CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 44 卷第 10 期 2022 年 10 月
Vol. 44, No. 10, Oct. 2022

文章编号:1007-130X(2022)10-1861-08

基于回译和比例抽取孪生网络筛选的汉越平行语料扩充方法*

王可超[1],郭军军[1,2],张亚飞[1,2],高盛祥[1,2],余正涛[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:回译作为翻译中重要的数据增强方法,受到了越来越多研究者的关注。其基本思想为首先基于平行语料训练基础翻译模型,然后利用模型将单语语料翻译为目标语言,组合为新语料用于模型训练。然而在汉-越低资源场景下,训练得到的基础翻译模型性能较差,导致在其上应用回译方法得到的平行语料中含有较多噪声,较难用于下游任务。针对此问题,构建基于比例抽取的孪生网络筛选模型,通过训练使得模型可以识别平行句对和伪平行句对,在同一语义空间上对回译得到的伪平行语料进行筛选去噪,进而得到更优的平行语料。在汉越数据集上的实验结果表明,所提方法训练的模型的性能相较基线模型有显著提升。

关键词:汉越平行语料扩充;回译;数据增强;比例抽取;孪生网络

中图分类号:H085　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2022.10.018

A Chinese-Vietnamese parallel corpus expansion method based on back translation and proportional extraction siamese network screening

WANG Ke-chao[1], GUO Jun-jun[1,2], ZHANG Ya-fei[1,2], GAO Sheng-xiang[1,2], YU Zheng-tao[1,2]

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

Abstract: As an important data enhancement method in translation, back translation has attracted more and more researchers' attentions. The basic idea is to first train a basic translation model based on parallel corpus, then use the model to translate monolingual corpus into the target language, and combine it into a new corpus for model training. However, in the Chinese-Vietnamese low-resource scenario, the performance of the basic translation model obtained by training is poor, which results in the parallel corpus obtained by applying the back translation method on it contains more noise and is difficult to use for downstream tasks. In response to this problem, a siamese network screening model based on proportional extraction is constructed. Through training, the model can identify parallel sentence pairs and pseudo-parallel sentence pairs, and filter and denoise the pseudo-parallel corpus obtained by back translation in the same semantic space, thereby obtaining a better parallel corpus. The test results on the Chinese-Vietnamese data set show that the proposed method significantly outperforms the baseline system.

Key words: Chinese-Vietnamese parallel corpus expansion; back translation; data enhancement; pro-

* **收稿日期**:2020-12-07;**修回日期**:2021-02-23
基金项目:国家自然科学基金(61732005,61761026,61866020,61672271,61762056,61972186);国家重点研发计划(2019QY1801,2019QY1802,2019QY1800)
通信作者:高盛祥(gaoshengxiang.yn@foxmail.com)
通信地址:650500 云南省昆明市昆明理工大学信息工程与自动化学院
Address: Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, P. R. China

portional extraction;siamese network

1 引言

神经机器翻译 NMT(Neural Machine Translation)[1,2]是自然语言处理领域的研究热点。相较于已经取得极大进步的资源丰富型神经机器翻译,低资源神经机器翻译由于缺少高质量的双语语料,效果并不理想。因此,如何高效地扩充语料规模,成为低资源神经机器翻译研究中亟需解决的问题。针对此问题,研究人员提出了多种数据增强方法,通过有限的语料资源扩充双语语料规模。早期的工作主要利用人工的方式进行语料扩充,但效率较低。近年来,随着深度学习技术的发展,利用深度学习方法来扩充双语语料成为有效途径。基于深度学习的数据增强方法主要分为生成式和抽取式。生成式数据增强方法包括:回译(back-translation),将目标端的单语语料通过反向翻译模型扩充为伪平行双语语料;词或单元的替换,通过各种手段替换句子中部分单元(词或短语)来扩充语料;加入枢轴语言,充分利用源-枢轴-目标语言间丰富的对齐语料来提升源-目标语言对的机器翻译性能。抽取式数据增强方法主要通过计算跨语言语义相似度,从可比语料(篇章对齐)中抽出伪平行语料。通过这几种方法,可以大规模扩充双语语料库的规模。

汉语-越南语作为典型的低资源语言对,其平行语料获取难度很大。传统的回译方法中,首先基于小规模平行语料训练基础翻译模型,在此基础上将越南语翻译为对应的汉语句子、组合成新语料再次投入训练。但是,由于用于训练基础翻译模型的平行语料规模和质量欠佳,造成基础翻译模型训练并不充分,若只是直接在该模型上通过回译方法进行语料扩充,得到的伪平行语料会含有过多的噪声,如表 1 所示。

Table 1 Comparison between back-translation generated translation and standard translation

表 1 回译生成译文与标准译文对比

原文	Ba người chúng tôi leo vào trong và rồi con tàu tách mình ra khỏi trạm không gian để rơi vào bầu khí quyển.
回译译文	我们 三个 人 爬 到 地上 , 然后 我们 离开 了 我们 的 飞船 , 从 太空 到 大气 中 , 进入 大气 中 。
标准译文	我们 三个 人 爬 进 舱内 , 然后 飞船 把 我们 带 出 空间站 坠入 大气层 。

表 1 中通过回译得到的汉语译文偏离了原句的意思,且有明显的语义逻辑错误,若要构建用于训练机器翻译模型的双语语料库,必须要过滤掉这种句对。本文将回译和伪平行句对抽取的方法相结合,通过计算跨语言句对间的语义相似度,对生成的语料进行筛选,以获得高质量双语语料。具体来说,本文首先利用回译的方法,将大规模的单语语料扩充为伪平行语料;然后结合回译数据的特点,对传统基于双向长短时记忆 Bi-LSTM(Bidirectional Long Short-Term Memory)孪生网络的句对抽取模型进行了改进,改进后的模型将平行语料和伪平行语料混合后对模型进行训练,使模型能更好地分辨平行句与伪平行句,从而抽取出质量更高的伪平行句,以构建用于汉越神经机器翻译的语料库。

2 相关工作

神经机器翻译是目前机器翻译领域内最热门的研究方法,在资源充足的语言对翻译上,神经机器翻译的性能已经明显超过了统计机器翻译[3],但在低资源神经机器翻译上,神经机器翻译的效果还有待提升[4]。用来训练低资源神经机器翻译模型的平行语料相对较少,导致翻译效果欠佳,因此如何获取高质量的双语语料,成为提高低资源神经机器翻译的一种关键性技术。近年来,国内外相关研究人员针对低资源语种的伪平行语料扩充方法进行了广泛研究,并取得了一系列成果。

目前应用最广泛的语料扩充方法是回译。它利用反向的翻译模型,将目标端语言的数据翻译成源端语言的数据,通过这一方法来构造伪平行双语数据来训练正向翻译模型。回译最早是由 Sennrich[5]等提出的,文中提出了 2 种方式来比较回译的性能。第 1 种方法在只有目标语言句子 y 的前提下,将源语言对应的句子设置为空,将句对 $(dummy,y)$ 将其加入到平行语料中进行训练,可以看成是翻译模型和语言模型多任务训练;第 2 种方式为回译,用训练好的目标语言到源语言的翻译模型翻译目标语言句子 y,得到伪平行句对 (x',y),将其加入到平行句对中一起训练。因为 y 是高质量的单语语句,而 x' 中可能包含一些〈UNK〉字符或者错误的句法等,其质量较差。这样训练可以想象成去噪声形式的训练。在有噪声的情况下,

训练 x(源语言)→y(目标语言)方向的翻译模型尽量还能翻译好,以此提升泛化性能。回译已经有了越来越多的扩展方法。He 等[6]提出了对偶学习的方法,将回译扩展为在 2 个翻译方向上训练 NMT 系统,利用源语言与目标语言的单语数据来同时提升 2 个方向的翻译模型;Hoang 等[7]提出了迭代回译的思想,通过使用回译的数据构建更好的翻译模型,再使用这个更好的翻译模型对数据进行回译,重复此过程以达到迭代的效果。数据增强的方法还有词或单元的替换。比如 2017 年 Fadaee 等[8]提出了一种增强语料的方法,首先在规模较大的单语语料上训练出语言模型,然后用语言模型找到句子中可以被低频词替换的高频词的位置并完成替换。通过这种单词替换,增加了训练语料中低频词出现的次数,从而增强神经机器翻译对低频词的理解能力。而蔡子龙等[9]将句子中最相似的单元进行位置上的对调,以此形成新的语料,改变的是语料中句子的结构信息而非语料中的词频信息。此外,Wei 等[10]提出了随机替换、随机插入、随机交换和随机删除的方法,为低资源神经机器翻译的数据增强技术开拓了新的思路,也提升了低资源 NMT 的性能。还有一种增强方法是加入枢轴语言。此类方法通过引入大语种丰富的对齐语料作为枢轴语言来充分提升小语种神经机器翻译的性能。Ren 等[11]提出,在大语种之间的翻译过程中将小语种作为中间隐变量引入,将该翻译过程拆分为两个经由小语种的翻译过程,如 X、Y 为两个大语种,它们之间有大量双语数据,Z 作为小语种,它和 X、Y 之间均只有少量双语数据,为了提升 X→Z 和 Y→Z 的翻译性能,可以用此方法来进行优化。

在抽取式语料扩充方法的研究中,Cristina 等[12]研究了从 NMT 系统编码器获得的句子表示中检测新的平行句对,通过比较余弦相似度来进行平行句和非平行句的区分。Grover 等[13]提出了一种利用连续向量表示的方法,在使用 Luong 等[14]提出的双语词嵌入模型学习单词表示后,再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。而 Grégoire 等[15]使用单一端到端模型估计可比语料中 2 个句子平行的条件概率分布,取得了更好的效果。

对汉越语言对来说,回译能够快速而有效地扩充汉越平行语料规模,然而,单独使用回译方法生成的伪平行语料质量较差,在实际应用中难以用于下游任务,若直接用于训练翻译模型,可能会降低翻译系统的性能[16]。针对此问题,本文结合回译和平行句对抽取方法对数据进行扩充和清洗。之前工作中,由于大多数句对抽取方法是针对可比语料特点进行训练的,所以本文在此基础上结合回译数据的特点对句对抽取方法进行了改进,使其可以对伪平行语料进行更有效的筛选。本文方法将伪平行语料与平行语料进行混合,用于训练句对抽取模型,以提升模型抽取出的平行句对的比例,使其能够分辨出平行句对与伪平行句对,进而从回译生成的伪平行语料中筛选出高质量的伪平行句对。

3 基于回译和比例抽取孪生网络筛选的伪平行句对抽取方法

3.1 整体框架

本文方法首先利用回译的基本思想,将大规模的越南语单语数据利用基础翻译模型翻译得到汉越伪平行双语数据。但是,由于汉越平行语料规模有限,训练得到的基础翻译模型(翻译方向:越→汉)性能一般,进而导致扩充的伪平行语料中部分句对质量不佳,无法更有效地推进后续工作。本文通过混合小规模平行语料和回译生成的大规模伪平行语料,训练一个基于比例抽取的 Bi-LSTM 孪生网络,使得该网络可以识别出混合语料中的平行句对。该句对抽取模型通过孪生网络将汉越句对映射到同一语义空间下,计算句对之间的语义相似度,并按相似度得分从高到低排列句对,取出相似度高于设定阈值的句对。在训练过程中,将平行句对和伪平行句对混合,并加标签区分,通过最大化抽取出的平行句对与抽取前平行句对的比值来训练模型,使得模型经过训练后,可以精确地识别原始平行句对。具体而言,抽取的句对结果中,平行句对优先排序,紧接其后的为最接近平行句对的伪平行句对,最后为质量较差的伪平行句对。因此模型在具有识别原始平行句对能力的同时,也能从混合语料中抽取出高质量的伪平行句对,以达到对伪平行数据进行筛选的目的。整体的框架如图 1 所示。其中,D'_1 指抽取出的原始平行句对,$count(D'_1)$ 表示抽取出的原始行句对的数量;$count(D_1)$表示总的原始平行句对的数量。

3.2 基于回译的伪平行句对生成

回译首先需要一部分高质量的平行语料来训练基础翻译模型,然后用这个翻译模型将大规模的目标端单语数据回译生成伪平行数据。在本文方

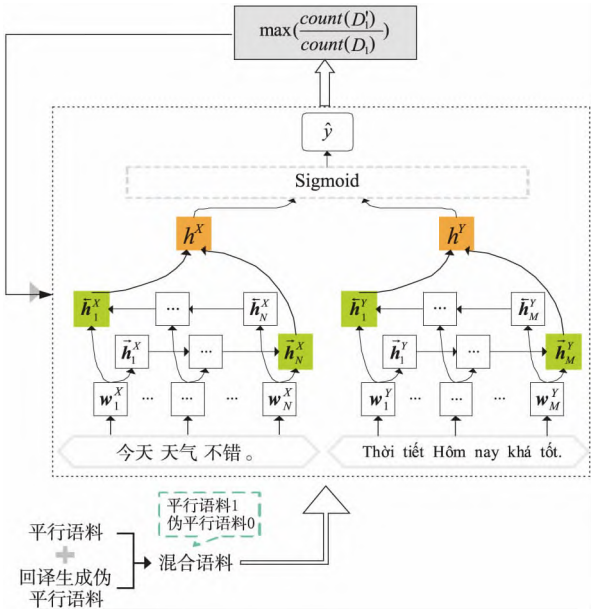

Figure 1 Overall framework

图 1 整体框架图

法中,设置汉语为源语言 X,越南语为目标语言 Y。首先利用爬虫爬取到一定规模的汉越平行句对,并对这些句对进行预处理,人工筛选掉质量不高的句对,使其符合本文训练的要求。将预处理后的汉越平行语料 $D_1=\{x^{(t)},y^{(t)}\}_{t=1}^{T}$ 用来训练越→汉基础翻译模型 $M_{Y\to X}$,然后将大规模的越南语单语语料 $Y=\{y'^{(m)}\}_{m=1}^{M}$ 通过 $M_{Y\to X}$ 回译生成汉语句子 $X'=\{x'^{(m)}\}_{m=1}^{M}$,构建为汉越伪平行语料库 $\widetilde{D}=\{x'^{(m)},y'^{(m)}\}_{m=1}^{M}$。

3.3 基于比例抽取的 Bi-LSTM 孪生网络平行句对抽取方法

基于回译的方法将大规模的目标端单语数据扩充为伪平行数据后,还需要进行数据筛选的工作。本文使用一个基于比例抽取的 Bi-LSTM 孪生网络来实现数据筛选任务。

Bi-LSTM 通过学习句对之间的跨语言语义来估计它们互为翻译的可能性。该句子抽取模型使用共享权值的孪生网络[17],利用双向 LSTM[18,19]句子编码器将句子在共享向量空间中进行连续的向量表示,然后源句和目标句的表示被输入到一个带 Sigmoid 输出层的前馈神经网络中,计算它们为平行句对的条件概率,将相似度高于设定阈值的句对抽取出来。

3.3.1 语句编码

以源端句子的 Bi-LSTM 的编码为例(目标句子的编码用 Y 替换 X)。在每个 t 时刻,词表 V^X 中的整数索引 k 定义的词表示为 one-hot 向量 $\boldsymbol{w}_k^X\in\{0,1\}^{|V^X|}$。该向量与词嵌入矩阵 $\boldsymbol{E}^X\in\mathbf{R}^{|V^X|\times d_e}$ 相乘,获得该词的连续向量表示 $\boldsymbol{w}_t^X\in\mathbf{R}^{d_e}$,如式(1)所示:

$$\boldsymbol{w}_t^X=\boldsymbol{E}^{X^{\mathrm{T}}}\boldsymbol{w}_k^X \tag{1}$$

式(1)得到的表示用作前向和后向 Bi-LSTM 编码器的输入,即 $\overrightarrow{\boldsymbol{h}}_t^X$ 和 $\overleftarrow{\boldsymbol{h}}_t^X$。前向 LSTM 读取变长的句子,并从第一个词到最后一个词更新其递归状态,从而创建一个固定大小的句子连续向量表示 $\overrightarrow{\boldsymbol{h}}_N^X\in\mathbf{R}^{d_h}$,同样,后向 LSTM 反向处理该句子。具体如式(2)和式(3)所示:

$$\overrightarrow{\boldsymbol{h}}_t^X=\phi(\overrightarrow{\boldsymbol{h}}_{t-1}^X,\boldsymbol{w}_t^X) \tag{2}$$

$$\overleftarrow{\boldsymbol{h}}_t^X=\phi(\overleftarrow{\boldsymbol{h}}_{t+1}^X,\boldsymbol{w}_t^X) \tag{3}$$

其中,$\phi(\cdot)$代表 LSTM。使用 2 个方向上的最后一个递归状态的串联作为第 i 个句子的最终表示,如式(4)所示:

$$\boldsymbol{h}_i^X=[\overrightarrow{\boldsymbol{h}}_{i,N}^X,\overleftarrow{\boldsymbol{h}}_{i,1}^X] \tag{4}$$

3.3.2 句对信息匹配

对源语句和目标语句进行编码之后,通过使用它们的元素乘积和元素差异的绝对值来量化源语句和目标语句之间的匹配信息,得到匹配向量,如式(5)和式(6)所示:

$$\boldsymbol{h}_i^{(1)}=\boldsymbol{h}_i^X\odot\boldsymbol{h}_i^Y \tag{5}$$

$$\boldsymbol{h}_i^{(2)}=|\boldsymbol{h}_i^X-\boldsymbol{h}_i^Y| \tag{6}$$

通过将匹配向量馈送到具有 Sigmoid 输出层的前馈神经网络来估计句子平行的条件概率,如式(7)和式(8)所示:

$$\boldsymbol{h}_i=\tanh(\boldsymbol{W}^{(1)}\boldsymbol{h}_i^{(1)}+\boldsymbol{W}^{(2)}\boldsymbol{h}_i^{(2)}+\boldsymbol{b}_1) \tag{7}$$

$$p(y_i=1\mid\boldsymbol{h}_i)=\sigma(\boldsymbol{v}\boldsymbol{h}_i+\boldsymbol{b}) \tag{8}$$

其中,$\sigma(\cdot)$是 Sigmoid 函数,$\boldsymbol{W}^{(1)}\in\mathbf{R}^{d_f\times d_h}$,$\boldsymbol{W}^{(2)}\in\mathbf{R}^{d_f\times d_h}$,$\boldsymbol{v}\in\mathbf{R}^{d_f}$,$\boldsymbol{b}_1\in\mathbf{R}^{d_f}$,$\boldsymbol{b}$ 是模型参数,$p(y_i=1)$表示第 i 个句对平行的概率,$\boldsymbol{d}_f$ 是前馈神经网络隐藏层的大小。通过最小化句对的交叉熵损失来训练模型,如式(9)所示:

$$L_1=\sum_{i=1}^{n(1+m)}(y_i\log\sigma(\boldsymbol{v}\boldsymbol{h}_i+\boldsymbol{b})+(1-y_i)\log(1-\sigma(\boldsymbol{v}\boldsymbol{h}_i+\boldsymbol{b}))) \tag{9}$$

如果句对的概率大于或等于决策阈值 ρ,则将其分类为平行,否则为不平行,如式(10)所示:

$$\hat{y}_i=\begin{cases}1, p(y_i=1\mid\boldsymbol{h}_i)\geqslant\rho\\ 0, p(y_i=1\mid\boldsymbol{h}_i)<\rho\end{cases} \tag{10}$$

其中 n 和 m 分别表示源泉语句和目标语句的个数。

将句子平行的条件概率作为句对之间的相似度,然后对该相似度进行从高到低排列,抽取出大于设定阈值的句对,用于训练一个能抽取出较高质量伪平行句对的句对抽取模型。

3.3.3 基于比例的损失函数改进

传统基于 Bi-LSTM 孪生网络筛选伪平行句对的方法是在可比语料上实现的,而本文是对回译生成的大规模伪平行语料进行筛选,所以本文方法在结合回译语料的基础上,对传统基于 Bi-LSTM 孪生网络方法做了一定的改进。

在模型训练阶段,本文方法不再用平行语料和随机生成负例来训练模型,而是将平行句对与伪平行句对按比例混合来训练模型,目的是使模型更好地识别出原始平行句对,在抽取过程中尽可能多地将原始平行句对抽取出来,如式(11)所示:

$$\theta=\arg\max\left(\frac{count(D'_1)}{count(D_1)}\right) \tag{11}$$

通过最大化 $count(D'_1)$ 和 $count(D_1)$ 的比例,使得训练后的模型可以从混合语料中精准地识别并抽取出原始平行句对。

为了使平行句对抽取比例对模型产生积极的影响,本文定义了另外一个损失函数,如式(12)所示:

$$L_2=1-\frac{count(D'_1)}{count(D_1)} \tag{12}$$

最终的损失函数由 L_1 和 L_2 共同决定,如式(13)所示:

$$L=\lambda L_1+(1-\lambda)L_2 \tag{13}$$

其中,λ 是超参数,通过人工设定,用于调节 L_1 和 L_2 的权重。

3.3.4 语料设置

训练开始之前,将汉越平行语料 D_1 与回译生成的伪平行语料 $\widetilde{D}$ 混合得到 $D^*=D_1\cup\widetilde{D}$,并加标签予以区分,在平行句对后加标签"1",伪平行句对后加标签"0"。加标签后的句对表示形式如表 2 所示。

将混合语料输入到基于比例抽取的 Bi-LSTM 句对抽取模型中,训练句对抽取模型,使模型能精准地分辨出平行句对和伪平行句对。

Table 2 Representation of sentence pairs after being labeled and mixed

表 2 加标签混合后句对的表示形式

句对	标签
什么 会 让 你 紧张 ?	1
cái gì mới thật sự làm bạn phải lo ngại ?	1
他们 都 是 精神病 医生 。	0
Họ đều là bác sỹ tâm lý.	0
人类 平等 的 革命 是 可能 发生 的 。	1
Cuộc cách mạng về quyền bình đẳng của con người có thể xảy ra.	1
明天 我们 要 去 冬天 吗 ?	0
Chúng ta sẽ đi vào rạng đông ngày mai nhé?	0

4 实验与分析

4.1 实验模型设置

翻译模型:为了验证本文方法的有效性,首先基于Transformer翻译模型进行了在汉-越任务上的训练,作为 baseline 翻译模型。在语料方面,通过网络爬虫工具爬取汉越双语语料,并经过初步的筛选,删掉标点符号过多或无效字符的句子,并删掉越南语中短于 5 个词和长于 50 个词的句子及其对应的汉语句子(因为句对过短或过长对于模型训练的收益不大);然后使用 jieba 分词工具对汉语句子进行分词,经过人工的精准校对和筛选,得到了 200 000 平行句对。从中分别随机抽取出 2 000 个句对作为 baseline 的验证集和测试集,剩余的作为训练集,初始的实验数据具体如表 3 所示。

Table 3 Experimental data of baseline model

表 3 baseline 模型实验数据

	训练集	验证集	测试集
汉-越	196 000	2 000	2 000

本文使用清华大学的开源 Transformer 翻译模型 THUMT,在参数设置上,将 batch size 设置为 512,train step 设置为 50 000,汉语词表大小为 41 000,越南语词表大小为 32 000,训练过程中每 2 个周期更新一次模型的参数,每训练 2 000 步,对模型进行一次评估,最后保存评估得分最高的 3 个中间模型,使用 BLEU(本文统一使用 *BLEU*4)作为评测指标。在汉→越和越→汉的 2 个翻译方向上分别对模型进行了训练,实验结果如表 4 所示。

Table 4 Experimental results of the baseline model

表 4 baseline 模型实验结果

	汉→越	越→汉
*BLEU*4	20.62	20.51

通过网络爬取大规模的越南语单语数据,并像之前设置一样删掉过短或过长的句子,选取其中的 600 000 单语句子。将训练的越→汉的基础翻译模型用于回译,将目标端越南语单语句子回译生成源端汉语句子,最终构成规模为 600 000 的伪平行语料库。

句对抽取模型:对之前初步校对过的 200 000

平行句对进行人工筛选,选出其中质量较高的50 000,从伪平行数据中选取200 000,将2部分混合作为句对抽取模型的训练集。从平行数据的剩余部分中分别抽取1 000个句对作为验证集和测试集。该实验数据中,汉语词表大小为50 000,越南语词表大小为35 000。

为了评估所训练模型的性能,本文使用精度P(Precision)、R召回率(Recall)和$F1$值作为评价指标。精度是指所有抽取出的句对中真实平行句对的比例,召回率是指被抽取出的真实平行句对占测试集中所有平行句对的比例,而$F1$值是精度和召回率的调和平均值。

Bi-LSTM中词嵌入层的维度设为512,前馈神经网络中的隐藏层有256个隐藏单元,训练过程中的学习率设置为0.000 2,训练5个epoch,train step为36 000,抽取的阈值设为0.98,λ设为0.7。模型的训练结果如表5所示。

Table 5 Training results of the proposed model

表5 本文模型训练结果 %

指标	P	R	$F1$
值	81.45	65.97	72.90

4.2 实验结果分析

为了验证本文所提方法的有效性,本文主要通过实验研究了伪平行语料和基于比例抽取的句对抽取方法对汉越神经机器翻译的影响。首先,对比了仅使用汉越平行语料$M_{X\to Y}$、汉越平行语料加回译生成伪平行语料$M^{*}_{1\,X\to Y}$和$M^{*}_{2\,X\to Y}$(400 000和135 000)、汉越平行语料加抽取(*Grégoire*等[15]提出基于孪生网络的传统抽取方法$M^{**}_{1\,X\to Y}$和基于比例抽取方法$M^{**}_{2\,X\to Y}$)后的伪平行语料这几种不同语料库训练的汉越神经机器翻译模型的实验,其中传统抽取模型和本文方法的抽取模型的阈值都设为0.99,结果如表6所示。

Table 6 Experimental results of different methods on different datasets

表6 不同方法在不同数据集上的实验结果

句对抽取模型	语料规模	词表大小		$BLEU4$
		汉语	越南语	
baseline $M_{X\to Y}$	150 000	41 000	32 000	20.62
$M^{*}_{1\,X\to Y}$(+回译)	150+400 000	62 000	60 000	19.84
$M^{*}_{2\,X\to Y}$(+回译)	150+135 000	52 000	55 000	20.35
$M^{**}_{1\,X\to Y}$(+回译+传统抽取)	150+top135 000	55 000	56 000	21.23
$M^{**}_{2\,X\to Y}$(+回译+比例抽取)	150+top135 000	55 000	57 000	21.76

通过上述实验发现,仅通过将伪平行语料与平行语料混合来直接训练翻译模型,不但没有提高模型的性能,反而会降低$BLEU4$值。这是由于用来训练回译基础模型$M_{Y\to X}$的汉越平行语料规模不足,导致用基于伪平行语料来直接训练正向的汉越翻译模型$M_{X\to Y}$时反而会引入更多的噪声,从而降低翻译模型的$BLEU4$值得分。通过基于传统的Bi-LSTM孪生网络方法对伪平行句对进行抽取后,可以有效筛选掉平行程度较低或含有过多噪声的句对,对比传统的抽取方法,本文提出的基于比例抽取的方法对翻译模型性能有更明显的提升,$BLEU4$值增加了1.14。

4.3 验证实验

本节对基于比例抽取Bi-LSTM孪生网络方法有效性进行验证。实验中的平行语料为人工校对过的高度平行的50 000汉越平行语料,将回译生成的400 000伪平行语料与这部分平行语料混合,并用标签区分它们,在平行句对后加标签"1",伪平行句对后加标签"0"。通过加标签区分混合语料中的平行和伪平行句对,可以直观地看到模型抽取出的平行句对数和伪平行句对数。将这个混合的语料库作为句对抽取模型的输入语料,通过改变模型抽取句对时的阈值,可以得到不同规模的伪平行语料。具体的实验结果如图2所示。

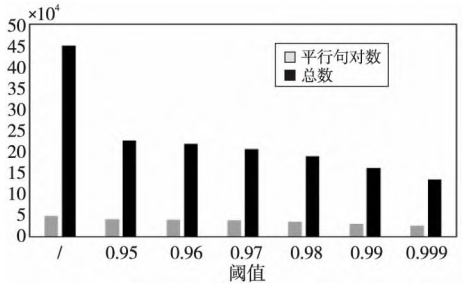

Figure 2 Sentence pair statistics under different thresholds

图2 不同阈值下的句对统计

由图2可知,当阈值设为0.95时,抽取出的混合语料的数量骤减到原来的一半,这说明伪平行语料中有大量含噪声的句对。当逐步提升阈值时,被抽取出的句对数量也随之减少,平行句对所占的比例也就越来越高,这也验证了本文模型的有效性。

为了继续验证抽取出的句对对神经机器翻译的影响,用上述通过不同阈值抽取出的句对分别对翻译模型进行训练,实验结果如图3所示。

通过对比不同阈值下抽取伪平行句对的结果可知,当句对抽取模型抽取出的原始平行数据占比

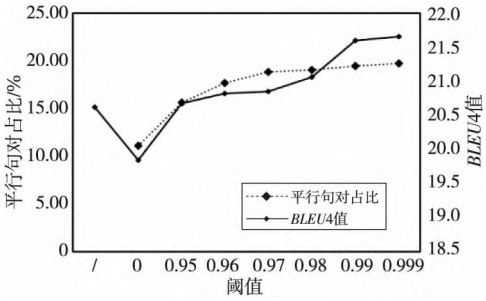

Figure 3 Relation of the proportion of parallel sentence pairs and the *BLEU*4 value

图 3 平行句对占比与 *BLEU*4 值的关系图

越高时,构成的混合语料库的质量越高,对神经机器翻译模型的提升越大。在阈值设置为 0.999 时,平行句对占比约为 20%,此时得到的 *BLEU*4 值最大为 21.76,相比只用平行语料训练的 baseline 提高了 1.14。

此外,为了探究训练数据是否加标签对本文方法的影响,分别用加标签和不加标签的训练语料进行了一组对比实验,实验结果如表 7 所示。

Table 7 Verification of label validity

表 7 验证标签的有效性

方法	*F*1 值	*BLEU*4 值
加标签混合语料的模型训练方法	72.90	21.76
无标签混合语料的模型训练方法	69.55	21.35

实验表明,训练数据中加入标签的方法有效地提升了句对提取模型的准确率,并且抽取出的语料对翻译模型的性能也有进一步的提升。

4.4 译文对比分析

为验证用基于回译和比例抽取孪生网络筛选方法构建语料库对神经机器翻译性能的影响,本文还用不同语料库训练的模型分别翻译同一语句进行对比分析,翻译结果如表 8 所示。

表 8 中 3 个用不同语料库训练的翻译模型分别翻译 2 个汉语句子,括号内为得到的越南语句子的中文参考翻译。通过对比分析可知,baseline 模型 $M_{X\to Y}$ 和模型 $M^{*}_{X\to Y}$(+回译)得到的句子在语义上与原句出现了较大的偏差,例如对原文 1 翻译的前半句,前 2 个译文都存在明显的语义不通,背离了语言逻辑。而本文方法训练的模型的翻译虽然还没有完全符合原意,但具有较高的逻辑性和流利度,较好地保留了原句的语义。

Table 8 Comparison of translations results generated by different models

表 8 不同模型生成译文的对比

原文 1	他 的 手 就 像 一 个 非常 老 的 人 的 手 , 但 他 仍然 是 一 个 孩子 。
baseline $M_{X\to Y}$	Anh ấy cầm tay như một người đàn ông rất cũ , và anh ấy vẫn là một đứa bé. (他像个老男人一样握着手，他仍然是个婴儿。)
$M^{*}_{X\to Y}$ (+回译)	Mắt của ông giống như đôi bàn tay của một người đàn ông rất già mà ông vẫn còn là một đứa trẻ.(他的眼睛就像是一个很小的老人的手，他还是个孩子。)
$M^{**}_{X\to Y}$ (+回译+比例抽取)	Bàn tay của anh ta giống như tay của một người rất già , và anh ta vẫn còn là một đứa trẻ.(他的手就像一个很老的男人，他还是个孩子。)
原文 2	他 做 了 一个 简单 的 计算 : 一个 小时 内 他 可以 伪造 30 份 假 文件 。
baseline $M_{X\to Y}$	Ông ta làm một phép tính đơn giản : ông ta có thể làm giả giấy tờ giả trong 30 tiếng đồng hồ .(他做了一个简单的计算：他可以在 30 小时内伪造伪造的文件。)
$M^{*}_{X\to Y}$ (+回译)	Ông làm một phép tính đơn giản : Trong vòng một giờ ông có thể làm một giấy tờ giả.(他做了一个简单的计算：一个小时之内他可以制作一张假纸。)
$M^{**}_{X\to Y}$ (+回译+比例抽取)	Ông làm một phép tính đơn giản : Trong một giờ ông có thể làm được 30 giấy tờ giả.(他做了一个简单的计算：一个小时内，他可以制作 30 张假纸。)

5 结束语

针对汉越神经机器翻译模型训练中平行语料不足的问题,本文提出了一种对语料进行扩充的方法。首先通过回译的方法,将越南语单语数据扩充为伪平行句对,利用基于比例抽取的 Bi-LSTM 孪生网络删除含有过多噪声的句对,同时抽取出相似度高的句对,用于构建汉越双语语料库。在句对抽取过程中,通过将平行句对混入伪平行句对中来指导抽取的过程。实验表明,基于此方法构建的语料库可以有效地提升汉越神经机器翻译的性能。在未来的工作中,我们会对翻译模型做更多的探索,以消除回译过程中产生的噪声,从而进一步提高汉越神经机器翻译的性能。

参考文献:

[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proc of International Conference on Neural Information Processing Systems, 2014: 3104-3112.

[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]//Proc of the International Conference on Learning Representations, 2015: 865-881.

[3] Liu Qun. Survey on statistical machine translation[J]. Journal of Chinese Information Processing, 2003, 17(4): 1-12. (in Chinese)

[4] Zoph B, Yuret D, May J, et al. Transfer learning for low-resource neural machine translation[C]//Proc of the Confe-

rence on Empirical Methods in Natural Language Processing, 2016:1568-1575.

[5] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data[C]//Proc of the 54th Annual Meeting of the Association for Computational Linguistics, 2016:86-96.

[6] He D, Xia Y C, Qin T, et al. Dual learning for machine translation[C]//Proc of International Conference on Neural Information Processing Systems, 2016:820-828.

[7] Hoang V C D, Koehn P, Haffari G, et al. Iterative back-translation for neural machine translation[C]//Proc of the 2nd Workshop on Neural Machine Translation and Generation, 2018:18-24.

[8] Fadaee M, Bisazza A, Monz C. Data augmentation for low-resource neural machine translation[C]//Proc of the 55th Annual Meeting of the Association for Computational Linguistics, 2017:567-573.

[9] Cai Zi-long, Yang Ming-ming, Xiong De-yi. Data augmentation for neural machine translation[J]. Journal of Chinese Information Processing, 2018, 32(7):30-36. (in Chinese)

[10] Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks[C]//Proc of the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, 2019:27-35.

[11] Ren S, Chen W H, Liu S J, et al. Triangular architecture for rare language translation[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics, 2018:56-65.

[12] Cristina E B, Ádám C V, Alberto B C, 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.

[13] Grover J, Mitra P. Bilingual word embeddings with bucketed CNN for parallel sentence extraction[C]//Proc of the 55th Annual Meeting of the Association for Computational Linguistics, 2017:11-16.

[14] Luong T, Pham H, Manning C D. Bilingual word representations with monolingual quality in mind[C]//Proc of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015:151-159.

[15] Grégoire F, Langlais P. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation[C]//Proc of the 27th International Conference on Computational Linguistics, 2018:1442-1453.

[16] Fadaee M, Monz C. Back-translation sampling by targeting difficult words in neural machine translation[C]//Proc of the Conference on Empirical Methods in Natural Language Processing, 2018:56-64.

[17] Chi Z M, Zhang B Y. A sentence similarity estimation method based on improved siamese network[J]. Journal of Intelligent Learning Systems and Applications, 2018, 10(4): 121-134.

[18] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11):2673-2681.

[19] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

附中文参考文献:

[3] 刘群. 统计机器翻译综述[J]. 中文信息学报, 2003, 17(4):1-12.

[9] 蔡子龙, 杨明明, 熊德意. 基于数据增强技术的神经机器翻译[J]. 中文信息学报, 2018, 32(7):30-36.

作者简介:

王可超(1995 -),男,山东潍坊人,硕士生,研究方向为机器翻译。**E-mail**:903695763@qq.com

WANG Ke-chao, born in 1995, MS candidate, his research interest includes machine translation.

郭军军(1987 -),男,山西吕梁人,博士,副教授,CCF 会员(A2270M),研究方向为自然语言处理、信息检索和机器翻译。**E-mail**:guojjgb@163.com

GUO Jun-jun, born in 1987, PhD, associate professor, CCF member(A2270M), his research interests include natural language processing, and information retrieval, and machine translation.

张亚飞(1981 -),女,河南洛阳人,博士,副教授,CCF 会员(38031M),研究方向为模式识别和自然语言处理。**E-mail**:zyfeimail@163.com

ZHANG Ya-fei, born in 1981, PhD, associate professor, CCF member(38031M), her research interests include pattern recognition, and natural language processing.

高盛祥(1977 -),女,云南洱源人,博士,副教授,CCF 会员(38040M),研究方向为自然语言处理、信息检索和机器翻译。**E-mail**:gaoshengxiang.yn@foxmail.com

GAO Sheng-xiang, born in 1977, PhD, associate professor, CCF member(38040M), her research interests include natural language processing, information retrieval, and machine translation.