doi:10.11835/j.issn.1000-582X.2021.01.006

结构特征一致性约束的双语平行句对抽取

毛存礼[a,b],高 旭[a,b],余正涛[a,b],王振晗[a,b],高盛祥[a,b],满志博[a,b]
(昆明理工大学 a.信息工程与自动化学院; b. 云南省人工智能重点实验室,昆明 650500)

摘要:平行句对抽取是解决低资源神经机器翻译平行语料不足的有效途径。基于孪生神经网络的平行句对抽取方法的核心是通过跨语言语义相似度判断 2 个句子是否平行,在相似的语言对上取得了非常显著的效果。然而针对英语–东南亚语言双语句对抽取任务,面临语言空间和句子长度存在较大差异,仅考虑跨语言语义相似度而忽略句子长度特征会导致模型对仅有语义包含关系但不平行句对的误判。笔者提出一种结构特征一致性约束的双语平行句对抽取方法,该方法是对基于孪生神经网络的双语平行句对抽取模型的扩展,首先通过多语言 BERT 预训练语言模型在嵌入层将两种语言编码到同一语义空间,以此缩小语义空间中语言的差异。其次分别对两种语言句子的长度特征进行编码,与孪生网络编码后的句子语义向量进行融合,增强平行句对在语义及结构特征上的表示,降低模型对语义相似但不平行句对的误判。在英–缅双语数据集上进行实验,结果表明提出的方法相比基线模型准确率提高了 4.64%,召回率提高了 2.52%,F_1 值提高了 3.51%。

关键词:双语平行句对;低资源语言;BERT 预训练;孪生网络;结构
中图分类号:TP391 **文献标志码**:A **文章编号**:1000-582X(2020)01-046-11

Extraction of bilingual parallel sentence pairs constrained by consistency of structural features

MAO Cunli[a,b], *GAO Xu*[a,b], *YU Zhengtao*[a,b], *WANG Zhenhan*[a,b], *GAO Shengxiang*[a,b], *MAN Zhibo*[a,b]
(a. Faculty of Information Engineering and Automation; b. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, P. R. China)

Abstract: Parallel sentence pair extraction is an effective way to solve the shortage of low-resource neural machine translation. The main method based on Siamese neural network is to judge whether two sentences are parallel through cross-language semantic similarity, which has achieved remarkable results on similar language pairs. However, for English- Southeast Asia language sentence pairs extraction tasks, there are not only great differences in language space but also great differences in sentence length. Considering only

收稿日期:2020-09-10
基金项目:国家自然科学基金重点资助项目(61732005);国家自然科学基金资助项目(61662041,61761026, 61866019, 61972186);云南省应用基础研究计划重点资助项目(2019FA023);云南省中青年学术和技术带头人后备人才资助项目(2019HB006)。
Supported by the Key Program of National Natural Science Foundation of China (61732005); the National Natural Science Foundation of China (61662041, 61866019,61761026, 61972186); the Key Project of Applied Basic Rearch Program of Yunnan Province (2019FA023); the Candidates of the Young and Middle Aged Academic and Technical Leaders of Yunnan Province (2019HB006).
作者简介:毛存礼(1977—),博士,副教授,主要从事自然语言处理、信息检索、机器翻译方向研究,(E-mail) maocunli@163.com。
通讯作者:余正涛,男,教授,博导,主要从事自然语言处理,信息检索,机器翻译方向研究,(E-mail) ztyu@hotmail.com。

cross-language semantic similarity and ignoring sentence length features will lead to misjudgment of non-parallel sentence pairs with only semantic inclusion. Therefore, this paper proposes a parallel sentence pairs extraction method constrained by consistency of structural features. The method is an extension of the model based on Siamese neural network. Firstly, using the multilingual BERT to embed the two languages into the same semantic space in the embedding layer, so as to reduce the language differences in the semantic space. Secondly, embedding the length features of sentences respectively, and fusing it with the semantic vectors of sentences encoded by Siamese networks to enhance the representation of parallel sentence pairs in semantic and structural features, so as to solve the misjudgment problem. We experiment on the English-Burmese data sets. The results show that the precision is increased by 4.64%, the recall is increased by 2.52%, and the F_1 value is increased by 3.51% compared with the baseline.

Keywords: parallel sentence; low-resource; BERT pretrain; siamese network; structural

基于互联网抽取平行句对是解决神经机器翻译中面临平行语料稀缺的有效途径。与英语、汉语这样资源丰富语言相比，东南亚语资源稀缺，直接获取大规模的平行句对十分困难。互联网中存在大量的英语和东南亚语言的双语数据，但这些数据大多是主题相关、内容相似的双语可比文档。准确抽取双语可比语料中平行句子，对开展英语–东南亚神经机器翻译研究具有重要价值。

计算句子相似度是解决问答系统[1]、平行句抽取的主要方法。针对双语平行句对抽取任务，传统的方法可以利用 Wikipedia 提供的附加注释和特征使用自动生成词典模型提高平行句对抽取任务的准确性[2]，但是这类方法不仅受限于双语词典的规模，且不能很好的从语义上计算两种不同语言间的相似度，导致抽取效果不理想。

随着深度学习的发展，研究员开始使用深度学习方法抽取平行句子。采用端到端的方式学习句子的向量表示，不需要其他的结构和特征，在英语–法语数据集上取得良好效果。英语和法语在语言特征上属于相似语言，在进行向量表示时可以直接映射到同一语义空间中，而英语与东南亚语言之间语言差异性较大，映射到语义空间中会导致句子语义信息的丢失，得到效果会明显降低，同时，英语和东南亚语言的句子长度存在差异，模型仅考虑语义信息，忽略句子长度特征会导致对仅有语义包含关系但不平行句对的误判，以英语和缅甸语为例，如图 1 所示。

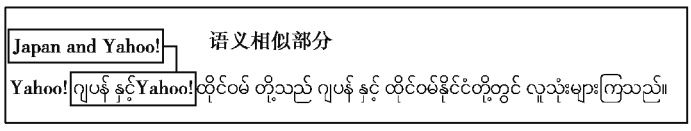

图 1　具有语义包含关系但不平行的句对示例

Fig. 1　Examples of semantic similar but not parallel

图 1 中，缅甸语句中有和英语句子相同的语义部分，所以在语义相似度计算时，句子对的相似度会很高，但实际上，两者属于不平行的句子。同时，对英–缅句子长度比的数据集进行了统计分析，英语与缅甸语的长度比在 0.4～0.7 之间的占比为 80%，因此，不考虑句子间长度特征时，由上述问题及句对本身的结构差异性，模型会出现错误分类。

因此针对因英语和东南亚语言的语义空间和长度特征差异导致模型误判问题，提出了结构特征一致性约束的平行句对抽取方法，结构特征一致性包括减少语义和长度上的差异，使两种语言能够保持一致的结构特征。首先使用 BERT(BERT, bidirectional encoder representations from transformers)[3] 预训练模型得到良好的英语和东南亚语言的词向量并结合基于 Bi-LSTM 的孪生网络将两种语言在语义空间中共享，同时分别对句子的长度特征进行编码并融合到句子向量中，以提高模型的准确率。

贡献主要有以下 2 点：

1)基于多语言 BERT 预训练模型并结合孪生网络将两种语言编码到同一语义空间,缩小语义空间中语言的差异,解决英语和东南亚语言空间差异较大问题,保持语义空间的一致。

2)对句子的长度特征进行编码,与孪生网络编码后的句子语义向量进行融合,有效增强结构特征上的表示,以解决模型对仅有语义包含关系但不平行的句对误判的问题,保持结构特征的一致。

1 相关研究工作

目前从双语可比语料中抽取平行句对的方法主要有三类:1)基于句子的特征信息、基于词典以及信息检索的方法;2)基于枢轴的方法;3)基于深度学习的方法。

第一类是基于句子的特征信息、基于词典以及信息检索的方法,例如 Munteanu 等人[4]通过使用一种信号处理的方法分析可能相似的句子对,检测源句子中的哪些片段被翻译为目标句子中的片段,这种方法甚至可以从非平行的语料库中提取相对有用的机器翻译训练数据。Zhao 等人[5]提出一种自适应方法,根据句子长度模型和基于词典的模型在最大似然的规则下进行组合,从双语可比新闻语料中提取平行句对,同时使用已抽取的平行句对迭代更新翻译词典,以获得更好词汇覆盖率和翻译概率参数估计。Munteanu[6]训练最大熵分类器判断是否是互译句子,并从大型中文,阿拉伯文和英文非平行报纸语料库中提取平行数据,Alberto 等人[7]通过余弦和跨语言信息检索中的长度因子来计算句子对之间的相似性,从而对齐来自维基百科的特定域并行文档;Tillmann 等人[8]直接在句子级别对大量候选句子对进行评分。通过一个对称评分函数实现句子级别提取。Chu 等人[9]通过笛卡尔乘积生成所有可能的句子对,并过滤掉不满足条件的,保留尽可能匹配的句子对,使用少量平行句对训练分类器,从候选句中识别平行句对。以上方法虽证明了抽取平行句对的有效性,但是由于依赖人工提取特征及词典等外部信息,无法取得更好效果。

第二类是利用枢轴的方法,主要利用机器翻译将源语言翻译成目标语言,再获取平行句对,如 Ann 等人[10]基于现有翻译系统,将源语言翻译成目标语言得到候选句子,然后对候选句子对进行打分排序,获得平行句子。Afli[11]等提出了一种基于多模态可比语料库的机器翻译并行数据提取方法,以及从多模态语料库中提取双语并行句子对或短语对的方法。Bouamor 等人[12]通过将多语言句子级嵌入,并与神经机器翻译和监督分类配对的混合,分类法语–英语语料库中的平行句子对。通过双语分布式表示模型学习的每个源–目标句子对的连续向量对目标翻译候选进行过滤,使用神经机器翻译系统或二进制分类模型选择最佳翻译。以上方法虽然有效但是依赖机器翻译的性能,不适用于低资源语言。

第三类是基于深度学习方法,利用深度学习对数据表征学习在机器翻译、情感分类等任务上取得较好的效果,已成为自然语言处理任务的主流方法。针对双语平行句抽取的任务开展大量研究。Grégoire 等人[13]第一次提出使用深度学习方法抽取平行句对,使用双向递归神经网络学习句子向量表示,由于相似语言间可以共享词表,结构相似因此在英语–法语的平行句对抽取任务上表现良好。Ramesh 等人[14]使用端到端的孪生网络构造了一个分类器,选择了少量平行句对作为正样本以及大量的负样本训练模型,然后提取英语–塔达米尔语平行句子。Grover 等人[15]训练模型以获取双语单词嵌入,然后在 2 个句子的单词之间创建相似度矩阵并将句子分类。相似语言拥有部分同源词,可以共享语义空间,且在句子结构上也基本相似,无需考虑长度特征问题。而针对英语–东南亚语言这样的差异性较大的语言对,没有同源词,在语义空间中差异较大且句子存在结构差异,上述方法并不完全适合,因此提出结构特征一致性约束的双语平行句对抽取方法,通过多语言 BERT 预训练语言模型得到较好的初始向量并共享双语语义空间,融合结构特征解决句子对语义相似但长度不平衡的问题,最终达到结构特征的一致性。实验表明,提出的方法在东南亚低资源语言的平行句对抽取任务上取得良好效果。

2 结构特征一致性约束的平行句对抽取模型

2.1 模型架构

模型架构主要分为两部分:第一部分是多语言 BERT 预训练语言模型,第二部分是融合结构特征的孪生网络,主要方法框架如图 2 所示(以英语–缅甸语为例)。在多语言预训练 BERT 部分首先使用多语言

BERT 模型预训练英语和缅甸语的词向量；其次，在嵌入层得到孪生网络的输入，通过孪生神经网络将英语和缅甸语句子表示共享到同一语义空间；同时对英语、缅语句子的长度特征进行编码，与孪生网络的编码后的向量融合，使模型既考虑了语义信息，又考虑了英语和缅甸语的结构特征；最后计算 2 个向量的相似程度。

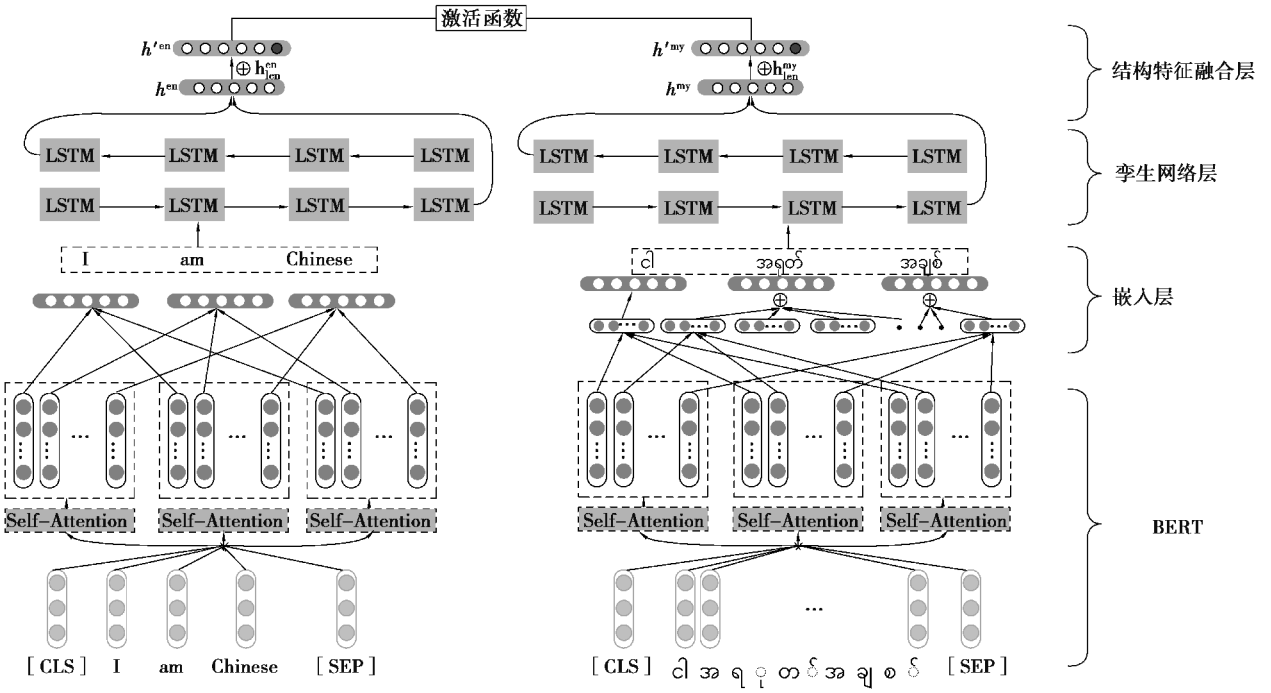

图 2　模型架构

Fig. 2　Model Structure

2.2　基于多语言预训练语言模型的双语词嵌入

预训练词向量（词嵌入）是通过训练得到词的向量化表示，Ma 等人[16]通过 Glove[17]、word2vec[18]的实验证明了预训练词向量方法与随机初始化词向量方法相比，具有显著改进。在跨语言词向量表示时，相似语言对可共享部分词表，所以在孪生神经网络中直接使用随机初始化方式可获得较好效果。但针对英语和东南亚语言这样语言差异较大的语言对，随机初始化词向量的方法不能将两种语言在同一语义空间中表示，直接影响模型性能。BERT 的网络架构基于多层 Transformer[19]结构构建的，在最近的研究中证明 Transformer 结构可以获得更好学习效果[20]并在特征表示取得较好的效果，广泛应用到摘要生成[21]等自然语言处理任务，是目前主流方法。由于语言差异性较大，使用随机初始化生成的词向量在语义上会有很大差别，为得到较好向量表示并共享语义空间及考虑上下文语义信息，使用多语言预训练语言模型（MBERT，multilingual bidirectional encoder representations from transformers）预训练英语和缅甸语词向量。目前 MBERT 包括了 104 种语言，其中包括缅甸语、越南语和泰语，生成的词向量能够共享同一语义空间，最近研究表明 MBERT 在跨语言任务的有效性[22-23]。然而 MBERT 模型对英语和东南亚语的训练方式不同，英语以空格分词并基于子词切分，因此在 Embedding 层可以直接使用 MBERT 预训练的词向量；东南亚语言由于构词特点不同，采取向量表示方法也会不同。缅甸语是基于字符构词，因此，在 Embedding 层进行缅甸语嵌入时有两种方法，一种直接使用缅甸语字符嵌入组成缅甸语句子向量表示，这种方法并没有考虑缅甸语字符之间的组合信息，通过字符嵌入得不到充分的语义信息。

笔者首先通过MBERT生成缅甸语文本中构成每个词语的各字符级向量，然后在Embedding层将字符级向量进行组合得到词向量的表示。如表1所示是多语言MBERT的词表。例如缅甸语句子‘ငါအရုတ်အချိန်’中的语义单词‘တရုတ်’是由‘အ’,‘ရ’,‘ု’,‘တ’,‘်’5个字符组合而成的。而在缅甸语的文本中，不同的上下文信息，字符的组合也会有所不同。在例句的语义场景中，‘အ’,‘ရ’,‘ု’,‘တ’,‘်’通过字符向量的相加将会组合成正确的目标词‘တရုတ်’的词嵌入。这样，可以更准确获得缅甸语词语的语义信息，通过字符和音节的组合在保留上下文信息以及语义信息的同时得到最终单词的词嵌入表示，最终得到句子表示；与缅甸语不同，越南语构词的主要特点是每一个音节都是一个有意义的单位，即越南语的最小语义单元是音节，可以独立使用，因此在生成向量表示时，基于音节得到词向量的表示；泰语与缅甸语类似，单独字符和音节没有实际的语义含义，需将字符组合成词向量表示。

表1 BERT东南亚语词汇表

Table 1 Vocabulary in BERT

缅甸语	က	ဖ	ထ	ဗ	အ	ဝး	၄	၌	ဂ	…
	ခ	ဥ	ဒ	ယ	ဣ	ဇ	၅	၍	ဃ	
越南语	Ă	ă	Ą	ą	Ć	ć	Č	č	Ď	…
	đ'	Đ	đ	Ē	ē	Ě	ě	Ė	ė	
泰语	ก	ข	ค	ฆ	ง	จ	ฉ	ช	ฌ	…
	ญ	ฎ	ฐ	ฑ	ฒ	ณ	ด	ท	ธ	

2.3 融合结构特征及孪生网络的双语句对分类

基于深度学习的平行句对分类方法的本质是将两种语言的句子在同一个语义空间中表示成向量然后计算2个向量的相似度。孪生神经网络由2个结构相同的神经网络构成，2个网络共享权值，主要应用于分类任务。为了得到共享语义空间中双语句子的向量表示，采用孪生神经网络结构对双语句子进行编码表示。孪生网络得到的向量仅考虑了语义信息而忽略了结构特征，导致仅有语义包含关系但不平行句对的相似度过高，因此研究融合了结构特征双语句对分类。

下面以英语–缅甸语为例介绍融合结构特征的孪生网络的双语句对分类方法。使用基于Bi-LSTM(bi-directional long short-term memory)的孪生网络结构对预训练模型得到的英语和缅甸语词向量作为输入，分别经过Bi-LSTM层将英语、缅甸语句子前向和后向2个方向最后状态的向量做拼接以表示句子向量。以缅甸语为例，比如缅甸语句子‘ဒါကလျှို့ဝှက်ချက် တစ်ခု မဟုတ်ပါဘူး’，在预训练模块后得到向量表示$[\boldsymbol{E}_1^{my},\boldsymbol{E}_2^{my},\cdots,\boldsymbol{E}_N^{my}]$，$N$表示句子单词的个数。之后经过Bi-LSTM编码，前向LSTM编码依次输入词向量得到$\{\boldsymbol{h}_{f0},\boldsymbol{h}_{f1},\cdots,\boldsymbol{h}_{fN}\}$，后向LSTM编码从后向前输入词向量得到$\{\boldsymbol{h}_{bN},\cdots,\boldsymbol{h}_{b1},\boldsymbol{h}_{b0}\}$，$N$表示单词的个数，$\boldsymbol{h}_{fi}$和$\boldsymbol{h}_{bi}$表示$i$时刻的隐向量，$\boldsymbol{h}_{fN}$和$\boldsymbol{h}_{b0}$表示前向和后向编码最后时刻的隐向量表示，将$\boldsymbol{h}_{fN}$和$\boldsymbol{h}_{b0}$拼接得到该句子的编码向量$\boldsymbol{h}$

$$\boldsymbol{h}=[\boldsymbol{h}_{fN};\boldsymbol{h}_{b0}]。\tag{1}$$

通过MBERT预训练语言模型，可以使英语和缅甸语中相同语义的词向量在语义空间中相互接近，但Bi-LSTM层中因为参数的不同，不能保证共享语义空间。共享权重的孪生网络可以在相同参数的Bi-LSTM网络下，将英语和缅甸语的向量表示最大程度共享到同一语义空间，提高跨语言句子表示的准确性，提高模型对跨语言句子的语义相似度计算准确性。通过孪生神经网络输出，得到英语和缅甸语句子向量$\boldsymbol{h}^{en}$和$\boldsymbol{h}^{my}$。

在融合句子长度特征方面，使用随机初始化方式将英语和缅甸语的句子的长度(即句中单词的个数)编码为向量，分别得到$\boldsymbol{h}_{len}^{en}$和$\boldsymbol{h}_{len}^{my}$，然后与孪生网络编码后的英语、缅甸语句子语义向量$\boldsymbol{h}^{en}$和$\boldsymbol{h}^{my}$进行融合，增强平行句对在语义及结构特征上的表示

$$\boldsymbol{h}'^{en}=\boldsymbol{h}^{en}\oplus\boldsymbol{h}_{len}^{en},\tag{2}$$

$$\boldsymbol{h}'^{my}=\boldsymbol{h}^{my}\oplus\boldsymbol{h}_{len}^{my},\tag{3}$$

为了衡量 2 个向量之间的相似程度，将 $\boldsymbol{h}'^{en}$ 和 $\boldsymbol{h}'^{my}$ 通过向量的点积和向量差的模表示句子相似程度的向量，计算 2 个句子平行的概率。

$$\boldsymbol{h}^1 = \boldsymbol{h}'^{en} \odot \boldsymbol{h}'^{my}, \tag{4}$$

$$\boldsymbol{h}^2 = \left| \boldsymbol{h}'^{en} - \boldsymbol{h}'^{my} \right|, \tag{5}$$

$$\boldsymbol{h}^3 = \tan \boldsymbol{h}(\boldsymbol{W}^1 \boldsymbol{h}^1 + \boldsymbol{W}^2 \boldsymbol{h}^2 + b), \tag{6}$$

$$p(y=1 \mid \boldsymbol{h}^3) = \sigma(\boldsymbol{W}^3 \boldsymbol{h}^3 + c), \tag{7}$$

式中：σ 是 sigmoid 函数；$\boldsymbol{W}^1$、$\boldsymbol{W}^2$、$\boldsymbol{W}^3$、b、c 是模型参数；p 是 2 个向量是平行句对的概率；$y=1$ 代表 2 个句子是平行句对。

对于模型训练的目标是最小化交叉熵损失函数

$$\text{loss} = -\sum_{i=1}^{m} y_i \log(\boldsymbol{W}^3 \boldsymbol{h}_i^3 + c) - (1 - y_i) \log(1 - \sigma(\boldsymbol{W}^3 \boldsymbol{h}_i^3 + c)), \tag{8}$$

m 代表句子对的数量。

3　实验结果及分析

为了验证提出方法的有效性，使用英–缅数据集作为实验数据，以使用孪生神经网络的平行句对抽取模型作为基线模型。

3.1　实验数据集

目前针对缅甸语的开源数据集较少，从亚洲语言树库等网站中收集了部分语料，其余的由人工构建。在的实验中，将英–缅平行语料分成训练集、验证集和测试集，具体如表 2 所示：

表 2　英–缅平行语料规模

Table 2　English-Burmese Parallel Corpus Scale

数据集	Numbers/sentence pairs/k
训练集	490
验证集	1
测试集	3

在训练集中，英语词汇大约共有 100 k 条，缅甸语词汇大约有 45 k 条，格式如下所示：

They assists schools in Yangon, Nay Pyi Taw and Mandalay.
They assists the teachers from universities and colleges to get access to banking systems in order to draw their salaries.

သူတို့သည် ရန်ကုန် နေပြည်တော် နှင့် မန္တလေး မှ ကျောင်းများကို ကူညီ ပေးသည်။
သူတို့သည် တက္ကသိုလ်နှင့် ကောလိပ် များမှ ဆရာ ဆရာမ များကို သူတို့၏ လစာ အတိုင်း ဘဏ် စနစ်ဖြင့် အသုံးပြုမှု ရနိုင်ရန် ကူညီခဲ့သည်။

为了增强分类模型的健壮性，即让模型更好的学习平行句对的分类任务，在使用平行句对作为正样本的同时，随机采样生成负样本（非平行句对），并为每个平行句对和非平行句对生成标签 y，$y=1$ 即为平行，$y=0$ 为不平行。因此，对每一对数据，是由三元组构成的，即英语（源语言）–缅甸语（目标语言）–标签（0 或 1）。正负样本的比例设置为 1∶5，针对训练集来说，则共有 2 944 896 个句子对用来训练模型；6 000 个句子对作为验证集，18 000 个句子对作为测试集。

3.2　实验环境及参数设置

实验的神经网络模型是基于 Tensorflow 实现的，Intel（R）Xeon（R）Gold 6132 CPU @ 2.60GHz，NVIDIA Corporation GP100GL GPU 的服务器上进行。具体实验环境如表 3 所示：

表3 实验环境

Table 3 Experimental Environment

实验环境	Configuration
操作系统	Linux
显卡 GPU	1
显存	16 G
Tensorflow 版本	1.14.0
编程语言	Python3.6

针对基准模型,使用一个单层双向 LSTM 的模型。基线模型的词向量维度设置为 512 维,全连接的隐藏层具有 256 个单元,Batch size 设置为 128,训练轮次为 15 个 epoch。

方法的实验参数设置如表 4 所示。

表4 实验参数设置

Table 4 Training Parameter Setting

参数	Configuration
词嵌入	768
Bi-LSTM 层数	1
隐藏层单元	256
批次大小	128
Dropout	0.2
学习率	1e-4
迭代轮次	15

3.3 实验评价标准

采用精确率、召回率和 F_1 值来评价提出方法的模型性能。精确率(Precision)是真正抽取的平行句子对与所有抽取的句子对的比例;召回率(Recall)是真正抽取的平行句子对与数据集中所有平行句子对的比例,F_1 值是精确度和召回率的调和平均值,具体公式如下所示

$$\text{Precision}=\frac{\text{真正抽取的平行句子对}}{\text{真正抽取的平行句子对}+\text{抽取的非平行句子对}}\ ;$$

$$\text{Recall}=\frac{\text{真正抽取的平行句子对}}{\text{真正抽取的平行句子对}+\text{数据集中未被抽取的平行句子对}}\ 。$$

$$F_1=\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}。$$

3.4 实验结果及分析

实验一:与基线模型对比

为了比较模型的性能,与使用机器学习的分类模型以及 Bi-LSTM 模型(基线模型)做了比较,实验结果如表 5 所示。

1)机器学习方法:传统的支持向量机(SVM)和线性回归(LR)模型。

2)神经网络方法:使用 Bi-GRU、LSTM、Bi-LSTM 的网络结构训练英–缅平行句对抽取模型。

3）方法：融合预训练语言模型及结构特征的英–缅平行句对抽取模型。

从表 5 中可以看出，使用深度学习方法的 Bi-LSTM 模型与机器学习的支持向量机模型（SVM）和线性回归（LR）分类模型相比具有更好效果，主要原因是 Bi-LSTM 模型可以更好的学习句子向量的特征，并且孪生网络将 2 种语言共享到同一语义空间中可一定程度解决跨语言的问题而机器学习方法无法解决跨语言问题使效果明显下降；基线模型的效果为 77.33%，而 MBERT＋Bi-LSTM 的方法的 F_1 值达到了 79.90%，提高了 2.57%。研究的方法 F_1 值达到了 80.84%，比基线模型提高了 3.51%，实验结果表明，在不相关语言对之间直接将 2 种语言共享语义空间会丢失语义信息从而导致模型性能下降，融合 MBERT 预训练模型可以提高平行抽取模型性能，而提出的结构特征一致性约束平行句对抽取方法达到了最好的效果。

为了探究实验在训练过程中不同迭代次数的效果，图 3、4 分别展示了随着迭代次数的增加，F_1 值和损失的变化情况，并与基线模型做了对比。

表 5　与基线模型对比

Table 5　Compared to Baseline Model　　%

Model	P	R	F_1
SVM	62.19	68.17	65.04
LR	57.51	62.78	60.03
LSTM	76.16	73.06	74.58
Bi-GRU	78.72	73.53	76.03
Bi-LSTM	79.27	75.48	77.33
MBert＋Bi-LSTM	82.42	77.53	79.90
研究方法	83.91	78.00	80.84

从图 3、4 中可以看出，基线模型和 MBERT＋Bi-LSTM 模型在第 14 个 epoch 都表现最好，F_1 值分别达到了 87.11%，90.09%，损失值分别达到了最低 0.005 243，0.003 311，方法同样在第 14 个 Epoch 时表现最好，F_1 值达到了 90.44%，损失值最低为 0.003 252，比较模型在最好训练轮次时的结果，证明方法与基线模型相比有明显的提升。

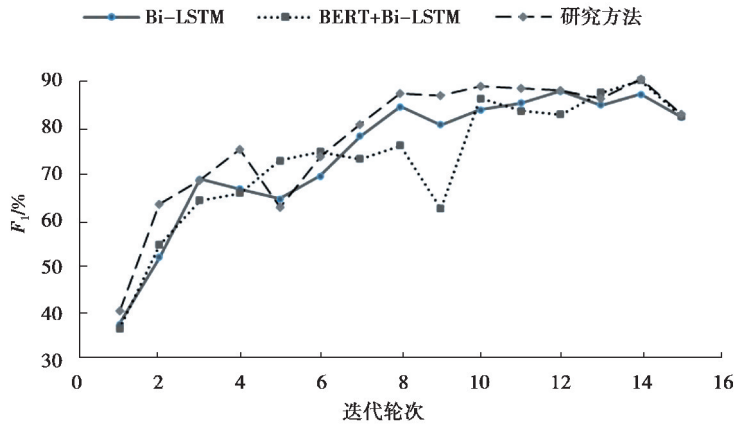

图 3　F_1 值随迭代次数的变化情况

Fig. 3　F_1 score changes with the number of iterations

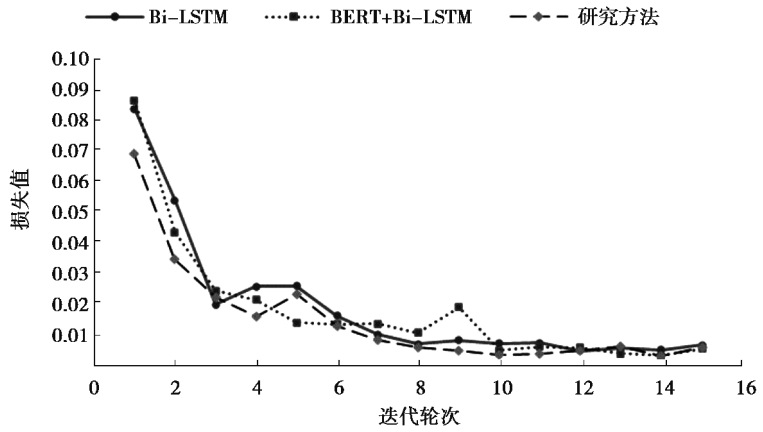

图4 损失值随迭代次数的变化情况

Fig. 4 Loss changes with the number of iterations

实验二:不同预训练词向量方法对比

在预训练的方法上,与使用随机初始化词向量的方法以及传统的 Word2Vec 方法做了对比,实验结果如表 6 所示

表6 不同预训练方法的对比

Table 6 Comparison of Different Pre-training %

方法	P	R	F_1
随机初始化词向量	79.27	75.48	77.33
Word2vec	81.34	75.61	78.37
BERT	82.42	77.53	79.90

从表 6 中可看出,使用传统的词向量训练方法对模型性能有一定提升,使用 MBERT 预训练词向量的方法达到了最好效果。随机初始化的向量并不能充分的表示语义信息,使用 Word2vec 的方法可以保留相对完整的语义信息,而 MBERT 的方法是在大规模的语料上训练,使用双向 Transformer 编码结构,不仅保留完整的语义信息,还考虑了上下文信息,所以结合模型的效果更好。

实验三:预训练词向量调优实验对比

为了探究词向量在平行句对抽取任务中是否微调对模型性能的影响,比较直接使用 MBERT 预训练词向量和初始化词向量并在任务中进行微调的实验结果,如表 7 所示。

表7 词向量微调对实验结果的影响

Table 7 Influence of Fine-tuning on Results %

Method	*P*	*R*	F_1
预训练	81.13	78.30	79.69
预训练+微调	83.91	78.00	80.84

实验结果表明,模型在训练过程中结合任务对词向量的微调可以得到适合任务的更好向量表示,使模型更准确分辨是否是平行句子。

实验四:融合结构特征的实验

在抽取平行句对实验中,为证明融合结构特征方法的有效性,对从维基百科中获取句子分别进行打分,得到效果对比如图 5—6 所示。

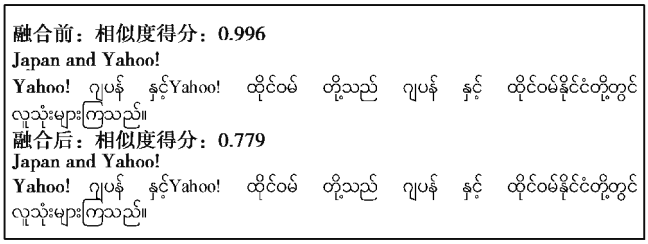

图 5　包含关系句对融合前后相似度比较

Fig. 5　Comparison of Similarity of Semantic Inclusion Sentences

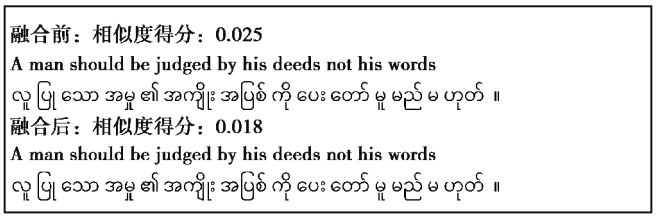

图 6　非平行句对融合前后相似度比较

Fig. 6　Comparison of Similarity of Non-parallel Sentences

从图 5 中可以看出，针对于仅有语义包含关系但不平行的句对，不融合结构特征的相似度得分为 0.996，融合了结构特征，相似度得分为 0.779，降低了模型对这类句对的误判率，同时图 6 中表明，针对不平行句对的相似度影响不大。

4　结　论

针对英语和东南亚语语言差异较大的平行句对抽取问题，对传统方法进行扩展，提出结构特征一致性约束的抽取方法，通过对两种语言语义以及长度特征约束，减少两种语言差异。实验结果表明，提出的模型优于基线模型，并且在东南亚低资源语言上具有推广性。

参考文献：

[1] 曹建文，万福成.面向自动问答系统的问句相似度计算研究[J].重庆大学学报，2019，42(9)：114-122.
Cao J W, Wan F C. Question similarity computing methid for antomatic question answering system[J]. Journal of Chongqing University, 2019, 42(9): 114-122.(in Chinese)

[2] Smith J, Quirk C, Toutanova K. Extracting parallel sentences from comparable corpora using document level alignment [C]// Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics. Los Angeles, California, June 1-6, 2010. Stroudsburg: ACL, 2010: 403-411.

[3] Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [J]. ArXiv: Computation and Language, 2018.

[4] Munteanu D S, Marcu D. Extracting parallel sub-sentential fragments from non-parallel corpora[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, July 17-21, 2006. Stroudsburg: ACL, 2006: 81-88.

[5] Zhao B, Vogel S. Adaptive parallel sentences mining from web bilingual news collection[C]// 2002 IEEE International Conference on Data Mining, 2002. Proceedings. Maebashi City, Japan, December 9-12, 2002. Piscataway: IEEE, 2002: 745-748.

[6] Munteanu D S, Marcu D. Improving machine translation performance by exploiting non-parallel corpora[J]. Computational Linguistics, 2005, 31(4): 477-504.

[7] BarróN-Cedeno A, Espana-Bonet C, Boldoba J, et al. A factory of comparable corpora from wikipedia[C]// Proceedings of the Eighth Workshop on Building and Using Comparable Corpora. Beijing, China, July 30, 2015. Stroudsburg: ACL,

2015：3-13.

[8] Tillmann C，Xu J. A simple sentence-level extraction algorithm for comparable data[C]// Proceedings of Human Language Technologies：The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics，Companion Volume：Short Papers. Boulder，Colorado，May 31-June 5，2009. Stroudsburg：ACL，2009：93-96.

[9] Chu C，Dabre R，Kurohashi S. Parallel sentence extraction from comparable corpora with neural network features[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation(LREC'16). Portorož，Slovenia，May 23-28，2016. ELRA：LREC，2016：2931-2935.

[10] Irvine A，Callison-Burch C. Combining bilingual and comparable corpora for low resource machine translation[C]// Proceedings of the eighth workshop on statistical machine translation. Sofia，Bulgaria，August 8-9，2013. Stroudsburg：ACL，2013：262-270.

[11] Afli H，Barrault L，Schwenk H. Multimodal comparable corpora as resources for extracting parallel data：Parallel phrases extraction[C]// Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya，Japan，October 14-19，2013. Asian Federation of Natural Language Processing：IJCNLP，2013：286-292.

[12] Bouamor H，Sajjad H. Parallel sentence extraction from comparable corpora using multilingual sentence embeddings[C]// Proc. Workshop on Building and Using Comparable Corpora. Miyzaki，Japan May 8，2018. Stroudsburg：ACL，2018：298-305.

[13] GréGoire F，Langlais P. A deep neural network approach to parallel sentence extraction[J]. ArXiv Preprint ArXiv：1709.09783，2017.

[14] Ramesh S H，Sankaranarayanan K P. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora[J]. ArXiv Preprint ArXiv:1806.09652，2018.

[15] Grover J，Mitra P. Bilingual word embeddings with bucketed cnn for parallel sentence extraction[C]// Proceedings of ACL 2017，Student Research Workshop. Vancouver，Canada，July 30-August 4，2017. Stroudsburg：ACL，2017：11-16.

[16] Ma X，Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. ArXiv preprint ArXiv：1603.01354，2016.

[17] Pennington J，Socher R，Manning C D. Glove：Global vectors for word representation[C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha，Qatar，October 25-29，2014. Stroudsburg：ACL，2014：1532-1543.

[18] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[J]. ArXiv Preprint ArXiv:1301.3781，2013.

[19] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]// Advances in neural information processing systems. Long Beach，CA，USA，Dec 4-9，2017. CA，USA：IEEE，2017：5998-6008.

[20] 许晓泓,何霆,王华珍,等.结合 Transformer 模型与深度神经网络的数据到文本生成方法[J].重庆大学学报,2020,43(7)：91-100.

Xu X H，He T，Wang H Z，et al. Research on data-to-text generation basecl on transformer model and deep neural network[J]. Journal of Chongqing University，2020，43(7)：91-100.(in Chinese)

[21] Liu Y. Fine-tune BERT for extractive summarization[J]. arXiv preprint arXiv:1903.10318，2019.

[22] Pires T，Schlinger E，Garrette D. How multilingual is Multilingual BERT[J]. ArXiv Preprint ArXiv:1906.01502，2019.

[23] Wu S，Dredze M. Beto，bentz，becas：The surprising cross-lingual effectiveness of bert[J]. ArXiv Preprint ArXiv：1904.09077，2019

（编辑 侯 湘）