文章编号：1003-0077(2021)09-0058-08

基于语义自适应编码的汉-越伪平行句对抽取方法

郭军军[1,2]，田应飞[1,2]，余正涛[1,2]，高盛祥[1,2]，闫婉莹[1,2]

(1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500)

摘 要：伪平行句对抽取是缓解汉-越低资源机器翻译中数据稀缺问题的关键任务，同时也是提升机器翻译性能的重要手段。传统的伪平行句对抽取方法都是基于语义相似性度量，但是传统基于深度学习框架的语义表征方法没有考虑不同词语语义表征的难易程度，因此导致句子语义信息不充分，提取到的句子质量不高，噪声比较大。针对此问题，该文提出了一个双向长短期记忆网络加语义自适应编码的语义表征网络框架，根据句子中单词表征难易的不确定性，引导模型使用更深层次的计算。具体思路为：首先，对汉语和越南语句子进行编码，基于句子中单词语义表征的难易程度，自适应地进行表征，深度挖掘句子中不同单词的语义信息，实现对汉语和越南语句子的深度表征；然后，在解码端将深度表征的向量映射到统一的公共语义空间中，最大化表示句子之间的语义相似度，从而提取更高质量的汉-越伪平行句子。实验结果表明，相比于基线模型，该文提出的方法在 F_1 得分上提升 5.09%，同时将提取到的句子对用于训练机器翻译模型，实验结果表明翻译性能的显著提升。

关键词：数据稀缺；语义表征；自适应编码

中图分类号：TP391 **文献标识码**：A

Pseudo-Parallel Sentence Pair Extraction for Chinese-Vietnamese Based on Semantic Adaptive Coding

GUO Junjun[1,2], TIAN Yingfei[1,2], YU Zhengtao[1,2], GAO Shengxiang[1,2], YAN Wanying[1,2]

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: Pseudo-parallel sentence pair extraction is a key method to improve the performance of low-resource machine translation such Chinese-Vietnamese. Existing methods based on deep learning framework do not consider the difficulty of semantic representation of different words, which leads to insufficient semantic information of sentences, low quality of extracted sentences and high noise. To solve this problem, this paper proposes a semantic representation network framework of bidirectional LSTM plus semantic adaptive coding. The specific idea is to encode Chinese and Vietnamese sentences first, and adaptive representation is carried out to deeply mine the semantic information of different words in the sentence to realize the depth representation of Chinese and Vietnamese sentences. Then the vector of depth representation is mapped to a unified common semantic space to maximize the semantic similarity between the sentences for higher quality Chinese-Vietnamese pseudo-parallel sentences. The experimental results show that the model improves F_1 score by 5.09%, which is better than the baseline model.

Keywords: data scarcity; semantic representation; adaptive encoding

0 引言

平行语料库的规模和质量对机器翻译的性能存在重要的影响。但平行语料库的构建成本很高，因为其需要相关领域人员的专业知识，在面对低投资语言时，平行语料的稀缺性问题显得更加严重，其中，汉语到越南语就属于低资源语言。通过大量的文本研究发现，在汉语和越南语之间存在着丰富的汉-越段落级可比语料库，这里可比语料库指的是主

收稿日期：2020-01-09 **定稿日期**：2020-04-24

基金项目：国家自然科学基金(61732005,61672271,61761026,61762056,61866020)；国家重点研发计划(2019QY1802)

题对齐但非句子对齐的多语言文本的集合。如表 1 所示，汉语和越南语段落级的可比语料库中存在语义非常接近的句子对。伪平行句对抽取的目的就是从这些大量的可比语料库中挖掘出语义更加接近、质量更高的句子，以增加数据的数量和所覆盖域的范围。

表 1　汉-越可比语料库中存在语义非常相似的句子

汉语	越南语
农业和农村发展部的单位还不止此地组织了许多工作团，以学习和交流在全国各省市之间在河内和越南之间发展农业生产的经验。***地区积极与农业和农村发展部各部门协调，组织会议，研讨会和论坛，以联系农产品的安全生产和消费。*** 同时，与市内各部门，部门，区，镇人民委员会和社会政治组织协调，加强安全农产品消费的联系。	Chưa dừng lại ở đó, các đơn vị thuộc Sở NN&PTNT đã xây dựng tổ chức nhiều đoàn công tác đi học tập trao đổi kinh nghiệm tại các tỉnh, thành phố trong cả nước về phát triển sản xuất nông nghiệp giữa thành phố Hà Nội và các địa phương. ***Chủ động phối hợp với các đơn vị của Bộ NN&PTNT tổ chức hội nghị, hội thảo, diễn đàn kết nối sản xuất, tiêu thụ nông sản an toàn.*** Đồng thời, phối hợp với các sở, ngành, UBND các quận, huyện, thị xã, các tổ chức chính trị xã hội trên địa bàn thành phố tăng cường công tác kết nối tiêu thụ sản phẩm nông sản an toàn.

如何从段落级的可比语料库中筛选出候选句子是本文的首要工作。首先，构建一个汉-越平行词典，然后，利用词典对段落级可比语料库进行预筛选，得到伪平行候选句子。例如，在表 1 中，汉语单词“积极”和越南语单词“Chủ động”是平行的；汉语单词“协调”和越南语单词“phối hợp”是平行的，当两个句子中的常见单词分别对齐之后，就将它们作为伪平行候选句子，因为这些单词只需进行一次编码，比较容易得到语义信息；但对于比较难的单词，例如，汉语单词“与”和越南语单词“với”就比较难以提取单词语义信息，需要进行更深层次的思考，深度挖掘词级语义信息。

传统的解决思路是利用词嵌入的方式对齐句子中的短语，然后提取平行句子。例如，Benjamin[1] 等提出了一种无须依赖任何文档级信息即可从一对大型单语语料库中提取伪平行句子的新方法，该方法首先利用词嵌入有效评估了数万亿个候选句子对，然后使用分类器进行查找，提升了神经机器翻译的性能；Minh Thang Loung[2] 等使用双语单词嵌入模型学习单词表示后，使用相似矩阵上的卷积神经网络对一对句子是否对齐进行分类，从而提取平行句子；Sanjika[3] 等探索了三种短语对齐方法来检测词嵌入在可比较句子中的并行短语对，当出现大量短语对候选时，可检测平行短语对。虽然这些方法取得了一定的性能，但传统的神经网络构建的深度学习体系结构不能够充分学习单词语义之间的依赖关系，导致句子表征能力不足，语义信息不充分，提取到的句子噪声较大，质量不高。

本文针对上述问题，并结合汉-越句子特性，受 Zhang[4] 等人思想的启发，提出了一个基于语义自适应编码的伪平行句提取系统，在编码端设置了一个思考模块，根据句子中单词表征难易的不确定性，引导模型使用更深层次的计算，进一步对源语言和目标语言进行句子上下文特征提取，深度挖掘句子中不同单词的语义信息，精炼每个时间步长的表示，并将深度表征的句子向量映射到统一的公共语义空间中，最大化表示句子之间的语义相似性。在本文方法中，我们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，所以我们把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。

1　相关工作

从可比语料库中提取伪平行句子并构建平行语料库提升机器翻译性能，对于低资源语言来说是行之有效的一种方法，最理想的方法是手工进行抽取，但是这样的成本比较高。利用统计机器翻译和神经机器翻译两种方法从可比语料库中抽取句子是比较有效的方法，也分别有学者进行了研究。在统计机器翻译方法中，Rauf[5] 等人提出了用统计机器翻译的方法翻译可比语料库的源语言部分，并将这些翻译作为查询，从可比语料库的目标语言方面进行信息检索生成平行语料库，提高了统计机器翻译的性能；而 Rauf[6] 等的工作是翻译可比语料库的源语言，然后与目标句子比较，以在目标语言找到候选句子；Alberto Barron Cedeno[7] 等提出了一种从维基百科自动提取域内可比语料库的模型，可以自动提取单语和可比较的文章集，并按需为语言对和领域提供一键式生成并行语料库，改善了机器翻译质量，并将其应用于特定领域的语料库。虽然上述方法取得了一定的成就，但是需要在翻译模型性能比较好

的基础上才能进行,同时在信息检索技术中存在词语语义信息不足的问题。

而在神经机器翻译方法中,Chehui Chu[8]等提出了一种基于深度学习架构的平行句对抽取模型,该模型包括一个候选平行句对筛选器和一个平行句对判别器。基于维基百科数据的实验结果表明,本文所提方法在平行句对抽取的准确性和统计机器翻译性能上均优于已有基线模型。而 Francis Gregoire[9]等是基于双向递归神经网络分别对源语言和目标语言进行编码,然后经过分类器区分源句子和候选目标句子是否平行;Cristina Espana Bonet[10]等通过测量翻译之间的相似度以及语义相关和语义不相关的句子对来评估语言对的质量和有效性,然后结合上下文向量和相似性度量在可比语料库中识别平行句子,达到了预期的效果;Juryong Cheon[11]等提出了一种基于语言资源查找相似句子的方法,用于从维基百科构建英语和韩语之间的平行语料库;Resnik[12]等提出了一种基于 HTML 从 Web 中提取相似文档的方法;Talvensaari[13]等提出了一种利用主关键词从源语言到目标语言的翻译词找到相似文档的方法。综上所述,这些方法都是从句子级扩充训练数据,然后构建高质量的平行语料库。虽然他们的方法都能很好地抽取伪平行句子,改善机器翻译的性能,但都是针对资源丰富语言(如英语-法语),而在低资源语言(如汉语-越南语)上性能则较差,同时提取到的句子噪声较大。本文在神经网络模型的基础上,引入了语义自适应编码的方法,更深层次地挖掘汉语和越南语的语义特征,然后比较它们之间的语义相似度,从而提取更高质量的汉-越伪平行句子,提升低资源下汉-越神经机器翻译的性能。

2 基于语义自适应编码的伪平行句对抽取模型

针对汉-越伪平行数据抽取的问题,本文在语言模型中增加语义自适应编程模块,提升模型对句子的语义表征能力,进而提高伪平行句对抽取的性能,最终提升机器翻译的性能。本文方法主要包括双语同步编码器,分别对源语言和目标语言进行编码,设 $x=\{x_1,x_2,\cdots,x_m\}$ 表示源语言句子,m 为源语言句子的长度,$y=\{y_1,y_2,\cdots,y_n\}$ 表示目标语言句子,n 为目标语言句子的长度。其中,编码器是堆叠的两层长短期记忆网络(lony short-term memory,LSTM),思考模块由竖向 LSTM 组成,同时对源语言和目标语言深度挖掘语义信息,以改善单词表示与句子上下文连接之间的文本信息,然后将深度表征的文本信息映射到统一的公共语义空间中,最大化表示句子之间的相似性。

2.1 语义自适应编码模型

本节首先介绍语义自适应编码模型,该模型结构由编码器、思考模块和预测模块组成。模型结构体系如图 1 所示。

2.1.1 编码器

编码器由两层 LSTM 堆叠成一个基本的编码单元,依次从源句和目标句中接收每个单词的单词嵌入矩阵 $\boldsymbol{W}_x\in R^{d\times|V_x|}$ 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。在每个时刻内,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 $\boldsymbol{w}_k^S\in\{0,1\}^{|V|}$,该独热向量与词嵌入矩阵 $\boldsymbol{E}^{ST}\in\mathbb{R}^{|V_x|\times d_e}$ 相乘,以获得该标记的连续向量表示 $\boldsymbol{w}_i^S$,其用作 BiLSTM 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循环状态,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,作为思考模块的输入。前向递归状态和后向递归状态分别计算如式(1)~式(4)所示。

$$\boldsymbol{w}_i^S=\boldsymbol{E}^{S^T}\boldsymbol{w}_k^S \tag{1}$$

$$\vec{\boldsymbol{h}}_i^S=\varphi(\vec{\boldsymbol{h}}_{i-1}^S,\boldsymbol{w}_i^S) \tag{2}$$

$$\overleftarrow{\boldsymbol{h}}_i^S=\varphi(\overleftarrow{\boldsymbol{h}}_{i+1}^S,\boldsymbol{w}_i^S) \tag{3}$$

$$\boldsymbol{h}_i=[\vec{\boldsymbol{h}}_i^S;\overleftarrow{\boldsymbol{h}}_i^S] \tag{4}$$

其中,$\boldsymbol{E}$ 表示单词嵌入,$\varphi(\cdot)$是 LSTM 模块。

2.1.2 思考模块

(1) 深度语义自适应思考模块

在语义自适应思考模型中,根据句子最大长度定义竖向 LSTM 思考模块,每个输入单词分别对应每一竖向 LSTM 思考模块,在竖向 LSTM 思考模块中,进行竖向的信息交流,当要更新状态 s_i 时,用上一时刻的状态 s_{i-1} 根据式(5)进行计算,模型进入思考模式,如图 2 所示,该模型的关键特征就是在预测标记前自适应地选择思考模块的深度。

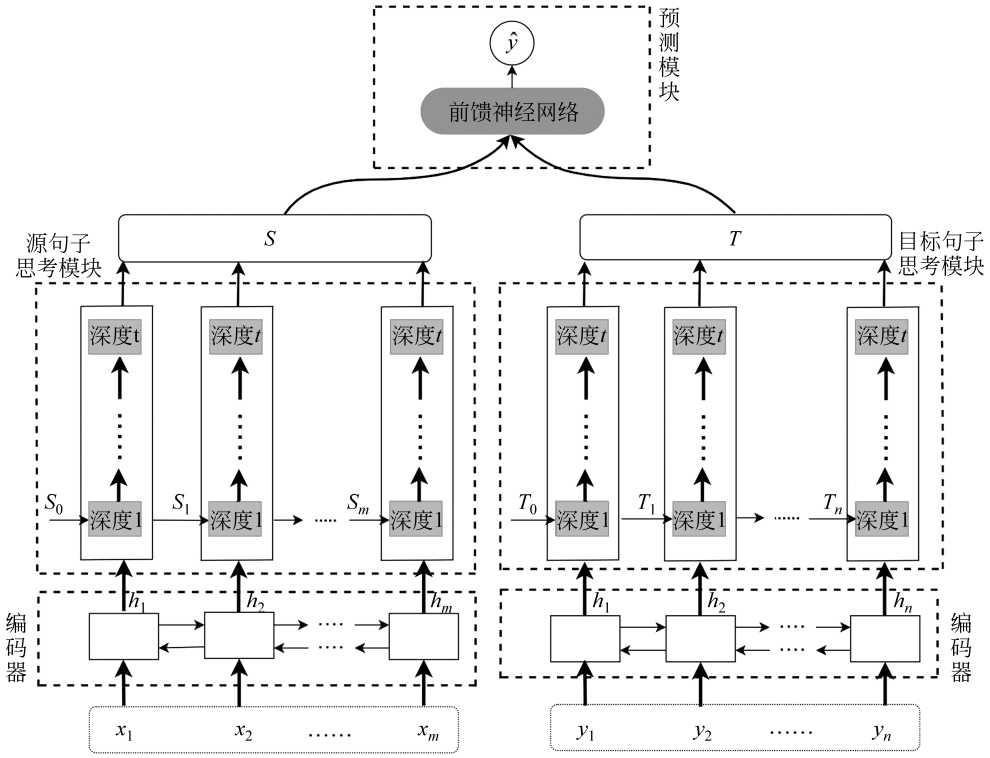

图 1　语义自适应编码结构体系

$$\boldsymbol{s}_i = f_{\mathrm{LSTM}}([\boldsymbol{e}_i;\boldsymbol{h}_i;\boldsymbol{c}_i;\mathrm{flag}],\boldsymbol{s}_{i-1}) \tag{5}$$

其中，$\boldsymbol{e}_i$ 是每个单词的词嵌入向量；$\boldsymbol{h}_i$ 是每个时刻的隐藏状态；s_0 初始化为 h_0；flag 是我们连接到输入嵌入的标记，取 1 或 0，其中 0 表示停止思考，1 表示继续进行思考。[·，·]表示向量的拼接。$\boldsymbol{c}_i$ 是上下文向量，由式(6)进行计算：

$$\boldsymbol{c}_i = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{h}_i \tag{6}$$

其中，$\boldsymbol{\alpha}_i$ 是全局注意力向量，$\boldsymbol{h}_i$ 是每个时刻的隐藏状态。为了表示清晰，我们只列出了源句子的思考状态 s_i，对于目标句子的思考状态 t_j 也采用同样的方式进行计算。

(2) 语义相似性度量模块

为了评估每个词在句子中的语义相似度，状态 s_i 进一步发送到语义相似性度量模块，借鉴 Zhang[4] 等人的处理思路，本文使用方差来表征模型的不确定性 $U_n(\boldsymbol{s}_i)=\mathrm{Var}(q)$，语义相似度计算模块如式(7)、式(8)所示。

$$U_n(\boldsymbol{s}_i) = \min(\gamma, \mathrm{Var}(q))/\gamma \tag{7}$$

$$q = \{p(\hat{y}_i = y_{i+1} \mid \boldsymbol{s}_i) \mid \theta_i\}_{i=1}^{F} \tag{8}$$

其中，p 是根据思考状态 s_i 预测下一个词 y_{i+1} 的概率，θ_i 是第 i 个前向传递中受 Dropout 影响的所有扰动参数的集合，F 为思考次数，在本文中参

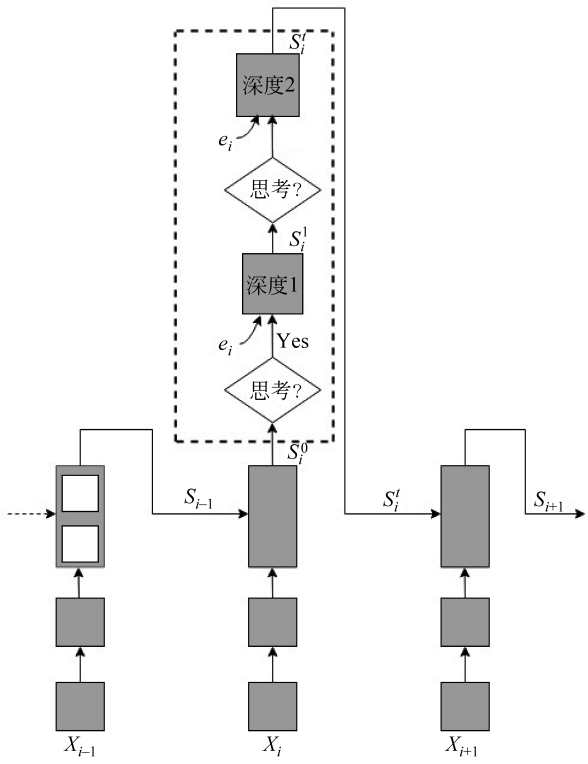

图 2　语义自适应思考模块

为简洁起见，仅显示 s_i 的思考模式以及前面两个思考深度，最大思考深度为 $t=3$。

数 $\gamma=0.15$。模型连续地从动作空间 $A=\{\text{Ponder}, \text{Stop}\}$ 中做出决策，每次选择 $a\in A$。如果方差大于或等于规定的决策阈值，模型就停止思考，如式(9)所示。

$$R(a \mid \boldsymbol{s}_i^{(t)})=\begin{cases}0 & \text{if } U_n(\boldsymbol{s}_i^{(t)})\geqslant\rho\\ 1 & \text{if } U_n(\boldsymbol{s}_i^{(t)})<\rho\end{cases} \tag{9}$$

如果方差小于规定的决策阈值，模型重新进入思考，再次使用式(5)进行状态更新，同时对输入使用相同的词嵌入 e_i，如式(10)所示。

$$\boldsymbol{s}_i^{(t)}=f_{\text{LSTM}}([\boldsymbol{e}_i;\boldsymbol{h}_i;\boldsymbol{c}_i;\text{flag}],\boldsymbol{s}_{i-1}^{(t-1)}) \tag{10}$$

其中，$\boldsymbol{s}_i^0=\boldsymbol{s}_i$，flag=1。模型将继续思考，直到其选择停止或达到 $t=3$ 的极限。

2.2 预测模块

预测模块由一个带 sigmoid 的前馈神经网络组成，首先把思考模块的每个时刻的输出状态进行拼接，得到源句和目标句的最终表示形式，将最终表示形式输入到前馈神经网络，该神经网络计算它们平行的概率，如式(11)、式(12)所示。

$$\boldsymbol{u}_t=\tanh(\boldsymbol{W}^{(1)}s_i+\boldsymbol{W}^{(2)}t_j+\boldsymbol{b}) \tag{11}$$

$$p(y_t=1 \mid \boldsymbol{u}_t)=\sigma(\boldsymbol{v}\boldsymbol{u}_t+\boldsymbol{b}) \tag{12}$$

其中，s_i 和 t_j 分别表示源句子和目标句子的最终表示形式，$\sigma(\cdot)$ 是 sigmoid 激活函数，$\boldsymbol{W}^{(1)}\in R^{d_f\times d_h}$，$\boldsymbol{W}^{(2)}\in R^{d_f\times d_h}$，$\boldsymbol{v}\in R^{d_f}$，$b\in R^{d_f}b$ 分别是模型参数，d_f 是前馈神经网络中隐含层的大小。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。

$$\hat{y}_t=\begin{cases}1 & \text{if } p(y_t=1 \mid \boldsymbol{u}_t)\geqslant\rho\\ 0 & \text{otherwise}\end{cases} \tag{13}$$

3 实验和结果

为了评估本文方法从可比语料库中抽取伪平行句子的有效性，我们进行了多次实验比较。在 3.1 节中，将介绍进行实验的数据集；3.2 节介绍实验设置；3.3 节中介绍了实验的结果并分析模型在提取汉-越伪平行句子方面的性能；3.3.4 节中将提取到的汉-越伪平行句子作为训练数据，评估机器翻译系统的性能。

3.1 数据集

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，所以我们利用汉-越平行词典从汉语到越南语段落级可比语料库中预筛选了 20 万对的汉语到越南语候选句子以及从维基百科中随机抽取了 10 万对的汉语到越南语候选句子，组成了训练数据集。表 2 列出了训练数据的规模。

表 2 实验数据集

	汉语数据集	越南语数据集
训练集	30 万	30 万
测试集	2 000	2 000
验证集	2 000	2 000

3.2 实验设置

本文模型利用 PyTorch 编写实现，单词嵌入维度和隐藏单元数均为 512，批处理大小设置为 64，最大思考步数设置为 3，学习率设置为 0.000 3，采用指数级的学习率衰减策略，Droupt 设置为 0.3，使用 Adam 优化器。为了评估模型的性能，本文分别用精度、召回率和 F_1 得分进行评估。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；F_1 得分是精度和召回率的调和平均值。

3.3 实验结果

3.3.1 语义自适应编码模型的实验结果及分析

为了验证模型的性能，分别在精度、召回率和 F_1 得分上与基线模型进行了实验比较。表 3 列出了本文模型以及基线模型的实验结果，其中，BiLSTM 是基线模型，+Pondering 是本文方法，表示在基线模型的基础上引入了语义自适应思考模块。

表 3 模型在汉语到越南语数据集上的精度(P)、召回率(R)和 F_1 分数

类型	P/%	R/%	F_1/%
RNN	58.95	47.43	52.57
LSTM	59.16	51.32	54.96
Bi-RNN	83.72	71.90	77.36
Bi-LSTM	86.35	74.74	80.14
Bi-LSTM+BERT	86.50	81.94	84.16
Ours	**92.86**	**85.91**	**89.25**

实验结果表明，在汉-越数据集上，本文模型的得分优于基线模型。由于 BiRNN 和 BiLSTM 神经

网络模型能够对句子进行双向编码，进一步提取上下文特征信息，所以在平行句对提取中得分比 RNN 和 LSTM 高；其中，Bi-LSTM + BERT 是利用 BERT 模型分别对源语言和目标语言进行预训练得到词向量作为模型的输入，所以相比其他基线模型性能有所提升。而本文模型在 Bi-LSTM 神经网络模型的基础上引入了语义自适应思考模块，所以与传统的神经网络模型相比，在基线模型的基础上分别在精度上提升 6.36%，召回率上提升 3.97%，F_1 得分上提升 5.09%，均优于原论文中的模型。由于在基线模型的基础上引入了语义自适应思考模块，所以相比之下会增加一部分额外的计算和参数量，但能有效地提升模型提取伪平行句对的性能。

3.3.2　模型提取到的汉-越伪平行句子

本文的最终目的是从可比语料库中提取高质量的汉语到越南语的伪平行句子，从而提升汉-越神经机器翻译的性能。表 4 显示了模型提取到的汉-越伪平行句子对。

表 4　提取到的汉-越伪平行句子示例

汉语	越南语
农业部门制定了详细的计划和措施，建立山区农业促进系统。	Ngành nông nghiệp đã xây dựng chương trình cụ thể và những biện pháp xây dựng hệ thống khuyến nông ở miền núi.
越南企业将不得不面对国内市场和国外市场的竞争。	Các doanh nghiệp Việt Nam sẽ phải đương đầu cạnh tranh cả trên thị trường trong và ngoài nước.
中国粮食产量大幅增长的主要原因是，与去年相比，种植面积增加了 66.6 万公顷。	Yếu tố chính khiến sản lượng lương thực của Trung Quốc tăng mạnh là do diện tích gieo trồng tăng 666.000 ha so với năm trước.

可以看出，从可比语料库中提取到的汉-越伪平行句子意思更加相近，在语义上也更相关。由于模型中的思考模块能够更深层次地挖掘句子中的语义信息，最大化表示句子之间的相似性，所以提取到的句子质量更高。

3.3.3　单词思考次数分析

由于不同单词在句子中的语义表征的难易程度不一样，导致句子的语义信息不充分，所以需要进行多次思考，充分提取单词的上下文语义信息。图 3 展示了表 4 中第 2 个示例句子中每个单词的不同思考次数(最大思考次数为 3)。

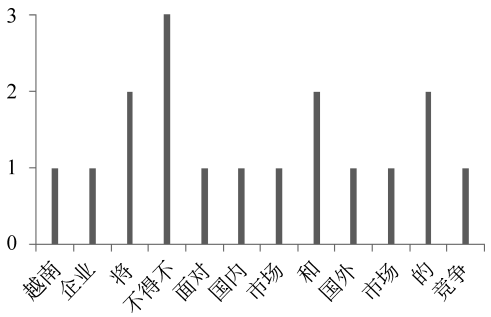

图 3　不同单词的思考次数

从图中可以看出，单词“越南”“企业”等在句子中比较容易进行上下文语义表征，所以只需进行一次思考；而像单词“不得不”在句子中对其进行上下文语义表征就比较困难，需要进行多次思考，深度挖掘语义信息，才能达到规定的阈值。

3.3.4　机器翻译比较

(1) 数据集

为了验证模型提取到的句子对机器翻译性能的影响，本文将提取到的句子对作为训练数据集，其中，训练集大小为 20 万对汉语-越南语句子，测试集和验证集大小均为 2 000。每个句子的最大长度设置为 80，我们使用 BLEU 值进行评分。

(2) 机器翻译系统评估

本文选择了目前比较主流的神经网络模型 Seq2Seq+Attention[14] 作为机器翻译模型，分别将基线模型和本文模型提取到的句子对作为训练数据集，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 个周期，其中，句子对的批量大小为 64。同时根据概率分数按降序对系统提取的句子对进行排序，分别以{10 万，15 万，20 万}不同数量的句子对进行训练。表 5 为基线模型和本文模型方法提取到的汉-越伪平行句子训练的机器翻译系统获得的 BLEU 分数。

表 5　本文模型和基线模型提取到的不同数量的汉-越伪平行句对在神经机器翻译模型上获得的 BLEU 分数

# Pairs	Model	BLEU
		Seq2Seq+Attention
10 万	Baseline	15.23
	BiLSTM+Pondering	15.57(+0.34)

续表

#Pairs	Model	BLEU
		Seq2Seq+Attention
15 万	Baseline	15.68
	BiLSTM+Pondering	16.30(+0.62)
20 万	Baseline	15.94
	BiLSTM+Pondering	17.92(+1.98)

注：Pairs 是训练集中句子对的数量。

实验结果表明，在训练集中添加系统提取到的句子对子集会带来显著的收益，分别通过 BiLSTM+Pondering 模型和基线模型提取到的句子对数量为 10 万时，翻译系统的 BLEU 得分分别为 15.57 和 15.23，提高了 0.34，优于基线模型；而当两个模型提取的句子对数量增加到 20 万时，翻译系统的 BLEU 得分分别为 17.92 和 15.94，比基线模型提高了 1.98。这些结果证实，本文模型提取到的句子对的质量较高，表明可比语料库中存在语义高度相似的伪平行句子。同时可以降低决策阈值 ρ，以便提取更大尺寸的语料库。

4 总结与未来工作

本文提出了一种基于语义自适应编码的双向循环神经网络模型，从可比语料库中提取汉-越伪平行句子，构建新的平行语料库。该模型是在神经网络模型的基础上增加语义自适应编码模块，能够根据阈值信息进一步选择思考的深度，同时深度挖掘句子的语义信息，尽可能地比较文本集合中句子的相似性，提高伪平行句子的质量，然后改善低资源下的神经机器翻译性能。实验结果表明，本文的方法优于基线模型，并且提取到的汉-越平行句子语义更加相近，噪声更小。

在这项工作中，该文主要针对汉语到越南语低资源语言来进行的，我们发现从相似文本集合中提取平行句子扩充语料库是比较有效的途径之一。由于可用于训练的汉语到越南语的数据目前相对比较少，因此在未来的工作中，我们会进一步增加训练集的数据，提高模型的性能，同时将会从多模态的角度去考虑，即将图像信息集成到文本中，使得模型能同时关注文本和图像，提取更高质量的平行数据，改善低资源下机器翻译的性能。

参考文献

[1] Benjamin M, Atsushi F. Efficient extraction of psedo-parallel sentence from raw monolingual data using word embedding[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017: 392-398.

[2] Minh Thang Luong, Hieu Pham, Christopher D. Manning Bilingual word representations with monolingual quality in mind[C]//Proceedings of NAACL Workshop on Vector Space Modeling for NLP, Denver, United States, 2015: 151-159.

[3] Sanjika Hewavitharana, Stephan Vogel. Extracting parallel phrases from comparable data[C]//Proceedings of the 4th Workshop on Building and Using Comparable Corpora, 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, Association for Computational Linguistics, 2011: 61-68.

[4] Xiang Zhang, Shizhu He, Kang Liu, et al. AdaNSP: Uncertainty-driven adaptive decoding in neural semantic parsing [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4265-4270.

[5] Rauf S A, Schwenk H. Parallel sentence generation from comparable corpora for improved SMT[J]. Machine Translation, 2011, 25(4): 341-375.

[6] Sadaf Abdul Rauf, Holger Schwenk. On the use of comparable corpora to improve SMT performance [C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2009: 16-23.

[7] Alberto Barron Cedeno, Cristina Espana Bonet, Josu Boldoba. A factory of comparable corpora from Wikipedia[C]//Proceedings of the 8th Workshop on Building and Using Comparable Corpora, 2015: 3-13.

[8] Chenhui Chu, Raj Dabre, Sadao Kurohashi. Parallel sentence extraction from comparable corpora with neural network features[C]//Proceedings of the 10th International Conference on Language Resources and Evaluation, Paris, France, European Language Resources Association, 2016.

[9] Francis Gregoire, Philippe Langlais. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine tranlation[J]. arXiv preprint arXiv: 1806. 05559v2, 2018.

[10] Cristina Espana Bonet, Adam Csaba Varga, et al. An empirical analysis of NMT-derived interlingual embeddings and their use in parallel sentence identification[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1340-1350.

[11] Juryong Cheon, Member, nonmember: Automatically extracting parallel sentences from wikipedia using sequential matching of language resources[J]. IEICE Transactions on Information and Sytems E100, 2017: 405-408.

[12] Resnik P, Smith N A. The web as a parallel corpus [J]. Computational Linguistics, 1994, 29(3) 349-380.

[13] Talvensaari T. Effects of aligned corpus quality and size in corpus-based CLIR[C]//Proceedings of the 30th Europear Conforerce on Advances in Information Retrieval, 2008: 114-125.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762v 56, 2017.

[15] Alexis Conneau, Guillaume Lample, et al. Word translation without parallel data[J]. arXiv preprint arXiv: 1710.04087, 2018.

[16] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, et al. Massively multilingual word embeddings[J]. arXiv preprint arXiv: 1602.01925v2, 2016.

[17] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, et al. Loss in translation: Learning bilingual word mapping with a retrieval criterion[J]. arXiv preprint arXiv: 1804.07745v3, 2018.

[18] Guillaume Lample, Myle Ott, Alexis Conneau, et al. Phrase based and neural unsupervised machine translation[J]. arXiv preprint arXiv: 1804.07755v2, 2018.

[19] Tomas Mikolov, Quoc V Le, Ilya Sutskever. Exploiting similarities among languages for machine translation[J]. arXiv preprint arXiv: 1309.4168v1, 2013.

[20] Po Yao Huang, Frederick Liu, Sz Rung Shiang, et al. Attention based multimodal neural machine translation [C]//Proceedings of the 5th Confererce on Machire Translation ACL. Berlin, Germany, 2016: 639-645.

[21] Hideki Nakayama, Noriki Nishida. Zero-resource mchinne translation by multimodal encoder-decoder nework with multimedia piot[J]. arXiv preprint arXiv: 1611.04503v1, 2016.

[22] Ozan Caglayan, Fethi Bougares. Multimodal attention for neural machine translation[J]. arXiv preprint arXiv: 1609.03976 v1, 2016.

[23] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, et al. Unsupervised machin e translation using monolingual corpra only[J]. arXiv preprint arXiv: 1711.00043v1, 2017.

郭军军(1987—)，博士，硕士生导师，主要研究领域为自然语言处理、机器翻译、机器学习。
E-mail: guojjgb@163.com

余正涛(1970—)，通信作者，教授，博士，博士生导师，主要研究领域为自然语言处理、信息检索、机器翻译、机器学习。
E-mail: ztyu@hotmail.com

田应飞(1997—)，硕士研究生，主要研究领域为自然语言处理、机器翻译。
E-mail: 2467312988@qq.com