ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 449－457
DOI:10. 16337/j. 1004-9037. 2020. 03. 008
© 2020 by Journal of Data Acquisition and Processing
http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于短语成分表示的中文关系抽取

刘娜娜[1,2]，程　婧[1,2]，闵可锐[3]，康　昱[4]，王　新[1,2]，周扬帆[1,2]

(1. 复旦大学计算机科学技术学院，上海，201203;2. 上海智能电子与系统研究院，上海，201203;3. 上海秘塔网络科技有限公司，上海，200135;4. 微软亚洲研究院，北京，100080)

摘　要：关系抽取是自然语言处理的重要研究内容，短语成分结构则是学界普遍认为能对关系抽取有重要影响的特征信息。然而目前短语成分应用于关系抽取任务时没有明显效果。这主要有两个原因：短语成分分析模型的泛化能力较差，会在关系抽取上造成错误传播，从而影响了它对关系抽取的有效性；关系抽取任务上使用短语成分特征的方式存在缺陷，即丧失短语成分分析学习到的句子结构信息，或者加大其对关系抽取的错误影响。本文在提升短语成分分析效果的基础上，提出了基于短语成分表示的中文关系抽取方法。该方法将短语成分分析模型学习到的文本表示嵌入到关系抽取模型中，从而提升关系抽取的性能。本文在公开的中文关系抽取数据集上验证了该方法的有效性。

关键词：短语成分表示；中文关系抽取；特征融合；短语成分分析

中图分类号：TP391.1;TP183　　**文献标志码**:A

Chinese Relation Extraction Based on Constituency Representation

LIU Nana[1,2], CHENG Jing[1,2], MIN Kerui[3], KANG Yu[4], WANG Xin[1,2], ZHOU Yangfan[1,2]

(1. School of Computer Science, Fudan University, Shanghai, 201203, China;2. Shanghai Institute of Intelligent Electronics & Systems, Shanghai, 201203, China;3. META SOTA, Shanghai, 200135, China;4. Microsoft Research, Beijing, 100080, China)

Abstract: Relation extraction is an important research in the natural language processing (NLP) area. The constituency grammar information, which is widely believed by the academic community, has an important influence on relation extraction. However, there is no obvious effect when the phrase syntactic tree is applied to the relation extraction task. There are two main reasons for this: First, the generalization ability of the constituency parser is poor, which will cause error propagation and then affect its effectiveness in the relation extraction; Second, there are flaws in the way of the use of the phrase syntactic features in the relation extraction task, that is the phrase syntactic structure information learned by the constituency parser is lost, or the wrong influence on the relation extraction is increased. This paper proposes a Chinese relation extraction method based on constituency vector representation to solve the above two problems. The method embeds the text representation learned by the constituency parser into the relation extraction model, thereby improving the relation extraction performance. This paper validates the method on a public Chinese relation extraction data set.

Key words: constituency vector representation; Chinese relation extraction; feature combination; constituency parser

---

基金项目：国家自然科学基金(61702107)资助项目；赛尔网络下一代互联网技术创新(NGII20180611)资助项目。
收稿日期：2019-08-20；**修订日期**：2019-12-09

引　言

关系抽取是信息抽取的关键内容之一，是自然语言处理的重要研究内容，也是知识图谱构建的关键技术。在关系抽取任务中，如果两个实体距离较远，中间有很多词汇干扰，则会严重干扰关系抽取的精度。比如，“周恩来（1898年3月5日—1976年1月8日），原籍浙江绍兴，1898年3月5日生于江苏淮安。”，在这个句子中，“周恩来”和“淮安”之间有较多词汇干扰，往往会影响现有的序列模型的效果。如果能获取这个句子的短语成分表示，则能在结构上去除这些干扰。用短语成分分析确定各成分之间的关系，学界普遍认为[1-2]这种结构化的表示能够帮助关系抽取任务。目前的关系抽取多是直接以词语、词性、实体及实体间的距离等作为特征向量[3-4]，采取端到端的模型，输出最终关系分类的结果，以往也曾经融合短语成分特征[2,5-8]，但在深度学习任务上没有明显效果。这主要有两方面的原因：(1)以往在关系抽取任务上适用短语成分特征，短语成分分析的性能是很大的一个阻碍因素。由于标注数据的获取较为困难，目前已知的中文短语成分分析的研究工作能达到的最高精度在86%～91%(F_1-score)[9-10]，然而一旦在开放领域，精度会严重下跌，甚至超过10个百分点。这个精度下生成的短语成分结构树本身有错误，继而在关系抽取任务上造成错误传播。(2)以往短语成分分析和关系抽取任务的融合主要有两种方式。首先，这两种方式均是根据短语成分分析模型得到完整树结构后再做处理。对于之后的处理，一种是根据完整树结构得到一些离散化特征，结合抽取的词法特征等一起放入机器学习模型中[2,5,6]，显然这种离散的方式割裂了短语成分解析树各成分之间的关联；另一种是根据得到的短语成分解析树直接构建树结构模型[7-8]，这种方式虽然充分利用了树结构的特性，但也加大了短语成分分析带来的错误影响，同时模型也更复杂。

针对上述所说的两个问题，本文提出了一种基于短语成分表示的中文关系抽取方法。对于第1个问题，本文采用短语成分分析模型[11]，简称为mparser(A minimal span-based neural constituency parser，一个最小化的基于跨度的神经句法分析器)，并在更大数据集上进行训练。针对第2个问题，本文提出了一种新的短语成分和关系抽取的融合方式。就像ELMo[12]用输入整句的一个编码函数来表示句子，使用上述mparser编码输出作为整个句子的短语成分表示(Constituency parsing to vector，Cons2vec)，再把这个短语成分表示迁移到关系抽取中，与词语、实体距离表示拼接到一起表示整个句子，之后再将这个向量表示注入分段卷积神经网络和注意力网络(A sentence-level attention-based piecewise convolutional neural network (CNN) for distant supervised relation extraction，PCNN_ATT)[4,13]，这样得到的短语成分表示能更完整地学习到句子的结构表示，减少解码过程中为优化损失函数而造成的信息偏离。

1　相关工作

目前国内外关于关系抽取的研究以英文为主，主流的研究方法包括有监督关系分类[2,5,14]、无监督关系发现[15-16]、基于知识库的远程监督关系抽取[4,6,14,17-21]和实体关系联合抽取方法[22-23]等。有监督方法最简单易用，精度也最高，但是其训练数据需要大量精准的有标签数据，具有很大的局限性，因此越来越多的学者关注远程监督关系抽取的研究及改进。使用远程监督方法能够快速构造海量的关系数据集，但是其中不可避免地存在大量的噪音，因此国内外许多学者主要研究工作就是围绕如何去噪进行，比如使用多实例学习[17-18]、注意力机制[4]、多语言学习[13]、强化学习[19-20]以及对抗式学习[21]等多种方式相结合。中文关系抽取由于缺乏有效的数据集，研究相对较少。而文献[13]提供了一个远程监督的中英文对照关系抽取数据集，同时为中文关系抽取提供了一个新的基准线。

短语成分分析作为自然语言处理的基础研究领域，对下游自然语言处理任务尤其是信息抽取领域能够提供结构信息上的支撑。近几年神经网络的深入发展，也给短语成分分析带来了很多改进。其中许多分析方法都采用编码-解码结构，编码器读取输入的句子表示，得到整合后的向量，再经过解码器构建出一个有标签的解析树[9,11,24]。这些工作的不同之处在于编码器和解码器使用多样的模型来表示，

比如编码器使用循环神经网络(Recurrent neural network, RNN),长短期记忆网络(Long short term memory, LSTM)以及双向长短期记忆网络(Bidirectional LSTM,BiLSTM),甚至自注意力(Self-attention)机制来表示。国内外关于短语成分分析的研究工作也是以英文为主,偶有少量工作有中文的对比实验,比如腾讯AI实验室提出的完全序列到序列模型(Sequence to sequence, seq2seq)的短语成分分析模型,就在中文关系抽取数据集上做了验证。就目前调研到的工作来看,在短语成分分析数据集上效果最好的单模型工作是mparser[11],后面有提升效果的工作则是在使用BERT外部预训练模型后才有稍明显的改进[9]。因此,为了得到短语成分表示,本文遵循的是文献[11]的mparser模型。

短语成分应用到关系抽取中主要有两种方式:(1)根据得到的短语成分解析树,按照某种规则生成离散的句法解析特征,与词法特征一起作为短语成分分析模型的特征向量输入。这种方式多用于传统机器学习方法,比如逻辑回归分类器、支持向量机分类器或最大熵[2,6]等。在数据量很小且对时间性能要求较高的场景下,这种方法很常用。但它的缺点也很明显,除了引言中所说原因造成精度不够外,还需要大量的特征抽取工作。(2)根据短语成分解析树,直接构建树结构模型,最流行的是tree-LSTM结构[7-8]。短语成分解析树的每个节点都是一个LSTM单元,每个单元仍然有3个门结构,只是每个门结构计算时是由所有子节点集成计算共同表示当前节点,递归地实现树结构。如果一个节点的子节点是叶子节点,那它的输入就是叶子节点向量表示,再通过一次线性计算和激活函数得到当前节点的表示。如果该节点的子节点都是非叶子节点,那它的输入就是所有非叶子节点的隐藏状态表示。

尽管人们普遍认为树结构模型对于像短语成分分析这样树结构的表达非常合理,并且对于下游关系抽取等任务的结构信息学习和去除冗余也非常有意义,但文献[25]也通过多组实验侧面表明,序列模型能发现隐藏的树结构。本文的工作就是使用短语成分分析模型中充分学习到结构信息的序列表示,将其应用到关系抽取中,在不影响原模型学习到的特征的前提下,加强对结构信息的学习,从而提升关系抽取的效果。

2 短语成分析模型和关系抽取模型

本节主要介绍本文工作涉及到的短语成分分析模型(mparser)和关系抽取模型(PCNN-ATT)。

2.1 mparser模型

本文遵循的是文献[11]的mparser模型,它使用编码-解码模型。如图1所示,这个模型使用的编码器可分解为两部分描述。第1部分是输入词语和词性后,通过BiLSTM得到句子的双向表示。具体地,编码器的输入是词语表示序列($\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3,\cdots,\boldsymbol{w}_t$)和词性标注序列($\boldsymbol{p}_1,\boldsymbol{p}_2,\boldsymbol{p}_3,\cdots,\boldsymbol{p}_t$)拼接而成($\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_t$),即有

$$\boldsymbol{x}_i=[\boldsymbol{w}_i,\boldsymbol{p}_i] \tag{1}$$

式中:第1个和最后1个是START和STOP标签,表示句子的开始和结束,i表示序列的第i个位置。输入的向量表示$\boldsymbol{x}_i$经过BiLSTM得到前向表示$\boldsymbol{f}_i$和后向表示$\boldsymbol{b}_i$。

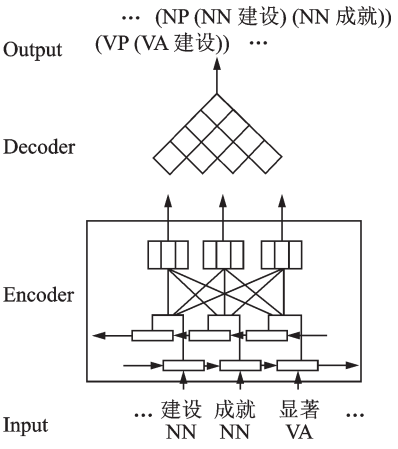

图1 mparser模型整体网络结构说明

Fig.1 Network structure of the mparser model

第2部分是根据序列的双向表示,得到跨度表示span(i,j),即从位置i到j这一范围的表示,可理解为获得句子从i到j的可能的短语成分标签。具体地,式(2,3)描述了跨度表示的具体实现,即有

$$\text{span}(i,j)=[\boldsymbol{f}_j-\boldsymbol{f}_i,\boldsymbol{b}_i-\boldsymbol{b}_j] \tag{2}$$

$$\boldsymbol{S}_{\text{span}}=\boldsymbol{V}g(\boldsymbol{W}\text{span}(i,j)+\boldsymbol{B}) \tag{3}$$

该工作使用两种解码器，即基于动态规划的传统图解析算法和基于贪心思想的自顶向下解析算法。本文使用易于理解且时间复杂度更低的自顶向下解析算法作解码器。该算法的整体思想是，对给定的序列范围指派一个标签，然后选定一个切分点，将该范围切分成左右两个子序列，重复此过程直到序列不能再被切分。选择一种代价最小的切分方法和标注方法，构造出最终的短语成分解析树。由于仅使用解码器训练短语成分分析模型，这里不做详细介绍。

2.2 PCNN-ATT模型

PCNN-ATT[4]是目前已知远程监督中文关系抽取数据集上效果最好的模型，也是基准模型，所以使用此模型作为本文工作的基础模型。图2是该模型的网络结构，总的来说，就是输入一组句子和对应的一对实体，最终输出它们关系的概率表示，仍然分为两部分描述。第1部分是句子编码部分，给定一个句子s，经过一个卷积网络和分段的池化层后，再经过一个非线性层，得到句子的分布式表示r'。第2部分是句子级别的注意力机制，前文得到的一组句子的表示r'对应学习到一个注意力权重α，最终关系$r=\sum ar'$，详细的注意力计算模型和损失函数参见文献[3]。

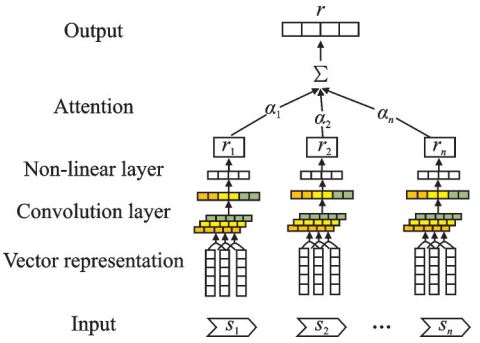

图2 PCNN-ATT模型的网络结构

Fig.2 Network structure of the PCNN-ATT model

3 短语成分表示的获取和嵌入

本节首先提出一种句子结构信息的表示方式：短语成分表示，然后采取一种新的短语成分结构和关系抽取的融合方式，将短语成分表示嵌入到关系抽取中，从而改进中文关系抽取的效果。图3是短语成分表示和关系抽取融合的整体结构示意图。输入是经过分词和词性标注的词语序列，一方面，它经过改进后的短语成分分析模型，得到短语成分表示。另一方面，根据输入序列得到对应的词向量表示和词语相对位置向量表示。然后将得到的短语成分表示和词向量、词语位置向量拼接到一起，经过卷积神经网络和注意力机制，最终得到更精准的关系分类结果。

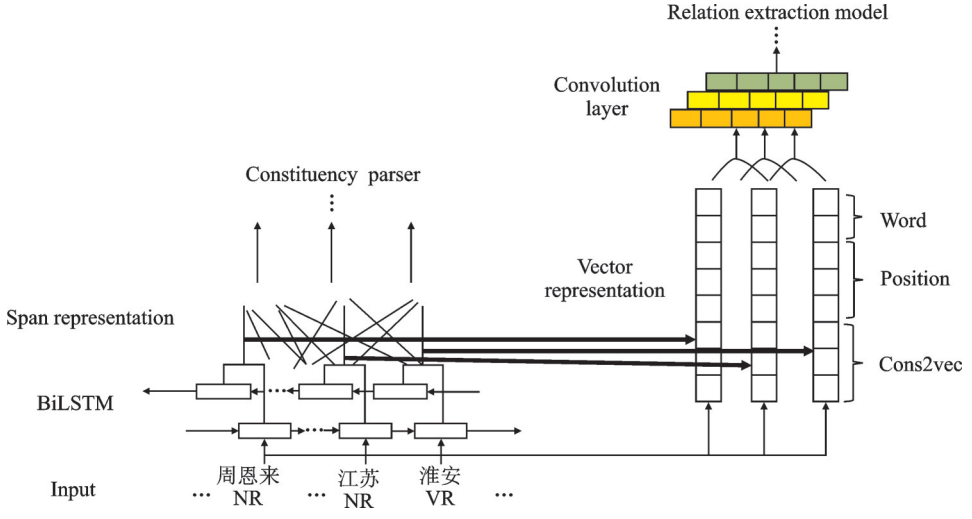

图3 短语成分表示和关系抽取融合结构的示意图

Fig.3 Combined structure of constituency vector representation and relation extraction

3.1 短语成分表示获取

为了获得泛化能力更强的短语成分分析模型,本文在更大的数据集上对模型mparser进行训练。然后将模型的BiLSTM双向隐状态输出作为短语成分表示,这样做能充分学习到句子结构信息。本文调研并选择了单系统最优的短语成分分析模型mparser[11],该工作初始用于英文短语成分分析,模型使用编码-解码结构,首先使用BiLSTM学习句子的表示,然后使用基于贪心思想的自顶向下解析算法,得到最优的树结构来表示短语成分解析树。经过算法调优和实验验证,发现其在中文短语成分分析上有更好的表现,优于目前已知工作的效果。之后,将其在更大的数据集上进行适配训练,最终得到预测能力更强的中文短语成分分析模型(c7parser)。

为了充分表示句子结构信息,从而改进关系抽取的效果,使用该模型的BiLSTM双向输出来表示句子结构,称之为短语成分表示。就像ELMo[12]用输入整句的一个编码函数来表示句子,本文也用这样的方式表示句子的结构信息。输入一个句子$\boldsymbol{s}$,得到$\boldsymbol{s}'\in\mathbf{R}^{(d_c\times 2)\times|\boldsymbol{s}|}$,其中$d_c$是每个前向LSTM单元或后向LSTM单元的隐状态输出维度,$|\boldsymbol{s}|$是句子的长度。

3.2 短语成分表示嵌入

获取到短语成分表示后,将其和词语分布式表示、词语位置表示一起,作为关系抽取的特征输入到关系抽取模型PCNN-ATT中。PCNN-ATT[3]是目前中文关系抽取数据集mnre[13]上的最好模型,因此选择这个模型作为本文训练的模型。PCNN-ATT基于多实例学习和句子级别的注意力机制。输入一组句子和一对实体,先映射得到词语分布式表示和词语位置表示,然后经过分段卷积神经网络和注意力层网络,最后得到关系分类结果。

词语分布式表示 词向量表示已经为大家熟知,成为自然语言处理任务的标配。具体使用的是谷歌提供的预训练词向量矩阵$\boldsymbol{V}\in\mathbf{R}^{(d_a\times|V|)}$,给定一个句子$\boldsymbol{s}$,由$t$个词语组成$(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3,\cdots,\boldsymbol{w}_t)$,每个词语从前述词向量矩阵中找到该词语的分布式表示$\boldsymbol{w}\in\mathbf{R}^{(d_a)}$,$d_a$为词向量的维度。

词语位置表示 关系抽取任务中,常将句子中每个词语到两个实体的相对位置单独表示,用以帮助CNN辨别该词语距离两个实体有多远。对于前述句子$\boldsymbol{s}$,每个词语的位置向量表示$\boldsymbol{p}\in\mathbf{R}^{(d_b\times 2)}$,$d_b$为位置表示的维度。

为了尽可能减少对模型的依赖,同时不影响短语成分表示的效果,将短语成分表示的每个timestep的隐状态直接与词语表示、词语位置表示拼接到一起。这样词语$\boldsymbol{w}$的最终低维向量表示为$\boldsymbol{w}\in\mathbf{R}^{(d=d_a+d_b\times 2+d_c\times 2)}$。

4 实验验证

本节主要通过实验数据验证本文的主要工作:(1)通过将合适的短语成分分析模型适配到中文短语成分分析任务上,中文短语成分分析的效果得到提升,F_1-score达到89.47%;(2)本文提出的基于短语成分表示的中文关系抽取方法,在中文关系抽取数据集mnre上效果有明显改进。由于这2个工作在2个数据集上进行训练,本节实验描述也分为两部分。

4.1 短语成分分析

对于短语成分分析,本文使用的是CTB5和CTB7数据集,用来训练不同性能的短语成分分析模型。这2个数据集分别来自Penn Chinese Treebank (CTB)版本5(CTB5)[26-27]和版本7(CTB7)[28],在CTB5上,使用标准的数据切分方式[10]。对于CTB7,为了更好地学习和测验预测能力,采用类似CTB5的切分方式。为了显示训练出的2个模型的扩展能力,本文使用了完全相同的测试集。遵照一般标

准[10],测试集的分词仍然使用数据集提供的标准分词,词性标注使用stanford词性标注器标注的结果。表1中给出了2个数据集的统计信息。

表1 短语成分分析模型训练数据集

Table 1 Dataset used by constituency parser

Constituency parser	Train set	Dev set	Test set
CTB7	44 434	2 634	4 375
CTB5	18 104	352	4 375(348)

表2是短语成分分析模型的效果对比,其中c5parser是使用CTB5数据集训练得到的模型,括号中的89.47%是测试集为348句样本与其他文献保持一致的情况下最终的F_1-score值。84.04%是使用扩展测试集CTB7数据得到的结果。c7parser是本文使用CTB7数据集训练得到的模型,86.49%则是使用CTB7测试集的F_1-score值。从表中结果可以看到,使用mparser在CTB5数据集上训练得到的模型效果已经是目前工作中最好的,F_1-score达到89.47%。但是当将CTB7的测试集作为开放领域的更广泛测试集来验证该模型的预测能力时,模型的F_1降到了84.04%,下降了5%左右。在CTB7数据集上训练该模型得到c7parser。同样的测试集下,F_1-score达到86.49%,因为训练数据集的数据分布更广泛,训练得到的模型预测能力也更强。表3是短语成分分析模型使用的超参数。

表2 短语成分分析模型的效果

Table 2 Performance of the constituency parser

Constituency parser	F_1-score/%
Socheretal[2011][29]	71.12
Zhangetal[2013][30]	84.43
Zhengetal[2015][31]	84.22
Dyeretal[2016][32]	84.60
Fernández-Gonzálezetal[2018][33]	86.80
Wangetal[2015][34]	86.60
c5parser	84.04(89.47)
c7parser	86.49

表3 短语成分分析模型的超参数

Table 3 Hyper-parameters of the constituency parser

Hyper-parameter	Value
Word embedding dimension	100
Pos dimension	50
LSTM dimension	50
Dropout rate	0.4
Batch size	2

4.2 中文关系抽取

本文使用的中文关系抽取数据集来自清华大学林衍凯等[13]公开的中英文双语关系抽取数据集,这是目前最大的中文关系抽取数据集。这个数据集中,中文实例是中文百度百科对齐wikidata生成的,英文实例是英文wikipedia对齐wikidata生成的。数据集中wikidata的关系事实分成3部分,分别用来作为训练集、验证集和测试集,包括NA(两个实体之间没有关系)在内,总共有176种关系,100多万条语句。表4是其中中文数据集的统计信息。遵循PCNN_ATT[3]的工作,本文也使用PR曲线作为评估指标。PR曲线就是以查准率Precision和查全率Recall为轴,取不同阈值画的一条曲线。曲线下的面积称为PR-auc,auc越大,或者曲线越接近右上角(查准率和查全率均为1),模型就越好。

表4 关系抽取模型训练数据集描述

Table 4 Dataset used by relation extraction

Relation extraction	Train set	Dev set	Test set
Sentence	940 595	82 699	167 224
Relation fact	42 536	2 192	4 326
Relation		176	

图4是中文关系抽取模型的对比效果。图中蓝线表示本文提出的基于短语成分表示的中文关系抽取模型的PR曲线(Cons7zh),橙线表示基准模型PCNN-ATT的PR曲线,可以看到Cons7zh的PR曲线

更靠近右上角,且几乎完全覆盖PCNN-ATT的PR曲线,验证了本文将短语成分表示嵌入到关系抽取确实有提升效果。

最后本文通过样例直观地说明短语成分表示对关系抽取的影响。表5列举了测试集中的3个实例,最后2列是本文模型和PCNN-ATT[3]分别预测的关系。可以看出,3个句子都是长句,且2个实体间的距离较远,中间有较多干扰。其中第1句和第3句,PCNN-ATT将这种情况的关系预测为NA,即实体间没有关系,而本文的模型则预测正确。对于第2句,原文将其预测为"主权国"关系,而实际上是"国籍国"。因此,从直观上也可以发现,短语成分表示的引入,不仅能有效去除长句中中间词汇的干扰,还能帮助区分更细粒度的关系。

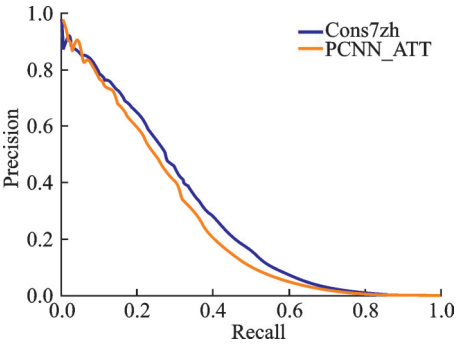

图4 关系抽取性能比较

Fig.4 Performance comparison of relation extraction models

表5 关系抽取的一些示例

Table 5 Some cases of relation extraction

No.	Sentence	Entity 1	Entity 2	Cons7zh relation	PCNN-ATT relation
1	1942年2月16日,金日成的儿子金正日出生在吉林长白山抗日游击队的营地。	金正日	长白山	P19(出生地)	NA
2	1859年——法国军官阿尔弗雷德·德雷福斯出生(1935年逝世)。	阿尔弗雷·德雷福斯	法国	P27(国籍国)	P17(主权国)
3	在德国十一月革命期间,他参加了汉堡工人士兵代表苏维埃的活动 。	汉堡	德国	P17(主权国)	NA

5 结束语

本文主要针对短语成分分析技术应用于关系抽取,以提升关系抽取效果这一问题,分析了现有方法存在的缺陷,针对性地提出了改进方法——基于短语成分表示的中文关系抽取方法,实验结果表明该方法确实提升了中文关系抽取的效果。本文针对中文关系抽取任务做出的改进是比较通用的方法,适用于有监督或者远程监督方法,采用的公开数据集也是远程监督数据集,但是本文方法还有待进一步优化。首先,由于加入了短语成分表示的提取过程,一旦数据量比较大,算法的整体执行时间就会增加;其次,加入了短语成分表示之后,也许可以通过优化目标函数等方式进一步提升训练效果。除此之外,短语成分分析应用于关系抽取还有很多值得探讨的地方,这些都是本文继续关注的方向。

参考文献:

[1] ZHOU Guodong, SU Jian, ZHANG Jie, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd Annual Meeting of the Association for Compu-tational Linguistics (ACL'05). Stroudsburg:ACL, 2005: 427-434.

[2] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction[C]//Proceedings of the ACL Interactive Poster and Demonstration Sessions. Stroudsburg: ACL, 2004: 178-181.

[3] 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. Stroudsburg:ACL,

2015.

[4] LIN Yankai, SHEN Shiqi, LIU Zhiyuan, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:ACL, 2016.

[5] SUCHANEK F M, IFRIM G, WEIKUM G. Combining linguistic and statistical analysis to extract relations from web documents[C]//Proceedings of the 12th ACM SIGKDD International Conference on knowledge Discovery and Data Mining. New York: ACM, 2006: 712-717.

[6] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of ACL-AFNLP. Stroudsburg: ACL. 2009: 1003-1011.

[7] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016.

[8] SHENG T K, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Stroudsburg: ACL, 2015: 1556-1566.

[9] KITAEV N, KLEIN D. Constituency parsing with a self-attentive encoder[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 2675-2685.

[10] LIU Lemao, ZHU Muhua, SHI Shuming. Improving equence-to-sequence constituency parsing[C]//Proceedings of the Thirty-Second {AAAI} Conference on Artificial Intelligence, (AAAI-18, the 30th Innovative Applications of Artificial Intelligence(IAAI-18, and the 8th {AAAI} Symposium on Educational Advances in Artificial Intelligence (EAAI-18). Menlo Park: AAAI, 2018: 4873-4880.

[11] STERN M, ANDREAS J, KLEIN D. A minimal span-based neural constituency parser[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 818-827.

[12] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2018: 2227-2237.

[13] LIN Yankai, LIU Zhiyuan, SUN Maosong. Neural relation extraction with multi-lingual attention[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 34-43.

[14] ZENG Daojian, LIU Kang, LAI Siwei, et al. Relation classification via convolutional deep neural network[C]//Proceedings of the 25th International Conference on Computational Linguistics. New York: ACM, 2014.

[15] HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2004: 415-422.

[16] CHEN Jinxiu, JI Donghong, TAN C L, et al. Unsupervised feature selection for relation extraction[C]//Proceedings of the International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2005: 262-267.

[17] HOFFMANN R, ZHANG C, LING Xiao, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]//Proceedings of ACL. Stroudsburg: ACL, 2011: 541-550.

[18] SURDEANU M, TIBSHIRANI J, NALLAPATI R, et al. Multi-instance multi-label learning for relation extraction[C]//Proceedings of EMNLP-CoNLL. Stroudsburg: ACL,2012: 455-465.

[19] FENG J, HUANG M, ZHAO Li, et al. Reinforcement learning for relation classification from noisy data[C]//Proceedings of the Thirty-Second {AAAI} Conference on Artificial Intelligence, (AAAI-18, the 30th Innovative Applications of Artificial Intelligence(IAAI-18, and the 8th {AAAI} Symposium on Educational Advances in Artificial Intelligence (EAAI-18). Menlo Park: AAAI, 2018.

[20] ZENG Xiangrong, HE Shizhu, LIU Kang, et al. Large Scaled relation extraction with reinforcement learning[C]//Proceedings of the Thirty-Second {AAAI} Conference on Artificial Intelligence,(AAAI-18). Menlo Park: AAAI, 2018.

[21] QIN Pengda, XU Weiran, WANG W Y. Distant supervision relation extraction via deep reinforcement learning[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 2137-2147.

[22] ZHENG Suncong, WANG Feng, BAO Hongyun, et al. Joint Extraction of entities and relations based on a novel tagging scheme[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 1227-1236.

[23] ZENG Xiangrong, ZENG Daojian, HE Shizhu, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 506-514.

[24] CHEN Danqi, MANNING C D. A fast and accurate dependency parser using neural networks[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 740-750.

[25] BOWMAN S R, MANNING C D, CHRISTOPHER P. Tree-structured composition in neural networks without tree-structured architectures[C]//Proceedings of the {NIPS} Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches Co-located with the 29th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2015: 37-42.

[26] XUE Naiwen, XIA Fei, CHIOU F D, et al. The penn Chinese treebank: Phrase structure annotation of a large corpus[J]. Natural Language Engineering, 2005, 11(2): 207-238.

[27] PALMER M, CHIOU F, XUE Nianwen, et al. Chinese treebank 5.0[EB/OL]. (2005-01-15)[2019-08-10]. https://catalog.ldc.upenn.edu/LDC2005T01,

[28] XUE Nianwen, JIANG Zixin, ZHONG Xiuhong, et al. Chinese treebank 7.0[EB/OL]. (2010-11-16)[2019-08-10]. https://catalog.ldc.upenn.edu/LDC2010T07.

[29] SOCHER R, LIN C C, MANNING C, NG A Y. Parsing natural scenes and natural language with recursive neural networks [C]//Proceedings of the 28th International Conference on Machine Learning (ICML-11). Washington: ICML, 2011: 129-136.

[30] ZHANG Meishan, ZHANG Yue, CHE Wanxiang, et al. Chinese parsing exploiting characters[C]//Proceed-ings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13). Stroudsburg: ACL, 2013.

[31] ZHENG Xiaoqing, PENG Haoyuan, CHEN Yi, et al. Character-based parsing with convolutional neural network[C]//Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015). Stroudsburg: ACL, 2015: 1054-1060.

[32] DYER C, KUNCORO A, BALLESTEROS M, et al. Recurrent neural network grammars[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. California: HLT-NAACL, 2016: 199-209.

[33] FERNÁNDEZ-GONZÁLEZ D, GÓMEZ-RODRÍGUEZ C. Faster shift-reduce constituent parsing with a non-binary, bottom-up strategy[J]. Artificial Intelligence, 2019, 275: 559-574.

[34] WANG Z, MI H, XUE N. Feature optimization for constituent parsing via neural networks[C]//Proceeding of Rewriting Techniques & Applications, International Conference. Como, Italy: [s.n.], 2015.

作者简介:

刘娜娜(1994-),女,硕士研究生,研究方向:关系抽取,E-mail: 17210240161@fudan.edu.cn。

程婧(1993-),女,硕士研究生,研究方向:文本分类。

闵可锐(1987-),男,博士,研究方向:信息检索、机器翻译和信息抽取。

康昱(1988-),**通信作者**,男,博士,研究方向:数据驱动的服务智能、基于数据分析方法提升云计算服务效能。

王新(1973-),男,教授,博士生导师,研究方向:新一代互联网体系结构、移动网络、网络编码。

周扬帆(1979-),男,副教授,博士生导师,研究方向:系统软件与软件工程。

(编辑:刘彦东)