智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合多层次特征的中文语义角色标注

王一成， 万福成， 马宁

引用本文:
王一成, 万福成, 马宁. 融合多层次特征的中文语义角色标注[J]. 智能系统学报, 2020, 15(1): 107–113.
WANG Yicheng, WAN Fucheng, MA Ning. Chinese semantic role labeling with multi–level linguistic features[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(1): 107–113.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201910012

DOI: 10.11992/tis.201910012

融合多层次特征的中文语义角色标注

王一成[1,2], 万福成[1], 马宁[2]

(1. 西北民族大学 中国民族语言文字信息技术教育部重点实验室, 甘肃 兰州 730030; 2. 西北民族大学 甘肃省民族语言智能处理重点实验室, 甘肃 兰州 730030)

摘 要: 随着人工智能和中文信息处理技术的迅猛发展, 自然语言处理相关研究已逐步深入到语义理解层次上, 而中文语义角色标注则是语义理解领域的核心技术。在统计机器学习仍占主流的中文信息处理领域, 传统的标注方法对句子的句法及语义的解析程度依赖较大, 因而标注准确率受限较大, 已无法满足当前需求。针对上述问题, 对基于 Bi-LSTM 的中文语义角色标注基础模型进行了改进研究, 在模型后处理阶段结合了 Max pooling 技术, 训练时融入了词法和句式等多层次的语言学特征, 以实现对原有标注模型的深入改进。通过多组实验论证, 结合语言学辅助分析, 提出针对性的改进方法从而使模型标注准确率得到了显著提升, 证明了结合 Max pooling 技术的 Bi-LSTM 语义角色标注模型中融入相关语言学特征能够改进模型标注效果。

关键词: 自然语言处理; 语义角色标注; 深度学习; Bi-LSTM; 语言学特征; 后处理层; Max pooling

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2020)01−0107−07

中文引用格式: 王一成, 万福成, 马宁. 融合多层次特征的中文语义角色标注 [J]. 智能系统学报, 2020, 15(1): 107–113.

英文引用格式: WANG Yicheng, WAN Fucheng, MA Ning. Chinese semantic role labeling with multi-level linguistic features[J]. CAAI transactions on intelligent systems, 2020, 15(1): 107–113.

Chinese semantic role labeling with multi-level linguistic features

WANG Yicheng[1,2], WAN Fucheng[1], MA Ning[2]

(1. Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730030, China; 2. Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu 730030, China)

Abstract: With the rapid development of artificial intelligence and Chinese information processing technology, studies relating to natural language processing have reached the level of semantic understanding gradually, while Chinese Semantic Role Labeling is the key technology in the semantic understanding field. Traditional tagging methods depend heavily on the parsing degree of sentence syntax and semantics, so the accuracy of tagging is limited. Aiming at the above problems, this paper improves the basic model of Chinese Semantic Role Labeling based on Bi-LSTM. To solve the above problem, the Max pooling technology is combined in the post-processing stage of the model, and multi-level linguistic features such as lexical item and sentence pattern are integrated into the training to further improve the original annotation model. Through a number of experimental demonstrations, combined with linguistic assistant analysis, targeted improvement methods are proposed to improve the accuracy of model annotation. It is proved that the Bi-LSTM semantic role labeling model combined with Max pooling technology can improve the effect of model annotation by incorporating relevant linguistic features.

Keywords: natural language processing; semantic role labeling; deep learning; Bi-LSTM; linguistic characteristics; post-processing layer; Max pooling

语义角色标注 (semantic role labeling, SRL) 是自然语言处理的重要技术, 这一技术的出现在很大程度上优化了现有的语言信息处理系统的性能。与深层语义分析相比, 语义角色标注具有问题清晰, 标注简便, 易于呈现等特点, 在信息检索、问答系统、机器翻译等多种自然语言处理领

收稿日期: 2019−10−11.
基金项目: 国家自然科学基金项目 (61602387, 61762076).
通信作者: 万福成. E-mail: wanfucheng@126.com.

域具有广阔的应用前景，并对今后进行深层语义分析以及篇章理解的研究具有重要意义。

近几年来，由于计算机技术与中文信息处理技术的快速发展，两种技术实现了高度的融合。在中文句法、语义处理方面不断取得新的突破性进展，中文语义角色标注已成为承上启下的关键技术，对其准确率的要求也随之提高，增加了深入研究语义角色标注技术的必要性。经过 10 余年的发展，中文语义角色标注取得了一系列的进展，尤其是基于机器学习方法的研究更是进展颇多。目前深度学习在语音识别领域的应用取得了巨大成功，该技术在文字信息处理领域的应用成为的一大研究热点。本文根据已有语义角色序列标注的研究现状，提出一种基于深度学习的中文语义角色标注改进方法，对现有基础特征进行选择性扩充，添加多层次语言学特征进行对比分析，经过多组对照测试达到准确率的提升，以期对语义角色标注研究提供借鉴。

语义角色标注是当前语义分析处理中切实可行的一种实践方案，随着统计机器学习方法在自然语言处理领域异军突起，很多大规模且具有语义信息的语料资源被建立，大大加快了基于特征学习的语义角色标注方法的实用化步伐。在基于机器学习的英文语义角色标注研究方面，Prandhan 等[1-2] 将支持向量机的机器学习方法应用于语义角色标注中，获得了较好的效果；Blunsom[3] 将更先进的机器学习模型——最大熵马尔可夫模型引入该领域，取得较好标注结果；Cohn 等[4] 则首次成功地将条件随机场应用到语义角色标注中；随着近两年人工智能的兴起，深度学习方法开始应用于该领域，Collobert 等[5] 将深度神经网络运用于框架语义角色标注，该方法减缓了传统机器学习方法应对复杂特征的人工干预力度，取得了理想的标注结果。随后，多层的神经网络也开始引入该领域，Socher 等[6] 采用神经网络单元与树结构编码器结合的方式进行标注，Yin 等[7] 则直接使用多层 CNN 模型进行标注。此类方法带来两大问题，一是模型层数增多，加重了梯度消失/爆炸的问题，二是不同层的神经网络单元需要学习不同层的语义知识，导致模型冗余。由于 LSTM 模型能够有效的改善梯度消失等问题，因此，Zhou 等[8] 采用 LSTM (long short-term memory) 模型[9] 进行语义角色标注，在模型训练过程中加入少量的词法特征，获得了相对理想的实验结果。

汉语方面，经过 10 余年的发展，语义角色标注在序列标注模型的实用方面取得了显著的进展。在 CoNLL 2004 大会中，首次将语义角色标注确立为主题，并在浅层句法分析理论基础上开展。于江德等[10] 结合英文语义角色标注研究实现了以短语或命名实体作为标注单元，利用条件随机场模型进行语义角色的标注；随着近年来，深度学习不断在自然语言处理领域拔得头筹，王臻等[11-12] 在 2014 年尝试将多层网络结构的深度学习模型应用于中文语义角色的识别和分类，虽然其实验效果与传统机器学习标注相比仍有较大差距，但该研究为深度学习算法应用于该领域提供了借鉴。在此基础上，该团队于 2015 年再次尝试将双向循环神经网络算法应用于该领域，该方法避免了大量的复杂特征提取，同时能够较好地利用标注序列中的信息。为解决多层神经网络带来的信息传递不畅以及网络层过多导致梯度消失/爆炸问题，王明轩等[13] 提出在多层 LSTM 模型单元内部设置信息连通的“直梯单元”，使标注信息能够快速地在不同层之间传递，而李天时等[14] 利用外部记忆单元构造出一个轻量级的单层 RNN 模型，该轻量级模型具有训练简便、标注效率高等优点，但准确率却接近多层次网络模型。杨耀文[15] 则在神经网络模型中引入词的分布表征和 Dropout 惩罚机制，极大地缓解了神经网络模型过度拟合的问题，明显提升了系统的标注性能。在基于规则的语义角色标注方面[17-18] 也进展显著。本研究也借鉴了相关文献 [19-22] 的模型构建方法，因篇幅有限，不详细介绍。

总体来说，由于可供训练的中文语义角色标注的语料资源还比较有限，另外汉语本身所带来的一些不同与英语的特点 (例如，汉语的目标动词不容易确定；汉语自动分析的基础模块，如分词、词性标注、句法分析等限制；等等)，导致中文语义角色标注的发展要相对曲折一些，因此中文语义角色的标注性能还有很大可提升空间。

1　模型构建

语义角色标注模型构建的主体思路是在一定规模的语料库中人工标出各种施事、受事、结果、方式等语义角色，运用深度学习方法从已标注完成的大规模语料中进行数据训练，通过提取各类语义角色在不同句子中的概率规则，对新语料中各语义角色进行概率最大的预估标注。对于标注模型来说，角色识别和角色分类是核心步骤，因此本文采用双向长短记忆网络 (Bi-directional long

short-term memory, Bi-LSTM) 算法从 Embedding 层获取角色识别和角色分类的核心高阶特征, 通过添加多种语言学特征使得原有模型的序列标注性能得到进一步提升。

1.1　理论方法

本方法采用词序列标注的标记策略, 使用神经网络分类器对句子中各类语义角色同时进行识别和标注。在后处理阶段, 利用 CNN 中的池化层对特征进行采样, 剔除冗余的特征信息。预测了可匹配的全部语义角色后, 采用简单的后处理规则去识别找不到匹配的语义角色成分, 保留预测概率最高的语义角色。

在标注模型选取上, 本文主要考虑了当前主流基于深度学习的序列标注模型——Bi-LSTM 模型。LSTM 是基于循环神经网络 (recurrent neural net-work, RNN) 的一种改进模型, 该神经网络模型具备超强的非线性拟合能力, 模型训练时将实例通过高阶、高纬度异度空间的复杂非线性变换, 映射得到一个低维的序列模型。相较于传统的机器学习模型, 无法灵活的添加自定义特征, 对于复杂语料中出现的语义角色不完全可分情况标注性能较差, 无法充分考虑序列内元素之间的相关信息等劣势, 由于 LSTM 设计特性, 其极大地改善了传统机器学习方法的不足, 能够较好地兼顾元素在序列中前后的顺序关系, 非常适合建模复杂的非线性序列数据, 如文本数据。综合考虑以上因素, 将采用 Bi-LSTM 模型进行研究, 实现标注性能的更大提升。

1.2　标注模型

融合多层次语言学特征的语义角色标注模型采取通用的标记方法, 采用最大池化法 (Max pooling) 的 Bi-LSTM 标注模型将问题转化成以词为基本标注单元的词序列标注问题, 融入多类型语言学特征, 构建深度学习标注模型进行自动标注训练, 之后选取参照语言学特征进行对比研究。本模型主要由预处理层、Bi-LSTM 层和后处理层架构组成, 其模型主体架构如图 1 所示。

1.2.1　预处理层

在输入层送入一段序列后, 经过预处理层将序列内的每一个词映射表示成所对应的词向量, 并送入 Bi-LSTM 层。假设输入句 A 包含 n 个词, $A=\{x_1,x_2,\cdots,x_n\}$, x_i 代表输入句中第 i 个词, 利用词向量矩阵 $\boldsymbol{E}_w$ 来获得词向量。$\boldsymbol{v}^w$ 表示词汇大小。通过式 (1), 可以将一个词 x_i 转变为词向量 $\boldsymbol{e}_i$:

$$\boldsymbol{e}_i = \boldsymbol{E}_w \boldsymbol{v}^i \tag{1}$$

式中: $\boldsymbol{v}^i$ 是向量 $\boldsymbol{v}^w$ 的绝对值距离。通过以上预处理, 初始序列片段句将以词向量的形式进入 Bi-LSTM 层网络。

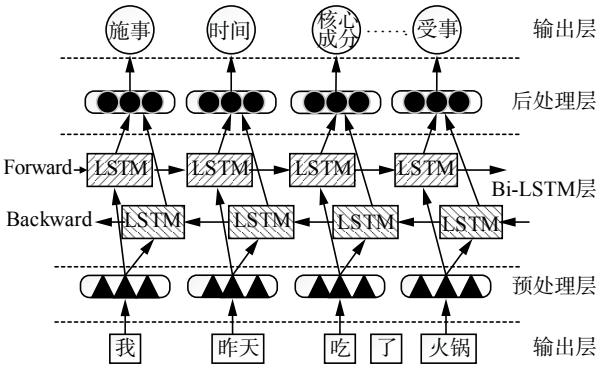

图 1　融合多层次语言学特征的语义角色标注模型架构图

Fig. 1　Semantic role labeling model architecture diagram incorporating multi-level linguistic features

1.2.2　Bi-LSTM 层

Bi-LSTM 层的基本思想是将一段训练句子同步加入到向前和向后两个循环神经网络 (RNN) 中, 并且这两个 RNN 训练出的单元同时指向一个 Max pooling 层接口。这种双向结构能够给 Max pooling 层提供输入句子中每一个词较为充分的上下文相关信息。其 LSTM 层的网络框架如图 2 所示。其中, 时序 t 对应的当前输入词为 X_t, 细胞状态为 C_t, 隐含层状态为 h_t。其训练过程可理解为通过遗忘和记忆单元处理当前时序状态下的新元素信息, 将影响因子较大的信息保留并传递给下一时序状态的细胞, 过滤影响因子较小的信息, 并输出该时序状态下的隐含层状态 h_t。依次迭代后, 就可得到与句子序列相对应的隐含层时序状态 $\{h_0,h_1,\cdots,h_{n-1}\}$。

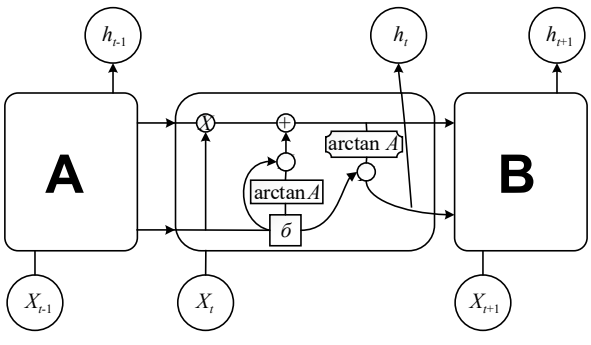

图 2　LSTM 层的网络框架示意图

Fig. 2　Schematic diagram of the network framework of the LSTM layer

Bi-LSTM 是由向前推算 (Forward pass) 与向后推算 (Backward pass) 双向组合而成, 其对于语义角色标注任务中普遍存在的上下文信息具有较好的兼顾。第 i 个词的隐含层状态由前向隐含层 $\overrightarrow{h}_1$ 与后向隐含层 $\overleftarrow{h}_1$ 进行异或得到, 如式 (2):

$$h_i = \left[\overrightarrow{h_i} \oplus \overleftarrow{h_i}\right] \tag{2}$$

1.2.3　后处理层

在语义角色标注过程中，句子序列内部的特征位置信息至关重要，比如施事一般位于句首、受事位于句尾、核心成分介于二者之间等，这些特征的位置信息对于角色识别及分类非常重要。而 Max pooling 技术的优势则在于：1) 标注序列中主要特征出现的位置在经过模型训练后仍可保持特征位置信息和旋转不变。2) 在神经网络模型训练中可用来减少网络参数，降低模型复杂度，减少训练的迭代次数。3) 对特征进行池化操作时，能够显著减少各滤波器的参数个数以及固定特征向量对应的神经元个数。因此，本模型在后处理层引入 Max pooling 技术。其特征向量与神经元映射关系如图 3 所示。

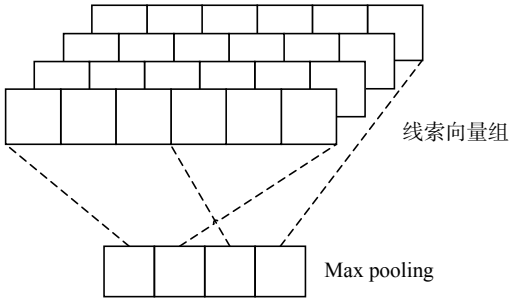

图 3　特征向量与神经元映射示意图

Fig. 3　Schematic diagram of eigenvectors and neuron mapping

2　实验

目前 LSTM 模型在序列标注的文本信息处理任务中取得了良好的效果，该方法不仅克服了原始卷积神经网络模型因不考虑句子序列内词序关系导致的句子向量表示与原始句子语义相冲突，而且可显著提升了长句式中元素初始部分的记忆能力。

本实验首先以公开的中文句法标注语料为基础，确定本实验使用的标记集，筛选并构建序列标注的中文语义角色标注语料库作为实验语料；然后调整语料的格式，构建并训练基于 Bi-LSTM 的 SRL 模型；在初始标注模型的基础上通过修改或添加多组新特征，逐步对原始模型进行改进训练，最后进行模型测评，分析对比得出相关结论。

2.1　语料选取

在中文语义角色标注语料方面，由于缺乏不同领域的大规模训练语料，导致在领域适应问题上，各类型语义角色标注方法未有较好的突破，因此本研究仅考虑单一领域的标注问题。实验选用面向新闻领域的清华大学依存句法树库语料作为原始语料，在此基础上进行加工，参照宾大中文树库的短语句法信息标注标准，实现了中文语义角色标注语料库的构建。在构建语义角色标注语料库的过程中，除了保留谓词划分、语义角色识别等传统语义角色标注语料库构建方面的要求外，还融入了词法、句法等相关语言学特征。

本实验构建的语料库题材是新闻语料，语义描述全面且颗粒度适中。经过筛选，共得到语料 22 000 句，其中训练语料 20 000 句，测试语料 2 000 句。语料库中主要语义角色统计数据如表 1 所示。

表 1　主要语义角色出现频数统计

Table 1　Frequency statistics of the main semantic roles

语义角色	频数	语义角色	频数
核心谓词	21 981	时间	2 997
施事	8 188	连接	11 288
受事	10 692	介词	8 621
程度	4 074	方位	3 539
处所	3 651	原因	372

词性粗颗粒度的序列标注语料构建具体过程如图 4 所示，从原始语料中抽取一个句子序列，可以看到其中含有多列特征，需要对这些标注信息进行筛选。

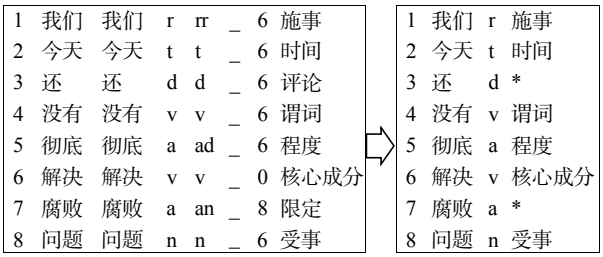

图 4　词性粗颗粒度语料构建示意图

Fig. 4　Schematic diagram of corpus construction of part-of-speech coarseness

2.2　模型参数

本实验使用 Bi-LSTM 模型提取句子序列的特征属性，将 Bi-LSTM 的隐藏层输出并做最大池化，得到所有可能的标注结果 Y^*。将 Y^* 输入后处理层的判别函数中，判别后输出最大概率的标注结果 Y。其标注模型的实验流程如图 5 所示。

实验中各个参数设置如下：dropout 值为 1，元素向量的维度大小为 50，学习率为 30%，学习速度下降梯度为 0.1，学习速度指数最大下降次数为 4，隐含层神经元的个数为 50，标注序列最大词汇数为 50。

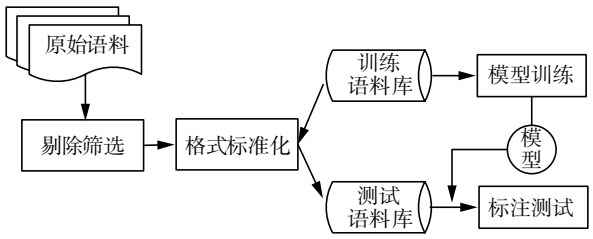

图 5　基于 Bi-LSTM 模型的实验流程图

Fig. 5　Experimental flow chart based on Bi-LSTM model

2.3　实验对比

实验 1　本实验对比基线 LSTM、Bi-LSTM、Bi-LSTM + Max pooling 3 种模型对同一的语义角色标注语料进行模型训练及测试，模型测试结果对比如表 2 所示。

表 2　各模型在实验中取得的最优结果

Table 2　the best results obtained by each model in the experiment

模型	词准确率/%	句子准确率/%
基线LSTM	53.95	8.34
Bi-LSTM	65.24	13.72
Bi-LSMT + Max pooling	78.75	18.25

实验以 LSTM 模型作为参考模型，在相同语料情况下，通过对比 3 种模型的词准确率和句准确率指标，同时对标注处理中效果较差的部分进行修改，表明 Bi-LSTM + Max pooling 技术相结合的模型标注性能最佳。因此，实验 2、3 将在 Bi-LSTM + Max pooling 模型的基础上进行。

实验 2　在 Bi-LSTM + Max pooling 模型的基础上，修改语料中的词性特征，使其采取粗−细颗粒度的划分方式，保持其他特征相同，分别进行 3 组模型训练的对照实验，模型测试结果对比如表 3 所示。通过 3 组不同词性颗粒特征的实验结果对比表明，不同词性颗粒特征对模型标注的准确率有较大影响，词性细颗粒度的标注结果最佳。

表 3　不同词颗粒特征标注结果对比

Table 3　Comparison of labeling results of different word granularity features

准确率	粗颗粒度	细颗粒度	粗−细颗粒度
词标注准确率/%	78.96	80.14	77.54
核心成分准确率/%	73.81	74.53	71.63
句子的准确率/%	18.25	20.10	17.58

实验 3　通过融入句式特征，探索能否进一步提升标注性能。

在对实验 2 的测试结果中错误句分析时发现：短句子中的非核心成分标注出错率较高，而长句子中核心成分的标注出错率较高。因此，实验设想融入长句和短句的句式判别来进一步探讨，通过句式判别程序在语料库中添加句式判别特征列，其中句式特征的阈值设定如表 4 所示。通过模型训练和测试，其测试结果对比如表 5 所示。

表 4　句式阈值判别表

Table 4　Sentence threshold judgment table

句式类型	元素个数
短句	$X \leqslant 5$
中句	$6 \leqslant X \leqslant 10$
长句	$11 \leqslant X \leqslant 15$
超长句	$X > 15$

表 5　融入句式特征的标注结果对比

Table 5　Comparison of labeling results with sentence features

准确率	融入句式特征	对照组
词标注准确率/%	82.96	80.14
核心成分准确率/%	77.81	74.53
句子的准确率/%	22.14	20.10

实验结果表明：对于训练语料中出现的语义相似、句式一致的短句非核心成分与长句核心成分的标注准确率有一定程度的提升。

将得到的实验最终结果与前人的语义角色标注最好结果进行对比如表 6 所示。

表 6　最终实验结果与前人结果对比

Table 6　the final experimental result is compared with the previous results

实验团队	总体准确率/%
Pradhan[16]	77.30
Zhou[8]	81.07
本文方法	82.96

2.4　实验分析

本实验是对深度学习的中文语义角色标注模型的一次改进尝试，构建的语料中既含有大量动词性谓词的句子，同时也包含大量名词性谓词的句子，接近真实的语言环境。实验 1 验证了不同种类的 CNN 模型对标注准确率的影响较大，选取合适的标注模型至关重要。融合 Max pooling 技术的 Bi-LSTM 模型不仅克服了传统 CNN 模型不考虑句子序列内词序关系导致的句子向量表示与原始句子语义相冲突，而且可以显著提升长句式中元素初始部分的记忆能力。为了验证不同词性颗粒度与标注准确率相关性的假设，遂进行了粗

颗粒度、细颗粒度、粗−细颗粒度 3 组对照实验，实验结果表明：不同粗细的词性颗粒度对标注准确率有一定影响，细颗粒度要比粗颗粒度的训练模型有更好的标注结果，但当尝试将粗、细颗粒度组合训练时，模型标注性能却不升反降。对比模型训练日志发现，随着词性颗粒度的复杂化，语义角色标签数量呈指数递增，标签数量的增多将直接反映于特征个数的增加，而特征个数过多随之带来模型收敛速度变差。测试结果的分析表明：1) 训练模型生成过程中出现了大量冗余或无关特征。2) 因词性粗细颗粒度的不同，模型在标注人名、组织机构名等命名实体时产生了切分歧义。从侧面说明，对于线性序列分类标注来说，并非特征越详细越好，特征过多，容易导致信息冗余，增加系统负担，拖累模型的整体标注准确率。对其标注结果进行错误分析，发现在句型相似的情况下，其短句中非核心成分以及长句中的核心成分出错率较高，进而猜想通过添加长短句标签，来提升句式及语义相似序列的标注准确率。实验 3 则验证了这一猜想，通过增加句式阈值特征，使得模型对长短句出错率较高部分的标注准确率有了不同程度的提升。

实验表明，每融入一个新特征都会对实验结果产生不同程度的影响。此外，模型进行标注预测时，可能会产生一些意料之外的预测结果 (如多个核心成分、超越边界、依赖边交叉等)。针对这些问题的解决将是我们下一步任务的改进重点。

3　结束语

语义角色凭借其展示简明、易于标注、应用广泛等特点，使其作为连接句法与语义层的研究关键点。本文在基于 Bi-LSTM 的中文语义角色标注模型构建时，结合了 Max pooling 技术，模型训练时尝试融入多类别的语言学特征，相较于卷积神经网络的语义角色标注基线方法，该方法是对传统神经网络标注方法的深度改进。实验结果表明，在 Bi-LSTM 模型中融入 CNN 的 Max pooling 技术能够有效提升标注准确率。此外，通过针对性的添加新特征，证实了对模型标注性能有一定提升作用。在接下来的研究工作中，我们将重点探究在模型中融入能够体现结构化的高阶特征，并将其与基于机器学习的线性序列标注进行组合，制定细化的角色判别规则以及引入语义相似度计算等深入处理，使模型能够更好、更快地识别语义角色，从而获得模型性能更大提升。

参考文献：

[1] PRADHAN S, HACIOGLU K, KRUGLER V, et al. Support vector learning for semantic argument classification[J]. Machine Learning Journal, 2005, 60(1/2/3): 11–39.

[2] PRADHAN S, WARD W, HACIOGLU K, MARTIN J, et al. Semantic role labeling using different syntactic views[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA, 2005: 581–588.

[3] BLUNSOM P. Maximum entropy markov models for semantic role labelling[C]//Proceedings of Australasian Language Technology Workshop 2004. Sydney, Australia, 2004: 109–116.

[4] COHN T, BLUNSOM P. Semantic role labelling with tree conditional random fields[C]//Proceedings of the 9th Conference on Computational Natural Language Learning. Ann Arbor, Michigan, 2005: 169–172.

[5] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]//Proceedings of the 25th Inter- national Conference on Machine Learning. New York, USA, 2008: 160–167.

[6] SOCHER R, HUANG E H, PENNINGTON J, et al. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection[C]//Proceedings of the Advances in Neural Information Processing Systems. Granada, Spain, 2011: 801–809.

[7] YIN W P, SCHÜTZE H. Convolutional neural network for paraphrase identification[C]//Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, 2015: 901–911.

[8] ZHOU J, XU W. End-to-end learning of semantic role labeling using recurrent 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, 2015: 1127–1137.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[10] 于江德, 樊孝忠, 庞文博, 等. 基于条件随机场的语义角色标注 [J]. 东南大学学报, 2007, 23(3): 361–364.
YU Jiangde, FAN Xiaozhong, PANG Wenbo, et al. Semantic role labeling based on conditional random field[J]. Journal of southeast university, 2007, 23(3): 361–364.

[11] WANG Zhen, JIANG Tingsong, CHANG Baobao, et al. Chinese semantic role labeling with bidirectional recur-

rent neural networks[C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1626–1631.

[12] 王臻, 常宝宝, 穗志方. 基于分层输出神经网络的汉语语义角色标注 [J]. 中文信息学报, 2014, 28(6): 56–61.
WANG Zhen, CHANG Baobao, SUI Zhifang. Chinese semantic role labeling based on neural network with optimized output layer[J]. Journal of Chinese information processing, 2014, 28(6): 56–61.

[13] 王明轩, 刘群. 基于深度神经网络的语义角色标注 [J]. 中文信息学报, 2018, 32(2): 50–57.
WANG Mingxuan, LIU Qun. A simple and effective deep model for semantic role labeling[J]. Journal of Chinese information processing, 2018, 32(2): 50–57.

[14] 李天时, 李琦, 王文辉, 等. 基于外部记忆单元和语义角色知识的文本复述判别模型 [J]. 中文信息学报, 2017, 31(6): 33–40.
LI Tianshi, LI Qi, WANG Wenhui, et al. Paraphrase identification with external memory and SRL knowledge[J]. Journal of Chinese information processing, 2017, 31(6): 33–40.

[15] 杨耀文. 基于神经网络模型的汉语框架语义角色识别 [D]. 山西大学, 2016.
YANG Yaowen. Identification of Chinese FrameNet semantic role based on neural networds model[D]. Shanxi University, 2016.

[16] Sameer Pradhan, Kadri Hacioglu, Wayne Ward, et al. Semantic role chunking combining complementary syntactic views[C]//Proceedings of the Conference on Computational Natural Language Learning, 2005:217–220.

[17] 何保荣, 邱立坤, 孙盼盼. 基于句式与句模对应规则的语义角色标注 [J]. 中文信息学报, 2018, 32(4): 59–65.
HE Baorong, QIU Likun, SUN Panpan. Semantic role labeling based on correspondence rules between syntactic pattern and semantic pattern of sentences[J]. Journal of Chinese information processing, 2018, 32(4): 59–65.

[18] 杨凤玲, 周俏丽, 蔡东风, 等. 结合短语结构句法的语义角色标注 [J]. 中文信息学报, 2018, 32(6): 1–11.
YANG Fengling, ZHOU Qiaoli, CAI Dongfeng, et al. Semantic role labeling combined with phrase structure prasing[J]. Journal of Chinese information processing, 2018, 32(6): 1–11.

[19] 谢先章, 王兆凯, 李亚星, 等. 基于卷积神经网络的跨领域语义信息检索研究 [J]. 计算机应用与软件, 2018, 35(08): 73–78.
XIE Xianzhang, WANG Zhaokai, LI Yaxing, et al. Cross-domain semantic information retrieval based on convolutional neural network[J]. Computer applications and software, 2018, 35(08): 73–78.

[20] 王策, 万福成, 于洪志, 等. 基于 Bi-LSTM 和 Max Pooling 的答案句抽取技术 [J]. 吉林大学学报(信息科学版), 2019, 37(4): 390–398.
WANG Ce, WAN Fucheng, YU Hongzhi, et al. Answer sentence extraction technology based on Bi-LSTM and Max Pooling[J]. Journal of jilin university(information science edition), 2019, 37(4): 390–398.

[21] WANG Y C, WAN F C, MA N, et al. Research on chinese semantic role labeling with hierarchical syntactic clues[C]//Proceedings of the 3rd International Conference on Economics and Management, Education, Humanities and Social Sciences. Suzhou, China, 2019: 190–196.

[22] 万福成. 基于改进混沌分区算法的模糊信息抽取 [J]. 计算机应用研究, 2019, 36(10): 2952–2954, 2970.
WAN Fucheng. Fuzzy information extraction based on improved chaotic partition algorithm[J]. Application research of computers, 2019, 36(10): 2952–2954, 2970.

作者简介:

王一成, 硕士研究生, 主要研究方向为自然语言处理、自动问答。

万福成, 副教授, 主要研究方向为自然语言处理、机器翻译、信息抽取、自动问答。主持和参与国家级、省部级项目 10 余项。获得专利及软件著作权 10 余项。出版著作 4 部, 发表学术论文 20 余篇。

马宁, 教授, 主要研究方向为自然语言处理、计算机应用。主持及参与国家自然科学基金项目 3 项。出版学术著作 1 部, 发表学术论文 40 余篇。