

基于特征融合的无标复句关系识别

杨进才, 马晨, 肖明

引用本文

杨进才, 马晨, 肖明. [基于特征融合的无标复句关系识别](#)[J]. 计算机科学, 2023, 50(11A): 221100065-6.

YANG Jincan, MA Chen, XIAO Ming. [Relation Recognition of Unmarked Complex Sentences Based on Feature Fusion](#) [J]. Computer Science, 2023, 50(11A): 221100065-6.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于边缘引导的多尺度医学影像分割方法](#)

Medical Image Segmentation Based on Multi-scale Edge Guidance

计算机科学, 2023, 50(11A): 220900059-7. <https://doi.org/10.11896/jsjcx.220900059>

[基于语义注意力的医学图像超分辨率方法](#)

Medical Image Super-resolution Method Based on Semantic Attention

计算机科学, 2023, 50(11A): 221200107-6. <https://doi.org/10.11896/jsjcx.221200107>

[一种基于因果推理的垃圾分类方法](#)

Novel Method for Trash Classification Based on Causal Inference

计算机科学, 2023, 50(11A): 220800218-6. <https://doi.org/10.11896/jsjcx.220800218>

[接诉即办智能派单业务调度算法研究](#)

Study on Scheduling Algorithm of Intelligent Order Dispatching

计算机科学, 2023, 50(11A): 230300029-7. <https://doi.org/10.11896/jsjcx.230300029>

[基于LSTM神经网络的QPSK智能接收机设计](#)

Design of QPSK Intelligent Receiver Based on LSTM Neural Network

计算机科学, 2023, 50(11A): 230200219-5. <https://doi.org/10.11896/jsjcx.230200219>

基于特征融合的无标复句关系识别

杨进才¹ 马晨¹ 肖明²

¹ 华中师范大学计算机学院 武汉 430079

² 华中师范大学语言与语言教育研究中心 武汉 430079

摘要 无标复句因缺少关联词的辅助,其关系识别为自然语言处理中的一项较为困难的任务。将词性特征融入到词向量中,训练得到含有外部特征的词向量表示,通过组合 BERT 模型与 BiLSTM 模型,将字向量、词向量、词性向量结合进行训练,并在特征融合层添加 BiLSTM 模型捕获的极性特征信息以及 CNN 模型捕获的依存句法特征信息。实验结果表明,该方法在汉语复句分类上取得了较好的效果,与基准模型相比在宏 F1 值与微 F1 值上均有提升,在顶层分类上取得了 83.67% 的微 F1 值,在第二层分类上取得了 68.28% 的微 F1 值。

关键词: 无标复句;BERT;特征融合;深度学习

中图法分类号 TP391

Relation Recognition of Unmarked Complex Sentences Based on Feature Fusion

YANG Jincai¹, MA Chen¹ and XIAO Ming²

¹ School of Computer Science, Central China Normal University, Wuhan 430079, China

² Research Center for Language and Language Education, Central China Normal University, Wuhan 430079, China

Abstract Unlike marked complex sentences, which lack the assistance of relation words, the identification of unmarked complex sentences is a difficult task in natural language processing. Integrating part of speech features into word vectors, and the word vector representation containing external features is obtained by training. By combining the BERT model and the BiLSTM model, the word vector and the part-of-speech vector are combined for training, and the polar feature information captured by BiLSTM model and the dependency syntax feature information captured by CNN model are added to the feature fusion layer. Experimental results show that the methods of adding features and combining multiple deep learning models can achieve better results in classification of Chinese complex sentences. Compared with the benchmark model, the macro F1 value and micro F1 value are improved. The best classification effect achieves 83.67% micro F1 value in the top layer classification and 68.28% micro F1 value in the second layer classification.

Keywords Unmarked complex sentence, BERT, Feature fusion, Deep learning

1 引言

复句在现代汉语中占据了十分重要的地位,关联词是用来连接各个分句构成复句的语法成分,但在汉语复句中,分句中的关联词可以部分省略或全部省略^[1],关联词全部省略时复句即为无标复句。无标复句关系识别旨在识别两个分句之间的语义关系,为下游任务提供帮助,例如文本摘要^[2]、信息抽取^[3]等,为上层的篇章分析^[4]打好基础,该任务在自然语言处理中具有重大意义。

目前,对汉语有标复句以及部分关联词缺省的非充盈态有标复句的研究已取得不少成果,而对关联词全部缺省的无标复句的研究还非常缺乏。在有标复句中,可以依据出现的关联词搭配来确定复句类别,而无标复句中因缺少连接分句的关联词的帮助,使无标复句关系的识别比有标复句更具挑战性。

例如 1 中缺省了关联词“虽然,但是”,但分析前后分句所

表达的语义,可以判断出其为转折关系。例 2 中缺省了关联词“既然,就”,但可以根据前后分句的语义,推测出后一分句为前一分句的结果,因此为因果关系。

例 1 地震无情,人间有爱。

例 2 有很多事情你无法改变,只能去适应。

2 相关研究

复句是联结句子与篇章的桥梁,根据是否存在连接词,篇章关系识别可以分为显式篇章关系识别和隐式篇章关系识别两大类^[5]。无标复句的关系识别对应于隐式篇章关系识别。

相比英文,目前汉语无标复句句料资源缺乏,已有的研究工作较少,并且大部分借鉴了英文任务中采用的方法,例如 CNN, RNN, LSTM 等。

Kim^[6]首次将卷积神经网络应用到句子分类任务中,并取得了较好的效果。Zeng^[7]使用 CNN 模型,加入单词位置特征、名词词汇特征等来实现关系分类。Cai 等^[8]基于 CNN

基金项目:国家社科基金(19BYY092)

This work was supported by the National Social Science Fund of China(19BYY092).

通信作者:杨进才(jcyang@mail.ccnu.edu.cn)

与 LSTM,提出了一种 BRCNN 模型架构并获取了双向文本语义信息,在句级别上解决了关系分类问题。Fan 等^[9]使用 BiLSTM 模型结合句法特征在隐式篇章关系识别上得到了比基准模型更好的效果。

上述均为对英文语料的研究,在对中文语料的研究中,Rönnqvist 等^[10]将循环神经网络(RNN)以及注意力机制应用到句间关系分析任务中,取得了较好的泛化能力。Tian 等^[11]于 2019 年提出了一种基于 BiLSTM 的多任务学习架构,联合显式与隐式语料解决了句间关系分析的问题,与 Ji 等^[12]的方法相比在平均 F1 值上有了提升,是对分句间关系、复句间关系、句群间关系的整体研究,并没有专门针对分句间的关系进行研究。Wang 等^[13]则通过使用 BERT 模型并结合主动学习的数据增强,在中文隐式篇章关系识别的三分类中取得了 72.1% 的宏 F1 值与 79.4% 的微 F1 值。

句子中每个词都具有不同的词性,在词性中也包含着一些重要的语义信息。为了学得句中更深层次的特征信息,可以通过标注词性的方法来区分词在句子结构或语义表达中的不同作用^[14-16],以帮助模型提取相应的特征。Zhang 等^[17]中通过引入词性特征构建词性向量来改善文本表示,提升了模型分类的准确率。

基于自然语言是由词汇和句法组成的事实,Shen 等^[18-19]将树结构的句法特征集成到循环神经网络,取得了较标准 LSTM 更好的效果。Peng 等^[20]将句法分析器融入到神经网络语言模型中,句法树监督的语言模型取得了较无监督隐式的语言模型更好的效果。

在汉语复句分句间关系的识别方面,Sun 等^[21]在神经网络框架基础上引入了句内注意力机制,模型的宏 F1 值达到了 85%。Yang 等^[22]采用卷积神经网络并融合关系词特征,在二句式非充盈态复句上的识别准确率达到 97%。以上研究是对有标复句进行的顶层三分类。

3 语料资源

本文选用的是邢福义先生提出的三分体系,第一层三大语义的关系为因果类、转折类、并列类,第二层包含 12 种语义关系,即因果句、假设句、目的句、推断句、条件句、转折句、让步句、假转句、并列句、连贯句、递进句、选择句。自建的自然无标语料库主要来源于 3 个部分。

1) THUCNEWS 中文文本分类数据集,它是清华大学根据新浪新闻 RSS 订阅频道 2005—2011 年间的历史数据筛选过滤生成的,包含 74 万篇新闻文档。

2) 哈尔滨工业大学中文篇章关系语料 HIT-CDTB, HIT-CDTB 语料内容来源于广播新闻、杂志和网络等,其中定义了显式和隐式两种关系体系,并根据不同的粒度,将篇章关系的语义结构类型分为六大类。针对每一篇原始语料,标注的结果分为了 3 类:分句句间关系、复句句间关系和句群句间关系。由于该语料分类体系不同于与本文选用的三分体系,因此在 HIT-CDTB 标注的隐式句间关系语料中挑选出部分能直接对应到三分体系的复句,并将其作为自然无标复句句料。

3) 考虑到大部分研究都是基于新闻语料的,而现实生活中,人们在日常交流的口语化表达时也存在无标复句的情况,因此引入部分中文近代小说的语料,如《平凡的世界》《活着》《白鹿原》《围城》等。

表 1 无标复句句料各类别的分布

Table 1 Distribution of various categories of unmarked complex

sentence corpus			
三分类	数据量	十二类	数据量
因果类	5 014	因果句	3 528
		假设句	446
		推断句	320
		目的句	538
		条件句	182
转折类	1 037	转折句	815
		让步句	217
		假转句	5
并列类	5 949	并列句	3 437
		连贯句	1 054
		递进句	1 422
		选择句	36

将上述复句句料通过 jieba 分词后将含有准关联词的复句全部除去,最终得到无标复句 205 914 条。通过人工方式标注 11 021 条,并将其作为数据集,结合 HIT-CDTB 中已标注的复句句料 979 条,得到自然无标复句数据 12 000 条,各类复句数量如表 1 所列。将标注后余下的 193 914 条作为测试集部分,在实验中随机抽取 1 400 条复句用于验证模型的效果。

4 模型与实验

4.1 模型的测试

在常用的深度学习模型中,CNN,RNN,LSTM 以及其变体在语言学方面都得到了广泛的运用。使用 CNN,RNN,LSTM,BiLSTM 模型对无标复句进行训练,在随机选取的 1 400 条测试数据上的顶层三分类评估值如表 2 所列,训练时长如表 3 所列。

表 2 常用模型对无标复句的分类效果

Table 2 Classification results of common models for unmarked complex sentences

(单位:%)			
指标	模型	顶层分类	第二层分类
宏 F1	RNN	55.42	11.10
	CNN	56.32	21.44
	LSTM	55.64	20.08
	BiLSTM	57.61	21.49
微 F1	RNN	67.57	50.46
	CNN	68.00	50.46
	LSTM	67.29	0.61
	BiLSTM	69.64	50.75

表 3 常用模型的训练时长

Table 3 Training time of common models

(单位:s)				
模型	CNN	RNN	LSTM	BiLSTM
时长	85	90	105	136

观察实验结果可以发现,在常用文本分析神经网络中 BiLSTM 取得了最佳的分类效果。BERT 模型可以做到并发执行,能在多个不同层次提取关系特征,进而更全面地反映句子语义。因此,在下一步实验中,将 BiLSTM 与 BERT 模型结合,并分析添加特征后的模型分类效果。

4.2 特征提取

深度学习模型能够自动从词向量中提取特征,但一些对关系类别有影响的特征却无法捕获,需要在模型中进行融入。

4.2.1 词性特征

一词多义的现象在汉语中普遍存在,具体例句如下:

例3 你能跟我说说事情的经过(名词,意为经历的过程)吗?

例4 今天我经过(动词,意为从某处过)你家门口。

例5 到现在大约经过(介词,意为通过)六、七人之手

例6 经过比较(动词,意为辨别事物的相同属性异同或高低),我认为第一种方式更合适。

例7 今天天气比较(副词,意为表示具有一定程度)冷。

例句中的词语“经过”与“比较”在不同的语境中分别表示不同的含义。传统的词向量模型并不能有效地区分一词项的不同语义,从而无法为多义词建模,使得词嵌入时将同一词项的不同语义编码为相同的向量表示,造成了模型对语句语义理解的偏差。

Stanford CoreNLP 是斯坦福(Stanford) NLP 组开发的 NLP 工具,提供工具来实现分词、分句、词性标注、句法分析等许多常见的 NLP 任务。使用 Stanford CoreNLP 获取词性标签序列,然后通过神经网络对输入句子进行编码,获取核心词和依存词,从而提升模型的性能^[9]。选用 StanfordNLP 工具的 pos_teg 方法获取词性后将词项与词性进行组合,得到“词项-词性”的形式,使用 Word2Vec 进行训练,得到含有词性的词向量表示。

训练选取的向量维度为 100 维和 50 维度,训练的迭代次数为 50,70,100,通过实验分析,维度 100 训练 50 次得到的向量效果最佳。将训练得到的词性向量拼接在普通词向量后,得到 300 维含有词性向量增强的向量表示。

4.2.2 极性特征

在隐式篇章关系识别中,极性特征的融入有助于提高准确率^[23]。直观上看,转折类复句前后分句表达的情感是相反的,而因果类复句中因和果所表达出的情感极性往往是一致的,并列类复句分句间的情感关系通常是一致或者增强。

首先收集了多组情感词典,包括知网、台湾大学、清华大学以及 bosonNlp 情感词典,将它们合并后去除其中的特殊字符、人名、单词、数字,最终得到包含 110065 词的情感词典。

给每个词赋一个情感极性值,一些强调性的内容,例如“非常”“极为”将会有较大的极性值,这些内容往往是句中重点表达的内容,通过极性值也有助于模型从语义的角度发现近义词与反义词。部分词汇的极性值如表 4 所列。

表 4 部分词汇的情感倾向以及极性值

Table 4 Emotional tendencies and polarity values of some words

词语	情感倾向	极性值
很	正向	0.530243
非常	正向	0.992129
好	正向	2.073294
不好	负向	-1.388997
开心	正向	2.612342
悲伤	负向	-1.215332
天晴	正向	1.082743
下雨	负向	-1.240289

4.2.3 依存句法特征

复句经过分词后,模型通过语序来对词语建立联系,通常是关注与当前词相邻的词。而依存句法特征可以将分词后得到的词语建立起联系,这种联系有时可以跨越多个词语,使距离较远的词关联起来,从而使模型更好地理解前后分句间的

语义关系。

依存句法指出词语在句法上的搭配关系,这种搭配关系与语义相关联,同时也有利于神经网络模型理解复句结构,有助于模型找出句子的中心词、名词中心词和动词中心词。因此通过获取依存句法特征,可以更好地捕获语义信息,更有效地理解自然语言。

例如“额度紧张,只能试试看”,使用 StanfordNLP 工具的 dependency_Parse 方法获取到的依存关系三元组为: [(‘ROOT’,0,2), (‘nsubj’,2,1), (‘punct’,2,3), (‘advmod’,6,4), (‘aux:modal’,6,5), (‘conj’,2,6), (‘punct’,2,7)]。

依存关系如图 1 所示,从图 1 中可以发现“紧张”和“试试看”作为中心词连接着两个分句,并且可以看到词语之间的依存关系,例如‘nsubj’表示的是名词主语。

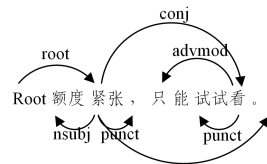


图 1 例句的依存句法关系图

Fig. 1 Dependency syntax diagram of example sentence

4.3 基于 BiLSTM 的训练结果

BiLSTM 模型在复句三分类上各小类的 F1 值以及模型整体的 F1 值如表 5 所列。通过表 5 可以看出,在 BiLSTM 模型上,添加外部特征后模型在各小类以及整体的评估值上均有提升,其中句法特征的引入使模型取得了在宏 F1 值与微 F1 值上最佳的效果。同时,该模型对因果类以及转折类的复句具有较好的识别效果,在因果类上取得了最佳的 F1 值(69.91%),在转折类上的最佳 F1 值为 37.65%,整体宏 F1 值上的最高值客运量为 61.21%,微 F1 值为上的最高值为 70.00%。添加词性特征的模型取得了在并列类上的最佳分类效果,F1 值为 76.65%。通过研究发现,通过外部特征的引入,在数据量较少的转折类上,模型的效果得到了很大的提升。

表 5 BiLSTM 模型顶层三分类的 F1 评估值

Table 5 Experimental results of BiLSTM model in top layer classification

模型	因果类	转折类	并列类	宏 F1	微 F1
BiLSTM	67.91	29.07	75.86	57.61	69.64
BiLSTM+词性特征	67.85	37.23	76.65	60.58	69.86
BiLSTM+极性特征	68.73	33.68	74.96	59.12	69.43
BiLSTM+句法特征	69.91	37.65	76.07	61.21	70.00
BiLSTM_CNN+词性+极性+句法特征	68.53	37.50	75.32	60.45	69.57

(单位:%)

BiLSTM 模型在第二层类上各小类的 F1 值以及模型整体的 F1 值如表 6 所列。通过分析表 6 中的数据可以得出,添加外部特征后,模型在整体的宏 F1 值与微 F1 值上都有提升,其中添加词性特征的模型取得了最佳的宏 F1 值 25.54%,较基准模型相比提升了约 3%,同时添加 3 种特征的模型取得了最佳的微 F1 值 52.82%,与基准模型相比提升了约 2%。在对复句语义关系进行进一步细分的第二层分类下,从总体来看,同时添加 3 种特征的模型取得的效果最好。与不添加

特征的模型相比,在多个小类上的识别能力均有提升,特别是目的句和转折句,这两类复句的数据量较少,添加外部特征的方法可以很有效地帮助模型识别出数据集中分布较少数据类型。未添加特征的模型在并列类复句中的

并列句、递进句这两类数据较多的复句上 F1 值较高,在数据量较少的复句上 F1 值较低,其原因可能是模型并未充分学得其中的语义信息,倾向于将复句识别为数据量较多的类型。

表 6 BiLSTM 模型第二层分类的 F1 评估值

Table 6 Experimental results of BiLSTM model in the second layer classification

(单位:%)

类别	BiLSTM	BiLSTM+ 词性特征	BiLSTM+ 极性特征	BiLSTM+ 句法特征	BiLSTM_CNN+词性+ 极性+句法特征
因果	54.42	59.54	59.20	58.40	61.08
目的	5.88	31.46	27.27	32.73	33.33
推断	0.00	0.00	0.00	0.00	0.00
假设	22.02	24.10	0.00	3.45	8.45
条件	0.00	0.00	0.00	8.70	0.00
转折	8.70	23.93	13.46	8.89	19.15
让步	0.00	0.00	0.00	0.00	0.00
并列	66.91	63.12	64.89	63.10	62.81
连贯	51.66	56.04	60.73	55.36	62.56
递进	27.03	11.73	25.39	7.92	18.49
宏 F1	21.51	24.54	22.81	21.69	24.17
微 F1	50.75	51.10	51.46	50.89	52.82

4.4 基于 BERT 的训练结果

BERT 在许多语言处理任务中都取得了很好的成绩,由于中文处理方面 BERT 是基于字级别的,在中文领域,词比字蕴含的信息更丰富,语义更充分。因此,本文将 BERT 模型与 BiLSTM 模型进行结合,融合后的模型将考虑到句子的词特征和字特征,并融入词向量特征,同时使用 CNN 网络对二维矩阵形式的句法特征进行信息提取。

通过jieba分词来获取词性,首先通过 Embedding 层将其转换为词向量,然后输入至 BiLSTM 模型。使用情感极性特征值对 Embedding 层赋初值,分词后的词序列通过 Embedding 层转换为极性值序列,然后输入至 BiLSTM 模型以提取特征。对于句法特征,使用二维的 CNN 网络,后经过 Pool 层,由 Flatten 层进行展开。模型通过 Linear 层将得到的特征向量维度压缩后在特征融合层进行叠加,最终输出预测结果,模型结构如图 2 所示。

实验中首先将文本处理为字序列并输入 BERT 模型,

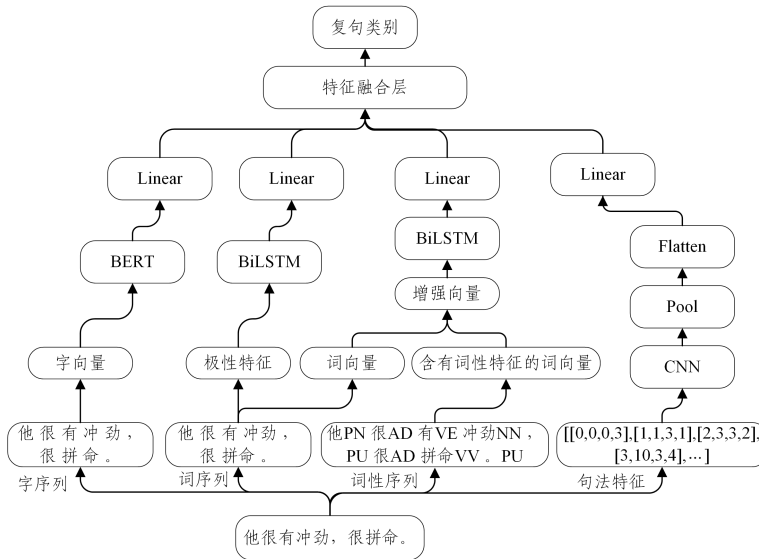


图 2 融入外部特征的 BERT_BiLSTM_CNN 模型的结构

Fig. 2 Structure of BERT_BiLSTM_CNN network with feature fusion

表 7 BERT 模型与融合模型在三分类上的评估值

Table 7 Experimental results of Bert and advanced model in top layer classification

(单位:%)

模型	因果类	转折类	并列类	宏 F1	微 F1
BERT	80.72	48.73	87.36	72.27	81.64
BERT_BiLSTM	80.28	55.00	86.75	74.01	81.72
BERT_BiLSTM+词性特征	80.93	58.02	86.86	75.27	82.66
BERT_BiLSTM+极性特征	83.62	57.63	87.29	76.18	83.67
BERT_CNN+句法特征	81.62	63.32	87.08	77.34	82.97
BERT_BiLSTM_CNN+词性+极性+句法特征	80.50	57.40	87.72	75.20	82.34

基于 BERT 与融入外部特征模型在三分类上的各小类的 F1 值与模型整体的 F1 值如表 7 所列。首先使用 BERT 模型对自然无标复句进行训练,观察实验结果可以发现,BERT 模型的识别效果远高于 BiLSTM 模型,但在同样的训练集上 BERT 模型训练的时长接近 4 小时,远长于普通神经网络的训练时长。如果能在 BERT 模型中结合使用 BiLSTM 模型对复句信息进行捕获,可以在增加少量训练时长的情况下提升模型的分类效果。在不添加外部特征,仅结合使用 BERT 与 BiLSTM 模型的情况下,虽然因果类和并列类的 F1 值有微量的下降,但模型整体的宏 F1 值和微 F1 值均有提升,因为改进后的模型在数据量较少的转折句上取得了更好的成绩,提升了约 6%,证实了改进模型 BERT_BiLSTM 的有效性。

然后引入词性、极性、句法这 3 个外部特征,观察表 7 所列的实验结果可以发现,融入外部特征后,模型的效果得到了进一步的提升,其中添加极性特征的模型取得了在因果类上的最佳 F1 值 83.62% 以及最佳的整体微 F1 值 83.67%。添加句法特征的模型在转折类上取得了最佳的 F1 值 63.32%,较基准 BERT 模型提升了约 15%,在整体的宏 F1 上取得了最高值 77.34%,较基准 BERT 模型提升了约 5%。同时添加 3 种特征的模型取得了在并列类上的最佳分类效果,F1 值为

87.72%。词性特征虽然未能在基于 BERT 的模型中得到突出表现,但词性特征的引入,在顶层三分类以及模型整体的 F1 值上较未添加词性特征时都有提升,只是没有极性特征与句法特征更为明显。

基于 BERT 与融入外部特征的模型在第二层分类上,各小类的 F1 值以及整个分类任务的宏 F1、微 F1 值如表 8 所列。观察实验结果可以发现,在第二层分类上,当未添加特征时,BERT_BiLSTM 模型较基准的 BERT 模型在因果句、目的句、转折句、并列句、递进句这 5 个小类以及模型整体的微 F1 值上有所提升,但在假设句、条件句、连贯句以及模型整体的宏 F1 值上有所下降。添加外部特征后,模型在之前未曾识别出的推断句上取得了 17% 左右的 F1 值,虽然效果仍然不好,但从无法识别出推断句,到能够识别出一部分,已经有很大的提升。添加词性特征的模型取得了在并列句以及整体的最佳微 F1 值,添加极性特征的模型取得了在推断句上的最佳 F1 值,而同时添加 3 种外部特征的 BERT_BiLSTM_CNN 模型,在 9 类能识别出的复句中取得了 6 种小类上的最佳 F1 值以及模型最佳的宏 F1 值,特别是在目的句与假设句的识别上,F1 值分别提升了 6% 和 8%。由此可以证实融入词性特征、极性特征以及依存句法特征有助于神经网络更好地理解自然语言。

表 8 BERT 模型与融合模型在第二层分类上的 F1 评估值

Table 8 Experimental results of BERT and advanced model in the second layer classification

(单位:%)

类别	BERT	BERT_BiLSTM	BERT_BiLSTM+ 词性特征	BERT_BiLSTM+ 极性特征	BERT_CNN+ 句法特征	BERT_BiLSTM_CNN+ 词性+极性+句法特征
因果句	72.39	72.58	74.53	74.22	74.94	75.40
目的句	52.89	58.33	57.60	54.17	57.14	58.33
推断句	0.00	0.00	17.02	17.78	0.00	9.30
假设句	57.38	53.45	61.02	61.11	60.34	65.38
条件句	58.82	41.18	20.00	45.16	48.28	55.00
转折句	52.38	54.86	54.55	55.91	46.63	56.63
让步句	0.00	0.00	0.00	0.00	0.00	0.00
并列句	77.08	78.12	79.25	76.59	77.21	77.06
连贯句	75.19	73.77	75.59	74.24	74.90	76.63
选择句	0.00	0.00	0.00	0.00	0.00	0.00
递进句	36.23	38.85	35.71	38.40	34.65	41.01
宏 F1	43.85	42.83	43.21	45.23	43.10	46.79
微 F1	66.36	66.86	68.28	67.64	67.78	67.57

4.5 实验结果

基准模型与组合模型在第二层分类上的评估值如图 3 所示。观察图 3 可以发现,3 种特征的引入都能在一定程度上提升模型的分类效果,在微 F1 值上均优于未添加特征的模型。同时添加 3 种特征的方法,取得了最高的宏 F1 值,较基准模型 BERT 提升了约 3%,较基准模型 BiLSTM 提升了约 25%。

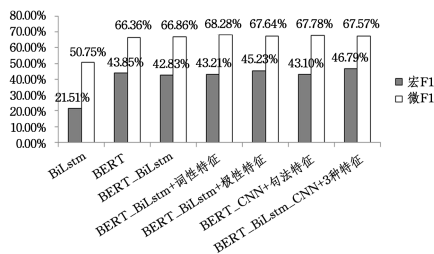


图 3 各模型在第二层分类上的评估值

Fig. 3 Evaluation value of each model on the second layer classification

结束语 本文研究汉语无标复句关系分类任务,基于 BiLSTM,CNN 与 BERT 模型,提出了融合词性特征、极性特征与句法特征的 BERT_BiLSTM_CNN 无标复句关系分类模型。实验结果表明,文本所构建的系统能够获得比单独使用 BiLSTM 或 BERT 模型更好的分类效果,并且外部特征的引入能进一步提升模型的性能。

从实验结果可以看出,对于无标复句关系分类任务目前的方法取得的效果仍不够理想。下一步工作中,还会尝试引入现有的有标复句资源,挖掘其潜在的与无标复句相似的特征来辅助模型对无标复句的分类,以及在模型上做进一步的改进来取得更好的效果。

参 考 文 献

[1] XING F Y. Research on Chinese Complex Sentences[M]. Beijing: The Commercial Press, 2001: 1-37.
 [2] COHAN A, DERNONCOURT F, KIM D S, et al. A Discourse-Aware Attention Model for Abstractive Summarization of Long

- Documents[C]//Proceedings of NAACL. 2018;615-621.
- [3] TANG J, LIN H, LIAO M, et al. From Discourse to Narrative: KnowledgeProjection for Event Relation Extraction[C]// 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). 2021;732-742.
- [4] LIU Y, LI S, ZHANG X, et al. Implicit Discourse Relation Classification via Multi-Task Neural Networks[J]. AAAI Press, 2016, 30(1):2750-2756.
- [5] SUN J, LI Y C, ZHOU G D, et al. Research of Chinese Implicit Discourse Relation Recognition[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1):111-117.
- [6] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Doha, Qatar: Association for Computational Linguistics, 2014;1746-1751.
- [7] ZENG D J, LIU K, LAI S W, et al. Relation Classification via Convolutional Deep Neural Network[C]// The 25th International Conference on Computational Linguistics: Technical Papers (COLING 2014). 2014;2335-2344.
- [8] CAI R, ZHANG X D, WANG H F. Bi-directional recurrent convolutional neural network for relation classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016;756-765.
- [9] FAN Z W, ZHANG M, LI Z H. BiLSTM-based Implicit Discourse Relation Classification Combining Self-attention Mechanism and Syntactic Information[J]. Journal of Computer Science, 2019, 46(5):214-220.
- [10] RÖNNQVIST S, SCHENK N, CHIARCOS C. A Recurrent Neural Model with Attention for the Recognition of Chinese Implicit Discourse Relations[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017;256-262.
- [11] TIAN W H, GAO Y Q, HUANG H W, et al. Implicit Discourse Relation Analysis Based on Multi-task Bi-LSTM[J]. Journal of Chinese information processing, 2019, 33(5):47-53.
- [12] JI J H, ZHANG M Y, QIN B, et al. The Chinese Discourse Parser[J]. Journal of Jiangxi Normal University (Natural Science), 2015(2):124-131.
- [13] WANG T S, LI P F, ZHU Q M. Chinese Implicit Discourse Relation Recognition Based on Data Augmentation[J]. Journal of Computer Science, 2021, 48(10):85-90.
- [14] TIAN J, ZHU D J, LONG H. Chinese Short Text Multi-Classification Based on Word and Part-of-Speech Tagging Embedding [C]// Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, 2018;1-6.
- [15] KALARANI P, SELVA B S. Sentiment Analysis by POS and JointSentiment Topic Features Using SVM and ANN[J]. Soft-Computing, 2019, 23(16):7067-7079.
- [16] HEH Y, ZHENG J, ZHANG Z P. Text Sentiment Analysis Combined with Part of Speech Features and Convolutional Neural Network[J]. Computer Engineering, 2018, 44(11):209-214, 221.
- [17] ZHANG X H, LIU L X, DAI X Y, et al. CNN_BIGRU Text Classification Model Based On Part of Speech Features[J]. Computer Applications and Software, 2021, 38(11):155-161.
- [18] SHEN Y, LIN Z, HUANG C W, et al. Neural Language Modeling by Jointly Learning Syntax and Lexicon[C]// International Conference on Learning Representations. 2018.
- [19] SHEN Y, TAN S, SORDONI A, et al. Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks [C]// International Conference on Learning Representations. 2019.
- [20] PENG H, SCHWARTZ R, SMITH N A. PaLM: A Hybrid Parser and Language Model[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;3644-3651.
- [21] SUN K L, DENG D H, LI Y, et al. Inner-Attention Based Multi-Way Convolutional Neural Network for Relation Recognition in Chinese Compound Sentence[J]. Journal of Chinese information processing, 2020, 34(6):9-17, 26.
- [22] YANG J C, WANG Y Y, CAO Y, et al. Relation Classification of Non-Saturated Chinese Compound Sentence via Feature Fusion CNN[J]. Computer Systems and Applications, 2020, 29(6):224-229.
- [23] LI S, KONG F, ZHOU G D. Recognizing PDTB Style Implicit Discourse Relations [J]. Journal of Computer Science, 2016, 30(4):81-89.



YANG Jincai, born in 1976, doctor, professor, doctoral supervisor, is a member of China Computer Federation. His main research interests include advanced database and information system, Chinese information processing, artificial intelligence and

natural language processing.