文章编号：1003-0077(2021)03-0043-08

融合 EMD 最小化双语词典的汉—越无监督神经机器翻译

薛明亚[1,2]，余正涛[1,2]，文永华[1,2]，于志强[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：神经机器翻译在平行语料充足的任务中能取得很好的效果，然而对于资源稀缺型语种的翻译任务则往往效果不佳。汉语和越南语之间没有大规模的平行语料库，在这项翻译任务中，该文探索只使用容易获得的汉语和越南语单语语料，通过挖掘单语语料中词级别的跨语言信息，融合到无监督翻译模型中提升翻译性能；该文提出了融合 EMD(Earth Mover's Distance)最小化双语词典的汉—越无监督神经机器翻译方法，首先分别训练汉语和越南语的单语词嵌入，通过最小化它们的 EMD 训练得到汉越双语词典，然后再将该词典作为种子词典训练汉越双语词嵌入，最后利用共享编码器的无监督机器翻译模型构建汉—越无监督神经机器翻译方法。实验表明，该方法能有效提升汉越无监督神经机器翻译的性能。

关键词：无监督学习；EMD；汉语—越南语；神经机器翻译

中图分类号：TP391 **文献标识码**：A

Chinese-Vietnamese Unsupervised Neural Machine Translation Based on EMD Minimal Bilingual Dictionary

XUE Mingya[1,2], YU Zhengtao[1,2], WEN Yonghua[1,2], YU Zhiqiang[1,2]

(1. Faculty 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: Neural machine translation (NMT) has achieved good results in tasks with sufficient parallel corpora, but often has poor results in translation tasks with scarce resources. To address NMT between Chinese and Vietnamese without large-scale parallel corpus, we explore the use of easily available Chinese and Vietnamese monolingual corpora by mining cross-language information at the word level. A Chinese-Vietnamese unsupervised neural machine translation method that incorporates Earth Mover's Distance(EMD) to minimize bilingual dictionaries is proposed. First, monolingual word embeddings for Chinese and Vietnamese are trained independently, and a Chinese-Vietnamese bilingual dictionary is obtained by minimizing their EMD. The dictionary is then used as a seed dictionary to train the Chinese-Vietnamese bilingual word embeddings. Finally, the shared encoder unsupervised machine translation model is applied to construct a Chinese-Vietnamese unsupervised neural machine translation. Experiments show that this method can effectively improve the performance of Chinese-Vietnamese unsupervised neural machine translation.

Keywords: unsupervised learning; Earth Mover's Distance; Chinese-Vietnamese; neural machine translation

收稿日期：2019-12-19 **定稿日期**：2020-02-11

基金项目：国家重点研发计划(2019QY1801)；国家自然科学基金(61732005，61672271，61761026，61762056，61866020)；云南省高新技术产业专项(201606)

0 引言

随着越南与我国的交流与合作越来越密切，机器翻译是跨语言信息交流比较有效的方式之一，研究汉越机器翻译有着非常重要的应用前景。

神经机器翻译（neural machine translation，NMT）[1-2]是近几年提出的机器翻译方法，并且NMT翻译质量已经在多个语言对上超过统计机器翻译[3]，成为主流的翻译方法。然而NMT需要大规模的平行语料才能取得较好的效果，当训练数据不足时，会导致翻译质量不佳[4]。汉语和越南语之间的平行语料稀少且不容易获取，所以汉—越机器翻译是典型的低资源语言机器翻译。但是汉语和越南语有大量的单语语料，本文探索只利用单语语料的汉—越无监督NMT，这对于其他低资源语言的机器翻译的研究也具有重要的理论和应用价值。

近年来，国内外相关研究人员针对无监督机器翻译的方法进行了大量研究，并取得了一系列成果。目前，无监督机器翻译的研究方法主要有基于对抗学习（generative adversarial networks，GAN）的无监督机器翻译和基于共享编码器的无监督机器翻译。Lample等人[5]提出将两种不同的单语语料库句子映射到同一空间，通过学习用这两种语言重建共享特征空间，仅利用单语语料实现无监督NMT。Artetxe等人[6]对模型进行修改，先预训练无监督的双语词嵌入，采用共享编码器和分别解码的方式提出了仅仅使用单语语料的无监督NMT。Yang等人[7]提出权重共享的无监督机器翻译模型，相较于共享编码器模型强化了每种语言的自身特点和内部特征，以此提高翻译质量。Lample等人[8]结合NMT和基于短语的统计机器翻译效果，可以得到进一步提升无监督NMT的效果。Lample等人[9]提出跨语言模型预训练，用于初始化查找表来提升预训练的跨语言词嵌入的质量，对无监督机器翻译模型的性能有显著提高。他们从相近语言的单语语料中利用同源词作为初始跨语言信息或者数字对齐，然后扩展学习实现无监督NMT。汉、越语言差异性较大，汉、越之间没有可以利用的同源词，所以利用语言同源词的方法在汉—越语言对上不可行，而Artetxe等人提出共享编码器的无监督NMT是在无监督的双语词嵌入的基础上实现的，符合语言差异性较大的特点。本文在Artetxe等人的工作上进行延伸，通过提升无监督双语词嵌入质量来提升汉—越无监督NMT的质量。

在只使用汉语和越南语单语语料的无监督机器翻译中，要直接实现机器翻译较难，但获取双语词典相对较容易，因此本文考虑从汉、越单语语料中先训练汉越双语词典，然后利用汉越双语词典作为种子词指导训练较高质量的双语词嵌入，从而提高汉越无监督NMT质量。Zhang等人[10]提出利用语言的词嵌入空间分布的相似性，使用EMD最小化的方法训练双语词典，整个过程只使用单语语料的无监督训练方式，且质量可以与有监督的方式相媲美，符合汉、越语言差异性较大的特点。所以本文提出融合EMD最小化双语词典的汉—越无监督NMT。

本文方法首先将汉语和越南语单语的词嵌入空间视为两个分布，通过最小化它们之间EMD距离训练汉—越双语词典，在不需要汉越双语信息的情况下，训练得到没有同源词语言的汉越双语词典。然后将汉越双语词典作为种子词典，利用自学习的方法训练汉—越双语词嵌入，在共享编码器模型上实现汉—越无监督NMT。

1 融合EMD最小化双语词典的汉—越无监督NMT模型

1.1 模型结构

本文提出的方法是在Artetxe等人共享编码器的基础上融合了基于EMD最小化的无监督双语词典，比原模型具有更强的挖掘汉语和越南语单语语料中跨语言信息的能力。模型架构如图1所示，该模型遵循Bahdanau等人[1]提出的具有注意机制的标准的编码器和解码器，由一个共享编码器和两个解码器组成，两个解码器分别对应源语言和目标语言。编码器端为双层双向循环神经网络（BiGRU），解码器端为双层循环神经网络（UniGRU）。关于注意力机制，本文使用Luong等人[11]提出的全局注意力方法和一般对齐函数。在编码器端，使用预训练的汉—越双语词典和双语词嵌入，接受输入序列并生成与语言无关的表征。而解码器端的词嵌入会随着训练不断更新，通过两个解码器进行训练和翻译。

对于汉语（L1）中的每个句子，系统交替两个步骤训练：去噪，它优化了用共享编码器对句子的噪声编码进行编码的概率，并用L1解码器重建它；回

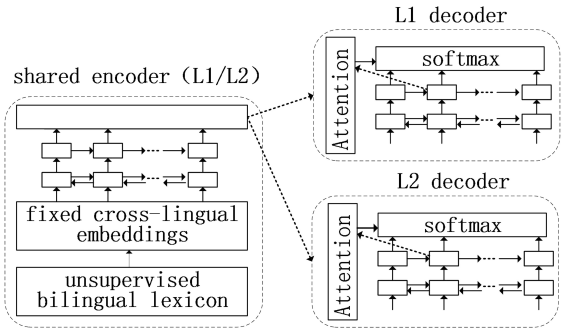

图 1　融合 EMD 最小化双语词典的汉—越无监督 NMT 模型

译，并进行动态反向翻译，它以推理模式翻译句子（用共享编码器对其进行编码并用越南语（L2）解码器进行解码），然后用共享编码器优化编码该翻译语句的概率，并用 L1 解码器恢复原始句子。模型的训练在 L1 和 L2 中的句子之间交替进行。

（1）**对偶结构**　虽然 NMT 系统通常是针对特定的翻译方向而构建的（如汉语→越南语或越南语→汉语），但本文利用机器翻译的双重性质[12-13]同时处理两个方向（如汉语↔越南语）。

（2）**共享编码器**　类似于 Ha 等人[14]、Lee 等人[15]和 Johnson 等人[16]，本文的系统是由两种语言共享的一个编码器。即汉语和越南语使用同一个编码器进行编码。该共享编码器旨在将两种语言表示成与语言无关的形式，然后将每个解码器解码成与其对应的语言。

（3）**预训练固定的双语词嵌入**　虽然大多数 NMT 系统随机初始化其词嵌入并在训练期间更新它们，但在编码器中使用预先训练的跨语言词嵌入，这些词嵌入在训练过程中保持不变。编码器具有与语言无关的单词级表示，并且它只需要学习如何组合来构建较大短语的表示。对于系统中提到的无监督的双语词典和双语词嵌入，将在下文中详细介绍。

1.2　基于 EMD 最小化训练的无监督双语词典

双语词典的获取大体可以分为三个步骤：第一步，将两种语言中的每个词表示为向量；第二步，为两种语言的向量空间建立联系，得到共有的双语向量空间；第三步，在双语向量空间中进行查找，获取双语词典。

首先使用 Word2Vec[17]训练汉语和越南语单语词嵌入，完成第一步。词嵌入分布如图 2 所示，图中所示的汉语和越南语的词嵌入是分别独立地在各自语言的单语语料上训练得到的，可以看出两种语言的单语词嵌入空间表现出近似的同态性，这意味着存在线性映射能够近似地连接这两个空间。Mikolov 等人[18]利用种子词典来学习这个线性映射，然而，本文希望完全不使用双语监督信号，因此需要设计一个方法来学习这个映射，并且这个方法不能依赖于种子翻译词对这种级别的监督信号。在生成对抗网络[19]的基础上，把词嵌入的跨语言映射学习建模成一个对抗游戏，张檬等人[20]成功实现了不使用任何双语监督信号联系两种语言的词嵌入空间，使得单纯的基于非平行语料的双语词典构建成为可能。在此基础上本文使用最小化它们的 EMD 训练得到无监督的汉越双语词典。

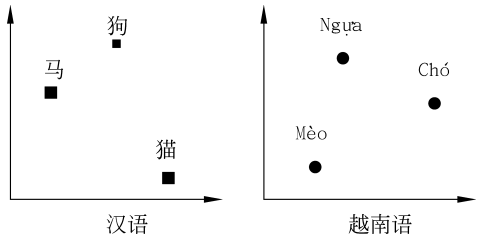

图 2　汉语和越南语的单语词嵌入空间

将图 2 中的圆点视为土堆，方块视为坑洞，它们的大小代表土堆的体积和坑洞的容积，或者说相应的权重。在图 3(a)中，所有的权重都相等。在这个

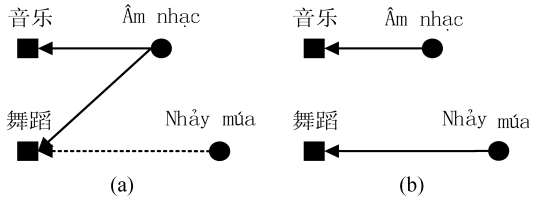

图 3　Hubness 问题

设定下，希望用最小的整体代价来移动土堆填满坑洞，而代价是由移动土堆的距离和体积的乘积衡量的。图 3(b)中的箭头代表了这个示例下的最优移动方案，而这个方案正好可以视为词汇翻译的结果。从单个词语看，由于"Âm nhạc"土堆中的泥土已经全部用来填"音乐"坑洞，它将不会去干涉"舞蹈"坑洞，从而由"Nhảy múa"土堆负责填满"舞蹈"坑洞。从整个词语集合看，整体移动代价的最小化使得可以考虑全局的信息，从而克服最近邻查找的局部性，以应对 hubness 问题[21]。上述比喻代表全局带权匹配思想，在数学上可以用 EMD 来实现，它的名字正是来源于上述的比喻。其对应如下的线性规划问题，如式(1)所示。

$$\begin{aligned}&\min\sum_{i=1}^{V_t}\sum_{j=1}^{V_s}W_{ij}C_{ij}\\&s.t.W_{ij}\geqslant 0\\&\sum_{j=1}^{V_s}W_{ij}\leqslant t_i,\quad i\in\{1,\cdots,V_t\}\\&\sum_{i=1}^{V_t}W_{ij}\leqslant s_j,\quad j\in\{1,\cdots,V_s\}\end{aligned}\tag{1}$$

其中，V_s代表源语言词汇表大小，V_t代表目标语言词汇表大小，C_{ij}代表第i个土堆与第j个坑洞之间的距离，t_i代表第i个土堆的体积，s_j代表第j个坑洞的容积，W_{ij}为优化问题的决策变量，代表从第i个土堆转移到第j个坑洞的泥土体积，因此，目标函数即为最小化整体的移动代价。求解完成后，非零的W_{ij}值即代表第j个源语言词与第i个目标语言词之间存在翻译关系。实验为了能更好地发挥EMD处理一词多译现象的能力，将EMD引入双语词嵌入的训练过程中。在训练的目标函数中，EMD作为其中一项以正则的形式参与训练，使得训练得到的双语词嵌入能够更好地捕捉一词多译现象，其效果通过实验得到了印证[22]。

前面对抗学习的方法也可以放在这个框架下看待，因为对抗学习隐式地优化了Jensen-Shannon divergence[19]。但是对于词汇翻译的任务来说，可能有其他更好的分布距离供选择。由于EMD也是分布之间距离的一种度量，其对词汇翻译任务非常适合，所以考虑使用EMD作为词汇表级别的准则来指导线性映射的学习，即寻找一个映射G，使得源语言经过映射后的词嵌入分布与目标语言的词嵌入分布之间的EMD最小化，如图4所示。使用数学公式可以表示成式(2)的形式。

$$\min_G \mathrm{EMD}(p_{G(x)},p_y)\tag{2}$$

其中，$p_{G(x)}$代表经过G映射后的源语言词嵌入分布，p_y代表目标语言词嵌入分布。

在EMD的优化问题上，利用了Wasserstein GAN(WGAN)，它可以视为优化EMD的GAN变种，再结合将EMD代入式(2)进行优化，有效地最小化EMD，找到相应的映射。

1.3 融合EMD最小化双语词典的汉越双语词嵌入的学习

词嵌入映射 假设汉语和越南语的词嵌入矩阵分别为$\boldsymbol{X}$和$\boldsymbol{Y}$，$\boldsymbol{X}_{i*}$为源语言的第i个词的向量，$\boldsymbol{Y}_{j*}$为目标语言的第j个词的向量；词典$\boldsymbol{D}$为一个

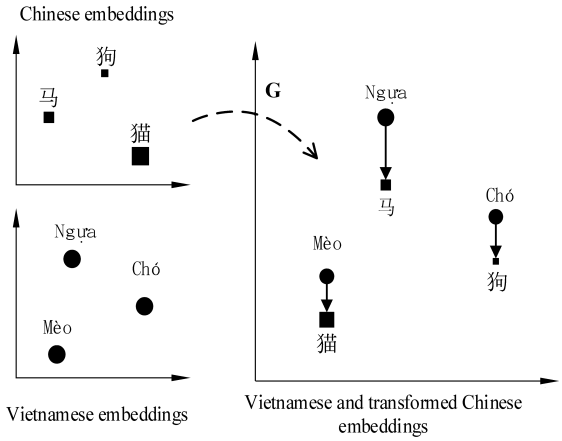

图4 EMD最小化学习

二进制的矩阵，当源语言第i个词与目标语言的第j个词对齐时，$\boldsymbol{D}_{ij}=1$。词映射的目标是找到一个映射矩阵$\boldsymbol{W}*$，使映射后的$\boldsymbol{X}_{i*}$和$\boldsymbol{Y}_{j*}$的欧氏距离最近，如式(3)所示。

$$\boldsymbol{W}^*=\arg\min_{\boldsymbol{W}}\sum_i\sum_j\boldsymbol{D}_{ij}\|X_{i*}\boldsymbol{W}-\boldsymbol{Y}_{i*}\|^2\tag{3}$$

对矩阵$\boldsymbol{X}$和$\boldsymbol{Y}$进行标准化和中心化，并将$\boldsymbol{W}$设置为正交矩阵后，上述求解欧氏距离的问题相当于最大化点积，如式(4)所示。

$$\boldsymbol{W}*=\arg\min_{\boldsymbol{W}}Tr(\boldsymbol{X}\boldsymbol{W}\boldsymbol{Y}^T\boldsymbol{D}^T)\tag{4}$$

其中，Tr表示矩阵的迹运算。可以求解得到最优解为$\boldsymbol{W}^*=\boldsymbol{U}\boldsymbol{V}^T$($U$、$V$表示两个正交矩阵)，经过奇异值分解，$\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{Y}=\boldsymbol{U}\sum\boldsymbol{V}^T$。鉴于矩阵$\boldsymbol{D}$是稀疏的，可以在线性时间内得到解。

词典自学习 映射后的源语言词的词嵌入与目标语言词的词嵌入在同一个空间内。根据最近邻检索的方法，为每个源语言词分配一个距离最近的目标语言词，将对齐的词对添加到词典中，再次进行迭代，直到收敛。

以图5为例，一开始词典中对齐的词对为(马-Ngựa，狗-Chó)，根据词典L1进行了一次映射，使得映射后的“马”与“Ngựa”以及“狗”与“Chó"之间的欧氏距离最近。然后在映射后的空间里，为其他词寻找距离最近的对应词，可以发现“猫”与“Mèo”的距离较近，因此把它也加入词典中。此时，尽管词典中包含了所有的词对，却并不是最佳的结果。将更新后的词典(马-Ngựa，狗-Chó，猫-Mèo)作为新的参考词典，重新进行欧氏距离的计算，得到了新的映射矩阵W^*，从而获得新的对齐结果。

训练结束用集束搜索(beam search)进行翻译，束的大小需权衡翻译的时间以及搜索的准确性来

确定。

融合基于 EMD 最小化训练的无监督双语词典，是将无监督获得的词典作为种子词典来提升词典自学习的效果，进而提升双语词嵌入的质量。

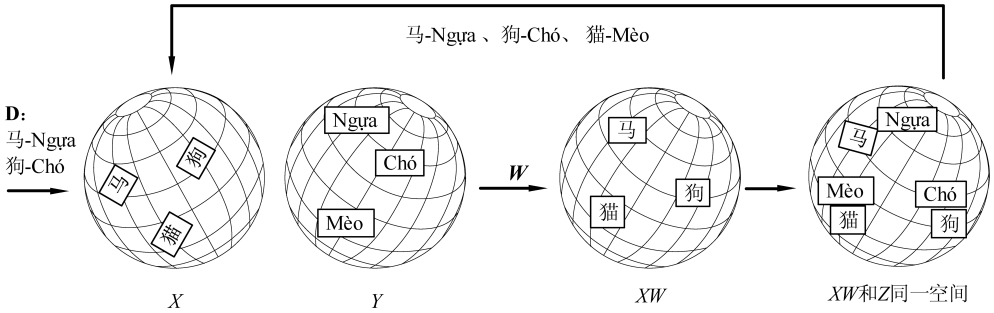

图 5　使用种子词典进行词映射过程的示意图

1.4　融合 EMD 最小化双语词典的汉-越无监督 NMT 模型的训练

在 Artetxe 等人[6]的实验中证明，在系统中加入去噪和回译有助于提升翻译质量，因此本文使用带有去噪和回译的共享编码器系统。

对汉语(L1)中的每个句子，该系统都通过两个步骤进行训练。去噪：如图 6(a)所示，其优化了用共享编码器对句子的噪声编码的概率，并用 L1 解码器重建它；回译：在推理模式(inference mode)下翻译该句子(使用共享编码器编码该句子，如图 6(b)中，使用越南语(L2)解码器进行解码)，然后利用共享编码器优化对译文句子进行编码和使用 L1 解码器恢复源句子的概率。交替执行这两个步骤对 L1 和 L2 进行训练，对 L2 的训练步骤和 L1 类似，如图 6(c)、图 6(d)所示。神经机器翻译系统通常用平行语料库进行训练，由于本例只有单语语料库，因此该监督式训练方法在本文的场景中行不通。但使用图 1 的系统架构，能够结合去噪和回译两种方法用无监督的方式训练整个系统。

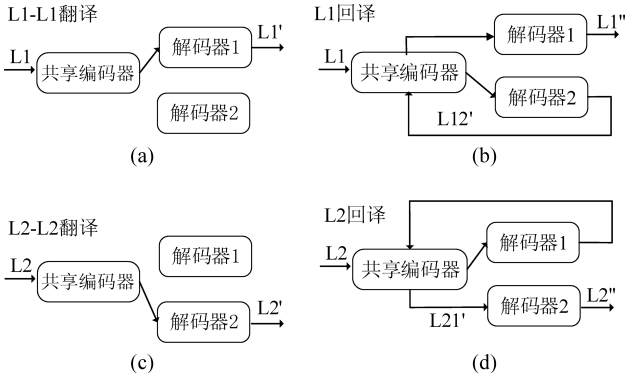

图 6　融合 EMD 最小化双语词典汉越无监督 NMT 模型训练的四个过程

去噪：由于使用了共享编码器，并利用了机器翻译的双重结构，本文的系统可以直接训练来重构输入句子。具体来说，系统使用共享编码器对给定语言的输入句子进行编码，然后使用该语言的解码器重构源句子。由于在共享编码器中使用了预训练的跨语言词嵌入，所以该编码器学习将两种语言的嵌入合称为语言独立的表征，每个解码器都应该学习将这类表征解码成对应的语言。在推理模式下，本文仅用目标语言的解码器替代源语言的解码器，这样系统就可以利用编码器生成的语言来独立表征生成输入文本的译文。

本文在输入句中引入随机噪声。这个想法是利用相同的自动编码器去噪原理，系统经过训练可以重建损坏的输入句子的原始版本。为此，通过在连续单词之间进行随机交换来改变输入句子的单词顺序。对于 N 个元素的序列，进行这种 $N/2$ 个随机交换。这样，该系统需要学习该语言的内部结构以恢复正确的词序。同时，通过阻止系统过分依赖输入序列的词序，可以更好地解释跨语言的实际词序差异。

回译：在系统中加入 Sennrich 等人[23]提出的回译方法。具体地，给定一种语言的输入句子，系统使用贪心解码在推断模式下将其翻译成另一种语言(即利用共享编码器和另一种语言的解码器)。这样，可以获得伪平行句子对，并训练系统从该合成翻译中预测原始句子。

需要注意的是，与使用独立模型一次反向翻译整个语料库的标准反向翻译相反，利用所提出的体系结构的双重结构，使用正在训练的模型即时反向翻译每个小批量句子。这样，随着训练的进行和模型的改进，它将通过反向翻译产生更好的合成句子

对,这有助于在随后的迭代中进一步改进模型。

在训练过程中,将培训目标从小批量交替到小批量。对语言L1(汉语)和L2(越南语),每次迭代将对L1进行一次小批量去噪,对L2进行另一次去噪,从L1到L2进行一次小批量的即时反向翻译,另一次从L2到L1。此外,在该模型的训练过程中还可以加入小的平行语料库,系统也可以通过组合这些步骤以直接预测该平行语料库中的翻译而以半监督方式训练,就像在标准NMT中一样。

加入平行语料的训练过程如图7所示,对无监督神经机器翻译模型进行监督训练,句子不用加噪声,像正常的监督模型一样训练,汉语到越南语(L1→L2)的翻译就是汉语不加噪声用共享编码器编码,用越南语解码器(解码器2)解码,解码出来的L2′与正确的越南语比较来指导模型训练,反之亦然。

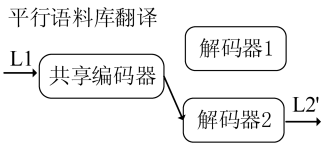

图7 平行语料在融合EMD最小化双语词典UNMT模型中的训练过程

2 实验与分析

2.1 实验数据及设置

从互联网上爬取的语料,经过清洗后,获得汉语单语语料5 800万句,越南语单语语料3 100万句,英语单语语料7 000万句,汉—越平行语料10万句对,英—越平行语料10万句对。对语料预处理,使用undertheseanlp越南语分词工具对越南语进行分词和词性标注,使用jieba分词工具对汉语进行分词和词性标注,使用NLTK工具对英语语料进行词性标注。使用Word2Vec训练单语词嵌入。汉语、越南语和英语分别都训练300维的词嵌入。

使用张檬等人[10]提出的方法训练双语词典,在Artetxe提出的方法上[6]融合EMD最小化的双语词典,将汉语和越南语单语词嵌入映射到共享空间。

本文使用Adam作为优化器,学习率为A=0.0002。在训练期间,使用丢失正则化,dropout为0.3,迭代30万步。

实验使用multi-bleu.perl脚本计算BLEU值作为评价指标。

实验主要分为以下五个部分:无监督基线模型在汉—越和英—越语言对上的翻译、融合EMD最小化双语词典的无监督NMT、在本文方法模型基础上再分别加入1万和10万平行语料、直接使用1万和10万的平行语料在GNMT和Transformer上的有监督模型训练。

2.2 实验结果

2.2.1 基于EMD最小化方法获取双语词典

使用Zhang等人[10]提出的方法训练双语词典,汉语、越南语和英语分别使用Word2Vec训练50维的词嵌入。对词出现的频率设置为大于等于1 000名词,实验结果如表1所示。

表1 基于EMD最小化的汉越双语词典生成数量表

数据/万	汉语 5 800	越南语 3 100	英语 7 000
词的数量	588 368	339 852	712 586
词频大于等于1 000的名词	4 563	3 721	6 437
生成词典数量/对	汉—越 2 722	英—越 2 953	

2.2.2 汉越机器翻译实验

无监督模型训练:使用单语语料训练翻译模型,无监督基线是应用基准模型训练汉—越和英—越的无监督翻译模型。本文的无监督方法是在基准实验上,融合EMD最小化双语词典的汉—越和英—越无监督NMT。

半监督模型训练:在本文提出方法的基础上再加1万和再加10万个平行句对进行实验。

监督模型训练:用上述半监督实验中加入的1万和10万个平行句对训练有监督GNMT和Transformer模型。

不同方法汉越机器翻译实验对比结果如表2所示。

从无监督基线和本文的无监督方法结果可以看出,本文的无监督方法,较基线系统在越—英方向上有2.19个BLEU值的提升。因为本文的方法能从单语语料中捕捉到更多跨语言信息,提升双语词嵌入的质量,从而进一步提升了翻译质量。在本文的无监督方法上加入1万的平行语料,汉—越达到10.02个BLEU值,越—汉达到了13.91个BLEU值,对比监督GNMT和Transformer,本文的半监督方法远高于用只有1万句对平行语料直接训练监督模型的效果。本文的半监督方法(加10万平行语

表 2　不同方法汉越机器翻译实验对比结果

训练方式	使用语料	汉—越	越—汉	英—越	越—英
无监督基线	仅使用单语语料	5.86	9.56	7.25	10.42
本文的无监督方法	仅使用单语语料	8.32	12.30	9.83	12.61
本文的半监督方法	加 1 万平行语料	10.02	13.91	11.29	13.76
	加 10 万平行语料	14.73	16.59	15.87	17.37
监督 GNMT	用 1 万平行语料	0.91	1.40	1.33	1.28
	用 10 万平行语料	9.86	11.91	11.15	13.01
监督 Transformer	用 1 万平行语料	1.02	1.07	1.25	1.35
	用 10 万平行语料	13.58	15.34	14.81	16.13

料)和监督 GNMT 以及监督 Transformer(用 10 万平行语料)的对比中可以看出,加入 10 万平行句对的时候,汉—越和越—汉两个方向均超过 Transformer 模型。实验结果进一步证明了,英—越方向上对比无监督基线和本文的无监督方法本文方法较基线系统模型有所提升。不同方法汉越无监督机器翻译实例分析如表所 3 所示。

从实验译文结果来看,虽然模型还存在学习偏差导致的翻译不准确的问题,但是本文方法的译文质量较基线系统有明显的提升。

表 3　不同方法汉越无监督机器翻译实例分析

源语言	译文	基线系统	本文方法译文
我们要去哪里?	Chúng ta sẽ đưa nó đến đâu ?	Chúng ta đi đâu?	Chúng ta đang đi đâu?
我们只看到了 DNA 的冰山一角。	Chúng ta chỉ mới thấy đỉnh của tảng bang DNA.	Chúng ta thấy phần nổi của tảng băng DNA.	Chúng tôi chỉ nhìn vào phần nổi của tảng băng DNA.
你我之间没有什么旧恨新仇,我说你几句完全是为你好,希望不要见怪。	Giữa tôi và anh chẳng có gì, tôi nói anh vài lời hoàn toàn là vì anh, hy vọng đừng chê bai.	Không có gì giữa bạn và tôi, tôi nói bạn là vì lợi ích của bạn, tôi hy vọng không ngạc nhiên.	Không có mối hận thù cũ và hận thù mới giữa bạn và tôi, tôi đã nói rằng bạn hoàn toàn vì lợi ích của bạn, tôi hy vọng bạn không đổ lỗi cho tôi.

3　总结

本文提出的融合 EMD 最小化双语词典的汉—越无监督神经机器翻译,是在语言差异性大、没有同源词可用的汉—越语言上,只使用汉语和越南语单语语料,利用 EMD 最小化的方法训练双语词典,在共享编码器的无监督翻译模型的基础上融合了 EMD 最小化的双语词典。本文方法在基线模型基础上,提高了从差异性较大单语语料中挖掘跨语言信息的质量,进而提升了无监督模型的翻译质量。实验结果显示,本文提出的方法对比基线系统在翻译质量上得到了有效提升。下一步的工作将探索在本文基础上融合句法信息,进一步提高汉越无监督神经机器翻译的质量。

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv: 1409.0473, 2014.

[2] Ilya Sutskever, Oriol Vinyals, Quoc V Le. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014: 3104-3112.

[3] Rico Sennrich, Barry Haddow, Alexandra Birch. Edinburgh neural machine translation systems for WMT16 [C]//Proceedings of the 1st Conference on Machine Translation, 2016: 371-376.

[4] Philipp Koehn, Rebecca Knowles. Six challenges for neural machine translation[C] //Proceedings of the 1st Workshop on Neural Machine Translation, Vancouver, 2017:28-39.

[5] Guillaume Lample, Ludovic Denoyer, Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only[J]. arXiv preprint arXiv: 1711.00043, 2017.

[6] Mikel Artetxe, Gorka Labaka, Eneko Agirre, et al. Unsu-pervised neural machine translation[J]. arXiv preprint arXiv:1710.11041, 2017.

[7] Zhen Yang, Wei Chen, Feng Wang, et al. Unsupervised neural machine translation with weight sharing

[J]. arXiv preprint arXiv:1804.09057, 2018.
[8] Guillaume Lample, Myle Ott, Alexis Conneau, et al. Phrase-based and neural unsupervised machine translation[C] //Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 5039-5049.
[9] Guillaume Lample, Alexis Conneau. Cross-lingual language model pretraining[J]. arXiv preprint arXiv: 1901.07291, 2019.
[10] Meng Zhang, Yang Liu, Huanbo Luan, et al. Earth mover's distance minimization for unsupervised bilingual lexicon induction [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, 2017:1934-1945.
[11] MinhThang Luong, Hieu Pham, Christopher D Manning. Effective approaches to attention-based neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015: 1412-1421.
[12] Di He, Yingce Xia, Tao Qin, et al. Dual learning for machine translation[C]//Proceedings of the International Conference on Neural Information Processing Systems, 2016: 820-828.
[13] Orhan Firat, Kyunghyun Cho, Yoshua Bengio. Multiway, multilingual neural machine translation with a shared attention mechanism[C]//Proceedings of the Association for Computational Linguistics, San Diego, 2016: 866-875.
[14] Thanh-Le Ha, Jan Niehues, Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder [J]. arXiv preprint arXiv:1611.04798, 2016.
[15] Jason Lee, Kyunghyun Cho, Thomas Hofmann. Fully characterlevel neural machine translation without explicit segmentation[C]//Proceedings of the Association for Computational Linguistics, 2017: 365-378.
[16] Melvin Johnson, Mike Schuster, Quoc V Le, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation[J]. Transactions of the Association for Computational Linguistics, 2016, 5: 339-351.
[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, 2013: 3111-3119.
[18] Tomas Mikolov, Quoc V. Le, Ilya Sutskever. Exploiting similarities among languages for machine translation[J]. arXiv preprint arXiv:1309.4168, 2013.
[19] Ian J Goodfellow, Jean Pouget Abadie, Mehdi Mirza, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, 2014: 2672-2680.
[20] Zhang Meng, Liu Yang, Luan Huanbo, et al. Adversarial training for unsupervised bilingual lexicon induction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017: 1959-1970.
[21] Angeliki Lazaridou, Georgiana Dinu, Marco Baroni. Hubness and pollution: Delving into crossspace mapping for zeroshot learning[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, 2015: 270-280.
[22] Georgiana Dinu, Angeliki Lazaridou, Marco Baroni. Improving zeroshot learning by mitigating the hubness problem [J]. Computer Science, 2014, 9284: 135-151.
[23] Zhang Meng, Liu Yang, Luan Huanbo, et al. Building Earth Mover's Distance on bilingual word embeddings for machine translation[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016: 2870-2876.
[24] Zhang Meng, Liu Yang, Luan Huanbo, et al. Inducing bilingual lexica from nonparallel data with Earth Mover's Distance regularization[C]//Proceedings of the 26th International Conference on Computational Linguistics, Osaka, 2016: 3188-3198.
[25] Rico Sennrich, Barry Haddow, Alexandra Birch. Improving neural machine translation models with monolingual data[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 86-96.

薛明亚(1992—),硕士研究生,主要研究领域为机器翻译。
E-mail: xuemy1227@outlook.com

文永华(1979—),博士研究生,讲师,主要研究领域为机器翻译。
E-mail: wyh194@163.com

余正涛(1970—),通信作者,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。
E-mail: ztyu@hotmail.com