文章编号：1003-0077(2023)04-0018-10

融合词性位置特征的多任务汉老双语短文本相似度计算方法

李炫达[1]，周兰江[1]，张建安[2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 战略支援部队 信息工程大学三院昆明大队，云南 昆明 650500)

摘 要：老挝语属于低资源语言，在有限的语料中获取更多的语义信息可以有效解决汉语和老挝语短文本相似度计算不准确的问题。多任务学习是有效获取语义信息的一种方法，该文对汉语和老挝语短文本特点进行研究后，提出一种融合词性位置特征的多任务汉老双语短文本相似度计算方法：首先，通过词性位置特征权重和 TF-IDF 权重加权表征双语短文本的同时，使用改进后的 TextRank 算法获取双语短文本的核心句；然后，通过带有自注意力机制的双向长短时记忆网络分别计算双语短文本的相似度与双语短文本对应核心句的相似度；最后，使用多任务学习方法，将双语短文本的核心句相似度计算作为辅助任务，获取更多的语义信息进行共享以提升汉老双语短文本相似度计算模型的性能。实验结果表明，该文提出的方法在有限的训练语料下取得了更好的效果，F_1 值达 76.16%。

关键词：汉语-老挝语；词性位置特征；多任务学习；短文本相似度计算

中图分类号：TP391 **文献标识码**：A

Multi-task Approach to Similarity Computing Method Between Chinese and Lao Combining Part-of-Speech and Position

LI Xuanda[1], ZHOU Lanjiang[1], ZHANG Jian'an[2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500 China;
2. Kunming Branch, No.3 College, PLA Information Engineering University, Kunming, Yunnan 650500, China)

Abstract: To acquire more semantic information from short texts in both Chinese and Lao, this paper studies the characteristics of short texts in Chinese and Lao and proposed to employ part-of-speech and position information is characteristics in a multi-task short text similarity computation method. Firstly, part-of-speech and position weights are augmented to the TF-IDF weight presentation of short texts in both languages, and then the improved TextRank algorithm is applied to obtain the core sentences of short texts in both languages. Subsequently, bi-directional long-short-term memory network with self-attention mechanism is applied to compute the similarity of short texts in both languages and similarity of corresponding core sentences, respectively. Finally, the multi-task learning method is applied for the similarity computing model of short texts in both Chinese and Lao, with the core sentence similarity computing of short texts in both languages as the auxiliary task,. The test results indicated that the method proposed in this paper achieved good performances with limited training linguistic materials (76.16%).

Keywords: Chinese-Lao; part of speech and position characteristics; multi-task learning; short text similarity computation

收稿日期：2020-08-08 **定稿日期**：2020-09-03

基金项目：国家自然科学基金(61662040)

0　引言

老挝是我国边境国之一，其语言老挝语属于稀缺资源型语言。汉老双语短文本相似度计算是指计算汉语和老挝语短文本间的语义相似程度，对汉老文本聚类、文本相关性判断、摘要提取等领域的研究有着重要作用。

短文本通常由多个句子构成，文本长度较短且特征词较为稀疏[1-2]，如何表征短文本的语义信息是计算短文本相似度的关键。在跨语言短文本相似度计算任务中，目前主要有传统方法和基于向量空间模型的方法。传统方法如基于语义词典的方法[3-5]、基于主题模型的方法[6-8]、基于语义特征工程的方法[9]等，这类方法大多需要依靠大量的语料数据和人工提取特征来表征短文本相似度，在语料资源较少或难以获取语言特征时效果较差；而在基于向量空间模型方法中，目前主流的方法如计算文本 TF-IDF 系数加权句向量表征文本的方法[10]、多特征权重加权的句向量表征文本的方法[11]、将短文本映射到向量空间后结合神经网络计算相似度的方法[12-15]等，这类方法的特点是将短文本映射到向量空间计算相似度，不依赖于人工抽取即可得到短文本特征，相似度计算的结果与短文本的语义向量表征有较大的关联性，因此在计算语言跨度较小或语料资源丰富的语言时，这类方法取得了更好的效果。与计算印欧语系[16-17]短文本相似度不同的是，汉语和老挝语具有较为明显的语言差异性[18-19]，因此针对汉老双语短文本相似度计算任务，采用结合语义特征工程的方法与向量空间模型的方法来进行研究。

与李卫疆等人[20]方法类似，本文使用基于特征工程的方法获取汉老双语短文本的特征后，融入汉老双语短文本的分布式表示，最终通过神经网络模型获取含有更多语义信息的特征向量，计算汉老双语短文本间的相似度分数。由于老挝语基础研究薄弱，目前仅在分词[21]和词性标注领域[22-23]取得了较好成果，因此本文在研究汉语和老挝语的语言特点的基础上，提出了一种根据单词词性及其位置信息来获取语义特征权重的方法，通过与短文本的 TF-IDF 权重系数结合，采用多权重加权的方式来获取含有更多语义信息的短文本分布式表示；在此基础上，在对汉老双语的差异性进行研究的同时，发现汉语和老挝语在语序结构上具有一定的相似性。因此采用多任务学习[24]的方法，通过改进后的 TextRank 算法获取汉老双语短文本的核心句，计算双语核心句间的相似度分数作为辅助任务，共享更多的语义信息以提高双语短文本相似度计算的准确性。

综上，本文在研究汉语和老挝语的语言特点的基础上，提出一种融合词性位置特征的多任务汉老双语短文本相似度计算方法。该方法分为主任务和辅助任务两部分：主任务在通过词性位置特征和 TF-IDF 方法得到句子加权权重系数后对短文本进行分布式表示，输入到带有自注意力机制（Self-Attention）的双向长短时记忆网络（BiLSTM）中计算汉老双语短文本的相似度分数；辅助任务在获取汉老双语短文本的核心句后，通过词性位置特征加权得到核心句的句向量作为全局向量，与核心句的词分布式表示进行拼接，输入到与主任务共享参数的网络模型中，计算双语短文本核心句的相似度分数，通过赋予辅助任务的损失函数权重以控制辅助任务共享的语义信息的重要性来提高模型性能。实验结果表明，与目前主流方法相比，本文方法在语料资源稀缺的情况下得到了更优的结果，模型的 F_1 值达到了 76.16%。

本文的主要贡献如下：

（1）提出一种根据汉老双语短文本的单词词性及其位置信息来获取短文本的语义特征权重的方法，使双语短文本的分布式表示得到更多的语义信息。

（2）采用多任务学习的思想，通过计算双语短文本的核心句相似度作为辅助任务，使模型在有限的语料下能得到更多的语义信息。

（3）通过对 TextRank 算法进行了改进，采用神经网络方法来计算相似度矩阵，有效提高算法获取核心句的准确性。

本文组织结构如下：第 1 节为相关工作，综述了目前短文本相似度计算的相关研究成果；第 2 节介绍了汉语和老挝语短文本的特点及异同；第 3 节介绍了本文使用的模型及其结构；第 4 节为本文模型的设置及实验；第 6 部分为总结与展望。

1　相关工作

传统的跨语言短文本相似度计算方法主要有基于语义词典的方法、基于主题模型的方法和基于语义特征工程的方法。

(1) 基于语义词典来计算短文本相似度的主要思想是,通过构建源语言与目标语言间的中间语义层,通过双语词典进行映射后计算双语文本间的相似度。例如,翟延冬等人[3]通过 WordNet 方法构建了语义中间层,采用 ST-CW 算法计算短文本间的相似度;Atoum 等人[4]结合 WordNet 方法与 Corpus 方法,通过加权词相似度来得到短文本相似度;霍文杰[5]使用英语作为中间层语言,通过 Wordnet 方法计算了汉语和老挝语文本的相似度。

(2) 基于主题模型(LDA)方法的思想是通过抽取文本的语义主题集合,构建主题概率分布来计算文本间的相似度。例如,王振振等人[6]在采用 LDA 主题模型建模语料库后,通过 Gibbs 抽样间接计算模型参数以挖掘文本内不同词与主题间的关系,最终计算文本的主题分布来得到文本间的相似度;He 等人[7]通过对 LDA 模型得到主题的基础上,学习主题特征权重进行加权,最终计算短文本间的相似度;张小川等人[8]在原有的 LDA 主题模型计算文本相似度方法的基础上,加入了语义特征和词语特征,进一步提高了文本相似度计算的准确性。

(3) 基于语义特征工程的方法主要是通过获取文本的特征来表征文本的相似度。例如,赵谦等人[9]将文本分割为句子单元,在提取句子依存句法特征的基础上加入情感特征,最后通过二元集合法将计算的句子相似度合并,得到文本间的相似度。

传统方法在计算文本相似度任务上虽然取得了不错的成果,但双语词典匹配的方法需要大量的双语词典资源来构建语义中间层,主题模型方法对特征词较为稀疏的短文本提取主题的效果有限,语义特征工程的方法依赖于人工抽取特征的效果。因此,近年来,对于跨语言短文本相似度计算的研究更多的是采用基于向量空间模型的方法,通过将短文本进行分布式表示,结合神经网络模型学习短文本的特征来计算相似度。例如,Yao 等人[12]通过并行卷积神经网络和 LSTM 网络作为编码器提取特征后计算了短文本间的相似度;Kenter 等人[13]在向量空间中结合外部语义知识表征短文本后,计算短文本间的相似度;Boom 等人[14]在融入语义背景知识特征的基础上,结合 TF-IDF 将短文本进行分布式表示,计算了短文本间的相似度;郭浩等人[15]将短文本映射到向量空间后,使用带有注意力(Attention)机制的卷积神经网络(CNN)和双向 LSTM 网络提取特征,计算了短文本间的相似度。

2 汉语-老挝语短文本的特点及异同

汉语和老挝语的语序结构相似,主要成分顺序均为主谓宾(SVO)[19],并且构成主要成分的实词词性及其在句子中所处的位置具有相似性[18]。从汉老句子主要成分的构成来看,汉语和老挝语的主语大多由名词、代词等词性构成,谓语主要由动词、形容词等词性构成,宾语的词性构成与主语的类似;从主要成分在汉老句子中的位置来看,对于拥有完整主要成分且不含从句的句子,主语往往位于句子的前部,谓语大多位于中部,而宾语大多位于尾部。如表 1 例句所示,例句均为主要成分完整且不含从句的句子,其中/n,/r,/v,/d,/p,/u,/m 分别为名词、代词、动词、副词、介词、助词和数词。

由于短文本通常文本长度较短且由多个句子组成,长度较长的复合句较少,因此通过对汉语和老挝语短文本的词性和位置信息进行识别,获取更多的实词语义信息,丰富双语短文本的语义表征。

表 1 汉老双语单词词性、位置及主要成分间的关系示例

例 1	汉语	那个/r 人/n(主语) 在/p 车里/n 睡着/v(谓语) 了/u
	老挝语	ฉาย/r ຄົນ/n(主语) ຢູ່ໃນ/p ລົດ/n ນັ້ນ/r ນອນຫລັບ/v(谓语)
例 2	汉语	路边/n(主语) 正好/d 停/v(谓语) 了/u 一辆/m 车/n(宾语)
	老挝语	ຕອນນີ້/n(主语) ມີ/v(谓语) ລົດຈອດ/n ຢູ່/p ແຄມທາງ/n(宾语)
例 3	汉语	老板/n(主语) 不/d 在/v(谓语) 办公室/n(宾语)
	老挝语	ນາຍຈ້າງ/n(主语) ຢູ່ ໃນ/v(谓语) ຫ້ອງການ/n(宾语)

3　融合词性特征的多任务汉老双语短文本相似度计算模型

3.1　模型结构

本文构建的模型分为主任务和辅助任务，其结构如图 1 所示，主要思路如下：

3.1.1　主任务

首先，对汉语和老挝语短文本进行分词和词性标注预处理，通过获取短文本中每个句子的词性位置特征权重系数后，与计算得到的 TF-IDF 系数累乘作为句子中每个单词的特征权重系数；其次，通过对每个句子按其对应的特征权重系数进行加权，得到短文本中每个句子的特征向量表示，将其组合得到短文本的最终语义分布式表示；最后，分别将短文本的语义表示向量输入到带有自注意力(Self-Attention)机制的双向长短时记忆网络(BiLSTM)中得到对应的双语短文本特征向量，通过计算向量间的相对积和相对差，将结果拼接传入全连接网络层得到汉老双语短文本间的相似度分数。

3.1.2　辅助任务

首先，对主任务中经过预处理后的汉老双语短文本使用改进后的 TextRank 算法提取对应的双语短文本核心句；其次，通过获取核心句的词性位置权重系数，加权得到每个句子对应的句向量作为核心句的全局特征向量，将其拼接在核心句分布式表示的最后以丰富句子的语义特征表示；最后，将含有全局语义信息的核心句分布式表示输入到共享参数的带有自注意力机制的双向长短时记忆网络中，对得到的双语核心句特征向量计算其相对积和相对差，将结果拼接后传入辅助任务的全连接网络层以计算核心句间的相似度分数，通过赋予辅助任务的损失函数权重，控制辅助任务经过网络层共享的语义信息的重要性来提高主任务模型的性能。

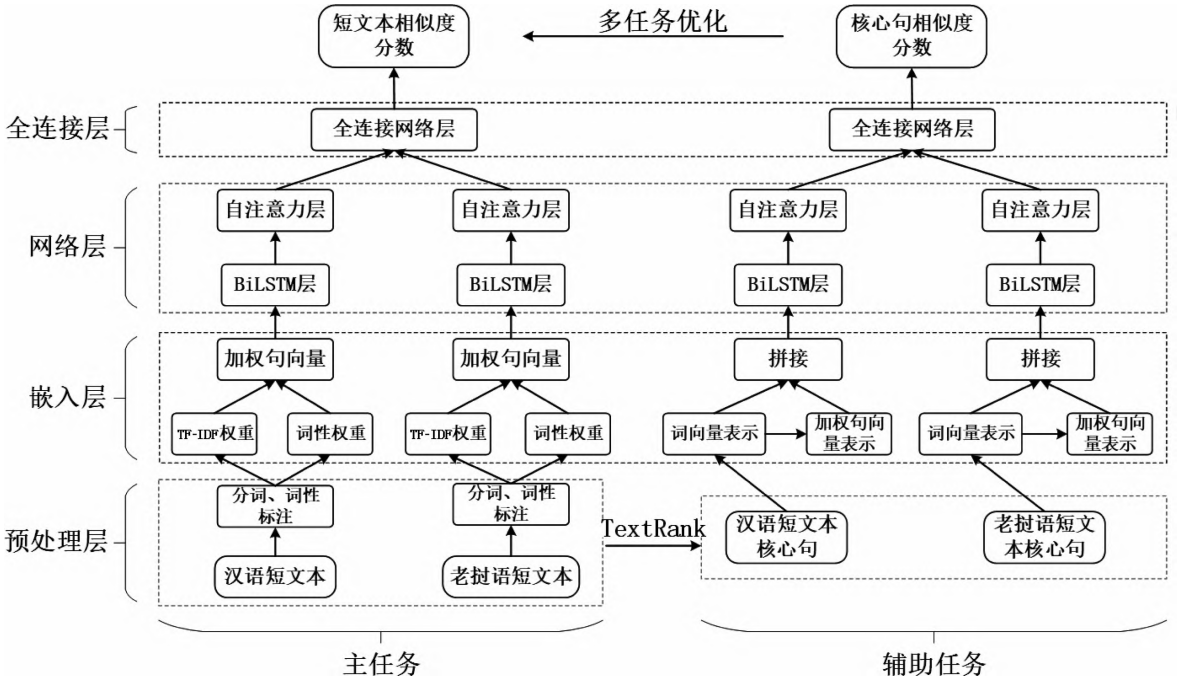

图 1　融合词性位置特征的多任务汉老双语短文本相似度计算模型结构图

如图 1 所示，本文模型分为主任务模型和辅助任务模型，模型由以下 4 个模块组成：

(1) 预处理层：对于主任务，预处理层是对输入的汉老双语短文本进行分词和词性标注处理；对于辅助任务，预处理层是通过改进后的 TextRank 算法对预处理后的双语短文本提取其对应的核心句。

(2) 嵌入层：对于主任务，嵌入层是结合汉老双语短文本的词性位置权重和 TF-IDF 权重，对双语短文本进行分布式表示；对于辅助任务，嵌入层是通过双语短文本核心句的词性特征权重加权得到句向量后，将其作为全局向量与核心句的分布式表示拼接，得到双语核心句的最终分布式表示。

(3) 网络层：网络层由带有自注意力机制的双向长短时记忆网络组成，在图 1 中，本文模型的主任务和辅助任务使用的 4 个网络层均为同一个共享参

数的网络层。

(4) 全连接层：对于主任务，全连接层的功能是计算经过网络层得到的汉老双语短文本特征向量间的相对积和相对差，拼接传输到全连接层得到双语短文本的相似度分数；对于辅助任务，通过网络层得到汉老双语短文本核心句特征向量后，计算其相对积和相对差，将结果拼接输入到全连接层，得到核心句的相似度分数，最终通过赋予损失函数权重以优化汉老双语短文本相似度计算模型的性能。

3.2 预处理层

3.2.1 分词及词性标注

在本文模型中，要计算汉老双语短文本的相似度，首先要对双语短文本进行分词和词性标注。汉语短文本采用 Stanford 开发的 CoreNLP 工具[①]进行分词和词性标注预处理，老挝语短文本采用实验室开发的老挝语分词工具[21]和词性标记工具[23]进行处理，标注完成后对词性标记进行统一化处理。

3.2.2 改进 TextRank 算法获取核心句

TextRank 算法是一种用于文本的基于图的排序算法，通过将汉老双语短文本分割为若干个句子并建立图模型，分别计算各个句子的权重作为分数，最终经过排序即可得到分数最高的句子，即汉老双语短文本分别对应的核心句。

对于给定汉老双语短文本 T_c 和 T_l，其中任意的短文本 $T=\{S_1,S_2,\cdots,S_n\}$，可以将其表示为有向有权图 $G=(S,E)$，其中 S 是短文本的句子集合，E 是图中的边集，ω_{ji} 为任意两个句子 S_i、S_j 之间边的权重。对于任意给定的句子 S_i，$\mathrm{In}(S_i)$ 为指向该句子的句子集合，$\mathrm{Out}(S_i)$ 为句子 S_i 指向的句子集合，则句子 S_i 的权重分数 W 定义为：

$$W(s_i)=(1-d)+d\times\sum_{S_j\in \mathrm{In}(S_i)}\frac{\omega_{ij}}{\sum_{S_k\in \mathrm{Out}(S_j)}\omega_{jk}}W(S_j) \tag{1}$$

其中，d 为阻尼系数，通常取值为 0.85，ω_{jk} 为句子间的权重，即通过余弦相似度计算得到句子间的相似度分数。通过迭代式(1)即可得到短文本中所有句子的得分，得分最高者为短文本核心句。

目前，针对 TextRank 算法的改进取得了一定的成果[25-27]，由式(1)可知，改进 ω 参数的计算方法可以有效提升计算句子权重分数的准确性，因此，在前人方法的基础上，本文结合神经网络模型对 TextRank 算法中计算句子相似度的方法进行改进，使用带有自注意力机制的双向长短时记忆网络来计算句子间的相似度，通过得到更加准确的相似度参数来提高算法计算核心句的准确性。由于计算句子相似度使用的模型结构框架与本文计算短文本相似度的模型框架相同，因此此处不做赘述，详细的模型结构将在 3.4 节、3.5 节介绍。

3.3 嵌入层

本文模型的嵌入层分为短文本嵌入层和核心句嵌入层，分别对应模型的主任务和辅助任务。

3.3.1 短文本嵌入层

短文本通常由多个句子组成，为了得到整个短文本的语义表征，需要对短文本中的句子进行词向量加权得到句向量。由于老挝语的语料资源稀缺，且基础研究较为薄弱，为了得到具有更多语义信息的短文本分布式表示，本文针对汉语和老挝语的语言特点，根据单词在短文本句子中所处的位置，获取带有位置信息的汉老双语词性特征。根据位置特征，对双语短文本中名词、代词、动词、形容词、副词和其他词性设置了相应的权重，通过关键词性获取更多语义信息的同时降低其他词性的干扰。

对于长度为 L 的句子，将其均匀划分后，对处于句子$[0,L/3]$部分的词，赋予名词、代词更高的权重以获取主语信息；对处于句子$(L/3,2L/3]$部分的词，赋予动词、形容词更高的权重以获取谓语信息；对处于句子$(2L/3,L)$部分的词，赋予名词、代词和动词更高的权重以获取句子的谓语和宾语信息。设置的词性位置权重表如表 2 所示。

表 2 词性位置权重表

	词性位置	名词	代词	动词	形容词	副词	其他词性
词性权重	$[0,L/3]$	0.75	0.75	0.5	0.5	0.1	0.01
	$(L/3,2L/3]$	0.65	0.65	0.6	0.6	0.1	0.01
	$(2L/3,L)$	0.7	0.7	0.5	0.5	0.1	0.01

给定汉老双语短文本 T_c 和 T_l，任意的双语短文本 $T=\{S_1,S_2,\cdots,S_n\}$，其中第 i 个句子 $S_i=\{w_1,w_2,\cdots,w_m\}$，则通过句子中的第 j 个单词 w_j 所在的位置 $q_j=j/L$，对应表 2 即可得到其词性位置权重 p_j，最终可得句子 S_i 对应的词性位置权重系数序列 $P_i=\{p_1,p_2,\cdots,p_m\}$。

在得到汉老双语短文本词性位置权重的基础

① http://corenlp.run/

上，与徐鑫鑫等人[11]提出的多加权权重方法类似，通过计算双语短文本中每个句子的 TF-IDF 权重系数，与词性位置权重加权表征短文本。对于任意的短文本 T，句子 S_i 中的第 j 个词语 w_j 的 TF-IDF 系数计算如式(2)～式(4)所示。

$$f_j = \frac{N(j \mid T)}{N(j)} \tag{2}$$

$$\mathrm{idf}_j = \log \frac{D}{N(T \mid j)+1} \tag{3}$$

$$\mathrm{tf}_j = f_j \times \mathrm{idf}_j \tag{4}$$

其中 f_j 为词语 w_j 在短文本 T 中的词频，idf_j 为词语 w_j 的逆文档频率，D 为总语料的短文本数，$N(j \mid T)$ 为词语 w_j 在短文本 T 中出现的次数，$N(j)$ 为词语 w_j 在总语料中出现的次数，$N(T \mid j)$ 为包含词语 w_j 的短文本数。将计算得到的词频和逆文档频率相乘即可得到词语 w_j 的 TF-IDF 权重系数 tf_j，最终可得句子 S_i 对应的 TF-IDF 权重系数序列 $\mathrm{TF}_i = \{\mathrm{tf}_1, \mathrm{tf}_2, \cdots, \mathrm{tf}_m\}$。

通过以上方法得到短文本 T 的 TF-IDF 权重系数和词性位置权重系数后，将其相乘作为本文最终加权句向量的权重系数 Y，如式(5)所示。

$$Y_i = P_i \times \mathrm{TF}_i \tag{5}$$

得到短文本 T 中每个句子的加权权重系数后，即可对句子进行词向量加权得到句向量。针对加权句向量时使用的双语词向量，本文在单独训练好汉语和老挝语词向量的基础上，采用 Artetxe 等人[28]的方法将汉语和老挝语词向量映射到共享的语义空间，得到汉老双语词向量，该方法可以有效减少汉老语言间的差异性。最终经过加权词向量，可得汉老双语短文本 T_c 和 T_l 的短文本分布式表示 $V_{\mathrm{tc}} = \{\mathrm{v}_{tc1}, \mathrm{v}_{tc2}, \cdots, \mathrm{v}_{tcn}\}$ 和 $\mathrm{V}_{tl} = \{\mathrm{v}_{tl1}, \mathrm{v}_{tl2}, \cdots, \mathrm{v}_{tln}\}$。

3.3.2　核心句嵌入层

针对汉老双语短文本对应的核心句，本文在使用 3.3.1 节预训练好的汉老双语词向量对其进行分布式表示的基础上，借鉴李霞等人[29]的思想，通过获取核心句的词性位置权重加权得到核心句的句向量后，将其拼接在核心句分布式表示的末尾，融入句子的全局语义信息以丰富句子的语义表示。

对于汉老双语短文本 T_c 和 T_l 对应的核心句 S_c 和 S_l，其中任意的核心句 $\mathrm{S} = \{\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_m\}$ 对应的词向量分布式表示为 $\mathrm{V} = \{\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_m\}$，通过获取句子的词性位置权重序列 $\mathrm{P} = \{\mathrm{p}_1, \mathrm{p}_2, \cdots, \mathrm{p}_m\}$ 后，加权得到句子的句向量表示 v^*，将其拼接在句子词向量分布式表示 V 的末尾以得到含有全局语义信息的核心句分布式表示 $\mathrm{V}^* = \{\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_m, \mathrm{v}^*\}$。最终双语短文本的核心句 S_c 和 S_l 的对应的分布式表示为 $\mathrm{V}_{sc}^* = \{\mathrm{v}_{sc1}, \mathrm{v}_{sc2}, \cdots, \mathrm{v}_{scm}, \mathrm{v}_{sc}^*\}$ 和 $\mathrm{V}_{sl}^* = \{\mathrm{v}_{sl1}, \mathrm{v}_{sl2}, \cdots, \mathrm{v}_{slm}, \mathrm{v}_{\mathrm{sl}}^*\}$。

3.4　网络层

通过嵌入层得到短文本与其对应的核心句分布式表示后，即可输入模型的网络层获取特征向量。模型的网络层由 BiLSTM 层和自注意力层构成，该网络结构与 3.2.2 节中改进后的 TextRank 算法在计算句子相似度时使用的网络结构相同。

3.4.1　BiLSTM 层

BiLSTM 网络由一个前向 LSTM 网络和一个后向 LSTM 网络组成，它可以有效获取含有上下文语义信息的特征向量，BiLSTM 网络的结构如图 2 所示。

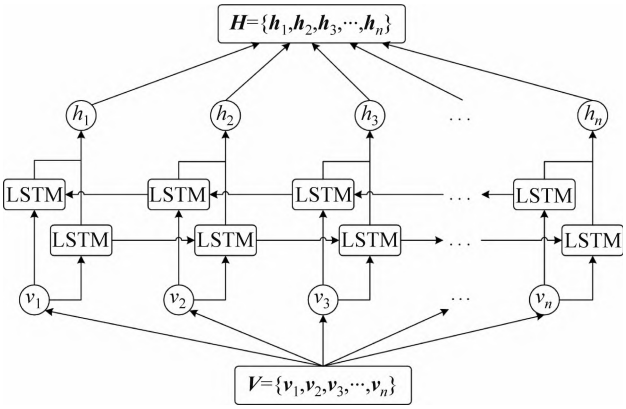

图 2　BiLSTM 网络的结构

对于给定的汉老双语短文本 T_c 和 T_l，经过预处理层和嵌入层后，分别得到汉语短文本和核心句的分布式表示 $V_{\mathrm{tc}} = \{v_{\mathrm{tc1}}, v_{\mathrm{tc2}}, \cdots, v_{\mathrm{tcn}}\}$，$V_{\mathrm{sc}}^* = \{v_{\mathrm{sc1}}, v_{\mathrm{sc2}}, \cdots, v_{\mathrm{scm}}, v_{\mathrm{sc}}^*\}$，以及老挝语短文本和核心句的分布式表示 $V_{\mathrm{tl}} = \{v_{\mathrm{tl1}}, v_{\mathrm{tl2}}, \cdots, v_{\mathrm{tln}}\}$，$V_{\mathrm{sl}}^* = \{v_{\mathrm{sl1}}, v_{\mathrm{sl2}}, \cdots, v_{\mathrm{slm}}, v_{\mathrm{sl}}^*\}$。将双语短文本的分布式表示 V_{tc} 和 V_{tl} 输入到 BiLSTM 网络中，分别得到前向 LSTM 的输出值 $\overrightarrow{H_{\mathrm{tc}}} = (\overrightarrow{h_{\mathrm{tc1}}}, \overrightarrow{h_{\mathrm{tc2}}}, \cdots, \overrightarrow{h_{\mathrm{tcn}}})$ 和 $\overrightarrow{H_{\mathrm{tl}}}$，与后向 LSTM 的输出值 $\overleftarrow{H_{\mathrm{tc}}} = (\overleftarrow{h_{\mathrm{tc1}}}, \overleftarrow{h_{\mathrm{tc2}}}, \cdots, \overleftarrow{h_{\mathrm{tcn}}})$ 和 $\overleftarrow{H_{\mathrm{tl}}}$ 后，按位置拼接即可分别得到汉老双语短文本的语义特征向量 $\boldsymbol{H}_{\mathrm{tc}}$ 和 $\boldsymbol{H}_{\mathrm{tl}}$，拼接如式(6)、式(7)所示。

$$\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n] \tag{6}$$

$$\boldsymbol{h}_i = [\overrightarrow{h_i}; \overleftarrow{h_i}] \in R^d \tag{7}$$

同理，将汉老双语核心句的分布式表示 V_{sc}^* 和 V_{sl}^* 输入到 BiLSTM 网络中，即可得到双语核心句

的语义特征向量 $\boldsymbol{H}_{sc}$ 和 $\boldsymbol{H}_{sl}$。

3.4.2 自注意力层

自注意力机制可以有效捕获输入序列的长距离依赖关系和学习序列的结构特点，本文处理的对象均为序列数据，使用自注意力机制可以有效获取更多的语义信息。与李卫疆等人[20]的做法类似，本文使用的自注意力层含有两层线性网络层 L_1 和 L_2，将 BiLSTM 网络层得到的汉老双语短文本特征向量 $\boldsymbol{H}_{tc}$，$\boldsymbol{H}_{tl}$ 输入到自注意力层，经过 L_1 和 L_2 网络层后可以分别得到特征向量对应的特征权重分数 a_{tc} 和 a_{tl}，将其与特征向量加权求和，即可得到含有长距离语义信息的汉老双语短文本特征向量 $\boldsymbol{C}_{tc}$ 和 $\boldsymbol{C}_{tl}$，自注意力层的计算如式(8)、式(9)所示。

$$\boldsymbol{a}=\mathrm{softmax}(\boldsymbol{\sigma}_{l2}\tanh(\boldsymbol{\sigma}_{l1}\boldsymbol{H})) \tag{8}$$

$$\boldsymbol{C}=\boldsymbol{a}\oplus\boldsymbol{H} \tag{9}$$

其中，$\boldsymbol{\sigma}_{l1}$，$\boldsymbol{\sigma}_{l2}$ 为线性网络层 L_1，L_2 学习得到的权重矩阵。同理可得汉老短文本核心句 $\boldsymbol{H}_{sc}$ 和 $\boldsymbol{H}_{sl}$ 对应的特征向量 $\boldsymbol{C}_{sc}$ 和 $\boldsymbol{C}_{sl}$。

3.5 多任务学习相似度表示

通过网络层得到汉老双语短文本和核心句的特征向量表示后，计算双语短文本特征向量间的相对积与相对差，将结果拼接输入到全连接网络层即可得到对应的相似度分数。双语短文本核心句的特征向量采用同样的方法进行处理。对经过网络层得到的双语短文本特征向量 $\boldsymbol{C}_{tc}$ 和 $\boldsymbol{C}_{tl}$，通过全连接网络层计算短文本间的相似度分数 g_t 如式(10)～式(13)所示。

$$G_1=\boldsymbol{C}_{tc}\oplus\boldsymbol{C}_{tl} \tag{10}$$

$$G_2=\boldsymbol{C}_{tc}\otimes\boldsymbol{C}_{tl} \tag{11}$$

$$G=\tanh(W^1G_1+W^2G_2+b) \tag{12}$$

$$g_t(y\mid G)=\mathrm{sigmoid}(WG+c) \tag{13}$$

其中，W^1，W^2，b，c 为模型参数，g_t 为介于 0 至 1 之间的相似度分数。计算短文本相似度时采用交叉熵(cross entropy)作为目标函数，如式(14)所示。

$$L_t=y\log(g_t)+(1-y)\log(1-g_t) \tag{14}$$

同理可得汉老双语核心句特征向量 $\boldsymbol{C}_{sc}$ 和 $\boldsymbol{C}_{sl}$ 经过全连接网络层计算得到的核心句相似度分数 g_s，计算核心句时使用的目标函数如式(15)所示。

$$L_s=y\log(g_s)+(1-y)\log(1-g_s) \tag{15}$$

在分别计算短文本相似度和核心句相似度时，由于共享网络层，因此主任务和辅助任务在计算相似度时获取的双语语义特征会以参数的形式进行共享。通过控制辅助任务共享的语义特征的重要性，使主任务得到更有效的语义信息，提高模型计算汉老双语短文本相似度的准确性与泛化能力。模型的总损失函数为 $L=L_t+\lambda L_s$，λ 为控制辅助任务信息重要性的参数。

4 实验及分析

4.1 实验设置及评价指标

4.1.1 实验数据与模型设置

本文使用的数据集分为两部分，其中第一部分为用于预训练双语词向量的数据集，汉语部分采用 Li 等人[30]预训练好的词向量(1.69 GB)，老挝语词向量是由老挝语维基百科①爬取的 140.56 MB 语料训练获得，在得到单语词向量的基础上，使用 Artetxe 等人[28]的方法将汉老单语词向量映射到共享语义空间得到高质量的双语词向量，词向量维度设置为 300 维；第二部分为训练模型使用的数据集，数据集来源为通过对中文维基百科②、老挝语维基百科、CRI 悦生活微信公众号爬取的篇章级对齐语料，在经过人工校对后最终得到 3 213 篇汉老双语平行对齐短文本，经过分词预处理后的短文本平均长度为 73 个单词。由于训练模型的语料较少，与 Grégoire 等人[31]的方法类似，本文以每对平行短文本的负样本数为 7 的比例来构建非平行语料库，最终得到 22 491 对非平行短文本，如表 3 所示。

表 3 汉老双语短文本数据集

数据集	双语短文本对数量
汉老对齐短文本数据集	3 213
汉老非对齐短文本数据集	22 491

实验在固定随机种子数的情况下采用 10 折交叉验证来训练模型，对实验结果取均值作为最终结果。每次训练使用的数据集划分结果如表 4 所示。

表 4 训练模型数据集划分结果

数据集	平行短文本数据集	非平行短文本数据集
训练集	2 892	20 242
测试集	321	2 249

模型采用 Python 语言及 Keras 框架来构建，优化器使用 Adam 算法，模型参数设置如表 5 所示。

① https://lo.wikipedia.org/wiki/ຫນ້າຫຼັກ

② https://wiki.hk.wjbk.site/wiki

表 5　模型参数设置

参数	值
Batch size	64
Learning rate	0.000 1
Dropout	0.3
LSTM dim	50
Dense dim	30
λ	0.2

4.1.2　评价指标

本文采用准确率 P(Precision)、召回率 R(Recall)、F_1 值三个参数来对模型进行评估，精确率可以有效评估模型的识别能力，召回率可以有效评估模型的查全能力，F_1 值可以体现模型的综合性能。准确率 P、召回率 R、F_1 值的具体计算如式(16)～式(18)所示。

$$P=\frac{\text{正确分类平行短文本的数量}}{\text{相似的短文本数量}} \tag{16}$$

$$R=\frac{\text{正确分类平行短文本的数量}}{\text{平行短文本的数量}} \tag{17}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\% \tag{18}$$

4.2　辅助任务损失函数权重设置

本文模型使用了多任务学习框架来获取更多的语义信息以提高汉老双语短文本相似度计算的准确性，为了探究辅助任务损失函数的权重 λ 对模型性能的影响，在控制其他参数的情况下进行了实验，实验结果如图 3 所示。通过实验结果可知，当本文辅助任务的损失函数权重 $\lambda=0.2$ 时，模型具有最佳的性能。

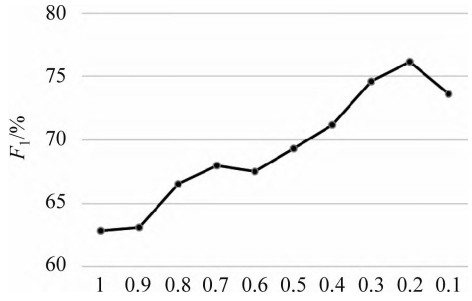

图 3　权重对模型性能的影响结果

4.3　模型对比实验

本文的模型是在常规 BiLSTM＋Self-Attention 模型的基础上，加入了汉老双语短文本词性位置特征权重与 TF-IDF 权重，同时使用多任务学习方法来获取更丰富的语义信息，提升模型计算双语短文本相似度的性能。为了探索本文使用方法对模型的影响，在不同的设置下训练了以下 6 个模型：

(1) BiLSTM＋Self-Attention 模型，作为本实验的基准模型(Base Model)，短文本的分布式表示方式采用加和平均词向量。

(2) 在基准模型的基础上采用汉老双语词性位置特征权重(Pos _weight)来加权表征短文本分布式表示。

(3) 在基准模型的基础上采用 TF-IDF 权重(Tfidf _weight)来加权表征短文本分布式表示。

(4) 在基准模型的基础上同时采用汉老双语词性特征权重和 TF-IDF 权重来加权表征短文本分布式表示。

(5) 在基准模型的基础上加入多任务学习方法(Multi _learning)，通过计算短文本的核心句相似度作为辅助任务。

(6) 在基准模型的基础上采用汉老双语词性特征权重和 TF-IDF 权重来加权表征短文本分布式表示，同时采用多任务学习方法，加入计算短文本核心句相似度的辅助任务，该方法即本文方法(Our)。

以上 6 个模型可以探索本文方法对模型的影响。在此基础上，为了验证本文方法的有效性，与另外 3 个具有代表性的短文本相似度计算模型进行了对比，分别为 Yao 等人[12]提出的 LSTM 编码器模型、Kenter 等人[13]提出的加权语义网络模型及郭浩等人[15]提出的 CNN＋BiLSTM 网络模型。以上共计 9 个模型均在同一语料下采用 10 折交叉验证方法进行实验，最终实验结果如表 6 所示。

表 6　不同模型性能对比结果　　(单位：%)

编号	本文模型	P	R	F_1
(1)	Base Model	71.56	71.12	71.34
(2)	＋Pos _weight	73.07	73.37	73.22
(3)	＋Tfidf _weight	72.23	71.73	71.98
(4)	＋Pos _weight＋tfidf _weight	74.10	73.65	73.65
(5)	＋Multi _learning	73.23	73.91	73.57
(6)	＋Pos _weight＋tfidf _weight ＋Multi _learning(Our)	76.84	75.49	76.16
对比模型		P	R	F_1
LSTM 编码器模型		75.16	74.52	74.84
加权语义网络模型		74.66	73.61	74.13
CNN＋BiLSTM 网络模型		73.11	71.61	72.35

由表 6 实验结果可知，采用权重加权的方法来表征短文本分布式表示，可以有效提升模型的性能。与基准模型相比，使用汉老双语词性位置特征权重加权的模型(2)的 F_1 值提升了 1.88%，说明通过本文提出的获取汉老双语词性位置特征权重的方法，可以得到更多的语义信息，有效提升模型的性能；相比之下，使用 TF-IDF 权重加权的模型(3)的 F_1 值仅提升了 0.64%，原因是短文本的特征词较为稀疏，通过计算短文本的 TF-IDF 权重获取的语义信息有限，但与基线模型使用的加和平均的方法相比，依旧可以获取更多的汉老双语短文本语义信息，模型(4)在同时使用汉老双语词性位置特征权重和 TF-IDF 特征权重加权时，与基准模型相比其 F_1 值提升了 2.31%，证明这一方法是有效的；此外，通过模型(5)的实验结果，可以有效说明多任务学习方法能获取更多的语义信息来提高模型的性能，加入多任务学习方法使基准模型的 F_1 值提升了 2.23%；最终，在基准模型的基础上通过本文的方法，模型的 F_1 值提升了 4.82%，说明本文方法可以有效提高汉老双语短文本相似度计算的准确性。

另一方面，再与前人方法做相比，LSTM 编码器模型的 F_1 值与本文模型相比低了 1.32%，这是由于 LSTM 编码器模型在对短文本分布式表示时，使用的方法为对词向量归一化处理，同时，该方法计算短文本相似度的结果更多地依赖于网络结构获取到的文本特征，因此在面对语言差异性较大的汉语和老挝语时，模型效果不是很好；加权语义网络模型与本文模型相比，F_1 值低了 2.03%，原因是加权语义网络模型在衡量短文本相似度时，是由多个维度的词向量来表征短文本的语义特征，以此来衡量短文本间的相似度，汉语和老挝语短文本的结构具有一定的相似性，与该方法相比本文方法可以挖掘出更多的语义信息，因此本文模型效果更优；在对比模型中，CNN+BiLSTM 网络模型与本文模型相比差距最大，F_1 值低了 3.81%，这是由于本文处理的汉老双语短文本平均长度较长，需要表征为句向量的形式减少语义信息的丢失，而 CNN+BiLSTM 网络模型仅使用了词向量来表示分布式短文本，在处理长度较长的短文本时会损失一部分语义信息，因此结果较差。

总之，由于汉老双语短文本资源较为稀缺，采用多任务学习方法和融入语言特征可以获取更多的语义信息，有效提升汉老双语短文本相似度计算模型的性能。

4.4 TextRank 算法改进对比实验

本文对传统 TextRank 算法进行了改进以获取更加准确的短文本核心句，作为多任务学习方法的辅助模型输入。为了验证本文改进 TextRank 算法的有效性，在与传统 TextRank 算法进行对比的同时，本节与目前具有代表性的两个改进方法进行了对比，分别为王子璇等人[26]提出的融入 WMD 语义相似度的 TextRank 算法和徐馨韬等人[27]提出的结合 Doc2Vec 和 K-means 的 DK-TextRank 算法，通过这两种算法获取汉老双语短文本的核心句后，使用本文模型计算汉老双语短文本的相似度。实验采用 10 折交叉验证方法，结果如表 7 所示。

表 7 不同 TextRank 算法对模型性能的影响

(单位：%)

模型	P	R	F_1
TextRank	74.03	71.57	72.78
TextRank+WMD	76.59	75.66	76.12
DK-TextRank	76.66	75.41	76.03
Our	76.84	75.49	76.16

由表 7 可知，针对汉老双语短文本相似度计算任务，使用 TextRank+WMD 算法的模型与本文方法差距最小，F_1 值仅低了 0.04%，原因是短文本的特征词稀疏，句子间联系不紧密，相比于 WMD 距离方法，用神经网络模型来计算相似度可以得到更好的结果，但 TextRank+WMD 算法在改进句子相似度计算的同时，还加入了位置特征、核心词特征和类别特征，因此取得了较好的结果；使用 DK-TextRank算法的模型与本文方法相比 F_1 值低了 0.13%，这是由于短文本中句子数量较少，在该方法中采用 K-means 聚类难以得到较好的结果，因此对性能有一定的影响；与传统 TextRank 算法相比，本文方法的 F_1 值高了 3.38%，说明本文改进 TextRank 的方法是有效的。

综上所述，本文通过改进 TextRank 算法的句子相似度计算方法，使用带有自注意力机制的双向长短时记忆网络来计算句子间的相似度值，可以有效提高 TextRank 算法计算句子权重分数的准确率，进一步提升模型的性能。

5 结论

本文根据汉语和老挝语短文本词性及位置的特

点，提出一种融合词性位置特征的多任务汉老双语短文本相似度计算方法，在获取汉老双语短文本词性位置特征权重的基础上加入 TF-IDF 权重共同表征短文本分布式表示，同时对 TextRank 算法改进后以获取汉老双语短文本的核心句，通过分别计算双语短文本的相似度和双语核心句的相似度，使用多任务学习思想将双语核心句相似度计算作为辅助任务共享更多的语义信息，有效提高计算汉老双语短文本相似度模型的性能。实验结果表明，与当前主流方法相比，本文提出的方法在语料资源有限的情况下取得了更优的结果，F_1 值达到了 76.16%。下一步将利用本文方法对汉老双语短文本的聚类进行研究。

参考文献

[1] SAHAMI M, TIMOTHY D H. A web-based kernel unction for measuring the similarity of short text snippets[C]//Proceedings of the 15th International Conference on World Wide Web, 2006: 377-386.

[2] 刘令强. 短文本相似度的关键技术研究[D]. 南宁：广西师范大学硕士学位论文，2016.

[3] 翟延冬，王康平，张东娜，等. 一种基于 WordNet 的短文本语义相似性算法[J]. 电子学报，2012，40(3)：617-620.

[4] ATOUM I, OTOOM A. Efficient hybrid semantic text similarity using WordNet and a corpus[J]. International Journal of Advanced Computer Ence & Applications, 2016, 7(9):124-130.

[5] 霍文杰. 汉老双语文本及句子相似度计算研究[D]. 昆明：昆明理工大学硕士学位论文，2018.

[6] 王振振，何明，杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学，2013，40(012)：229-232.

[7] HE X B, ZHONG N, CHEN J H. Weighted-LDA-TVM: Using a weighted topic vector model for measuring short text similarity[C]//Proceedings of the International Conference on Brain Informatics. Springer, Cham, 2019: 212-219.

[8] 张小川，余林峰，张宜浩. 基于 LDA 的多特征融合的短文本相似度计算[J]. 计算机科学，2018，45(9)：266-270.

[9] 赵谦，荆琪，李爱萍，等. 一种基于语义与句法结构的短文本相似度计算方法[J].计算机工程与科学，2018，40(07)：1287-1294.

[10] 詹志建，杨小平. 一种基于复杂网络的短文本语义相似度计算[J]. 中文信息学报，2016，30(4)：71-80.

[11] 徐鑫鑫，刘彦隆，宋明. 利用加权词句向量的文本相似度计算方法[J]. 小型微型计算机系统，2019，40(10).

[12] YAO L, PAN Z, NING H. Unlabeled short text similarity with LSTM encoder[J]. IEEE Access, 2018, 7: 3430-3437.

[13] KENTER, T, MAARTEN DE R. Short text similarity with word embeddings[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015: 1411-1420.

[14] BOOM C D, CANNEYT S V, BOHEZ S, et al. Learning semantic similarity for very short texts [C]//Proceedings of the 15th IEEE International Conference on Data Mining Workshops, 2016: 1229-1234.

[15] 郭浩，许伟，卢凯，等. 基于 CNN 和 BiLSTM 的短文本相似度计算方法[J]. 信息技术与网络安全，2019，38(6)：61-64,68.

[16] 吴建明. 印欧语与汉语的非人称句对比研究[J]. 长沙大学学报，2013，27(004)：112-114.

[17] 林立. 汉语复句与印欧语复句的异同[J]. 山东外语教学，1986(01)：37-42.

[18] 翁帕松.汉语和老挝语句法比较研究[J].汉字文化，2018,(3)：13-15.

[19] 罗芳玲. 汉语和老挝语主谓宾成分的特点及比较[J]. 出国与就业(就业版)，2011，000(016)：220-221.

[20] 李卫疆，李涛，漆芳. 基于多特征自注意力 BiLSTM 的中文实体关系抽取[J]. 中文信息学报，2019，33(10)：47-56,72.

[21] 何力，周兰江，周枫，等. 基于双向长短期记忆神经网络的老挝语分词方法[J]. 计算机工程与科学，2019,41(7)：1312-1317.

[22] 王兴金，周兰江，张金鹏，等. 融合词预测的半监督老挝语词性标注研究[J]. 小型微型计算机系统，2019,40(12)：2500-2505.

[23] 王兴金，周兰江，张建安，等. 融合词结构特征的多任务老挝语词性标注方法[J]. 中文信息学报，2019，33(11)：39-45.

[24] LUONG M T, LE Q V, SUTSKEVER I, et al. Multi-task sequence to sequence learning[C]//Proceedings of IcLR, 2017: 1-10.

[25] 张莉婧，李业丽，曾庆涛，等.基于改进 TextRank 的关键词抽取算法[J].北京印刷学院学报，2016，24(4)：51-55.

[26] 王子璇，乐小虬，何远标. 基于 WMD 语义相似度的 TextRank 改进算法识别论文核心主题句研究[J]. 数据分析与知识发现，2017，1(4)：1-8.

[27] 徐馨韬，柴小丽，谢彬，等. 基于改进 TextRank 算法的中文文本摘要提取[J]. 计算机工程，2019，045(003)：273-277.

[28] ARTETXE M, LABAKA G, AGIRRE E. Learning bilingual word embeddings with (almost) no bilingual data[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 451-462.

（下转第 33 页）

[2] 金永寿. 汉朝翻译理论研究现状与今后的研究方向[J]. 中国朝鲜语文，2020(6)：66-73.
[3] LI J，SUN A，HAN J，et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering，2020，34(1)：50-70.
[4] 何玉洁，杜方，史英杰，等. 基于深度学习的命名实体识别研究综述[J]. 计算机工程与应用，2021，57(11)：21-36.
[5] HUANG Z，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv：1508.01991，2015.
[6] MA X，HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics，2016：1064-1074.
[7] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems，2017：6000-6010.
[8] YAN H，DENG B，LI X，et al. TENER：Adapting transformer encoder for named entity recognition[J]. arXiv preprint arXiv：1911.04474，2019.
[9] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the NAACL，2019：4171-4186.
[10] 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.
[11] 赵丹丹，黄德根，孟佳娜等. 多头注意力与字词融合的中文命名实体识别[J]. 计算机工程与应用，2022，58(07)：142-149.
[12] 殷章志，李欣子，黄德根，等. 融合字词模型的中文命名实体识别研究[J]. 中文信息学报，2019，33(11)：95-100.
[13] ZHANG Y，YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics，2018：1554-1564.
[14] PARK S，MOON J，KIM S，et al. KLUE：Korean language understanding evaluation[J]. arXiv preprint arXiv：2105.09680，2021.

高君龙(1996—)，硕士，主要研究领域为自然语言处理。
E-mail：2020050072@ybu.edu.cn

赵亚慧(1974—)，通信作者，硕士，教授，主要研究领域为文本分析、智能计算、自然语言处理。
E-mail：903873610@qq.cn

崔荣一(1962—)，博士，教授，主要研究领域为智能计算、模式识别、机器学习、自然语言处理。
E-mail：cuirongyi@ybu.edu.cn

(上接第 27 页)

[29] 李霞，刘承标，章友豪，等. 基于局部和全局语义融合的跨语言句子语义相似度计算模型[J].中文信息学报，2019，33(06)：18-26.
[30] LI S，ZHAO Z，HU R，et al. Analogical reasoning on Chinese morphological and semantic relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics，2018：138-143.
[31] GRÉGOIRE F，LANGLAIS P. Extracting parallel sentences with bi-directional recurrent neural networks to improve machine translation[C]//Proceedings of the 27th International Conference on Computational Linguistics，2018：1442-1453.

李炫达(1995—)，硕士，主要研究领域为自然语言处理。
E-mail：lixuanda0@qq.com

张建安(1972—)，硕士，副教授，主要研究领域为信息安全和机器学习。
E-mail：zjaemail@163.com

周兰江(1964—)，通信作者，硕士，副教授，主要研究领域为信息检索、机器学习和自然语言处理。
E-mail：915090822@qq.com