计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(9):233−240 [doi: 10.15888/j.cnki.csa.008715]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BERT-FHAN 模型融合语句特征的汉语复句关系自动识别①

杨进才[1], 曹煜欣[1], 胡　泉[2], 蔡旭勋[2]

[1](华中师范大学 计算机学院, 武汉 430079)
[2](华中师范大学 人工智能教育学部, 武汉 430079)
通信作者: 杨进才, E-mail: jcyang@ccnu.edu.cn

摘　要: 复句关系是指复句分句之间的逻辑语义关系, 复句关系识别是对分句间语义关系的甄别, 是自然语言处理中的难点问题. 本文以有标复句为研究对象, 提出了一种 BERT-FHAN 模型, 该模型利用 BERT 模型获取词向量, 在 HAN 模型中融入关系词本体知识以及词性、句法依存关系、语义依存关系特征. 通过实验对提出的模型进行验证, BERT-FHAN 模型取得的最高宏平均 *F*1 值和准确率分别为 95.47% 与 96.97%, 表明了本文方法的有效性.

关键词: 复句关系识别; 词性; 句法依存; 语义依存; BERT 模型; HAN 模型

引用格式: 杨进才,曹煜欣,胡泉,蔡旭勋.基于 BERT-FHAN 模型融合语句特征的汉语复句关系自动识别.计算机系统应用,2022,31(9):233–240. http://www.c-s-a.org.cn/1003-3254/8715.html

Automatic Recognition of Chinese Compound Sentence Relation Based on BERT-FHAN Model and Sentence Features

YANG Jin-Cai[1], CAO Yu-Xin[1], HU Quan[2], CAI Xu-Xun[2]

[1](School of Computer Science, Central China Normal University, Wuhan 430079, China)
[2](Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China)

Abstract: The relations of compound sentences refer to the logical semantic relations between the clauses. The relation recognition of compound sentences is therefore the identification of semantic relations between clauses, and it is a difficult issue in natural language processing (NLP). Taking the marked compound sentences as the research object, this study proposes a BERT-FHAN model. In this model, the BERT model is employed to obtain word vectors, and the HAN model is used to integrate the ontology knowledge of relational words, as well as the characteristics of the part of speech, syntactic dependency relations, and semantic dependency relations. The proposed model is verified by experiments, and the result indicates that the highest macro average *F*1 value and accuracy of the BERT-FHAN model are 95.47% and 96.97%, respectively, which demonstrates the effectiveness of the method.

Key words: compound sentence relation recognition; part of speech; grammar dependency relation; semantic dependency relation; BERT model; HAN model

中文信息处理进程分为字处理、词处理、句处理、篇章处理 4 个阶段[1], 目前, 字处理和词处理方面的研究均取得了巨大的进展, 对句以及篇章的研究正在继续向前推进.

复句是由两个或两个以上的分句组成的句子[2], 汉语文本中复句占多数. 复句连接单句和篇章, 在篇章视野大范围内进行复句关系识别有助于加深对篇章句间语义关系的理解[3,4]. 因而可以广泛应用到机器翻译[5]、

① 基金项目: 国家社会科学基金 (19BYY092)
收稿时间: 2021-12-10; 修改时间: 2022-01-29; 采用时间: 2022-02-18; csa 在线出版时间: 2022-07-07

篇章分析[6]、自动问答[7]和信息抽取[8]等领域中.

复句的语义表达复杂, 复句的分类问题作为复句理论研究与应用研究的重要内容, 一直是汉语言学界关注的热点, 同时也是自然语言处理的难点. 目前, 在语言界最有影响的是《现代汉语》教材的分类和邢福义[2]在《汉语复句研究》(2001) 中提出的复句三分系统. 三分系统将复句划分为因果、转折、并列大类, 又将这 3 大类依次划分为因果、假设、推断、条件、目的, 并列、连贯、递进、选择, 转折、让步、假转 12 个二级类. 本文采用三分系统的 12 个二级类作为分类标准.

关系词 (关系标记) 用来连接复句的各个分句, 拥有关系词标志的复句被界定为有标复句[9]. 在有标复句中, 由于关系词的积极指向作用, 使得识别有标复句关系类别的准确率要高于无标复句[10,11]. 但在有标复句关系识别中存在如下困难: (1) 搭配使用的关系词部分缺省, 余下的关系词可对应多种类别; (2) 存在一部分跨类别的关系词.

例 1. 你不说, 我们<也>查得出你姓甚名谁! (吴强《红日》)

例 2. 条件不同, 面临的任务<也>不同. (《邓小平文选》)

在例 1 与例 2 中, 关系词均为“也”, 但对应的关系分别为让步与因果.

1 相关工作

从中文信息处理角度对复句类别自动识别的方法包括: 利用规则、结合规则和机器学习、利用深度学习的方法 3 类. 李艳翠等[12]以有标的清华汉语树库作为研究对象, 抽取显式和隐式的自动句法树的规则特征, 判定复句关系类别; 杨进才等[13]把已知的复句句法、关系词搭配等知识结合在一起, 以非充盈态二句式有标复句为研究对象, 鉴定复句所属关系类别; 杨进才等[14]探索复句字面及内部语法等特征, 并总结特征形成规则, 判断复句所属的关系类别. 随着深度学习方法研究的不断发展, 因其可以自动获取特征, 所以被应用在复句类别识别的研究中. 孙凯丽等[15]将 Bi-LSTM 模型学习到的句内注意力多路特征与 CNN 建模得到的复句局部特征结合, 使用 Inatt-MCNN 模型对复句进行因果、并列、转折 3 大关系类别识别. 孙凯丽等[16]使用 CNN 和 Bi-LSTM 相结合的 BCCNN 网络和词聚类算法来捕获单词间的相似特征, 从而辅助计算机识别复句的关系类别.

深度学习不需要人工操作, 能够自己独立研究复句语料中的特征. 然而, 在深度学习过程中融入已有的、显然的、人们主动选择的外部知识, 对模型而言依然具有吸引力[17]. Qin 等[18]以词性为句子特征, 联合词向量一起传输到 CNN 中, 来判断无标复句所属关系类别; 杨进才等[19]在 CNN 模型中融合关系词特征, 对非充盈态复句进行 3 类识别; 杨进才等[20]在 Transformer 网络中拼接关系词、词性的特征, 完成因果、假设、推断、条件、目的的因果类复句识别任务.

复句作为中文中出现频率最高的句子形态, 语言学界对复句的研究积累了丰富的知识. 前述的关系类别识别的利用规则、机器学习、深度学习的方法用到了这些语句特征, 本文探讨在深度学习模型中充分融入多种语句特征, 进行复句关系类别的识别.

2 复句文本特征表示

2.1 词向量表示

复句主要组成部分为词, 计算机将词处理成稠密的词向量表示[21], 词向量的表示效果影响着复句关系识别的准确率. 目前主流的词向量模型分为两类: 一类是以 Word2Vec[22]为代表的词向量模型, 它分为连续词袋模型 (CBOW) 和 Skip-gram 模型, 另一类是最新的 BERT 词向量模型[23].

Word2Vec 词向量模型利用输入的单词及其上下文信息, 在映射层中将信息整合, 由输出层输出对单词分析的结果. 但 Word2Vec 得到的词向量与对应的单词之间属于静态文本表示, 这种表示方式在解决一词多义问题上表现局限, 因而在某些任务中不能很好地动态优化. BERT 模型在中文处理方面有很大的优势, 它利用双向 Transformer 语言模型[24]进行预训练, 在不同单词间添加注意力机制将单词联合起来, 为解决中长期依赖问题提供了思路. 掩码语言模型和下一句预测是 BERT 的两种任务. 这两项任务使 BERT 不仅具备对目标句上下文进行预测的能力, 同时能够捕获句子间的深层语义关联. 因而, BERT 文本表示比 Word2Vec 表达的语义更丰富, 内容更全面. 将例 2 输入 BERT 预训练模型, 获得的输入表示如图 1 所示. 句子开头和结束分别使用 [CLS] 和 [SEP] 标志表示, 复句映射的向

量由词向量, 词在整个复句中的位置向量, 词在子句中的位置向量 3 部分组成.

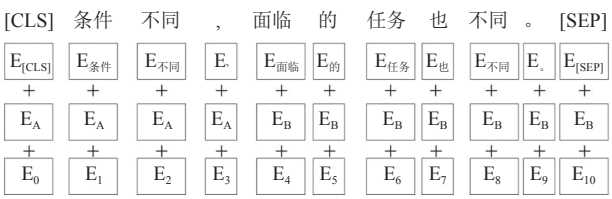

图 1 例 2 的 BERT 词向量组成

2.2 复句语句特征

2.2.1 候选关系词及类别特征

候选关系词 (准关系词) 是可以充当关系词的词, 当候选关系词在一条复句中能够将各分句联系起来时, 该候选关系词是关系词. 候选关系词的所属类别对识别复句的关系类别有着积极的指示作用. 因此, 候选关系词及其所属类别是至关重要的语句特征.

例 3. 法不仅有阶级性的一面, 而且有社会性的一面. (《人民日报》1981 年 01 月 27 日)

“不仅”“一面”“而且”“一面”, “一面”是复句存在的部分, 属于句中的方位名词, 它不是句子的关系词. 例 3 的两个分句由“不仅”“而且”连接, “不仅”“而且”属于递进关系类别. 例 3 是表示递进的复句.

2.2.2 词性特征

复句的各个单词均有与之相对应的词性, 词性反映了单词所具有的语法功能, 也约束了该词在复句中所充当的角色. 对例句 3 词性标注, 结果如图 2 所示.

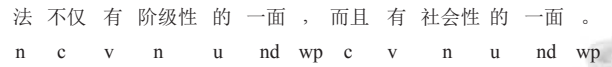

图 2 例 3 的词性标注图

例句 3 中“一面/一面”的词性都为 nd (方向名词), 它们属于复句的组成部分, 代表了句中的方位, 因此不是复句的关系词.

2.2.3 句法依存关系特征

法国语言学家特斯尼耶尔将句子中的词的关联构成句子的句法依存关系. 句法依存关系能够辨析复句中主、谓、宾、定、状、补的组成结构, 从复句的构成单元出发, 分析各个单元之间的相互关联, 加深对复句句法关系的理解. 5 条公理[25] 规定了句法依存关系, 复句中有且仅有一个独立核心成分, 其他单词都与支配词有句法依存关系. 复句核心词的句法依存特征为 HED (核心关系), 再依次抽取复句中其他单词与支配词之间的句法依存关系, 构成本文的句法依存关系特征. 将复句根据句法依存关系转换为相应的句法依存图, 例 3 分析结果如图 3 所示.

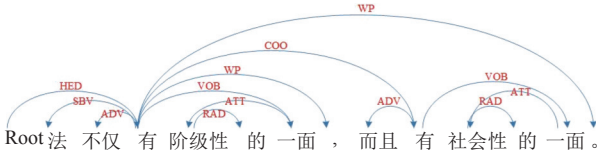

图 3 例 3 的句法依存分析图

候选关系词依次为“不仅”“一面”“而且”“一面”, 与它们的支配词之间的句法依存关系分别为“ADV (状中关系)”“VOB (动宾关系)”“ADV”“VOB”. 关系词与其支配词之间出现频率相对较高的句法依存关系是状中关系, 而 VOB 关系常见于动名词之间的句法依存连接. 从句法依存角度分析, 例 3 中的关系词为“不仅”“而且”, 它是一条表示递进关系的复句.

2.2.4 语义依存关系特征

语义依存和句法依存形似, 它们都是一种框架, 用以直观描述语言内部结构. 而不同的是, 语义依存采用单词的语义结构特征来阐述复句中单词彼此之间的关系, 它着重分析实词在句中的语义关联以及逻辑关联. 语义依存不会随着语句结构变化而变化, 它能够超越句子表层的句法结构, 更进一步得到句子的语义信息. 通过分析句子的语义依存关系, 能够明晰词汇在复句中所承担的语义角色. 语义依存关系中, 复句有且仅有一个核心词汇, 其他词汇与支配词间均有语义依存关系. 核心词的语义依存关系特征为 Root (根节点), 依次抽取复句中其他单词与其支配词之间的语义依存关系, 构成本文的语义依存关系特征. 将语义依存关系转换成相应的语义依存图, 例 3 分析结果如图 4 所示.

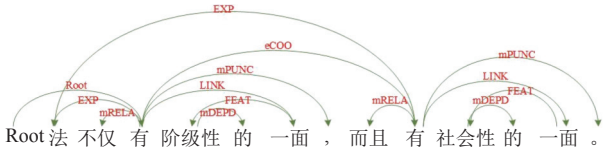

图 4 例 3 的语义依存分析图

由图 4 知, 候选关系词依次为“不仅”“一面”“而且”“一面”, 与它们的支配词之间的语义依存关系分别为“mRELA (关系标记)”“LINK (系事关系)”“mRELA”“LINK”. 关系标记是关系词与其支配词之间多见的语义依存关系, 而系事关系表示的是与事件相关联的客

体, 表示“一面”在例 3 中是方位名词, 不是句子的关系词. 因此, 例 3 句由“不仅”“而且”连接, 表示递进关系.

2.3 融合语句特征的文本表示

将单词的文本表示分别与候选关系词 (candidate relational words, CRW)、词性 (part of speech, POS)、句法依存关系 (grammar dependency relation, GDR)、语义依存关系 (semantic dependency relation, SDR) 排列组合得到的语句特征拼接作为复句的文本特征表示. 具体表示如下, 对于一个长为 L 的, 含有 n 个单词的句子, 第 i 个单词文本表示为 W_i (i=1,…, n), 其候选关系词特征为 CRW_i, 词性为 POS_i, 句法依存关系为 GDR_i, 语义依存关系记为 SDR_i.

候选关系词特征表示如式 (1) 所示, *flag* 为候选关系词是否为关系词的标志, *relation* 是候选关系词所属类别的关系矩阵; 融合上述多个特征的第 i 个单词的词向量 $VecW_i$ 表示如式 (2) 所示:

$$CRW = concat(flag, relation) \tag{1}$$

$$VecW_i = concat(W_i, CRW_i, POS_i, GDR_i, SDR_i) \tag{2}$$

在例 3 中, “不仅”的词向量为 [0.3145324], 候选关系词标志为 1, 候选关系词属于递进关系, *relation* 是递进关系的特征矩阵, 词性为“c”, 句法依存关系是“ADV”, 语义依存关系为“mRELA”. “不仅”融合多个特征的文本特征表示为 [0.3145324, 1, 递进, 5, 4, 7].

3 基于语句特征的 BERT-FHAN 模型

为了研究语句特征对复句关系识别的影响, 本文利用 BERT 预训练模型动态表示复句文本, 并在 HAN 神经网络中融入外部语言学知识, 得到 FHAN 模型. 进而构建 BERT-FHAN 模型, 该模型结构如图 5 所示.

3.1 词嵌入层

词嵌入层用机器能够识别的数字向量表示文本, 使用第 2.3 节提出的融合复句语句特征的文本表示方法, 获得每个单词的特征文本表示 $VecW_1$, $VecW_2$, …, $VecW_n$. 将它们依次输入单词注意力机制层, 帮助机器捕获语义知识.

3.2 单词注意力机制

使用双向 GRU 来获取单词的进一步表示, GRU 通过重置门、更新门模拟语言模型, 综合单词的上下文信息获取到每个单词的隐藏状态. 复句中第 i 个子句的第 j 个单词单词的前向隐藏状态$\overrightarrow{h_{ij}}$后向隐藏状态$\overleftarrow{h_{ij}}$, 其计算公式为:

$$\overrightarrow{h_{ij}} = \overrightarrow{GRU}(VecW_i) \tag{3}$$

$$\overleftarrow{h_{ij}} = \overleftarrow{GRU}(VecW_i) \tag{4}$$

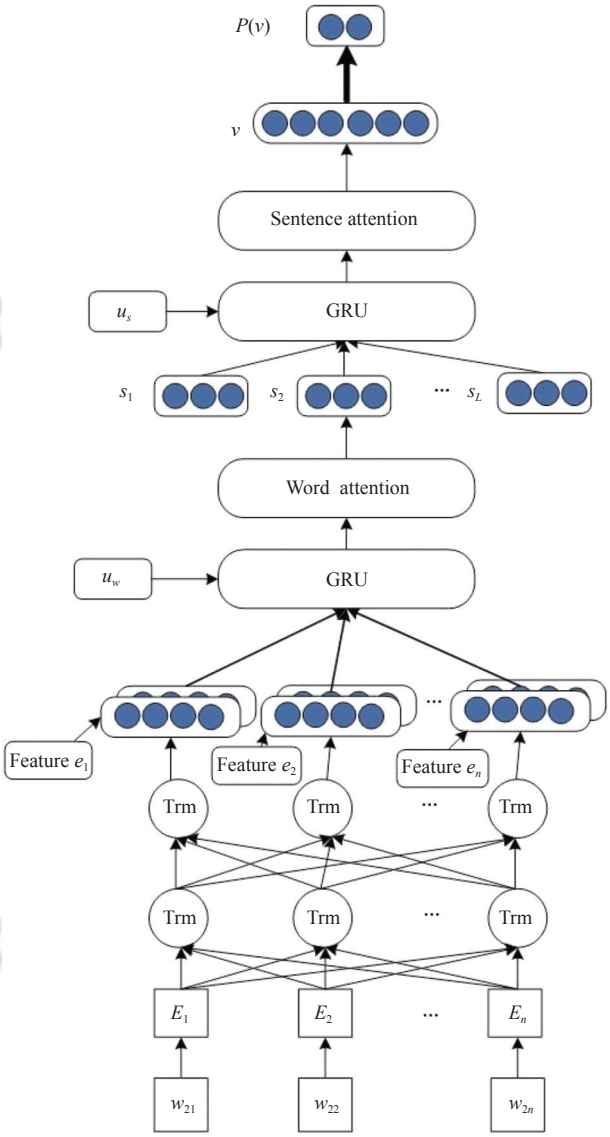

图 5 BERT-FHAN 模型结构图

通过单词的前后向隐藏状态得到单词的编码表示信息 h_{ij}:

$$h_{ij} = [\overrightarrow{h_{ij}}, \overleftarrow{h_{ij}}] \tag{5}$$

在复句关系识别过程中, 不是所有单词对任务都有影响, 因而引入注意力机制来提取对复句语义表示起作用的单词的隐藏表示 u_{ij}:

$$u_{ij} = \tanh(W_w h_{ij} + b_w) \tag{6}$$

计算与单词隐藏表示 u_{jt} 的相似性, 来判断单词的重要性, 通过 Softmax 得到单词的权重α_{ij}:

$$\alpha_{ij}=\frac{\exp(u_{ij}^{\mathrm{T}}u_w)}{\sum_j \exp(u_{ij}^{\mathrm{T}}u_w)} \tag{7}$$

sen_i 为 u_{ij} 和α_{ij}加权和, 它蕴含了分句 i 的信息:

$$sen_i=\sum_j \alpha_{ij}u_{ij} \tag{8}$$

3.3 句子注意力机制

句子级注意力机制和单词级注意力机制相似, 通过 GRU 获取复句中第 i 个子句的前、后隐藏状态$\overrightarrow{h_i}$、$\overleftarrow{h_i}$:

$$\overrightarrow{h_i}=\overrightarrow{GRU}(sen_i) \tag{9}$$

$$\overleftarrow{h_i}=\overleftarrow{GRU}(sen_i) \tag{10}$$

通过子句的前后向隐藏状态得到子句的编码表示信息 h_i:

$$h_i=[\overrightarrow{h_i},\overleftarrow{h_i}] \tag{11}$$

通过注意力机制来获取对复句语义表示起作用的子句隐藏表示信息 u_i, 将子句信息汇总得到复句的表示信息:

$$u_i=\tanh(W_s h_i+b_s) \tag{12}$$

通过计算与子句隐藏表示 u_i 的相似性, 来判断子句的重要性, 通过 Softmax 得到子句权重α_i, 最后获得复句的信息表示 com_sen:

$$\alpha_i=\frac{\exp(u_i^{\mathrm{T}}u_s)}{\sum_i \exp(u_i^{\mathrm{T}}u_s)} \tag{13}$$

$$com\_sen=\sum_i \alpha_i h_i \tag{14}$$

3.4 输出层

复句信息表示 com_sen 蕴含了复句 L 的所有信息, 通过 Softmax 激活函数得到复句类别的分类结果 $result$:

$$result=Softmax(W_c\cdot com\_sen+b_r) \tag{15}$$

4 实验与分析

4.1 数据集

汉语复句语料库 (the corpus of Chinese compound sentence, CCCS)[26] 是目前针对有标复句研究的规模最大的语料库, 它共收录了 65 万余条有标复句, 数据主要源自《人民日报》与《长江日报》. 我们在 CCCS 语料库中添加随机因子并排序后, 从中选择了 60 000 条复句, 构成一个新的用于标注关系类别的语料库, 简记为 CCCSRA (the corpus of Chinese compound sentence with relation annotation). 在 CCCSRA 语料库中, 各个类别的数据分布如表 1 所示. 将 CCCSRA 按照 14:3:3 的比例划分训练、测试、验证集.

表 1 CCCSRA 语料库数据分布表

关系类别	复句数目	比例 (%)
因果句类	13 334	22.22
假设句类	8 269	13.78
推断句类	663	1.11
条件句类	3 179	5.30
目的句类	1 507	2.61
并列句类	9 806	16.34
连贯句类	579	0.97
递进句类	6 962	11.60
选择句类	1 271	2.12
转折句类	3 391	5.65
让步句类	10 264	17.11
假转句类	775	1.29

4.2 实验参数

本文使用的是 BERT 预训练的 768 维词向量, 训练过程中为了使模型不产生过拟合的情况, 采用了 dropout 策略[27], 在神经网络中取舍. 同时, 实验借助 L2 正则项来提高模型的实际应用能力. 模型单词级、句子级注意力层的 GRU 的值设置为 300. 详细的参数如表 2 所示.

表 2 模型参数设置

参数	取值
词向量维度	768
batch_size	32
epoch	10
dropout	0.5
优化器	Adam
L2正则化系数	0.01
hidden_size	300
max_learning_rate	0.001
min_learning_rate	0.000 5

4.3 对比实验

为了验证 BERT-FHAN 模型的性能, 我们在 CCCSRA 数据集上设置了几个基线模型: ① TextCNN 模型[28], 通过卷积层网络来捕获句子的文本特征, 依靠固定的 filter 窗口抽取特征进行分类; ② 带自注意力机制的 Bi-LSTM 模型[29], 通过双向 LSTM 提取每个词语上下

文特征, 结合 attention 对每个词语加权求和, 使用 Softmax 激活函数进行输出; ③ Transformer 模型, 使用 encoder 模型, 通过位置编码获取单词相对位置信息, 使用 ReLU 激活函数进行输出; ④ Inatt-MCNN[30], 对复句语义编码, 添加注意力机制, 之后通过 CNN 获得局部特征信息, 通过 Softmax 得到输出结果; ⑤ HAN 模型[31], 通过多层注意力机制获得丰富的复句文本知识表示, 使用 Softmax 激活函数输出类别; ⑥ 结合 BERT 模型获取复句的动态词向量表示, 与 HAN 模型结合, 得到 BERT-HAN 模型.

4.4 实验结果分析

本文使用准确率 (Accuracy), 召回率 (Recall), 精确率 (Precision), *F*1 值作为评估标准. 从图 6 可以看出, 基线模型的准确率在 80% 以上, 这证明了实验中使用的深度学习模型均能够有效识别复句的关系类别. 其中, BERT-HAN 模型的性能优于其他模型, 4 项指标的值最高.

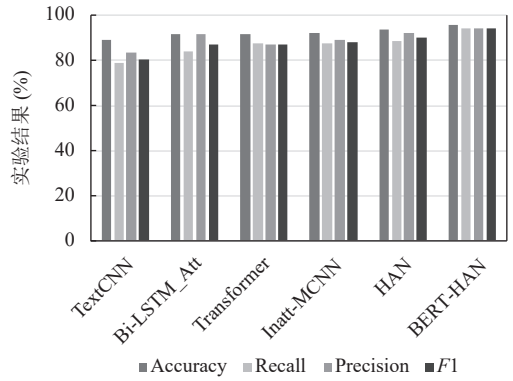

图 6 基线模型的实验结果

为了进一步研究句子特征和不同的文本表示对汉语复句关系识别的影响, 我们分别在 HAN 和 BERT-HAN 模型中融入 CRW、POS、GDR、SDR 排列组合的 15 种组合特征, 得到 FHAN 和 BERT-FHAN 模型. FHAN 和 BERT-FHAN 模型的实验准确率如图 7 所示. 在 BERT-FHAN 模型上的实验结果如表 3 所示.

从图 7 可知, 无论是 Word2Vec 还是 BERT 词向量表示方法, 融合不同语句特征, 模型训练后的准确率都在 90% 以上. BERT 文本表示方法的准确率在 Word2-Vec 基础上有所提升. 这是因为 BERT 模型得到的词向量是动态的, 可以随复句的上下文而变化, 提升了机器识别复句语义关系的能力.

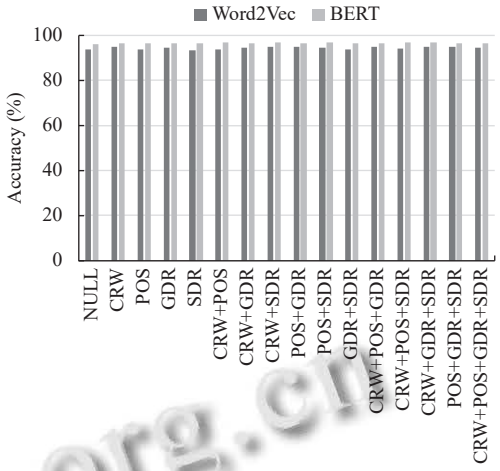

图 7 FHAN 和 BERT-FHAN 的实验准确率

表 3 BERT-FHAN 融合不同语句特征的实验结果 (%)

特征组合	准确率	召回率	精确率	*F*1值
NULL	95.97	94.12	94.43	94.23
CRW	96.42	94.67	95.31	94.94
POS	96.66	94.89	95.31	94.93
GDR	96.48	94.62	94.97	94.73
SDR	96.5	94.68	94.97	94.76
CRW+POS	**96.97**	**95.25**	**95.82**	**95.47**
CRW+GDR	96.66	94.99	95.18	95.05
CRW+SDR	96.74	95.11	94.85	94.96
POS+GDR	96.53	94.86	94.45	94.60
POS+SDR	96.71	95.02	95.41	95.16
GDR+SDR	96.59	94.63	95.32	94.89
CRW+POS+GDR	96.33	94.42	94.74	94.50
CRW+POS+SDR	96.88	95.12	95.46	95.21
CRW+GDR+SDR	96.71	95.14	95.50	95.25
POS+GDR+SDR	96.41	94.64	95.18	94.85
CRW+POS+GDR+SDR	96.37	94.72	94.67	94.66

从表 3 的实验结果可知, 融合了语句特征的实验结果与无语句特征的结果相比均有所提升. 当添加单个语句特征时, 实验准确度最高的是 POS (词性), 其次是 SDR (语义依存关系), 然后是 GDR (句法依存关系), 最后是 CRW (候选关系词). 在所有特征组合中, “CRW+POS”组合特征的效果最好, 其准确度可达 96.97%. 其次是融合“CRW+POS+SDR”组合特征, 它的准确率为 96.88%. 组合多个句子特征的模型通常比单句特征模型结果更好, 这是因为融合单个特征时, 复句的复杂语义会在现有单个特征基础上产生一些歧义特征. 而多个语句特征的组合可以有效地消除这些歧义, 从而更好地发现复句中的语义关联, 准确地识别复句之间的关系.

“CRW+POS+GDR+SDR”组合特征融合了所有语句特征, 但它的准确率与最佳组合特征相比降低了 0.6%. 这是因为不同语句特征相互干扰, 阻碍特征独立表达,

影响模型自动获取复句的内部特征; 除此之外, BERT 动态文本表示与外部特征以及模型都能捕获句子的语义关联, 三者内部共同作用, 结果却适得其反. 在 BERT-FHAN 模型的实验中, 以上所有特征组合的 $F1$ 值的变化幅度很小, 表明此方法稳定性较强, 抛开少数的误差情况, BERT-FHAN 模型能够正确判别复句关系.

在复句关系识别中, 一些关系词对应关系多种类别会给复句关系识别带来一定的困难. 经统计, CCCSRA 语料库中有 10 722 条复句的关系词对应多种关系类别, 占比为 17%. 由实验结果知, 融合不同语句特征的 BERT-FHAN 模型准确率均在 95% 以上, 证实了本文模型对含有一对多关系词的复句进行类别识别的有效性. 为进一步验证模型在复句类别识别中的两个困难问题上的适用性, 在融合“CRW+POS”特征的模型上, 统计出测试集中含有跨类别关系词的复句总数及模型识别的正确率, 结果如表 4 所示. 模型在这些含有跨复句类别的关系词的复句上, 识别的正确率均超过 87.5%, 统计结果表明文方法可适用在含有一对多关系词的复句关系识别任务上. 同时将测试集部分结果输出, 如表 5 所示. 语料中用缩写标注复句的关系类别, 模型将测试集结果输出, 数值按照三分系统二级类的顺序依次输出, 0 表示因果句, 1 表示假设句, 2 表示推断句, 依次类推. 由表 4 输出结果可知, 本文使用的模型能够正确输出测试集中含有这些关系词的有标复句的关系类别, 证明了本文方法的有效性.

表 4 测试集中跨类别关系词统计情况

标记名	总数 (条)	正确数量 (条)	正确率 (%)
一…就…	8	7	87.5
万一…(就/也)…	3	3	100
从而	57	55	96.49
…就/又/也/才/还…	1 235	1 207	97.73
于是	24	23	95.83
而/那么	182	177	97.25

表 5 测试集部分一对多关系词结果

类型	复句	描述	类别	结果
搭配使用的关系词部分缺省	正是国防落后, 近代中国才不断受到列强的欺凌.	关系词为“也”, “由于”缺省	yg	0
	这样, 才不会辜负那些真心球迷.	关系词为“也”, “只有”缺省	tj	3
	火势一直蔓延到家门口, 报警及时, 才免于一场火灾.	关系词为“也”, “幸亏”缺省	yg	0
	以色列先从黎巴嫩撤军, 整个气氛才会有变化.	关系词为“也”, “只有”缺省	tj	3
	你说, 小时候在乡下住过, 那么, 你所认识的植物一定比我多了.	关系词为“那么”, “既然”缺省	md	4
	对一个词的词义系统掌握的越全面、越完整, 那么, 对这个词的理解也就越深刻, 运用得也就越准确, 越灵活.	关系词为“那么”, “只有”缺省	tj	3
关系词跨复句多种类别	信一投进邮箱, 我就追悔莫及.	关系词“一…就…”一对多	lg	6
	往年, 在人民广场一开群众大会, 公安局就要宣布断绝交通.	关系词“一…就…”一对多	tj	3
	一说到行动风险, 他就立刻停止了公司的收购计划.	关系词“一…就…”一对多	yg	0
	过了那林, 船便弯进了灵港, 于是赵庄便真在眼前了.	关系词“于是”一对多	lg	6
	她浑身一震, 又紧闭上嘴, 于是, 唇边的深细皱纹, 又显现出来.	关系词“于是”一对多	yg	0
	雷磊第一个交了卷, 就匆匆忙忙地走了.	关系词“就/又/也/才/还”一对多	lg	6
	人活着, 就有希望.	关系词“就/又/也/才/还”一对多	tj	3

5 总结

本文提出 BERT-FHAN 模型, 进行复句关系类别识别. 实验结果表明, BERT-FHAN 模型在复句关系识别任务上相对于多个深度学习模型, 表现较好. 融入 15 种不同语句特征组合时, 实验结果较基线模型均有所提升, 其中, 融合候选关系词、词性语句特征得到的准确率最高. 充分证明了方法的有效性与适用性. 同时, 发掘出对关系类别有显著影响的语句特征, 弥补了深度学习模型对特征利用的不可解释的不足.

在今后的工作中, 我们将进一步挖掘复句语句特征, 研究在深度学习模型中更有效利用语言学研究的成果. 目前, 无标复句关系识别的正确率还很低, 我们将探索借助有标复句关系识别来进行无标复句关系识别的方法.

参考文献

1 严为绒, 徐扬, 朱珊珊, 等. 篇章关系分析研究综述. 中文信息学报, 2016, 30(4): 1–11.

2 邢福义. 汉语复句研究. 北京: 商务印书馆, 2001: 1–20.

3 孔芳, 王红玲, 周国栋. 汉语篇章理解研究综述. 软件学报, 2019, 30(7): 2052–2072. [doi: 10.13328/j.cnki.jos.005834]

4 胡超文, 杨亚连, 邬昌兴. 基于深度学习的隐式篇章关系识别综述. 计算机科学, 2020, 47(4): 157–163. [doi: 10.11896/jsjkx.190300115]

5 Mehta S, Ghazvininejad M, Iyer S, *et al*. DeLighT: Deep and light-weight transformer. arXiv: 2008.00623, 2021.

6 Liu X, Ou JF, Song YQ, *et al*. On the importance of word and sentence representation learning in implicit discourse relation classification. Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: IJCAI.org, 2020. 3830–3836.

7 Joshi M, Chen DQ, Liu YH, *et al*. SpanBERT: Improving pre-training by representing and predicting spans. arXiv: 1907.10529, 2020.

8 Bi KP, Jha R, W. Croft B, *et al*. AREDSUM: Adaptive redundancy-aware iterative sentence ranking for extractive document summarization. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Online: ACL, 2021. 281–291.

9 吴锋文. 基于关系标记的汉语复句分类研究. 汉语学报, 2011, 3: 63–73.

10 Huang HH, Chang TW, Chen HY, *et al*. Interpretation of Chinese discourse connectives for explicit discourse relation recognition. Proceedings of the 25th International Conference on Computational Linguistics. Dublin: ACL, 2014. 632–643.

11 Kishimoto Y, Murawaki Y, Kurohashi S. A knowledge-augmented neural network model for implicit discourse relation classification. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: ACL, 2018. 584–595.

12 李艳翠, 孙静, 周国栋, 等. 基于清华汉语树库的复句关系词识别与分类研究. 北京大学学报 (自然科学版), 2014, 50(1): 118–124.

13 杨进才, 陈忠忠, 沈显君, 等. 二句式非充盈态有标复句关系类别的自动标志. 计算机应用研究, 2017, 34(10): 2950–2953. [doi: 10.3969/j.issn.1001-3695.2017.10.016]

14 杨进才, 胡巧玲, 胡泉. 基于规则的有标复句关系的自动识别. 计算机科学, 2021, 48(S2): 124–129.

15 孙凯丽, 邓沌华, 李源, 等. 基于句内注意力机制多路 CNN 的汉语复句关系识别方法. 中文信息学报, 2020, 34(6): 9–17, 26. [doi: 10.3969/j.issn.1003-0077.2020.06.003]

16 孙凯丽, 李源, 邓沌华, 等. 基于词聚类 CNN 和 Bi-LSTM 的汉语复句关系识别方法. 计算机与数字工程, 2021, 49(8): 1588–1593. [doi: 10.3969/j.issn.1672-9722.2021.08.017]

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

18 Qin LH, Zhang ZS, Zhao H. Shallow discourse parsing using convolutional neural network. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning: Shared Task. Berlin: ACL, 2016. 70–77.

19 杨进才, 汪燕燕, 曹元, 等. 关系词非充盈态复句的特征融合 CNN 关系识别方法. 计算机系统应用, 2020, 29(6): 224–229. [doi: 10.15888/j.cnki.csa.007369]

20 杨进才, 曹元, 胡泉, 等. 基于 Transformer 模型与关系词特征的汉语因果类复句关系自动识别. 计算机科学, 2021, 48(S1): 295–298, 305.

21 赵京胜, 宋梦雪, 高祥, 等. 自然语言处理中的文本表示研究. 软件学报, 2022, 33(1): 102–128. [doi: 10.13328/j.cnki.jos.006304]

22 Goldberg Y, Levy O. Word2Vec explained: Deriving Mikolov *et al*.'s negative-sampling word-embedding method. arXiv: 1402.3722, 2014.

23 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.

24 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

25 刘海涛. 依存语法的理论与实践. 北京: 科学出版社, 2009: 23.

26 华中师范大学语言教育研究中心. 汉语复句语料库. http://linguist.ccnu.edu.cn/jiansuo/TestFuju.jsp. [2021-12-01].

27 Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.

28 Kim Y. Convolutional neural networks for sentence classification. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 1746–1751.

29 凡子威, 张民, 李正华. 基于 BiLSTM 并结合自注意力机制和句法信息的隐式篇章关系分类. 计算机科学, 2019, 46(5): 214–220. [doi: 10.11896/j.issn.1002-137X.2019.05.033]

30 Sun KL, Li Y, Deng DH, *et al*. Multi-channel CNN based inner-attention for compound sentence relation classification. IEEE Access, 2019, 7: 141801–141809. [doi: 10.1109/ACCESS.2019.2943545]

31 Yang ZC, Yang DY, Dyer C, *et al*. Hierarchical attention networks for document classification. Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 1480–1489.

(校对责编: 孙君艳)