文章编号：1003-0077(2022)02-0058-11

融合句子结构特征的汉老双语句子相似度计算方法

李炫达，周兰江，张建安

（昆明理工大学 信息工程与自动化学院，云南 昆明 650500）

摘　要：在低资源神经机器翻译中，双语平行句对是重要的数据资源，融合语言结构特点能够较好地解决双语句子由于语言差异性导致的句子相似度计算不准确问题。该文提出一种融合句子结构特征的汉老双语句子相似度计算方法。首先，通过该文提出的特征模板获取汉语和老挝语对应的句子结构特征，预训练含有句子结构特征的汉老双语词向量分布式表示，并使用双语词典将其映射到共享的语义空间，然后通过带有自注意力(self-attention)机制的双向长短时记忆网络(BiLSTM)获取句子的特征向量表示，最后分别计算双语向量的相对差和相对积，将结果拼接后传输到全连接网络层计算出相似度分数。实验结果表明，相比目前主流研究方法，该文方法在有限的语料下取得了更好的效果(F_1 值为 70.24%)。

关键词：汉语-老挝语；资源稀缺型语言；句子结构特征；双向长短期记忆网络；自注意力机制

中图分类号：TP391　　**文献标识码**：A

Sentence Similarity Metirc Between Chinese and Laotian Based on Syntax Feature

LI Xuanda, ZHOU Lanjiang, ZHANG Jian'an

(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: To construct bilingual parallel sentence pairs, this paper proposes a Chinese-Lao sentence similarity metric incorporating syntactic information. Firstly, the corresponding sentence structure of Chinese and Lao are obtained by the template proposed in this article. Secondly, the pre-trained representation of Chinese-Lao bilingual words with syntactic characteristics is mapped to a shared semantic space using a bilingual dictionary. Thirdly, the sentence representation is obtained through a Bi-directional Long Short-Term Memory (BiLSTM) network with a Self-Attention mechanism. Finally, the relative difference and relative product of the bilingual vectors are calculated and transmitted to the fully connected network layer to calculate the similarity score. Experimental results show that compared with the current mainstream research methods, the proposed method has achieved better results with limited corpus (F_1=70.24%).

Keywords: Chinese-Laotian; resource scarce language; sentence structure characteristics; BiLSTM; self-attention mechanism

0　引言

老挝与我国云南接壤，其语言老挝语属汉藏语系，在机器翻译中属于资源稀缺型语言。汉老双语句子相似度计算是指计算汉语和老挝语之间的句子语义相似程度，是抽取汉老双语平行句对的重要方法，在老挝语研究中具有非常重要的地位。

近年来，传统方法和基于神经网络模型的方法在跨语言句子相似度计算任务中均取得了很好的效果，然而目前主流的传统方法如基于双语词典匹配的方法[1-2]、基于特征工程的方法[3-4]等往往需要大规模的语料数据和提取大量的文本特征来表征句子相似度；基于神经网络模型的方法[5-8]虽然可以使用网络结构提取文本特征，通过计算特征向量间的距离来表征句子相似度，但在面对跨度较大的语言时，

收稿日期：2020-08-12　**定稿日期**：2020-09-03

基金项目：国家自然科学基金(61662040)

其使用网络结构提取特征的效果较差，因此对于语言结构差异性较大的语言，大多考虑在神经网络模型中融合传统方法文本特征。

目前已有的工作大多为使用基于特征工程的方法提取文本特征后，将其对应的特征向量与句子的分布式表示进行拼接以融合特征信息[9]，通过神经网络模型表征句子的相似度。老挝语的基础研究薄弱，目前还没有成熟的句法分析工具，难以使用传统方法提取特征，因此本文在研究了汉语和老挝语的句子结构异同后，构建了一种根据关键词性和位置信息来获取老挝语句子结构特征的特征模板，提出一种融合句子结构特征的汉老双语句子相似度计算方法。不同于目前提取特征向量再进行拼接的方法，由于本文特征模板提取句子结构特征需要确定词性和位置信息，因此需要先添加特征标记，再将含有特征标记的句子进行分布式表示，并映射到共享的语义空间，最后通过带有自注意力(self-attention)机制的双向长短时记忆网络(BiLSTM)模型得到汉老双语句子的相似度分数。实验结果表明，与目前主流方法相比，本文方法在有限的语料下具有更优的表现，模型的 F_1 值达到了 70.24%。

本文的主要贡献如下：

(1) 提出一种通过关键词性和位置信息来获取老挝语句子结构特征的特征模板。

(2) 将汉-老双语词嵌入映射到共享的语义空间，减少了汉、老语言间的差异性。

(3) 在 BiLSTM 网络中加入自注意力机制，有效提高跨语言句子相似度计算模型的效果。

本文组织结构如下：引言部分介绍本文的研究背景及目的，第 1 节为相关工作，综述双语句子相似度计算的相关文献；第 2 节介绍汉语和老挝语句子结构的异同；第 3 节介绍本文使用模型的结构；第 4 节为本文模型的设置与相关实验的结果；第 5 节为总结与展望。

1　相关工作

传统的双语句子相似度计算方法主要有以下三类方法。

(1) **基于双语词典匹配的方法**　这类方法的思想是使用双语词典将源语言和目标语言转换为中间层语言，通过计算词的相似度来衡量句子的相似性，如石杰等人[1]使用多语言版本的 WordNet 将汉语和泰语转换为英语，通过转换后文本的特征词匹配来计算相似度；闫红等人[2]通过 HowNet 的多义词消歧对句子中的词语进行处理，以词语相似度为基础计算了句子的相似度。

(2) **基于特征工程的方法**　这类方法的思想是通过抽取文本特征来表示句子的语义信息，从而计算句子间的相似度，如 Tian 等人[3]通过提取句子的序列特征、句法分析特征、句子对齐特征来表示句子语义信息，计算英语、阿拉伯语和西班牙语间的句子语义相似度；黄洪等人[4]利用依存句法分析方法得到句子中各成分的关系特征，以获取句子的核心词和关键词，通过词匹配的方法计算句子相似度。

(3) **基于机器翻译模型的方法**　这类方法的思想是将源语言翻译成目标语言来计算跨语言句子的相似度，如 Erdmann 等人[10]将双语维基百科的文章翻译为同一语言来计算文章的相似度，构建了双语词典；Wu 等人[11]将目标语言翻译为英语后，通过 WordNet 词典中层次树结构的非重叠信息计算了英语、阿拉伯语和西班牙语间的句子语义相似度。

传统方法虽然取得了不错的效果，但基于双语词典匹配的方法仍需要大量的双语词典资源来解决未登录词问题，特征工程的方法需要人工抽取大量的文本特征以保证句子语义信息的准确性，机器翻译模型的方法依赖于翻译的效果。随着深度学习的兴起，基于神经网络模型的跨语言句子相似度计算方法在无需传统特征的基础上取得了较好的结果[12-14]。Mueller 等人[5]提出了一种连体 LSTM 网络结构(Siamese LSTM)，通过将句子对输入到共享参数的 LSTM 网络，得到特征向量后计算向量间的曼哈顿距离表征句子对的相似度；李霞等人[6]分别运用卷积神经网络(convolutional neural network，CNN)和注意力机制(attention mechanism)得到每个句子的局部语义信息和全局语义信息，将其拼接后传输到全连接网络层，计算得到句子间的相似度分数；Chi 等人[7]将改进的连体 LSTM 网络与注意力机制结合，得到更加准确的句子语义向量，通过全连接网络层计算向量间的相对差与相对积来获得句子间的相似性分数。Chien 等人[8]通过学习转换矩阵将训练好的汉语词嵌入映射到英语词嵌入语义空间，然后计算汉语和英语句子的平均逐词相似度，从而获取平行句子对。

2　汉语-老挝语句子结构异同

老挝语的句子构成分为主要成分和次要成分，

主要成分指句子的主谓(或主谓宾)成分;次要成分指解释句子主要成分的附加部分,即定语、状语、补语等。汉语和老挝语的主要成分具有相同的顺序结构,均为主谓宾顺序(SVO),并且汉语和老挝语的主要成分通常由相同词性的单词构成[15],如表 1 所示的例句为经过词性标注和句子主要成分标注处理的句子,其中,/p、/r、/v、/u、/m、/n、/a 分别表示介词、代词、动词、助词、数词、名词和形容词性标记;Subject,Verb,Object 分别表示句子的主语、谓语和宾语。通过表 1 可知,具有完整主谓宾结构的汉老双语句子,其主谓宾在句子中具有相同或相近的位置,并且通常由相近词性的单词来构成主谓宾成分;缺少宾语结构的汉老双语句子,其主语和谓语具有相同或相近的位置,并且同样由相近词性的单词来构成主谓成分。

汉语和老挝语的主语都可以由名词、代词等词性充当,并且在句子中处于相同的位置;谓语由动词、形容词等词性充当,并且谓语都位于主语之后;宾语构成的词类一致,并且都位于谓语之后。因此对于老挝语,可以通过句子中的名词、代词、动词和形容词以及其在句子中对应的位置来识别老挝语句子的主要成分,提取句子的结构特征。

表 1 汉语-老挝语句子结构示例

例句 1:	
汉语:	他/r (Subject) 特别/ad 喜欢/v (Verb) 运动/n (Object)
老挝语:	ລາວ(Subject)ໂດຍສະເພາະ/ad ແມ່ນ/v ມັກ/v (Verb) ກິລາ/n (Object)
例句 2:	
汉语:	芯片/n 价值/n (Subject) 最高/a (Verb)
老挝语:	ຄ່າ/n ຊິ້ງຊຸດ/n (Subject) ຂອງ/p ຊິ/a (Verb)

3 融合句子结构特征的汉老双语句子相似度计算模型

3.1 模型结构

本文构建模型的基本思路如下:首先对汉语和老挝语的平行句对进行分词和词性标注预处理,通过汉语句法分析工具和本文提出的老挝语句子结构特征标记模板分别获取汉、老句子的句子结构特征,加入特征标记;其次,预训练含有特征标记的汉语和老挝语词向量分布式表示,使用双语种子词典将汉老双语词嵌入映射到共享的语义空间,通过带有自注意力机制的双向长短时记忆网络(BiLSTM)获取含有长距离语义信息的双语句子对特征向量表示;最后,分别计算双语特征向量的相对差和相对积,将结果拼接后传输到全连接网络层计算出相似度分数,模型的结构如图 1 所示。

本文模型由以下部分构成:

(1) **预处理层**:对给定的汉语、老挝语双语句子进行分词和词性标注,分别使用 CoreNLP 工具和本文提出的特征模板对汉语和老挝语添加句子结构特征标记。

(2) **词嵌入层**:输入预处理好的具有句子结构特征标记的汉老双语句子对,利用预训练的方式映射在共享语义空间的双语词向量进行转换,得到对应的词向量序列。

(3) **BiLSTM 层**:针对句子训练的问题,是一个典型的序列到序列的问题,BiLSTM 可以较好地捕捉到句子之间的特征[16],将汉老双语句子对应的词向量序列输入到 BiLSTM 网络中,得到含有双向语义信息的特征向量。

(4) **自注意力层**:自注意力层可以有效捕获长距离语义特征[17]。将含有双向语义信息的特征向量传输到自注意力层中,得到含有长距离语义信息的汉老双语句子特征向量。

(5) **全连接层**:将得到的汉老双语句子特征向量分别进行按位减和按位乘操作,把结果进行拼接后传输到全连接网络层中计算得到汉老句子对的相似度分数。

3.2 老挝语句子结构特征标记模板

老挝语是一种缺少语料资源的稀缺语言,由于缺少成熟的句法分析工具,无法直接获取句法特征向量。本文在对汉语和老挝语句子结构进行研究后,发现汉老双语句子成分相似[15],并且具有相同的主谓宾结构(SVO),因此可以通过关键词性和位置信息在原句中添加句子成分标记,获取句子结构

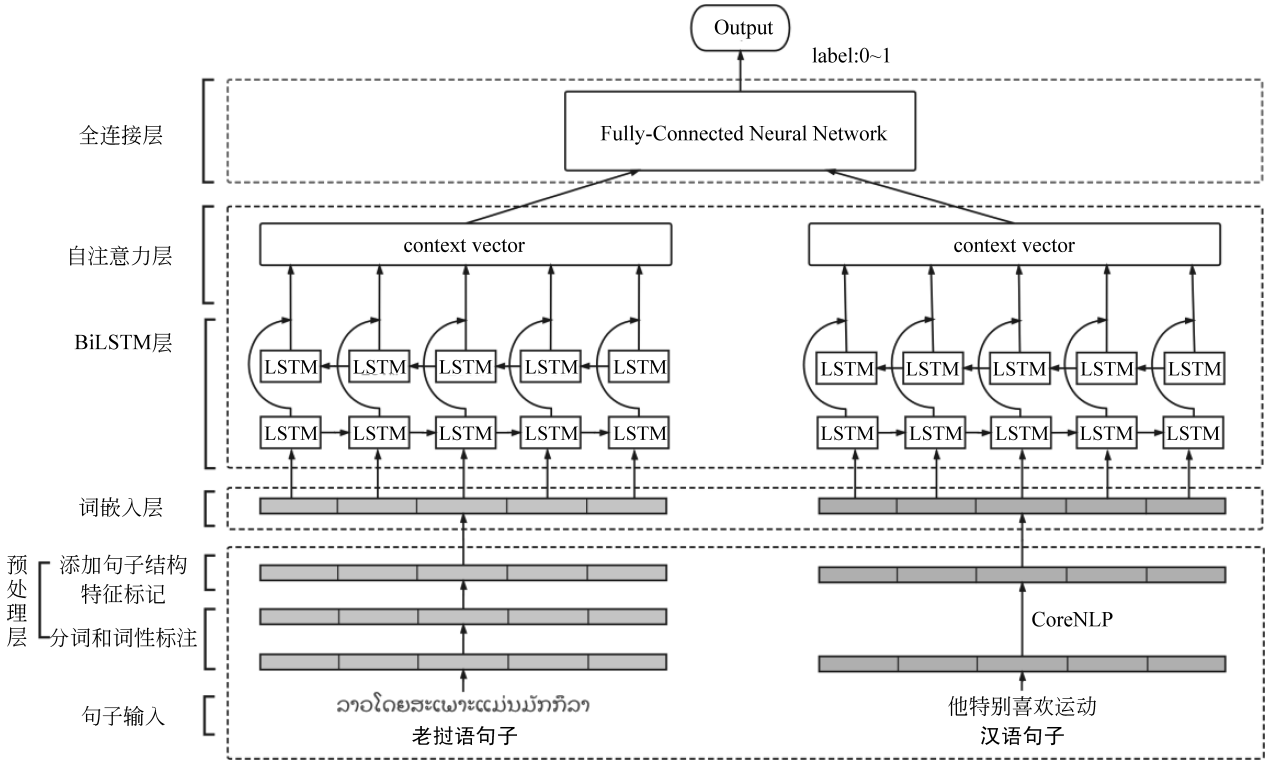

图 1　融合老挝语句子结构特征的汉老双语句子相似度计算模型图

特征。使用实验室开发的老挝语分词工具[18]和词性标注工具[19]对老挝语句子进行处理，保留句子中的名词、动词、形容词和代词词性，按以下规则构建特征标记模板来获取老挝句子结构标记：

（1）若老挝语句子保留的词性中拥有除动词和形容词词性以外的其他词性，则将句子中连续的动词和形容词词性视为一个成分，在末尾添加标记 verb；将老挝语句子中连续的名词、代词词性视为一个成分，在末尾添加标记 func_tag；

（2）若句子仅有一个 verb 标记且具有多个 func_tag 标记，则 verb 前的 func_tag 标记部分为主语成分，替换 func_tag 为 sub 标记；verb 后的 func_tag 为宾语成分，替换为 obj 标记；

（3）若句子仅有一个 verb 标记和一个 func_tag 标记，且 func_tag 位于 verb 之前，则把句子视为缺少宾语的主谓句，func_tag 为主语成分，将其替换为 sub 标记；

（4）不满足以上条件时，句子多为成分不全的简单句或具有从句的复杂句，使用特征标记模板难以获取句子结构特征，因此不做处理。

通过以上特征标记模板，可以对老挝语句子添加主语、谓语和宾语标记。对于汉语句子，本文使用 Stanford 开发的 CoreNLP① 工具对汉语句子进行句法分析，保留主、谓、宾成分，使用相同的标记进行处理。例如，对于例句“ລາວໂດຍສະເພາະແມ່ນມັກກິລາ”（他特别喜欢运动），使用特征标记模板标记老挝语句子的过程如图 2 所示，使用 CoreNLP 工具标记汉语句子的过程如图 3 所示。

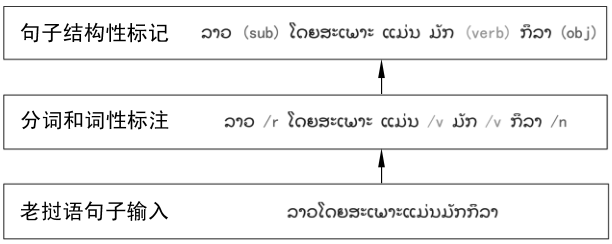

图 2　老挝语句子结构标记过程图

图 2 为使用特征模板对老挝语进行标记的过程，例句“ລາວໂດຍສະເພາະແມ່ນມັກກິລາ”首先经过分词处理，然后进行词性标注，并保留关键词性“/r”“/v”“/n”，最后根据特征模板规则将其替换为(sub)、(verb)、(obj)标记。

图 3 为使用 CoreNLP 对例句“他特别喜欢运动”进行标记的过程，首先经过分词和词性标记处理后，通过 CoreNLP 的句法分析得到句子的主谓宾成

① http://corenlp.run/

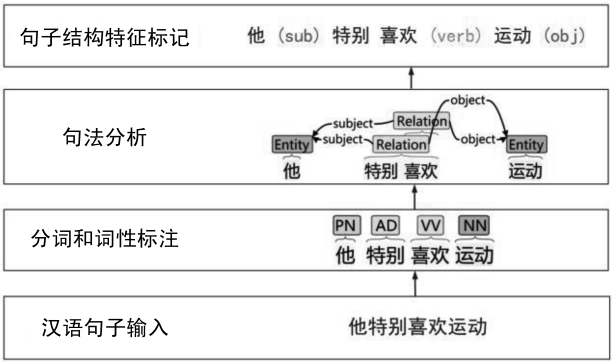

图3 汉语句子结构标记过程图

分,最后将多余句子成分标记去除后,替换为和老挝语相同的标记(sub)、(verb)、(obj)。

通过以上处理,即可在汉语和老挝语句子中加入特征标记。

3.3 含有句子结构特征的汉老双语词向量分布式表示

词向量分布式表示可以将单词映射到低维空间中,不同的维度可以表征不同的语义信息。对于跨度较大的语言,通常将不同语言的词嵌入映射到相同的向量空间中,保证单语言下的语义不变性,同时确保具有相同语义的词非常接近。汉语和老挝语的语言差异性较大,因此在本模型中通过利用汉老双语种子词典映射的方式将汉语和老挝语映射到共享的语义空间。

对于分别预训练好且带有特征标记的汉语和老挝语词嵌入矩阵 $\boldsymbol{S}$、$\boldsymbol{T}$,与 Artetxe[20] 等人的方法类似,引入双语种子词典 M,通过 SVD 以自学习的方式和迭代算法学习线性转换矩阵,得到最佳映射矩阵 $\boldsymbol{W}^*$ 后对汉语词嵌入矩阵进行线性变换得到 $\boldsymbol{S}'$,即可将汉语和老挝语词向量映射在共享的语义空间,如式(1)、式(2)所示。

$$\boldsymbol{S}' = \boldsymbol{S}\boldsymbol{W}^* \tag{1}$$

$$\boldsymbol{W}^* = \arg\min_{W} \sum_i \sum_j M_{ij} \parallel S_i \boldsymbol{W} - T_j \parallel \tag{2}$$

其中,S_i 表示第 i 个汉语的词嵌入,T_j 表示第 j 个老挝语的词嵌入。随机抽取 100 对汉老双语词向量,映射前和映射后的词嵌入在二维空间下的分布如图 4、图 5 所示。

通过以上处理,即可将汉老双语分布式表示映射到共享的语义空间,缩小语言的差异性。

3.4 双向长短时记忆网络(BiLSTM)层

BiLSTM 通过一个正向顺序读取句子的

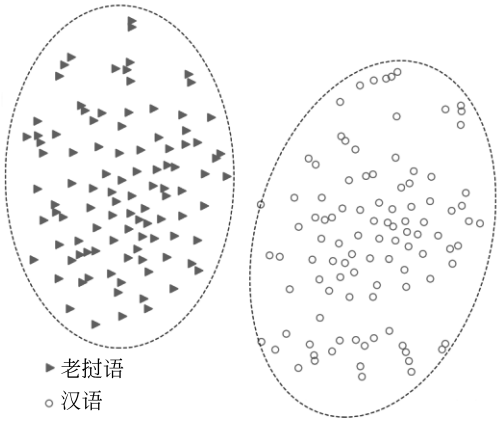

图4 映射前的汉老双语词嵌入图

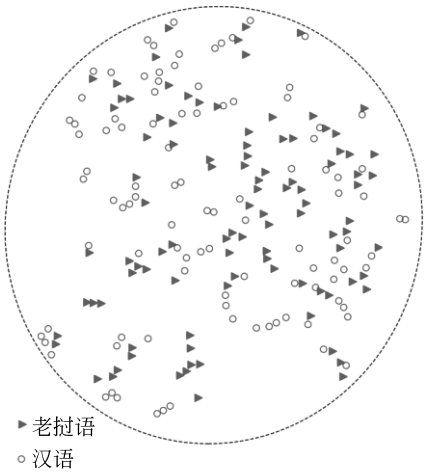

图5 映射后的汉老双语词嵌入图

LSTM 和一个反向顺序读取句子的 LSTM 来分别生成两个隐藏状态,将其拼接得到含有双向信息输出的网络结构。LSTM 的计算如式(3)~式(8)所示。

$$i_t = \text{sigmoid}(W_i x_t + U_i h_{t-1} + b_i) \tag{3}$$

$$f_t = \text{sigmoid}(W_f x_t + U_f h_{t-1} + b_f) \tag{4}$$

$$o_t = \text{sigmoid}(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$

$$u_t = \tanh(W_u x_t + U_u h_{t-1} + b_u) \tag{6}$$

$$c_t = i_t * u_t + f_t * u_{t-1} \tag{7}$$

$$h_t = o_t * \tanh(c_t) \tag{8}$$

其中,i_t 表示 LSTM 的输入门,f_t 表示遗忘门,o_t 表示输出门,h_t 为 LSTM 网络输出的隐藏状态;W_i、W_f、W_o、W_u、U_i、U_f、U_o、U_u是权重数据;b_i,b_f,b_o,b_u为偏置量。

对于给定的汉老句子对 S_1 和 S_2,将句子 S=$\{x_1, x_2, \cdots, x_T\}$中词的向量序列作为 BiLSTM 的输入,通过上述 LSTM 公式最终分别得到前向和后向的隐藏状态,即将汉语句子 S_1 和老挝语句子 S_2 输

入 BiLSTM 网络，分别得到汉语的前向隐藏状态 $\overrightarrow{H}_1=(\overrightarrow{h}_1,\overrightarrow{h}_2,\ldots,\overrightarrow{h}_T)$ 和后向隐藏状态 $\overleftarrow{H}_1=(\overleftarrow{h}_1,\overleftarrow{h}_2,\ldots,\overleftarrow{h}_T)$，老挝语的前向隐藏状态 $\overrightarrow{H}_2$ 和后向隐藏状态 $\overleftarrow{H}_2$，按位置拼接即可得到完整的汉语和老挝语输出序列 H_1 和 H_2，其中 $H=[h_1,h_T]$，隐藏状态拼接如式(9)所示。

$$h_i=[\overrightarrow{h}_i;\overleftarrow{h}_i]\in R^{2L} \tag{9}$$

通过以上处理，即可分别得到含有双向语义信息的汉老双语句子特征向量表示。

3.5　自注意力网络(Self-Attention)层

自注意力层是一般注意力机制(attention)的一种特殊情况[17]，与一般的注意力机制相比，自注意力机制可以无视词之间的距离而直接计算依赖关系，对于捕获句子长距离依赖关系和学习句子内部结构的特点具有更好的效果。本文处理的对象为汉老双语句子对，使用自注意力机制可以得到更加准确的句子特征表示。将 BiLSTM 网络层得到的汉老双语句子输出状态 H_1 和 H_2 分别输入到自注意力层，通过自注意力层学习词和特征的重要性，同时学习句子的序列信息，最终分别得到含有长距离语义信息的汉老双语句子对特征向量。自注意力层的计算如式(10)所示。

$$a=\mathrm{softmax}(w_{l2}\tanh(w_{l1}\boldsymbol{H})) \tag{10}$$

自注意力层的计算过程如图 6 所示，其中，$\boldsymbol{H}$ 表示 BiLSTM 网络层的输出结果，$\boldsymbol{H}\in R^{T\times j}$，$T$ 为句子长度，j 为 LSTM 单元的输出维度，w_{l1} 和 w_{l2} 为自注意力网络层学习得到的权重矩阵。通过将汉语和老挝语的输出结果 H_1 和 H_2 输出自注意力机制层，经过第一层线性网络层 L_1 和第二层线性网络层 L_2 计算后分别得到句子中词的特征权重分数 a_1 和 a_2，将其与对应的向量和加权求和，得到含有

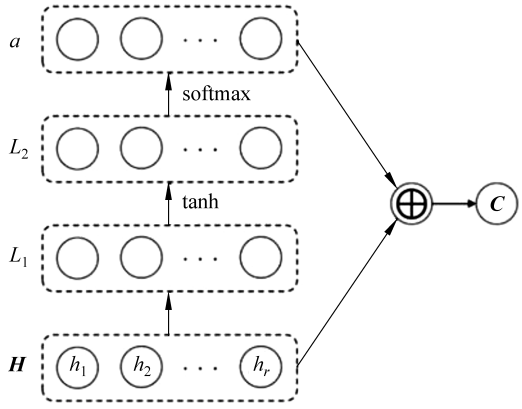

图 6　自注意力机制计算过程

长距离语义信息的汉老句子对特征向量 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$，计算如式(11)所示。

$$\boldsymbol{C}_i=a_i\oplus\boldsymbol{H}_i \tag{11}$$

通过自注意力层的计算，即可分别得到含有长距离语义信息的汉老双语句子特征向量表示。

3.6　汉老双语句子相似度表示

对于汉老双语句子对 S_1 和 S_2，通过 3.1～3.5 节所描述的方法获取含有长距离语义信息和句子结构信息的汉老双语句子语义表示向量 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$ 后，分别对其进行按位减和按位乘操作，捕获句子对间的匹配信息，将结果进行拼接后传输到全连接网络层，计算汉老双语句子对的相似度分数 p。具体计算如式(12)～式(15)所示。

$$P_1=\boldsymbol{C}_1\oplus\boldsymbol{C}_2 \tag{12}$$

$$P_2=\boldsymbol{C}_1\otimes\boldsymbol{C}_2 \tag{13}$$

$$P_s=\tanh(W^1P_1+W^2P_2+b) \tag{14}$$

$$p(y\mid P_s)=\mathrm{sigmoid}(W^sP_s+c) \tag{15}$$

其中，W^1，W^2，W^s，b，c 为模型参数，p 为取值介于 0 至 1 之间的相似度分数。模型采用交叉熵(cross entropy)作为目标函数，如式(16)所示。

$$L=y\log(p)+(1-y)\log(1-p) \tag{16}$$

通过以上公式，即可计算得到汉老句子对 S_1 和 S_2 的相似度分数 p。

4　实验及分析

4.1　实验设置与评价

4.1.1　实验数据与模型设置

本文使用的数据集分为预训练词嵌入矩阵的数据集和模型训练数据集，其中预训练词嵌入矩阵的数据集，汉语部分采用 Li 等人[21]预训练好的词向量(1.69 GB)，老挝语部分通过老挝语维基百科①爬取了 125.95 MB 单语语料，使用实验室开发的老挝语工具对其预处理后，采用 Artetxe[20]等人的方法预训练高质量的双语词嵌入，词嵌入维度设置为 300；在模型训练数据集中，对中文维基百科②和老挝语维基百科爬取了篇章级对齐语料，通过人工对齐和校对后得到 75 040 条平行句对，与 Grégoire 等人[22]的方法类似，本文以每个平行句子对的负样本

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

② https://wiki.hk.wjbk.site/wiki

数为7的比例来构建非平行语料库，并且限定生成的句子对之间的长度之比不大于2，最终构建了524 810条非平行句对，如表2所示。

表2 汉老双语数据集

数据集	句子对数量
汉老双语平行句对数据集	75 040
汉老双语非平行句对数据集	524 810

实验在固定随机种子数下使用10折交叉验证，将构建的汉老双语平行句对语料库的90%作为训练集，剩余的10%作为测试集分别训练10次，取实验结果的均值，每次训练使用的数据集划分如表3所示。

表3 数据集划分

数据集	平行句子对数量	非平行句子对数量
训练集	67 536	472 329
测试集	7 504	52 481

模型实现使用Python语言及Keras框架，表4列出了模型的实验参数设置。

表4 模型超参数

参数	值
Batch size	256
Learning rate	0.000 1
Dropout rate	30%
LSTM dimention	50
Dense dimention	30

4.1.2 评价指标

本文按照标准评价指标，统计了各种方法的准确率P和召回率R，在此基础上将各方法的F_1值作为衡量模型是否可以正确分类汉语-老挝语的平行句子的最终评价指标。采用0.5作为句子相似的判别阈值，当句子对的相似度分数大于0.5时即将其分为相似句子对。准确率P、召回率R、F_1值的具体计算如式(17)～式(19)所示。

$$P=\frac{\text{正确分类平行句对的个数}}{\text{相似的句对个数}} \tag{17}$$

$$R=\frac{\text{正确分类平行句对的个数}}{\text{平行句对的个数}} \tag{18}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\% \tag{19}$$

4.2 模型对比实验

本文使用的模型框架为带有自注意力机制的BiLSTM模型，在此基础上加入了句子结构特征来丰富句子语义表示。为了验证自注意力机制对模型的有效性，在不同设定下训练了四个模型，每个模型的设定如下：

(1) BiLSTM模型；

(2) 带有注意力机制(attention)的BiLSTM模型；

(3) 带有自注意力机制的BiLSTM模型；

(4) 加入句子结构特征(struct_tag)的带有自注意力机制的BiLSTM模型，即本文方法。

其中，设定(1)是本文的基准模型(Base Model)；设定(2)和设定(3)是为了比较不同注意力机制对模型性能的影响；设定(4)为本文方法。此外，与目前主流的3种跨语言句子相似度计算模型作了对比：

(1) **Siamese LSTM模型**[5]：将平行句对分别输入共享参数的LSTM网络提取句子对的特征向量，通过计算特征向量间的曼哈顿距离得到句子对的相似度分数。模型结构设置与超参数均与原文一致，LSTM隐状态维度为50维，优化算法选择Adadelta。

(2) **CNN+Self-Attention模型**[6]：对输入的平行句对分别运用CNN和自注意力机制(self-attention)得到每个句子的局部语义信息和全局语义信息，将其拼接后计算特征向量间的相对差和相对积，将结果拼接后传输到全连接网络层计算得到句子间的相似度分数。模型结构设置与超参数均与原文一致，其中，CNN卷积核设定为300，池化操作中的k设置为3，自注意力机制设置8个头，每个头的参数矩阵设置为16维，全连接层中第一层神经元节点设置为900，第二层设置为6。

(3) **LSTM+ Attention模型**[7]：对输入的句子对使用带有注意力机制的LSTM提取句子对的特征向量，计算特征向量间的相对差和相对积，将结果拼接通过全连接网络层计算相似度分数。模型结构设置与超参数均与原文一致，其中LSTM隐状态维度为50，dropout设置为0.2，损失函数中L2正则设置为0.000 1，优化算法使用Adam。

以上7个模型均在相同训练语料下采用10折交叉验证进行实验，并且固定随机种子数，实验结果如表5所示。

表5 不同模型对比结果

模型	P/%	R/%	F_1/%
Base Model	63.26	57.55	60.27

续表

模型	P/%	R/%	F_1/%
＋attention	65.53	66.78	66.15
＋Self-Attention	67.73	66.61	67.17
＋Self-attention＋struct_tag(本文方法)	**70.88**	**69.61**	**70.24**
Siamese LSTM	63.32	56.10	59.49
CNN＋Self-Attention	65.37	65.48	65.42
LSTM＋Attention	67.44	63.98	65.67

由表 5 可知，加入注意力机制可有效提升模型性能，与基准模型相比 F_1 值提升了 5.88%，这是由于注意力机制可以快速提取数据的重要特征，而自注意力机制作为注意力机制的改进，将注意力机制替换为自注意力机制后模型的 F_1 值进一步提升了 1.02%，原因是自注意力机制减少了对外部信息的依赖，可以更有效地捕获数据和特征的内部关联性。设定(2)和设定(3)训练的模型相比较，说明了自注意力机制在研究句子相似度任务上的有效性。此外，加入句子结构特征使模型的 F_1 值提升了 3.07%，说明设定(4)的特征方法对于汉老双语句子相似度的研究是有效的。

另一方面，Siamese LSTM 模型和 CNN＋Self-Attention 模型与本文模型相比 F_1 值分别低了 10.75% 及 4.82%。分析原因后发现 Siamese LSTM 模型的框架虽然对于跨语言句子相似度计算具有较好的适应性，并且 LSTM 网络可以在一定程度上捕获句子的特征信息，但对于高维度的特征向量，通过曼哈顿距离来度量相似性存在一定的误差；而 CNN＋Self-Attention 模型则是对同一语系或差异性较小的语言具有较好的效果，汉语-老挝语的语言跨度较大，虽然通过自注意力机制可以在一定程度上提取句子更加准确的语义特征，但 CNN 提取的汉老双语句子特征具有较大差异性，因此与本文方法相比该方法的实验结果较差。LSTM＋Attention模型相比本文模型的 F_1 值低了 4.57%，并且与模型(2)相比 F_1 值低了 0.48%，出现这一结果的原因是 BiLSTM 网络相比 LSTM 网络可以更好地进行句子建模，增加句子语义表示的准确性。

总结而言，在汉老双语句子相似度计算任务中，由于语言差异性较大，BiLSTM 网络相比于 LSTM 网络和 CNN 网络可以更好地对句子进行建模，并且加入自注意力机制和句子结构特征可以进一步提升模型效果。

4.3　特征标记方法对比实验

由 4.2 小节设定(4)训练的模型可知，使用特征模板获取句子结构特征可以有效提升模型性能。为了验证本文提出的特征模板的有效性，探索特征模板的不同标记方法对模型结果产生的影响，本节按以下设定额外训练了 7 个模型，并且与 3.2 节中的设定(3)和(4)做比较，具体设定如下：

(1) 带有自注意力机制的 BiLSTM 模型；

(2) 在设定(1)的基础上加入句子的主语特征标记(sub)；

(3) 在设定(1)的基础上加入句子的谓语特征标记(verb)；

(4) 在设定(1)的基础上加入句子的宾语特征标记(obj)；

(5) 在设定(1)的基础上加入句子的主语和谓语特征标记(sub＋verb)；

(6) 在设定(1)的基础上加入句子的主语和宾语特征标记(sub＋obj)；

(7) 在设定(1)的基础上加入句子的谓语和宾语特征标记(verb＋obj)；

(8) 在设定(1)的基础上加入句子的词性标记(pos_tag)；

(9) 在设定(1)的基础上加入完整的句子结构特征标记(sub＋verb＋obj)，用 struct_tag 表示，即本文方法。

在以上 9 个设定训练的模型中，设定(1)和设定(9)分别为 4.2 节中设定(3)和设定(4)训练好的模型。在本节中，设定(1)为验证特征标记对模型影响的基准模型；设定(2)、设定(3)、设定(4)和设定(5)、设定(6)、设定(7)是为了探索不同特征标记对模型的影响，以及探索不同组合的特征标记对提升模型性能的有效性；设定(8)和设定(9)则是比较了加入词性特征标记与句子结构特征标记对模型性能的影响。以上模型均使用同一训练语料采用 10 折交叉验证进行实验，并且固定随机种子数，实验结果如表 6 所示。

表 6　不同特征标记对模型性能的影响

模型	P/%	R/%	F_1/%
BiLSTM＋Self-Attention	67.73	66.61	67.17
＋sub	70.13	66.67	68.36
＋verb	69.92	66.03	67.92

续表

模型	P/%	R/%	F_1/%
+obj	69.89	66.87	68.35
+sub+verb	70.87	67.68	69.24
+sub+obj	70.37	68.84	69.60
+verb+obj	70.62	68.34	69.46
+pos_tag	67.53	60.39	63.76
+struct_tag（本文方法）	**70.88**	**69.61**	**70.24**

由实验结果发现，在加入一种特征标记的模型中[设定(2)、设定(3)、设定(4)]，加入主语标记(sub)的设定(2)对模型效果提升最大，与设定(1)的F_1值相比提升了1.19%；加入两种特征标记的模型中[设定(5)、设定(6)、设定(7)]，加入主语和宾语标记(sub+obj)的设定(6)对模型的性能提升最高，相比设定(1)的F_1值提升了2.43%；而加入完整句子结构特征(本文方法)的设定(9)取得了最好的效果，相比设定(1)的F_1值提升了3.07%。设定(2)和设定(6)在两组对比中得到了最好的效果，并且两者均未含有谓语标记(verb)，分析后发现原因是由于在句子结构中，谓语成分通常位于句子的中间或末尾，具有模糊的位置关系，通过本文提出的特征模板对老挝语的谓语成分进行标记存在一定的误差；而主语和宾语成分通常位于句子的两端，使用本文的特征模板可以较好地确定标记位置，因此设定(6)在加入两种特征标记的模型中F_1值提升最大。设定(8)在加入词性特征标记后相比未加入前的设定(1)，模型的F_1值反而降低了3.41%，得到这一结果的原因是由于汉语和老挝语虽然在句子的主要成分上具有一致的顺序结构(SVO)，但句子的次要成分具有差异性。例如，汉语的定语通常在主语之前，状语在主语之后，而老挝语则正好相反，仅添加词性标记反而使模型更难获取句子的特征信息。

总的来说，使用特征模板获取的句子结构特征对汉老双语句子相似度计算任务是个十分有效的方法，可以弥补语料资源稀缺对模型性能的影响。

4.4 词嵌入映射方法对比实验

为了减少汉老双语的语言差异性，与Artetxe[20]等人提出的方法类似，本文采用弱监督映射方法将双语词嵌入映射到共享的语义空间。为了验证方法的有效性，本节与目前主要使用的无监督和监督映射的方法[23-24]做对比，其中无监督映射方法指通过自学习方式学习线性变换矩阵进行映射[24]，监督映射方法指使用较大双语词典学习映射矩阵的方法[25]。将未经过词嵌入映射的模型作为基准模型(Base Model_2)，分别使用unsupervised、supervised和semi_supervised代表无监督、监督和弱监督映射方法，其中弱监督映射方法即本文方法。实验结果如表7所示，模型均在同一数据集下采用10折交叉验证进行实验，并且固定随机种子数，超参数均使用原文中参数，监督映射方法和弱监督映射方法使用的映射词典为同一种子词典(836对常用词)。

表7 不同词嵌入映射方式对模型性能的影响

模型	P/%	R/%	F_1/%
Base Model_2	67.37	66.36	66.86
+unsupervised	69.42	68.10	68.75
+supervised	68.62	67.91	68.26
+semi_supervised（本文方法）	**70.88**	**69.61**	**70.24**

由结果可知，在使用了词嵌入映射后模型的性能均获得了提升，与基准模型相比，监督映射方法(supervised)的提升最小，F_1值仅提升了1.4%，而无监督映射方法(unsupervised)的F_1值提升了1.89%，得到这一结果的原因是监督映射的方法需要在较大规模的双语词典下才能取得较好的效果，而由于老挝语资源稀缺，目前仅拥有小规模的词典，因此效果较差；无监督映射的方法不需要种子词典，而是通过线性变换学习转换矩阵，因此取得了一定的效果。弱监督映射(semi_supervised)的方法取得了最好的效果，F_1值提升了3.38%，原因是该方法仅需要较小的种子词典即可学习到效果较好的转换矩阵，并且由于汉语和老挝语的语言差异较大，仅通过无监督映射学习存在一定的困难，因此与无监督方法相比，弱监督方法取得了最好的效果。

综上所述，对于汉语和老挝语的句子相似度计算，通过使用双语词嵌入映射的方法可以有效缩小语言间的差异性，提升模型的性能。

5 结论

本文根据汉语和老挝语句子结构的特点提出一种融合句子结构特征的汉老双语句子相似度计算方法，在将双语词嵌入映射到共享语义空间缩小语言差异性的基础上，通过加入句子结构特征有效提高了汉老双语句子相似度计算模型的性能。实验结果

表明，本文方法在有限的训练样本下效果明显优于目前的主流方法，F_1 值达到了 70.24%。下一步将考虑利用该方法提取汉老双语句子对，融入机器翻译和其他老挝语相关的自然语言处理工作中来提升效果。

参考文献

[1] 石杰，周兰江，线岩团，等.基于 WordNet 的中泰文跨语言文本相似度计算[J].中文信息学报，2016，30(4)：65-70.

[2] 闫红，李付学，周云.基于 HowNet 句子相似度的计算[J].计算机技术与发展，2015，25(11)：53-57.

[3] Tian J，Zhou Z，Lan M，et al.Ecnu at semeval-2017 task 1：leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity[C]//Proceedings of the 11th International Workshop on Semantic Evaluation，2017：191-197.

[4] 黄洪，陈德锐.基于语义依存的汉语句子相似度改进算法[J]. 浙江工业大学学报，2017，045(001)：6-9.

[5] Mueller J，Thyagarajan A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence，2016：2786-2792.

[6] 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.

[7] Chi Z，Zhang B. A sentence similarity estimation method based on improved Siamese network[J]. Journal of Intelligent Learning Systems and Applications，2018，10(4)：121-134.

[8] Chien C Y，Chang C H，Wei C P. Bilingual parallel sentence extraction from comparable corpora[C]//Proceedings of the Conference on Computational Linguistics and Speech Processing，2019：167-181.

[9] 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.

[10] Erdmann M，Finch A，Nakayama K，et al. Calculating wikipedia article similarity using machine translation evaluation metrics[C]//Proceedings of the IEEE Workshops of International Conference on Advanced Information Networking and Applications. 2011：620-625.

[11] Wu H，Huang H，Jian P，et al. BIT at SemEval-2017 task 1：using semantic information space to evaluate semantic textual similarity[C]//Proceedings of the 11th International Workshop on Semantic Evaluation，2017：77-84.

[12] Zhuang W L，Chang E. Neobility at SemEval-2017 Task 1：an attention-based sentence similarity model[J]. arXiv preprint arXiv：1703.05465，2017.

[13] Shao Y. HCTI at SemEval-2017 Task 1：use convolutional neural network to evaluate semantic textual similarity[C]//Proceedings of the 11th International Workshop on Semantic Evaluation，2017：130-133.

[14] He H，Gimpel K，Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2015：1576-1586.

[15] 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：就业教育，2011，(016)：220-221.

[16] Gers F. Long short-term memory in recurrent neural networks[D].Verlag nicht ermittelbar，2001.

[17] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems，2017：5998-6008.

[18] 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(07)：1312-1317.

[19] 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.

[20] 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.

[21] Li S，Zhao Z，Hu R，et al. Analogical reasoning on chinese morphological and semantic relations[J]. arXiv preprint arXiv：1805.06504，2018.

[22] Joulin A，Bojanowski P，Mikolov T，et al. Loss in translation：learning bilingual word mapping with a retrieval criterion[J]. arXiv preprint arXiv：1804.07745，2018.

[23] Barnes J，Klinger R，Walde S S. Projecting embeddings for domain adaptation：joint modeling of sentiment analysis in diverse domains[J]. arXiv preprint arXiv：1806.04381，2018.

[24] Artetxe M，Labaka G，Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings[J]. arXiv preprint arXiv：1805.06297，2018.

[25] Artetxe M，Labaka G，Agirre E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2016：2289-2294.

[26] Grégoire F, Langlais P. Extracting parallel sentences with bidirectional 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

CCL 2022 系统展示征集

“第二十一届中国计算语言学大会”(The 21st China National Conference on Computational Linguistics, CCL)将于 2022 年 10 月 14—16 日在江西南昌举行,会议主办单位为中国中文信息学会,承办单位为江西师范大学。中国计算语言学大会创办于 1991 年,是中国中文信息学会的重要会议。经过 30 多年的发展,CCL 被广泛认为是国内自然语言处理领域最权威的、最具影响力的学术会议会议。作为中国中文信息学会的旗舰会议,CCL 聚焦于中国境内各类语言的智能计算和信息处理,为研讨和传播计算语言学最新学术和技术成果提供了最广泛的高层次交流平台。

自然语言处理是学术界与产业界最紧密合作的研究领域之一。在学术报告、论文交流等活动之外,会议专门设立了系统展示环节,努力提供一个供学术界和产业界交流前沿研究进展、展示技术应用的平台。系统展示环节欢迎各项与自然语言处理相关的系统进行展示,包括**但不限于**以下几种类型:

- 自然语言处理原型系统
- 以语言技术为核心的应用系统
- 以文本为中心的多模态应用系统
- 用于辅助计算语言学研究的软件或工具
- 数据可视化和数据标注软件或工具

系统展示环节将在会议期间进行,无需发表论文,只需提供系统介绍海报。具体展示安排随会议组织一并通知。本次 CCL 大会将评选并颁发最佳系统展示奖,欢迎大家报名参与!

为了更好地组织好系统展示环节,请在 2022 年 7 月 15 日之前与会议组织方(wnzhang@ir.hit.edu.cn 或 pengm@whu.edu.cn)联系。组委会将根据场地及参会人员情况确定最终的参展单位和展示方案。