

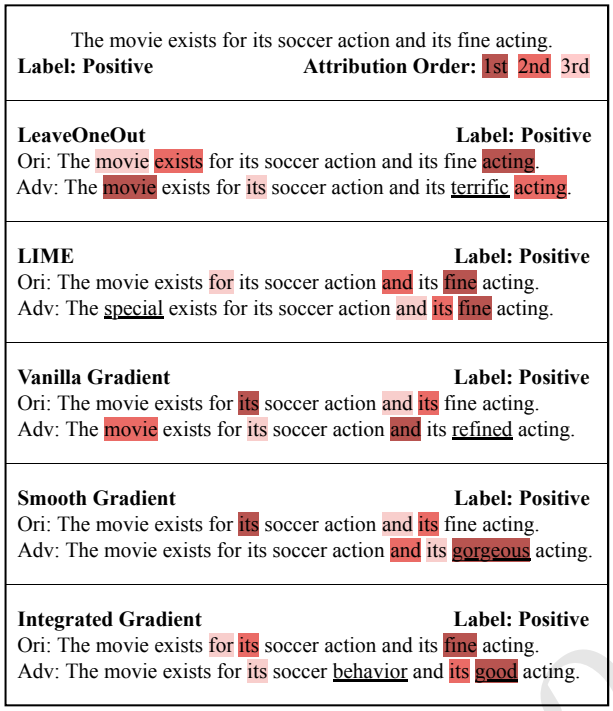


Figure 1: An example of the result of our adversarial attack. We select a sentence from SST-2 and show the adversarial examples for explanation method **Vanilla Gradient** (Simonyan et al., 2013). `Ori` and `Adv` stand for original sentence and corresponding adversarial example respectively. We show the three most important tokens and sign them in different colors.

no need for the transparency of the model framework. However, we could not directly extend the current adversarial attack on the explanation methods. In our explanation stability test setting, the attack method should ensure the original prediction model has unchanged outputs for the adversarial examples, but the explanations vary, which is obviously different from the target of the common textual adversarial attacks. Thus, the main challenge is, for such adversarial examples, how to ensure the explanations are different but the outputs of the original model are the same. To this end, we modified the target of the standard textual adversarial attack to keep the prediction label of the adversarial examples unchanged. At the same time, we define two criteria to measure the difference between two explanations and add them respectively to the score function. Such explanation difference measurements are used to help the judgment of the adversarial examples' qualities in the attacking procedure.

Finally, we put the attack on five typical feature attribution explanation methods. Experimental results show their performance on *stability*. We find perturbation-based explanation methods perform better on *stability* than gradient-based methods. All of the source code and data will be available soon.

2 Related Work

2.1 Feature Attribution Explanation Method

Feature attribution explanation methods score each token of the input based on its contribution to the prediction label. We can easily find the key tokens according to the attribution value. These explanation methods can be simply classified as below two categories: perturbation-based methods and gradient-based methods.

Perturbation-based get the attribution score by perturbing the input sequence: **LIME** (Ribeiro et al., 2016) sampled enough new sequences from the neighbor of the input sequence and fit the output logits of these sampled sequences by a linear function, the coefficients of the fitted function are the attribution

not black-box such as gradient-based methods. Whether the explanation method is black-box has nothing to do with our black box attack method.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

score for each token. **Leave-one-out** (Li et al., 2016) observed the probability change on the predicted class when erasing some certain word and the value of probability change is the attribution score for the removed word. Gradient-based methods compute the attribution score according to the gradient of the input: **Vanilla Gradient** (Simonyan et al., 2013) simply computed the gradient of the loss with respect to each token. **Smooth Gradient** (Smilkov et al., 2017) added small Gaussian noise to every embedding and take the average gradient value as the final attribution score for each token. **Integrated Gradient** (Sundararajan et al., 2017) integrated the gradient along the path from a sequence of all-zero embeddings to the original input and take the integral value as the attribution score.

2.2 Evaluation of Explanation Methods

Recently, a collection of explanation methods has emerged exploring to interpret neural networks. To compare these explanation methods, various explanation metrics have been proposed. Faithfulness refers to how accurately the explanation reflects the true reasoning process of the model (Herman, 2017; Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020). Plausibility refers to how convincing the explanation is to humans by comparing explanations that generated by explanation methods and human annotated explanations (Atanasova et al., 2020; DeYoung et al., 2019). Besides, readableness measures whether human could understand the explanations (Molnar, 2020) and compactness requires a explanation should be short or selective (Miller, 2019; Jiang et al., 2021). However, these evaluation metrics ignore whether the explanation method is reliable.

To evaluate the reliability of existing explanation methods, *consistency* and *stability* have been proposed. However, *consistency* is quite different from *stability* actually. To evaluate *consistency*, existing studies usually modified original model to generate different explanations when the inputs and outputs keep unchanged. Jain and Wallace (2019) modified the attention value and maintain the output unchanged to illustrate attention is not explanation. Heo et al. (2019) applied adversarial model manipulation to generate different explanations. Slack et al. (2020) aims to sample based explanation methods. They modified the original classifier into two parts: original classifier for original instances and another model for instances in neighbor. Wang et al. (2020) construct a new model which has similar outputs with original model but definitely different gradient. They added this model on original model and the added model shows similar prediction but totally different gradient-based explanations. Indeed, they all try to modified the original model to generate different explanations. However, for *stability*, we just put perturbation on inputs not on model, which is extremely different with *consistency*.

For *stability*, though existing works defined its specific meanings, only a few work design corresponding experiments to evaluate the performance of *stability*. (Ghorbani et al., 2019) applied pixel-level perturbations to evaluate the stability. However, pixel-level perturbations can not be easily transferred in NLP. In NLP only (Ding and Koehn, 2021) evaluated this property by manually constructing similar instances, which is much time-consuming and expensive. Therefore, in this paper, we automatically construct similar instances by learning from textual adversarial attack.

3 Formulation

In this section, we first introduce the basic information of the common textual adversarial attack in Section 3.1. Then we introduce how to formulate explanation adversarial attack in Section 3.2.

3.1 Textual Adversarial Attack

Formally, suppose that a sentence $x_k = \omega_1\omega_2\cdots\omega_n$, where ω_i is the i-th word in x_k. For a given classifier $P(y|x)$ and label set $Y = (y_1, y_2, ..., y_m)$, the model prediction y_k for x_k can be formulated as $y_k = arg\max_{y\in Y} P(y|x_k)$. The target is to find $x_k^{'}$, which can be formulated as:

$$\begin{aligned} & x_k^{'} \\ s.t. \quad & y_k \neq y_k^{'}, \left\| x_k^{'} - x_k \right\| < \epsilon \end{aligned} \tag{1}$$

where $x_k^{'}$ is the adversarial example of x_k. The core constraint is to ensure the difference between x_k and $x_k^{'}$ is small enough. In this paper, we ensure the semantics of x_k and $x_k^{'}$ to be as similar as possible,

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

which has been shown more imperceptible for human (Zhang et al., 2020).

3.2 Explanation Adversarial Attack

Feature attribution explanation method can generate an explanation $e_k = (s_1, s_2, \cdots, s_n)$ according to x_k and its prediction y_k, where s_i is the attribution score of ω_i. Therefore, the target is to find $x_k^{'}$, which can be formulated as follow:

$$\begin{aligned} & x_k^{'} \\ s.t. \qquad & e_k \neq e_k^{'}, y_k^{'} = y_k, \left\| x_k^{'} - x_k \right\| < \epsilon \end{aligned} \tag{2}$$

We also follow the common textual adversarial attack to keep the semantics of x_k and $x_k^{'}$ to be similar. And the most important difference is an extra constraint $y_k^{'} = y_k$, we must ensure this constraint should be satisfied because of the definition of *stability*. Obviously, the constraint is contrary to the target of common textual attack, where $y_k^{'} \neq y_k$. By contrast, our target is to ensure the explanations are different. Therefore, we will define how to measure explanation difference in the following section.

4 Attack Method

According to Section 3.2, we need to measure the explanation difference. Therefore, we propose two metrics in Section 4.1. Then we present our detailed attack strategies to attack existing explanation methods in Section 4.2.

4.1 Measuring the Explanation Difference

For feature attribution methods, people usually do not care the specific attribution score of each token but the relative importance ranks of these tokens. Therefore, we consider the rank differences between explanations. We can easily get the corresponding rank sequence R_k for explanation E_k in descending order, where $R_k = (r_1^k, r_2^k, ..., r_n^k)$, r_i^k stands for the descending rank of the i-th token in x_k. We can also get the corresponding position sequence $P_k = (p_1^k, p_2^k, ..., p_n^k)$ via `argsort`, p_i^k stands for the index of the i-largest attribution score in x_k. Based on this, we design two quantitative criteria to measure the difference between explanations.

Rank-count: In this setting, we compute the number of positions whose rank has changed:

$$d_{count}(E_i, E_j) = \sum_{k=1}^{n} ||r_k^i - r_k^j||_0 \tag{3}$$

where $||\cdot||_0$ refers to the L0 norm.

Rank-topk: In this setting, we compute the size of intersection set of two position set of the top-k rank. The top-k set for e_i is the first k elements of position sequence r_i: $E_{topk}^i = \{p_1^i, p_2^i, ..., p_k^i\}$.

$$d_{topk}(E_i, E_j) = |E_{topk}^i \cap E_{topk}^j| \tag{4}$$

where $|\cdot|$ refers to the size of a set.

For example, given $E_1 = \{0.1, 0.5, 0.3, 0.2\}$ and $E_2 = \{0.6, 0.3, 0.4, 0.2\}$. We get the rank sequence $R_1 = \{3, 0, 1, 2\}$ and $R_2 = \{0, 2, 1, 3\}$, then we can get the position sequence $P_1 = \{1, 2, 3, 0\}$ and $P_2 = \{0, 2, 1, 3\}$. Accordingly, we compute $d_{count}(E_1, E_2) = 3$ and $d_{topk}(E_1, E_2) = 2$ when $k = 3$.

4.2 Attack Strategies

Word-substitution based textual adversarial attack methods usually consist of two main steps: determining substitution order and selecting substitution words. In different steps, we employ different strategies. To determine the substitution order, we modify Samanta and Mehta (2017) as an example. To select substitution words, we utilize OpenHowNet (Qi et al., 2019) as the substitution resource (Zang et al., 2020). Notably, other word-substitution based adversarial attack methods (Ren et al., 2019; Alzantot et al., 2018; Zang et al., 2020) are also applicable.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

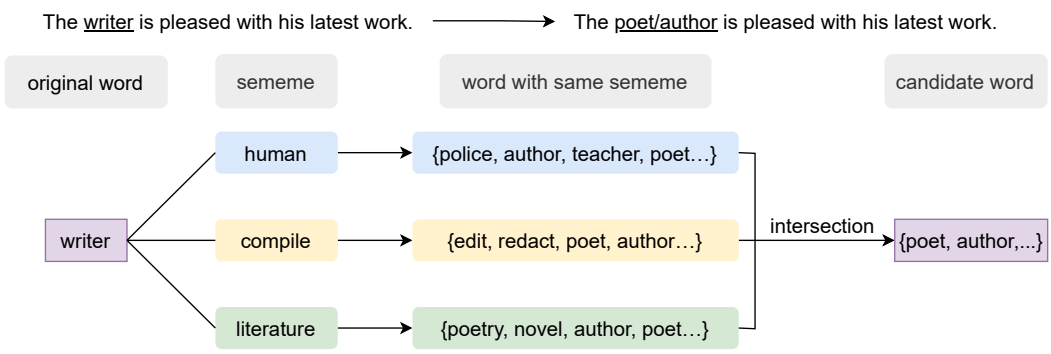


Figure 2: An example of how to construct candidate substitution word set for the word `writer` by its sememes `human`, `compile` and `literature`.

4.2.1 Determining Substitution Order

Formally, for a sentence $x = \omega_1\omega_2\cdots\omega_i\cdots\omega_n$, to determine the substitution order, we compute the word saliency WS_i for token ω_i first. To compute WS_i , we should get $\hat{x}_i = \omega_1\omega_2\cdots\ \mathbf{0}\ \cdots\omega_n$ by replacing ω_i with $\mathbf{0}$.

$$WS_i = P(y_{ori}|x) - P(y_{ori}|\hat{x}_i) \tag{5}$$

where y_{ori} refers to the original output label. We calculate the word saliency WS_i for all $\omega_i \in x$ and then we sort all of the tokens in descending order based on their saliency value. Then we substitute the words in this order (Samanta and Mehta, 2017).

4.2.2 Selecting Substitution Words

We construct candidate substitution set via sememes and utilize OpenHowNet (Qi et al., 2019) as the resource. Sememe is the minimum semantic unit of language (Bloomfield, 1926) and the sememes of one word can composite the meaning of this word. Therefore, words that have the same sememe can substitute for each other (Zang et al., 2020). As shown in Figure 2, when we want to find substitution words for the original word `writer`. We utilize OpenHowNet to get its sememes `human`, `compile` and `literature`. Then we get three word sets that has these three sememes respectively. Finally, we compute the intersection of these three word sets and get the substitution word `poet` and `author` for the original word `writer`. According to Qi et al. (2019) and Zang et al. (2020), when we replace the word with the obtained substitution word, the semantic of the original sentence would not change.

After getting substitution set for the original word by above method, we still have to choose which word to substitute the original word. Therefore, we also need a quantitative criterion to help us to find the most suitable substitution word from the whole substitution set. Specifically, we define our score function as follow:

$$score(x_1, x_2) = d(e_1, e_2) \times (1 - ||y_1 - y_2||_0) \tag{6}$$

where $d(e_1, e_2)$ represent the explanation difference for x_1, x_2 and we directly employ the Equation (3) and Equation (4). y_1, y_2 are the prediction label for x_1, x_2. We directly force the labels must be same, otherwise the score would be zero.

With this score function, we can get the substitution word ω_i^* for ω_i in $x_i = \omega_1\omega_2\cdots\omega_i\omega_n$. This process can be formulate as follow:

$$\omega_i^* = \arg\max_{\omega_i \in L_{\omega_i}} score(x, x_i^{'}) \tag{7}$$

where $x_i^{'} = \omega_1\omega_2\cdots\omega_i^{'}\cdots\omega_n$ and L_{ω_i} is the candidate set for the word ω_i. Finally, ω_i^* is the substitution word for ω_i is x.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

5 Experiments

5.1 Datasets and Models

Following previous explanation studies (DeYoung et al., 2019; Atanasova et al., 2020), we also select sentiment analysis as the target task. In specific, we choose SST-2 (Socher et al., 2013) and IMDB (Maas et al., 2011) as the test benchmark dataset and select the base version of BERT (Devlin et al., 2018) and BiLSTM (Conneau et al., 2017) as the target model.

For BERT, we utilize the base version of BERT. For BiLSTM, the hidden states are 256-dimensional and we utilize the 300-dimensional pre-trained Glove (Pennington et al., 2014) word embeddings. Our reproduced BERT can achieve accuracy of 91.28% and 91.36% on SST-2 and IMDB respectively. And BiLSTM can achieve accuracy of 85.50% and 90.38% on SST-2 and IMDB respectively.

To improve evaluation efficiency, we randomly sample 500 correctly classified instances with the length of 10-100 from the test set.

5.2 Explanation Methods

We select five classical feature attribution explanation methods in the two mainstream types to conduct our experiments:

A. Perturbation-based Explanation Method:

LIME (Ribeiro et al., 2016) sampled enough sentences from the neighbor of the input and fit the output logits of these samples by a linear function. The coefficients of the obtained linear function is the corresponding attribution scores.

LeaveOneOut (LOO) (Li et al., 2016) observed the probability change on the predicted class when erasing each word one by one and take this change value as the atribution score.

B. Gradient-based Explanation Method:

VanillaGradient (VG) (Simonyan et al., 2013) simply computed the gradient of the model loss with respect to the token and multiply with its embedding as its corresponding attribution score.

$$a_i = x_i \cdot \frac{\partial f(x_i)}{\partial x_i} \tag{8}$$

SmoothGradient (SG) (Smilkov et al., 2017) added small Gaussian noise to every embedding N times and average these N VanillaGradient value as the final attribution score.

$$a_i = \frac{1}{N}\sum_{i=1}^{N}(x_i + \mathcal{N}(0,1)) \cdot \frac{\partial f(x_i + \mathcal{N}(0,1))}{\partial(x_i + \mathcal{N}(0,1))} \tag{9}$$

where $\mathcal{N}(0,1)$ refers to the Gaussian noise.

IntegratedGradient (IG) (Sundararajan et al., 2017) integrated the gradient along the path from a basic sequence $x_i^{'}$ to the original input x_i and take the integral value a_i as the attribution.

$$a_i = (x_i - x_i^{'}) \int_{\alpha=0}^{1} \frac{\partial f(x_i^{'} + \alpha \times (x_i - x_i^{'}))}{\partial \alpha} d\alpha \tag{10}$$

Specifically, it is time-consuming to compute intergral value. To improve computation efficiency, we divide the integral area into K parts and obtain the approximate value of a_i (Sundararajan et al., 2017).

$$a_i = (x_i - x_i^{'}) \sum_{m=1}^{K} \frac{\partial f(x_i^{'} + \frac{m}{K} \times (x_i - x_i^{'}))}{\partial x_i} \times \frac{1}{K} \tag{11}$$

5.3 Experimental Settings and Results

5.3.1 Explanation Similarity

Firstly, we fix m modified words to generate corresponding adversarial examples whose explanations are the most different. Then we use explanation similarity to evaluate the stability of explanation methods.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Dataset	Explanations	m=1			m=2			m=3		
			change↓	*spearman*↑	*inte*↑	*change*↓	*spearman*↑	*inte*↑	*change*↓	*spearman*↑	*inte*↑
BERT	SST-2	LIME	**79.87**	**0.80**	**3.87**	**84.03**	**0.78**	**3.81**	**86.52**	**0.76**	**3.75**
		LOO	89.13	0.64	3.14	92.62	0.62	3.09	94.12	0.61	3.03
		VG	92.99	0.48	2.83	95.65	0.45	2.71	97.11	0.42	2.64
		SG	92.86	0.55	2.92	95.71	0.53	2.87	96.70	0.52	2.83
		IG	86.79	0.71	3.45	90.01	0.69	3.38	91.69	0.67	3.37
	IMDB	LIME	**84.60**	**0.92**	**4.23**	**88.65**	**0.90**	**4.08**	**90.04**	**0.88**	**3.87**
		LOO	90.10	0.84	3.48	93.47	0.79	3.12	95.22	0.76	2.91
		VG	92.75	0.79	3.23	95.44	0.73	2.88	96.65	0.69	2.66
		SG	92.48	0.82	3.29	95.26	0.76	2.89	96.60	0.73	2.67
		IG	85.49	0.91	4.07	89.58	0.89	3.90	91.37	0.87	3.81
BiLSTM	SST-2	LIME	**71.18**	**0.81**	**4.02**	**80.38**	**0.74**	**3.78**	**84.22**	**0.68**	**3.63**
		LOO	75.76	0.77	3.89	84.07	0.71	3.70	86.96	0.67	3.60
		VG	78.20	0.75	3.78	85.04	0.62	3.52	88.50	0.56	3.36
		SG	77.83	0.77	3.85	84.49	0.68	3.55	87.21	0.64	3.40
		IG	73.55	0.79	3.99	81.73	0.72	3.75	85.39	0.67	3.61
	IMDB	LIME	**81.44**	**0.90**	**4.24**	**86.36**	**0.86**	**4.07**	**88.25**	**0.84**	**3.92**
		LOO	84.96	0.86	4.11	89.48	0.82	3.91	90.78	0.81	3.85
		VG	86.25	0.85	3.72	90.42	0.80	3.41	91.88	0.77	3.27
		SG	86.22	0.86	4.08	90.00	0.81	3.89	91.45	0.79	3.80
		IG	82.80	0.88	4.21	87.41	0.84	4.02	89.19	0.83	3.89

Table 1: Results of similarity of explanations between original instances and their adversarial examples by replacing m words for BERT and BiLSTM. *change* is defined as the percentage of positions whose corresponding ranks have changed. *spearman* is the spearman's rank order correlation between two explanations. *inte* is defined as the size of the intersection of the 5 most important tokens before and after perturbation.

More stable explanation methods could get higher explanation similarity. In specific, we employ three specific criteria including *change*, *spearman* and *inte*. *change* refers to the percentage of positions whose corresponding rank has changed, *spearman* refers to the spearman's rank order correlation efficient between the ranks of two explanations (Spearman, 1961), and *inte* refers to the size of the intersection of the 5 most important tokens before and after perturbation (Ghorbani et al., 2019). Table 1 presents the experimental results of the five explanation methods that conduted on BERT and BiLSTM on the two datasets SST-2 and IMDB.

To evaluate *stability*, following its definition, we should ensure the same output and keep semantics of adversarial examples unchanged. For output consistency, we test the consistency of predictions between all test instances and their adversarial examples, which can achieve 100%. It means our methods satisfy the requirement of the same outputs. As for input semantic consistency, we perform human evaluation to check the semantic similarity between the adversarial example and the original example. Specifically, We invite 4 postgraduates score ranges 1 to 3 according to the semantic similarity between original instances and their adversarial examples. Scores of 1,2 and 3 indicate low, medium and high semantic similarity, respectively. Higher scores mean better consistency. Table 2 shows the results of human evaluation. These results show that our generated examples could keep semantics unchanged. Therefore, our experiment satisfies the definition of *stability* and the experimental results in Table 1 are convincing.

From the experimental results in Table 1, we find the *stability* performance of the five typical explanation methods keep same on different models and different datasets. And the *stability* performance (from good to bad) of these explanation methods is as follow: **LIME**, **Integrated Gradient**, **LeaveOneOut**, **Smooth Gradient**, **Vanilla Gradient**.

According to the results for different m in Table 1, when we replace more words, explanation difference obviously increases. However, from the human evaluation results in Table 2, we find the semantic consistency also decreases as m increases. Therefore, one thing must be pointed out, to satisfy the semantic consistency of input, we should control the modification rate when we evaluate the *stability* of explanation methods.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Dataset	Explanation	m=1	m=2	m=3
BERT	SST-2	LIME	2.75	2.48	2.23
		LOO	2.74	2.46	2.18
		VG	2.73	2.42	2.12
		SG	2.74	2.44	2.14
		IG	2.75	2.47	2.21
	IMDB	LIME	2.82	2.67	2.41
		LOO	2.79	2.63	2.36
		VG	2.77	2.60	2.34
		SG	2.77	2.61	2.33
		IG	2.80	2.65	2.39
BiLSTM	SST-2	LIME	2.76	2.48	2.25
		LOO	2.73	2.44	2.19
		VG	2.72	2.41	2.13
		SG	2.72	2.44	2.16
		IG	2.75	2.46	2.23
	IMDB	LIME	2.81	2.67	2.37
		LOO	2.75	2.47	2.18
		VG	2.74	2.44	2.15
		SG	2.74	2.46	2.16
		IG	2.75	2.50	2.22

Table 2: Results of human evaluation. The human evaluation score is not an objective metric and the higher score does not stand for the better method. We list it here just to show the adversarial examples in Table 1 keep the semantic unchanged.

5.3.2 Attack Success Rate

Secondly, following the common textual adversarial attack, We design a series of success conditions to check the attack success rate for different explanation methods. Combining with the finding in Section 5.3.1 that we should control the modification rate when evaluating *stability*, we set the maximum modification rate 20%. And existing textual adversarial attack also usually control the modification rate less than 20% (Ren et al., 2019; Alzantot et al., 2018; Zang et al., 2020).

Then we illustrate our formulated success conditions. We utilize the quantitative criteria introduced in Sec 4.1 and then define the success conditions as $d_{count} > \alpha * length$ and $d_{topk} < \beta$ for different α, β. $d_{count} > \alpha * length$ refers to the proportion of positions whose ranks have changed in should bigger than α and we select α from $\{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$. $d_{topk} < \beta$ refers to the size of intersection of the top-5 important tokens should smaller than β and we choose β from $\{1, 2, 3, 4, 5\}$. Obiviously, bigger α and smaller β mean more difficult success conditions, and a smaller attack success rate on the same condition means a more stable explanation method. Given a sentence, if achieving the success condition with the modification rate less than 20% , we define this is a successful attack. Otherwise, when the success condition can not be achieved even on the maximum modification rate, we define this is a unsuccessful attack. Then we calculate the corresponding attack success rate on all examples.

Figure 3 shows the results of BERT on SST-2. Under the two type of success conditions, we find the relative rank of the five explanation methods appears the same. And more difficult success condition would cause lower attack success rate. The *stability* performance (from good to bad) is the same as the results in §5.3.1: **LIME**, **Integrated Gradient**, **LeaveOneOut**, **Smooth Gradient**, **Vanilla Gradient**.

In summary, in our different experiment settings (Table 1 and Figure 3), all experimental results consistently show that the *stability* performance (from good to bad) of the five methods is as follows: **LIME**, **Integrated Gradient**, **LeaveOneOut**, **Smooth Gradient**, **Vanilla Gradient**. Besides, we also observe perturbation-based methods have better performance on *stability* than gradient-based methods.

6 Discussion

Beyond the above experiments, our discussions would address the following research questions:

- **RQ1** How do the evaluation results change when replacing the two steps in the proposed attack

CCL 2022

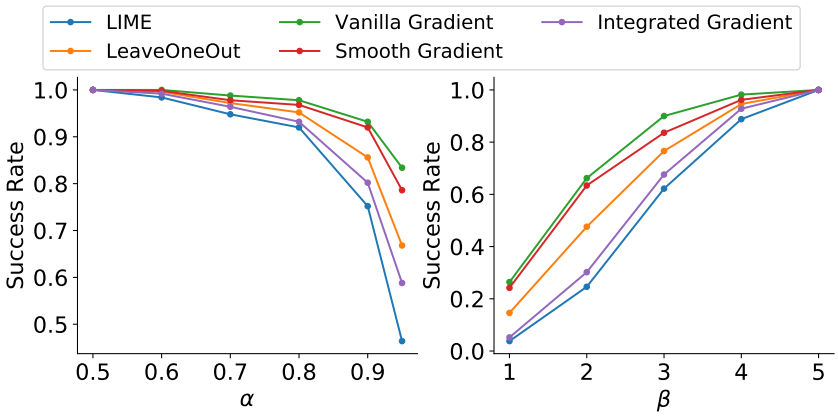


Figure 3: Success rate for different success conditions. Left part shows the condition $d_{count} > \alpha * length$ for $\alpha \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$. Right part shows the condition $d_{topk} < \beta$ for $\beta \in \{1, 2, 3, 4, 5\}$. Success rate is the percentage of instances whose explanation difference could satisfy the condition. Bigger α and smaller β indicate more different explanations. A smaller success rate on the same success condition indicates a more stable method.

	m=1						m=2					
	change↓		*spearman*↑		*inte*↑		*change*↓		*spearman*↑		*inte*↑	
	ori	rand	ori	rand	ori	rand	ori	rand	ori	rand	ori	rand
LIME	**79.87**	**76.00**	**0.80**	**0.84**	**3.87**	**4.03**	**84.03**	**82.71**	**0.78**	**0.79**	**3.81**	**3.89**
LOO	89.13	84.25	0.64	0.76	3.14	3.48	92.62	90.40	0.62	0.69	3.09	3.25
VG	92.99	89.82	0.48	0.62	2.83	3.20	95.65	94.58	0.45	0.55	2.71	2.99
SG	92.86	89.13	0.55	0.65	2,92	3.23	95.71	94.20	0.53	0.55	2.87	2.99
IG	86.79	79.39	0.71	0.80	3.45	3.89	90.01	86.12	0.69	0.75	3.38	3.69

Table 3: Results of explanation similarity for BERT on SST-2. `ori` refers to the results based on the word substitution order in §4.2.1 and `rand` refers to the results based on the random substitution order.

strategy with othe existing methods?

- **RQ2** How can we improve the stability of explanation methods?

6.1 Correlation Analysis Between The Two Attack Steps and The Evaluation Results

To address **RQ1**, we modify the two steps in Section 4.2 to conduct experiments in the following parts:

Effect of Substitution Order To verify whether the other substitution order is effective to evaluate the *stability* of explanation methods, we utilize a random order to replace the substituion order in Section 4.2.1. Specifically, following experiments settings in Section 5.3.1, we select SST-2 and conduct experiments on BERT model. To improve efficiency, we only choose $m = 1$ and $m = 2$.

Table 3 shows the corresponding results. Compare to results in Table 1, all of the attack performance have dropped. In specific, for same explanation method on same setting, the *change* metric decreases and the $spearman$ and $inte$ metrics both increases, which stands for the higher explanation similarity. And this is consistent with the common textual adversarial attack, which has been shown the random order would much decrease the attack performance (Ren et al., 2019). Besides, we find the *stability* performance of these five explanation methods still keep same as the previous findings.

Effect of Substitution Set To verify whether the other substitution set is effective, we utilize WordNet (Miller, 1995) to construct substitution word set. We can easily find synonyms for a given word via WordNet. Following experiments settings in Section 5.3.1, we select IMDB and conduct experiments on BiLSTM model. To improve efficiency, we also only choose $m = 1$ and $m = 2$.

Similar to replacing the substitution order with random order, the attack performance also drop. And the *stability* performance of these five explanation methods also keep same.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	m=1						m=2					
	change↓		*spearman*↑		*inte*↑		*change*↓		*spearman*↑		*inte*↑	
	ori	WN	ori	WN	ori	WN	ori	WN	ori	WN	ori	WN
LIME	**81.44**	**78.89**	**0.90**	**0.92**	**4.24**	**4.41**	**86.36**	**83.21**	**0.86**	**0.89**	**4.07**	**4.09**
LOO	84.96	82.18	0.86	0.89	4.11	4.18	89.48	86.32	0.82	0.85	3.91	3.98
VG	86.25	83.79	0.85	0.87	3.72	4.02	90.42	88.14	0.80	0.83	3.41	3.85
SG	86.22	83.72	0.86	0.87	4.08	4.14	90.00	87.97	0.81	0.84	3.89	3.95
IG	82.80	79.97	0.88	0.90	4.21	4.27	87.41	84.56	0.84	0.87	4.02	4.05

Table 4: Results of explanation similarity for BiLSTM on IMDB. `ori` refers to utilizing OpenHowNet to construct substitution set and `WN` refers to utilizing WordNet to construct substitution set.

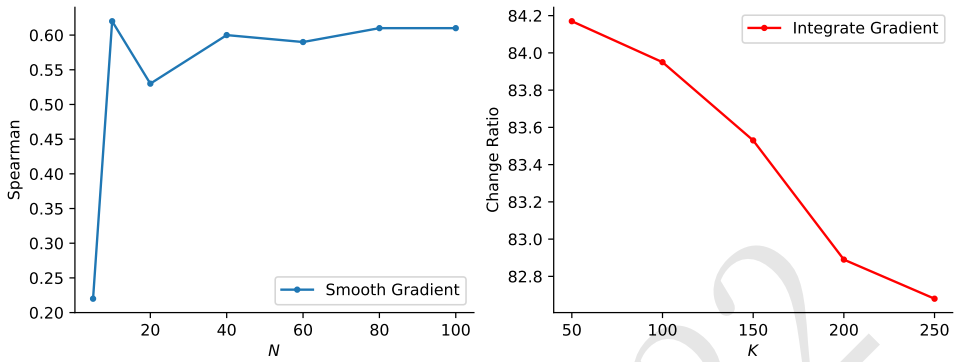


Figure 4: The left figure shows the relation between Spearman's rank order correlation and the number of the added noise M in **Smooth Gradient**. The right figure shows the relation between change ratio and the number of the divided parts K in **Integrated Gradient**.

In summary, our evaluation frame is independent to the specific substitution order and how to construct substitution set. These specific steps only influence the attack performance and could get the similar results of existing explanation methods when evaluating *stability*.

6.2 Simply Improving Stability of Explanation Method

To address **RQ2**, we try to explore how to improve the *stability* of two explanation methods.

Adding more noise We explore the influence of the number of the added noise N (Equation (9)) in Smooth Gradient. We select Spearman's rank order correlation as the evaluation metric. Figure 4 (left) shows the results. We find adding appropriate noises is useful and adding more noises is not meaningful.

More robust mechanism Integrated Gradient is a more robust mechanism compared to Simple Gradient and Smooth Gradient, because it satisfy *sensitivity* and *implementation invariance* these two important axiom (Sundararajan et al., 2017). We explore the influence of the divided parts K in Equation (11). Figure 4 (right) shows the results of change rate. We find adding the number of the divided parts K is useful. The bigger K is, the more accurate the integral value is, which means more robust mechanism. Therefore, more robust mechanism could improve the *stability* of explanation methods.

Therefore, we can try to add appropriate noises and seek more robust mechanisms to make explanation methods more stable. And we take the further exploration of improving *stability* as our future work.

7 Conclusion

This paper proposes a new evaluation frame to evaluate the *stability* of typical feature attribution explanation methods via adversarial attack. Various experimental results on different experimental settings reveal their performance on *stability*, which also show the effectiveness of our proposed evaluation frame. We also conduct experiments to show the proposed frame is dependent of specific step. Therefore, we hope the proposed evaluation frame could be applied to evaluating the *stability* of feature attribution explanation methods in the future and attract more research on this important but often overlooked property.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

8 Limitations

The proposed evaluation frame only focus on the rank of the feature attribution explanation methods. These explanation methods also provide specific attribution scores and these scores may further refine the proposed frame.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61922085, 61831022, 61906196), the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006) and the Youth Innovation Promotion Association CAS. This work was also supported by Yunnan provincial major science and technology special plan projects, under Grant:202103AA080015.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. *arXiv preprint arXiv:2009.13295*.

Joost Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160*.

Leonard Bloomfield. 1926. A set of postulates for the science of language. *Language*, 2(3):153–164.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

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*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. *arXiv preprint arXiv:2104.05824*.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, pages 2925–2936.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Zhongtao Jiang, Yuanzhe Zhang, Zhao Yang, Jun Zhao, and Kang Liu. 2021. Alignment rationale for natural language 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 5372–5387, Online, August. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

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.

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marko Robnik-Šikonja and Marko Bohanec. 2018. Perturbation-based explanations of prediction models. In *Human and machine learning*, pages 159–175. Springer.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

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*, pages 1631–1642.

Charles Spearman. 1961. The proof and measurement of association between two things.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of nlp models is manipulable. *arXiv preprint arXiv:2010.05419*.

CCL 2022

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 932-944, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Dynamic Negative Example Construction for Grammatical Error Correction using Contrastive Learning

Junyi He, Junbin Zhuang, and **Xia Li***
Guangzhou Key Laboratory of Multilingual Intelligent Processing,
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China
{junyihe,junbinzhuang,xiali}@gdufs.edu.cn

Abstract

Grammatical error correction (GEC) aims at correcting texts with different types of grammatical errors into natural and correct forms. Due to the difference of error type distribution and error density, current grammatical error correction systems may over-correct writings and produce a low precision. To address this issue, in this paper, we propose a dynamic negative example construction method for grammatical error correction using contrastive learning. The proposed method can construct sufficient negative examples with diverse grammatical errors, and can be dynamically used during model training. The constructed negative examples are beneficial for the GEC model to correct sentences precisely and suppress the model from over-correction. Experimental results show that our proposed method enhances model precision, proving the effectiveness of our method.

1 Introduction

Grammatical error correction (GEC) (Chollampatt and Ng, 2018; Kaneko et al., 2020; Kiyono et al., 2019) aims at correcting texts with different types of grammatical errors into natural and correct forms. It is an important research topic for both natural language processing and language education.

Most of the current GEC systems are developed for correcting writings by learners of English as a second language (Brockett et al., 2006; Chen et al., 2020a; Chollampatt and Ng, 2018; Kaneko et al., 2020; Kiyono et al., 2019). However, GEC for native writings is also worth exploring, as texts written by native speakers may also contain grammatical errors that should be corrected for enhancement of writing quality. Currently, it is not feasible to train a GEC model specifically for correcting native writings because GEC data containing native writings are not sufficient. Therefore, native writings are often corrected by GEC models that are trained on GEC data consisting of writings by non-native speakers such as the Lang-8 (Mizumoto et al., 2011) and NUCLE (Dahlmeier et al., 2013) datasets. However, the error type distribution, error density and fluency are inconsistent between the writings by non-native and native speakers. Therefore, those GEC models may over correct sentences and produce a low precision of error correction (Flachs et al., 2020). In terms of this issue, contrastive learning (CL) (Chen et al., 2020b; Chen et al., 2020c; Gao et al., 2021; Liu and Liu, 2021) can be incorporated to help alleviate the over correction behaviour of the GEC models. The core idea is to take the over-corrected sentences as negative examples, and to effectively avoid or alleviate the problem of over correction by increasing the distance between the anchor sentence and the negative examples. So the focus is on how to construct effective negative examples for training the GEC models effectively. Previous studies about GEC models mainly focus on data augmentation for generating pseudo parallel training pairs as complement for the current insufficient GEC training data (Zhao et al., 2019; Takahashi et al., 2020), or focus on improving the correction performance with a variety of model architectures (Awasthi et al., 2019; Stahlberg and Kumar, 2020; Sun and Wang, 2022), few of them focus on improving the performance of the GEC models with contrastive learning. To the best of our knowledge, Cao et al. (2021) is the only recent work

*Corresponding author: xiali@gdufs.edu.cn
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

for that. In their work, they propose two approaches for constructing negative examples. First, they treat the beam search candidates produced by an off-the-shelf GEC model as negative examples. They find that many output candidates generated by beam search contain erroneous edits, and the constructed negative examples help suppress the trained GEC model from producing erroneous edits. Second, the source sentence is treated directly as a negative example if it contains grammatical errors. Their intuition is that there should be differences between the corrected output sentence and the source sentence, otherwise the GEC model fails to detect any grammatical errors in the source sentence. The negative examples constructed in this way suppress the trained GEC model from outputting the erroneous source sentences as they are without any modifications.

Although the aforementioned study produces good performance, we believe that there are still two points that can be improved: (1) The negative examples constructed with beam search may not be sufficient. Many beam search output candidates are the same as the target sentence and cannot be used as negative examples. That leads to a small number of the generated negative examples. In addition, the beam search candidates may contain unrealistic grammatical errors with a small number of error types, limiting the diversity of grammatical errors in the generated negative examples. As a result, the low diversity of the negative examples makes the GEC model less easier to learn to distinguish the negative examples from the anchor, which limits the improvement of error correction performance brought by contrastive learning. (2) They construct the negative examples with their negative example construction methods before model training. As a result, the GEC model can only be able to see a fixed set of negative examples in each iteration during training, which may reduce the generalization ability of the GEC model.

To this end, we propose a dynamic negative example construction method for grammatical error correction using contrastive learning. The proposed method contains a negative example construction strategy that makes use of realistic grammatical error patterns produced by humans to generate sufficient negative examples with more diverse grammatical errors. With the constructed negative examples, the GEC model can learn to correct sentences precisely and be suppressed from over-correction. Moreover, the proposed strategy is simple and lightweight, enabling it to be applied dynamically during the training process. In this manner, our method enhances the generalization ability of the GEC model.

The main contributions of this work are as follows:

(1) We propose a dynamic negative example construction method for grammatical error correction using contrastive learning. The proposed method can construct sufficient negative examples with diverse grammatical errors, and can be dynamically applied during model training. The constructed negative examples are beneficial for the model to correct sentences precisely and suppress it from over-correction.

(2) We conduct extensive experiments on the public CWEB dataset that contains native writings, and compare our proposed method with existing GEC studies focusing on negative example construction. Experimental results show that our proposed method indeed enhances model precision and suppresses the GEC model from over-correction.

2 Related Work

In this section, we briefly review different GEC methods, including the early rule-based methods, the widely used methods based on machine translation or BERT, and the recently proposed GEC methods using contrastive learning.

GEC Methods based on Rules. Early GEC models are mostly rule-based pattern recognizers or dictionary-based linguistic analysis engines (Macdonald, 1983; Richardson and Braden-Harder, 1988; Sidorov et al., 2013; Sidorov, 2013). These rule-based methods require a set of pre-defined error recognition rules to detect grammatical errors in the input sentences. Once a certain span in the input sentence is matched by a certain rule, the error correction system provides a correction for the matched error.

GEC Methods based on Machine Translation. GEC models based on machine translation have been proposed to "translate" wrong sentences into correct sentences. Brockett et al. (2006) use a noisy channel model in conjunction with a statistical machine translation model for error correction. Felice et al. (2014) propose a hybrid GEC model that integrates grammatical rules and a statistical machine

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

translation model. They also adopt some techniques such as type filtering. Zheng and Ted (2016) apply a neural machine translation model with the attention mechanism to GEC. In addition, they also introduce a method that uses a combination of an unsupervised alignment model and a word-level translation model to solve the problem of sparse and unrecognized words. Chollampatt et al. (2018) integrate four convolutional neural translation models combined with a re-scoring mechanism. Kiyono et al. (2019) construct the GEC model with Transformer and use many data augmentation techniques.

GEC Methods based on BERT. Many studies have introduced the Transformer-based deep bidirectional language model BERT (Devlin et al., 2019) into GEC, hoping that the correction performance can be improved with the help of its rich language knowledge and deep textual understanding ability. Awasthi et al. (2019) modify the BERT structure to predict the edit operation of each word in the source sentence by sequence tagging. Then they apply the edit operations to the source sentence to construct its correct form. Chen et al. (2020a) use BERT to predict and annotate error spans in the input sentence, and subsequently rewrite the annotated error spans with a sequence model. Kaneko et al. (2020) first fine-tune BERT with GEC data, then feed the output representations of the fine-tuned BERT to the GEC model as additional information for error correction.

GEC Methods based on Contrastive Learning. Contrastive learning (Chen et al., 2020b; Chen et al., 2020c; Gao et al., 2021; Liu and Liu, 2021) is a discriminative self-supervised learning method used to enhance the feature representation ability of deep learning models. For any training example, contrastive learning requires automatic construction of examples that are similar (positive examples) and dissimilar (negative examples) to an anchor. And during training, the model needs to reduce the distance between the anchor and the positive examples in the representation space, while to increase the distance between the anchor and the negative examples.

Cao et al. (2021) try to use contrastive learning to improve the error correction ability of the GEC model. Since constructing positive examples is difficult for GEC, they propose a margin-based contrastive loss, which only requires to construct negative examples and does not requires to construct positive examples. Their work is the most similar to ours. In view of the limitations of their negative example construction method, we propose a dynamic negative example construction method to better address the over correction problem of the GEC model in low error density native writings.

3 Our Method

In this section, we first describe the overall architecture of our method in Section 3.1. Then, we will detail the proposed negative example construction strategy in Section 3.2, and the proposed dynamic mechanism in Section 3.3.

3.1 Overall Architecture

As mentioned above, the purpose of this work is to incorporate contrastive learning into the GEC model to effectively alleviate the problem of over correction by increasing the distance between the anchor sentence and the negative examples. We illustrate our method in Figure 1, which consists of two components: the negative example construction component and the contrastive learning component.

Given a training pair $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$ indicates the source sentence that may contain grammatical errors, x_i is the i^{th} word, m is the source sentence length, and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ indicates the target sentence that is grammatically correct, y_j is the j^{th} word, n is the target sentence length. The goal of the GEC task is to correct sentence $\boldsymbol{x}$ into sentence $\boldsymbol{y}$.

For the negative example construction component, we use the proposed negative example construction strategy to construct K negative examples $\tilde{\boldsymbol{Y}} = \{\tilde{\boldsymbol{y}}_{\boldsymbol{1}}, \tilde{\boldsymbol{y}}_{\boldsymbol{2}}, \cdots, \tilde{\boldsymbol{y}}_{\boldsymbol{K}}\}$ for the training pair $(\boldsymbol{x}, \boldsymbol{y})$. Each negative example $\tilde{\boldsymbol{y}}_{\boldsymbol{k}}$ is constructed as follows. First, several words in the target sentence $\boldsymbol{y}$ are randomly selected by a noising probability p. For each selected word y_j, a noised word y'_j is generated by the negative example construction strategy. The generated noised word y'_j is used to replace the selected word y_j. After replacing all selected words, the modified target sentence is treated as a constructed negative example $\tilde{\boldsymbol{y}}_{\boldsymbol{k}}$.

For the contrastive learning component, we first input the source sentence $\boldsymbol{x}$ into the GEC model and

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

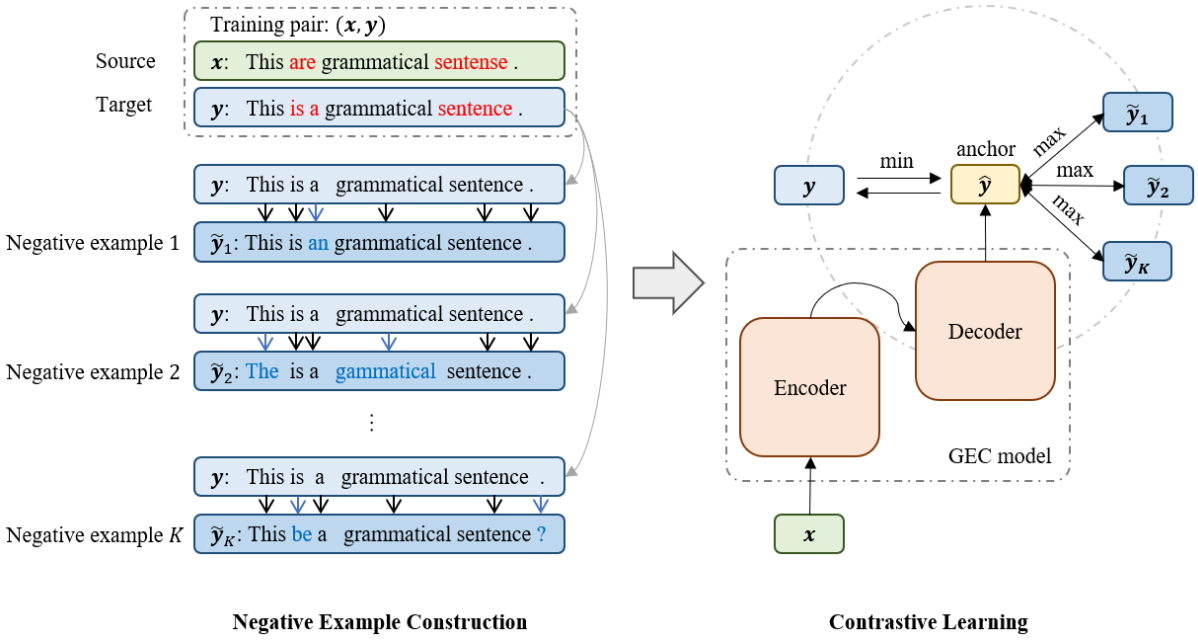


Figure 1: Overall architecture of our proposed method. Our method consists of two components: the negative example construction component and the contrastive learning component. For the negative example construction component, we use the proposed negative example construction strategy to construct K negative examples $\tilde{\boldsymbol{Y}} = \{\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \cdots, \tilde{\boldsymbol{y}}_K\}$ for the training pair $(\boldsymbol{x}, \boldsymbol{y})$. For the contrastive learning component, we first input the source sentence $\boldsymbol{x}$ into the GEC model and obtain the decoder output $\hat{\boldsymbol{y}}$. Then, we treat the decoder output $\hat{\boldsymbol{y}}$ as an anchor, and maximize the distance between the anchor $\hat{\boldsymbol{y}}$ and the negative examples $\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \cdots, \tilde{\boldsymbol{y}}_K$ constructed by our proposed negative example construction strategy.

obtain the decoder output $\hat{\boldsymbol{y}}$. Then, we treat the decoder output $\hat{\boldsymbol{y}}$ as an anchor, and maximize the distance between the anchor $\hat{\boldsymbol{y}}$ and the negative examples $\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \cdots, \tilde{\boldsymbol{y}}_K$ constructed by our proposed negative example construction strategy.

3.2 Negative Example Construction Strategy

In this section, we detail the proposed negative example construction strategy. It contains three schemes: realistic scheme, random scheme and linguistic scheme. In most cases, we use the realistic scheme for constructing negative examples. When the realistic scheme is not applicable, we use the random scheme or the linguistic scheme instead. We demonstrate examples using our proposed strategy in Table 1.

The realistic scheme makes use of the realistic grammatical error patterns produced by human beings for constructing negative examples. Realistic grammatical error patterns are effective for introducing realistic grammatical errors into error-free text and have been used in previous GEC studies for data augmentation (Li and He, 2021) and error-aware BERT fine-tuning (He et al., 2021). In this study, we utilize them for generating sufficient negative examples with more diverse grammatical errors.

Specifically, we first extract all realistic grammatical error patterns {WRONG:CORRECT} from the training data, where WRONG indicates an erroneous word in a sentence and CORRECT indicates its correction. Then, we reverse their key-value pairs into the form of {CORRECT:WRONG} for negative example construction. When constructing a negative example, for each word y_j selected from the target sentence $\boldsymbol{y}$, we randomly choose one of the {CORRECT:WRONG} patterns whose key CORRECT is y_j, and use the value WRONG as the noised word of y_j for replacement. The intuition is that we replace a correct word with a wrong word.

In practice, however, a word y_j randomly selected from the target sentence may not be one of the keys in the available {CORRECT:WRONG} pairs, such as an out-of-vocabulary word. To handle this case, we

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Target sentence y	We are exploring negative example construction strategies.
Realistic Scheme	We is exploring negative example construction strategy.
Random Scheme	We are exploring Title example fill strategies.
Linguistic Scheme	
-synonym	We are exploring passive example building strategies.
-inflection	We are explored negative example constructed strategies.
-function word	They are exploring negative example construction strategies.
-case	we are exploring Negative example construction strategies.
-misspelling	We air exploding negative example construction strategies.

Table 1: Demonstration of our proposed negative example construction strategy, which contains three schemes. Note that each linguistic transformation in the linguistic scheme is demonstrated separately for clarity. In practice, one of the linguistic transformations will be randomly selected.

propose two additional schemes as compromise:

1) Random Scheme. Such a particular word is replaced by another word sampled from the vocabulary of the dataset in a uniform distribution.

2) Linguistic Scheme. Such a particular word is replaced by one of the five linguistic transformations described below. These linguistic transformations are used by some GEC studies to mimic realistic grammatical errors (Takahashi et al., 2020; Li and He, 2021; He et al., 2021).

∘ *Synonym Transformation.* Replacing y_j with one of its synonyms. It is helpful for generating word misuse errors (noun errors, verb errors, adjective errors, etc.) commonly appeared in writings, such as misusing *"situation"* as *"condition"*.

∘ *Inflection Transformation.* Replacing y_j with one of its inflections. It imitates inflection misuse in the writings, such as misusing noun declension and verb conjugation. E.g., using the present tense of *"is"* where the past tense *"was"* is required.

∘ *Function Word Transformation.* Replacing y_j with another function word that belongs to the same function word category of y_j. It imitates the improper function word uses in writings, such as misusing *"at"* as *"in"* and misusing *"to"* as *"towards"*.

∘ *Case Transformation.* Replacing y_j with one of the three case patterns: lowercase, uppercase, and capitalize. It mimics the case errors made frequently by native English speakers due to their carelessness, such as lower-casing country names, city names and abbreviations.

∘ *Misspelling Transformation.* Replacing y_j with one of its 10 most similar words. It mimics the misspelling errors commonly appeared in writings by the native English speakers due to carelessness or rapid typing with keyboards.

3.3 Dynamic Construction

Many contrastive learning studies (Chen et al., 2020b; Chen et al., 2020c; Gao et al., 2021) have proved that the variety of negative examples is beneficial for improving the performance of the trained model. In our method, the proposed negative example construction strategy is based on rules, and the operations required for constructing negative examples are merely random sampling and replacement. Therefore, they are lightweight and consume little time, enabling them to be dynamically applied during the training process.

As shown in Figure 2, we depict the proposed dynamic negative example construction, and compare it with the static negative example construction. In the figure, $(\boldsymbol{x}, \boldsymbol{y})$ denotes a training pair, where $\boldsymbol{x}$ denotes the source sentence and $\boldsymbol{y}$ denotes the target sentence. $\tilde{\boldsymbol{Y}}$ denotes the constructed negative examples. f_{static} denotes the static negative example construction strategy and f_{dynamic} denotes our proposed dynamic negative example construction strategy.

With static construction (the higher part of the figure), the negative examples $\tilde{\boldsymbol{Y}}$ (blue) for the training pair $(\boldsymbol{x}, \boldsymbol{y})$ are constructed before the training process. And during training, the same set of negative examples constructed ($\tilde{\boldsymbol{Y}}$) is used in each iteration. On the contrary, with dynamic construction (the

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

lower part of the figure), different sets of negative examples are constructed dynamically for the training pair $(\boldsymbol{x}, \boldsymbol{y})$ in each iteration during training. Specifically, in iteration 1, a set of negative examples $\tilde{\boldsymbol{Y}}_{\mathbf{1}}$ (orange) are constructed with the negative example construction strategy f_{dynamic}. Similarly, another set of negative examples $\tilde{\boldsymbol{Y}}_{\mathbf{2}}$ (yellow) are constructed in iteration 2. In this manner, for the same training pair, dynamic construction enables the model to see different sets of negative examples in different iterations during training, and significantly increases the variety of the negative examples.

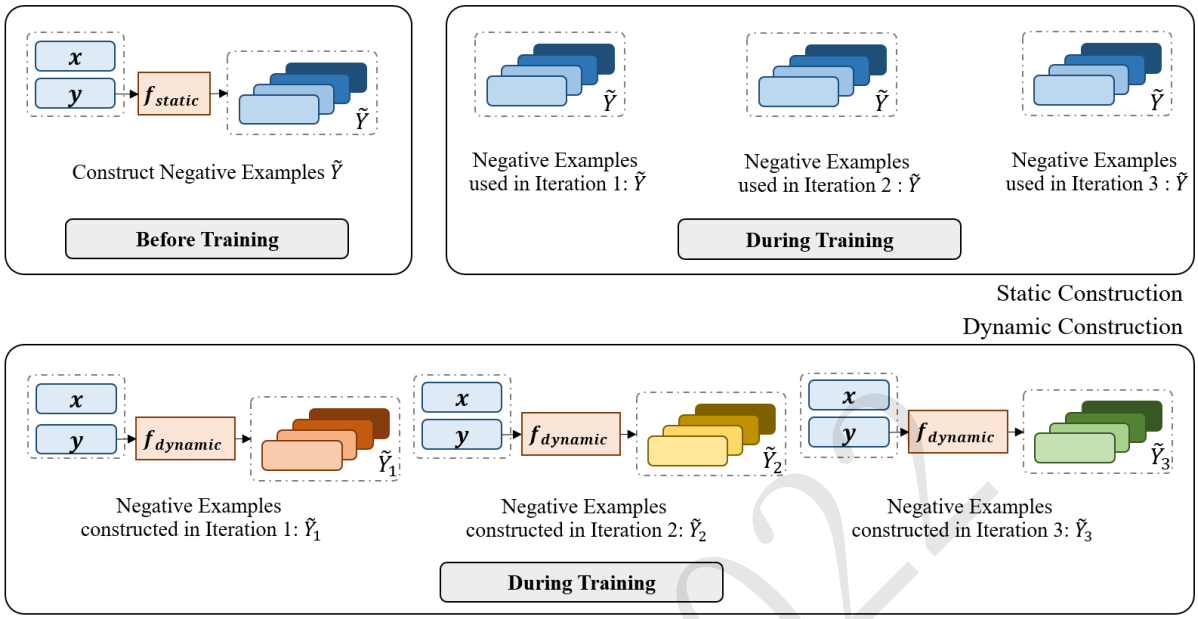


Figure 2: Demonstration of the proposed dynamic negative example construction and its comparison with the static negative example construction.

3.4 Model Training

Following Cao et al. (2021), we use the weighted sum of the negative log likelihood loss L^{NLL} and a margin-based contrastive loss L^{CL} as the training loss L for each training pair $(\boldsymbol{x}, \boldsymbol{y})$ to optimize model parameters, as in Equation 1. α is a weighting parameter that controls the relative importance of the two losses. During training, the negative log likelihood loss L^{NLL} (Equation 2) increases the similarity between the model output $\hat{\boldsymbol{y}}$ and the target sentence $\boldsymbol{y}$. And the contrastive loss L^{CL} (Equation 3) discourages the model from generating each negative example $\tilde{\boldsymbol{y}}_{\boldsymbol{k}}$ that contains grammatical errors. K is the number of constructed negative examples, and γ is the margin.

$$L = \alpha \cdot L^{\text{NLL}} + (1 - \alpha) \cdot L^{\text{CL}} \tag{1}$$

$$L^{\text{NLL}} = -\boldsymbol{y} \log \hat{\boldsymbol{y}} \tag{2}$$

$$L^{\text{CL}} = \frac{1}{K} \sum_{k=1}^{K} \max(-\boldsymbol{y} \log \hat{\boldsymbol{y}} + \tilde{\boldsymbol{y}}_{\boldsymbol{k}} \log \hat{\boldsymbol{y}} + \gamma, 0) \tag{3}$$

4 Experiments

4.1 Datasets

We use the CWEB dataset (Flachs et al., 2020) for experiments. It contains low error density writings from native English speakers and includes two domains. The G domain (CWEB-G) contains writings with a higher number of grammatical errors, and the S domain (CWEB-S) contains more professional writings with fewer grammatical errors.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945–957, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

The CWEB dataset only contains development data and test data but no training data. Following previous studies (Cao et al., 2021; Flachs et al., 2020), we extract the first 1,000 samples of CWEB-G and the first 1,000 of CWEB-S from the original development data and combine them to form the training data, which are used for training models and extracting realistic grammatical error patterns. The remaining of the original development data are taken as new development data for obtaining the best model during training. The original test data of CWEB are left unchanged, with which we evaluate the trained GEC models. We use ERRANT (Bryant et al., 2017) to calculate precision, recall and $F_{0.5}$ for evaluating the correction performance of the GEC models. Statistics of the dataset are shown in Table 2.

The grammatical errors and their corresponding corrections are annotated by two annotators. When training the model, we only use the corrections from annotator 1 as target sentences. When evaluating the trained model on test data, we calculate the scoring performance against each annotator and take the average for report.

Splits	Original		Derived	
Train	-	-	2,867	1,862
Dev	3,867	2,862	1,000	1,000
Test	3,981	2,864	3,981	2,864

Table 2: Statistics of the CWEB dataset. **Original** is the statistics of the original dataset and **Derived** is the statistics after splitting the development set into training and development data.

4.2 Experiment Settings

We use Transformer-big (Vaswani et al., 2017) as the model architecture. Following Cao et al. (2021), we use the pre-trained weights of GEC-PD (Kiyono et al., 2019) to initialize the GEC model. We use the Adam (Kingma and Ba, 2014) optimizer with the learning rate set to 3e-5. We train the model for 10 epochs and validate it after each epoch on the development set. Model weights of the smallest validation loss is used as the best model for evaluation on the test set. We construct $K = 4$ negative examples for each training pair and set the noising probability p to 0.15. We run 3 times with different seeds for each experiment and take the average of the 3 runs for report to reduce randomness.

4.3 Compared Models

We compare our method with several strong baselines to prove the effectiveness of the proposed method:

Direct Inference. Making predictions on CWEB test data directly with an off-the-shelf GEC model developed for correcting writings by learners of English as a second language, without further training on CWEB training data. In experiments, we use GEC-PD (Kiyono et al., 2019) for this purpose.

NLL. The model is first initialized with the weights of the GEC-PD model. Then, it is trained on the training data merely with negative log-likelihood (i.e., without contrastive learning) and evaluated on the test data.

$\mathbf{CL_{2021}}$. The model propsoed by Cao et al. (2021). They first initialize the model with the weights of GEC-PD. Then, they train the model on the training data with their contrastive learning method, and evaluate the trained model on the test data.

4.4 Overall Results and Analysis

The overall experimental results are shown in Table 3. **Direct Inference** are the results by the GEC-PD (Kiyono et al., 2019) model without further training on the training set. **NLL** are the results of the GEC model initialized with the weights of GEC-PD and trained merely using the negative log-likelihood loss without contrastive learning. **$\mathbf{CL_{2021}}$** are the results reported in the paper of Cao et al. (2021). **Ours (Realistic+Rand)** are the results of our proposed method with realistic scheme & random scheme, and **Ours (Realistic+Ling)** are the results of our proposed method with realistic scheme & linguistic scheme. **Average** are the average results of CWEB-G and CWEB-S. The best scores of each column are shown in bold.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	CWEB-G			CWEB-S			Average		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Direct Inference	21.18	23.01	21.45	17.27	15.76	16.92	19.22	19.38	19.18
NLL	40.46	18.93	32.76	36.78	16.66	29.51	38.62	17.79	31.14
$\mathbf{CL_{2021}}$	37.21	**23.15**	33.03	36.30	**20.40**	**31.34**	36.76	**21.78**	32.19
Ours (Realistic+Rand)	41.37	19.80	33.80	38.06	17.08	30.48	39.71	18.44	32.14
Ours (Realistic+Ling)	**42.42**	19.06	**33.89**	**39.10**	16.80	30.82	**40.76**	17.93	**32.36**

Table 3: Overall experiment results. **Direct Inference** are the results by the GEC-PD (Kiyono et al., 2019) model without further training on the training set. **NLL** are the results of the GEC model initialized with the weights of GEC-PD and trained merely using the negative log-likelihood loss without contrastive learning. $\mathbf{CL_{2021}}$ are the results reported in the paper of Cao et al. (2021). **Ours (Realistic+Rand)** are the results of our proposed method with realistic scheme & random scheme, and **Ours (Realistic+Ling)** are the results of our proposed method with realistic scheme & linguistic scheme. **Average** are the average results of CWEB-G and CWEB-S. The best scores of each column are shown in bold.

First, it is shown that the results of **Direct Inference** with GEC-PD are low. It's average $F_{0.5}$ is 19.18. And its average precision is only 19.22, which is lower than other results by a large margin. That supports the finding that the GEC model developed for correcting writings by learners of English as a second language indeed produces low performance on the writings by native English speakers due to the low error density (Flachs et al., 2020).

Second, we find that after training GEC-PD on the CWEB training data with our proposed method, the results are improved. Specifically, the average $F_{0.5}$ of our proposed **Realistic+Ling** (32.36) is higher than **NLL** (31.14) by 1.22, and higher than $\mathbf{CL_{2021}}$ (32.19) by 0.17.

Third, we can also see that our method significantly boosts the precision of the GEC model. For example, the precision of **Realistic+Ling** in CWEB-G and CWEB-S are 42.42 and 39.10, which are 1.96 and 2.32 higher than **NLL**, 5.21 and 2.80 higher than $\mathbf{CL_{2021}}$. At the same time, it also produces the highest average precision (40.76). The higher precision of our GEC model illustrates that the grammatical errors detected by the model indeed are erroneous, rather than accurate. In the task of grammatical error correction, a GEC model with an ability to **accurately** correct the detected grammatical errors (higher precision) is more preferred than one with an ability to detect many grammatical errors but fail to correct them (higher recall). This is also reflected by the evaluation metric $F_{0.5}$ of GEC, which values the precision twice as the recall. Therefore, our proposed method is beneficial for enhancing the correction performance of the GEC model, as it indeed makes the model correct the detected grammatical errors precisely and suppresses the model from over-correction.

Finally, the results also show that **Realistic+Ling** produces higher average precision (40.76) and average $F_{0.5}$ (32.36) than **Realistic+Rand** (39.71 and 32.14). It proves that the pseudo grammatical errors generated by the linguistic transformations are beneficial and effective for the construction of negative examples, which leads to a better GEC model.

5 Discussion and Analysis

5.1 Case Study

In order to demonstrate that our proposed negative example construction strategy can indeed generate sufficient negative examples with realistic and diverse grammatical errors, we extracted one training pair from the CWEB dataset accompanied by their corresponding negative examples constructed by Cao et al. (2021)'s method and those constructed with our proposed method, as shown in Table 4. In the training pair, there is a case error (*"allow"* → *"Allow"*), which is coloured red. In the negative examples, noises introduced by the negative example construction methods are coloured blue.

We can see that the first, third and fourth negative examples constructed by Cao et al. (2021)'s method are the same as the source sentence. The second example contains an insertion error (*"pick"* → *"pick*

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945–957, Nanchang, China, October 14 – 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Source sentence	allow them to pick some coloring sheets that you can print for them.
Target sentence	Allow them to pick some coloring sheets that you can print for them.
Negative examples by Cao et al. (2021)	allow them to pick some coloring sheets that you can print for them. Allow them to pick up some coloring sheets that you can print for them. allow them to pick some coloring sheets that you can print for them. allow them to pick some coloring sheets that you can print for them.
Negative examples by our method	Allow them to pick some coloring sheets that you can print with them. Allow them to pick some coloring piece that you can prunt for them. Allow them to pick sum coloring sheets that you can print for them. allow them to pick some coloring sheets that you can print in them.

Table 4: Case study. We extracted one training pair from the CWEB dataset accompanied by their corresponding negative examples constructed by Cao et al. (2021)'s method and those constructed with our proposed method. Grammatical errors in the training pair are coloured red. Noises in the negative examples introduced by the negative example construction methods are coloured blue. The negative examples constructed with out proposed method are more sufficient with more diverse grammatical errors.

Method	**Static**			**Dynamic**		
	P	**R**	**$F_{0.5}$**	**P**	**R**	**$F_{0.5}$**
Realistic+Rand	39.90	17.74	31.78	39.71	18.44	**32.14**
Realistic+Ling	38.75	18.73	31.77	40.76	17.93	**32.36**

Table 5: Scoring performance comparison of the proposed Realistic+Rand and Realistic+Ling methods with static and dynamic construction respectively.

up"). Obviously, these negative examples do not contain diverse and realistic grammatical errors, which is not helpful for the model to learn to correct properly from contrastive learning. On the other hand, the negative examples constructed using our proposed method contain a large number of diverse and realistic grammatical errors. For instance, the first example contains a preposition error (*"for"* → *"with"*). The second example contains a synonym error (*"sheets"* → *"piece"*) and a misspelling error (*"print"* → *"prunt"*). From the negative examples with diverse and realistic errors, the GEC model can better learn to correct sentences precisely through contrastive learning.

5.2 Effect of Dynamic Construction

As mentioned above, our proposed dynamic negative example construction can increase the variety of the negative examples during model training. In this section, we investigate the effect of dynamic construction by comparing the scoring performance of the proposed Realistic+Rand and Realistic+Ling methods with static and dynamic construction respectively.

The experimental results are shown in Table 5. The left half shows the results of static construction, while the right half shows the results of dynamic construction. The results of each negative example construction method are the average of CWEB-G and CWEB-S. The higher results between the static and dynamic construction of each method are bolded.

As shown in the table, the dynamic results are generally higher than the static results. Specifically, the $F_{0.5}$ of static Realistic+Rand is 31.78, while that of the dynamic one is 32.14, with a performance gap of 0.36. The $F_{0.5}$ of static Realistic+Ling is 31.77, while the dynamic one is 32.36, with a large performance gap of 0.59. It proves that by increasing the variety of the negative examples during training, dynamic construction indeed increases the variety of the negative examples, thereby avoiding overfitting and enhancing the generalization ability of the GEC model.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

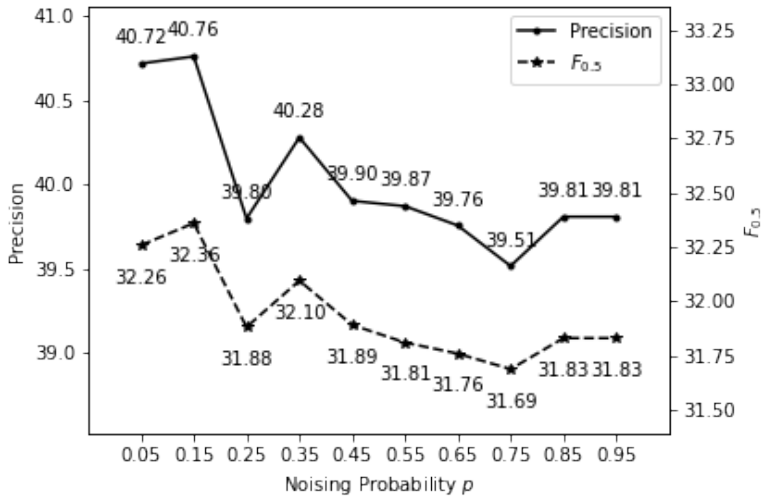


Figure 3: Scoring performance of the GEC model at different values of p, from 0.05 to 0.95 at a 0.1 interval. The dynamic Realistic+Ling strategy is used for constructing negative examples in the experiment. The experiment results are obtained from averaging the results of CWEB-G and CWEB-S. When p is set to 0.15, the scores reaches the highest (precision=40.76, $\mathrm{F}_{0.5}$=32.36). As p gradually increases, the precision and $\mathrm{F}_{0.5}$ drop gradually.

5.3 Effect of the Noising Probability

When constructing a negative example with our proposed negative example construction strategy, a noising probability p should be determined to randomly select words from the target sentence for replacement. In this section, we analyze the impact of different values of p on the correction performance. Specifically, we construct negative examples with the proposed dynamic Realistic+Ling strategy according to different values of p, from 0.05 to 0.95 at a 0.1 interval. The precision and $F_{0.5}$ at each probability are shown in Figure 3, which are obtained from averaging the results of CWEB-G and CWEB-S.

The results show that when p is set to 0.15, the scores reaches the highest (precision=40.76, $\mathrm{F}_{0.5}$=32.36). As p gradually increases, the precision and $\mathrm{F}_{0.5}$ drop gradually. The reason may be that as p increases, more words are selected from the target sentence for replacement. Therefore, the negative examples constructed are more different from the target sentence. The greater the difference between the target sentence and the negative example, the easier it is for the GEC model to compare their differences, and the smaller the improvement in the error correction ability of the model obtained from contrastive learning.

6 Conclusion

In this paper, a dynamic negative example construction method for grammatical error correction using contrastive learning is proposed. The proposed method constructs sufficient negative examples with diverse grammatical errors dynamically during model training. The constructed negative examples are beneficial for the GEC model to correct sentences precisely and suppress the model from over-correction. Experimental results show that our proposed method enhances the correction precision significantly. In this study, positive example construction strategy is not proposed for grammatical error correction using contrastive learning, as it is hard to construct sentences that are morphologically different from but semantically identical to the target sentence. One possible solution for that may be utilizing data augmentation. In future work, we will investigate this topic in depth.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61976062).

References

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. 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 4260–4270, Hong Kong, China, November. Association for Computational Linguistics.

Chris Brockett, William B. Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia, July. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, July. Association for Computational Linguistics.

Hannan Cao, Wenmian Yang, and Hwee Tou Ng. 2021. Grammatical error correction with contrastive learning in low error density domains. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4867–4874, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Mengyun Chen, Tao Ge, Xingxing Zhang, Furu Wei, and Ming Zhou. 2020a. Improving the efficiency of grammatical error correction with erroneous span detection and correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7162–7169, Online, November. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020c. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, 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.

Mariano Felice, Zheng Yuan, Øistein E. Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24, Baltimore, Maryland, June. Association for Computational Linguistics.

Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. Grammatical error correction in low error density domains: A new benchmark and analyses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8467–8478, Online, November. 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.

Junyi He, Xia Li, Han Su, Xinyin Chen, Hao Yang, and Minghao Chen. 2021. Ea-mlm: Error-aware masked language modeling for grammatical error correction. In *2021 International Conference on Asian Language Processing (IALP)*, pages 363–368.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online, July. Association for Computational Linguistics.

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

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. 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 1236–1242, Hong Kong, China, November. Association for Computational Linguistics.

Xia Li and Junyi He. 2021. Data augmentation of incorporating real error patterns and linguistic knowledge for grammatical error correction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 223–233, Online, November. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. 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 1065–1072, Online, August. Association for Computational Linguistics.

Nina H Macdonald. 1983. Human factors and behavioral science: The unix™ writer's workbench software: Rationale and design. *Bell System Technical Journal*, 62(6):1891–1908.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Stephen D. Richardson and Lisa C. Braden-Harder. 1988. The experience of developing a large-scale natural language text processing system: Critique. In *Second Conference on Applied Natural Language Processing*, pages 195–202, Austin, Texas, USA, February. Association for Computational Linguistics.

Grigori Sidorov, Anubhav Gupta, Martin Tozer, Dolors Catala, Angels Catena, and Sandrine Fuentes. 2013. Rule-based system for automatic grammar correction using syntactic n-grams for English language learning (L2). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 96–101, Sofia, Bulgaria, August. Association for Computational Linguistics.

Grigori Sidorov. 2013. Syntactic dependency based n-grams in rule based automatic english as second language grammar correction. *International Journal of Computational Linguistics and Applications*, 4(2):169–188.

Felix Stahlberg and Shankar Kumar. 2020. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online, November. Association for Computational Linguistics.

Xin Sun and Houfeng Wang. 2022. Adjusting the precision-recall trade-off with align-and-predict decoding for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Dublin, Ireland, May. Association for Computational Linguistics.

Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. 2020. Grammatical error correction using pseudo learner corpus considering learner's error tendency. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–32, Online, July. 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.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, San Diego, California, June. Association for Computational Linguistics.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. 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 156–165, Minneapolis, Minnesota, June. Association for Computational Linguistics.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 945-957, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

SPACL: Shared-Private Architecture based on Contrastive Learning for Multi-domain Text Classification

Guoding Xiong[1], **Yongmei Zhou**[1,2*], **Deheng Wang**[1], **Zhouhao Ouyang**[3]
[1]School of Cyber Security, Guangdong University of Foreign Studies, Guangzhou 510006, China
[2]Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou 510006, China
[3] School of Computing, University of Leeds, Wood-house Lane, Leeds, West Yorkshire, LS2 9JT, United Kingdom
1184467802@qq.com,yongmeizhou@gdufs.edu.cn
1148684516@qq.com,tal-darim@foxmail.com

Abstract

With the development of deep learning in recent years, text classification research has achieved remarkable results. However, text classification task often requires a large amount of annotated data, and data in different fields often force the model to learn different knowledge. It is often difficult for models to distinguish data labeled in different domains. Sometimes data from different domains can even damage the classification ability of the model and reduce the overall performance of the model. To address these issues, we propose a shared-private architecture based on contrastive learning for multi-domain text classification which can improve both the accuracy and robustness of classifiers. Extensive experiments are conducted on two public datasets. The results of experiments show that the our approach achieves the state-of-the-art performance in multi-domain text classification.

1 Introduction

Text classification is one of the most basic tasks among the many tasks of Natural Langugae Processing(NLP). In recent years, the research work of text classification has produced a large number of applications and achieved remarkable results. With the continuous release of a large number of pretrained language models in recent years, such as BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019) and other pretrained models, text classification problems have been able to achieve good results on the basis of neural network and pretrained models. However, most text classification problems are highly domain-dependent in that the meaning of the same word may transform in different domains. For example, the word apple expresses the fruit in kitchen review (e.g., I have shifted to an apple for lunch), while in electronics review, it means a brand of electronic products(e.g., I can't understand how apple sell so much ipod video). A common strategy, training multiple classifiers for different domains, is used to solve above problems. However, text data in reality often have characteristics of multiple domains and the cost of labeling a large number of multi-domain data is too high. Therefore, it is very important and practical meaningful to improve the accuracy of text classification in multiple related domains. Multi-domain text classification(MDTC) (Li and Zong, 2008) is proposed to solve above problems, it aims to utilize textual information in different domains to improve the performance of model architecture, but there is no need to train a separate classifier for each domain. In recent years, deep learning has been widely used in MDTC problems, and has achieved excellent results (Wu and Huang, 2015; Wu and Guo, 2020). The method used in most studies is shared-private architecture. Private modules are used to capture domain-specific knowledge for each domain, and shared modules are used to capture domain-invariant knowledge (Liu et al., 2016).However, these researches only pay attention to how to obtain the shared knowledge of multiple domains and domain-specific knowledge better, but ignore the representation of the samples in the representation space. In order to solve the problems above, in this paper, we propose Shared-Private Architecture based on Contrastive Learning(SPACL), which uses contrastive learning to improve the representations of different types of samples in the representation space, thereby improving the performance level of downstream tasks. Different from previous studies,

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

our architecture can not only use conditional adversarial training to extract domain-invariant features, but also generate better sample representations for MDTC.

The contributions of this paper are summarized as follows: 1)In order to strengthen the alignment representations of data in different domains, we propose a shared-private architecture based on contrastive learning for multi-domain text classification which can improve both the accuracy and robustness of the text classfier. 2)We adopt a conditional adversarial network to interact domain-shared features and classification labels, which can be better adapted to multi-domain text classification. 3)Experiments are carried out on two public multi-domain datasets, and the experimental results compared with multiple baselines show that our proposed model architeture has achieved state-of-the-art results.

2 Related Work

2.1 Multi-domain text classification

Multi-domain text classification was proposed first to improve performance through fusing training data from multiple domains (Li and Zong, 2008). The biggest challenge of this task is that the same text may has different implications in different domains, and the cost of labeling each domain is too costly.

Some early studies mainly used domain transfer learning techniques for MDTC. The structural correspondence learning (SCL) algorithm was proposed to select source domains most likely to adapt well to given target domains (Blitzer et al., 2007). Pan et al. (2010) proposed a spectral feature alignment (SFA) method to align domain-specific words from different domains into unified clusters, with the help of domain-independent words as a bridge. Wu and Huang (2015) proposed a novel approach based on multi-task learning to train sentiment classifiers for different domains in a collaborative way. Liu et al. (2015) proposed a multi-task deep neural network(MTDNN) for learning representations across multiple tasks, not only leveraging large amounts of cross-task data, but also benefiting from a regularization effect that leads to more general representations to help tasks in new domains. Liu et al. (2017) proposed an adversarial multi-task learning framework, alleviating the shared and private latent feature spaces from interfering with each other.

The most recent prior works on MDTC include Meta Fine-Tuning(MFT) for multi-domain text classification (Wang et al., 2020). Dual Adversarial Co-Learning(DACL) for Multi-Domain Text Classification (Wu and Guo, 2020), Conditional Adversarial Networks(CAN) for Multi-Domain Text Classification (Wu et al., 2021a) and Mixup Regularized Adversarial Networks(MRAN) for Multi-Domain Text Classification (Wu et al., 2021b). MFT uses meta-learning and domain transfer technology to learn highly transferable knowledge from typical samples in various domains. Both DACL and CAN leverage adversarial training to obtain the shared domain features. MRAN adopts the domain and category mixup regularizations to enrich the intrinsic features in the shared latent space and enforce consistent predictions in-between training instances. However, these methods ignore the distance of samples in the feature space when learning multi-domain feature representations, which is an important guideline to help classification. Furthermore, they did not consider the interaction between the extracted features and class labels, which is often important to improve their correlation. Different from the above studies, the work our proposed further advances the line of study by deploying contrastive learning. It can also model the interactions between shared domain features and classes to enhance their representations through a conditional adversarial network. We assume that data in various domains is insufficient, and make full use of data from multiple domains to improve overall system performance.

2.2 Contrastive Learning

Recently, related researches show that contrastive learning is an effective self-supervised learning method. Chen et al. (2020) proposed simple framework for contrastive learning of visual representations(SimCLR) to improve the quality of the learned representations by contrastive learning. Meng et al. (2021) present a self-supervised learning framework, COCO-LM, that pretrains Language Models by COrrecting and COntrasting corrupted text sequences. Giorgi et al. (2020) present Deep Contrastive Learning for Unsupervised Textual Representations(DeCLUTR) to enclose the performance gap between unsupervised and supervised pretraining for universal sentence encoders. One of the key aspects of con-

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

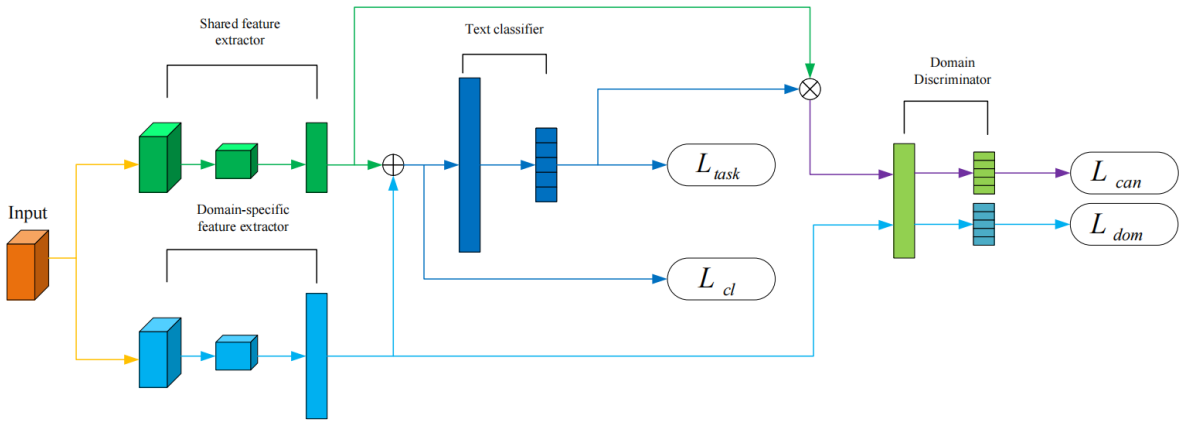


Figure 1: The overall model atchitecture of SPACL. A shared feature extractor is used to capture the shared knowledge. Each domain-specific extractor is trained to extract the domain-specific knowledge. A domain classifier is trained to predict the domain label of the input sample. A text classifier is trained to predict the class of samples and calculate the loss of contrastive learning. L_{task} is the loss function of text classification. L_{cl} is the loss fuction of contrastive learning. L_{dom} is the loss function of domain classfication. L_{can} is the conditoning adverserial loss function which extracts the shared knowledge across domains.

trastive learning is the sampling of positive pairs. Gao et al. (2021) add dropout noise to keep a good alignment for positive pairs. Fang et al. (2020) uses data augmentation to generate positive pairs from the original sentences.

We develop our model architecture with contrastive learning. In our experiments, we select a sample and combine it with itself to get a positive pair. And then combine it with other different kinds of samples to get negative pairs. A contrastive loss is used to control the distance between samples of different classes in the sample space so that enhance the ability of the text classifier.

3 Methodology

3.1 Model Architecture

In this paper, we consider MDTC tasks in the following settings.

Specifically, there exists M domains $\{D_i\}_{i=1}^{M}$. The labeled training collection of the m-th domain is denoted by $\mathrm{X}_1^{\mathrm{m}} = \left\{\left(x_j^m, y_j^m\right) \mid \mathrm{j} \in [1, N_1^m]\right\}$, where x_j^m and y_j^m are the input texts and the label of the j-th sample of the m-th domain. $N_l^m$is the total number of the labeled samples of the m-th domain. The unlabeled training collection of the m-th domain is denoted by $\mathrm{X}_u^{\mathrm{m}} = \{(x_k^m) \mid \mathrm{k} \in [1, N_u^m]\}$, where x_k^m and N_u^m are the input texts of the j-th sample and the sample size of the m-th domain. N_L represents the amount of labeled data for all domains and N_U represents the amount of unlabeled data for all domains. The goal of MDTC is to improve the overall system performance by utilizing the training sets of M domains. The classification performance of the system is measured by the average classification accuracy across M domains.

3.2 Domain-specific Representation Learning

In order to ensure the validity of our extracted domain-specific features, we add a simple and effective domain discriminator D_d, which takes the extracted domain-specific features as input and outputs the predicted domain category, so as to optimize the domain discrimination ability. The h_p is the output of the domain-specific extractor for the given instance X. The domain classifier $D_d\left(h_p; \theta_d\right) \rightarrow \hat{d}$ maps the domain-specific feature representation to a domain label prediction. θ_d denotes the parameters of the domain classifier D_d. The discriminator D_d is trained to minimize the prediction loss on labeled and unlabeled instances of multiple domains:

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$$L_{\mathrm{dom}} = -\frac{1}{N_U + N_L} \sum_{\mathrm{m}=1}^{\mathrm{M}} \sum_{\mathrm{j}=1}^{\mathrm{N_l^m}+\mathrm{N_u^m}} \mathrm{d_j^m} \log \hat{\mathrm{d}}_{\mathrm{j}}^{\mathrm{m}} + \left(1 - \mathrm{d_j^m}\right) \log \left(1 - \hat{\mathrm{d}}_{\mathrm{j}}^{\mathrm{m}}\right) \tag{1}$$

where the $\hat{d}$ is prediction probaiblities of domain labels of domain discriminator D_d and the d is the true domain label of input text.

3.3 Conditional Adversarial Network

Motivated by some previous works of domain separation learning (Bousmalis et al., 2016; Shen et al., 2018), we adopt a conditional adversarial network for SPACL to extract domain shared features. After the domain-specific learning, we freeze the parameters θ_d of the domain discriminator D_d to ensure that the discriminator has good domain recognition capabilities. At the same time, in order to ensure that the features we extract can express shareability across domains, we also adopt a negative entropy loss so that the domain classifier cannot accurately identify the domain of the shared-representation the input text.

The h_s is the output of the shared extractor F_s for the given instance X. The h_s is the output of the shared extractor F_s for the given instance X. The h_p is the output of the shared extractor F_p for the given instance X. The final joint representations h is the concatenated vector of private features h_p and shared features h_s. The text classifier C outputs the probability distribution of the prediction labels which are denoted as h_c.The domain classifier $D_d\left(h_c \otimes h; \theta_{\mathrm{d}}\right) \rightarrow \hat{d}$ maps the joint feature representation h and the class prediction h_c to a domain label $\hat{d}$. The loss can be defined as:

$$L_{\mathrm{can}} = \frac{1}{N_U + N_L} \sum_{\mathrm{m}=1}^{\mathrm{M}} \sum_{\mathrm{j}=1}^{\mathrm{N_l^m}+\mathrm{N_u^m}} \mathrm{d_j^m} \log \hat{\mathrm{d}}_{\mathrm{j}}^{\mathrm{m}} + \left(1 - \mathrm{d_j^m}\right) \log \left(1 - \hat{\mathrm{d}}_{\mathrm{j}}^{\mathrm{m}}\right) \tag{2}$$

where $h_c \otimes h$ denotes the cross-covariance of the two vectors which is calculated by multilinear conditioning (Long et al., 2018).

3.4 Contrastive Learning

Intuitively, we hope that the distance between the final joint representation vectors of samples of different categories is as far as possible, so as to make the final text classifier C easier to distinguish. Therefore, we adopt a contrastive learning approach to generate better joint representation vectors. Specifically, assuming that given a batch of samples, we will sample a pair of positive examples and other sets of negative examples in the batch. The class label of every sample denotes y. Given a final joint representation h_i of a sample, from a batch we can get an positive pair (h_i, h_{pos}) and other negative sample pairs $\{(h_i, h_{neg}) \mid h_i \in y, h_{neg} \notin y\}$.

The loss of contrastive learning is defined as:

$$L_{\mathrm{cl}} = -\frac{1}{N_b} \sum_{\mathrm{i}=1}^{N_b} \log \frac{\exp\left(\mathrm{sim}\left(h_i, h_{\mathrm{pos}}\right)\right)}{\sum_{\{(h_{\mathrm{i}}, h_{neg}) | h_i \in y, h_{neg} \notin y\}} \exp\left(\mathrm{sim}\left(h_i, h_n\right)\right)} \tag{3}$$

where $\mathrm{sim}(\mathrm{u}, \mathrm{v}) = \mathrm{u}^{\mathrm{T}} v / \|u\|_2 \|v\|_2$ denotes the cosine distance between the two vectors u and v. N_b and $\|.\|_2$ denote the number of batch size and the L2 norm.

3.5 Objective Function

The multi-domain text classification task is a binary classification task.Therefore, we define the task loss is :

$$L_{\mathrm{task}} = -\frac{1}{N_L} \sum_{\mathrm{m}=1}^{\mathrm{M}} \sum_{\mathrm{j}=1}^{\mathrm{N_l^m}} \mathrm{y_j^m} \log \hat{\mathrm{y}}_{\mathrm{j}}^{\mathrm{m}} + \left(1 - \mathrm{y_j^m}\right) \log \left(1 - \hat{\mathrm{y}}_{\mathrm{j}}^{\mathrm{m}}\right) \tag{4}$$

The text classifier C takes the final joint representation as input, and outputs the prediction labels which denote $\hat{\mathrm{y}}$.

The final loss function is the combination of above losses:

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$$L = L_{\text{task}} + L_{\text{dom}} + \alpha L_{\text{can}} + \beta L_{\text{cl}} \tag{5}$$

where α and β are hyperparameters for balancing different losses.

4 Experiment

4.1 Dataset

We evaluate SPACL on two standard datasets in our experiments: the Amazon review dataset (Blitzer et al., 2007) and the FDU-MTL dataset (Liu et al., 2015). The Amazon review dataset contains reviews in four domains: books, DVDs, electronics, and kitchen. The data for each domain has 1000 positive samples and 1000 negative samples. This dataset is already preprocessed into a bag of features (unigrams and bigrams) which loses word order information. The FDU-MTL datasets contains a total of 16 domains: books, electronics, DVDs, kitchen, apparel, camera, health, music, toys, video, baby, magazine, software, sport, IMDB, and MR. Each domain of FDU-MTL dataset contains a development set of 200 samples, a test set of 400 samples, a training set of about 1400 samples, and about 2000 unlabeled samples.

4.2 Baselines

To evaluate SAPCL, we compare it with the following baselines.

The multi-task learning with bidirectional language (MT-BL) method utilizes extraction of task-invariant features by leveraging potential information among related tasks, which improves the performance of a single task (Yang and Shang, 2019). The multinomial adversarial network (MAN) learns features that are invariant across multiple domains by resorting to its ability to reduce the divergence among the feature distributions of each domain (Chen and Cardie, 2018). This method trains the domain discriminator by two loss functions: the least square loss (MAN-L2) and the negative log-likelihood loss (MAN-NLL). Dual adversarial co-learning (DACL) deploys dual adversarial regularizations to align features across different domains, aiming to improve the classifiers' generalization capacity with the learned features (Wu and Guo, 2020). Conditional adversarial networks (CANs) introduce a conditional domain discriminator to model the domain variance in both shared feature representations and class-aware information simultaneously and adopts entropy conditioning to guarantee the transferability of the shared features (Wu et al., 2021a). The collaborative multi-domain sentiment classification (CMSC) train the models by three loss functions: the least square loss (CMSC-LS), the hinge loss (CMSC-SVM), and the log loss (CMSC-Log) (Wu and Huang, 2015). The adversarial multi-task learning for text classification (ASP-MTL) alleviates the shared and private latent feature spaces from interfering with each other (Liu et al., 2017). All the comparison methods use the standard partitions of the datasets. Thus, we cite the results from (Wu and Huang, 2015; Liu et al., 2017; Chen and Cardie, 2018; Yang and Shang, 2019; Wu and Guo, 2020; Wu et al., 2021a) for fair comparisons.

Domain	CMSC-LS	CMSC-SVM	CMSC-Log	MAN-NLL	MAN-L2	DACL	CAN	SPACL(proposed)
Books	82.10	82.26	81.81	82.98	82.46	83.45	83.76	**84.65**
DVD	82.40	83.48	83.73	84.03	83.98	**85.50**	84.68	85.20
Elec.	86.12	86.76	86.67	87.06	87.22	87.40	**88.34**	88.20
Kit.	87.56	88.20	88.23	88.57	88.53	90.00	90.03	**90.10**
Avg	84.55	85.18	85.11	85.66	85.55	86.59	86.70	**87.03**

Table 1: MDTC results on the Amazon review dataset

4.3 Experimental Setting

In our experiment, we set the hyperparameters α=0.001, β=0.1. The experiment uses the Adam optimizer with the learning rate of 0.0001. The vector size of the shared feature extractor is 64 while the vector size of the domain-specific feature extractor is 128. The dropout rate is 0.5. ReLU is the activation function.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Domain	MT-BL	ASP-MTL	MAN-L2	MAN-NLL	SPACL(proposed))
books	89.0	84.00	87.6	86.8	**90.2**
electronics	**90.2**	86.80	87.4	88.8	90.0
dvd	88.0	85.50	88.1	**88.6**	88.5
Kitchen	**90.5**	86.20	89.8	89.9	90.0
apparel	87.2	87.00	87.6	87.6	**88.0**
camera	89.5	89.20	91.4	90.7	**91.2**
health	**92.5**	88.20	89.8	89.4	90.2
music	**86.0**	82.50	85.9	85.5	**86.0**
toys	**92.0**	88.0	90.0	90.4	91.1
video	88.0	84.5	89.5	**89.6**	88.7
baby	88.7	88.20	90.0	**90.2**	89.9
magazine	92.5	92.20	92.5	**92.9**	92.5
software	**91.7**	87.20	90.4	90.9	89.5
sports	**89.5**	85.7	89.0	89.0	88.2
IMDb	88.0	85.5	86.6	87.0	**88.7**
MR	75.7	**76.7**	76.1	**76.7**	76.5
AVG	88.6	86.1	88.2	88.4	**88.7**

Table 2: MDTC results on the FDU-MTL dataset

Method	Book	DVD	Electronics	Kitchen	AVG
SPACL w/o C	83.10	83.05	85.10	86.20	84.36
SPACL w/o CL	84.10	82.50	84.00	85.05	83.90
SPACL w/o D	83.05	80.01	82.05	83.17	82.07
SPACL(full)	**84.65**	**85.20**	**88.20**	**90.10**	**87.03**

Table 3: Ablation study on the Amazon review dataset

The batch size is 128. MLP feature extractors are the feature extractor of the experiment on the Amazon review dataset with an input size of 5000. MLP feature extractor is composed of two hidden layers, with size 1,000 and 500, respectively. CNN feature extractor with a single convolutional layer is the feature extractor of the experiment on the FDU-MTL review dataset. Each CNN feature extractor uses different kernel sizes (3, 4, 5) with input size of 1000. Text classifier and discriminator are MLPs with one hidden layer of the same size as their input (128 + 64 for text classifier and 128 for discriminator).

4.4 Results

We conduct the experiments on the Amazon review dataset and FDU-MTL dataset following the setting of (Chen and Cardie, 2018). A 5-fold cross-validation is conducted on the Amazon review dataset. All data is divided into five folds: three folds are used as the training set, one fold is used as the validation set, and the remaining one fold is used the test set. The experimental results on the Amazon review dataset are shown in Table 1 and the results on the FDU-MTL dataset are shown in Table 2. The best performance is shown in bold.

From Table 1, we can see that our proposed SPACL architecture is able to achieve the best average accuracy across multiple domains on the Amazon review dataset. This suggests our proposed model architecture is more effective than other baselines. From the experimental results on FDU-MTL in the Table 2, the average accuracy of our proposed SPACL is superior to the other methods. The experimental results once again demonstrate the effectiveness of our proposed method.

The reasons for the above results are as follows: 1)Our model utilizes a conditional adversarial network to correlate the extracted shared features and predicted class labels, thereby improving the overall

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

generalization performance of the model architecture. 2)Our model architecture expands the distance between samples of different classes in the sample space and the distance of samples of the same class through the method of comparative learning. Therefore, our model performs better at multi-domain text classification tasks.

4.5 Ablation Study

To validate the contribution of conditional adversarial networks and contrastive learning in our model architecture, we conduct extensive ablation experiments on the Amazon review dataset. In particular, we studied two kinds of ablation variants: (1)SPACL w/o C, the variant model architecture of our SPACL without conditional adversarial learning on shared feature extractor; (2)SPACL w/o CL, the variant model architecture of our SPACL without contrastive learning on the final joint representation; (3)SPACL w/o D, the variant model architecture of our SPACL without domain-specific representation learning; The ablation experiment results are shown in the Table 3, where we can see all variants of produce poor results, the full model architecture provides the best performance. Therefore, this validated our model architecture of the components in the presence of necessity. From the results of the ablation experiments, we can see that using contrastive learning to improve the sample representation benefits the performance of our model.

5 Conclusion

In this paper, we proposed a shared-private architecture based on contrastive learning to use across different domains of all the available resources for multi-domain text classification. The model architecture expands the distance between shared-representations of samples of different categories in the sample space by introducing contrastive learning, thereby further improving the discriminative ability of the model architecture. In addition, the model architecture uses a conditional adversarial network to establish the correlation between domain shared features and classification prediction labels which improves the overall performance of the model architecture. The experimental results on two benchmarks show that the SPACL model architecture can effectively improve the performance of the system on the multi-domain text classification task. In the future, we will explore a better solution to transfer knowledge from different domains for multi-domain text classification.

Acknowledgements

This work has been supported by the Ministry of education of Humanities and Social Science project under Grant No.19YJAZH128 and No.20YJAZH118, the Science and Technology Plan Project of Guangzhou under Grant No.202102080305.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems*, 29.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. *arXiv preprint arXiv:1802.05694*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Mengting Hu, Yike Wu, Shiwan Zhao, Honglei Guo, Renhong Cheng, and Zhong Su. 2019. Domain-invariant feature distillation for cross-domain sentiment classification. *arXiv preprint arXiv:1908.09122*.

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*.

Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In *Proceedings of ACL-08: HLT, Short Papers*, pages 257–260.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval.

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, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

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*.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. Meta fine-tuning neural language models for multi-domain text mining. *arXiv preprint arXiv:2003.13003*.

Yuan Wu and Yuhong Guo. 2020. Dual adversarial co-learning for multi-domain text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6438–6445.

Fangzhao Wu and Yongfeng Huang. 2015. Collaborative multi-domain sentiment classification. In *2015 IEEE international conference on data mining*, pages 459–468. IEEE.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021a. Conditional adversarial networks for multi-domain text classification. *arXiv preprint arXiv:2102.10176*.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021b. Mixup regularized adversarial networks for multi-domain text classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7733–7737. IEEE.

Qi Yang and Lin Shang. 2019. Multi-task learning with bidirectional language models for text classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Low-Resource Named Entity Recognition Based on Multi-hop Dependency Trigger

Jiangxu Wu*
wujx27@mail2sysu.edu.cn

Peiqi Yan*
yanpeiqiswu@163.com

Abstract

This paper introduces DepTrigger, a simple and effective model in low-resource named entity recognition (NER) based on multi-hop dependency triggers. Dependency triggers refer to salient nodes relative to an entity in the dependency graph of a context sentence. Our main observation is that triggers generally play an important role in recognizing the location and the type of entity in a sentence. Instead of exploiting the manual labeling of triggers, we use the syntactic parser to annotate triggers automatically. We train DepTrigger using an independent model architectures which are Match Network encoder and Entity Recognition Network encoder. Compared to the previous model TriggerNER, DepTrigger outperforms for long sentences, while still maintain good performance for short sentences as usual. Our framework is significantly more cost-effective in real business.

1 Introduction

Named Entity Recognition (NER) aims to detect the span from text belonging to the semantic category such as person, location, organization, etc. NER plays a core component in many NLP tasks and is widely employed in downstream applications, such as knowledge graph (Ji, 2021), question answering (Molla, 2004) and dialogue system (Peng, 2020). The deep-learning based approaches have shown remarkable success in NER, while it requires large corpora annotated with named entities. Moreover, in many practical settings, we wish to apply NER to domains with a very limited amount of labeled data since annotating data is a labor-intensive and time-consuming task. Therefore, it is an emergency to improve the performance of the deep-learning based NER model with limited labeled data.

Previous work in low-resource NER mainly focused on meta-learning (Snell, 2017), distantly supervision (Yang, 2018), transfer learning (Lin, 2017),et al. Recently, (CLin, 2020) proposed an approach based on entity trigger called *TriggerNER*. The key idea is that an entity trigger is a group of words that can help explain the recognition process of an entity in a sentence. Considering the sentence "Biden is the president of _", we are able to infer that there is a country entity on the underline according to "the president of". In this case, "the president of" is a group of triggers. Experiments reveal that the performance of utilizing 20% of the trigger-annotated sentences is comparable to that of exploiting 70% of conventional annotated sentences. However, crowd-sourced entity trigger annotations, which suffer from the same problem as traditional annotation, require labor costs and expert experience.

Inspired by attribute triggers in Attribute Extraction (Huang2, 2017), this paper presents an alternative approach to automatically annotate the trigger in a sentence by utilizing the syntactic parse. Fig. 1 is the dependency parse result of the sentence "Alice was born in Beijing", the relation "nsubj:pass" shows that the subject of "born" is "Alice". According to the meaning of "born", we are capable of inferring that "Alice" is a person entity. Inspired by this fact, we propose a novel model, namley *DepTrigger*, which explore the structures of dependency trees and utilize the syntactic parser to annotate trigger in a sentence.

* Equal contribution

©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 966-972, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

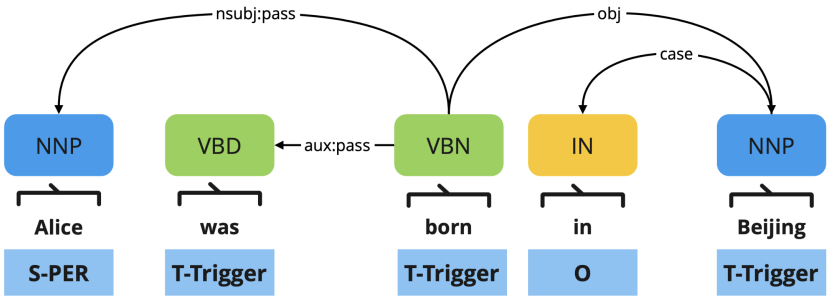


Figure 1: The dependency parse results of "Alice was born in Beijing", "S-PER" is entity label, "T-Trigger" is trigger label, "O" denotes others.

Naturally, we propose a simple yet effective framework for low-resource NER, namely *DepTriggerNER*. It includes a trigger semantic matching module (Trigger Match Network) and a sequence annotation module (Entity Recognition Network). The DepTriggerNER adopts two-steps pipeline mode: 1) we first trains the Trigger Match Network module for learning trigger representation; and 2) we combine trigger representation to train the Entity Recognition Network module. Our main contribution includes the new proposed "DepTrigger" model, which reduces the cost and complexity by using a syntactic parser to automatically annotate trigger.

We evaluate DepTrigger on CoNLL2003 (Erik, 2003) and BC5CDR (Li, 2016), where DepTrigger outperforms the TriggerNER model on BC5CDR but slightly under-performs on CoNLL2003. Compared to TriggerNER, DepTrigger is particularly useful in its ability to automatically produce annotated triggers. Besides, the independent model architectures have a better performance. Our results suggest that DepTrigger is a promising alternative to the TriggerNER in low-resource NER tasks.

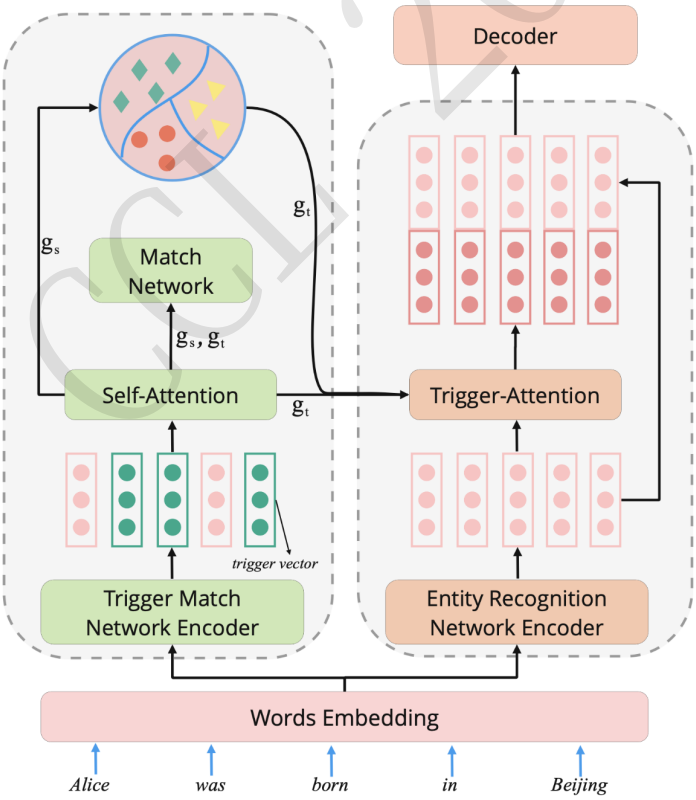


Figure 2: The framework of DepTriggerNER. The left is the Trigger Match Network. The right is the Entity Recognition Network. The circle in the upper left corner is Trigger Pattern Prototype, it is a look-up table generated by Trigger Match Network after training.

2 Model

In this section, we present the framework of DepTriggerNER in Fig. 2. Compared to TriggerNER, there are three main differences: (1) instead of crowd-sourced, we harness the syntactic parser to annotate trigger automatically; (2) we omit the trigger classification network; and (3) the Trigger Match Network encoder and the Entity Recognition Network encoder are independent of each other.

This section is organized as follows. We first describe how to use the syntactic parser to annotate dependency trigger in section 2.1. We then introduce Trigger Match Network and Entity Recognition Network in section 2.2 and section 2.3, respectively.

2.1 DepTrigger

DepTrigger are prominent nodes relative to an entity in the context sentence dependency graph. We apply Stanford CoreNLP to the sentences to obtain dependency paths. The dependency paths is a directed graph with words as nodes and dependencies as edges. Fig. 1 shows the dependency parse results of the sentence "Alice was born in Beijing". In Fig. 1 , "born" is connected with the entity "Alice" by relation "nsubj:pass", so that "born" is a DepTrigger. Words have a one-hop relationship with entities are called primary triggers, and words have a two-hop relationship with entities are called secondary triggers.

2.2 Trigger Match Network

Each entity contains a group of DepTrigger, which form a trigger pattern. We assume that each sentence has an entity and contains a trigger pattern. In the training stage, the Trigger Match Network aims to learn the representation of trigger patterns and sentences. In the inference stage, the trigger pattern representation with similar semantics to the sentence representation will be selected from the Trigger Pattern Prototype.

In Fig. 2, each sentence is first transformed into a vector by the Words Embedding module. Then, the hidden state matrix is obtained through the Trigger Match Network Encoder. The self-attention layer is used to obtain sentence representation $\vec{g}_s$ and trigger pattern representation $\vec{g}_t$, (Lin, 2017) defined as follows:

$$\vec{\alpha}_s = Softmax\left(W_2 \times tanh\left(W_1 \times H\right)\right) \tag{1}$$

$$\vec{g}_s = \vec{\alpha}_s H \tag{2}$$

$$\vec{\alpha}_t = Softmax\left(W_2 \times tanh\left(W_1 \times M\right)\right) \tag{3}$$

$$\vec{g}_t = \vec{\alpha}_t M \tag{4}$$

W_1 and W_2 are the trainable parameters. H and M represent the hidden state matrix of the sentence and the hidden state matrix of DepTrigger, respectively.

The Match Network calculates the distance between trigger pattern representation and sentence representation. The matching loss function (CLin, 2020) is defined as follows:

$$L = \begin{cases} ||\vec{g}_s - \vec{g}_t||_2^2, t \in s \\ max\left(0, m - ||\vec{g}_s - \vec{g}_t||_2^2\right), t \notin s \end{cases} \tag{5}$$

$||\cdot||_2$ is L2-norm distances, m is margin. $t \in s$ indicates trigger pattern representation and sentence representation matches well while $t \notin s$ is on the contrary. We create negative samples by randomly matching trigger pattern representation and sentence representation in a batch.

2.3 Entity Recognition Network

Entity Recognition Network is similar to most deep-learning based NER models and consists of encoder and decoder. However, the Entity Recognition Network has been added a trigger-attention layer. Note that the parameters of Trigger Match Network are frozen when training Entity Recognition Network.

In training, each sentence passes through the Trigger Match Network Encoder and the Entity Recognition Network Encoder, respectively. Then, $\vec{g}_t$ is obtained from the self-attention layer. In the trigger-attention layer, $\vec{g}_t$ is used to calculate the weight of each vector in the Entity Recognition Network

Proceedings of the 21st China National Conference on Computational Linguistics, pages 966-972, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Encoder's outputs as follows (Luong, 2015):

$$\vec{\alpha} = Softmax\left(\vec{v} \times tanh\left(U_1 \times H + U_2 \times \vec{g}_t\right)\right) \tag{6}$$

$$H' = \vec{\alpha} H \tag{7}$$

$U_1, U_2, \vec{v}$ are model parameters, and H is the Entity Recognition Network Encoder's outputs matrix. Finally, we concatenate the matrix H with the trigger-enhanced matrix H as the input $([H; H'])$ fed into the decoder.

2.4 Inference

After training, each sentence in the training set is re-input into Trigger Match Network to obtain trigger pattern representation. We then save these representations in memory, shows as the Trigger Pattern Prototype in Fig. 2. In the inference stage, We first obtain sentence representations $\vec{g}_s$ through Trigger Match Network and then retrieve the semantic similarity vector $\vec{g}_t$ from Trigger Pattern Prototype. Vector $\vec{g}_t$ is used as the attention query in Entity Recognition Network.

3 Experiments

3.1 Experiments Setup

Dataset	#Class	#Sent	#Entity
CoNLL'03	4	14986	23499
BC5CDR	2	4560	9385

Table 1: Data statistics.

CoNLL2003 (Erik, 2003) and BC5CDR (Li, 2016) are used to evaluate our model. The statistics of these datasets are shown in Table. 1. We choose BiLSTM-CRF (Ma, 2016) and TriggerNER (CLin, 2020) as baseline models. TriggerNER is the first trigger-based NER model. We choose BiLSTM as encoder and CRF as decoder in our model. To ensure a fair comparison, we use the same codebase and words embedding from GloVE (Pennington, 2014), which used in baseline model. The hyper-parameters of the model are also the same. Our code and data are released [0].

We choose BIOES tagging schema for non-triggers, and triggers are all labeled with "T-trigger". In order to make the model learn the relation between entity and its trigger better, we repeat a sentence N times, and N is the number of entities in the sentence. Each sentence retains one entity and its trigger, other entities are marked as non-entities.

CoNLL 2003					BC5CDR				
#sent	BiLSTM-CRF	#trig	Trigger-NER	Ours	#sent	BiLSTM-CRF	#trig	Trigger-NER	Ours
5%	69.04	3%	75.33	**77.42**	5%	71.87	3%	61.44	**63.37**
10%	76.83	5%	80.2	**80.26**	10%	72.71	5%	66.11	**66.92**
20%	81.3	7%	**82.02**	81.3	20%	69.92	7%	67.22	**69.27**
30%	83.23	10%	**83.53**	82.96	30%	73.71	10%	70.71	**71.42**
40%	84.18	13%	**84.22**	83.26	40%	72.71	13%	71.87	**73.17**
50%	84.27	15%	**85.03**	83.86	50%	75.84	15%	71.89	**74.35**
60%	85.24	17%	**85.36**	84.32	60%	75.84	17%	73.05	**75.08**
70%	86.08	20%	**86.01**	84.53	70%	76.12	20%	73.97	**76.44**

Table 2: F1 score results. "#sent" denotes the percentage of the sentences labeled only with entity label, "#trig" denotes the percentage of the sentences labeled with both entity label and trigger label.

[0]https://github.com/wjx-git/DepTriggerNER

Proceedings of the 21st China National Conference on Computational Linguistics, pages 966-972, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

3.2 Results

As shown in Tabel. 2, Our model achieves a similar performance as TriggerNER. More detailed, our model performs better on BC5CDR than TriggerNER, but slightly worse on CoNLL2003. We explain this phenomenon in terms of the number of triggers each entity has. Fig. 3 shows the ratio of the number of sentences with the number of triggers an entity has in each dataset. The two yellow curves are very close when the abscissa value is greater than 3, and the yellow dotted line is larger than the solid line when the abscissa value is less than 3. This fact demonstrates that on CoNLL2003 the number of triggers annotated by our method is less than TriggerNER. In the two blue curves, the solid blue line is larger than the dashed line when the abscissa value is greater than 4, and the opposite is true when the abscissa value is less than 4. This shows that the number of triggers annotated by our method is more than TriggerNER on BC5CDR. We believe that an entity is easier to recognize when it has more triggers, which would explain why our model performs better on BC5CDR and slightly worse on CoNLL2003.

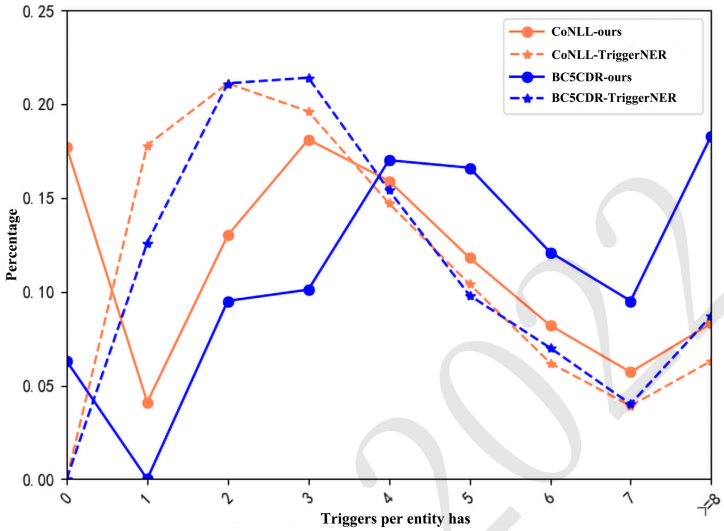


Figure 3: Ratio of the number of sentences with the number of triggers each entity has in the dataset. The X-axis is the number of triggers of a entity has, and the Y-axis is the percentage. The solid lines represent the trigger of ours. The yellow line represents CoNLL datasets.

We analyzed the sentence length distribution in the two datasets to further understand why we annotate fewer triggers in CoNLL and more in BC5CDR than in TriggerNER. The statistical results of sentence length distribution in Table 3, show that sentences are shorter in the CoNLL dataset and longer in the BC5CDR dataset. From Table 3 and Figure 3, it can be concluded that our method can label more triggers in long sentences but fewer triggers in short sentences compared to manual marking in TriggerNER. Therefore, our method is more suitable for datasets with longer sentences.

Datasets	1~10	10~25	25~50	50~
CoNLL	52.32%	27.33%	19.93%	0.42%
BC5CDR	5.7%	50.64%	37.54%	6.51%

Table 3: Statistical results of sentence length distribution

In our model, Trigger Match Network encoder and Entity Recognition Network encoder are independent, which is different from the TriggerNer. The main purpose of Trigger Match Network is to learn the representation of trigger patterns, and Entity Recognition Network is to learn entity representation. So we think we can not get an advantage by combining Trigger Match Network and Entity Recognition Network because they need to capture specific information. That is inspired by (Zexuan, 2021), and they observe that the contextual representations for the entity and relation models essentially capture specific information, so sharing their representations hurts performance.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 966-972, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

#trig	CoNLL 2003		BC5CDR	
	merge	separate	merge	separate
3%	76.36	**77.42**	61.3	**63.37**
5%	79.38	**80.26**	66.15	**66.92**
7%	80.37	**81.3**	68.02	**69.27**
10%	81.58	**82.96**	70.93	**71.42**
13%	82.55	**83.26**	72.7	**73.17**
15%	83.03	**83.86**	73.25	**74.35**
17%	83.51	**84.32**	74.95	**75.08**
20%	83.81	**84.53**	75.08	**76.44**

Table 4: Comparative experiment F1 score results. **merge** means to merge Trigger Match Network encoder and Entity Recognition Network encoder. **separate** means to separate Trigger Match Network encoder and Entity Recognition Network encoder. The best results are in **bold**.

We do a comparative experiment to test the performance of our model for merging and separating, respectively, while leaving everything else unchanged. The experimental results are shown in Table. 4, **merge** means to merge Trigger Match Network encoder and Entity Recognition Network encoder. Separate means to separate Trigger Match Network encoder and Entity Recognition Network encoder. It shows that the performance is better when the Trigger Match Network encoder and Entity Recognition Network encoder are independent.

In order to compare the influence of primary and secondary trigger words on the model, we backup two datasets of CoNLL, and only the primary triggers are labeled in one dataset, and only the secondary trigger words are labeled in the other dataset, do the same for BC5CDR. Table. 5 shows the F1 score on these datasets. Compared primary and secondary trigger, there is no evident show that one is better than the other. Combined with table 1 and table 4, the effect of using the primary trigger and the secondary trigger at the same time is significantly better than that of using them alone. .

#trig	CoNLL 2003		BC5CDR	
	primary	secondary	primary	secondary
3%	**63.4**	62.35	**52.3**	50.92
5%	**66.3**	**66.3**	54.17	**55.84**
7%	**70.37**	69.44	**58.92**	57.33
10%	**74.02**	73.44	**60.32**	60.24
13%	74.86	**74.91**	61.35	**62.01**
15%	**76.2**	75.46	64.26	**64.25**
17%	**77.36**	76.33	**64.51**	64.26
20%	**77.55**	77.53	65.94	**66.69**

Table 5: Comparative experiment of primary and secondary trigger

4 Conclusion and Future Work

We have introduced dependency trigger to incorporate trigger information into NER method. The core of our method is using syntactic parser to automatically label the trigger of entities. Our model performs well for long sentences, while maintain similar performance as TriggerNER for short sentences. Thanks to automatically annotate trigger of entities, our framework is more practical in the real business. Future work with DepTrigger includes: 1) adjusting our model to encoder based on language model; 2) making a further analysis of trigger type; 3) developing models for improving the performance on short sentences.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 966-972, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

References

Ji. Shaoxiong, Pan. Shirui, Cambria. Erik, Marttinen. Pekka and Yu. Philip. 2021. *A Survey on Knowledge Graphs: Representation, Acquisition and Applications*. IEEE Transactions on Neural Networks and Learning Systems.

Diego. Mollá, Menno. van. Zaanen, Daniel. Smith. 2004. *named entity recognition for question-answering system*. Proceedings of INTERSPEECH 2004 - ICSLP, Jeju Island, Korea.

Baolin. Peng, Chunyuan. Li, Jinchao. Li, Shahin. Shayan-deh, Lars. Liden, Jianfeng. Gao. *SOLOIST: Few-shot Task-Oriented Dialog with A Single Pre-trained Auto-regressive Model*. arxiv.

B. Yuchen. Lin, DH. Lee, M. Shen, R. Moreno, X. Huang, P. Shiralkar, X. Ren. *few-Shot Named Entity Recognition: A Comprehensive Study*. arxiv.

B. Yuchen. Lin, DH. Lee, M. Shen, R. Moreno, X. Huang, P. Shiralkar, X. Ren. *TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Jake. Snell, Kevin. Swersky, Richard. Zemel. *Prototypical networks for few-shot learning*. arxiv.

Yaosheng. Yang, Wenliang. Chen, Zhenghua. Li, Zhengqiu. He, Min Zhang. *Distantly supervised NER with partial annotation learning and reinforcement learning*. Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe. New Mexico, USA.

Lifu. Huang, Avirup. Sil, Heng. Ji, Radu. Florian. *Improving Slot Filling Performance with Attentive Neural Networks on Dependency Structures*. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark.

Dian. Yu, Heng. Ji. *Unsupervised Person Slot Filling based on Graph Mining*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany.

Feifei. Zhai, Saloni. Potdar, Bing. Xiang, Bowen. Zhou. *Neural Models for Sequence Chunking*. Proceedings of Thirty-First AAAI Conference on Artificial Intelligence.

Sang. Erik, Meulder. Fien. *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL.

Jiao. Li, Yueping. Sun, Robin. J. Johnson, Daniela. Sciaky, Chih-Hsuan. Wei, Robert. *Biocreative v cdr task corpus: a resource for chemical disease relation extraction*. Database.

Thang. Luong, Hieu. Pham, Christopher. D. Manning. *Effective approaches to attention-based neural machine translation*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal.

Jeffrey. Pennington, Richard. Socher, Christopher. Manning. *Glove: Global vectors for word representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.

Xuezhe. Ma, Eduard. Hovy. *End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany.

Zhouhan. Lin, Minwei. Feng, Cicero. Nogueira. dos Santos, Mo. Yu, Bing. Xiang, Bowen. Zhou, Yoshua. Bengio. *A structured self-attentive sentence embedding*. Proceedings of the 5th International Conference on Learning Representations.

Zexuan. Zhong, Danqi. Chen. *A Frustratingly Easy Approach for Entity and Relation Extraction*. NAACL.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 966-972, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Fundamental Analysis based Neural Network for Stock Movement Prediction

Yangjia Zheng, Xia Li*, Junteng Ma, Yuan Chen
School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China
{yjzheng,xiali,junteng.ma,yuanchen}@gdufs.edu.cn

Abstract

Stock movements are influenced not only by historical prices, but also by information outside the market such as social media and news about the stock or related stock. In practice, news or prices of a stock in one day are normally impacted by different days with different weights, and they can influence each other. In terms of this issue, in this paper, we propose a fundamental analysis based neural network for stock movement prediction. First, we propose three new technical indicators based on raw prices according to the finance theory as the basic encode of the prices of each day. Then, we introduce a coattention mechanism to capture the sufficient context information between text and prices across every day within a time window. Based on the mutual promotion and influence of text and price at different times, we obtain more sufficient stock representation. We perform extensive experiments on the real-world StockNet dataset and the experimental results demonstrate the effectiveness of our method.

1 Introduction

Stock Movement Prediction aims to predict the future price trend of a stock based on its historical price or related information. Stock price prediction can help investors, ordinary users and companies to predict the stock trend in the future, which has good application value.

The high randomness and volatility of the market make the task of Stock Movement Prediction a big challenge (Adam et al., 2016). However, with the development of neural network technology, stock movement prediction has achieved good results in recent years (Nelson et al., 2017; Hu et al., 2018; Xu and Cohen, 2018; Feng et al., 2019a; Sawhney et al., 2020; Tang et al., 2021; Zhao et al., 2022). Based on fundamental and technical analysis, existing methods can be roughly grouped into two categories, namely methods based on price factors only and methods based on price and other factors (e.g., news of the stock.). Nelson et al. (2017) used the LSTM(Hochreiter and Schmidhuber, 1997) network to predict future stock price trends based on historical price and technical analysis indicators. Feng et al. (2019a) used the adversarial training as perturbations to simulate the randomness of price variables, and trained the model to work well with small but intentional perturbations. They also extracted 11 related price features to effectively help the model to predict future changes.

According to the Efficient Market Hypothesis (EMH) (Fama, 1970), price signals themselves cannot capture the impact of market accidents and unexpected accidents, while social media texts such as tweets could have a huge impact on the stock market. Based on this idea, different models have been proposed to model relevant news texts to improve the overall performance of stock movement prediction. Hu et al. (2018) proposed to use the hierarchical attention mechanism to predict the trend of stocks based on the sequence of recent related news. Xu and Cohen (2018) integrated signals from social media which reflected the opinions of general users and used Variational Autoencoder(VAE) to capture the randomness of prices and the importance of different time steps by adding temporal attention. Sawhney et al. (2020) introduced a novel architecture for efficient mixing of chaotic temporal signals from financial data, social

*Corresponding author: xiali@gdufs.edu.cn
©2022 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

media, and inter stock relationships in a hierarchical temporal manner through Graph Attention Neural Network.

Although previous studies have achieved good results, whether it is a purely technical approach based on historical prices or a fundamental approach based on multiple factors such as prices and news, they can be improved in terms of the full integration of the two important factors of texts and prices. We found that previous works usually encode news and prices separately according to time series, and then fuse them through simple concatenation operation, similar to the work of Sawhney et al. (2020). In fact, in practice, prices on a given day can be influenced by different news at different times (e.g., previous day or after two days). Similarly, some news about a stock on a given day may be influenced by stock prices at different times. As is shown in Figure 1, if we can capture the context information of each price and text by different days, we can get more sufficient information for predicting the stock price accurately.

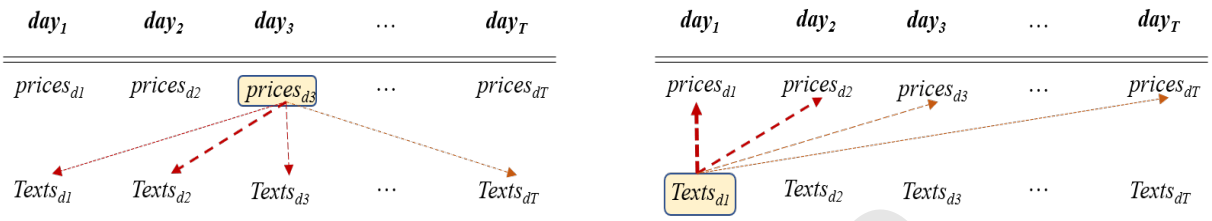


Figure 1: Contexts of prices and texts across the history captured by co-attention. Left shows each price representation of one day captures context of all news about the stock from day_1 to day_T by different attention weights. Right shows each texts representation of one day captures context of all prices about the stock from day_1 to day_T by different attention weights.

To this end, in this paper, we propose a fundamental analysis based neural network for stock movement prediction. More specifically, we first use Bi-GRU to encode the original texts of each day. Then, we use text-level attention to get a text representation of each day. As for the prices of each day, we use the existing 11 indicator features and 3 indicators we proposed in this paper as price representation of each day. Then we use the coattention mechanism (Xiong et al., 2016) to capture more information between texts and prices across every day within a time window. Finally, we incorporate a Bi-GRU to encode the fully integrated texts and prices representation according to the time window, so that it can obtain various prices and text-related information of the stock, and obtain the final effective representation of the stock.

The contributions of this work are as follows:

- We propose a fundamental analysis based neural network for stock movement prediction. The model introduces the coattention mechanism into text and price features of a stock to learn the effective context information of them. The method can obtain sufficient stock representation based on the mutual promotion and influence of texts and prices at different times.
- We also introduce three technical indicators based on raw prices in the financial field as their input features to better reflect the fluctuation information of the market. We perform multiple experiments on the StockNet dataset and the results demonstrate the effectiveness of our model.

2 Related Work

In this section, we will review the related work about stock movement prediction from technical analysis based approach and fundamental analysis based approach.

2.1 Technical Analysis based Approach

Technical analysis based approach is to predict the trend of a stock based on its historical price features such as close price and movement percent of price, which follows the assumption that future price changes are the result of historical behavior. Most recent stock movement prediction methods are based on deep learning. Among them, recurrent neural networks such as LSTM and GRU have become a

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

key part for capturing the temporal patterns of stock prices. This is beacuse they can further capture long-term dependencies in time series. Nelson et al. (2017) used LSTM networks to study future trends, predicting stock prices based on historical stock prices and technical analysis indicators. These indicators are mathematical calculations designed to determine or predict the characteristics of a stock based on its historical data. A total of 175 technical indicators are generated each period, and they are designed to represent or predict a very different set of characteristics of a stock, like the future price, volume to be traded and the strength of current movement trends. Feng et al. (2019a) proposed to use adversarial training and add perturbations to simulate the randomness of price variables, and trains the model to work well with small but intentional perturbations. In addition, they extracted 11 related price features that effectively help the model predict future changes. Feng et al. (2019b) proposed the Temporal Graph Convolution (TGC) model combining historical prices for predicting movement of stock, which dynamically adjusts the predefined firm relations before feeding them into Graph Convolution Network (GCN) (Kipf and Welling, 2017). As LSTM struggles to capture extremely long-term dependencies, such as the dependencies across several months on financial time series. Transformer-based employs multi-head self-attention mechanism to globally learn the relationships between different locations, thus enhancing the ability to learn long-term dependencies. Ding et al. (2020) proposed various enhancements for Transformer-based models, such as enhancing locality of Transformer with Multi-Scale Gaussian Prior, avoiding learning redundant heads in the multihead self-attention mechanism with Orthogonal Regularization and enabling Transformer to learn intra-day features and intra-week features independently with Trading Gap Splitter. However, in reality, it is often difficult to find clear pattern of change from the market historical data. Furthermore, it fails to reveal the rules governing market volatility beyond stock price data.

2.2 Fundamental Analysis based Approach

Efficient Market Hypothesis tells that textual information can be used to extract implicit information for helping predict the future trend of stock prices, such as financial news and social media. Fundamental analysis based approach is able to capture information that is not available in traditional price-based stock prediction. A hybrid attention network (Hu et al., 2018) is proposed to predict stock trends by imitating the human learning process. In order to follow three basic principles: sequential content dependency, diverse influence, and effective and efficient learning, the model builds news-level attention and temporal attention mechanisms to focus on key information in news, and applies self-paced learning mechanisms to automatically select suitable training samples for different training stage improves the final performance of the framework. Different from the traditional text embedding methods, Ni et al. (2021) proposed Tweet Node algorithm for describing potential connection in Twitter data through constructing the tweet node network. They take into account the internal semantic features and external structural features of twitter data, so that the generated Tweet vectors can contain more effective information. Financial news that does not explicitly mention stocks may also be relevant, such as industry news, and is a key part of real-world decision-making. To extract implicit information from the chaotic daily news pool, Tang et al. (2021) proposed News Distilling Network (NDN) which takes advantage of neural representation learning and collaborative filtering to capture the relationship between stocks and news. Xie et al. (2022) conducted adversarial attacks on the original tweets to generate some new semantically similar texts, which are merged with the original texts to confuse the previously proposed models, proving that text-only stock prediction models are also vulnerable to adversarial attacks. This also reflects that the model obtained only by text training is less robust, so it is still necessary to incorporate knowledge such as relevant historical price features and the relationship between stocks to better improve the performance of the model.

Therefore, some studies fuse price and text data to build models, and even add the relationship between stocks to improve the performance of the model. A novel deep generation model that combines tweets and price signals is proposed by (Xu and Cohen, 2018). They introduced temporal attention to model the importance of different time steps and used Variational Autoencoder(VAE) to capture randomness of price. Recent studies have attempted to simulate stock momentum spillover through Graph Neural

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Networks(GNN). Sawhney et al. (2020) introduced an architecture for efficient mixing of chaotic temporal signals from financial data, social media, and inter stock relationships in a hierarchical temporal manner. Cheng and Li (2021) proposed a momentum spillover effect model for stock prediction through attribute-driven Graph Attention Networks (GAT) (Veličković et al., 2017), and the implicit relations between stocks can be inferred to some extent. Zhao et al. (2022) constructed a market knowledge graph which contains dual-type entities and mixed relations. By introducing explicit and implicit relationships between executive entities and stocks, dual attention network is proposed to learn stock momentum overflow features.

Since stock prices have temporal characteristics, that is, the price of a day will be affected by the price and news text of previous days, in this paper, we propose to use coattention mechanism to obtain the context information of stock prices and news text under different timestamp, so as to improve the final representation of the stock and the prediction performance.

3 Our Method

3.1 Task Definition

Similar to the previous work Xu and Cohen (2018), we define the stock movement prediction task as a binary classification problem. Given a stock s, we define the price movement of the stock from day T to $T+1$ as:

$$Y_{T+1} = \begin{cases} -1, & p^c_{T+1} < p^c_T \\ 1, & p^c_{T+1} \geq p^c_T \end{cases} \quad (1)$$

where p^c_T represents adjusted closing price on day T, -1 represents stock price goes down and 1 represents the stock price goes up. The goal of the task is to predict the price movement Y_{T+1} of a stock s according to its historical prices collections P and news text collections L in a time sliding window of T days, where $P = \{P_1, P_2, ..., P_i, ..., P_T\}$, $L = \{L_1, L_2, ..., L_j, ..., L_T\}$, where P_i is the price features of the stock s on day i and L_j is the news text collection of the stock s on day j .

3.2 Overall Architecture

The whole architecture of our method is shown in Figure 2. As is shown in Figure 2, we first encode raw text for each stock across every day over a fixed time window. As for the price, the existing price features and the three new proposed indicators are concatenated together as the price representation. Then richer information will be captured by our introduced coattention mechanism. In order to obtain the integrated information of various prices and texts within the time window, we adopt a Bi-GRU for final encoding.

In the following sections, we will describe text and price features encoding in Section 3.3 and 3.4. And we will introduce temporal fusion to handle prices and text in Section 3.5 and introduce global fusion by sequential modeling in Section 3.6. Finally, model training will be introduced in Section 3.7.

3.3 Text Encoding

As each text contains rich semantic information, we use a Bi-GRU to encode the text and get the representation of each text in one day. Besides, different texts within the same day about the same stock may also be different (e.g., one text contains important information about the stock while other texts don't have valuable information about the stock.). For addressing that, we use a soft-attention operation to get the weighted representation of the texts of one day.

Following the work of Xu and Cohen (2018), we incorporate the position information of stock symbols in texts to handle the circumstance that multiple stocks are discussed in one single text. Given stock s contains K number of related texts on day m, which is denoted as $L_m = \{l_{m1}, l_{m2}, ..., l_{mi}, ...l_{mK}\}$, where l_{mi} denotes the i-th text of stock s on day m. For each text $l_{mi} = \{w_1, w_2, ..., w_n\}$, suppose that the location where the stock symbol appears first is denoted as z, we use two GRUs to encode the words sequence from w_1 to w_z to get the hidden representations $\overrightarrow{h_f}$ and words sequence from w_z to w_n to get the hidden representations $\overleftarrow{h_b}$, respectively. We use the average of the last hidden states of the two GURs $\overrightarrow{h_z}$ and $\overleftarrow{h_z}$ as the hidden representation of the text $h_{l_{mi}}$:

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

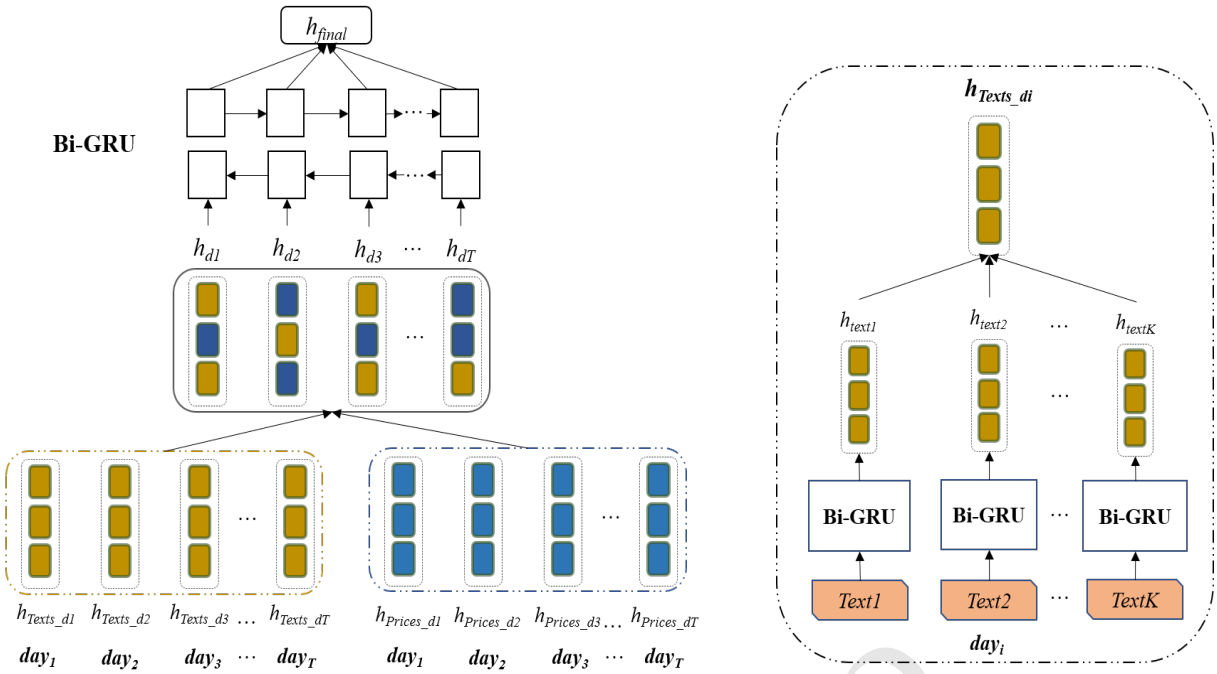


Figure 2: Overview architecture of our method.

$$\overrightarrow{h_f} = \overrightarrow{GRU}\left(e_f, \overrightarrow{h_{f-1}}\right) \tag{2}$$

$$\overleftarrow{h_b} = \overleftarrow{GRU}\left(e_b, \overleftarrow{h_{b+1}}\right) \tag{3}$$

$$h_{l_{mi}} = \left(\overrightarrow{h_z} + \overleftarrow{h_z}\right)/2 \tag{4}$$

Where e_f,e_b is the word embedding using pre-trained Global Vectors for Word Representation(GloVe) (Pennington et al., 2014) for words of the text, $f \in [1, \ldots, z], b \in [z, \ldots, n]$. After that, we can get all the text representations M_i=[$h_{l_{m1}}$,$h_{l_{m2}}$,...,$h_{l_{mK}}$]. Since the text quality is different, we use a text-level attention mechanism to identify texts that could have a more substantial impact on the market every day, and finally obtain a final representation of all texts. The calculation formula is as follows:

$$u_K = \tanh(M_i W_m + b_m) \tag{5}$$

$$\alpha_K = softmax(u_K W_u) \tag{6}$$

$$h_{Texts\_dm} = \sum_K \alpha_K h_{l_{mK}} \tag{7}$$

where α_K is the attention weight, W_m and W_u are the parameters to be learned, b_m is the bias terms. h_{Texts_dm} is the representation of the news text of stock s on m-th day(day_m). According to the time sliding window defined previously, the text data in the window is finally recorded as $H_t = [h_{Texts\_d1}, h_{Texts\_d2}, ..., h_{Texts\_dT}]$.

3.4 Price Features

As mentioned in Section 2.2, the models that predict stock trends only based on text are often fragile, while price features have been shown to effectively reflect market volatility. In this paper, we introduce three new relevant price features to be used in our method. The three new technical indicators are from financial domain and are used to describe fluctuation of stock, namely Average True Range (ATR) (Bruni, 2017), Bias Ratio (BIAS) and Momentum (MTM) (Lin et al., 2017). The detailed calculation of the three indicators is shown in Table 1. We describe the tree indicators as follows:

CCL 2022

- **ATR:** ATR is a volatility indicator that was developed by Wilder (1978) and is used to measure the volatility or the degree of price movement of security. It was originally designed for commodity trading, which is frequently subject to gaps and limit moves. As a result, ATR takes into account gaps, limit moves, and small high-low ranges in determining the true range of a commodity, and it also applies to the stock market.

- **BIAS:** BIAS is the deviation between the closing price and moving average. When the stock price moves drastically to deviate from the trend, the possibilities for a pullback or rebound increase; When the stock price movement does not deviate from the trend, it is likely that the trend will continue.

- **MTM:** MTM is an indicator that shows the difference between today's closing price and the closing price n days ago. Momentum generally refers to the continued trend of prices. Momentum shows a trend, staying positive for a sustained uptrend or negative for a sustained downtrend. An upward crossing of zero can be used as a signal of buying, and a downward crossing of zero can be used as a signal of selling. How high the indicator is (or how low when negative) indicates how strong the trend is.

Features	Calculation
ATR	$EMA(max(high_t, close_{t-1}) - min(low_t, close_{t-1}), n)$
BIAS	$\frac{close_t}{\sum_{i=0}^{4} close_{t-i}/5} - 1$
MTM	$close_t - close_{t-1}$

Table 1: The three price features.

Following previous work, We adopt 11 temporal price features based on the raw price (Feng et al., 2019a), denoted as $F_1 = \{p_1, p_2, ..., p_{11}\}$ and our proposed three new price features, denoted as $F_2 = \{p_{atr}, p_{bias}, p_{mtm}\}$, as our final price features. The two are concatenated together to get the final price features of m-th day, recorded as $h_{Prices\_dm} = [F_1, F_2]$. According to the time sliding window defined above, the price features in the window are finally recorded as $H_p = [h_{Prices\_d1}, h_{Prices\_d2}, ..., h_{Prices\_dT}]$.

3.5 Temporal Fusion by Coattention Neural Network

After Section 3.3 and Section 3.4, the coding features of price and text were obtained as H_p and H_t respectively. To effectively blend text and price, we use the coattention mechanism (Xiong et al., 2016) to learn the fusion between text and price to obtain richer implicit information. First, we use a nonlinear projection layer to convert the dimension of the price feature into the same dimension as the text with the following formula:

$$H'_p = \tanh\left(H_p W_p + b_p\right) \tag{8}$$

Applying the coattention mechanism to focus on both text and price, and learn about fusion. We first compute an affinity matrix that contains the corresponding affinity scores of all prices hidden states and texts hidden state pairs. Then the affinity matrix is normalized by Softmax, attention weights are generated for text features by row, and attention weights of price features are generated by columns. The calculation formula is as follows:

$$L = H_t \left(H'_p\right)^T \tag{9}$$

$$A_t = softmax(L) \tag{10}$$

$$A_p = softmax\left(L^T\right) \tag{11}$$

CCL 2022

Next, we calculate the attention context of price features based on the attention weight of text features. The calculation formula is as follows:

$$C_t = A_t H'_p \tag{12}$$

Meanwhile, we compute the attention context of the text features as $A_p H_t$ based on the attention weights of the price features. Following Xiong et al. (2016), we also calculate $A_p C_t$ which maps text feature encoding into the space of price feature encoding. The calculation formula is as follows:

$$h_d = A_p \left[H_t, C_t\right] \tag{13}$$

Where h_d is interdependent representation of the text and the price. The [ ] denotes for concatenation operation.

3.6 Global Fusion by Sequential Encoding

We input h_d obtained from Section 3.5 into the bidirectional GRU to obtain the hidden states for each time t. To capture past and future information as its context, we connect the hidden states from the two directions to construct a two-way encoding vector h_i with the following formulas:

$$\overrightarrow{h_i} = \overrightarrow{GRU}\left(h_d\right) \tag{14}$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}\left(h_d\right) \tag{15}$$

$$h_i = \left[\overrightarrow{h_i}, \overleftarrow{h_i}\right] \tag{16}$$

In addition to its own information, h_i also contains information about its adjacent contexts. In this way, we encoded its time series. Since news releases on different dates contributed unequally to stock trends, we employed soft attention mechanism which is calculated as follows:

$$o_i = \tanh\left(h_i W_h + b_h\right) \tag{17}$$

$$\beta_i = softmax\left(o_i W_o\right) \tag{18}$$

$$h_{final} = \sum_i \beta_i h_i \tag{19}$$

where β_i is the attention weight, W_h and W_o are the parameters to be learned, b_h is the bias terms. Finally, h_{final} is input into a classic three-layer preceptron (MLP) to predict the future trend of stocks.

3.7 Model Training

We use cross entropy for model training, which is calculated by equation (20), where N is the total number of stocks, y_i^t and $\hat{y_i^t}$ represent the ground truth and predict stock trend of stock i at t day, respectively.

$$l = -\sum_{i=1}^{N} \sum_t y_i^t ln(\hat{y_i^t}) \tag{20}$$

4 Experiments

4.1 Dataset

We use the SotckNet[1] dataset (Xu and Cohen, 2018) to train and validate the model. The dataset contains historical data on the high trading volumes of 88 stocks in the NASDAQ and NYSE stock markets. We annotate the samples based on the movement percent of the adjusted closing price of stock, and label the samples as up and down when movement percent $\geq 0.55\%$ or $\leq -0.5\%$, respectively. We split the dataset temporarily with 70/20/10, leaving us with date ranges from 2014-1-1 to 2015-8-1 for training, 2015-8-1 to 2015-10-1 for validation and 2015-10-1 to 2016-1-1 for testing. Similarly, we adjusted trading days by removing samples with missing prices or texts and further aligned data for all trading day windows to ensure that data is available for all trading days in all windows.

[1]https://github.com/yumoxu/stocknet-dataset

CCL 2022

4.2 Experiment Settings

We use a 5-day trading day sliding window to build the samples. Following the setting of Xu and Cohen (2018), we set the maximum number of texts in a day to 30, and each text has a maximum of 40 words. Glove word embedding was also used to initialize words into 50-dimensional vectors. We train the model using the Adam optimizer, with an initial learning rate set to 5e-5. The bidirectional GRU hidden dimensions for encoding tweets and sequential modeling were set to 100 and 64, respectively. Each model is trained for 40 epochs with a batch size of 32. We report the best average test performance of the model on the validation set at 5 different runs.

Following previous studies (Xu and Cohen, 2018; Sawhney et al., 2020), we use Accuracy (Acc), F1 score, and Matthews Correlation Coefficient (MCC) as evaluation metrics for this classification task.

4.3 Compared Models

To demonstrate the effectiveness of our proposed model, we compare the results with the following comparative models.

- **RAND.** A simple predictor to make random guess about the rise and fall.
- **ARIMA.** Autoregressive Integrated Moving Average, an advanced technical analysis method using only price signals. (Brown, 2004).
- **Adversarial LSTM.** Feng et al. (2019a) proposed a deep model using an adversarial attention LSTM mechanism, which exploits adversarial training to simulate randomness during model training. They propose the use of adversarial training to improve the generalization of neural network prediction models, since the input feature for stock prediction is usually based on stock price, which is essentially a random variable that naturally changes over time. They added perturbations to their stock data and trained the model to work well with small but intentional perturbations.
- **RandForest.** Pagolu et al. (2016) implemented a sentiment analysis model based on Twitter data. The authors used Word2vec to analyse the polarity of sentiments behind the tweets and directly assessed tweets related to stock and tried to predict the price of the stock for the next day.
- **TSLDA.** A new topic model, Topic Sentiment Latent Dirichlet Allocation (TSLDA), which can obtain new feature that captures topics and sentiments on the documents simultaneously and use them for prediction of the stock movement (Nguyen and Shirai, 2015).
- **HAN.** A hybrid attention network that predicts stock trends by imitating the human learning process. Follows three basic principles: sequential content dependency, diverse influence, and effective and efficient learning. The model includes news-level attention and temporal attention mechanisms to focus on key information in news (Hu et al., 2018).
- **StockNet.** A Variational Autoencoder (VAE) to encode stock inputs to capture randomness and use temporal attention to model the importance of different time steps (Xu and Cohen, 2018). We compare with the best variants of StockNet.
- **MAN-SF.** Multipronged Attention Network (MAN-SF) jointly learns from historical prices, tweets and inter stock relations. MAN-SF through hierarchical attention captures relevant signals across diverse data to train a Graph Attention Network (GAT) for stock prediction. And the study considers one pre-built graph from Wikidata (Sawhney et al., 2020).

4.4 Experimental Results

We conduct several experiments to evaluate the performance of our method. In this section, we analyze the benchmark performance and the results of our model on the StockNet dataset. The experimental results of the different models are shown in Table 2.

First, we compare the first three baseline models presented in this paper. All three baseline methods use only historical price information, although Adversarial LSTM with more representative features and

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	Acc	F1	MCC
RAND	50.9	50.2	-0.002
ARIMA (Brown, 2004)	51.4	51.3	-0.021
Adversarial LSTM (Feng et al., 2019a)	57.2	57.0	0.148
RandForest (Pagolu et al., 2016)	53.1	52.7	0.013
TSLDA (Nguyen and Shirai, 2015)	54.1	53.9	0.065
HAN (Hu et al., 2018)	57.6	57.2	0.052
StockNet (Xu and Cohen, 2018)	58.2	57.5	0.081
MAN-SF (Sawhney et al., 2020)	60.8	60.5	0.195
ours	**62.6**	**61.1**	**0.228**

Table 2: The results of different models.

training with adversarial learning achieved better performance. Our model clearly exceeds these three methods in each evaluation indicator.

Second, our model is compared to models that only use textual information, such as RandForest, TSLDA, and HAN. Our model also significantly outperforms these three methods, outperforming the best-performing HAN by 5, 3.9, and 0.176 in Acc, F1, and MCC, respectively. So far, we can find that the performance of the model using only price or text is not satisfactory enough.

Finally, compared to StockNet, which also uses texts and prices, our model is 4.4, 3.6 and 0.147 higher on Acc, F1 values and MCC, respectively. Compared to another MAN-SF using the same data, our model contains no additional knowledge of stock relations. But the result still demonstrates that our model is 1.8, 0.6, and 0.033 higher than the MAN-SF on Acc, F1 values, and MCC, respectively. Overall experimental results demonstrate the effectiveness of the proposed model.

4.5 Ablation Study

In order to better demonstrate the different effects of components of our method, we conduct ablation studies to investigate the different contribution of coattention mechanism and the three proposed financial indicators. The results are shown in Table 3. We mainly design two variants: ours w/o coattention and ours w/o ATR-BIAS-MTM.

For w/o coattention, we change the method of learning effective implicit information between price and text from the coattention mechanism to the direct concatenation of the two. This model drops 1.7, 0.7 and 0.014 compared to the full model on Acc, F1 value and MCC, respectively, proving that the coattention mechanism can effectively improve the performance of the model and obtain richer information between price and text.

For w/o ATR-BIAS-MTM, We remove the three features proposed earlier in this paper and only use the 11 features proposed in previous studies(Feng et al., 2019a). The experimental results of the model decreased by 0.3, 0.3 and 0.007 on Acc, F1 values and MCC, respectively, which also prove that these three features help the performance of the model by reflecting the volatility of the market. Here we take ATR as an example to analyze, it can simply be understood as the expectations and enthusiasm of traders. Large or increasing volatility indicates that traders may be prepared to continue buying or selling stocks during the day. A reduction in volatility indicates that traders are not showing much interest in the stock market.

Model	Acc	F1	MCC
ours	**62.6**	**61.1**	**0.228**
w/o coattention	60.9	60.4	0.214
w/o ATR-BIAS-MTM	62.3	60.8	0.221

Table 3: The ablation study of our method.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4.6 Case Study

As mentioned before, we use the coattention mechanism in the model to capture richer information, which in turn help to learn more precise attention weights of intra-day tweets (Tweet-level attention) and inter-day of time slide window (Temporal attention). In order to investigate how the coattention mechanism guides the learning of attention weights, we conducted a case study on a sample of $FB(FaceBook) between Nov 5^{th} and Nov 9^{th}, 2015, which is finally used to predict the rise or fall of Nov 10^{th}, 2015.

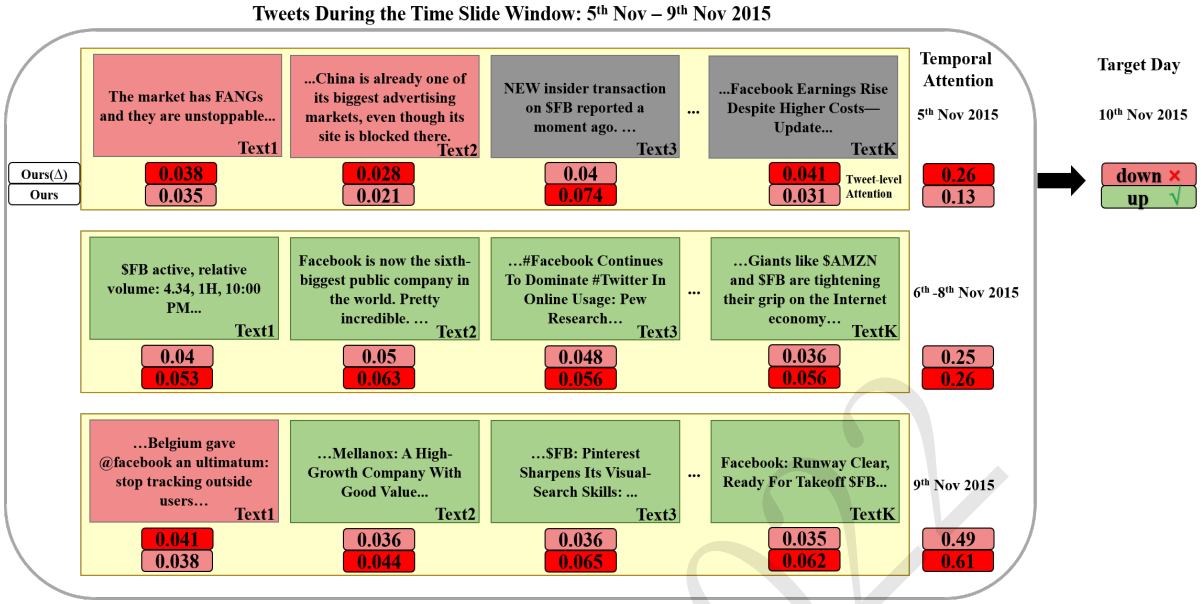


Figure 3: Text-level and Temporal attention weights learned by Ours and Ours(Δ) (as mentioned ours w/o coattention) on a sample of $FB (FaceBook). Numbers represent weight values and darker colors indicate greater weights. Text on green, red and grey backgrounds represent signals with positive, negative and neutral respectively.

As shown in Figure 3, a row represents a day. For example, the first row represents texts of 5^{th}. And we use the trading day alignment, because the 7^{th} and 8^{th} are weekends, so the text data for the three days from the 6^{th} to the 8^{th} were merged together. Each rectangle inside each row represents the content of a text. All texts within a day are denoted as $[Text1, Text2, \ldots, TextK]$. And We present the attention weights learned by our model (Ours) and without coattention mechanism (denoted as Ours(Δ)).

First, we can see that the closer to the target day, the more weight Ours gives to that day. This is also in line with the laws of the real world, and the newer news can have a greater impact. Specifically, Ours pays more attention to the positive signals from the 6^{th} to the 9^{th}. On the 5^{th}, it pays too much attention to a neutral $Text3$ whose impact is uncertain. However, because of giving it a lower weight on the day, it can help its correct prediction for the rise. On the contrary, Ours(Δ) has a greater weight than Ours on the 5^{th}. At the same time, the tweet texts with negative signals in the 5^{th} and 9^{th} are more concerned by Ours(Δ), and finally make a wrong prediction.

Next we analyze the texts for each day in more detail. For a more intuitive understanding, we artificially add different background colors to each rectangle to represent different tendencies of the text, such as green, red and grey backgrounds representing signals with positive, negative and neutral respectively. On the 5^{th} day, we can see that Ours(Δ) has higher attention than Ours on the two negative texts $Text1$ and $Text2$. During the period from the 6^{th} to the 9^{th}, Ours gives a higher weight value to the texts with positive signals than Ours(Δ), such as the $Text2$ from the 6^{th} to the 8^{th} and the $TextK$ of the 9^{th}, which all reflect the good development prospects of FaceBook. In particular, Ours has a smaller weight than Ours(Δ) on the $Text1$ with negative influence in 9^{th}. Although this negative news appears on the day closest to the target prediction, because the model combined with coattention can fuse the information of the entire window, and analyzes that Facebook stock is still showing an upward trend in general.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

The observation shown in Figure 3 indicates that the coattention mechanism can guide the model to pay more attention to texts with tendencies and can effectively model the temporal. With more accurate attention weights, Ours can capture more effective representation, thus it can achieve better performance than Ours(Δ).

5 Conclusion

To effectively fuse texts and prices to predict future stock movements, in this paper, we propose a fundamental analysis based neural network for stock movement prediction. Our model introduces the coattention mechanism to capture richer implicit information between text and price as a better representation of a stock. We also introduce three new technical indicators in the financial field as price features. We perform the extensive experiments on the StockNet dataset and the experimental results show the effectiveness of our proposed method. In the future, we plan to use more data other than stock prices, such as financial reports, relationships between stock, to better capture market dynamics. In addition, extracting features that can better reflect trend changes is still a direction worth exploring.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61976062).

References

Klaus Adam, Albert Marcet, and Juan Pablo Nicolini. 2016. Stock market volatility and learning. *The Journal of finance*, 71(1):33–82.

Robert Goodell Brown. 2004. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

Renato Bruni. 2017. Stock market index data and indicators for day trading as a binary classification problem. *Data in brief*, 10:569–575.

Rui Cheng and Qing Li. 2021. Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 55–62.

Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Jian Guo. 2020. Hierarchical multi-scale gaussian transformer for stock movement prediction. In *IJCAI*, pages 4640–4646.

Eugene F Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019a. Enhancing stock movement prediction with adversarial training. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5843–5849. International Joint Conferences on Artificial Intelligence Organization, 7.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019b. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems (TOIS)*, 37(2):1–30.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 261–269.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Tao Lin, Tian Guo, and Karl Aberer. 2017. Hybrid neural networks for learning the trend in time series. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence*, pages 2273–2279.

David MQ Nelson, Adriano CM Pereira, and Renato A De Oliveira. 2017. Stock market's price movement prediction with lstm neural networks. In *2017 International joint conference on neural networks (IJCNN)*, pages 1419–1426. IEEE.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. 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 1354–1364.

Huihui Ni, Shuting Wang, and Peng Cheng. 2021. A hybrid approach for stock trend prediction based on tweets embedding and historical prices. *World Wide Web*, 24(3):849–868.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE.

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.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Shah. 2020. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426.

Tsun-Hsien Tang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Retrieving implicit information for stock movement prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2010–2014.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *6th International Conference on Learning Representations*.

J Welles Wilder. 1978. *New concepts in technical trading systems*. Trend Research.

Yong Xie, Dakuo Wang, Pin-Yu Chen, Xiong Jinjun, Sijia Liu, and Oluwasanmi Koyejo. 2022. A word is worth a thousand dollars: Adversarial attack on tweets fools stock prediction. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Yu Zhao, Huaming Du, Ying Liu, Shaopeng Wei, Xingyan Chen, Huali Feng, Qinghong Shuai, Qing Li, Fuzhen Zhuang, and Gang Kou. 2022. Stock movement prediction based on bi-typed and hybrid-relational market knowledge graph via dual attention networks. *arXiv preprint arXiv:2201.04965*.

CCL 2022

Proceedings of the 21st China National Conference on Computational Linguistics, pages 973-984, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China