基于枢轴的汉越联合训练神经机器翻译

高盛祥[1,2]，刘畅[1,2]，余正涛[1,2*]，黄继豪[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：越南语是一种典型的低资源语言，为了缓解汉越机器翻译面临资源稀缺的问题，提出一种基于枢轴的汉越联合训练神经机器翻译方法。首先利用小规模的汉越平行语料训练翻译模型得到汉语和越南语的词向量表征，再将以英语作为枢轴语言的汉语-英语，英语-越南语翻译模型进行联合训练。汉语-英语，英语-越南语翻译模型的汉语、越南语的向量表示与汉越模型得到的汉语、越南语的向量表示计算优化从而进行汉越联合训练。实验结果表明，本文方法将汉越平行语料与汉英、英越平行语料结合起来进行联合训练，充分利用了英语枢轴语料提升了汉越机器翻译性能。相比基线系统，基于枢轴的汉越联合训练神经机器翻译模型提升了 1.81 个 BLEU 值。

关键词：神经机器翻译；低资源；汉越；枢轴；联合训练；

中图分类号：TP 391　　文献标志码：A

机器翻译是用来进行大规模语言翻译的有效工具。近年来，中国与越南的交流与合作越来越密切，而机器翻译是跨语言信息交流较为有效的方式，因此研究汉越机器翻译有着非常重要的应用价值。神经机器翻译是 Sutskever 等人[1]在 2014 年提出的一种机器翻译方法,目前主流的神经机器翻译模型都采用编码器-解码器的架构，其中编码器把源语句编码成一个连续表示，解码器把这个连续表示解码成目标语。编码器和解码器之间一般通过注意力机制[2](Attention Mechanism)连接,通过不断训练神经网络从而得到源语言映射到目标语言的翻译模型。神经机器翻译在拥有大规模平行语料的语言对上已经取得了良好的翻译性能，在许多翻译任务中取得了令人瞩目的成绩，但 Zoph 等人[3]表明在低资源的场景下，神经机器翻译的翻译质量低于统计机器翻译（Statistical Machine Translation, SMT），传统的 SMT 研究中，Cohn 等人[4]提出利用枢轴语言作为桥接点，将源语言和目标语言相关联作为一种研究方法。目前，多数研究者在努力探索提升低资源神经机器翻译性能的方法，但在汉越这种低资源语

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）；云南高科技人才项目（201606）；云南省重大科技专项（201902D080019）；云南省基础研究计划（201901S070057，2018FB104）；昆明理工大学省级人培项目（KKSY201703005）

* **通信作者：**ztyu@hotmail.com

言对上，它受到汉越平行语料库的规模与质量的影响，导致汉越机器翻译性能不佳。

1 相关工作

在低资源情况下的翻译任务中，训练数据稀缺对构建性能较好的机器翻译系统带来很大挑战。为了缓解汉越机器翻译面临的资源稀缺问题，目前解决思路侧重于利用枢轴语言来

善低资源机器翻译的性能，其主要方法分成以下三类：第一类是将源语言数据通过枢轴语言的加入，间接地通过两步翻译生成源语言-目标语言的平行语料，如 Li[4]等人在日汉专利平行语料构建中，利用英语作为枢轴语言，借助 Google 英汉翻译引擎将已有的日英平行语料中的英文句子翻译成汉语，获得日汉平行语料，从而提高了日汉机器翻译的性能；第二类是通过枢轴语言间接的训练源语言-目标语言的机器翻译模型。为减小利用枢轴语言翻译过程中的翻译误差提出了利用源-枢轴语言和枢轴-目标语言的平行语料的三种预训练方法从而提升了低资源的神经机器翻译性能，如 WU 和 Utiyama 等人[6、7]提出轴语言的方法，使用轴语言桥接源语言和目标语言，利用存在的源语言-枢轴语言和枢轴语言-目标语言的平行语料库，分别训练源语言到枢轴语言和枢轴语言到目标语言的翻译模型；Ren 等人[8]提出了一种三角训练结构，用大语种之间丰富的双语数据帮助小语种的翻译，主要思想是将小语种作为中间的隐变量，引入到大语种之间的翻译中，将大语种之间的翻译分解为经由小语种的两个步骤，之后用 EM 方法进行迭代，优化这两个步骤中对应的翻译模型。Wu 和 Johnson 等人[9、10]则提出了一个多语言翻译系统，该系统被应用于谷歌在线翻译系统，该文献指出多语种之间共享注意力机制可以类似于中轴语言。Liu 等人[11]以英语或汉语为中心语言的枢轴机器翻译考察了加入轴语言入后机器翻译的在共享任务的性能提升。上述研究是在不同维度上通过共享语言之间的表示来实现多语言之间的翻译，改进了低资源下的翻译性能；第三类是使用枢轴语料进行联合训练，如 Zheng 等人[12] 分别针对零资源神经机器翻提出了一种最大期望似然估计准则训练的方法，实现在无平行语料情况下的直接翻译建模，缓解了传统方法分段解码所面临的错误传播问题。Zhang 等人[13]利用单语语料，使用 EM 算法优化源语言到目标语言和目标语言到源语言神经机器翻译模型，从而更好地利用单语言数据进行神经机器翻译。Chen 等人[14]提出了使用枢轴语言进行神经机器翻译的联合训练，使用了三种链接方式来桥接源语言-枢轴语言和枢轴语言-目标语言这两种模型，使他们在训练过程中能够相互作用。

以上方法训练所得到的神经机器翻译模型均能提升低资源下的机器翻译任务性能，但是采用枢轴语言进行机器翻译训练过程中，源语言-枢轴语言，枢轴语言-目标语言的模型训练

过程中会因为多语言输入而产生噪声。汉越神经机器翻译是一种典型的低资源场景下的神经机器翻译，其训练语料稀缺，但是却存在大量汉英、英越平行语料，所以汉越神经机器翻译适用于枢轴的方法。为了提升汉越神经机器翻译性能并且利用到小规模的汉越平行语料，我们提出了基于枢轴的汉越联合训练神经机器翻译，其基本思想是，先使用小规模的汉越平行语料训练神经机器翻译模型来得到汉越词语在语义空间上的表示信息，再将其与英语作为枢轴语言的汉语-英语，英语-越南语翻译模型进行联合训练。在联合训练中汉语-英语，英语-越南语翻译模型的汉语、越南语的向量表示与汉越模型得到的汉语、越南语的向量表示计算优化，提升低资源场景下汉越机器翻译的效果。

2 基于枢轴的汉越联合训练神经机器翻译

2.1 基于枢轴的神经机器翻译

神经机器翻译模型是将源语言句子表示成一个固定向量，但是固定长度的向量不能充分表达出源语言句子语义信息与上下文的关系。注意力机制能让神经网络在解码中，重点关注输入中相关度较高的信息，基于注意力机制的神经机器翻译先将源语言句子编码为向量序列，然后在生成目标语言时，通过注意机制动态寻找与生成该词相关的源语言词语信息，大大增强了神经网络机器翻译的表达能力。在这个神经机器翻译模型训练中，我们给定源语言单词的序列表示为 $\mathrm{x}=\left(x^1,...,x^n\right)$，目标语言单词的序列表示为 $y=\left(y^1,...,y^n\right)$,源语言-目标语言的平行语料库表示为 $\mathrm{D_{x,y}}=\left\{<\mathrm{x}^n,y^n>\right\}_{n=1}^{N}$。我们用 $P\left(\mathrm{y}|\mathrm{x};\theta_{x\to y}\right)$ 表示一个基于注意力机制的神经机器翻译模型 Bahdanau 等人[15]，其中 $\theta_{\mathrm{x}\to y}$ 是模型参数。模型可以使用最大似然估计表示为：

$$\hat{\theta}_{\mathrm{x}\to y}=\arg\max_{\theta_{x\to y}}\left\{\mathcal{L}(\theta_{x\to y})\right\} \tag{1}$$

其对数似然函数可表示为：

$$\mathcal{L}\left(\theta_{\mathrm{x}\to y}\right)=\sum_{\mathrm{n}=1}^{N}\log P\left(y^{(n)}|\mathrm{x}^{(n)};\theta_{x\to y}\right) \tag{2}$$

Wu [5] 等人提出使用轴语言的方法，假设存在枢轴语言 $\mathrm{z}=\left(z^1,...,z^n\right)$。源语言-枢轴语言的语料库 $\mathrm{D_{x,z}}=\left\{<\mathrm{x}^n,z^n>\right\}_{n=1}^{N}$，枢轴语言-目标语言的语料库 $\mathrm{D_{z,y}}=\left\{<\mathrm{z}^n,y^n>\right\}_{n=1}^{N}$。使用轴语言桥接源语言和目标语言。利用存在的源语言-枢轴语言和枢轴语言-目标语言的平行语料库，分别训练源语言到枢轴语言和枢轴语言到目标语言的翻译模型可表示为：

$$\hat{\theta}_{\mathrm{x}\to z}=\arg\max_{\theta_{x\to z}}\left\{\mathcal{L}(\theta_{x\to z})\right\} \tag{3}$$

$$\hat{\theta}_{z\to y} = \arg\max_{\theta_{z\to y}} \left\{ \mathcal{L}(\theta_{z\to y}) \right\} \tag{4}$$

其对数似然函数可表示为：

$$\mathcal{L}\left(\theta_{x\to z}\right) = \sum_{n=1}^{N} \log P\left(z^{(n)} | x^{(n)}; \theta_{x\to z}\right) \tag{5}$$

$$\mathcal{L}\left(\theta_{z\to y}\right) = \sum_{n=1}^{N} \log P\left(y^{(n)} | z^{(n)}; \theta_{z\to y}\right) \tag{6}$$

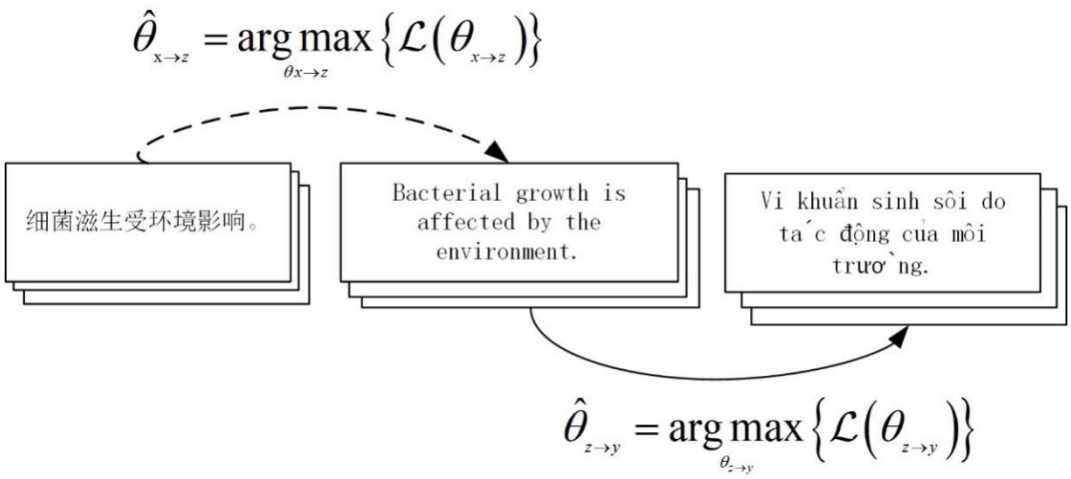

Figure 1: Translation process of a pivot-based machine translation model

图 1：基于枢轴的汉越神经机器翻译示意图

2.2 基于枢轴的汉越联合训练神经机器翻译

为了缓解训练中枢轴语言带来的误差传播问题，我们采取了联合训练的方法。使用汉英，英越平行语料来对汉越进行联合训练从而提升汉越神经机器翻译性能。其中神经机器翻译联合训练可表示为：

$$J(\theta_{x\to z}, \theta_{z\to y}) = \mathcal{L}(\theta_{x\to z}) + \mathcal{L}(\theta_{z\to y}) \tag{7}$$

其中 $\mathcal{L}(\theta_{x\to z})$ 与 $\mathcal{L}(\theta_{z\to y})$ 表示的是汉语-英语，英语-越南语的似然函数。在联合训练中，为了降低枢轴语言带来的传播误差，所以词在语义空间中词的表示要一样，我们定义 $v^{wz}{}_{x\to z}$ 是汉语-英语的词表中的英语，$v^{wz}{}_{z\to y}$ 是英语-越南语的词表中的英语，用 $w \in (v^{wz}{}_{x\to z} \cap v^{wz}{}_{z\to y})$ 表示 w 是汉语-英语，英语-越南语词表中共有英语的词，λ 是超参数，则基于枢轴联合训练的注意力机制神经机器翻译模型可以表示为：

$$J(\theta_{x\to z}, \theta_{z\to y}) = \mathcal{L}(\theta_{x\to z}) + \mathcal{L}(\theta_{z\to y}) + \lambda R(\theta_{z\to y}, \theta_{z\to y}) \tag{8}$$

$$\mathcal{J}(\theta_{x\to z},\theta_{z\to y})=\mathcal{L}(\theta_{x\to z})+\mathcal{L}(\theta_{z\to y})+\lambda\sum_{w\in(v^{wz}_{x\to z},v^{wz}_{z\to y},)}\|\theta^{wz}{}_{x\to z}-\theta^{wz}{}_{z\to y}\|^2 \tag{9}$$

为了提升基于枢轴的汉越联合训练神经机器翻译的性能，我们使用小规模的汉越平行语料训练了汉越神经机器翻译模型，得到了汉语与越南语的向量表示，并且把它们加入到汉英，英越的联合训练中。其中 $\hat{w}$ 是汉越词典与 w 中的共有词英语对应的汉语与越南语。

$$\mathcal{J}(\theta_{x\to z},\theta_{z\to y})=\mathcal{L}(\theta_{x\to z})+\mathcal{L}(\theta_{z\to y})+\lambda R(\theta_{x\to z},\theta_{z\to y},\ \theta_{x\to y}) \tag{10}$$

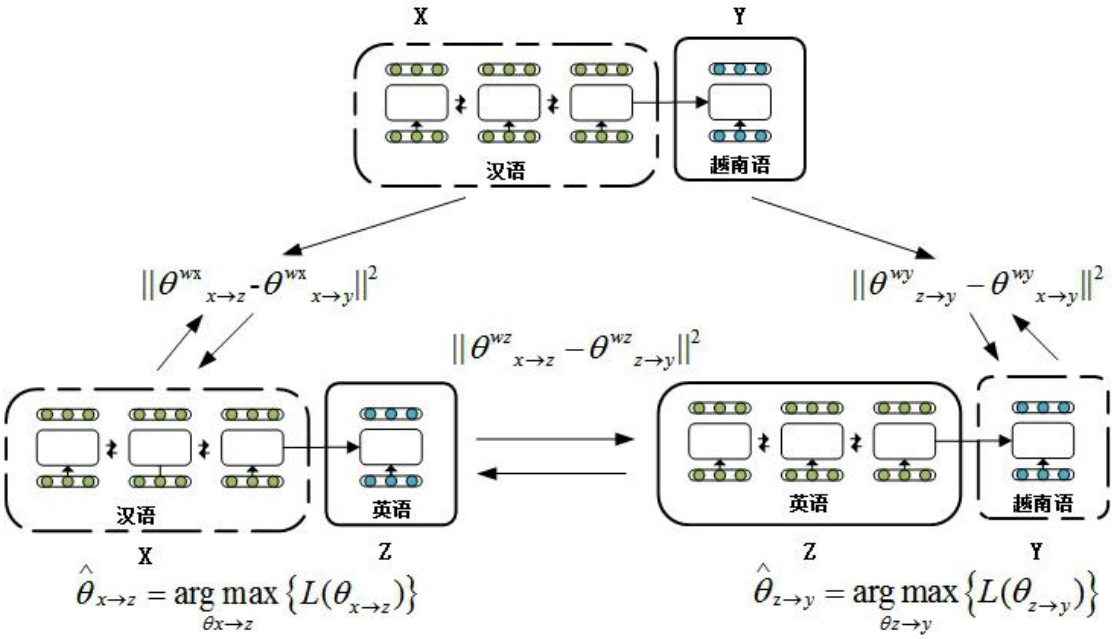

图 2：基于枢轴的汉越联合训练神经机器翻译训练流程图

Figure 2: Pivot-based Chinese-Vietnamese joint training neural machine translation training flowchart

由公式（5）可得在训练过程中基于枢轴的汉越联合训练神经机器翻译模型的最大似然估计表示为：

$$\hat{\theta}_{x\to z},\hat{\theta}_{z\to y}=\arg\max_{\theta_{x\to y}}\{\mathcal{L}(\theta_{x\to z},\theta_{z\to y})\} \tag{11}$$

源语言到枢轴语言的模型的参数 $\theta_{x\to z}$ 的偏导数可以计算为：

$$\frac{\partial\mathcal{J}(\theta_{x\to z},\theta_{z\to y})}{\partial\theta_{x\to z}}=\frac{\partial\mathcal{L}(\theta_{x\to z})}{\partial\theta_{x\to z}}+\lambda\frac{\partial R(\theta_{x\to z},\theta_{z\to y},\ \theta_{x\to y})}{\partial\theta_{x\to z}} \tag{12}$$

源语言到枢轴语言的模型的参数 $\theta_{z\to y}$ 的偏导数可以计算为：

$$\frac{\partial\mathcal{J}(\theta_{x\to z},\theta_{z\to y})}{\partial\theta_{z\to y}}=\frac{\partial\mathcal{L}(\theta_{z\to y})}{\partial\theta_{z\to y}}+\lambda\frac{\partial R(\theta_{x\to z},\theta_{z\to y},\ \theta_{x\to y})}{\partial\theta_{z\to y}} \tag{13}$$

3 实验

3.1 实验数据

本文在低资源翻译场景下进行汉语–越南语和越南语–汉语的神经机器翻译实验。实验训练语料规模为:汉越平行语料规模 10 万句对，英越平行语料规模 70 万句对，汉英语料规模 1000 万句对。具体的实验数据集如表 1 所示。在训练之前对实验数据进行了过滤乱码与分词处理，其中汉语分词采用结巴分词，越南语分词采用 Underthesea-Vietnamese NLP 工具。

表 1 实验数据集表

Tab. 1 Experimental Data Set Table

数据集	训练集	验证集	测试集
汉英	10M	10K	20K
英越	700k	4K	5K
汉越	100k	1K	2K

3.2 实验设置

为了评估基于枢轴的汉越联合训练神经机器翻译方法的有效性我们设置了以下几组对比实验。实验选取了六个基线系统，分别是基于统计机器翻译的 Moses[16]、基于 OPENNMT[17]框架的 Transformer、Convolutional Neural Networks（CNN）、基于注意力机制的 GNMT [18]、传统的枢轴机器翻译、李等人[19]提出的迁移学习实现的藏汉神经机器翻译模型 Nmt+trans、利用单语数据进行回译（Back Translation,BT）[20]与本文的方法（不使用汉越语料）、本文的方法+CV（使用汉越语料）进行翻译效果的对比。

Moses 训练中，我们使用了 Mgiza[21]训练词对齐，利用 Lmplz[22]训练 3-gram 的 Language Model（LM）。CNN 中编码器设置为 10 层的卷积神经网络，解码器则采用 LSTM 网络，批次大小为 64，卷积核大小设置为 3。GNMT 中隐藏层数量设置为 2，“num_units”设置为 128，“dropout”设置为 0.2。Back Translation 在实验前需要对数据进行预处理，首先对训练数据进行 tokenization 处理，将句子长度在 50 个词以上的句对过滤，使用谷歌（Google）开源模型 Transformer，“dropout”设置为 0.1，批次大小为 64。Transformer 机器翻译模型、Nmt+trans、传统的枢轴机器翻译与本文的方法采用基于 OPENNMT[23]框架的 Transformer，使用的词表设置为 32000 个词，实验均在单卡 GPU 服务器上进行，句子的最大长度设置为 50，“transformer_ff”设置为 2048，“label_smoothing ”设置为 0.1，“attention head”

设置为 2，“dropout”设置为 0.2，隐藏层数量设置为 2，词嵌入维度设置为 256，“batch_size”设置为 128，学习率设置为 0.2。优化器选择 Adam，其参数设置为$\beta_1 = 0.9$、$\beta_2 = 0.99$、$\varepsilon = 1e-8$。实验中使用 BLEU 值作为评测指标。传统的枢轴机器翻译，即采用分步训练的方法，先训练一个汉语翻译成英语的神经机器翻译模型再训练一个由英语翻译成越南语的神经机器翻译模型，最后利用测评语料汉语通过二次解码的方法，分步翻译由汉语翻译成越南语，传统的枢轴翻译方法由于经过两次编码解码翻译误差较大，因为汉越语料规模较少，汉越神经机器翻译模型训练不充分，对于词频较低的词语，翻译性能不好，所以在联合训练中，只取词频大于 Top_k=30 的词汇。

3.3 实验结果及对比分析

表 2 中给出的是基线系统与基于枢轴的汉越联合训练神经机器翻译在汉语-越南语和越南语-汉语两个翻译方向上的模型的 BLEU 值对比结果。

表 2 不同模型的 BLEU 值对比结果

Tab. 2 Comparison of BLEU values of different models

模型	汉语-越南语	越南语-汉语
Moses	16.39	16.21
CNN	16.87	16.35
GNMT	14.21	16.47
Transformer	17.35	17.02
传统的枢轴方法	18.16	17.79
Nmt+trans	17.98	17.65
Back Translation	18.03	17.95
本文的方法	18.75	18.12
本文的方法+CV	19.16	18.64

从实验结果对比看，基线模型中 Transformer 模型的 BLEU 值高于其它基线模型，这说明基于 Transformer 的汉越神经机器翻译框架中的遮蔽注意力机制可以更好地对目标语言进行翻译；汉越双语神经机器翻译上，本文采用的方法其效果明显优于基线系统，其中本文方法对比 Moses 方法在汉语-越南语翻译方向上提升了 2.77 个 BLEU 值，在越南语-汉语方向

上提升了 2.43 个 BLEU 值，这说明基于汉越神经机器翻译的方法比统计机器翻译更好。对比 Transformer 方法在汉语-越南语翻译方向上提升了 1.81 个 BLEU 值，在越南语-汉语翻译方向上提升了 1.62 个 BLEU 值。对比传统的枢轴方法在汉语-越南语翻译方向上提升了 1 个 BLEU 值，在越南语-汉语方向上提升了 0.33 个 BLEU 值。对比 Nmt+trans 在汉语-越南语翻译方向上得到 1.18 个 BLEU 值提升，越南语-汉语翻译方向上得到个 0.99 个 BLEU 值提升。对比 Back Translation 在汉语-越南语翻译方向上提升了 1.13 个 BLEU 值，在越南语-汉语翻译方向上提升了 0.69 个 BLEU 值。说明在实验过程中，使用小规模汉越语料训练汉越神经机器翻译模型得到汉越词的语义表示信息，再将汉越的语义表示信息与汉英，英越翻译模型进行联合训练，从而提升翻译模型的性能，同时也验证了本文方法的有效性。

表 2 中给出的是基线系统与基于枢轴的汉越联合训练神经机器翻译在汉语-越南语翻译方向上译文的对比示例。

表 3 不同模型的译文示例

Tab. 3 Translation Examples of Different Models

源语言句	这场 **比赛** 从 上午 一直 持续 到 下午 。
译文	**Cuộc thi đấu** này từ buổi sáng kéo dài đến tận chiều.
Transformer	**Trò chơi** kéo dài từ sáng đến chiều.
本文的方法+CV	Cuộc thi **đấu** diễn ra từ sáng đến chiều.
源语言句	生活 就 像 一杯 **白开水** ，你 每天 都在 喝 ，不要 羡慕 别人 喝 的 饮料 有 **各种 颜色** ，其实 未必有 **你的** 白开水 解渴
译文	Cuộc đời giống như một cốc **nước sôi**, anh ngày nào cũng uống nước, đừng ghen tị với người khác uống đồ uống **có nhiều màu sắc**, thực ra chưa chắc đã có màu trắng nước giải khát **bạn**.
Transformer	Cuộc sống là giống như một ly **nước**,anh uống mỗi ngày , không phải là để ghen tị với những người khác uống **một thức uống màu sắc**, trong thực tế, không phải là nước của **bạn** để quench khát.
本文的方法+CV	Cuộc sống giống như một ly **nước sôi**. anh uống nó mỗi ngày nước. Đừng ghen tị với những đồ uống mà người khác uống **có nhiều màu sắc**. Thực tế, nước đun sôi của bạn có thể không làm dịu cơn khát của **bạn**.

从表 3 的第一组句子中可以看出 Transformer 的译文出现了语句不准确的现象，翻译错了比赛 “ Cuộc thi đấu ”，相比之下本文的方法+CV 更加准确。在第二组句子中，Transformer 的译文比起第一组数据，出现了更多漏译的情况，例如， 白开水“ nước sôi”、很多颜色“ có nhiều màu sắc”等。由于漏翻的词汇在基线模型的训练语料中出现的次数较少，神经机器翻译模型无法很好的学习低频词的语义表示，从而出现了漏翻的情况。而本文采用基于枢轴的联合训练方法再使用小规模的汉越语料的前提下，还使用了英语桥接汉语和越南语，提升了汉越翻译任务的性能。本文方法虽然还存在翻译不充分的问题，但是在汉越神经机器翻译上对比基线系统，其产生的译文准确度更高。

4 结论

针对汉越平行语料缺失导致翻译模型性能不佳的问题，本文提出了基于枢轴的汉越联合训练神经机器翻译方法。通过利用小规模汉越语料训练汉越神经机器翻译模型得到汉越词的语义表示信息，再将汉越的语义表示信息与汉英，英越翻译模型进行联合训练。实验结果表明，该方法能够提升低资源场景下汉越神经机器翻译性能，在汉语-越南语的翻译方向上达到了 19.16 的 BLEU 值，相比较于基线模型均有明显的提升。在下一步的工作中，我们将研究在现有基础上融入汉越的词对齐信息及枢轴词典等，从而提升越南语的翻译性能。

参考文献：

[1] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks. networks[C]//Advances in Neural Information Processing Systems.New York: Massachusetts Institute of Technology Press, 2014: 3104-3112

[2] Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg:ACL,2015: 1412-1421.

[3] Zoph B, Yuret D, May J, et al. Transfer learning for low-resource neural machine translation[J]. arXiv preprint ,2016,arXiv:1604.02201

[4] Cohn T , Lapata M . Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora[C]// Acl, Meeting of the Association for Computational Linguistics, June, Prague, Czech Republic. DBLP, 2007.

[5] Li X, Meng Y, Yu H. Improving Chinese-to-Japanese Patent Translation Using English as Pivot

Language[C]//Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation.Indonesia:Faculty of Computer Science,Universitas Indonesia . 2012: 117-126.

[6] Wu H, Wang H. Pivot language approach for phrase-based statistical machine translation[J]. Machine Translation, 2007, 21(3): 165-181.

[7] Utiyama M, Isahara H. A comparison of pivot methods for phrase-based statistical machine translation[C]// The Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference. Stroudsburg: ACL, 2007: 484-491.

[8] Ren S, Chen W, Liu S, et al. Triangular Architecture for Rare Language Translation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: (Volume 1: Long Papers).56-65.

[9] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint ,2016,arXiv:160908144,

[10] Johnson M , Schuster M , Le Q V , et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5:339-351.

[11] Liu C H , Silva C C , Wang L , et al. Pivot Machine Translation Using Chinese as Pivot Language[M]. Springer, Singapore, 2018.

[12] Zheng H, Cheng Y, Liu Y. Maximum Expected Likelihood Estimation for Zero-resource Neural Machine Translation[C]//International Joint Conference on Artificial Intelligence.California.Morgan Kaufmann. 2017: 4251-4257.

[13] Zhang Z , Liu S , Li M , et al. Joint Training for Neural Machine Translation Models with Monolingual Data[J]. 2018.

[14] Cheng Y, Yang Q, Liu Y, et al. Joint Training for Pivot-based Neural Machine Translation[C]//26th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2017: 3974-3980.

[15] Bahdanau D , Cho K , Bengio Y . Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer ence, 2014.

[16] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation[C]// 55th Annual Meeting of the Association for Computational Linguistics, Companion volume proceedings of the demo and poster sessions. Stroudsburg: ACL, 2007: 177-180.

[17] Klein G, Kim Y, Deng Y, et al. OpenNMT: Open-Source Toolkit for Neural Machine Translation[C]// 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations. Stroudsburg: ACL, 2017: 67-72.

[18] Thang Luong and Eugene Brevdo and Rui Zhao. Neural Machine Translation (seq2seq) Tutorial.2017. https://github.com/tensorflow/nmt.

[19] 李亚超, 熊德意, 张民, 江静, 马宁, 殷建民. 藏汉神经网络机器翻译研究[J]. 中文信息学报, 2017, 31(6): 103-109.

[20] Sennrich R, Haddow B, Birch A, et al. Improving neural machine translation models with monolingual data[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Long Papers), 2016: 86-96

[21] Gao Q, Vogel S. Parallel implementations of word alignment tool[C]//Software Engineering, Testing, and quality assurance for natural language processing. Stroudsburg: ACL, 2008: 49-57.

[22] Heafield K, Pouzyrevsky I, Clark J H, et al. Scalable modified Kneser-Ney language model estimation[C]// 51th Annual Meeting of the Association for Computational Linguistics, Short Papers. Stroudsburg: ACL, 2013: 690-696.

[23] Klein G, Kim Y, Deng Y, et al. OpenNMT: Open-Source Toolkit for Neural Machine Translation[C]// 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations. Stroudsburg: ACL, 2017: 67-72.

Chinese-Vietnamese Joint Training Neural Machine Translation Based on Pivot

GAO Shengxiang[1,2], LIU Chang[1,2], YU Zhengtao [1,2*], HUANG Jihao [1,2]

(1.College 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)

Abstract: Vietnamese is a typical low-resource language. In order to alleviate the scarcity of resources faced by Chinese-Vietnamese machine translation, a method of Chinese-Vietnamese

neural machine translation based on joint training is proposed. First, a small-scale Chinese-Vietnamese parallel corpus is used to train the translation model to obtain Chinese and Vietnamese word vector representations, and then the Chinese-English and English-Vietnamese translation models with English as the pivot language are jointly trained. In the joint training, the Chinese-Vietnamese vector representations of the Chinese-English and English-Vietnamese translation models are optimized with the Chinese and Vietnamese vector representations obtained from the Chinese-Vietnamese model. Experimental results show that the method in this paper effectively combines Chinese-Vietnamese parallel corpus with Chinese-English and English-Vietnamese parallel corpus for joint training, which improves the Chinese-Vietnamese machine translation performance to a certain extent.

Keywords: Neural Machine Translation; Low Resources; Chinese-Vietnamese; Pivot; Joint Training;

中图分类号： TP391 **文献标码:** A