ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 3,May 2021,pp. 605－620
DOI:10. 16337/j. 1004-9037. 2021. 03. 018

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

面向中文关系抽取的句子结构获取方法

杨卫哲[1,2], 秦永彬[1,2], 黄瑞章[1,2], 王　凯[1,2], 程华龄[1,2], 唐瑞雪[1,2], 程欣宇[3,4], 陈艳平[1,2]

(1. 贵州大学省部共建公共大数据国家重点实验室,贵阳 550025;2. 贵州大学计算机科学与技术学院,贵阳 550025;3. 贵州省智能医学影像分析与精准诊断重点实验室,贵阳 550025;4. 贵州省智能人机交互工程技术研究中心,贵阳 550025)

摘　要:在关系抽取中,神经网络模型是目前最常用的技术之一,然而现有神经网络模型很少考虑句子中两个实体之间的结构特征。该文针对关系抽取任务的特点,提出了基于神经网络模型的句子结构获取方法。该方法通过对关系实例中两个实体的位置进行特殊标记,使神经网络模型能够有效捕获句子中关于实体的结构信息。为了验证方法的有效性,分别采用两种主流的神经网络模型进行对比实验,实验结果表明,该方法在ACE 2005中文关系抽取数据集上的抽取性能得到显著提升,超出对比工作约11个百分点,表明该方法能有效提升关系抽取任务的性能。

关键词:关系抽取;结构特征;自然语言处理;实体标记

中图分类号:TP391　　**文献标志码**:A

Sentence Structure Acquisition Method for Chinese Relation Extraction

YANG Weizhe[1,2], QIN Yongbin[1,2], HUANG Ruizhang[1,2], WANG Kai[1,2], CHENG Hualing[1,2], TANG Ruixue[1,2], CHENG Xinyu[3,4], CHEN Yanping[1,2]

(1. State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China; 2. College of Computer Science and Technology, Guizhou University, Guiyang 550025, China; 3. Key Laboratory of Intelligent Medical Image Analysis and Precision Diagnosis of Guizhou Province, Guiyang 550025, China; 4. Guizhou Intelligent Human-Computer Interaction Engineering Technology Research Center, Guiyang 550025, China)

Abstract: Neural network model is one of the most commonly used techniques in relation extraction. However, the existing neural network models seldom consider the structural features between two entities in a sentence. Based on the characteristics of relation extraction task, this paper proposes a sentence structure acquisition method on neural network model. In this method, the positions of two entities in relation instance are marked so that the neural network model can effectively capturethe structural information about the entities in sentences. In order to verify the effectiveness of the proposed method, two mainstream neural network models are used for comparative experiments. Experiments show that the performance is improved significantly on ACE 2005 Chinese corpus. The result has

基金项目:国家自然科学基金通用联合基金重点(U1836205)资助项目;国家自然科学基金重大研究计划(91746116)资助项目;国家自然科学基金(62066007,62066008)资助项目;贵州省科技重大专项计划(黔科合重大专项字[2017]3002)资助项目;贵州省科学技术基金重点(黔科合基础[2020]1Z055)资助项目;贵州省教育厅青年科技人才成长项目(黔教合KY字[2017]137)资助项目;贵州省科技计划项目(黔科合基础[2018]1082)资助项目。
收稿日期:2020-01-16;**修订日期**:2020-12-25

exceeded the comparison work by approximately 11 percentage points. That proves that this method can significantly improve the performance of relation extraction.

Key words: relation extraction; structural features; natural language processing; entity marking

引　　言

关系抽取(Relation extraction,RE)是自然语言处理(Natural language processing,NLP)领域中的一项基础任务,该任务旨在识别并抽取文本中已知实体之间的语义关系,在中英文等多种语言中已有广泛研究。例如:对于句子实例“任正非是华为公司总裁”,关系抽取系统能自动识别出实体“任正非”和“华为公司”之间的语义关系是“雇佣”。关系抽取的研究成果在信息抽取[1]、问答系统[2-3]、智能问答[4]和知识图谱构建[5]等任务中都有广泛应用,然而关系抽取任务的研究主要基于句子级别,实例文本中包含的文字数量通常较少,用仅有的文本信息难以获取足够的特征来支撑抽取实体之间的语义关系,造成了严重的特征稀疏问题。如何利用有限的文本信息获取句子结构特征以解决特征稀疏问题,从而支撑关系抽取任务是本文研究的重点。

首先,从语言层面来看,中西方思维方式存在差异,语言产生的背景环境不同等特点使得中英文的语言结构不尽相同。语言有分析型和综合型之分,分析型语言语序固定,综合型语言语序灵活。英语是综合型语言,在语言结构上较重视形式分析和逻辑推理,语法严格,从句形式较多,句子一般较长;中文属于分析型语言,语序较为固定,没有曲折变化,其词语组合成句依靠语序和虚词,短句较为常见[6]。英语等印欧语系语言的基本结构单位是词,汉语的基本结构单位是字,所以在中文的处理中,文字组合的结构要求更为严格,如何将字组合成符合文本语境和语义的词非常关键。例如,对于句子实例“南京市长江大桥位于南京市鼓楼区”,抽取实体关系时需要使关系抽取系统能够对文本中实体对进行感知,得到完整且有效的实体表示,即组合成实体词的文字应尽量正确搭配,避免产生歧义和错义。然而,中文基本结构单位的不同组合易使结果产生歧义和错义,因此对文字组合的结构性有较高要求。如上句实例,基本结构单位组合的实体结果可能如下:“南京市”、“南京市长”,“江大桥”,“长江大桥”……。关系抽取系统若无法正确地感知实例中实体对的位置及内容,就会对关系抽取结果产生影响。根据人类知识库,句子实例中实体“南京市”与“长江大桥”之间存在一种地点之间的包含关系。若“江大桥”的文字组合结果被关系抽取系统认为其具有人名类实体的相关属性信息,而“南京市鼓楼区”被关系抽取系统判定为具有地点类实体的相关属性信息,则系统会抽取到两实体之间存在一种“人名-地点”的关系。可见,不同的句子结构能够使系统的识别结果发生变化。相反,将上述实例句用英文表达为“Nanjing Yangtze River Bridge is located in Gulou District, Nanjing”,由于其语法严格,关系抽取系统会根据神经网络学习到的词向量表示,将“Nanjing”、“Yangtze River Bridge”、“Gulou District”等基本结构单位的组合识别为具有实体属性信息的词组,不易产生类似于中文中“南京市”、“南京市长”这样会造成实体关系识别产生差异的组合结果。上述对比可知基本结构单位的不同组合使得文本实例有不同的句子结构,从而影响实体关系识别的结果。

此外,实体类型信息也会影响实体对关系的识别。相同的基本结构单位组合得到的实体在不同语境下可能有着不同的实体类型,在关系识别的时候就能体现出实体类型对结果的影响。例如:实例1,“中华人民共和国的开国大典在北京隆重举行”;实例2,“北京当局宣布中国政府会永远站在中国人民身边”。二者都包含由基本结构单位“北”和“京”组成的实体“北京”。但是实例1中的“北京”是一个带有地点属性类型的实体,实例2中的“北京”是一个带有组织属性类型的实体。显然,如果没有实体类型说明,关系抽取系统会将由相同基本结构单位组成但在不同语境下代表了不同含义的实体认为是同一

实体,这是不合适的。因此,实体类型属性是实体信息中很重要的特征之一,能使神经网络有效获取句子结构特征。

最后,关系抽取中实体间的结构特征也是影响抽取结果的因素之一,这里用实体间的相对位置关系来表示实体对结构特征,即实体1相对于实体2是处于其之前还是之后的位置描述。例如:句子实例3"人类是能够制造并使用工具的高级生物",此句中存在实体对"人类"和"工具"。理想情况下,关系抽取系统应能够识别出实体"人类"与"工具"之间存在"制造"的关系。但相反,对于实体"工具"与"人类"之间却不能存在"制造"类实体关系。可知,实体的相对位置信息对于获取以实体对为中心的句子结构特征是有效的。

本文的主要贡献是基于句子结构特征和语义信息对实体关系抽取有重大影响的前提,提出了一种面向关系抽取的句子结构特征获取方法。通过使用实体类型、实体相对位置和实体边界构造标记符及分隔符来标记和分隔实体的方式突出文本结构,获取句子的结构特征,增强神经网络对实体关系学习的能力。利用循环神经网络(Recurrent neural networks,RNN)的原理,结合Attention机制[7],在基于变换器的双向编码器表征技术(Bidirectional encoder representations from transformers, BERT)预训练语言模型[8]以及卷积神经网络(Convolutional neural networks ,CNN)模型上都能使神经网络对实例文本的实体对有更深刻的语义认识,达到提高中文关系抽取性能的目的。实验数据显示,实体类型及相对位置信息标记的确能够获取以实体对为中心的句子结构特征,取得更好的中文关系抽取性能,在ACE 2005中文关系抽取数据集上使得关系抽取性能F_1提升9%~11%。

1 相关工作

关系抽取的研究通常包括基于规则、有监督学习、半监督学习和无监督学习等方法。早期的关系抽取研究方法主要基于语法规则,通过分析句子的语法结构,找出尽可能多的在指定语法规则中出现的实体对,将其作为实体关系发生的依据。但是该方法需要人工制定较为严谨的规则,制定方法较为复杂,规则制定依赖较强的领域语言文学专业知识和语法知识,领域性强,普适性低。

在关系抽取中应用的传统机器学习方法依赖于人工选择的大量特征,对研究者的领域知识有较高要求。而深度学习技术能够通过大量的训练数据自动获取分类特征,因此成为关系抽取研究领域的重要技术,其目的在于利用计算机建立模仿人脑分析学习机制的人工神经网络。构建人工神经网络是深度学习研究中的重要部分,以其独特的结构组成和数据处理方式应用于许多领域并取得了显著成效。在中文自然语言处理中,Liu等[9]首次使用CNN模型在ACE 2005数据集[10]上研究关系抽取任务,提出了用同义词编码方式表示具有相同语义的单词,取得良好效果,但该模型没有池化层,受噪声影响比较明显。此外,基于特征工程或核函数的传统关系抽取方法也取得了很大进步[11-12],但这类方法需要人工设计较为复杂且优良的特征,可移植性差。除了在公共领域研究实体关系,以达到普遍支持下游任务的目的以外,在特定领域的实体关系抽取研究也十分有必要。文献[13]研究了司法领域的数据融合与分析应用,其中关系抽取的研究可在司法领域知识图谱构建中有效应用。

在关系抽取任务中,实体对将句子分为5个部分,是一种独特的句子结构。在该任务的研究过程中,曾有许多研究者考虑过使用句子结构特征。王长有等[14]提出了一种基于句子结构特征的领域术语上下位关系获取方法,该方法通过分析句法结构,融合句子结构特征进行关系抽取。Chen等[15]提出了一种多通道深度神经网络的方法,利用句子结构特征获得同一词语的不同表示。Socher等[16]提出了一种使用矩阵-递归神经网络模型(Matrix-vector recursive neural networks, MV-RNN)的方法来

进行关系抽取。该方法在关系抽取任务中考虑了句子的句法信息和语法结构信息，但是未能将以实体对为中心的句子结构信息表达利用起来。

在捕获句子结构特征的研究中，最简单的是Damashek[17]提出的n-gram特征提取方法。通常将n设置为1，2，或3，当n变大时，可能会捕捉到噪声特征，导致关系抽取性能降低。n-gram特征可以与语义或语法特征（如：潜在主题、位置特征）结合来生成组合特征[18-19]，使得特征分布改变，得到偏斜分布，有助于改善实验性能。刘娜娜等[20]基于中文短语成分结构对关系抽取有重大影响的理论，提出使用短语成分分析模型MPARSER和关系抽取模型PCNNATT获取短语结构提高中文关系抽取的性能。

序列模型（隐马尔科夫模形（Hidden markov model，HMM），条件随机（Conditional random field，CRF），长短期记忆网（Long short-term memory，LSTM）等）常用于建模词语之间的依赖关系。给定一个句子，序列模型可以得到一个最大的标签序列以指定句中的语言单元，通过假定标签之间的高阶相关性，达到有效捕捉句子结构信息的目的。但是序列模型通常难以捕捉到具有长期依赖关系的文本相关性，即两个间隔很远的语言单元很难互相影响。有些自然语言处理任务需要得到的信息可能分散在整个较长的文档中，故很难用序列模型捕获到全局信息，而且序列模型也不适用于某些嵌套型信息抽取任务，如嵌套命名实体识别[21]。

解析树和依赖树是获取更细粒度句子结构的建模方法，从句法理论出发，提供了一种自然语言研究方法。基于解析树的方法广泛应用于句子级别的信息抽取任务[22-24]，该方法存在的问题主要是不正确的分块或解析，数据的零碎嘈杂异构等特点可能导致性能不佳。因此，相对于解析整个句子，在特定语言单元中解析局部依赖的上下文关系对关系抽取更为有效。在深度学习模型架构中，字的位置向量特征和解析树都可以获得句子的结构特征。关系抽取研究中常用的位置向量是根据实例文本中每个字与两实体的相对距离产生的，将其坐标转换为分布式表示，最后将位置向量与字向量直接拼接送入神经网络。Zeng等[25]使用卷积神经网络，首次提出了使用位置特征（Position features，PF）编码当前词与目标词对的相对距离，同时说明位置特征是比较有效的特征。

例如，在ACE 2005中文数据集中有这样一个句子："南韩总统金大中以及高层人士"，句中两个实体分别是"南韩"，"金大中"，称为目标词。在BERT中，每个字的位置向量是由该字在句中出现的位置下标决定的，如例句中的"南韩总统"4个字所对应的位置分别为"0"，"1"，"2"，"3"，以此类推。而关系抽取研究中，常用的位置特征获取方法是根据当前字与两标记实体的相对位置决定的。实例文本中除两实体以外的所有字相对于两个实体均存在两个距离值，规定实体左边取负值，在实体右边取正值，将这两个相对距离作为位置向量映射的自变量，位置向量是映射的因变量，即向量值。如例句中"南韩总统"所对应的位置分别为"(0，－4)"、"(0，－3)"、"(1，－2)"、"(2，－1)"，以此类推随后将其映射成一个低维度的向量$\boldsymbol{d}_1$和$\boldsymbol{d}_2$，然后从位置向量的查找表中将位置映射为向量表示，最终将两者串联得到最后的位置特征$\boldsymbol{PF}=[\boldsymbol{d}_1,\boldsymbol{d}_2]$。可通过文中实体词与非实体词之间的相对位置关系来获取句子的结构特征。

在使用位置向量获取句子结构特征的基础上，许多研究者提出了更多神经网络结构。Sahami等[23]使用卷积神经网络，直接拼接词向量和位置向量。Santos等[26]首先从词向量和位置向量中探究句子的表示形式，然后将每一个句子的表示与矩阵相乘去对每一个关系类别进行预测评分。Zeng等[27]提出了一个分段获取句子结构信息的方法，利用句子中的两个实体将句子分割，对每部分内容进行池化操作，获取更多特征。另外，注意力机制方法也多应用于位置向量获取信息[28]。2018年，Devlin等[8]提出的BERT采用表义能力更强的双向Transformer网络结构训练语

言模型。

BERT预训练语言模型直接训练一个位置向量用于保留字的位置信息,每个位置随机初始化一个向量,加入模型训练,然后得到一个包含位置信息的位置向量,最后BERT将位置向量和字向量直接拼接。本文在中文关系抽取的实验上引入了CNN神经网络模型和基于Transformer的BERT预训练语言模型,使用实体类型标记的方法获取句子的结构特征。实验显示本文方法有效提高了中文关系抽取的性能,在ACE 2005中文数据集上取得了0.898的F_1值。

2 句子结构获取方法

本文所采用的数据集是ACE 2005中文数据集[10],其数据以分层结构存储于文件中。对于分层结构的数据,最自然的表示方法是使用树表示整篇文章,树中的节点表示实体、关系等元素,故采用树形结构解析的方法从文件中获取关系提及实例、实体内容、实体类型和关系类型等信息。

首先从大量非结构化文本中提取出带有两个或两个以上实体的句子,然后将整篇存储了句子实例文本、实体内容和实体中心词等诸多元素信息的文档进行树形解析,提取出数据集中标注了实体关系类型的句子作为正例使用。与整个文档的交互(读取和写入文件)是在树及元素节点级别上完成的,使用树形结构解析ACE 2005数据集中的文档,从中提取出句子实例文本、实体内容、实体类型和关系类型等信息。在取得整篇文章之后使用逗号、句号、冒号、分号、问号5种标点符号将其分割为句子,将包含了实体对且未在正例中出现的实例文本作为负例使用。

获取到实例文本之后,将句子中多余的空格、换行等在实验中认为无意义的字符予以割除。最后按照句中实体对出现的相对位置不同,将其分为“嵌套”和“独立”两种类型的相对位置关系。“嵌套”关系包括了“实体1出现在实体2内部”“实体2出现在实体1内部”两种位置关系;“独立”关系包括了“实体1出现在实体2之前”“实体2出现在实体1之前”两种位置关系。最终从数据集中整理得到数据为形如“实体1 实体2 实体1类型 实体2类型 实体提及句子 关系类型”的实例集合。据统计,在ACE 2005中文数据集的关系抽取实例中,包含“嵌套”实体对的实例文本占总数据量的29.80%,这说明了对能够凸显实体对的句子结构信息抽取方法的研究很有意义。

关系抽取实例文本中包含的实体个数并不固定,舍弃文本中少于两个实体的句子,当句子中实体个数大于等于2时,按照其实体的不同顺序两两组合,该实体对间可能存在某种语义关系,或者不存在。在关系抽取模型中,可由实体存在的上下文语境,结合多种特征识别该实体对在当前语境中蕴含的实体关系类型。可用特征包括位置、词性等。在关系抽取研究中,句子级别和文档级别的研究是当前较为主流的两个分支,句子级别的关系抽取研究更加广泛,但是句子实例中包含的词量一般较少,可利用的原子特征不足,特征稀疏问题十分严重。因此,怎样利用神经网络获取到句子中以实体对为中心的句子结构特征是研究目的,借此可以解决句子级关系抽取中特征稀疏的问题。关系抽取实例文本中包含两个或两个以上已识别出的实体,以其中要识别其关系类型的两实体边界为分隔,将文本句子分为“左、实体1、中、实体2、右”,共5个部分,用实体类型标记符对两实体的开始和结束进行标记。

对于实体类型标记及分隔方法,设置了如图1(b~f)所示的5种不同的实体标记构成方法。实验中,除了使用实体类型标记符标记实体的开始和结束边界,用于获取句子中以实体对为中心的结构特征。还使用了实体对复用的方法,将实例文本中的实体对前置至句首处,并用字符“0”对复用实体对进行分隔,加强实体语义信息的表达能力,使得神经网络模型获取以实体对为中心的句子结构特征。

(1) 原对照文本。图1(a)所示文本处理方法为原始文本输入,不对句子做任何处理,保留句子中实

体及除标点符号以外的其他全部内容作为对照实验。

(2) 实体对复用。图1(b)所示的文本处理方法是将句子中的实体对前置至句首,起到实体对语义增强的作用。

(3) 实体对标记。图1(c)所示的文本处理方法为将句子中被识别其关系的实体对分别使用由实体类型和实体相对位置构成的实体标记符作标记处理。这种处理方法是本研究中获取句子结构特征的主要方法。

(4) 实体对标记与复用结合。图1(d~f)所示的方法是将(2,3)中的实体对标记与复用相结合的组合方法,同时利用实体对增强的语义信息和实体对标记的方法获取句子的结构特征。

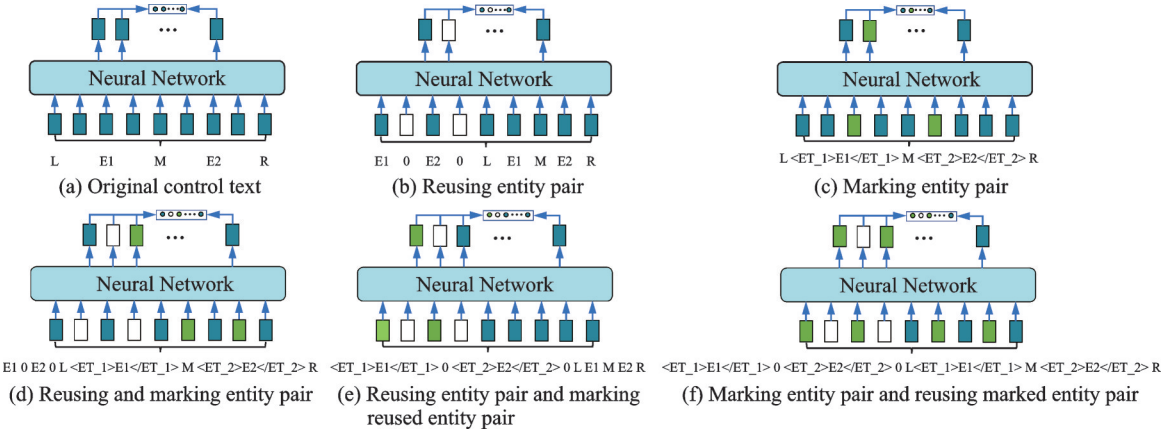

图1 实体标记及分隔方法

Fig.1 Method of entity marking and separation

图1中的符号“L”“E1”“M”“E2”“R”“<ET_1>”“</ET_1>”“<ET_2>”“</ET_2>”分别代表的含义为,L:实例文本中实体1左边部分的文本;E1:实体1;M:实例文本中实体1与实体2之间部分的文本;E2:实体2;R:实例文本中实体2右边部分的文本;“<ET_1>”:实体1的开始标记;“</ET_1>”:实体1的结束标记;“<ET_2>”:实体2标记;“</ET_2>”的结束标记。且图1中不同颜色的方块代表了不同的输入与输出,一一对应。蓝色方块表示一般文本内容的输入与输出,绿色方块表示实体标记符的输入与输出,白色方块表示实体分隔符的输入与输出。

在实体标记的处理过程中,当实体1与实体2的相对位置关系为“嵌套”时为较为特殊的实体相对位置关系。首先将处于内部的实体使用标记符标记,然后将外围的实体进行标记,此时即内部的实体标签已经被外围的实体标签包裹住。当出现两实体完全重合的情况,先后进行实体1、2的标记,以达到获取句子结构特征的目的。

以图1(c)所示的实体对标记方法为例:当实例文本满足一般格式,即句子中存在两个实体,且实体1左边的部分(L)存在,实体1与实体2之间的部分(M)不为空,实体2右边的部分(R)也存在,则原始语句S可以表示为

$$S=(s_1,s_2,\cdots,s_i,s_{i+1},\cdots,s_{i+k},s_{i+k+1},\cdots,s_j,s_{j+1},\cdots,s_{j+t},s_{j+t+1},\cdots,s_n) \tag{1}$$

式中:$s_{i+1},\cdots,s_{i+k}$和$s_{j+1},\cdots,s_{j+t}$分别表示实例文本中的两个实体;$s_1,s_2,\cdots,s_i$表示实例文本中实体1的左边文本;$s_{i+k+1},\cdots,s_j$表示实例文本中实体1与实体2之间的文本;$s_{j+t+1},\cdots,s_n$表示实例文本中实体2右边文本。将句子S用本节中提出的实体类型标记方法(图1(c))处理为

$$S\_=(s_1,s_2,\cdots,s_i,E\_T1_{\text{Begin}},s_{i+1},\cdots,s_{i+k},E\_T1_{\text{End}},s_{i+k+1},\cdots,s_j,E\_T2_{\text{Begin}},s_{j+1},\cdots,s_{j+t},E\_T2_{\text{End}},s_{j+t+1},\cdots,s_n) \tag{2}$$

式中:实体1的开始和结束标记分别用E_T1_{Begin}和E_T1_{End}表示,实体2的开始和结束标记分别用E_T2_{Begin}和E_T2_{End}表示,用于表示实体边界。本实验中使用的实体边界标记符是由实体类型及实体相对位置复合而成,作为实例文本中要抽取其关系的实体对,实体类型必会影响实体间语义关系,故使用实体类型对实例文本中的实体对进行标记。实例文本中两个实体的位置顺序会影响实例文本的结构组成,故使用阿拉伯数字"1"和"2"对两实体之间的相对位置关系进行标记说明。定义其标记规则为:以<实体1类型_1>和</实体1类型_1>作为实体1的开始和结束,以<实体2类型_2>和</实体2类型_2>作为实体2的开始和结束。实验所用数据集ACE 2005中包括了VEH、WEA、GPE、FAC、PER、LOC、ORG等7个类别的实体类型。

图1(b~f)所示的5种实体标记及分隔方式是相互并行的同级关系,是不同的句子结构获取方式,但结构获取方法的核心思想是相同的——利用实体类型及相对位置关系复合而成的实体标记符在神经网络中获取句子结构特征。图1(b,d~f)这种实体对复用的方法是为了避免当句子中的"L""M"或者"R"3部分中有为空的情况下,句子长度可能会受到影响,当句子长度变短时,句中的语义信息就会减少,造成特征损失,特征稀疏问题更加严重,影响关系抽取性能,故而采用此方法增强实体语义,获取以实体对为中心的句子结构特征。以实体类型及实体相对位置关系复合作为实体标记符,以"0"作为分隔符的方式为句中的实体对赋予无实际语义信息的边界标识符号,能够有效获取句子中以实体对为中心的句子结构特征,支持实体关系抽取任务。

3 模型结构

本部分重点介绍实体标签标记部分的研究。本文提出一种利用实体类型和实体相对位置等信息对关系抽取实例文本中的实体对进行标记和分隔、通过神经网络获取句子中以实体对为中心的句子结构特征的方法。本文认为句子结构特征能有效提高中文关系抽取任务的性能,故将这种获取句子结构特征的方法在CNN和BERT等神经网络模型上进行了实验和验证。实验结果表明,该方法可以有效提升中文关系抽取性能。

将实体标记方法与神经网络模型结合抽取句子结构特征,其过程如图2所示。图2描述了使用实体标记实例文本进行关系抽取研究的整个过程,将其分为"实例输入""实体标签标记""神经网络""结果输出"4个部分。

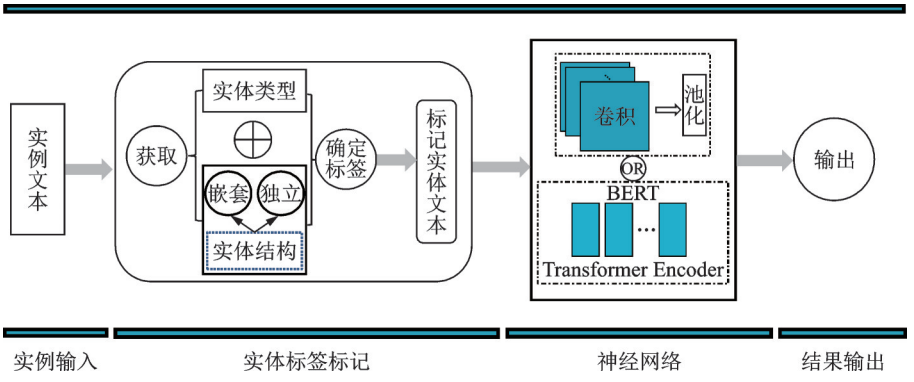

图2 模型总图

Fig.2 Model overview

实例输入:从数据集中获取的带有实体对的句子。

实体标签标记:根据句子中实体的类型及位置关系进行标签标记。

神经网络:将标签标记的文本输入神经网络,获取句子结构特征,进行关系抽取,在研究中分别使用了CNN和BERT两种神经网络模型。

结果输出:得到关系识别的最终结果。

3.1 实体标记下的CNN模型

CNN是一种包含卷积计算且具有深度结构的前馈神经网络,是深度学习代表算法之一,在自然语言处理和计算机视觉领域中均有广泛应用。CNN的结构组成包括了输入层、隐藏层和输出层。通常隐藏层中又包括了卷积层、池化层和全连接层等3种常见的神经网络构筑层,通过CNN可获取文本抽象特征。

本文采用的CNN网络模型结构由输入层、词嵌入层、卷积层、池化层、全连接层及输出层组成,模型结构如图3所示。其神经网络的各部分都各司其职,共同作用抽取抽象特征,用于关系抽取任务。该模型的重点和特别之处在于输入层对关系实例文本进行的处理,首先获得实体类型,然后根据文本中实体对间的位置关系等信息确定该实体对的实体标签类型,最后对实例文本中的实体对进行标记,利用神经网络获取句子结构特征。

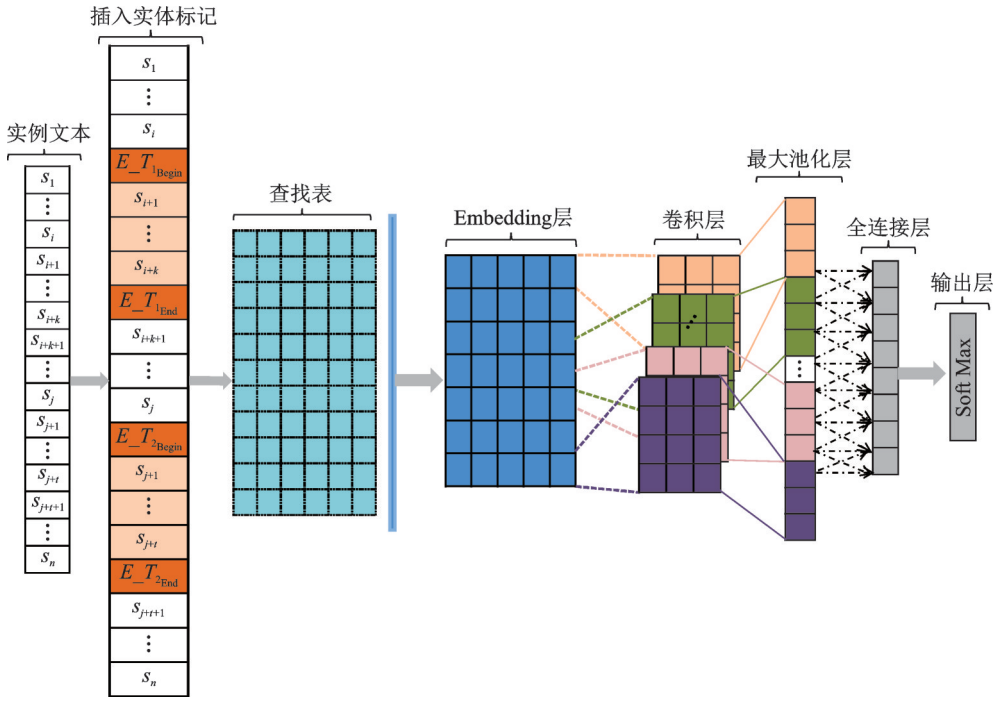

图3 实体标记的CNN模型结构图

Fig.3 Architecture of entity-marked CNN model

在CNN的输入层,使用式(1)中S代表包含实体对的实例文本,其中$s_i(i\in(1,n))$表示句子S中第i个字。通过第2章表述的实体标记方法对实体对进行标记,将实例S标记成为$S_$的形式,来获取句子结构特征。

字嵌入层中,通过字向量查找表 $\boldsymbol{W}_e$ 对实例文本 $S\_$ 中每一个字进行向量映射,得到字表示。以 $\boldsymbol{x}_i \in \mathbf{R}^H$ 作为 s_i 的字向量表示,H 代表向量的维度,该网络结构中将 H 设为300。如果字典的大小为 V,则 $\boldsymbol{W}_e \in \mathbf{R}^{H \times V}$ 为向量映射层的参数矩阵。由于CNN需要固定输入文本的长度,故对长于或短于固定长度的实例文本分别进行裁剪或扩充。假设实例文本预定义长度为 L,将句子 $S\_$ 映射为向量序列,表示为 $\boldsymbol{X}=[x_1, x_2, \cdots, x_L]$。将该过程形式化表示为

$$\boldsymbol{X} = \text{Embedding}(S) = [x_1, x_2, \cdots, x_L] \tag{3}$$

在神经网络的卷积层对向量矩阵 $\boldsymbol{X}$ 进行卷积操作,该模型在卷积层运用了15个7×7的卷积核,通过这样的方法获得文本的抽象化特征,对该序列进行的卷积操作形式化表示为

$$\boldsymbol{c}_i = f_c(\boldsymbol{W}_C^{\mathrm{T}} \cdot \boldsymbol{X}^{\mathrm{T}} + b) \tag{4}$$

上述过程中 $\boldsymbol{W}_c \in \mathbf{R}^{K \times H}$ 是卷积运算的滤波器,b 是一个偏置值,f_c 是一个非线性函数(如:双曲正切函数)。$\boldsymbol{c}_i$ 的维度为 H,当对 $[x_1, x_2, \cdots, x_L]$ 迭代卷积运算时,此步形式化表示为

$$c = \text{Conv}(x) \tag{5}$$

CNN中卷积运算能够有效捕捉句子局部特征,它将一个 K-gram 向量矩阵 $[x_i, x_{i+1}, \cdots, x_L]$ 映射到一个高阶特征表示中,能够学习到 K-gram 的语义或句法信息 $[w_i, w_{i+1}, \cdots, w_L]$。为了获取对关系抽取任务更有效的特征,在神经网络的池化层中,对每一个卷积结果 c 都实现最大池化操作,可将其形式化表示为

$$p = \text{Pooling}(c) = \text{Max}(c_1, c_2, \cdots, c_L) \tag{6}$$

在池化层之后,利用全连接层进行全局调节,Softmax层输出各类别上的概率值。该过程可形式化表示为

$$y = \text{Soft max}(\text{conn}(p)) \tag{7}$$

综上,在给出标记实例文本 $S\_$ 的前提下,利用传统的CNN进行关系抽取的整体过程可以表示为

$$y = \text{Soft max}(\text{Conn}(\text{Pool in}g(\text{Conv}(\text{Embedding}(S\_))))) \tag{8}$$

3.2 实体标记下的BERT语言模型

BERT是在大规模语料库上训练所得到的预训练语言模型。它会根据文本中字出现的位置不同,给每一个字赋予位置向量,根据其在句中出现的位置不同具有不同的语义信息结合字向量、文本向量和位置向量作为模型输入[8,29]。

近年来,研究人员使用深度神经网络在大量非特定领域的文本数据集上训练语言模型,得到在横向任务上效果较好的预训练语言模型,然后在预训练语言模型的基础上针对特定领域纵向训练。比较典型的语言模型是对于一个给定的文字字符串,从左到右计算下一个字出现的概率,如式(9)所示[30],S 表示特定排列的字串 $[w_1, w_2, \cdots, w_m]$,其中 $w_i (i \in (1, m))$ 表示字串中的第 i 个字,由贝叶斯公式可得字串 S 出现的概率表示公式为

$$p(S) = p(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} p(w_i | w_1, w_2, \cdots, w_{i-1}) \tag{9}$$

本文中的方法主要是通过实体类型及实体相对位置关系等信息作为实体的边界标记,用分隔符将实体对分隔处理,最后利用神经网络获取句子的结构信息。相对于不添加实体标记符的文本实验,该方法在CNN及BERT语言模型上均有较好的表现,使得关系抽取性能得到大幅度提升。BERT模型结构如图4所示。

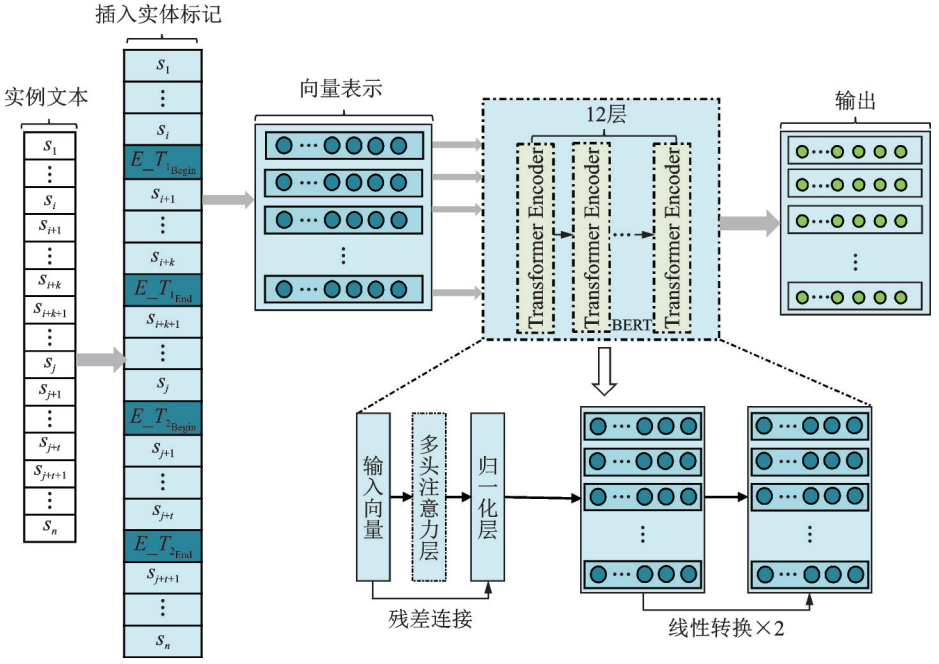

图4 实体标记的BERT模型结构图

Fig.4 Architecture of entity-marked BERT model

4 实验与分析

4.1 实验数据集

本文选取ACE(Automatic Content Extraction)2005中文数据集[10]进行关系抽取的测试和分析。该数据集是开放信息抽取任务上的公共数据集,根据关系抽取任务的特点及需求,筛选掉不规范的文档(句子级关系抽取任务基于实例文本中包含两个或两个以上实体的事实,进而使用关系抽取系统抽取两个实体之间的语义关系,故舍弃数据集中实体个数小于2的句子,以及不包含关系实例的文档。),实验1共用到628个文档。该数据集共包含6个实体关系类别。由于实体对间的关系是有方向的,例如:实体对(A,B)在实例文本中存在"PART-WHOLE"关系,但是反过来实体对(B,A)在数据集中就不存在这样的关系,故给予(B,A)实体对关系标记为"Negative"。或者句中出现的实体之间不存在任何关系,同样将这样的实体文本标记为"Negative",称为负例。

现有关系抽取任务的研究分为句子级别与文档级别[31-32]两大类,当前的研究工作主要聚焦于句子级关系抽取,也就是根据句子呈现的短文本内容识别实体关系。ACE 2005数据集中的实例文本是以句子的形式存储于树形结构文本中的。在取得整篇文章之后按照逗号、句号、冒号、分号、问号5种标点符号将其分割为句子,将包含了实体对且未在正例中出现的实例文本作为负例,将其设定为负例实例文本。经过筛选,最终得到的实验数据有107 384个包含两个或两个以上实体的例句,其中包括9 244条正例和98 140条负例。

4.2 实验设置

实验中,将107 384条数据按照6:2:2的比例切分为训练集、验证集和测试集。由于着重解决特征稀疏问题,同时也要控制特征噪声问题,所以文本固定长度的设定尤为重要。首先对实例文本长度进行分析,发现长度在15~35的实例频数在2 000以上,长度在5~15、40~50以及100~200区间的实例频

数在1 000～2 000区间内,为了减少特征噪声对模型预测的影响,同时降低因特征稀疏造成的性能下降等问题,将句长设置为180。实验证明这样的参数设置对ACE 2005中文数据集上的关系抽取性能有好的影响。数据集中实例文本长度分布如图5所示。

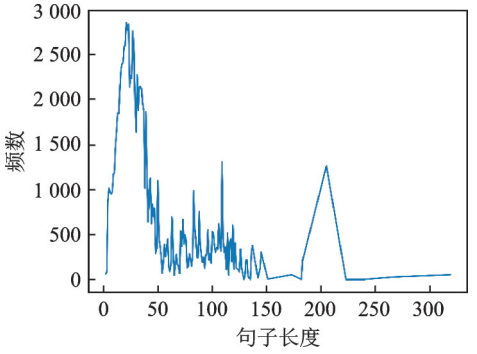

图5 ACE 2005中文数据集句子长度分布折线图

Fig.5 Lengths of sentences in ACE 2005 Chinese dataset

采用NLP技术常用评测指标准确率(P),召回率(R),综合评价指标(F_1)等对实验结果进行分析和判定。P和R是广泛用于信息检索和统计学分类领域的两个度量值,用来评价结果的质量。在关系抽取任务中,P是模型预测结果中预测正确的关系类别数与预测总数的比率,衡量预测系统的准确性;R是指模型预测结果中每个类别的正确数与被预测数据中该类别出现的个数的比率,衡量检索系统的查全率。综合评价指标F_1=正确率P × 召回率R × 2 /(正确率P +召回率R),F_1是正确率和召回率的调和平均值,是综合二者指标的评估指标,用于综合反映整体性能的指标。

4.3 实验结果与分析

本文使用第2章中描述的句子结构获取方法进行实验研究,利用语义表达能力较强的BERT预训练语言模型和能自动获取抽象特征的CNN模型分别在ACE 2005中文数据集进行关系抽取,实验性能如表1所示。

表1 实验结果

Table 1 Experimental results

实验组	BERT宏平均			CNN宏平均		
	准确率P	召回率R	F_1	准确率P	召回率R	F_1
1	0.841	0.769	0.803	0.750	0.615	0.675
2	0.889	0.838	0.863	0.772	0.619	0.687
3	**0.926**	**0.871**	**0.898**	**0.821**	**0.743**	**0.780**
4	0.924	0.869	0.896	0.842	0.718	0.775
5	0.922	0.870	0.895	0.798	0.748	0.772
6	0.920	0.868	0.893	0.844	0.723	0.779

对表1中实验数据加以详细分析。

(1) 第1组实验对照。该组实验不对实例文本中的实体对做特殊标记或说明,未凸显句子结构特征。

(2) 第2组实验,实体对复用。将实例文本中的实体对复制并前置至句首,使用“0”字符进行实体分隔,通过实体语义增强的方式获取句子结构特征。在BERT中,表义能力较强的预训练字向量使得实例文本获得更好的语义表示,性能有较大提升。CNN中采用随机初始化的向量表示,其表义能力较弱,因此实体对复用方法在CNN上没有取得性能的大幅度提升。但实体对复用能在神经网络下提升关系抽取的性能,此方法适用于在语义表示能力更强的文本表示的情况下来获取句子结构特征。

(3) 第3组实验,实体标记。实体标记方法是本文阐述的获取句子结构特征的核心方法,也是研究

动机。实体标记符由实体类型,实体相对位置复合而成,并通过字符“<”和“/”结合区分实例文本中实体的开始与结束。由引言对实体类型及实体相对位置关系的论述可知该类型的实体标记方法在关系抽取中的重要性,能使神经网络获取句子结构特征,改善句子级别关系抽取任务中特征稀疏问题带来的不良影响。结果表1中的“2”“3”行分别对应“实体对复用”和“实体对标记”的模型性能。第3组实验相对于第2组实验,BERT上有3%的提升,相对于第1组实验有近10%性能的提升,使得性能达到最高。在CNN中,由于实体对复用未能使神经网络获取到较强的实体对语义表达,性能提升不明显,但实体对标记方法却能在CNN中同样获得10%的性能提升。因此可以看出,利用实体类型及实体相对位置说明得到的实体对标记在获取句子结构特征,提升句子级中文关系抽取任务的性能中占据主导地位。

(4) 对于第4~6这3组实验,由于实体标记方法已经最大限度地利用其结构特征改善特征稀疏问题,在解决特征稀疏带来的不良影响方面已尽其所能,无法通过实体标记及实体复用结合的方法使得关系识别性能得到更大幅度的提升,甚至会因为特征过多且并非优质特征而带来噪声对性能产生少许不良影响,导致实验性能下降。实体对标记方法已经使得神经网络获取到以实例文本中实体对为中心的句子结构特征,且该特征能够显著提升实验性能,使得此方法下的特征获取达到瓶颈。因此,实体对标记与实体对复用方法进行结合以后,导致性能比单独使用实体标记方法的性能稍微差一点。

通过观察表1中的实验结果,最明显的比对来自于第1组(未使用任何标记和分隔)和第3组(使用实体类型与实体相对位置信息标记)。在神经网络上使用实体类型和实体相对位置复合标记符对实例文本中的实体对作标记,能够使实验性能相对于未使用实体标记的方法提高9%~11%。说明在实体标记中引入实体类型和实体相对位置等信息能使得神经网络学习到文本中以实体对为中心的句子结构特征和句子语义特征,更好地表达文本特征,同时也验证了前文所述的基本结构单位的组合、实体类型、实体间的结构特征等信息的确能够在句子级中文关系抽取中有效获取句子结构特征和语义特征,解决句子级中文关系抽取中的特征稀疏问题。在该CNN模型下,使用的字向量由一定范围内随机生成的查找表得到,语义表示能力相对较差,故而性能相对于BERT较低。本文的创新点是利用实体类型及实体相对位置信息标记和分隔的方法,使用神经网络获取句子的结构特征,在ACE 2005数据集上取得更好的中文关系抽取性能。

在BERT模型和CNN模型中实验性能宏平均F_1最高的均是第3组实验,即单使用实体标记方法,得到每个关系类别上的P、R、F_1值分别如表2所示。

表2 各大类实验性能

Table 2 Experimental results on main relation types

关系类型	BERT			CNN		
	准确率P	召回率R	F_1	准确率P	召回率R	F_1
PHYS	0.885	0.705	0.785	0.627	0.436	0.515
ART	0.919	0.803	0.857	0.807	0.620	0.701
GEN-AFF	0.921	0.864	0.892	0.854	0.767	0.808
ORG-AFF	0.950	0.911	0.930	0.921	0.825	0.871
PART-WHOLE	0.885	0.929	0.906	0.818	0.857	0.837
PER-SOC	0.931	0.890	0.910	0.739	0.706	0.722
Negative	0.993	0.998	0.995	0.982	0.991	0.987
Macro-F1	0.926	0.871	0.898	0.821	0.743	0.780

为了进一步验证该实体标记方法的有效性,做了对比实验,结果见表3。

表3 模型对比

Table 3 Model comparison

方法	特征	F_1
树核[33]	实体类别、GPE角色、引用类型、LDC类型	0.670
树核[34]	实体大类、实体小类、词林词群	0.668
CNN[35]	字向量、位置特征	0.673
Att-BiLSTM[36]	字向量、实体标记特征	0.842
TRE[37]	字向量、实体复用特征	0.562
Ours (CNN)	字向量	0.675
Ours (CNN+结构特征)	字向量、结构特征	0.780
Ours (BERT)	字向量、位置特征	0.803
Ours (BERT+结构特征)	字向量、位置特征、结构特征	0.898

虞欢欢等[33]提出了一种基于树核函数的方法获取句子信息。利用实体语义信息,构造合一句法和实体语义关系树来有效捕获结构化信息和语义信息,从而提高关系抽取的性能,作者在ACE 2005中文数据集上F_1达到了0.67。

语义信息在实体间语义关系抽取任务中具有重要作用。刘丹丹等[34]利用基于树核函数的方法,以《同义词词典》为例子,探讨了词汇的语义信息对中文关系抽取的影响,证明了无论在实体类型是否已知的情况下,语义信息都能提高某些关系抽取的性能。文章作者在ACE 2005中文数据集上F_1达到了0.668。

Nguyen等[35]利用多核卷积网络来自动提取句子的特征,使用了预训练的词向量和位置向量作为卷积网络的输入。由于文中作者使用的数据集与本文使用的数据集不同,因此按照作者论文中的方法复现了他的方法,不过文中使用位置向量,我们未在复现的时候使用位置向量,除此之外,均使用文中说明的方法构建CNN网络模型,并用在ACE 2005中文数据集中,得到F_1为0.673。

Zhou等[36]依据LSTM网络能够从较长文本中获取上下文语义依赖信息的特点,提出了一种基于注意力的双向长短期记忆网络(Att-BiLSTM)来捕获句子中最重要的语义信息。同样,该文也使用了实体标记方法,与本文不同的是,该文作者实验中使用的方法是利用“<e1>”、“</e1>”等标记符作为文本中实体对的位置指示器,从而对文本中的实体对进行标记。为了对比本文实验方法,我们同样使用了Att-BiLSTM网络模型在ACE 2005中文数据集上进行关系抽取实验,其实验性能F_1可以达到0.841。LSTM神经网络善于记忆长文本之间的语义依赖关系,而中文数据集中,单个字这样的基本组成单位组合成词,然后词与词的连接使得句子长度较大,同时由于实体标记方法可以使得神经网络获取到句子的结构特征等重要信息,故取得较好的实验结果。

Alt等[37]使用了Transformer框架,将其应用在英文实体关系抽取任务上,使用SemEval2010-Task8数据集和TACRED数据集,并将其提出的模型框架取名为基于转换器的关系抽取模型(Transformer for relation extraction,TRE)。作者的主要方法就是使用了预训练深度语言模型,捕获实体之间的语义关系。除此之外,该文中也使用了将实体对复用前置至句首的方法来处理输入数据,使用“[start]”和“[cls]”作为句子的开始与结束边界,使用“[sep1]”和“[sep2]”作为前置实体对的结束标记,但文中并未做出与非处理数据的对比实验效果。该文中的实验有一个缺点是并未对实体嵌套以及实体位置重叠的情况进行说明和相应处理。最终在两个数据集上F_1分别达到了0.674和0.871。使用该

模型且未使用预训练中文词向量的情况下在ACE 2005中文数据集上F_1达到了0.562。

从表3可以看出，通过对文本中实体对添加由实体类型及实体相对位置关系复合而成的标记符的方法，利用神经网络获取句子的结构信息效果显著。表3中第2、第3行使用了基于树核的方法，在使用了大量原子特征的情况下，F_1取得0.67。

实验证明，本文提出的面向中文关系抽取的句子结构获取方法能显著提高中文关系抽取性能。虽然使用BERT模型得到的结果已经明显高于其他方法，但是本文方法的关键之处在于使用实体标记符使得神经网络获取句子的结构特征，可以使中文关系抽取的F_1提高9%～11%。从利用CNN网络模型在获取到句子结构特征与仅使用词向量特征实验的对比中发现，本文方法能显著提高关系抽取的性能。在BERT模型中，利用Attention机制使得句中的所有文字都能学习到相对于两实体的相关信息，能获取到句中不同维度的依赖特征，从而提升关系抽取性能。

5 结束语

中文实体语义关系抽取中，由于句子长度参差不齐，分析数据长度并合理设置相关参数可达到少增加噪声信息，少丢失有用信息的目的，进而提高实验性能。实体类型、实体相对位置和实体标记等在实体关系抽取任务中能有效获取句子中以实体对为中心的结构特征，提高关系抽取精度值。

实验证明，本文提出的面向中文关系抽取的句子结构获取方法，在仅使用BERT模型的情况下，在ACE 2005数据集上取得F_1为0.803的抽取性能，再使用实体类型对句中的实体对进行标记和分隔，能够使得性能提高到0.898。在CNN模型结构中，同样使用实体类型标记和分隔，能够使得性能提高10.5%。由于实体对是关系抽取任务要识别的中心对象，利用实体标记可以获得以实体对为中心的句子结构特征，且该标记方法中加入了可以获得实体对语义信息的实体类型，故而能够较好地获得句子结构信息和语义信息，进而提升关系抽取的性能。

另一方面还发现句子的结构信息对正确理解句子至关重要，在特征提取到一定程度的时候，关系抽取的性能达到瓶颈，很难得到大幅度提升，这是接下来的研究过程中要解决的关键问题，进一步提升关系抽取的性能。之后将着重研究如何利用神经网络来提取句子的结构信息，并探究原子特征与结构特征之间的融合方法，同时考虑实例文本中除了对实体对施加类型标记符以外，是否能够在实体以外文本上通过其他方法获取更多特征信息，更有效地解决短文本带来的特征稀疏问题，达到更好更充分的特征结合与利用。并探究如何使用有效的预训练词向量与CNN结合，更好地获取句子结构特征，以达到更好的抽取性能。

参考文献：

[1] BASETIANELLI E, CASTELLUCCI G, CROCE D, et al. Textual inference and meaning representation in human robot interaction[C]//Proceedings of the Joint Symposium on Semantic Processing. Trento, Italy:[s.n.], 2013:65-69.

[2] SURDEANU M, HARABAGIU S, WILLIAMS J, et al. Using predicate-argument structures for information extraction[C]// Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan: [s.n.], 2003:8-15.

[3] SHEN Dan, LAPATA M. Using semantic roles to improve question answering[C]//Proceedings of the 2007 joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Prague, Czech Republic: [s.n.], 2007: 12-21.

[4] 李兆兆. 基于语义理解的智能问答系统关键技术研究 [D]. 西安: 西安邮电大学, 2019.
LI Zhaozhao. Research on key technologies of intelligent question answering system based on semantic understanding[D]. Xi′an: Xi′an University of Posts and Telecommunications, 2019.

[5] 黄恒琪，于娟，廖晓，等. 知识图谱研究综述[J]. 计算机系统应用，2019，28(6):1-12.
HUANG Hengqi, YU Juan, LIAO Xiao, et al. Review on knowledge graphs[J]. Computer Systems and Applications, 2019, 28(6): 1-12.

[6] 王文霞, 刘红岩. 英汉立法语言信息结构对比及在翻译中的应用[J]. 青年文学家, 2009 (19): 115-116.
WANG Wenxia, LIU Hongyan. Comparison of the information structure of English and Chinese legislative language and its application in translation[J]. Young Literary Scholar, 2009(19): 115-116.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in neural information processing systems. Doha, Qatar: Association for Computational Linguistics, 2017: 5998-6008.

[8] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019.

[9] LIU Chunyang, SUN Wenbo, CHAO Wenhan, et al. Convolution neural network for relation extraction[C] // International Conference on Advanced Data Mining and Applications. Springer, Berlin: Heidelberg, 2013: 231-242.

[10] WALKER C, STRASSEL S, MEDERO J, et al. Ace 2005 multilingual training corpus[J]. Progress of Theoretical Physics Supplement, 2006, 110(110):261-276.

[11] NAND A, KAMBHATL A. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the 42nd Annual Meeting of the Association for Computation Linguistics, Association for Computational Linguistics. Stroudsburg, United States: [s.n.], 2004: 22.

[12] RAZVAN C B,RAYMOND J M. A shortest path dependency kernel for relation extraction[C]//Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics. Vancouver, Canada: [s.n.], 2005: 724-731.

[13] 秦永彬,冯丽,陈艳平,等."智慧法院"数据融合分析与集成应用[J]. 大数据,2019,5(3): 35-46.
QIN Yongbin, FENG Li, CHEN Yanping, et al. "Intelligent Court" data fusion analysis and integrated application[J]. Big Data Research, 2019,5(3):35-46.

[14] 王长有, 杨增春. 一种基于句子结构特征的领域术语上下位关系获取方法[J]. 重庆邮电大学学报: 自然科学版, 2014, 26 (3): 385-389.
WANG Changyou, YANG Zengchun. An acquisition method of domain-specific terminological hyponym based on structure features of sentence[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2014, 26 (3): 385-389.

[15] CHEN Yanping, WANG Kai, YANG Weizhe, et al. A multi-channel deep neural network for relation extraction[J]. IEEE Access, 2020, 8: 13195-13203.

[16] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces[C] // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: [s.n.], 2012: 1201-1211.

[17] DAMASHEK M. Gauging similarity with n-grams: Language-independent categorization of text[J]. Science, 1995, 267(5199): 843-848.

[18] CHEN Yanping, ZHENG Qinghua, CHEN Ping. Feature assembly method for extracting relations in Chinese[J]. Artificial Intelligence, 2015, 228: 179-194.

[19] QUOC L, TOMAS M. Distributed representations of sentences and documents[C]//International Conference on Machine Learning. Beijing, China:[s.n.], 2014: 1188-1196.

[20] 刘娜娜,程婧,闵可锐,等. 基于短语成分表示的中文关系抽取[J].数据采集与处理,2020,35(3):449-457.
LIU Nana, CHENG Jing, MIN Kerui, et al. Chinese relation extraction based on constituency representation[J]. Journal of Data Acquisition and Processing,2020,35(3): 449-457

[21] 付春元. 汉语嵌套命名实体识别方法研究[D].哈尔滨: 黑龙江大学, 2011.
FU Chunyuan. Research on Chinese nested named entity recognition method[D]. Harbin: Heilongjiang University, 2011.

[22] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3: 1083-1106.

[23] SAHAMI M, HEILMAN T D. A web-based kernel function for measuring the similarity of short text snippets[C] // Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland: [s.n.], 2006: 377-386.

[24] ZHAO Shubin, GRISHMAN R. Extracting relations with integrated information using kernel methods[C] // Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05). Michigan: [s.n.], 2005: 419-426.

[25] ZENG Daojian, LIU Kang, LAI Siwei, et al. Relation classification via convolutional deep neural network[C] // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland. Dublin City University and Association for Computational Linguistics, 2014: 2335-2344.

[26] SANTOS C N, XIANG Bing, ZHOU Bowen. Classifying relations by ranking with convolutional neural networks[C]//

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China:ACL,2015: 626-634.

[27] ZENG Daojian, LIU Kang, CHEN Yubo, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]// Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon, Portugal: [s.n.], 2015: 1753-1762.

[28] SHEN Yatian, HUANG Xuanjing. Attention-based convolutional neural network for semantic relation extraction[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan:[s.n.], 2016: 2526-2536.

[29] 杨飘, 董文永. 基于BERT嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(4): 40-45.
YANG Piao, DONG Wenyong. Chinese named entity recognition method based on BERT embedding[J]. Computer Engineering, 2020, 46(4): 40-45.

[30] CHEN Jiangning, DAI Zhibo, DUAN Juntao, et al. Naive bayes with correlation factor for text classification problem[C]// Proceedings of 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). [S.l.]: IEEE, 2019: 1051-1056.

[31] WEI C H, PENG YIFAN, LEAMAN R, et al. Overview of the BioCreative V chemical disease relation (CDR) task[C]// Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. Sevilla, Spain: [s.n.], 2015, 14.

[32] YAO Yuan, YE Deming, LI Peng, et al. DocRED: A large-scale document-level relation extraction dataset[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL 2019. Florence, Italy: [s.n.], 2019: 764-777.

[33] 虞欢欢, 钱龙华, 周国栋, 等. 基于合一句法和实体语义树的中文语义关系抽取[J]. 中文信息学报, 2010, 24(5): 17-23.
YU Huanhuan, QIAN Longhua, ZHOU Guodong, et al. Chinese semantic relation extraction based on unified syntactic and entity semantic tree[J]. Journal of Chinese Information Processing, 2010, 24(5): 17-23.

[34] 刘丹丹, 彭成, 钱龙华, 等.《同义词词林》在中文实体关系抽取中的作用[J]. 中文信息学报, 2014, 28(2): 91-99.
LIU Dandan, PENG Cheng, QIAN Longhua, et al. The effect of Tongyici CiLin in Chinese entity relation extraction[J]. Journal of Chinese Information Processing, 2014, 28(2): 91-99.

[35] NGUYEN T H, GRISHMAN R. Relation extraction: Perspective from convolutional neural networks[C]// Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver, Colorado: [s.n.], 2015: 39-48.

[36] ZHOU Peng, SHI Wei, TIAN Jun, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 207-212.

[37] ALT Christoph, HÜBNER Marc, HENNIG Leonhard. Improving relation extraction by pre-trained language representations [C]//Proceedings of the 2019 Conference on Automated Knowledge Base Construction, Amherst, Massachusetts: [s.n.], 2019:18.

作者简介:

杨卫哲(1996-),男,硕士研究生,研究方向:自然语言处理、关系抽取,E-mail: wzyang.gzu@foxmail.com。

秦永彬(1980-),**通信作者**,男,博士,教授,研究方向:大数据治理与应用、多源数据融合。

黄瑞章(1979-),女,博士,副教授,研究方向:数据融合分析、文本挖掘、网络挖掘和知识发现。

王凯(1995-),男,博士研究生,研究方向:自然语言处理、关系抽取。

程华龄(1996-),女,硕士研究生,研究方向:自然语言处理、关系抽取。

唐瑞雪(1987-),女,博士研究生,研究方向:自然语言处理、关系抽取。

程欣宇(1978-),男,硕士,副教授,研究方向:机器学习、机器视觉、软件工程与网络通信。

陈艳平(1980-),男,博士,副教授,研究方向:人工智能、自然语言处理。

(编辑:陈珺)