山西大学学报(自然科学版)42(4):770－776,2019
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.cnki.shanxi.univ(nat.sci.).2019.06.03.009

基于句子特征向量的汉-越伪平行句对抽取

翟家欣,高盛祥*,余正涛,文永华,郭军军
(昆明理工大学 信息工程与自动化学院,云南省人工智能重点实验室,云南 昆明 650500)

摘　要:从可比语料中抽取伪平行句对是翻译语料扩充的重要方法之一。汉-越机器翻译是典型的资源稀缺型机器翻译,提高汉越翻译语料的规模能够显著提升汉越神经机器翻译性能。文章提出基于句子特征向量的汉越伪平行句对抽取方法,该方法首先根据汉越句法特性,将汉越句法差异部分的词性融入嵌入层,再使用自我注意力机制的神经网络抽取句子特征,生成一个句子特征向量,用这个句子特征向量来判断汉越句对是否为伪平行句对,实现从汉-越可比语料中抽取汉-越伪平行句对。实验表明,文章所提方法能够有效地从汉越可比语料中抽取出汉越伪平行句对。

关键词:句子特征向量;自我注意力机制;伪平行句对抽取;汉越机器翻译
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:0253-2395(2019)04-0770-07

Chinese-Vietnamese Pseudo-parallel Sentence Pair Extraction Based on Sentence Feature Vector

ZHAI Jiaxin,GAO Shengxiang*,YU Zhengtao,WEN Yonghua,GUO Junjun
(*School of Information Engineering and Automation, Kunming University of Science and Technology, Artificial Intelligent Key Laboratory of Yunnan province, Kunming* 650500, *China*)

Abstract: Extracting pseudo-parallel sentence pairs from comparable corpus is one of the important methods for translation corpus expansion. Chinese-Vietnamese machine translation is a typical resource-poor machine translation. Increasing the scale of the Chinese-Vietnamese translation corpus can significantly improve the performance of the Chinese-Vietnamese translation neural machines. This paper proposes a method based on sentence eigenvectors for the extraction of Chinese and Vietnamese pseudo-parallel sentence pairs. Firstly, according to the syntactic features of Chinese and Vietnamese, the method integrates the part of speech of the syntactic difference between the Chinese and Vietnamese into the embedded layer. Then use the self-attention mechanism to extract the sentence features and generate a sentence feature vector. Finally, this sentence feature vector is used to judge whether the Chinese and Vietnamese sentence pairs are pseudo-parallel sentence pairs, and the Chinese-Vietnamese pseudo-parallel sentence pairs are extracted from the Chinese-Vietnamese comparable corpus. The experiments show that the method can effectively extract the Chinese-Vietnamese pseudo-parallel sentence pairs from the Chinese-Vietnamese comparable corpus.

* **收稿日期**:2019-06-03;**接受日期**:2019-06-15
基金项目:国家重点研究开发计划项目(2018YFC0830105;2018YFC0830100);国家自然科学基金(61732005;61672271;61761026;61762056;61866020;61972186)
作者简介:翟家欣(1993－),女,山东烟台人,硕士研究生,研究方向为自然语言处理。E-mail:yt1209zzz@163.com
* **通信作者**:高盛祥(GAO Shengxiang),E-mail:gaoshengxiang.yn@foxmail.com
引文格式:翟家欣,高盛祥,余正涛,等.基于句子特征向量的汉-越伪平行句对抽取[J].山西大学学报(自然科学版),2019,**42**(4):770-776. DOI:10.13451/j.cnki.shanxi.univ(nat.sci.).2019.06.03.009

Key words: sentence feature vector; self-attention mechanism; pseudo-parallel sentence pair extraction; Chinese-Vietnam machine translation

0 引言

基于数据驱动的机器翻译(统计机器翻译、神经机器翻译)对训练模型中的数据量有较高的要求,神经机器翻译在语料需求方面更加显著。在大规模语料训练的神经机器翻译中如英-法、汉-英等的神经机器翻译都已经取得十分不错的成绩。但是,对于资源稀缺、语料规模小的神经机器翻译(如汉-越神经机器翻译),翻译性能并不十分理想。因此,扩大汉越机器翻译语料规模能够有效提升汉越神经机器翻译性能。

用于训练统计机器翻译模型和神经机器翻译模型的数据主要是双语平行句对,也就是互译的双语句对,由于没有可以用于汉越机器翻译的公开数据集,并且网上可爬取的汉越平行句对资源较少,汉越平行句对语料的数据规模有限,导致汉越神经机器翻译效果不佳。同时,互联网上存在大量的汉越可比语料,如汉语和越南语的维基百科数据,在这些可比语料中往往存在汉越平行句对。本文的目的是从汉越可比语料中抽取汉越伪平行句对。

判断两个汉越句子是否平行主要是比较汉越句子的特征,句子特征向量往往能够包含句子的一些特征,因此本文提出了基于句子特征向量的汉越平行句对抽取。该方法主要是解决句子特征向量如何包含更多的句子特征问题,从而提高判断汉越句对是否平行的准确率。本文将汉越句对的一些外部特征加入到嵌入层,并采用自我注意力机制在汉越句对拼成的一个句子中进行计算,尽可能多地抽取句子本身的特征,最终生成一个句子特征向量并用其作为判断汉越句对是否平行的依据。

统计机器翻译和神经机器翻译中都已经有学者在抽取训练语料方面进行了研究,提高了机器翻译的性能。在统计机器翻译方面,Rauf 等[1]将目标语言用机器翻译翻译为源语言,利用跨语言信息检索技术在双语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,Benjamin 等[2]基于词嵌入在单语语料中抽取平行句对,提升了神经机器翻译性能,Wang 等[3]基于句向量筛选了领域外和领域内相关的平行句对,提高了领域内的机器翻译性能。

以上方法都有效地抽取了伪平行句对,提高了机器翻译的性能,但是他们大多是从词的级别去比较两个句子是否平行,这种做法不容易捕捉句子本身的一些特征,如词的同义关系、共同指代关系、句子语义关系等,抽取到的伪平行句对噪声较大。

近几年,已经有了许多对自我注意力机制工作的研究,并将其应用在自然语言处理应用中,取得了更好的效果,如文本分类、情感分析。Lin 等[4]首先提出了自我注意力机制,并将其应用于双向 LSTM 隐层,从而在情感分类、文本蕴含任务中都取得了不错的效果。Ashish 等[5]提出了 Transformers 翻译模型,Transformer 最基本的构成是一种更简单易懂的自我注意力机制,并且在模型训练速度上和翻译效果上都有了很大的提升。BERT 模型[6]是在 Transformer 基础上的预训练模型,BERT 使用了双向的 Transformer,产生的向量能够更好地包含句子特征,并刷新了 11 个 NLP 任务的最好性能。

自我注意力机制能够很好地捕捉到句子本身的特征,生成一个更能表征句子的句子特征向量,减少句子本身的特征在句子相似度比较时的影响,使抽取到的伪平行句对噪声较小。因此,本文在嵌入层融入汉越语言差异特性部分的词性信息,并基于自我注意力机制生成句子特征向量,使其更好地抽取句子特征,从而更好地在汉越可比语料中抽取汉越平行句对。

1 汉-越平行句对抽取模型

1.1 汉越句法差异

句法结构简单的汉语和越南语句子的语序基本一致,其句法成分最基本的排列顺序都是主动宾(SVO)或主动补(SVP)[7]。汉语和越南语最大的差异是这两种语言的修饰语(定语、状语)与中心语的排列顺序不同。相对于汉语,越南语具有修饰语后置的特性,如表 1 中汉语越南语的例子体现了汉越句法差异特性,主要为词序差异。汉越句法特征差异总结如下:

(1)汉语句子中修饰名词性成分的定语排在所修饰的中心语之前,越南语中定语位于中心语之后,但当定语为数词、量词或指代词时,汉越语序一致,都将其放到中心语之前;

(2)汉语中描写性定语的排列顺序为:1. 主谓短语;2. 动词(短语)/介词短语;3. 形容词短语及其他描写性短语;4. 形容词或描写性短语。即,汉语的描写性定语的排列顺序为:1-2-3-4-中心语;而越南语的描写性定语的排列顺序为:中心语-4-3-2-1;

(3)状语主要分为限制性状语和描写性状语。对于大多限制性状语,汉语与越南语的语序基本一致,位于句首或主语谓语之间;对于表示时间的限制性状语,若该状语是由名词性短语构成,则汉语通常将其放到主语之后,而越南语将其放到句首;若该状语是由介词性短语构成,通常则其位于越南语句子的句尾;

(4)汉语中的描写性状语通常位于主谓之间,而越南语中描写性状语较为少见,若越南语句子中有对应的状语形式存在,则其语序与汉语一致,否则,越南语中通常是使用补语与汉语句子中的描写性状语对应,其位置通常位于句尾。

表 1 汉越句法差异

Table 1 Differences Between Chinese and Vietnamese Sentences Structure

成分	汉语	越南语
定语	我的房间在三楼。	Phòng(房间)tôi(我的)ôtrên(在)t ângba(三楼).
定语为数词、量词或指代词	我买了一本书。	Tôi(我)đãmua(买了)môtcu ôn(一本)sách(书).
描写性定语	她是我见过的最美丽的女孩。	Côlà(她是)côgái(女孩)xinhđep(美丽的)nh â t(最)tôitùngth ây(我见过的).
状语	他今天没有来上课。	Hôm nay(今天)anh(他)không(没)đ ênlóp(上课).

从以上特征中可以看出,汉语和越南语主要在修饰语与中心语的顺序上有所不同,而修饰语与中心语主要是由动词、副词、形容词、名词构成,并且在汉语句子中,相同的词在不同的位置时词性可能会不同,标注词性后的词在句子中往往更具有可识别性。因此,我们将汉语和越南语中的动词、副词、形容词、名词的词性标注出来,作为汉越句法差异特征融入到嵌入层。

1.2 自我注意力机制

本文采用的自我注意力机制是 Transformer 翻译模型中使用的自我注意力机制,也是 Transformer 的基础单位。并且,在 Facebook 的 OpenAI GPT 模型[8]和 Google 的 BERT 模型工作中也证明了 Transformer 对于抽取句子特征的有效性,并可以将这个句子表征应用于更多的下游任务中,并取得更好的效果。

图 1 为 Transformer 中的自我注意力机制计算的流程图[5],这种结构取代了神经机器翻译模型中的循环神经网络结构和卷积神经网络结构。注意力机制是要得到目标语言句子中的元素和源语言句子中哪些信息更有关联,给更有效的信息更高的权重,自我注意力机制是一种特殊的注意力机制,它同样是计算当前元素和句子中哪些元素更有关联,只是进行计算的对象从源语言句子和目标语言句子变为了句子本身,因此自我注意力机制也叫作内部注意力机制。

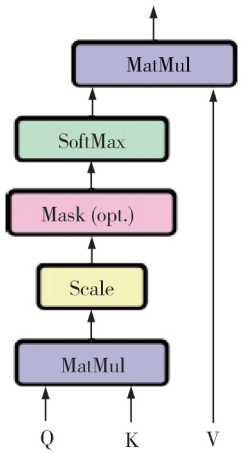

Fig. 1 Point product self-attention mechanism.

图 1 点积自我注意力机制

抽象地来说,源语言句子中的元素是由一系列<key,value>键值对组成,目标语言句子中的元素是由一系列 query 组成,那么对于注意力机制来说,计算目标语言句子中的 query 与各个 key 的相似性或者相关性,得到每个 key 对应 value 的权重系数,然后对 value 进行加权求和,即得到了最终的注意力机制结果。本质上,注意力机制是对目标语言句子中元素的 value 值进行加权求和,而 query 和 key 用来计算对应 value 的权重系数。那么我们可以得知,自我注

意力计算是在目标语言句子与源语言句子相同的情况下进行的注意力计算,即在一个句子内部进行注意力计算,此时 key、value 和 query 都是在同一个句子中。

图 1 中,注意力/自我注意力的计算主要分为 3 个步骤:

1. 计算 query 和每个 key 之间的相似性获得权重系数,常用的相似性函数包括点积、拼接、检测器等,本文使用的方法为点积;

2. 使用 softmax 函数正则化这些权重系数;

3. 最后将这些权重与相应的 value 一起加权求和获得最终的注意力结果。

公式(1)为本文使用的自我注意力计算公式:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d_k}}\right)\boldsymbol{V}\ , \tag{1}$$

公式 1 中 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别是句子中的 query、key、value,d_k 为 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 向量的维度。

1.3 句子特征向量表示

图 2 是本文提出的基于句子特征向量抽取汉越伪平行句对方法的模型框架,模型将伪平行句对抽取问题转化为一个二分类问题,使用 softmax 判断汉越句对否平行。模型的输入是汉语句子和越南语句子的拼接,先经过嵌入层后得到词嵌入,再经过自我注意力计算得到句子特征向量,最后使用 softmax 进行分类,得到最终的结果。

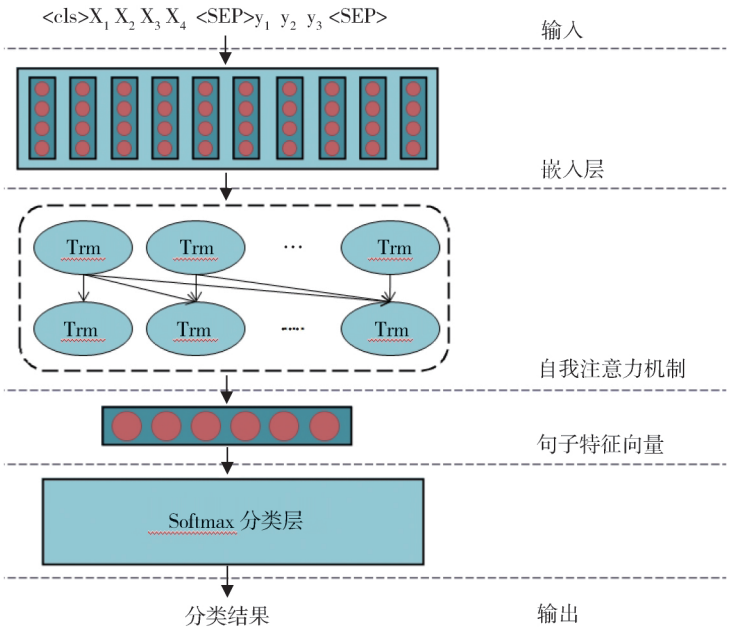

Fig. 2 Chinese-Vietnamese parallel sentence pair extraction model based on sentence feature vector

图 2 基于句子特征向量抽取汉越伪平行句对模型

在嵌入层中,由于下一步的自我注意力机制对一个序列进行计算时并不能捕获序列中的位置信息,也就是说两个相同元素组成的序列,尽管它们在元素的排列上有所不同,但是经过自我注意力机制得到的结果是相同的。因此通过在嵌入层加入位置嵌入层(Position Embedding)引入词的位置信息,解决自我注意力机制对词的位置不敏感的问题。同时,通过分句嵌入层(Segment Embedding),即用 $\boldsymbol{E}_A$ 和 $\boldsymbol{E}_B$ 两个特征来对两个句子进行区分,使模型能够区分两个句子。最后我们通过词性嵌入层(POS Embedding)来加入汉越句法差异特征部分的词性特征,主要为动词、名词、形容词和副词的词性,目的是让模型对这些词更敏感。图 3 为模型整个嵌入层的示意图。

图中的输入(Input)是汉语和越南语拼接成的一个大句子,并且在整个输入中加入[CLS]标签和[SEP]标签。每个输入的开始为标签[CLS],标签[SEP]作为汉越句子之间的间隔和结尾。输入的每个词最终的词嵌入为:

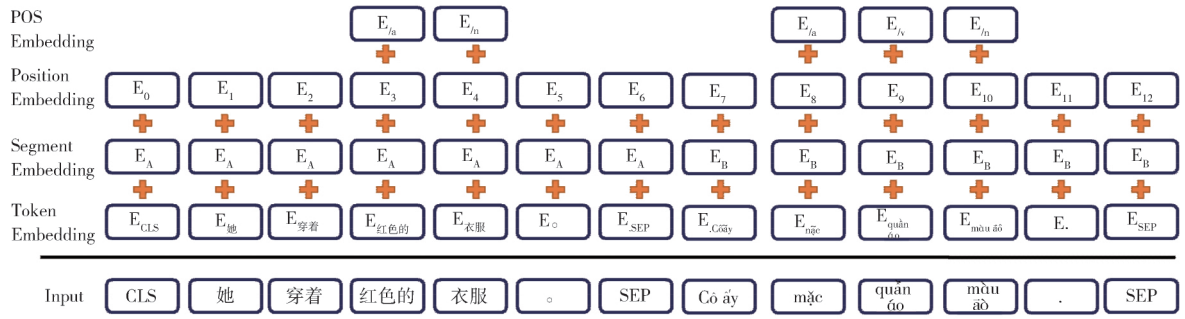

Fig. 3 Embedded layer of the model

图 3 模型嵌入层

$$\boldsymbol{E} = \boldsymbol{E}_{token} + \boldsymbol{E}_S + \boldsymbol{E}_P + \boldsymbol{E}_{POS}, \tag{2}$$

公式(2)中 $\boldsymbol{E}$ 为每个词的经过嵌入层输出的词嵌入,它是传统的词嵌入 $\boldsymbol{E}_{token}$、分句特征 $\boldsymbol{E}_S$、位置信息特征 $\boldsymbol{E}_P$、汉越句法差异特征部分的词性特征 $\boldsymbol{E}_{POS}$ 这四个向量之和。

在嵌入层中融入句子的外部特征,并用自我注意力在句子内部进行计算,最终得到的特征向量作为每个输入的汉越句子特征向量,我们用它来判断汉越句对是否平行。

2 实验

2.1 数据设置

本文将汉越平行句对抽取问题转化为分类问题,因此我们用从互联网上爬取的 12 万汉-越平行句对与 12 万不平行的汉-越句对作为汉-越伪平行句对抽取模型的训练数据,其中汉-越平行句对主要来自越南语学习网站,汉-越不平行句对是随机抽取中文维基百科与越南维基百科中的句子组合而成,并且每个句对后都有一个是否平行的标签。测试集由 5 000 汉越句对组成,其中 2 500 句对为汉越平行句对,2 500 句对为不平行句对。

表 2 为本文基于句子特征向量的汉越伪平行句对抽取模型的语料规模设置。

表 2 实验数据

Table 2 Experimental data

	汉越平行句对	汉越不平行句对
训练语料	120 k	120 k
测试语料	2.5 k	2.5 k

2.2 待分类汉越句对筛选

二分类判断汉越句对是否平行是将汉越可比语料库中的句子进行两两组合后再进行判断,而汉越可比语料规模较大,若汉越可比语料是由 10^6 的汉语句子和 10^6 的越南语句子组成,则需要进行 $10^6 \times 10^6$ 次分类计算,且句子特征向量计算复杂,这样就无法快速地从汉越可比语料中抽取汉-越伪平行句对。

为解决这个问题,我们使用以下方法先筛选出待分类的汉-越句对:

1)汉语和越南语处于不同的语言空间,为了方便计算,我们使用 Mikolov 等人[9]的方法先将预训练好的汉语词嵌入投影到越南语词嵌入空间,以便在同一空间表示汉语和越南语。

$$\boldsymbol{s} = \frac{1}{|S|}\sum_{i=1}^{|S|}\boldsymbol{e}_i^{emb}, \tag{3}$$

公式(3)为句嵌入的表示,其中,$|S|$ 为句子的长度,$\boldsymbol{e}_i^{emb}$ 是句子 S 第 i 个词在汉-越同一语言空间中的词嵌入。

$$S(x,y) = \Phi(\boldsymbol{x}^{emb}, \boldsymbol{y}^{emb}), \tag{4}$$

使用公式(4)计算汉语句子和越南语句子的相似度,其中 $\Phi(\boldsymbol{x}^{emb}, \boldsymbol{y}^{emb})$ 是句子 x 和句子 y 在同一语言空间下的余弦相似度。我们用构成句子词嵌入的平均来表示每个句子,即句嵌入 s,并用句嵌入计算每对汉越候选“平行”句对的相似度,得到一个分数 $S(x,y)$。并且为每个汉语句子保留 10 个最接近的越南语句子。

(2)平行的汉-越句对的长度比应在一定范围内,若汉-越句对的长度比过大或过小,那么他们平行的概率较低。我们根据汉-越平行语料统计出汉越平行句对的长度比范围,剔除超出这个范围的句对。

经过以上两个步骤,我们从汉越可比语料中筛选出更可能是互译关系的汉越句对,大大缩小了模型的计算次数,减少了时间复杂度。

2.3 实验结果

为验证自注意力机制和汉越语言差异特性能更多地获取句子特征,我们也进行了不加入汉越句法差异特性和基于 LSTM[10] 获取句子特性的实验,表 3 是不同方法的实验结果,评价指标为准确率。

表 3 不同方法的实验结果

Table 3 Experimental results of different methods

方法	准确率/%
LSTM	59.16
LSTM+POS	60.45
Self-attention	62.94
Self-attention+POS	63.32

从表 3 中的结果我们可以看出加入汉越句法差异特性部分的词性后,普遍提高了判断汉越句对是否平行的准确性;基于自我注意力机制[11]的平行句对分类模型要远远好于基于 LSTM 的平行句对分类模型,我们认为这是由于自我注意力机制对句子内部进行加权计算更能捕捉到句子本身的隐性特征,产生更好的句子表征。

我们最终的目的是在汉越可比语料中抽取汉越伪平行句对,因此我们在可比语料中进行了实验,表 4 为从汉越可比语料中抽取到的汉越伪平行句对展示。我们可以观察出,从可比语料中抽取到的汉越伪平行句对更多是句子意思相近,并不是完全平行的句对,噪声较大。

表 4 抽取到的汉越伪平行句对

Table 4 Extracted Chinese-Vietnamese pseudo-parallel sentence pairs

汉语	越南语
维基百科是自由内容、协同编辑以及多语言版本一个的网络百科全书项目	Wikipedia làmôtbách khoa toànthu'm ở vó'imucđíchchínhlàchophépmoingu'iđềucóthể viếtbàib ằngnhiềuloaingônngữ'trênInternet.(维基百科是一本开放的网络百科全书,其目的是允许每个人写文章,有多种语言。)
当前是全球网络上最大且最受大众欢迎的参考工具书,名列全球十大最受欢迎的网站,维基百科当前由非营利组织维基媒体基金会负责营运。	Wikipedia đanglàcôngtrìnhthamkh ả oviếtchungló'nnhâtvàph ổ biếnnhâttrên Internet, vàhiêntaiđu'o'cxếphangtrang web ph ổ biếnthú ' 5 trêntoànâu. Wikipedia thuôcvề ổ chú'c phi lo'inhuân Wikimedia Foundation.(维基百科目前是互联网上最大和最受欢迎的一般写作的参考,目前在全球排名第五。维基百科属于维基媒体基金会的非营利组织。)
越南社会主义共和国,通称越南,是位于东南亚中南半岛东端的社会主义国家。	ViêtNam,tênchínhthú'c: CônghòaXãhôich ủnghĩaViêt Nam làqu ốcgian ằm ở phía Đ ôngbánđ ả o Đ ôngDu'o'ngthuôckhuvu'c Đ ông Nam Á.(越南,越南社会主义共和国是一个位于东南亚东印度支那半岛的国家。)

3 结论

本文主要是解决从可比语料中抽取汉越伪平行句对的问题。这个问题本质上是跨语言进行句子的分类,本文直接将汉语句子与越南语句子进行拼接,作为一个整体输入,并且在其中融入汉越句法差异特性部分的词性,利用自我注意力机制生成句子特征向量,这些做法都是为了产生一个更能表征句子的句子特征向量。实验结果表明,融入汉越句法差异特性部分的词性和使用自我注意力机制都证明了所提方法的有效性。

参考文献:

[1] Rauf S A, Schwenk H. Parallel Sentence Generation from Comparable Corpora for Improved SMT[J]. *Machine Translation*, 2011, **25**(4): 341-375. 2011. DOI: 10.1007/s10590-011-9114-9.

[2] Benjamin M, Atsushi F. Efficient Extraction of Pseudo-Parallel Sentences from Raw Monolingual Data Using Word Embeddings[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics(ACL), July 30-August 4, 2017, Vancouver, Canada. 2017: 392-398. DOI: 10. 18653/v1/P17-2062.

[3] Wang R, Finch A, Utiyama M, *et al*. Sentence Embedding for Neural Machine Translation Domain Adaptation[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), July 30-August 4, 2017, Vancouver, Canada. 2017, **2**: 560-566. DOI: 10. 18653/v1/P17-2089.

[4] Lin Z, Feng M, Santos C N, *et al*. A Structured Self-attentive Sentence Embedding[Z/OL]. 2017. arXiv: 1703. 03130.

[5] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is All You Need[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017(NISP), December4-9, 2017, Long Beach, CA, USA. 2017: 6000-6010.

[6] Devlin J, Chang M W, Lee K, *et al*. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[Z/OL]. 2018. arXiv: 1810. 04805.

[7] **武氏河. 越南语与汉语的句法语序比较**[J]. **云南师范大学学报(对外汉语教学与研究版)**, 2005, **3**(6): 65-68. DOI: 10. 3969/j. issn. 1672-1306. 2005. 06. 115.
Vu T H. A Comparison Between Vietnamese and Chinese Syntactic Constituent Orders[J]. *Journal of Yunnan Normal University(Teaching and Research of Chinese as A Foreign Language)*, 2005, **3**(6): 65-68. DOI: 10. 3969/j. issn. 1672-1306. 2005. 06. 115.

[8] Radford A, Narasimhan K, Salimans T, *et al*. Improving Language Understanding with Unsupervised Learning[R]. Technical report, OpenAI, 2018.

[9] Mikolov T, Le Q V, Sutskever I. Exploiting Similarities Among Languages for Machine Translation[Z/OL]. 2013. arXiv: 1309. 4168.

[10] Zhou J, Xu W. End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL), July 26-31, 2015, Beijing, China. 2015, 1: 1127-1137.

[11] Li Q Y, Miao Z J. Self-attentional Convolution for Neural Networks[J]. *Journal of Physics: Conference Series*, 2019, **1169**(1): 012034. DOI: 10. 1088/1742-6596/1169/1/012034.