基于跨语言双语预训练及Bi-LSTM的汉-越平行句对抽取方法

刘畅[1,2]，高盛祥*[1,2]，余正涛[1,2]，黄于欣[1,2]，尤丛丛[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
liuxiaochang32@163.com, gaoshengxiang.yn@foxmail.com
ztyu@hotmail.com, huangyuxin2004@163.com, 1257767625@qq.com

摘要

汉越平行句对抽取是缓解汉越平行语料库数据稀缺的重要方法。平行句对抽取可转换为同一语义空间下的句子相似性分类任务，其核心在于双语语义空间对齐。传统语义空间对齐方法依赖于大规模的双语平行语料，越南语作为低资源语言获取大规模平行语料相对困难。针对这个问题本文提出一种利用种子词典进行跨语言双语预训练及Bi-LSTM（Bi-directional Long Short-Term Memory）的汉-越平行句对抽取方法。预训练中仅需要大量的汉越单语和一个汉越种子词典，通过利用汉越种子词典将汉越双语映射到公共语义空间进行词对齐。再利用Bi-LSTM和CNN（Convolutional Neural Networks）分别提取句子的全局特征和局部特征从而最大化表示汉-越句对之间的语义相关性。实验结果表明，本文模型在F1得分上提升7.1%，优于基线模型。

关键词： 汉-越；平行句对抽取；跨语言预训练；公共语义空间；Bi-LSTM

Chinese-Vietnamese Parallel Sentence Pair Extraction Method Based on Cross-lingual Bilingual Pre-training and Bi-LSTM

Chang Liu[1,2], Shengxiang Gao*[1,2], Zhengtao Yu[1,2],
Yuxin Huang[1,2], Congcong You[1,2]
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology Kunming 650500, China
liuxiaochang32@163.com, gaoshengxiang.yn@foxmail.com
ztyu@hotmail.com, huangyuxin2004@163.com, 1257767625@qq.com

Abstract

The extraction of Chinese-Vietnamese parallel sentence pairs is an important method to alleviate the scarcity of Chinese-Vietnamese parallel corpus data. Parallel sentence pair extraction can be converted into sentence similarity classification task in the same semantic space, the core of which is to achieve bilingual semantic space alignment. The traditional semantic space alignment method relies on large-scale bilingual parallel corpus, and it is relatively difficult for Vietnamese to obtain large-scale parallel corpus as a low-resource language. To address this problem, this paper proposes a bilingual dictionary for cross-lingual bilingual pre-training and Bi-LSTM (Bi-directional Long Short-Term Memory) Chinese-Vietnamese parallel sentence pair extraction method. Only a large number of Chinese-Vietnamese monolingual and a Chinese-Vietnamese

*通信作者：高盛祥，email地址：gaoshengxiang.yn@foxmail.com
基金项目：国家自然科学基金（61761026，61972186，61732005，61672271，61762056）；国家重点研发计划（Nos.2019QY1802，2019QY1801，2019QY1800）；云南省自然科学基金2018FB104；云南高科技人才项目（201606）；昆明理工大学省级人培项目（KKSY201703005）

seed dictionary are required for pre-training. By using the Chinese-Vietnamese seed dictionary to map the Chinese-Vietnamese bilingual to the common semantic space for word alignment.Then, Bi-LSTM and CNN (Convolutional Neural Networks) are used to extract the global and local features of sentences to maximize the semantic relevance between Chinese-Vietnamese sentence pairs. Experimental results show that the model in this paper improves F1 score by 7.1%, which is better than the baseline model.

Keywords: Chinese-English-Burmese , Low Resource Language , Multilingual Neural Machine Translation , Joint training , Semantic space mapping , Shared parameters

1 引言

平行语料库的规模和质量对于机器翻译的性能至关重要。在大规模语料中的机器翻译如汉-英神经机器翻译中都已经获得很好的结果。由于汉-越低资源语言很难获得足够多的平行句对，从而导致汉-越机器翻译性能较差。人工手动构建大规模高质量的平行语料库耗时耗力，通过大量的文本研究发现在同一段时间内报道的新闻或具有同一结构的网页都能获得大量的可比语料，从可比语料中抽取平行句对是扩充翻译语料的重要方法之一。本文的目的是从汉越可比语料中抽取汉越平行句对。

目前双语平行句对抽取的方法大致可以分为以下四类：首先是利用统计机器翻译和神经机器翻译方法从可比语料库中抽取平行句对是比较有效的。在统计机器翻译方面，Rauf等人(Rauf and Schwenk, 2011)的方法是将目标语言翻译成源语言，利用跨语言信息检索技术从可比语料库中抽取平行句对，提高了统计机器翻译的性能；在神经机器翻译方面，(Marie and Fujita, 2017; Choudhary et al., 2018)提出了基于词嵌入在大型单语语料库中抽取平行句对从而提升了神经机器翻译的性能。Utiyama等人(Masao Utiyama, 2013)经过两次机器翻译，首先将日语句子翻译得到n-best英语译文，再把英语译文翻译成汉语，构建中日平行语料库。这些方法都是通过有效抽取平行句对来提升机器翻译的性能，但需要在翻译模型性能比较好的基础上才能进行。

其次在基于特征工程方面，(Chuang et al., 2004; España-Bonet et al., 2017; Luong et al., 2015)提出了在双语词典信息的基础上结合了标点符号统计信息和词汇信息的双语平行文本对齐的方法；Gale等人(Gale and Church, 1991)介绍了一种基于字符长度的统计模型对齐平行文本中的句子的方法，识别一种语言的句子和另一种语言的句子之间的长度对应关系。Peng等人(Peng et al., 2010)提出了一种Fast-Champollion句子对齐算法，它结合了基于长度和基于词典信息，通过将输入的双语文本分割成小块进行对齐的过程，提升句子对齐的效果。Ann 等人(Masao Utiyama, 2013)基于现有的翻译系统，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，从而获得平行句子。Chu等人(Chu et al., 2016)从对齐的文章中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句子对，保留尽可能匹配的句子对，然后使用少量平行句对训练分类器，以从候选者中识别平行句对。Tillmann等人(Tillmann and Xu, 2009) 提出了一种用于可比数据的新颖句子对提取算法，直接在句子级别对大量候选句子对进行评分。通过一个简单对称评分函数实现句子级别的提取。但这些方法通常依赖于大量的与语言相关的特征知识，虽然证明了抽取平行句对的有效性，但是由于句对分类准确性不高，无法取得较好的效果。

然后在基于深度学习方面，Francis Gregoire 等人(Grégoire and Langlais, 2017)提出基于双向递归神经网络对源语言和目标语言分别进行编码，然后经过分类器区分源句子和候选目标句子是否平行；Munteanu等人(Munteanu and Marcu, 2005)提出一种利用最大熵分类器从大量可比语料中抽取平行句对的方法，从零开始构建了汉英翻译系统。Grover 等人(Grover and Mitra, 2017)训练模型以获取双语单词嵌入，然后在两个句子的单词之间创建相似度矩阵，并使用卷积神经网络（CNN）将句子分类。Bouamor等人(Bouamor and Sajjad, 2018)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方法，来分类法语-英语语料库中的平行句子对。首先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表示对目标翻译候选进行过滤。然后，使用神经机器翻译系统或二进制分类模型选择最佳翻译。它们能有效利用深度学习的方法从可比语料中抽取平行句对但它们在训练过程中需要大量双语平行句对。

最后在句子相似度计算的方面，(Cheon and Youngjoong, 2017; Azpeitia et al., 2018)提出了一种利用语言资源的顺序匹配在句子之间执行相似度计算从而查找相似句子的方法，从维基百科构建英语和韩语之间的平行语料库。Alberto 等人(Barrón-Cedeño et al., 2015)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于域的并行文档。这种方法是从句子级扩充训练数据，从而构建高质量的平行语料库但也都是针对资源丰富型语言（例如英语-法语），但在低资源语言（如汉语-越南语）上的性能较差，并且抽取出的句子噪声较大。

以上方法在预训练过程中均是利用大量的双语平行句对作支撑，但汉语和越南语都是独立派系的语言且汉越双语训练数据稀缺。通过大量的文本研究发现在同一段时间内报道的新闻或具有同一结构的网页都能获得大量的可比语料，因此如何从汉越可比语料库中获得平行句对具有重要意义。考虑到汉越双语平行句对很难获取而得到汉越单语句子相对容易，结合汉-越句子特性，受Francis Gregoire 等人(Grégoire and Langlais, 2018) 和Artetxe等人(Artetxe et al., 2016; Artetxe et al., 2017; Artetxe et al., 2018)思想启发，提出了一个基于跨语言双语预训练及Bi-LSTM的汉-越平行句对抽取方法，从汉越可比语料中抽取汉越平行句对，来提升低资源语言机器翻译的性能。其主要思想是在汉越双语预训练中将汉越双语句子映射到公共语义空间下，通过汉-越种子词典来缩小汉越双语在语义空间中的距离，从而加强汉越双语的语义相关性。在本文方法中针对的是汉语到越南语两种语言，由于汉语到越南语没有公开的数据集，因此考虑从维基百科文章中抽取的汉-越段落语料以及收集的汉-越段落语料添加到一个语料库中，以训练模型的性能。

2 基于汉-越双语预训练及Bi-LSTM的平行句对抽取模型

针对上文问题，提出一个基于汉-越双语预训练及Bi-LSTM的平行句对抽取方法，具体模型结构体系如图1所示。该模型主要分为三个部分。第一部分是基于汉-越双语预训练，第二部分是由Bi-LSTM和CNN组成的汉-越句子特征提取部分的编码器，第三部分是全连接层进行汉-越平行和句非平行句分类。

首先，将汉语-越南语跨语言双语词嵌入映射到公共的语义空间进行预训练，使得汉语-越南语的语义相似词在该空间中接近，增强汉语和越南语语义空间中的相关性。设x = $(x_1, x_2, ..., x_m)$ 表示表示输入的汉语单词, y = $(y_1, y_2, ..., y_n)$ 表示输入的越南语单词。在双语预训练中，汉-越种子词典在没有大规模平行语料情况下可以实现在汉越统一空间语义对齐，并以自学习的方式迭代地生成新词典。再利用汉-越种子词典来学习词嵌入并指导后面Bi-LSTM和CNN在公共语义空间进行统一编码。将训练好的词向量输入Bi-LSTM来获取单词前后信息特征，并用CNN来提取双语句子更深层语义特征。最后对汉语句子和越南语句子进行编码，通过使用元素乘积和元素绝对差将它们提供给全连接层，使用输出概率作为汉越句对是否为平行语句对的度量来捕获其匹配信息。

3 汉越跨语言词向量预训练

3.1 词向量预训练方法

在双语中，利用单独语料进行独立训练的方法如Mikolov等人(Mikolov et al., 2013)的word2vec(CBOW/Skip-grim)训练出有语义相似性的词嵌入向量。在各自语料上进行独立训练，导致两种语言词嵌入矩阵在分布上也是独立不相关。在汉语和越南语词向量表征中也是如此。双语词嵌入将两种不同语言的词映射到公共的语义空间，公共语义空间中每个单词嵌入之间的距离则暗示着一定的语义关系。这可以保证在单语语义不变性情况下确保具有两个相同语义的词在公共语义空间中的距离非常近，但双语词嵌入的学习都依赖于大规模平行语料库，这对于资源稀缺型语言对（汉语-越南语）是难以获得的。

我们在汉越跨语言词向量预训练中提出了一种自学习的方法。该方法利用了嵌入空间的结构相似性，结合基于汉语-越南语种子词典的映射技术，降低了汉语-越南语双语资源的需求。该自学习的方法框架先是对汉语和越南语在各自的单语语料库上进行独立训练，再通过线性变换来最小化汉越双语词典中的距离从而将汉语-越南语跨语言映射在同一语义空间。通常需要大规模双语词典进行训练，针对汉语-越南语难以获取大规模词典，跨语言预训练中将大型双语词典的需求减少到较小的种子词典，通过不断迭代使更新的种子词典来学习新的映射矩阵，直至收敛。汉-越跨语言双语词嵌入预训练具体细节如下图2所示:

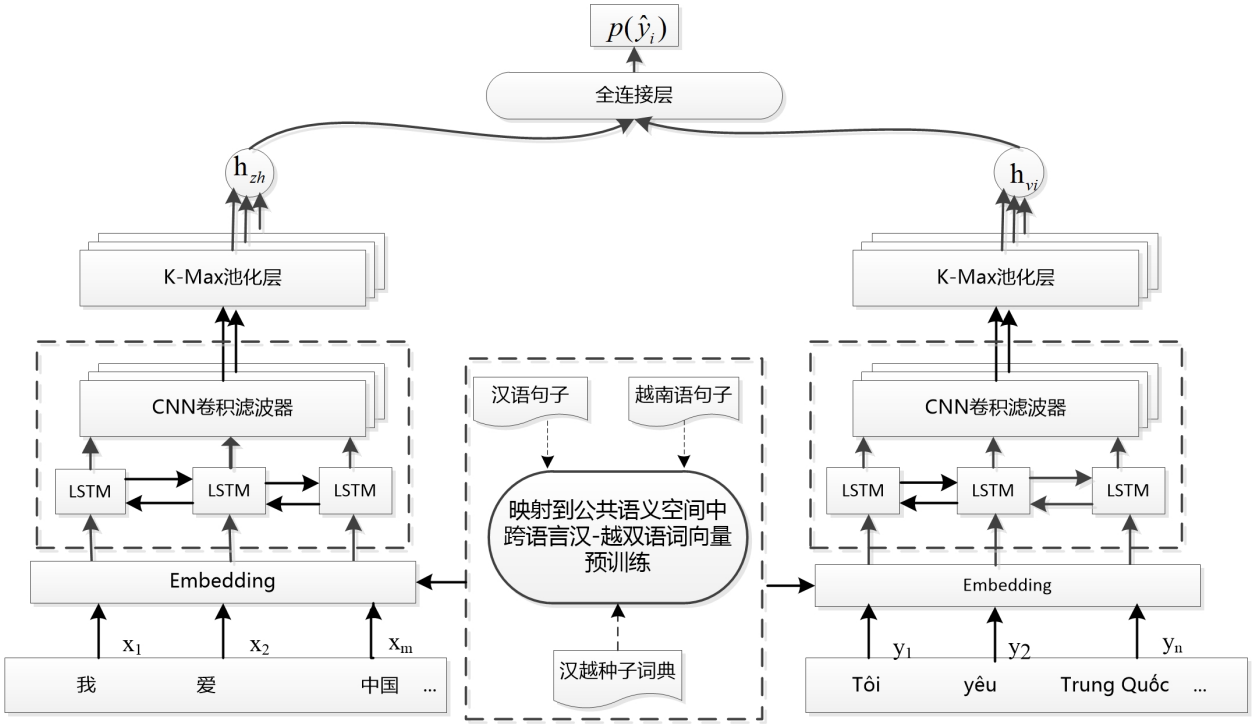

Figure 1: 基于汉-越双语预训练及Bi-LSTM的平行句对抽取模型

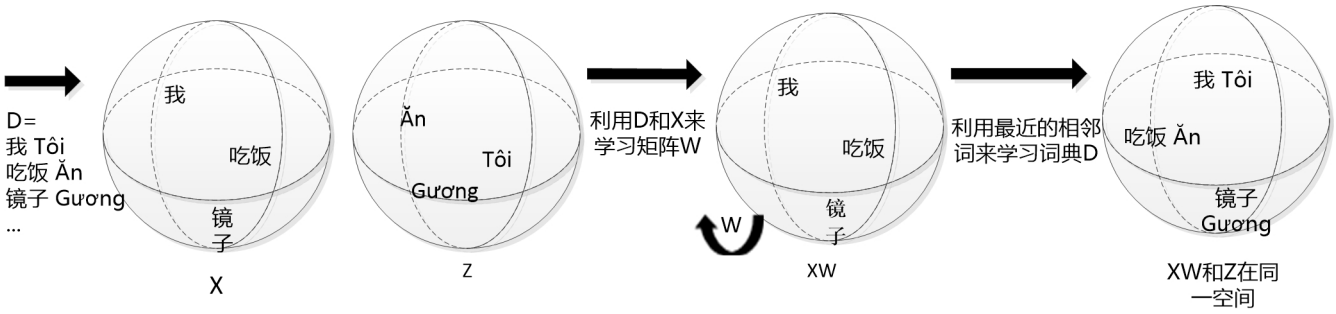

Figure 2: 汉-越跨语言双语词嵌入预训练过程

3.2 词向量预训练的基本步骤

首先，构建一个汉语和越南语同时映射的特征向量空间，汉语语料训练得到的词嵌入矩阵X，越南语语料中训练的词嵌入矩阵Z。将种子字典表示为一个二进制矩阵D。$D_{ij}=1$时代表越南语中的第j个单词是汉语中第i个单词的翻译。然后找到最佳映射矩阵W ,让汉语词向量和越南语词向量分布在同一个向量空间，使得映射汉语词嵌入$X_{i*}W$与越南语词嵌入Z_{j*}之间的欧几里德距离的平方和最小，映射矩阵:

$$W^* = \underset{W}{argmin} \sum_i \sum_j D_{ij} ||X_{i*}W - Z_{j*}||^2 \tag{1}$$

其中，预处理步骤对词嵌入矩阵X和Z进行长度归一化和平均居中，最后再进行一次归一化处理，并将W约束为正交矩阵即$WW^T = W^TW = I$，以强制执行汉语和越南语的单语不变性，防止单语性能的降低，同时可以产生更好的汉语-越南语跨原语言双语映射。在这种正交性约束下，最小化平方欧几里德距离就等于最大化点积，因此因此映射矩阵被定义为如下公式(2)所示:

$$W^* = \underset{W}{argmax}\, Tr(XWZ^TD^T) \tag{2}$$

其中，$Tr(\cdot)$表示主对角线上的所有元素之和，$W^* = UV^T$给出了此问题的最佳正交解，其中$X^TDZ = U\sum V^T$是X^TDZ的奇异值分解。由于字典矩阵D是稀疏的，这可以有效地在线性时间内对字典条目数进行计算。

获得了这个映射矩阵W之后，对于字典外的任何一个没有翻译的单词，可以根据映射后的空间余弦相似度来进行词对齐。在最近邻检索中，为每个源语言单词分配了目标语言中最接近的单词，我们将映射的源语言嵌入和目标语言嵌入之间的点积用作相似度度量。最后，通过矢量化相似矩阵XWZ^T并进行不断迭代计算，找到该矩阵的最大值，从而达到优化目标。

$$cos_{\text{dic}}(wx_i, z_j) = \frac{\sum_{i=1}^n wx_i z_j}{\sqrt{\sum_{i=1}^n (wx_i)^2 \sum_{j=1}^n (z_j)^2}} \tag{3}$$

4 基于Bi-LSTM和CNN公共语义空间编码

基于LSTM模型充分考虑了长距离单词之间的依赖性，并保留了诸如单词顺序之类的功能。同时CNN模型可以提取丰富的组合特征及卷积核的多样性。但是由于LSTM不使用反向单词编码信息，因此不能在双向单词编码中学习到语义信息特征，而Bi-LSTM可以考虑单词的双向编码。再使用CNN卷积并合并Bi-LSTM的输出以提取句子的关键语义特征。为了考虑上述特征，编码器由两层Bi-LSTM和CNN堆叠成一个基本的编码单元，依次从源语句和目标句中接受每个单词的单词嵌入矩阵$W_x \in R^{d\times|Vx|}$来输入单词x,其中d为单词嵌入向量的维数,V_x为所有输入单词的集合。每个时刻内，由词汇表V_x中的整数索引k定义的第i个句子中的标记表示为one-hot向量$w_k^S \in 0,1^{|V|}$,该one-hot向量与词嵌入矩阵$E^{S^T} \in R^{|Vx|\times d_e}$相乘，以获得该标记的连续向量表示$w_i^S$，其作为Bi-LSTM编码器的前向和后向循环状态的输入。前向LSTM读取变长句，并从第一个标记到最后一个标记更新其递归状态，从而创建一个固定大小的句子连续向量表示；后向LSTM反向处理该句子，然后将第二层相同位置上每个时间步长的两个方向的编码器输出都拼接在一起$h_i = [\overrightarrow{h}_i^S, \overleftarrow{h}_i^S]$，作为卷积神经网络的输入。前向递归状态和后向递归状态分别计算如下:

$$w_i^S = E^{S^T} w_k^S \tag{4}$$

$$\overrightarrow{h}_i^S = \phi(\overrightarrow{h}_{i-1}^S, w_i^S) \tag{5}$$

$$\overleftarrow{h}_i^S = \phi(\overleftarrow{h}_{i-1}^S, w_i^S) \tag{6}$$

$$h_i = [\overrightarrow{h}_i^S, \overleftarrow{h}_i^S] \tag{7}$$

其中E表示单词嵌入，$\phi(\cdot)$是LSTM模块。

原始的CNN由卷积层，池化层和全连接层组成。对于句子长度为n的句子，可以将它表示成$x_{1:n} = x_1 \oplus x_2 \oplus ... \oplus x_n$，$\oplus$表示全连接，$x_i \in \Re^d$表示的是第$i$个词向量，$d$表示的是词向量的维度。卷积运算的核心是对滑动窗口的大小的序列应用在过滤器上以产生新的特征，如下公式所示，

$$c_i = f(W \cdot x_{i:i+h-1} + b) \tag{8}$$

其中，$b \in \Re$是一个偏移向量，f是非线性函数（比如Sigmoid，ReLU）。长度为n的句子可以通过卷积层获得句子中任何连续单词序列的深层语义特征，如公式所示，

$$c = [c_1, c_2, ..., c_{n-h+1}] \tag{9}$$

本文将窗口大小为$F = [F(0)...F(\mathrm{m}-1)]$的卷积核与Bi-LSTM的输出向量进行卷积以获得特征向量，如公式所示:

$$c = tanh[(\sum_{i=0}^{m-1} h(t+i)^T F(i)) + b] \tag{10}$$

b是偏移向量，F和b是过滤器的参数。从典型的CNN结构可以看出，池化层构建在卷积层之上。在本文中，通过K-Max Pooling，每个滤波器最大值k会被保留，$\hat{c} = c_{k-max}$。

5 模型训练与分类

基于以上步骤，具有融合功能的Bi-LSTM 和CNN提取出源语句和目标句的语义特征，即C_i^S，C_i^T，然后使用元素积和绝对元素差来捕获它们的匹配信息，然后反馈到全连接的层以评估汉语-越南语句对相互翻译的可能性大小。具体公式如下:

$$C_i^a = C_i^S \odot C_i^T \tag{11}$$

$$C_i^a = |C_i^S - C_i^T| \tag{12}$$

$$C_i = tanh(W^a C_i^a + W^b C_i^b + b) \tag{13}$$

$$p(y_i|c_i) = \sigma(W^c c_i + c) \tag{14}$$

$$L = -\sum_{i=1}^{n(1+m)} y_i log\sigma(W^c h_i + c) - (1 - y_i) log(1 - \sigma(W^c h_i + c)) \tag{15}$$

其中$\sigma(\cdot)$是sigmoid激活函数W^a，W^b，W^c，b，c是模型参数，其中n是汉语句子的数量，m是候选越南语句子的数量。通过最小化标记的汉越句对的交叉熵作为损失函数来训练模型：对于预测，如果句子对的概率大于或等于设置的决策阈值ρ，则将其分类为平行；如果小于决策阈值ρ，则将其分类为不平行。

$$p(\widehat{y_i}) = \begin{cases} 0 & if \quad p(y_i = 1 \mid h_i) \geq \rho, \\ 1 & otherwise \end{cases} \tag{16}$$

6 实验与分析

6.1 实验数据

	汉越平行句对	汉越非平行句对
训练集	130k	130k
测试集	10k	10k
验证集	10k	10k

Table 1: 实验数据集表

本文将汉越平行句对抽取问题转化为二分类问题。由于在汉语到越南语低资源语言上，目前尚未找到用于训练的公开数据集，所以本文实验数据集的来源主要是从汉越新闻网站上检索和从维基百科dump获得的汉语-越南语翻译文章，该文章经过处理获取汉语句子315211，越南语单语句子316243，手动对齐以获得13万个汉越平行句对，使用VecMap工具训练高质量的跨语言双语词向量。同时基于每个平行句对的负采样样本数设置为1:1，随机构造了13万个汉越非平行句对，设置种子词典规模大小设置为3852个词条。同时为了衡量本文中汉越平行句对抽取模型分类器的性能，设置1万句汉越平行语料和1万句汉越非平行语料作为测试集。表1为本文基于跨语言双语预训练及Bi-LSTM的汉-越平行句对抽取模型的语料规模。

6.2 实验设置与评价指标

利用TensorFlow编写实现，单词嵌入维度和隐藏单元数均为300，隐藏层为1。批处理大小设置为64，训练的epochs为15，梯度截取设置为5.0，学习率设置为0.0002，Dropout设置为0.7-0.8，使用Adam优化器，激活函数采用sigmoid函数，损失函数为交叉熵损失函数。在评估指标方面，使用“精度”，“召回率”和“ F1值”作为衡量模型是否可以正确分类汉语-越南语是否为平行句子的指标。其中精度是所有提取的句子对中真正平行句子对的比例；召回率是测试集中所有平行句子对中真正平行提取的句子对的比例；F1值是精度和召回率的调和平均值。具体公式如下：

$$Precision = \frac{|TP|}{|TP + FP|} \tag{17}$$

$$Recall = \frac{|TP|}{|TP + FN|} \tag{18}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \times 100 \tag{19}$$

其中，TP是提取句子中真正平行的句对的数量，FP是提取句子中非平行句对的数量，FN是测试集中未被提取的平行句对的数量。

6.3 实验结果与分析

跨语言双语预训练使用Artetxe等人(Artetxe et al., 2017)提出的VecMap开源框架对加强汉越语言相关性及Bi-LSTM和CNN更好地捕获取句子上下文信息和局部信息，设计了以下三组对比实验，再通过上面的评价指标进行实验评价与分析。

实验一：为了验证预训练方法的有效性，设置阈值为0.90，将经过预训练的Bi-LSTM和CNN汉越平行句对抽取模型与不经过预训练的效果进行对比。我们还将仅使用Bi-LSTM抽取汉越双语平行句对的基线方法进行比较，同时，为了突出分类器构造比传统机器学习更深入学习具有更好的准确性，同时还比较了Munteanu D S等人(Munteanu and Marcu, 2005)提出的最大熵模型。具体实验结果见表2。 从表2中可以看出，在汉-越数据集上，本文模型的F1得分优于基线模型和其他模型。使用深度学习方法的Bi-LSTM模型与机器学习是支持向量机模型（SVM）和线性回归（LR）分类模型相比具有更好的效果，主要原因是Bi-LSTM模型可以更好的学习句子向量特征，并且孪生网络将汉越两种语言共享到同一语义空间中可以一定程度上解决跨语言的问题而机器学习方法无法解决跨语言的问题使效果明显下降。经过深度学习训练的特征提取分类器比最大熵模型具有更好的性能。其主要原因是神经网络能够自动学习并提取更好的特征。Bi-LSTM和CNN的结合优化于简单使用Bi-LSTM，是因为通过CNN可以获得更多的语义特征信息。基线模型的效果为63.6%,而本文方法的F1值达到了70.7%,与不做预训练和CNN特征提取相比提高了7.1%。经过跨语言预训练的模型比单独使用Bi-LSTM和CNN编码的效果要好是因为将汉-越两种语言映射到相同空间，语义相关性更好。

实验二：为了进一步证明本文提出的汉语-越南跨语言预训练方法的有效性，设置在阈值为0.9，做了一组将本文在词向量表征部分与word2vec(Mikolov et al., 2013)的词向量表征模型的对比实验，具体实验结果如表3所示。 从表3中可以看出，本文提出的预训练方法VecMap比word2vec在汉越双语抽取工作中的效果要好，其主要原因是VecMap是跨语言双语词向量预训练将汉越双语映射到公共语义空间训练加强汉越跨语言相关性，从而能抽取到更高质量的汉越双语平行句对。

方法	R(%)	P(%)	F1(%)
最大熵模型	54.2%	49.6%	51.8%
LR	57.9%	53.8%	55.7%
SVM	62.2%	57.3%	59.6%
LSTM	65.8%	59.7%	62.6%
Bi-LSTM	67.2%	60.5%	63.6%
BiLSTM+CNN	69.9%	61.4%	65.3%
本文方法(VecMap+BiLSTM+CNN)	75.6%	66.5%	70.7%

Table 2: 不同模型对比实验结果

方法	R(%)	P(%)	F1(%)
word2vec-BiLSTM-CNN	72.8%	64.2%	68.2%
本文法方法（VecMap）	75.6%	66.5%	70.7%

Table 3: 不同词向量表征方法对比实验结果

实验三：为了验证选取不同阈值时是否会影响模型的效果，为抽取到更高质量的汉越双语平行句对提供阈值参数基础，设置了在本文提出方法上不同阈值的对比实验，实验结果如表4所示。

不同的阈值M	R(%)	P(%)	F1(%)
M=0.8	77.3%	68.6%	72.7%
M=0.85	75.6%	66.5%	70.7%
M=0.90	73.9%	64.7%	68.9%

Table 4: 不同阈值对比实验结果

从表4中可以看出，不同的阈值M对实验结果的影响。其中，实验设置阈值参数越大，抽取汉越双语平行句对的F1分值反而越低。阈值M作为汉越双语平行句对抽取的判别值。

实验四：为了验证本文方法抽取出的汉-越平行句对对神经机器翻译模型性能的影响。本文选择了目前比较主流的神经网络模型Seq2seq+Attention(Vaswani et al., 2017)作为机器翻译模型，编码器和解码器的单词嵌入和循环状态的维度都设置为512，训练20个epochs，其中句对的批量大小为64。我们挑选了10万条汉-越平行句对作为基础训练集，并与添加了从可比语料中抽取的5万平行句对作对比，表5显示了不同规模数据的BLEU得分。从表5可以看出，在训练集中添加本文系统抽取到的5万平行句对后，翻译系统的BLEU得分分为15.89，提高了0.34，优于直接利用10万平行句对训练的翻译模型。实验结果证实了本文模型抽取到的汉越平行句对的质量，表明了可比语料库中存在大量语义空间相近的汉越平行句对。

数据规模	BLEU
100k	15.45
+（抽取50k）	15.89(+0.34)

Table 5: 平行句对对神经机器翻译性能的影响

7 结论

针对汉-越神经机器翻译数据稀缺的问题，本文提出了一种基于跨语言预训练及Bi-LSTM方法抽取汉越双语平行句对。在没有大规模汉越平行语料情况下，该方法利用汉越种子词典进行汉越跨语言预训练，将汉越双语表征到同一语义空间中，实现语义对齐。利用深度神经网络Bi-LSTM和CNN分别提起汉越句对的上下文信息和局部信息从而抽取出匹配度更高，噪声更小的汉越双语平行句。实验结果表明，经过跨语言预训练的平行句提取方法在准确率和召回率上高于基线模型，并且抽取到的汉-越平行句的语义更近。在未来的工作中，我们会探索该模型用于多语言平行句对抽取且在机器翻译的效果。

参考文献

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martınez Garcia. 2018. Extracting parallel sentences from comparable corpora with stacc variants. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, pages 48–52.

Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A factory of comparable corpora from wikipedia. In *Eighth Workshop on Building And Using Comparable Corpora*.

Houda Bouamor and Hassan Sajjad. 2018. H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proc. Workshop on Building and Using Comparable Corpora*.

Juryong Cheon and K. O. Youngjoong. 2017. Automatically extracting parallel sentences from wikipedia using sequential matching of language resources. *Ieice Transactions on Information And Systems*, E100.D(2):405–408.

Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. 2018. Neural machine translation for english-tamil. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2016. Parallel sentence extraction from comparable corpora with neural network features. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2931–2935, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Thomas C. Chuang, Jian Cheng Wu, Tracy Lin, Wen Chie Shei, and Jason S. Chang. 2004. Bilingual sentence alignment based on punctuation statistics and lexicon. In *Natural Language Processing-ijcnlp, First International Joint Conference, Hainan Island, China, March, Revised Selected Papers*.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef Van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

William A. Gale and Kenneth Ward Church. 1991. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*.

Jeenu Grover and Pabitra Mitra. 2017. Bilingual word embeddings with bucketed CNN for parallel sentence extraction. In *Proceedings of ACL 2017, Student Research Workshop*, pages 11–16, Vancouver, Canada, July. Association for Computational Linguistics.

Francis Grégoire and Philippe Langlais. 2017. A deep neural network approach to parallel sentence extraction.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Benjamin Marie and Atsushi Fujita. 2017. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Hitoshi Isahara Masao Utiyama. 2013. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Li Peng, Maosong Sun, and Xue Ping. 2010. Fast-champollion: A fast and robust sentence alignment algorithm. In *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*.

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*, 25(4):p.341–375.

Christoph Tillmann and Jian-ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 93–96, Boulder, Colorado, June. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.