基于孪生对比网络的汉语-东南亚语言多语言平行句对抽取

周远卓[1,2] 毛存礼[1,2] 沈 政[1,2] 张思琦[1,2] 余正涛[1,2] 王振晗[1,2]

摘 要 平行句对抽取应用在东南亚稀缺资源语言上性能不佳,主要原因在于缺少训练语料,导致句对抽取模型表征能力较差. 因此,文中提出基于孪生对比网络的汉语-东南亚语言多语言平行句对抽取方法,从模型结构、训练策略与数据三方面提升性能. 首先,提出孪生对比网络框架,将对比学习思想应用到孪生网络中,增强模型对平行句对的表征能力. 然后,引入相似语言联合训练策略,有效进行知识共享,提高模型的学习能力. 最后,通过多语言词替换的方式构造汉语-混合东南亚语言平行句对,为训练提供较充分的样本信息. 在汉语-泰语和汉语-老挝语数据集上的实验表明,文中方法可有效提升平行句对抽取性能.

关键词 平行句对抽取, 对比学习, 联合训练, 孪生网络

引用格式 周远卓,毛存礼,沈 政,张思琦,余正涛,王振晗. 基于孪生对比网络的汉语-东南亚语言多语言平行句对抽取. 模式识别与人工智能, 2023, 36(10): 931-941.

DOI 10.16451/j.cnki.issn1003-6059.202310006 **中图法分类号** TP 391.1

Siamese Contrastive Network Based Multilingual Parallel Sentence Pair Extraction between Chinese and Southeast Asian Languages

ZHOU Yuanzhuo[1,2], MAO Cunli[1,2], SHEN Zheng[1,2], ZHANG Siqi[1,2], YU Zhengtao[1,2], WANG Zhenhan[1,2]

ABSTRACT The poor performance of parallel sentence pair extraction application on Southeast Asian languages with scarce resources is primarily due to the weak representation capabilities of the sentence pair extraction models caused by the lack of training corpora. Therefore, a siamese contrastive network based multilingual parallel sentence pair extraction between Chinese and Southeast Asian languages is proposed to optimize model structure, training strategy and data. Firstly, a siamese contrastive network framework is employed, integrating contrastive learning concept into the siamese network to enhance the representation capability for parallel sentence pairs. Next, a strategy of joint training with similar languages is introduced to share knowledge effectively and improve the learning ability of the model. Finally, Chinese-mixed Southeast Asian parallel sentence pairs are constructed by multilingual word replacement, providing abundant sample information for training. Experiments on Chinese-Thai and Chinese-

收稿日期:2023-09-06;录用日期:2023-10-27
Manuscript received September 6, 2023;
accepted October 27, 2023
国家自然科学基金项目(No. 62166023,U21B2027,61972186)、云南省科技重大专项项目(No. 202103AA080015,202203AA080004,202302AD080003)、云南省基础研究计划项目(No. 202301AT070471)资助
Supported by National Natural Science Foundation of China(No. 62166023,U21B2027,61972186), Major Science and Technology Projects of Yunnan Province(No. 202103AA080015,202203AA080004,202302AD080003), Yunnan Fundamental Research Projects(No. 202301AT070471)
本文责任编委 马少平
Recommended by Associate Editor MA Shaoping

1. 昆明理工大学 信息工程与自动化学院 昆明 650504
2. 昆明理工大学 云南省人工智能重点实验室 昆明 650504
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504
2. Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming 650504

Lao datasets demonstrate that the proposed method effectively enhances the performance of parallel sentence pair extraction.

Key Words Parallel Sentence Pair Extraction, Contrastive Learning, Joint Training, Siamese Network

Citation ZHOU Y Z, MAO C L, SHEN Z, ZHANG S Q, YU Z T, WANG Z H. Siamese Contrastive Network Based Multilingual Parallel Sentence Pair Extraction between Chinese and Southeast Asian Languages. Pattern Recognition and Artificial Intelligence, 2023, 36(10): 931-941.

平行句对是构成互译关系的双语句子对,在机器翻译[1-3]、多语言大模型构建[4-5]、跨语言图像描述[6]、跨语言摘要[7]等任务上起到重要作用. 平行句对抽取旨在从可比语料中抽取平行句对,能有效扩充东南亚语言等低资源语言的双语平行语料. 现有的平行句对抽取方法在丰富资源语言对(如英语-法语、英语-德语)上有较好表现,因为大量的数据能够为平行句对抽取模型提供足够的泛化能力. 然而,在低资源语言对上,由于训练数据稀缺,平行句对抽取模型难以对句子进行有效表征. 因此,现有的平行句对抽取模型在东南亚语言上难以取得较优效果.

平行句对抽取方法主要分为有监督平行句对抽取方法和无监督平行句对抽取方法.

有监督平行句对抽取方法使用有标签平行双语数据训练句对抽取模型,模型性能高度依赖训练数据.

有监督平行句对抽取方法可分为基于统计规则的方法、基于神经网络的方法和利用机器翻译的方法. Smith 等[2]使用文档级对齐,提升平行句子对齐效果,并且使用维基百科的注释和词汇模型作为模型训练的额外特征. Grégoire 等[3]通过端到端神经网络模型判断句对是否平行. Grover 等[8]利用相似度矩阵,将不同长度的句子表征汇集到一个固定维度的矩阵中,并通过卷积神经网络(Convolutional Neural Network, CNN)判断句子是否平行. Bouamor 等[1]结合句子级嵌入、神经机器翻译和有监督分类方法,再利用神经机器翻译模型或二元分类模型选择最佳平行句子. Zhu 等[9]提出 PHAN(Parallel Hierarchical Attention Network),构造并行层次注意力神经网络,对单语句子和双语句子进行编码,然后使用分类器抽取并行句子.

上述方法改进有监督数据特征提取方法和网络结构,提升平行句对抽取性能,但受限于训练数据规模,有监督平行句对抽取方法在东南亚语言上难以建模高质量表征,无法实现较好的平行句对抽取效果.

无监督平行句对抽取方法引入外部知识或迁移学习,构建平行句对抽取模型,分为有训练双语词嵌入[10-11]的方法、微调预训练语言模型[12-14]的方法. Hangya 等[10]利用无监督方式训练的双语词嵌入针对双语平行关系进行建模,结合正则化相似性和动态阈值方法,确定候选句子对是否平行. Hangya 等[11]检测候选句对中连续的平行部分,利用两种语言中单词的余弦相似度抽取双语平行句对. Keung 等[12]将 mBERT(Multilingual Bidirectional Encoder Representations from Transformers)[4]生成的句子表示用于最近邻搜索,并使用自训练方式训练模型. Kvapilíková 等[13]仅依赖单语数据推导多语言句子嵌入,使用无监督机器翻译模型生成伪平行数据并微调 XLM(Cross-Lingual Masked Language Model)预训练模型,再利用 XLM 生成的多语言句子表示确定候选句子对是否平行. Tien 等[14]假设跨语言对齐策略是可迁移的,使用预训练模型 XLM-R[15]和对抗网络,在一个语言对的平行数据上学习对齐策略,再将该模型用于其它语言对的平行句对抽取.

上述方法通过融入外部知识与知识迁移等无监督方法提升平行句对抽取性能. 但是,由于存在任务差异性,无监督知识迁移构建的平行句对抽取方法存在较大的性能损失,不能直接有效应用在东南亚语言平行句对抽取任务上.

综上所述,如何有效利用额外语言特征和建模高质量表征以提升低资源平行句对抽取模型性能是本文面临的关键挑战.

东南亚语言具有特殊语言特点. 泰语和老挝语作为典型的东南亚稀缺资源语言,具有发音和语法构成上的相似性[16]. 例如:泰语句子

เธอเป็นเด็กผู้หญิงที่สวยงาม

和老挝语句子

ລາວເປັນເດັກຍິງທີ່ງາມ

的释义都为“她是个美丽的女孩”,分别对泰语句子

和老挝语句子进行分词,对应的分词结果和句子中每个词语的释义分别为

เธอ-她/เป็น-是/เด็กผู้หญิง-女孩/ที่-个/สวยงาม-美丽,

ເພິ່ນ-她/ເປັນ-是/ເດັກຍິງ-孩子/ທີ່-个/ງາມ-美好,

对其进行国际音标转写,转写后的泰语发音和老挝语发音分别为

ˈtʰɤː/peːna/deːkpʰuːhajiŋa/tʰiː/sowjaŋaːmˈ,

ˈpʰɤn/pen/dekɲiŋ/tʰiː/ŋaːmˈ,

由此可见泰语和老挝语具有很高的发音相似性.

此外,泰语和老挝语同属分析语言,遵循“主语+谓语+宾语(Subject-Verb-Object, SVO)”的语序规则且形容词后置,具有相似的句法规则和相同的词性.因此,基于这些语言相似特征能够有效实现知识共享,提高汉语-东南亚语言句对抽取模型的学习能力.

另一方面,对比学习是一种有效的表示学习方法,着重于学习同类实例之间的共同特征,区分非同类实例之间的不同之处.对比学习最早应用在计算机视觉领域,在无监督视觉表示学习中表现良好[17-19].无监督视觉表示学习认为良好的视觉表示应将同类实例在表示空间中建模在一起,同时远离其它实例表示.基于这种思想,视觉对比表示学习应使用图像转换方法(如裁剪、旋转、裁剪等),为每幅图像随机生成两个增强版本作为正实例,并通过对比学习,使它们在表示空间中更接近,从而优化视觉表示.

对比学习在自然语言处理任务中也具有广泛应用.在预训练语言模型领域,Zhang 等[20]提出 IS-BERT(Info-Sentence BERT),在 BERT 的基础上添加一维 CNN 层,通过最大化全局句子嵌入与其相应的局部上下文嵌入之间的互信息(Mutual Information, MI)训练 CNN. Fang 等[21]提出 CERT(Con-trastive Self-Supervised Encoder Representations from Transformers),采用与 MoCo(Momentum Contrast)[18]类似的结构,并使用反向翻译进行数据增强.

在跨语言研究领域,Wang 等[22]使用 Dual Momentum Contrast,对齐跨语言句子表示. Cheng 等[23]提出 CiCo(Cross-Lingual Contrastive Learning),通过对比学习的方式,将文本和手语视频映射至联合嵌入空间,同时学习识别细粒度的手语到单词的跨语言映射.陈庆宇等[24]提出基于伪孪生网络双层优化的对比学习算法(Contrastive Learning Based on Bi-level Optimization of Pseudo Siamese Networks, CLBO),将对比学习应用到伪孪生网络结构的教师-学生网络中,引导教师网络更好地指导学生网络学习.陈谨雯等[25]提出用于多跳阅读理解的双视图对比学习网络(Dual View Contrastive Learning Networks, DVCGN),将对比学习应用到多跳阅读理解任务上,通过节点级正负样本对比学习任务获取丰富的上下文互信息. Pan 等[26]提出 mRASP2,将对比学习引入多语言机器翻译领域,使用对比学习优化编码表示,将语义相同的句子拉近到表示空间中的相邻位置.

因此,在汉语-东南亚语言平行句对抽取任务中引入对比学习能够优化汉语表征和东南亚语言表征.

基于上述分析,本文提出基于孪生对比网络的汉语-东南亚语言多语言平行句对抽取方法.在训练数据有限的情况下,充分利用语言相似特征和高质量表示学习,提升平行句对抽取模型的性能.具体地,设计基于语言相似特征的数据增强方法,使用多语言词替换方式构造汉语-混合东南亚语言平行句对,进行数据扩充.提出孪生对比网络框架,使用 XLM-R[15]多语言预训练模型表征文本输入,并进行表征增强训练.引入联合训练策略,实现知识共享.最后,构建汉-泰和汉-老平行句对数据集,并在此基础上进行实验,结果表明本文方法可以有效提升汉语-东南亚语言平行句对抽取质量.

1　基于孪生对比网络的汉语-东南亚语言多语言平行句对抽取

本文构建基于孪生对比网络的汉语-东南亚语言多语言平行句对抽取方法,结构如图 1 所示.在数据增强、网络结构设计、知识共享三个层面改进抽取性能.

本文方法包含如下部分:基于语言相似特征的数据增强方法、孪生对比网络、相似语言联合训练策略.基于语言相似特征的数据增强方法通过多语言词替换方式构造汉语-混合东南亚语言平行句对,进行数据扩充.孪生对比网络引入预训练词嵌入,结合对比学习,实现平行句对抽取.相似语言联合训练策略选用具有相似性的东南亚语言构成的汉语-东南亚语言对进行联合训练,有效利用语言相似性,实现知识共享.

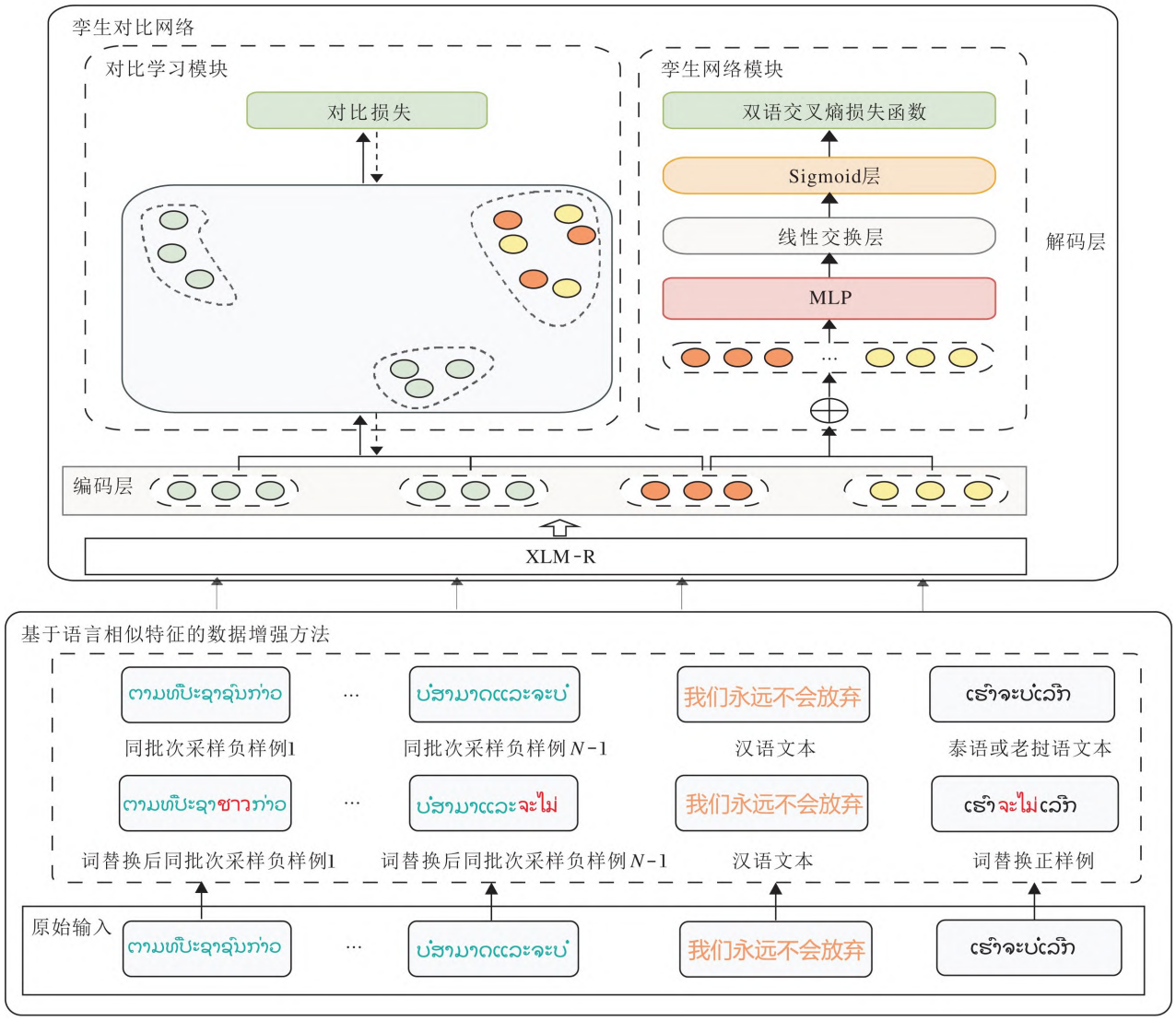

图 1 本文方法整体结构图

Fig. 1 Overall structure of the proposed method

1.1 任务定义

平行句对抽取任务通过计算双语句对相似度判断句对是否平行. 对于待判断的双语句对 (x,y),首先将双语句对分别送入编码层,建模双语句对 x、y 的句子表征:

$$h\_x = Encoder(x),$$

$$h\_y = Encoder(y).$$

获得 h_x 和 h_y 后,为了获取双语句子的相关性,计算双语句子表征的余弦相似度,利用余弦相似度对文本匹配关系进行打分,根据匹配分数

$$score = cosine\_similarity(h\_x, h\_y)$$

判断句对是否平行. 或将建模的双语句子表征直接送入二分类神经网络,通过分类网络得到句对是否平行的判断结果:

$$Result = Classifier(h\_x, h\_y).$$

1.2 孪生对比网络

孪生网络[27]是深度学习模型的一种,网络包含 2 个结构相同的子网络,可分别对 2 段文本序列进行编码,获取包含文本语义特征的 2 个固定长度的向量表示. Reimers 等[28]已证明孪生网络可以有效计算两段文本序列的语义相似度. 本文在孪生网络的基础上进行改进,引入预训练词嵌入,结合对比学习构建孪生对比网络. 孪生网络模块的优化目标是拉近平行句对的表征,对比学习模块的优化目标是拉远非平行句对的表征.

为了缓解训练数据稀缺问题,更有效地对文本进行语义表示,本文首先引入 XLM-R 预训练模型对文本进行编码,每段文本序列分别利用一个高维向

量进行表示.

对于给定的中文文本序列

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n),$$

其中 n 为序列长度,将其输入 XLM-R 进行编码,得到稠密的隐向量 $\boldsymbol{u} = F(\boldsymbol{x})$,其中 $F(\cdot)$ 表示 XLM-R 编码层.

对于给定的东南亚语言文本序列

$$\boldsymbol{y} = (y_1, y_2, \cdots, y_b),$$

其中 b 为序列长度,其处理过程与中文文本一致,利用 XLM-R 对其进行编码,得到稠密的隐向量 $\boldsymbol{v} = F(\boldsymbol{y})$,其中 $F(\cdot)$ 表示 XLM-R 编码层.

下面将编码表示分别输入孪生网络模块和对比学习模块,优化模型.

对于孪生网络模块,将平行句对的编码表示输入孪生网络模块,通过孪生网络对两段文本的语义表示进行特征匹配,计算语义相似度. 本文使用多层感知机(Multilayer Perceptron, MLP)学习匹配关系.

对于上述步骤得到的编码表示 $\boldsymbol{u}$、$\boldsymbol{v}$,首先将 $\boldsymbol{u}$、$\boldsymbol{v}$ 以及二者的差 $\boldsymbol{u} - \boldsymbol{v}$ 和按位相乘 $\boldsymbol{u} \times \boldsymbol{v}$ 进行拼接,将拼接结果 $[\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} \times \boldsymbol{v}]$ 输入 MLP 的线性变换层 $\boldsymbol{W}_1$,利用 tanh 激活函数提高模型表征能力,更好捕获文本隐向量表示中的匹配关系,由此得到包含文本匹配关系的隐向量:

$$\boldsymbol{h} = \tanh(\boldsymbol{W}_1[\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} \times \boldsymbol{v}]).$$

为了对 $\boldsymbol{h}$ 进行分类,本文将 $\boldsymbol{h}$ 输入一个线性变换层 $\boldsymbol{W}_2$,进行特征压缩,再通过一个 sigmoid 层对文本匹配关系进行打分,即

$$s = sigmoid(\boldsymbol{W}_2\boldsymbol{h}).$$

本文使用双语交叉熵损失对基础的孪生网络进行优化训练,损失函数如下:

$$L_b = -[a \ln s + (1 - a)\ln(1 - s)],$$

其中 a 为 $\boldsymbol{h}$ 对应的输入文本标签.

对于对比学习模块,通过同批次采样的方式构造负样例,并进行负样本对比学习. 对于给定的汉语-泰语或汉语-老挝语训练句对 (x_i, y_i),将中文句子和同批次的其它任意泰语和老挝语句对作为负样例对,其核心思想是将任意非平行的句对构造为负样例对. 在模型训练过程中,从所有的训练数据中随机采样 n 个句对作为一个批次的训练数据,即

$$sample = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}.$$

对于其中任意一个训练样本 (x_i, y_i) 中的中文句子 x_i,将 x_i 分别和同批次的其它 $n-1$ 个泰语或老挝语句子进行配对,得到 $n-1$ 个非平行句对 (x_i, y_j) 作为实验的负样本句子对,进行负样本对比学习训练,从而使非平行句对的语义差距尽可能大. 该方式构造的负样本训练对比损失如下:

$$L_1 = -\frac{1}{n-1}\sum_{\substack{j=1 \\ j \neq i}}^{n} \ln(1 - s_{i,j}),$$

其中 $s_{i,j}$ 表示负样本句子对 (x_i, y_j) 的语义相似度.

相比传统的对比学习方法需要同时构造正样例和负样例的情况,本文的孪生对比网络中孪生网络模块和正样本对比学习的优化目标一致,都是拉近平行句对的语义相似度,因此对比学习模块不需要单独设置正样本句子对.

1.3　基于语言相似特征的数据增强方法

为了利用东南亚语言的语言相似性,将其应用在数据增强方法上,本文有效扩充训练数据. 受对比学习正负样例构建的启发,本文提出基于语言相似特征的数据增强方法. 参照孪生对比网络的输入 (x_i, y_i),对汉语-东南亚语言双语句对中的东南亚语言进行多语言同义词替换,构建一部分伪数据正样例 (x_i, y_i'),和 1.2 节同样的方式,使用该部分数据构造另一部分伪数据负样例 (x_i, y_j'),将上述构造的伪数据送入孪生对比网络中进行训练,使模型获得更健壮的文本表征能力.

本文利用人工标注的泰语-老挝语双语词典对句子进行词替换. 对于任意一个训练样本 (x_i, y_i) 中的泰语或老挝语句子 y_i,首先对其进行分词,再对每个词在双语词典中进行检索,如果能找到对应的同义词,就在原句子中进行替换,得到新的泰语或老挝语句子 y_i',并和原句对中的中文句子 x_i 构成一个新的正样例训练数据 (x_i, y_i'). 泰语和老挝语在句法上具有较高的相似性,多语言词替换的方式并不会对泰语和老挝语词汇在句子中的位置信息造成干扰,伪数据词汇可以学习到和原始数据词汇相似的上下文语义信息,因此构造的伪数据具有较高的质量,保证模型在对新构造的伪数据进行编码时能准确对语义进行表征.

同时,为了构造更多的样例数据,提升模型训练效果,本文在多语言词替换构造的伪数据中也使用如 1.2 节的负样本构造方式一致的同批次数据构造方法,构造更多的负样例数据,因此,该方式构造的样本数据训练时的损失函数如下:

$$L_2 = L_b' + L_1',$$

其中,L_b'和 L_1'的计算方式可参考 L_b 和 L_1 的计算方式,其区别只在于训练数据不同,需要将 y_i(或 y_j)更换为 y_i'(或 y_j').

1.4　相似语言联合训练策略

本文针对东南亚语言相似性特点提出相似语言

联合训练策略,流程如图 2 所示. 充分利用东南亚语言间的相似性在训练中实现知识共享,增强模型的学习能力. 对给定的汉语-泰语或汉语-老挝语平行句对,首先通过 XLM-R 获取输入文本的编码表示,再将汉语的编码表示输入一端的编码器,将泰语和老挝语的编码表示输入另一端的参数共享编码器,利用多语言联合训练增强模型对泰语和老挝语的表征能力. 泰语和老挝语具有较高的相似性,共享泰语编码器和老挝语编码器能够使模型学习到更丰富的东南亚语言知识,建立更强的语言表征能力,此外,泰语和老挝语的相似特点也使得共享编码器后的模型在编码泰语和老挝语时语种间相互干扰较小,因此相似语言联合训练的方式能使模型性能得到有效提升.

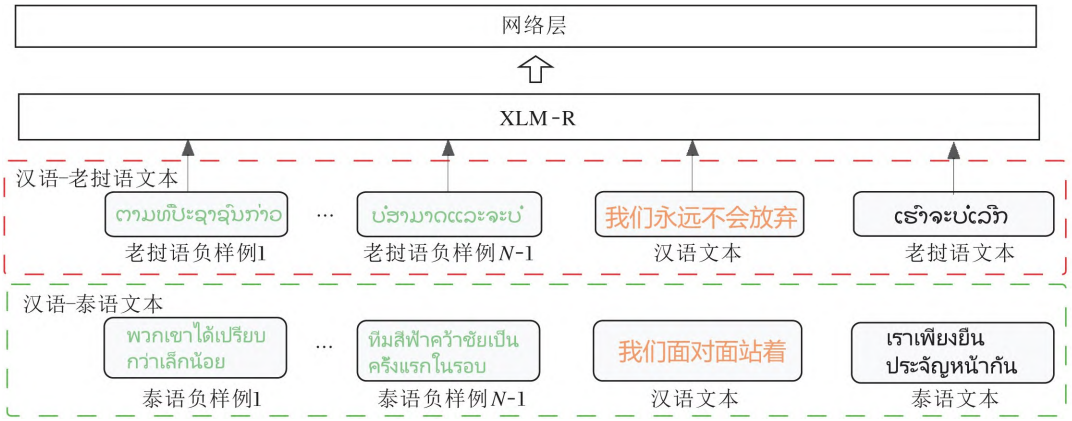

图 2 相似语言联合训练策略流程图

Fig. 2 Flow chart of similar language joint training strategy

在训练过程中,针对每个中文句子输入 x_i,分别有 (x_i, y_i) 的交叉熵损失 L_b,(x_i, y_j) 的对比损失 L_1,(x_i, y_i') 的交叉熵损失 L_b',(x_i, y_j') 的对比损失 L_1',则训练时的总损失函数为:

$$L = L_b + L_1 + L_2 = L_b + L_1 + L_b' + L_1'.$$

2 实验及结果分析

2.1 实验数据集

泰语和老挝语都属于东南亚低资源语言,网络上开源的汉语-泰语(简记为汉-泰)和汉语-老挝语(简记为汉-老)平行数据集稀缺,难以获取足够平行句对抽取模型训练的开源汉-泰平行数据和汉-老平行数据. 因此,本文首先从 OPUS[29] 和亚洲语言树库[30] 上获取少量的汉-泰和汉-老平行语料,通过爬虫技术结合人工标注,预先构建一定数量的汉-泰和汉-老双语平行句对数据集,满足平行句对抽取模型的训练需要. 训练数据的数据量统计和数据集划分如表 1 所示.

为了使用上述数据训练本文的孪生对比网络,增强平行句对抽取模型的文本表示能力,本文利用上述平行句对数据,按照 1∶1 的比例构建一部分负样例. 对每个汉语句子采用随机采样的方式配对一个不平行的泰语或老挝语句子,并对该句对赋予不平行标签,对原有平行句对赋予平行标签,即模型的每条训练数据都是一个句对三元组

〈汉语句子,泰语或老挝语句子,平行标签〉.

此外,还根据基于语言相似特征的数据增强方法构建等比例的伪数据.

表 1 训练语料信息

Table 1 Description of training corpus

	数据集	句对数量
汉-泰	训练集	196000
	验证集	2000
	测试集	2000
汉-老	训练集	96000
	验证集	2000
	测试集	2000

2.2 对比方法与评价指标

本文选择如下 7 种基线方法进行对比.

1) BiRNN(Bidirectional Recurrent Neural Network)[3]. 利用循环神经网络捕获文本上下文语义信息,实现文本的语义分类.

2) mBERT[4]. 利用多语言 BERT 预训练模型获取语义表示,计算双语句子的语义相似度,根据相似度判断是否平行.

3)SVM[31]. 分类领域中常见的二分类方法.

4)LR(Logistic Regression)[32]. 经典机器学习算法,回归模型的一种.

5)文献[33]方法. 利用多语言 BERT 预训练模型获取语义表示,再将语义表示送入 Bi-LSTM 中,基于语义相似度建模分类任务,实现对平行句对的判断.

6)LaBSE[34]. 谷歌提出的用于表示语言无关句子的模型. 应用 LaBSE 句子嵌入模型实现网络并行文本挖掘,通过 LaBSE 获取语言无关句子表示,对双语句子的语义相似度进行计算,判断句对是否平行.

7)E5(Embeddings from Bidirectional Encoder Representations)[35]. 基于弱监督对比学习训练文本嵌入表示. 利用 E5 获取语义表示,并计算语义相似度,实现对平行句对的判断.

本文使用精确率(Precision)、召回率(Recall)和 F1 值(F1)作为评价指标. 精确率是指所有抽取的句对中平行句对的比例,召回率是指从数据集上所有平行句对中抽取的平行句对的比例,F1 值是精确率和召回率的调和平均值. 评价指标计算公式如下:

$$P = \frac{|TP|}{|TP + FP|} \times 100\%,$$

$$R = \frac{|TP|}{|TP + FN|} \times 100\%,$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%,$$

其中,TP 表示真的正例,FP 表示假的正例,FN 表示假的反例,TN 表示真的反例.

2.3　实验设置

本文所有实验的深度学习模型都是基于 PyTorch 1.8.1 实现的,并且在单个 NVIDIA 3090 GPU 上进行实验. 学习率设置为 5e-6,参照 Vaswani 等[36]设置热启动 $warm\_steps = 500$ 的 warm-up 策略动态调整学习率,每个训练批次包含大约 32 个三元组.

2.4　实验结果

2.4.1　对比实验

本节将本文方法与 7 种基线方法进行对比,实验结果如表 2 所示:句对类型表示模型的句对抽取方向,汉-泰表示方法用于汉-泰平行句对抽取,汉-老表示方法用于汉-老平行句对抽取.

由表 2 可知,相比传统的机器学习方法 SVM 和 LR,本文方法在汉-泰和汉-老两个测试集上的抽取效果都取得较大提升,说明本文方法可以更好地从现有训练数据中学习语言的语义特征并实现良好的知识共享. 而传统的基于机器学习的方法依赖训练数据中的特征,泛化能力较差,由于泰语和老挝语都属于低资源语言,模型训练过程中缺乏充足的训练数据,难以学习到充分的语义表征,导致性能不佳.

表 2　各方法在平行句对抽取任务上的性能对比

Table 2　Performance comparison of different models for parallel sentence extraction

%

方法	句对类型	精确率	召回率	F1 值
SVM	汉-泰	72.31	75.67	73.95
	汉-老	57.17	54.50	55.80
LR	汉-泰	68.18	70.47	69.31
	汉-老	55.87	53.65	54.73
BiRNN	汉-泰	75.68	78.36	76.99
	汉-老	59.58	63.26	61.36
mBERT	汉-泰	77.48	75.46	76.45
	汉-老	56.43	54.37	55.38
文献[33]方法	汉-泰	84.38	83.43	83.90
	汉-老	65.46	67.72	66.57
LaBSE	汉-泰	93.74	86.90	90.19
	汉-老	94.87	91.55	93.18
E5	汉-泰	93.46	88.00	90.65
	汉-老	91.37	92.20	91.78
本文方法	汉-泰	**94.68**	**89.10**	**91.80**
	汉-老	**95.52**	**97.05**	**96.27**

相比现有的基于深度学习的平行句对抽取方法,本文方法在汉-泰和汉-老两个测试集上的抽取效果都达到最优值. 相比 BiRNN,本文引入预训练语言模型,设计孪生对比网络和基于语言相似特征的数据增强方法,提升模型表示能力,在汉-泰和汉-老句对抽取任务上分别提升 14.81% 和 34.91% 的 F1 值. 相比 mBert,本文通过孪生对比网络训练和数据增强方法对模型编码进行优化,在编码时获得更准确的跨语言表征,在汉-泰和汉-老句对抽取任务上分别提升 15.35% 和 40.89% 的 F1 值. 相比文献[33]方法,本文方法从多层面进行优化,学习到更强大的语义表征,具有更好的建模能力,在汉-泰和汉-老句对抽取任务上分别提升 7.88% 和 29.7% 的 F1 值. 相比建立语言无关句子向量的 LaBSE,本文方法针对泰语和老挝语进行优化,提升模型对泰语和老挝语的表征能力,在汉-泰和汉-老句对抽取任务上分别获得 1.61% 和 3.09% 的 F1 值提升. 相比同样基于对比学习训练的 E5,本文方法在预训练模型表征和对比学习之外还进行多层次优化,在汉-泰和汉-老句对抽取任务上分别提升 1.15% 和 4.49% 的 F1 值,达到更好的句对抽取性能.

综上所述,本文方法在汉-泰和汉-老句对抽取上均达到最优值,表明本文方法的有效性.

2.4.2 消融实验

为了探究在平行句对抽取模型中引入相似语言联合训练策略(简记为联合训练)、孪生对比网络(简记为对比网络)、基于语言相似特征的数据增强方法(简记为数据增强)的有效性,本文设置消融实验,在本文方法的基础上分别消除各模块.

上述 3 个模块在汉-泰、汉-老测试集上的实验结果如表 3 和表 4 所示. 训练集语料中标注汉-泰+汉-老是采用相似语言联合训练策略的实验结果.

由表 3 和表 4 可以看出,3 个模块都可以有效提升平行句对抽取模型效果. 相比完整方法,消除相似语言联合训练策略使方法在汉-泰和汉-老上分别下降 7.37% 和 24.71% 的 F1 值,消除孪生对比网络使方法在汉-泰和汉-老上分别下降 0.57% 和 3.13% 的 F1 值,消除基于语言相似特征的数据增强方法使方法在汉-泰和汉-老上分别下降 1.77% 和 3.52% 的 F1 值,而同时消除孪生对比网络和基于语言相似特征的数据增强方法使方法在汉-泰和汉-老上分别下降 5.75% 和 9.54% 的 F1 值.

表 3 在汉-泰测试集上的消融实验结果

Table 3 Results of ablation experiment on Chinese-Thai test set

%

组件部分	训练集语料	精确率	召回率	F1 值
本文方法	汉-泰+汉-老	**94.68**	**89.10**	**91.80**
- 联合训练	汉-泰	83.26	85.64	84.43
- 对比学习	汉-泰+汉-老	90.14	92.35	91.23
- 数据增强	汉-泰+汉-老	89.08	91.00	90.03
- 对比学习和数据增强	汉-泰+汉-老	80.87	91.95	86.05

表 4 在汉-老测试集上的消融实验结果

Table 4 Results of ablation experiment on Chinese-Lao test set

%

组件部分	训练集语料	精确率	召回率	F1 值
本文方法	汉-泰+汉-老	95.52	97.05	96.27
- 联合训练	汉-老	70.78	72.36	71.56
- 对比学习	汉-泰+汉-老	88.30	98.55	93.14
- 数据增强	汉-泰+汉-老	89.41	96.35	92.75
- 对比学习和数据增强	汉-泰+汉-老	78.57	96.80	86.73

分析消融实验结果可得如下结论.

1)相似语言联合训练策略对方法性能提升明显,并且在数据规模更小的汉-老句对上取得比汉-泰句对更大的性能提高,说明相似语言联合训练策略通过共享汉-泰和汉-老中的参数实现跨语言的知识共享,有效提高方法对泰语和老挝语的学习能力,显著提高模型性能.

2)孪生对比网络和基于语言相似特征的数据增强方法都能对方法性能进行有效提升,而利用语言相似特征的数据增强方法对方法性能的贡献更大,主要原因在于泰语和老挝语在句法上具有较高的相似性,构造的伪数据质量较高,将高质量伪数据送入模型进行训练能有效提升模型表征能力.

3)相比去除孪生对比网络和基于语言相似特征的数据增强模块,去除相似语言联合训练策略后的方法性能下降更明显,尤其是对于训练数据更少的汉-老任务. 这主要是因为汉-泰(20 万条)、汉-老(10 万条)训练数据稀少,在低资源场景下,性能更多受限于数据,此时增加不同的数据实例可以有效增强方法性能,而本文的相似语言联合训练策略可以等价为训练过程中针对汉-泰和汉-老任务的数据实例扩充. 对比汉-泰联合训练消融结果和汉-老联合训练消融结果可发现,数据实例的增加对更低资源的汉-老任务提升更大. 而对比联合训练消融结果和数据增强消融结果可发现,相比在原始数据上进行数据扩充的数据增强方法,引入不同数据实例的相似语言联合训练策略对性能提升更大.

2.4.3 不同词嵌入方法对比

为了验证本文将 XLM-R 预训练模型应用在平行句对抽取模型中的优越性,将 XLM-R 与随机初始化、Word2vec 及 mBert 在相同实验条件下进行对比实验.

测试集为泰语和老挝语时不同词嵌入方法的实验结果如表 5 和表 6 所示.

表 5 不同词嵌入方法在汉-泰测试集上的性能对比

Table 5 Performance comparison of different word embedding methods on Chinese-Thai test set

%

词嵌入方法	精确率	召回率	F1 值
随机初始化	52.36	53.68	53.01
Word2vec	66.46	73.48	69.79
mBert	81.52	85.28	83.35
XLM-R	94.68	89.10	91.80

由表 5 和表 6 可知,XLM-R 取得最优值. 相比随机初始化方法,使用预训练方法的模型效果显著提升,说明预训练词嵌入可有效提升模型表征能力. 相

比 Word2vec,多语言预训练模型取得更好效果,说明相比静态词嵌入,动态词嵌入方法可以获得更适合任务场景的词嵌入表示. 相比 mBert,XLM-R 效果更优,说明 XLM-R 可以获得更准确的语义表示.

表 6　不同词嵌入方法在汉-老测试集上的性能对比

Table 6　Performance comparison of different word embedding methods on Chinese-Lao test set

%

词嵌入方法	精确率	召回率	F1 值
随机初始化	56.65	59.72	58.14
Word2vec	62.93	68.51	65.01
mBert	86.27	88.32	87.28
XLM-R	95.52	97.05	96.27

2.5　实例分析

为了进一步验证本文方法能够优化模型学习能力,建模更好的表征以提升句对抽取效果,本节进行实例分析,对比本文方法和 XLM-R 抽取汉-泰和汉-老平行句对的结果,具体如表 7 和表 8 所示.

表 7　汉-泰平行句对抽取结果的实例分析

Table 7　Example analysis of Chinese-Thai parallel sentence pair extraction

泰语	พฤติกรรมอย่างนี้ทำลายความรู้สึกของการเล่นด้วยความยุติธรรม(这种行为违反了公平竞争的原则)
汉语	这种行为违反了规则
标签	不平行
XLM-R	平行
本文方法	不平行

表 8　汉-老平行句对抽取结果的实例分析

Table 8　Example analysis of Chinese-Lao parallel sentence pair extraction

老挝语	ນ້ຳໜັກຊຸ່ນນາຕົກຂອງນ້ຳເຄືອຂ່າຍນ້ອໂຣຈະໄດ້ຮັບຮັບການແກ້ໄຂໃນລັກສະນະອາດ(神经网络突触权重会以一种有序方式进行修改)
汉语	神经网络会以某种方式更新其突触权重
标签	不平行
XLM-R	平行
本文方法	不平行

由表 7 可见,在汉-泰实例上,泰语句子的意思与对应汉语意思较相近,本文方法有效建模泰语句子表示和汉语句子表示,并正确计算它们的相似度,将其判定为不平行句对. 相比本文方法,XLM-R 未利用语言特征和对比表示学习进行多层次优化,难以构建准确的“公平竞争”编码表示,从而导致模型的误判,将其判定为平行句对. 分析表明,本文方法成功增强模型的语义表征能力,可以对文本的相关度进行更准确地建模,从而提升模型抽取效果. 由表 8 也可得到相同结论.

3　结束语

针对现有的双语平行句对抽取方法依赖大规模标注语料库,应用在东南亚稀缺资源语言时效果不佳的问题,本文提出基于孪生对比网络的汉语-东南亚语言多语言平行句对抽取方法,充分挖掘东南亚语言相似性,构建基于语言相似特征的数据增强方法、孪生对比网络与相似语言联合训练策略,增强模型对汉语-东南亚语言的平行句对抽取能力. 实验表明本文方法可有效提升汉语-东南亚语言平行句对抽取质量,在汉语-泰语及汉语-老挝语的平行句对抽取任务上取得较高的 F1 值. 本文方法在数据增强、网络结构设计、知识共享三方面改进抽取性能,在有标注双语数据有限的情况下,有效利用语言相似特征和引入对比表示学习,提升平行句对抽取模型在低资源语言上的性能. 本文的工作为研究东南亚多语言及跨语言任务提供高质量的数据基础. 此外,本文方法也为低资源语言平行句对抽取提供一种思路. 尽管本文方法在汉语-东南亚语言多语言平行句对抽取任务上达到较好效果,但有监督学习方法对双语数据数量仍有一定要求. 而现实中大量的双语方向是极低资源或无资源的,今后将探索本文方法在极低资源语言对上的应用,引入单语自监督学习和强化学习方法,减少模型对双语训练数据的依赖,并进一步提升平行句对抽取性能.

参　考　文　献

[1] BOUAMOR H, SAJJAD H. H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings[C/OL]. [2023-05-16]. http://lrec-conf.org/workshops/lrec2018/W8/pdf/8_W8.pdf.

[2] SMITH J R, QUIRK C, TOUTANOVA K. Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment // Proc of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2010: 403-411.

[3] GRÉGOIRE F, LANGLAIS P. A Deep Neural Network Approach to Parallel Sentence Extraction[C/OL]. [2023-05-16]. https://arxiv.org/abs/1709.09783v1.

[4] PIRES T, SCHLINGER E, GARRETTE D. How Multilingual Is

Multilingual BERT? // Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2019: 4996-5001.

[5] XUE L T, CONSTANT N, ROBERTS A, *et al*. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer // Proc of the Conference of the North American Chapter of the Association for Computational Linguistics(Human Language Technologies). Stroudsburg, USA: ACL, 2021: 483-498.

[6] SONG Z J, HU Z Z, HONG R C. Embedded Heterogeneous Attention Transformer for Cross-Lingual Image Captioning [C/OL]. [2023-05-16]. https://arxiv.org/pdf/2307.09915.pdf.

[7] ZHU J N, WANG Q, WANG Y N, *et al*. NCLS: Neural Cross-Lingual Summarization // Proc of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: ACL, 2019: 3054-3064.

[8] GROVER J, MITRA P. Bilingual Word Embeddings with Bucketed CNN for Parallel Sentence Extraction // Proc of ACL 2017(Student Research Workshop). Stroudsburg, USA: ACL, 2017: 11-16.

[9] ZHU S L, YANG Y, XU C. Extracting Parallel Sentences from Nonparallel Corpora Using Parallel Hierarchical Attention Network. Computational Intelligence and Neuroscience, 2020. DOI: 10.1155/2020/8823906.

[10] HANGYA V, BRAUNE F, KALASOUSKAYA Y, *et al*. Unsupervised Parallel Sentence Extraction from Comparable Corpora // Proc of the 15th International Conference on Spoken Language Translation. Stroudsburg, USA: ACL, 2018: 7-13.

[11] HANGYA V, FRASER A. Unsupervised Parallel Sentence Extraction with Parallel Segment Detection Helps Machine Translation // Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2019: 1224-1234.

[12] KEUNG P, SALAZAR J, LU Y C, *et al*. Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings. Transactions of the Association for Computational Linguistics, 2020, 8: 828-841.

[13] KVAPILÍKOVÁ I, ARTETXE M, LABAKA G, *et al*. Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining // Proc of the 58th Annual Meeting of the Association for Computational Linguistics (Student Research Workshop). Stroudsburg, USA: ACL, 2020: 255-262.

[14] TIEN C, STEINERT-THRELKELD S. Bilingual Alignment Transfers to Multilingual Alignment for Unsupervised Parallel Text Mining // Proc of the 60th Annual Meeting of the Association for Computational Linguistics(Long Papers). Stroudsburg, USA: ACL, 2022: 8696-8706.

[15] CONNEAU A, KHANDELWAL K, GOYAL N, *et al*. Unsupervised Cross-Lingual Representation Learning at Scale // Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2020: 8440-8451.

[16] DING C C, UTIYAMA M, SUMITA E. Similar Southeast Asian Languages: Corpus-Based Case Study on Thai-Laotian and Malay-Indonesian // Proc of the 3rd Workshop on Asian Translation. Stroudsburg, USA: ACL, 2016: 149-156.

[17] CHEN T, KORNBLITH S, NOROUZI M, *et al*. A Simple Framework for Contrastive Learning of Visual Representations // Proc of the 37th International Conference on Machine Learning. San Diego, USA: JMLR, 2020: 1597-1607.

[18] HE K M, FAN H Q, WU Y X, *et al*. Momentum Contrast for Unsupervised Visual Representation Learning // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2020: 9726-9735.

[19] CHEN T, KORNBLITH S, SWERSKY K, *et al*. Big Self-Supervised Models Are Strong Semi-Supervised Learners // Proc of the 34th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2020: 22243-22255.

[20] ZHANG Y, HE R D, LIU Z Z, *et al*. An Unsupervised Sentence Embedding Method by Mutual Information Maximization // Proc of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2020: 1601-1610.

[21] FANG H C, WANG S C, ZHOU M, *et al*. CERT: Contrastive Self-Supervised Learning for Language Understanding [C/OL]. [2023-05-16]. https://arxiv.org/pdf/2005.12766.pdf.

[22] WANG L, ZHAO W, LIU J M. Aligning Cross-Lingual Sentence Representations with Dual Momentum Contrast // Proc of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2021: 3807-3815.

[23] CHENG Y T, WEI F Y, BAO J M, *et al*. CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2023: 19016-19026.

[24] 陈庆宇,季繁繁,袁晓彤. 基于伪孪生网络双层优化的对比学习. 模式识别与人工智能, 2022, 35(10): 928-938.
(CHEN Q Y, JI F F, YUAN X T. Contrastive Learning Based on Bilevel Optimization of Pseudo Siamese Networks. Pattern Recognition and Artificial Intelligence, 2022, 35(10): 928-938.)

[25] 陈谨雯,陈羽中. 用于多跳阅读理解的双视图对比学习网络. 模式识别与人工智能, 2023, 36(5): 471-482.
(CHEN J W, CHEN Y Z. Dual View Contrastive Learning Networks for Multi-hop Reading Comprehension. Pattern Recognition and Artificial Intelligence, 2023, 36(5): 471-482.)

[26] PAN X, WANG M X, WU L W, *et al*. Contrastive Learning for Many-to-Many Multilingual Neural Machine Translation // Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Long Papers). Stroudsburg, USA: ACL, 2021: 244-258.

[27] BROMLEY J, GUYON I, LECUN Y, *et al*. Signature Verification Using a "Siamese" Time Delay Neural Network // Proc of the 6th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 1993: 737-744.

[28] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks // Proc of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: ACL, 2019: 3982-3992.

[29] TIEDEMANN J. Parallel Data, Tools and Interfaces in OPUS //

Proc of the 8th International Conference on Language Resources and Evaluation. Cambridge, USA: MIT Press, 2012: 2214-2218.

[30] RIZA H, PURWOADI M, GUNARSO, *et al*. Introduction of the Asian Language Treebank // Proc of the Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques. Washington, USA: IEEE, 2016. DOI: 10.1109/ICSDA.2016.7918974

[31] PLATT J C. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Technical Report, MSR-TR-98-14. Cambridge, USA: Microsoft Research, 1998.

[32] DREISEITL S, OHNO-MACHADO L. Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review. Journal of Biomedical Informatics, 2002, 35(5/6): 352-359.

[33] 毛存礼,高 旭,余正涛,等.结构特征一致性约束的双语平行句对抽取.重庆大学学报, 2021, 44(1): 46-56.
(MAO C L, GAO X, YU Z T, *et al*. Extraction of Bilingual Parallel Sentence Pairs Constrained by Consistency of Structural Features. Journal of Chongqing University, 2021, 44(1): 46-56.)

[34] FENG F X Y, YANG Y F, CER D, *et al*. Language-Agnostic BERT Sentence Embedding // Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Long Papers). Stroudsburg, USA: ACL, 2022: 878-891.

[35] WANG L, YANG N, HUANG X L, *et al*. Text Embeddings by Weakly-Supervised Contrastive Pre-Training[C/OL]. [2023-05-16]. https://arxiv.org/abs/2212.03533.

[36] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention Is All You Need // Proc of the 31st International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2017: 6000-6010.

作者简介

周远卓,硕士研究生,主要研究方向为自然语言处理、机器翻译. E-mail:550132791@qq.com.
(**ZHOU Yuanzhuo**, master student. His research interests include natural language processing and machine translation.)

毛存礼(通信作者),博士,教授,主要研究方向为自然语言处理、信息检索、机器翻译. E-mail:maocunli@163.com.
(**MAO Cunli**(Corresponding author), Ph. D., professor. His research interests include natural language processing, information retrieval and machine translation.)

沈 政,硕士研究生,主要研究方向为自然语言处理、机器翻译. E-mail: 1591744723@qq.com.
(**SHEN Zheng**, master student. His research interests include natural language processing and machine translation.)

张思琦,博士研究生,主要研究方向为自然语言处理、机器翻译. E-mail:943714686@qq.com.
(**ZHANG Siqi**, Ph. D. candidate. Her research interests include natural language processing and machine translation.)

余正涛,博士,教授,主要研究方向为自然语言处理、信息检索、机器翻译. E-mail:ztyu@hotmail.com.
(**YU Zhengtao**, Ph. D., professor. His research interests include natural language processing, information retrieval and machine translation.)

王振晗,博士研究生,主要研究方向为自然语言处理、机器翻译. E-mail:wangzhenhan93@gmail.com.
(**WANG Zhenhan**, Ph. D. candidate. His research interests include natural language processing and machine translation.)