文献引用格式：张乐乐 , 郭军军 , 王繁 . 基于预训练语言模型及交互注意力的平行句对抽取方法 [J]. 通信技术 ,2021,55(4):443-452.
doi:10.3969/j.issn.1002-0802.2022.04.006

基于预训练语言模型及交互注意力的平行句对抽取方法 *

张乐乐[1,2]，郭军军[1,2]，王　繁[1,2]

（1. 昆明理工大学，云南 昆明 650504；2. 云南省人工智能重点实验室，云南 昆明 650504）

摘　要：从互联网可比语料中筛选高质量的平行句对，是提升低资源机器翻译性能的有效手段之一。针对该问题，融合预训练语义表征提出一种基于双向交互注意力机制的跨语言文本语义匹配方法，首先利用预训练语言模型分别获得源语言和目标语言的双语表征，其次基于双向交互注意力机制实现跨语言特征的空间语义对齐，最后基于多视角特征融合后的语义表征实现跨语言句对的关系判定。实验结果表明，所提方法优于已有的平行句对抽取模型。此外，借助抽取出的平行语料，机器翻译模型的性能得到了明显的改善。

关键词：预训练语言模型；交互注意力机制；平行句对抽取；语义匹配；低资源神经机器翻译
中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1002-0802(2022)-04-0443-10

Parallel Sentence Pair Extraction Method Based on Pre-Training Language Model and Cross Interactive Attention

ZHANG Lele[1,2], GUO Junjun[1,2], WANG Fan[1,2]

(1.Kunming University of Science and Technology, Kunming Yunnan 650504, China;
2.Artificial Intelligence Key Laboratory of Yunnan Province, Kunming Yunnan 650504, China)

Abstract: Selecting high-quality parallel sentence pairs from comparable Internet corpora is one of the effective means to improve the performance of low-resource machine translation. In response to this problem, this paper proposes a cross-language text semantic matching method based on a cross interactive attention mechanism by fusing pre-trained representations. First, it uses pre-trained language models to obtain bilingual representations of source language and target language. Then, it realizes the spatial semantic alignment of cross-language features based on the cross interactive attention mechanism. Finally, it realizes the cross-language based on the semantic representation after multi-view feature fusion judgment of the relationship of sentence pairs. Experimental results indicate that the method proposed in this paper is superior to the existing parallel sentence pair extraction models. In addition, with the help of the extracted parallel corpus, the performance of the machine translation model is significantly improved.

Keywords: pre-training language mode; cross interactive attention mechanism; parallel sentence pair extraction; semantic matching; low-resource neural machine translation

* **收稿日期：**2021-12-07；**修回日期：**2022-03-11　Received date:2021-12-07; Revised date:2022-03-11
基金项目：科技创新 2030-“新一代人工智能”重大项目（2020AAA0107904）多模态机器自动翻译；国家自然科学基金（61866020，61732005）；基于图像、文本特征对齐融合的多模态神经机器翻译问题的研究；云南省科技厅面上项目（2019FB082）
Foundation Item: Technological Innovation 2030-" New Generation Artificial Intelligence" Major Project (2020AAA0107904) Multimodal Machine Automatic Translation; National Natural Science Foundation of China (61866020, 61732005); Research on Multimodal Neural Machine Translation based on the alignment and fusion of image and text features; General Project of Yunnan Provincial Department of Science and Technology (2019FB082)

0 引 言

神经机器翻译模型的性能依赖大量高质量平行数据，主流语言对如英 – 德、英 – 法等已有丰富的平行语料库，因此在这些语言对上，机器翻译性能较高，已接近人类译者的水平[1]。然而，对于大量的非主流语言对，由于其不具备大规模高质量的平行句对资源，因此严重制约了机器翻译模型的性能[1–4]。此外，大量的工作证明了涵盖各种应用领域的网络资源可以作为扩展低资源平行数据的有效来源[5–11]。通过维基百科、影视字幕、双语新闻、同一结构的网页等可以获取大量在内容和形式上都具有可比性的可比语料。从可比语料中获取高质量的平行数据是缓解低资源平行数据稀疏的有效方法之一。近年来，大量针对低资源可比语料库的平行句对抽取方法取得了很好的效果[12–14]，证明了从中提取的平行数据可以有效地提高机器翻译的性能。

平行句对抽取任务基于语义相似性实现两种语言的匹配，核心在于实现双语语义空间的对齐，从而判别语义一致性，目的在于使用抽取到的平行句对作为训练数据，提升机器翻译等自然语言处理任务的性能。网络爬取的数据极其复杂，而且可比的数据并不一定都是直译的，需要根据句子的深层语义一致性实现平行句对的抽取。如表 1 所示，越南语 1 和越南语 2 在词级方面有极高的相似性，但存在巨大的语义偏差，基于统计的传统方法很难区分，因此基于语义空间对齐的方法应运而生。

表 1 双语可比数据样例

语 种	样 例
中 文	当你打开开关时，灯会亮起
越南语 1（平行）	Và khi bạn bật công tắc, đè n s á ng l ê n.
越南语 1（译文）	当你打开开关时，灯亮了
越南语 2（非平行）	Khi bạn rời khỏi công ty, đè n vẫn sang.
越南语 2（译文）	当你离开公司时，灯还亮着

传统的语义空间对齐方法使用不同神经网络结构学习不同语言句子的向量表示，并将其映射到一个共享的向量空间中，以判断跨语言句对的语义相似性；但是在解决存在大量噪声的网络资源数据时，受限于训练数据的数量和覆盖领域，传统的方法难以生成好的语义表征，进而影响语义对齐的效果。相反融合跨语言预训练语义表征可以很好地实现公共语义空间上双语语义的对齐，大量有关预训练的工作表明预先训练的模型有利于下游自然语言处理任务[15,16]。平行句对双语语义表征一般分为词级粒度语义表征和句子级粒度语义表征。针对词级表征，预先训练的词嵌入模型包括 Word2vec[17] 和 GloVe[18], 以及包含上下文信息的语境话模型 CoVe[19] 和 ELMo[20]。针对句子级表征，主要的预训练语言模型包括 OpenAI GPT[21]、ULMFiT[22] 和 BERT[15]。在大规模数据上预先训练的模型已经被证明可以学习通用语言表示，从而避免模型因为数据不足而引起的性能不佳的问题，同时因为不用从头训练模型，进一步解放了计算资源。本文使用预训练语言模型作为先验知识，对获得的语义表征基于双向交互注意力机制进行语义对齐。

为了解决包含噪声数据的语义表征和深层语义对齐问题，本文提出了基于预训练语言模型及双向交互注意力的跨语言文本语义匹配方法。该方法首先利用预训练语言模型的语义捕获能力，为输入的跨语言句子对生成更好的语境化语义表征，其次经过跨语言语义对齐层学习跨语言句子对的依赖关系，再次在跨语言语义融合层从多视角比较句子对的特征表示，最后经过语义预测层实现跨语言句对的关系判定。为验证本文方法的有效性，本文在人工构建的汉 – 越可比语料库上进行了实验，并且为了进一步验证模型对平行句对的捕获能力，在由 IWSLT15 英 – 越公共数据集构造的英 – 越可比语料库上进行了平行句抽取实验，验证了模型提取真实数据的有效性。实验结果表明，本文方法优于已有平行句对抽取方法。最后，本文将提取的平行数据添加到机器翻译语料中用于训练神经机器翻译模型，其性能获得很大的提升，证明所抽取数据可以有效提升下游任务的性能。

1 相关工作

平行句对抽取是缓解低资源机器翻译数据匮乏的主要手段之一，目前双语平行句抽取方法可以集中地分为基于统计的方法和基于深度学习的方法。

基于统计的方法依赖文档内部结构信息或语言学知识，包括文件出版日期、文件标题或文件结构等。Munteanu 等人[23] 提出使用出版日期和信息检索系统对齐报纸文章中的相似文档，通过单词重叠和句子长度比例选择候选句子对，再通过分类器从候选句子对中识别平行句对。Chuang 等人[24] 提出了基于标点符号统计和词汇信息对齐双语平行文本的新方法。Peng 等人[25] 结合了基于长度和基于词汇的算法，将双语文本切分为小块，优化句子的对齐效果。Rauf 等人[26] 提出使用统计机器翻译可比语料库的源语言，被翻译的部分作为查询从目标语

言进行信息检索抽取平行句对，显著提高了统计机器翻译系统的性能。

基于深度学习的方法利用双语句对的语义一致性实现平行句对抽取。Francis 等人 [27] 首次提出应用双向递归神经网络学习文本的通用表示，以及利用端到端的深度神经网络检测两种不同语言句子之间的翻译对等的方法来提取平行句对。Bouamor 等人 [1] 通过混合多语种句子级嵌入、神经机器翻译和监督分类，从可比语料库中提取平行句子。Hangya 等人 [28] 检测候选句子对的连续平行片段，基于源语言单词和目标语言单词的余弦相似性挖掘平行句对。Zhu 等人 [6] 结合连续词嵌入和深度学习方法，引入跨语言语义表示来诱导双语信号，从多语种网站抽取平行句对。Lison 等人 [29] 提出了结合语言和非语言的特征组合自动检测在线电影和电视字幕方法，从在线电影和电视字幕库中提取平行语料库。Bartholomaus[30] 通过分析主题和子主题的链接拓扑，检查维基百科多语言内容的哪一部分对于获取双语数据是可行的，并从中抽取平行句对构建平行语料库和特定领域的词汇表。

不同于以上方法，本文将预训练语言模型融入平行句对抽取过程中，借助预训练语言模型强大的表征能力获得更好的双语语义表征。受 Grégoire 等人 [27] 和 Yang 等人 [31] 的启发，本文提出了基于预训练语言模型及双向交互注意力的跨语言文本语义匹配方法，实现了公共语义空间中，跨语言句对的语义对齐，并利用跨语言句对的语义一致性判定抽取平行句对，并使用提取的平行句作为训练神经机器翻译模型的训练数据，有效改善低资源下神经机器翻译的性能。

2　基于预训练语言模型及双向交互注意力的平行句对抽取方法

针对包含噪声数据的语义表征和深层语义对齐问题，本文提出一种基于预训练语言模型及双向交互注意力的跨语言文本语义匹配模型（Pretrained Encoder Aligned Fusion Prediction Model，EAFP）。该模型主要包括基于预训练语言模型的跨语言文本编码模块、跨语言文本语义匹配模块、跨语言文本语义融合模块和跨语言语义预测模块这 4 个部分，模型结构体系如图 1 所示。

2.1　基于预训练语言模型的跨语言语义编码层

跨语言文本语义编码层分别对源语言和目标语言进行编码，长度为 s 的源语言句子序列表示为 $S_m=\{x_1,x_2,\cdots,x_s\}$，$m \in M$，长度为 t 的目标语言句子序列表示为 $T_n=\{y_1,y_2,\cdots,y_t\}$，$n \in N$，$M$ 和 N 表示句子总数。使用预训练的多语言 BERT 作为双语编码器，源语言和目标语言经过语义编码层分别表示为：

$$\boldsymbol{v}_{S_m}=BERT(S_m) \tag{1}$$

$$\boldsymbol{v}_{T_n}=BERT(T_n) \tag{2}$$

式中：$\boldsymbol{v}_{S_m} \in \boldsymbol{R}^{s\times d}$，$m \in M$ 为经过编码后源语言的向量表示；$\boldsymbol{v}_{T_n} \in \boldsymbol{R}^{t\times d}$，$n \in N$ 为经过编码后目标语言的向量表示；d 为源语言和目标语言句子中单词的词向量维度，生成的向量作为下一层跨语言语义对齐层的输入。

2.2　基于双向交互注意力的跨语言语义对齐层

受不同语言的特性影响，两种语言的语序并不是完全对应的。考虑到自注意力机制不受单词间的所在位置影响，直接计算单词对之间的语义相关性。本文在此基础上采用改进的双向交互注意力机制捕获跨语言文本间的语义交互关系，将源语言与目标语言映射到公共语义空间进行空间语义对齐。与 Vaswani 等人 [32] 的工作一样，本文使用并行的多个注意力头使模型关注不同层面的语义信息。

在得到源语言和目标语言的语义表示 $\boldsymbol{v}_{s_i}$ 和 $\boldsymbol{v}_{t_j}$ 后，为了公式的简洁性，本文之后的公式使用 $\boldsymbol{v}_s$ 作为源语言句语义表征的通用表示，$\boldsymbol{v}_t$ 作为目标语言语义表征的通用表示。源语言到目标语言方向的注意力计算过程：

$$Attention(\boldsymbol{v}_s\boldsymbol{W}_i^1,\boldsymbol{v}_t\boldsymbol{W}_i^2,\boldsymbol{v}_t\boldsymbol{W}_i^3)=softmax\left(\frac{\boldsymbol{v}_s\boldsymbol{W}_i^1\boldsymbol{v}_t^{\mathrm{T}}\boldsymbol{W}_i^2}{\sqrt{d_k}}\right)\boldsymbol{v}_t\boldsymbol{W}_i^3 \tag{3}$$

$$\boldsymbol{v}_s^{'}=MultiHead(\boldsymbol{v}_s,\boldsymbol{v}_t,\boldsymbol{v}_t)=Concat(head_1,\cdots,head_h)\boldsymbol{W}^o \tag{4}$$

$$Where\,head_i=Attention(\boldsymbol{v}_s\boldsymbol{W}_i^1,\boldsymbol{v}_t\boldsymbol{W}_i^2,\boldsymbol{v}_t\boldsymbol{W}_i^3) \tag{5}$$

式中：i 代表第 i 个注意力头；$\boldsymbol{W}_i^1,\boldsymbol{W}_i^2,\boldsymbol{W}_i^3$ 为第 i 个头对应的参数矩阵；h 为注意力头的个数，在本文中设置为 8，每个注意力头的维度为 64；$\boldsymbol{W}^o$ 为最后拼接所有注意力头结果做线性投影的参数矩阵；$\boldsymbol{v}_s'$ 为源语言编码向量经过跨语言语义对齐层的输出。

目标语言到源语言的注意力计算过程：

$$Attention\left(\boldsymbol{v}_t\boldsymbol{W}_i^1,\boldsymbol{v}_s\boldsymbol{W}_i^2,\boldsymbol{v}_s\boldsymbol{W}_i^3\right)=softmax\left(\frac{\boldsymbol{v}_t\boldsymbol{W}_i^1\boldsymbol{v}_s^{\mathrm{T}}\boldsymbol{W}_i^2}{\sqrt{d_k}}\right)\boldsymbol{v}_s\boldsymbol{W}_i^3 \tag{6}$$

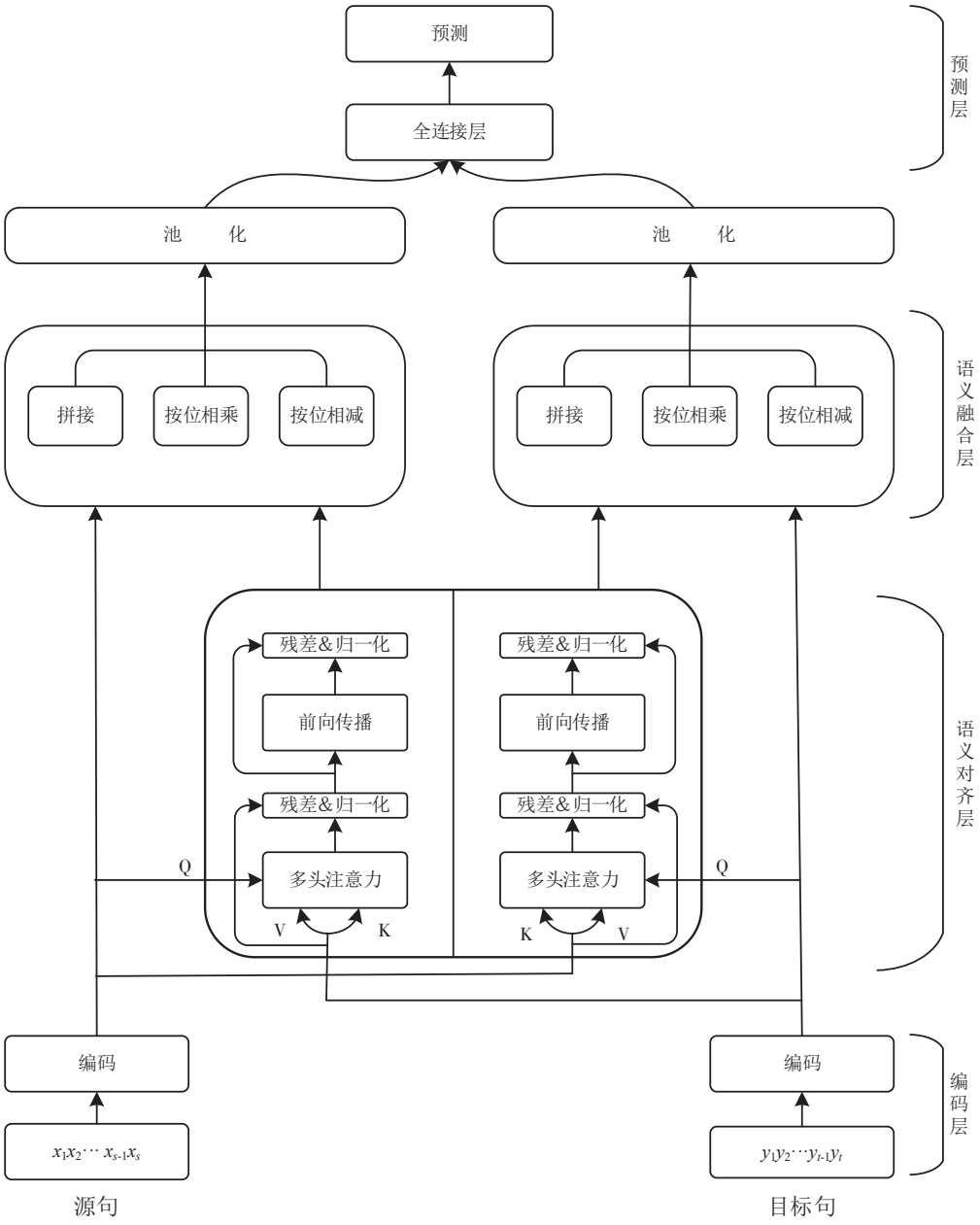

图 1 基于预训练语言模型双向交互注意力的跨语言文本语义匹配模型

$$\boldsymbol{v}_t^{'} = MultiHead\left(\boldsymbol{v}_t,\boldsymbol{v}_s,\boldsymbol{v}_s\right) = Concat\left(head_1,\cdots,head_h\right)\boldsymbol{W}^o \quad (7)$$

$$Where\,head_i = Attention(\boldsymbol{v}_t\boldsymbol{W}_i^1,\boldsymbol{v}_s\boldsymbol{W}_i^2,\boldsymbol{v}_s\boldsymbol{W}_i^3) \quad (8)$$

式中：$\boldsymbol{v}_t'$ 为目标语言编码向量经过跨语言语义对齐层的输出。

2.3 跨语言语义融合层

跨语言语义融合层从多个视角比较语义向量的全局表示以及对齐表示的相似性。本文设计 3 种不同的融合策略，分别为对原始语义信息和对齐后的语义信息进行拼接、按位相减、按位相乘，得到 3 个不同的语义特征向量，然后将所有的向量矩阵投影到同一空间，得到最终的语义融合层输出。源语言经过语义融合层输出 $\boldsymbol{V}_s$ 计算过程：

$$\boldsymbol{V}_{s1}=G_1([\boldsymbol{v}_s;\boldsymbol{v}_s']) \quad (9)$$

$$\boldsymbol{V}_{s2}=G_2([\boldsymbol{v}_s;\boldsymbol{v}_s-\boldsymbol{v}_s']) \quad (10)$$

$$\boldsymbol{V}_{s3}=G_3([\boldsymbol{v}_s;\boldsymbol{v}_s \odot \boldsymbol{v}_s']) \quad (11)$$

$$\boldsymbol{V}_s=G([\boldsymbol{V}_{s1};\boldsymbol{V}_{s2};\boldsymbol{V}_{s3}]) \quad (12)$$

式中：G_1、G_2、G_3 和 G 分别为具有独立参数的单层前馈神经网络；⊙代表对应元素相乘。特征向量之间的差异性由两者的差值衡量，乘法运算用以突出两者的相似性。目标语言经过语义融合输出结果 $\boldsymbol{V}_t$ 的计算与此一致，故省略其公式。

2.4　跨语言语义预测层

语义融合层的输出经过最大池化操作进行特征压缩，获得的源语言和目标语言向量表示 $\boldsymbol{O}_s$，$\boldsymbol{O}_t$ 作为语义预测层的输入。得到跨语言句子对的语义相关性概率分布：

$$\hat{y}=H([\boldsymbol{O}_s;\boldsymbol{O}_s]) \tag{13}$$

式中：H 代表多层前馈神经网络；$\hat{y}=\mathbb{R}^C$ 代表所有类别的概率分数；C 为类别的数量。之后，根据概率分布 $\hat{y}=\text{argmax}\ x_i\hat{y}_i$, 区分输入的源语言和目标语言句子对是否为平行句子。训练目标是最小化训练数据集的交叉熵：

$$L=-\sum_{i=1}^{n}y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i) \tag{14}$$

式中：y_i 为真实标签；$\hat{y}_i$ 为预测结果。最后，在得到跨语言句子对的预测结果后，将最小化预测结果和真实结果之间的交叉熵作为损失函数来训练模型。

3　实验与分析

本文使用自行构建的汉语 – 越南语可比语料库和 IWSLT15 英文 – 越南语可比语料库训练模型。本节内容安排：3.1 节介绍数据集；3.2 节介绍负采样细节；3.3 节介绍评价指标；3.4 节介绍实验参数设置；3.5 节介绍实验结果；3.6 节介绍消融实验；3.7 节介绍案例分析。

3.1　数据集

网络爬取的汉 – 越可比语料库，目前在汉语到越南语低资源语言对上，缺乏公开使用的汉 – 越数据集。基于此，本文收集并构建了一个汉 – 越平行语料库。数据来源包括维基百科、双语新闻网站、电影字幕等，在经过数据清洗和对齐后，用作模型训练的正样本。在训练模型时，为了保持样本数量的平衡，为每个正样本构造一个对应的负样本，由正负样本构成的可比语料库用来训练模型。同时为了验证本文提出的平行句对抽取模型的性能，本文在构建汉 – 越语料库中手动选择日常表达和新闻文本数据作为测试集和验证集。表 2 为实验的语料规模。

IWSLT15 英语 – 越南语可比语料库 : 本文使用标准的英语 – 越南语机器翻译数据集来验证本文方法在公共数据集上提取真实平行句的性能。在原始数据集的基础上，为训练集、验证集、测试集按照 1 ： 1 的比例构造负样本。扩充后的数据集作为完整的可比语料库。可比语料库上的实验结果表明，该方法可以有效地识别数据集中的平行句对。具体的数据规模如表 2 所示。

表 2　实验数据集规模

数据集	汉 – 越平行句对	汉 – 越非平行句对	英 – 越平行句对	英 – 越非平行句对
训练集	14 万	14 万	13 万	13 万
验证集	0.5 万	0.5 万	0.1 万	0.1 万
测试集	0.5 万	0.5 万	0.1 万	0.1 万

3.2　负采样

在训练过程中，本文使用包含 n 个平行句对的平行语料库。这些平行句作为本文训练集中的正样本，对于每对平行句，本文随机抽样生成负样本，因此本文的训练数据由 $2n$ 个三元组组成 (S_i^s,S_i^T,y_i)。S_i^s 代表源语言句子，S_i^T 代表目标语言句子。y_i 是表示 S_i^s 和 S_i^T 之间翻译关系的标签。当源语言句与目标语言句为平行句时，y_i 为 1，反之为 0。

3.3　评价指标

平行句对抽取任务可以看作自然语言处理领域中的一项基础问题，即文本二元分类问题。使用精度（Precision，P）、召回率（Recall，R）和 $F1$ 值作为分类模型的评价指标。精度代表平行句子在所有提取的句子中所占的比例。召回率表示在所有平行句子中所占的被分类正确的平行句子的比例。$F1$ 值代表精度和召回率的调和平均值。具体公式为：

$$P=\frac{|TP|}{|TP+FP|} \tag{15}$$

$$R=\frac{|TP|}{|TP+FN|} \tag{16}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{17}$$

式中：TP 为提取出来真正平行句的数量；FP 为被错认为平行句的数量；FN 为被错认为非平行句的数量。

3.4　实验设置

本文中的模型使用 Pytorch 框架编写实现，BERT 使用 12 层每层 768 维隐藏单元的隐藏层。注意头的数量是 12。模型训练了 10 个时期。在训练过程中使用 Adam[33] 作为优化器，批次大小设置为 128，学习率是 0.000 5，设置 dropout 为 0.2 来防止过拟合。

3.5 实验结果

为了测试基于预训练语言模型及双向交互注意力的平行句抽取模型的性能，本文进行了一系列实验，本节对实验结果进行展示和分析，内容安排为：第 3.5.1 节分析不同编码方式对模型性能的影响，第 3.5.2 节分析真实场景下的平行句对抽取结果，第 3.5.3 节分析不同数据规模下的抽取效果，第 3.5.4 节分析将所抽取数据使用到机器翻译任务中对译文质量的影响。

3.5.1 不同编码方式的实验结果及分析

为了研究不同编码方式对模型性能的影响，验证文中所用预训练语言模型的有效性，在本节中使用不同的网络作为编码层，比较不同编码方式下模型平行句对抽取的效果。首先，使用传统的循环神经网络（Recurrent Neural Network, RNN）、长短期记忆网络（Long-Short Term Memory，LSTM）、门控循环单元（Gate Recurrent Unit,，GRU）、卷积神经网络（Convolutional Neural Networks，CNN），以及双向网络（BiRNN、BiLSTM、BiGRU）等替换模型中的预训练语言模型作为模型编码层。其次，在编码中引入注意力机制，包括将经双向网络 BIGRU 生成的源语言和目标语言表征通过注意力机制进行特征交互的 BIGRU+ATT，以及基于自注意力网络的 Transformer。最后，使用其他多语言预训练模型进行平行句对抽取，包括针对跨语言相似表征的 Sentence-BERT[34] 和解决多语言句子嵌入表征的 LASER。在保持其他设置不变的情况下，使用包含 28 万正负样本的汉 – 越数据分别训练不同的对比模型，并通过精度（P），召回率（R）和 $F1$ 值比较不同模型的性能。具体实验结果如表 3 所示，其中，对比模型均使用获得最佳性能的参数。

表 3 不同模型在汉 – 越小规模可比语料库数据的实验结果 %

模 型	P	R	$F1$
RNN	73.25	95.78	83.01
LSTM	72.68	95.34	82.48
GRU	73.89	93.70	82.62
CNN	71.12	96.20	81.78
BIRNN	77.56	96.84	86.14
BILSTM	74.56	94.30	83.28
BIGRN	77.90	96.36	86.15
BIGRU+ATT	76.91	94.66	84.78
Transformer	74.49	88.98	81.09
Sentence-BERT	87.70	94.12	90.79
LASER	79.76	95.46	86.90
EAFP	94.36	94.46	94.41

表 3 中的实验结果表明，双向网络结构在性能上优于单向网络结构的模型，原因在于双向网络可以更好地表征输入文本的上下文信息。基于 Transformer 的模型 $F1$ 值比基于 RNN 模型下降了 1.92%，笔者认为是因为其复杂的网络结构导致模型参数过多，受限于训练数据的规模，模型并没有得到充分训练，尚未达到拟合。基于预训练模型的性能大幅优于其他模型的性能，笔者将这归因于数据规模对神经网络的性能起着至关重要的作用。预训练语言模型使用了海量的数据为单词表示学习更好的语境化向量表示，从而为输入文本生成更好的语义表征。与预训练模型 Sentencebert 和 Laser 相比，本文的模型有较为明显的优势，主要原因在于本文方法在得到预训练语义表征后，进一步进行了语义的对齐和融合，因此取得了更好的结果。

3.5.2 真实场景下的抽取结果

为验证模型与下游任务的适配能力，本文在包含 26 万句对的英 – 越公共数据集上模拟真实的应用场景，进一步鉴别模型识别平行句对的能力。具体实验结果如表 4 所示。

表 4 不同模型在英 – 越可比语料库实验结果 %

模 型	P	R	$F1$
RNN	96.59	98.19	97.38
LSTM	95.61	97.87	96.73
GRU	96.35	97.95	97.15
CNN	89.76	98.19	93.79
BIRNN	97.12	98.34	97.73
BILSTM	97.35	98.42	97.88
BIGRN	97.02	97.63	97.33
BIGRU+ATT	79.47	96.14	87.01
Transformer	98.65	97.63	98.14
Sentence-BERT	94.67	96.69	95.67
LASER	85.12	93.85	89.27
EAFP	98.44	99.68	99.06

从表 4 的实验结果可以看出，在英 – 越可比语料中，本文模型依然优于其他方法，取得了最好的效果。这说明本文模型拥有较强的语义信息捕获能力，可有效捕获平行句对。值得注意的是，在数据规模相似的情况下，表 4 中的结果明显优于表 3。笔者推测，相较于中 – 越句对，英语与越南语同属一个语系，拥有更相近的语言形式，使得语义空间的对齐拥有更强的可操作性。结合表 3 可知，本文模型在不同语言环境中表现出更强的鲁棒性，拥有更强的泛化能力。

3.5.3　不同数据规模对比实验

为了进一步研究数据规模对模型性能的影响，本文在汉 – 越语言对的数据集中添加了 20 万的训练数据重新训练各模型，表 5 的实验结果显示了各模型在较大规模数据集下的性能。

表 5　不同模型在汉 – 越较大规模可比语料库的实验结果 %

模　型	*P*	*R*	*F*1
RNN	82.71	94.98	88.42
LSTM	78.07	93.32	86.01
GRU	85.23	94.56	89.65
CNN	75.51	94.30	83.50
BIRNN	83.32	96.16	89.28
BILSTM	84.65	94.58	89.34
BIGRN	87.63	95.56	91.42
BIGRU+ATT	76.82	93.70	84.42
Transformer	85.39	94.62	89.77
Sentence-BERT	90.40	95.56	92.91
LASER	91.30	95.54	93.37
EAFP	95.52	97.63	96.57

实验结果表明，各种模型的性能都有了不同程度的提高，其中，基于 Transformer 等复杂结构的模型提升效果更加明显，超越了基于循环神经网络等结构简单的模型，这证明了本文在 3.5.1 节中关于复杂模型对数据规模依赖性的猜想，即结构复杂的模型因参数量更大而对训练数据的规模更加敏感，在小规模数据的情况下难以达到理想的效果。相反，其在大规模数据下则能实现更好的拟合，学习更多的上下文信息，从而生成更好的文本表征。与其他模型的提升幅度相比，这种模型显示出更强的稳健性，究其原因是其在小规模数据中表现出的强竞争力，证明基于预训练模型方法对数据规模的鲁棒性，可以在更少的数据量上达到理想的效果。

3.5.4　机器翻译性能评估

本文的目标是通过过滤网络数据来扩展平行语料库的规模，扩宽覆盖领域，从而提高低资源机器翻译模型的性能。为了验证通过本文的方法提取的平行句对机器翻译模型性能的影响，本文在两个低资源语言对上构建了神经机器翻译模型，分别为汉语 – 越南语和英语 – 越南语。本文使用 Facebook 开源的 pytorch 版本的 fairseq 框架训练神经机器翻译（Neural Machine Translation，NMT）模型，并通过 NMT 模型的双语互译质量辅助工具（Bilingual evaluation understudy，BLEU）评分来评估其质量。翻译模型由 6 层编码器 – 解码器的序列到序列结构组成。

表 6　数据规模对机器翻译性能的影响

数据规模	BLEU 得分	
	汉语 – 越南语	英语 – 越南语
13 万	20.21	30.86
18 万 (+5 万)	20.85(+0.64)	31.30(+0.44)
23 万 (+5 万)	21.21(+1.00)	31.52(+0.66)
28 万 (+5 万)	21.48(+1.27)	31.91(+1.05)
33 万 (+5 万)	22.04(+1.83)	32.15(+1.29)
38 万 (+5 万)	22.32(+2.11)	32.34(+1.48)

表 6 中的结果显示了不同规模的训练数据下的 BLEU 得分。实验结果表明，加入提取到的 25 万平行句后，汉 – 越机器翻译系统的 BLEU 分数从 20.21 增加到 22.32。英 – 越机器翻译系统的 BLEU 分数从 30.86 增加到 32.34。这进一步证明了本文的方法可以有效地提取平行句对，可以用于扩展多语言平行语料库。本文方法为缓解资源紧缺的神经机器翻译系统缺乏训练数据的问题提供了有效的解决方案，提取出的语义一致的平行句对有利于神经机器翻译系统性能的提升。

3.6　消融实验

为了分析模型不同模块对抽取结果的影响，更好地理解不同部分在模型中的具体效用，本文进行了一系列消融实验。实验中，本文针对不同的组件对主模型进行删除简化，或改变某一模块的策略得到不同的变种模型。实验结果展示在表 7 中。表 7 中的几类模型为改变部分主模型结构得到的消融模型，具体细节介绍如下文所述。

（1）EP 模型：删除语义对齐层和语义融合层，具体研究核心组件语义对齐层和语义融合层对模型性能的影响。输入的跨语言文本经过语义编码层之后，将生成的语句表征直接送入语义预测层，仅利用多语言预训练语言模型自身对不同语言的表征差异性完成对跨语言句对的分类。

（2）EAP 模型：删除语义融合层，具体研究从多视角融合编码层和对齐层语义信息，全局比较特征向量的相似性和差异性的语义融合层对模型性能的影响。输入的跨语言文本首先经过语义编码层得到语义表征，并进一步经语义对齐层进行交叉表征，学习相互之间的依赖关系，最后进入语义预测层进行分类。

（3）EAFP–A 模型：与本文方法不同的是，将语义对齐层的学习策略替换为遵循 Parikh 等人 [35] 方法的一种简单的注意力机制，在更少的模型参数下学习跨语言文本的相互依赖关系。

（4）EAFP–B 模型：与本文方法不同的是，语义对齐层的学习策略替换为自注意力（self–attention）机制，加强跨语言文本本身的语义表征，不考虑彼此之间的文本交互关系。

表 7 消融模型实验结果 %

模 型	*P*	*R*	*F*1
EP	83.24	92.42	87.59
EAP	87.10	90.13	88.59
EAFP–A	93.88	95.70	94.78
EAFP–B	92.33	96.52	94.38
EAFP	95.52	97.63	96.57

从表 7 中的结果可以看出，文中所用方法在精确率、召回率和 *F*1 值方面均优于其他变种模型，证明了基于预训练语言模型及双向交互注意力的平行句对抽取方法的有效性。其中，删除语义对齐层和语义融合层之后的 EP 相较于本文模型 EAFP，模型精确率下降了 12.28%，*F*1 值下降了 8.98%。精确率的大幅下降表明，在原始语义特征相差较大时，被区分为负类的样本数大幅增加，说明仅使用原始预训练语言模型输出的语义表征对深层语义特征的区分效果不佳，并不能有效区分不同语言之间的语义差异性，需要对其进行进一步的对齐融合。与此相比，当添加了语义对齐层和语义融合层之后，模型性能有了显著的提高。EAP 模型相比于 EP，*F*1 值有了明显的提升，原因在于语义融合层从多视角比较语义的局部表示和对齐表示，可以更好地区分语义特征的相似性和差异性，保证了结果的准确性。最后，改变语义对齐层学习策略后的 EAFP–A 和 EAFP–B 相较于本文模型，*F*1 值分别下降了 1.79% 和 2.19%，表明基于双向交互的注意力机制可以更好地学习跨语言句子之间的交互信息，且较大模型有助于语言特征的准确表达。

总的来说，模型各模块在保证模型性能方面发挥着不同的作用，对语义的表达和最终有效抽取句对都必不可少，也再次说明本文所提出的基于预训练语言模型及双向交互注意力的平行句对抽取方法的强大效果。

3.7 案例分析

本文提出一个具体的实例分析，来展示序列间的结果在本模型中的演变过程。依据注意力权重体现模型所学习的语义交互关系，从汉 – 越验证集中选择一对例句。源语言句是“当你打开开关时，灯会亮起。”，目标句是“V à khi bạn bật công tắc , đ è n s á ng l ê n .”。图 2 显示了跨语言对齐层中注意力分布的可视化结果（方程 3）。图中颜色的深浅代表了词与词的语义的相关程度。语义相关性越强的两个单词对应的颜色越深。

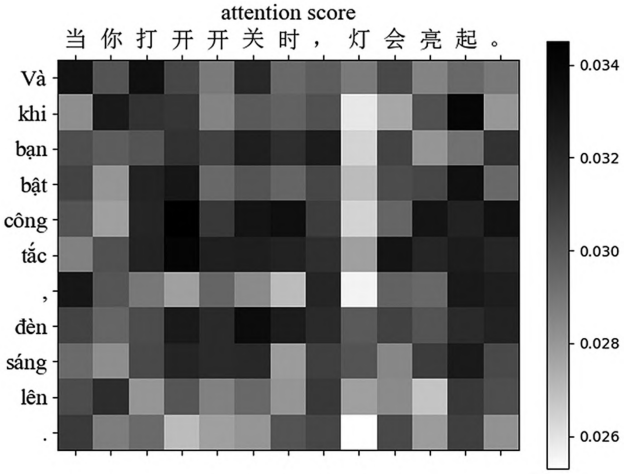

图 2 注意力权重可视化结果

在图 2 中，可以看到“打开”与“công（手动）”“tắc（转变）”“đ è n（灯）”语义关系较强，“亮起”与“khi（当……的时候）”“bật（打开）”“công（手动）”语义关系较强。基于双向交互注意力机制可以学习到双语句对的语义交互关系，通过对源语言和目标语言单词更有效地进行空间语义对齐，可以帮助判别跨语言句子的语义关系。

4 结 语

在网络资源存在大量噪声数据的前提下，本文提出了一种基于预训练语言模型及双向交互注意力的跨语言文本语义匹配方法。使用预训练语言模型，在资源有限的情况下，为跨语言句对生成语境化的语义表示。利用双向交互注意力在公共语义空间中对跨语言句对进行语义对齐，最后得到跨语言句对的关系判定，实现了从可比语料库中提取深层语义一致的双语平行句子扩充双语平行语料库，进而缓解了资源匮乏的语言对缺乏训练数据的问题。实验结果表明，该方法优于其他模型，本文提取的平行句对进一步提高了低资源神经机器翻译的性能。在未来的工作中，笔者希望将此方法扩展到其他非主流语言的研究中。

参考文献：

[1] BOUAMOR H,SAJJAD H.H2@BUCC18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings [C]//the Eleventh International Conference on Language Resources and

Evaluation,2018: 7–12.

[2] LEPAGE Y.Quasi–parallel corpora:Hallucinating translations for the Chinese–Japanese language pair[C]//11th Workshop on Building and Using Comparable Corpora,2018:33.

[3] MARIE B,FUJITA A.Efficient extraction of pseudo–parallel sentences from raw monolingual data using word embeddings[C]//the 55th Annual Meeting of the Association for Computational Linguistics,2017: 392–398.

[4] ZHANG B L,NAGESH A,KNIGHT K.Parallel corpus filtering via pre–trained language models[C]//the 58th Annual Meeting of the Association for Computational Linguistics,2020:8545–8554.

[5] HEWAVITHARANA S,VOGEL S.Extracting parallel phrases from comparable data[C]//Workshop on Building & Using Comparable Corpora: Comparable Corpora & the Web. Association for Computational Linguistics,2011:549–573.

[6] ZHU S L,LI X,YANG Y T,et al.A novel deep learning method for obtaining bilingual corpus from multilingual website[J].Mathematical Problems in Engineering, 2019,2019:1–7.

[7] WOŁK K,MARASEK K.Building subject–aligned comparable corpora and mining it for truly parallel sentence pairs[J].Procedia Technology,2014,18:126–132.

[8] SRIDHAR V K R,BARBOSA L,BANGALORE S.A scalable approach to building a parallel corpus from the web[C]//Interspeech 2011,2011:1–4.

[9] RAMESH S H,SANKARANARAYANAN K P.Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora[C]// the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Student Research Workshop,2018:112–119.

[10] LING W,MARUJO L,DYER C,et al.Crowdsourcing high–quality parallel data extraction from twitter[C]// the Ninth Workshop on Statistical Machine Translation, 2014:426–436.

[11] CHU C H,NAKAZAWA T,KUROHASHI S.Constructing a Chinese–Japanese parallel corpus from wikipedia[C]// LREC,2014:642–647.

[12] OTERO P G,LOPEZ I G.Wikipedia as multilingual source of comparable corpora[C]//the 3rd Workshop on Building and Using Comparable Corpora,2010:21–25.

[13] PATRY A,LANGLAIS P.Identifying parallel documents from a large bilingual collection of texts:Application to parallel article extraction in Wikipedia[C]//the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web,2011:87–95.

[14] BARRÓN–CEDEÑO A,ESPAÑA–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,2015:3–13.

[15] DEVLIN J,CHANG M W,LEE K,et al.BERT: Pre–training of deep bidirectional transformers for language understanding[C]//the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies:4171–4186.

[16] PETERS M E,AMMAR W,BHAGAVATULA C,et al.Semi–supervised sequence tagging with bidirectional language models[C]//the 55th Annual Meeting of the Association for Computational Linguistics,2017:1756–1765.

[17] MIKOLOV T,SUTSKEVER I,0010 K C,et al.Distributed representations of words and phrases and their compositionality[C]//the 26th International Conference on Neural Information Processing Systems,2013:3111–3119.

[18] PENNINGTON J,SOCHER R,MANNING C.Glove:Global vectors for word representation[C]//the 2014 Conference on Empirical Methods in Natural Language Processing,2014:1532–1543.

[19] BRYAN M,JAMES B,CAIMING X,et al.Learned in translation: Contextualized word vectors[C]//the 31st International Conference on Neural Information Processing Systems,2017:6297–6308.

[20] PETERS M,NEUMANN M,IYYER M,et al.Deep contextualized word representations[C]//the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2018:2227–2237.

[21] ALEC R,KARTHIK N,TIM S,et al.Improving language understanding by generative pre–training[EB/OL].[2021–12–02].https://s3–us–west–2.amazonaws.com/openaiassets/research–covers/languageunsupervised/language understanding paper.pdf.

[22] HOWARD J,RUDER S.Universal language model fine–tuning for text classification[C]//the 56th Annual Meeting of the Association for Computational

Linguistics,2018:328–339.
[23] MUNTEANU D S,MARCU D.Improving machine translation performance by exploiting non–parallel corpora[J].Computational Linguistics,2005,31(4):477–504.
[24] CHUANG T C,WU J C,LIN T,et al.Bilingual sentence alignment based on punctuation statistics and lexicon[C]//International Conference on Natural Language Processing,2004:224–232.
[25] DINU G,LAPATA M.Topic models for meaning similarity in context[C]//COLING 2010,23rd International Conference on Computational Linguistics,2010:250–258.
[26] ABDUL RAUF S,SCHWENK H.Parallel sentence generation from comparable corpora for improved SMT[J].Machine Translation,2011,25(4):341–375.
[27] GRÉGOIRE F,LANGLAIS P.Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation[C]//the 27th International Conference on Computational Linguistics,2018:1442–1452.
[28] HANGYA V,FRASER A.Unsupervised parallel sentence extraction with parallel segment detection helps machine translation[C]//the 57th Annual Meeting of the Association for Computational Linguistics,2019:1224–1234.
[29] LISON P,DOGRUOZ A S.Detecting machine–translated subtitles in large parallel corpora[C]//11th Workshop on Building and Using Comparable Corpora,2018:25–32.
[30] BARTHOLOMAUS W.Identifying bilingual topics in wikipedia for efficient parallel corpus extraction and building domain–specific glossaries for the japanese–english language pair[C]//11th Workshop on Building and Using Comparable Corpora,2018:15.
[31] YANG R,J ZHANG,GAO X,et al.Simple and effective text matching with richer alignment features[C]//the 57th Annual Meeting of the Association for Computational Linguistics,2019:4699–4709.
[32] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C]//the 31st Conference on Neural Information Processing Systems,2017:1–11.
[33] KINGMA D,BA J.Adam: A method for stochastic optimization[C]//International Conference on Learning Representations,2015:1–15.
[34] REIMERS N,GUREVYCH I.Sentence–BERT: Sentence embeddings using siamese BERT–networks[C]// the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,2019:3982–3992.
[35] PARIKH A,TCKSTRM O,DAS D,et al.A decomposable attention model for natural language inference[C]// the 2016 Conference on Empirical Methods in Natural Language Processing,2016:2249–2255.

作者简介：

张乐乐（1994—），男，硕士，主要研究方向为自然语言处理；

郭军军（1987—），男，博士，副教授，主要研究方向为自然语言处理、信息检索、机器翻译；

王　繁（1996—），男，硕士，主要研究方向为自然语言处理。