低资源跨语言嵌入初始化的神经机器翻译方法

韩越[1,2]，宜年[1,2]，艾山•吾买尔*[1,2]，汪烈军[1,2]，
刘胜全[1,2]，吐尔根•依布拉音[1,2]
（1. 新疆大学信息科学与工程学院，新疆维吾尔自治区，乌鲁木齐 830046；
2. 新疆多语种信息技术重点实验室，新疆维吾尔自治区，乌鲁木齐 830046；）

摘要：低资源下，用预训练的词嵌入初始化端到端模型的编码器和解码器嵌入层是神经机器翻译的实用技巧。通用的做法是在大规模单语数据上训练词嵌入并用其初始化 RNN 神经机器翻译模型的嵌入层。然而，现有的机器翻译模型通常采用基于 BPE 切分的子词作为输入，并使用 Transformer 做为翻译模型。因此，该文首先比较了基于子词嵌入的初始化方法在不同翻译模型上的性能，并针对在 Transformer 上预训练子词嵌入初始化方法表现不佳的问题，提出使用跨语言嵌入初始化翻译模型嵌入层的方法。提出的方法在 5 个语言对 10 个方向上平均提升 0.78 个 BLEU 值，最高提升 1.19，最低 0.39，并在多语言场景下也观察到 0.37 个 BLEU 值的提升。

关键词：子词嵌入；跨语言嵌入；机器翻译；预训练；Transformer

中图分类号：TP391 **文献标识码：**A

0 前言

近年来，神经机器翻译模型都使用句子级对齐的平行语料进行训练，这在大规模平行语料上显示出卓越的性能。然而在低资源上，模型因为无法获得单词良好的语义表示，翻译性能不佳。为解决这一问题，学者们试图利用更容易获得的大规模单语语料来获取大量的语言学知识，将其融入机器翻译。例如：使用在源端和目标端单语语料库上训练的语言模型来初始化神经机器翻译模型的编码器和解码器[1]。在大规模单语语料库上训练句子的上下文表示，并将其融入神经机器翻译模型[2]。在大规模单语语料库上训练任务无关的词嵌入，并初始化神经机器翻译模型编码器和解码器的嵌入层[3-5]。

使用预训练词嵌入的方法无疑为单词引入了更好的表示，从而提高机器翻译的质量。现有的预训练词嵌入初始化神经机器翻译模型嵌入层的方法均是使用单词级别的词嵌入初始化基于 RNN[6]翻译模型的编码器和解码器嵌入层。然而当前的神经机器翻译系统为缓解 OOV 问题，大都使用 BPE 切分的子词作为输入，并且使用最先进的 Transformer[7]模型作为神经机器翻译模型。因此，实验首先对用于机器翻译模型训练的平行语料进行 BPE 切分，在此基础上训练子词嵌入，使用该嵌入初始化 RNN 翻译模型和 Transformer 翻译模型编码器、解码器的嵌入层，观察平行语料上训练的子词嵌入初始化翻译模型嵌入层的方法对不同翻译模型带来的影响。最后，针对预训练子词嵌入初始化 Transformer 翻译模型嵌入层性能不佳的问题提出使用跨语言子词嵌入初始化神经机器翻译模型嵌入层。

通常有两种获得跨语言词嵌入的方法：一种是使用平行语料、可比语料或种子词典作为监督信号的监督方法[8-11]，另一种是无需使用任何词典或平行语料的无监督方法[12-13]。本文使用公开的 Vecmap[1]进行有监督[11]和无监督[13]的跨语言词嵌入学习，并初始化神经机器翻译模型编码器、解码器嵌入层。跨语言词嵌入初始化方法相比于基线模型在 5 个语言对 10 个方向上带来了平均 0.78 个 BLEU 的提升，最高带来了 1.19 个 BLEU 的提升，并在多语言场景中观察到了可比的结果。通过对实验结果进行分析，跨语言嵌入初始化翻译模型嵌入层的方法在机器翻译解码的过程中提高了源语言和目标语言对齐单词之间的注意力得分。

本文的主要贡献有以下几个方面：

(1)探索了基于平行语料训练的子词嵌入初始化翻译模型嵌入层的方法对不同翻译模型的性能影响。

(2)比较了多个单语词嵌入算法训练的单语

[1] https://github.com/artetxem/vecmap

基金项目：国家语委科研项目（ZDI135-54），国家自然科学基金项目(No. 61662077），国家重点研发计 (No.2017YFB1002100)，国家语委重点项目(ZDI135-54)，国家自然科学基金项目(61662077)，自治区“天山创新团队计划”申报书(编号:202101642)

嵌入用于初始化翻译模型嵌入层的适用性。

(3)针对子词嵌入初始化翻译模型嵌入层的方法在 Transformer 翻译模型上性能不佳的问题，提出使用跨语言子词嵌入初始化翻译模型嵌入层的方法。

(4)比较了监督和无监督方法训练的跨语言子词嵌入对 Transformer 翻译模型的性能影响。

(5)在多个语言上，以及多语言场景下进行实验，证明了跨语言子词嵌入初始化翻译模型嵌入层方法的可靠性。

本文的剩余内容组织结构如下：第 1 节介绍了与本文研究内容相关的工作。第 2 节介绍了本文提出的跨语言子词嵌入初始化翻译模型的方法。第 3 节介绍了实验所用数据，实验中所用模型的参数配置，以及实验方法及结果分析。第 4 节对本文的工作进行总结，并提出下一步的研究方向。

1 相关工作

预训练词嵌入由于能够利用大规模未标记语料库，在序列标注[14]，文本分类[15]，机器翻译[3-5]等自然语言处理任务中都扮演着重要角色。在机器翻译中，通常的做法是基于大规模单语语料训练词嵌入，用该嵌入初始化神经机器翻译模型的编码器和解码器嵌入层从而提高机器翻译的性能。

2016 年 Venugopalan 等人[16]提出使用 Glove 算法在单语语料库上训练词嵌入并初始化解码器的嵌入层。2017 年 Neishi 等人[3]将平行语料库上训练的单语词嵌入初始化翻译模型编码器和解码器的嵌入层改善原始的机器翻译模型性能。2017 年 Gangi 等人[4]提出在源语言的单语数据上训练单词嵌入，并提出三种不同的策略将该词嵌入融入到翻译模型。该方法在低资源场景下表现出了巨大潜力。2018 年 Qi 等人[17]分析了什么时候以及为什么预训练词嵌入对神经机器翻译模型有用，并且尝试了与本文类似的对词嵌入进行先验对齐的方法，然而他们并没有观察到有益的结果。2019 年 Li 等人[5]比较了不同粒度级别上训练的词嵌入对神经机器翻译的影响，并表明更细粒度的嵌入有助于提高翻译模型的性能。

虽然以上预训练嵌入初始化方法提升了翻译模型的性能，但是源语言和目标语言的嵌入之间缺乏交互性，降低了不同语言上语义相近的单词的注意力得分，导致了不正确的翻译，损害了翻译性能。先前的研究中将先验的词对齐信息融入到注意力机制，以获得更准确的翻译结果[18-19]，或将源和目标嵌入连接起来，从而更好的关注相关的源和目标词[20]。受此启发，提出使用跨语言子词嵌入初始化机器翻译模型编码器和解码器的嵌入层。

我们使用公开的 Vecmap 训练跨语言嵌入，Vecmap 既可以使用监督的方式又可以使用无监督的方式训练跨语言嵌入。

监督方法学习跨语言词嵌入：给定 X 和 Z 以及种子词典 D，X 和 Z 分别表示给定词典的两种语言的词嵌入矩阵。X_{m*} 和 Z_{m*} 分别表示词嵌入矩阵的第 m 行，对应词典中的第 m 项。基于监督的方法旨在寻找一个映射矩阵 W,使得 XW 接近 Z，目标函数如式(1)：

$$\arg\min_{W} \sum_i \|X_{m*}W - Z_{m*}\|^2 \quad (1)$$

无监督方法学习跨语言词嵌入：给定词汇表中所有单词的相似度矩阵，每个单词都有不同的相似度分布。而在不同的语言中，语义上等价的单词应该具有相似的分布。根据这一观察，Vecmap 等人归纳出原始的单词对集合。根据单词对集合，学习词嵌入之间的映射关系，并利用这一映射关系，获得更多的单词对，进一步学习映射关系。通过将这种初始化词典方案和自学习方法相结合无监督的学习跨语言词嵌入。自学习伪代码如表 1 所示。

表 1 自学习方法伪代码

Tab1 Pseudocode of self-learning methods

自学习算法
INPUT：X（源语言嵌入）
INPUT：Y（目标语言嵌入）
INPUT：D（种子词典）
Repeate LEARN_MAPPING(X,Y,D) ⟶ W LEARN_DICTIONARY(X,Y,W) ⟶ D Until Converge

2 基于跨语言嵌入的神经机器翻译方法

跨语言词嵌入是指不同语言的词嵌入处于同一个词嵌入空间中，使得语义相同却来自不同语言的词嵌入具有相似的向量表示。这显然有助于基于 Transformer 的解码器计算“编码器-解码器注意力”。因此，本文提出使用跨语言子词嵌入初始化翻译模型嵌入层，使其更好的学习源和目标之间的对应关系。

使用跨语言嵌入初始化神经机器翻译模型嵌入层的基本流程如图 1 所示：

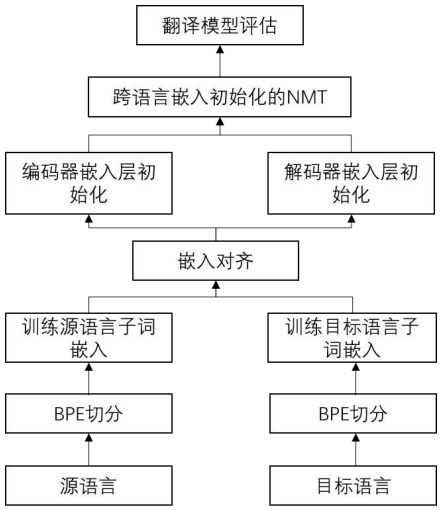

图 1 跨语言嵌入初始化神经机器翻译流程图

Fig1 Flow chart of initializing neural machine translation using cross-language embedding

(1)BPE 切分：对平行语料使用 subword-nmt 进行 BPE 切分，将单词 W 切分为子词单元 $(w_1, w_2, \cdots w_n)$,其中 w_1, w_2分别是单词 W 的子词单元。

(2)训练子词嵌入：对 BPE 切分后的源语言和目标语言语言分别训练子词嵌入。为比较不同单语词嵌入算法对翻译模型性能的影响，分别使用 FastText 以及 Word2Vec 训练单语子词嵌入。对于 FastText 和 Word2Vec，分别使用 CBOW 和 Skip-gram 算法。

(3)嵌入对齐：利用跨语言词嵌入训练工具 Vecmap 对训练的源语言子词嵌入和目标语言子词嵌入进行对齐得到跨语言子词嵌入，分别使用监督和无监督的方式进行对齐。

(4)嵌入层初始化：使用对齐后的源语言子词嵌入和对齐后的目标语言子词嵌入分别初始化编码器和解码器的嵌入层。

(5)模型训练：进行跨语言子词嵌入初始化的神经机器翻译模型训练。

(6)翻译模型评估：对跨语言子词嵌入初始化的翻译模型进行评估。评估不同对齐方式的跨语言子词嵌入初始化的翻译模型，以及跨语言子词嵌入初始化的模型在多个语言对上，以及多语言场景下的性能。

其中对于第 3 步，监督的跨语言词嵌入对齐方法需要种子词典。由于实验中是对子词嵌入进行对齐，没有现有的子词对齐的词典。为了获得该词典，使用 FastAlign[2]工具对平行语料进行子词对齐。与实际情况相符，得到的对齐存在“一对一”，“一对多”、“多对一”的现象，即一个源单词对应一个目标单词，一个源单词对应多个目标单词，多个源单词对应一个目标单词的情况。为保证准确率，选择“一对一”的词对构建种子词典。

3 实验

3.1 实验数据及处理

先前的研究表明，预训练词嵌入对低资源下的神经机器翻译效果明显。因此，选择在英语(EN)和捷克语（CS）、德语(DE)、爱沙尼亚语(ET)，维吾尔语(UY)和汉语(ZH)，乌兹别克语（UZ）和汉语（ZH）的低资源上进行实验。并在多语种机器翻译时增加了土耳其语和英语的实验。我们对于不同来源的数据进行实验，以验证该方法的普适性。捷克语-英语的训练数据集来源于 WMT16 新闻机器翻译共享任务可用数据集的 Common Crawl 数据集。使用的测试数据集为 newstest2015，验证集为 newstest2016。爱沙尼亚语-英语的语料来自于 WMT18 中的 Rapid corpus of EU press releases，测试集为 newstest2018，验证集为 newsdev2018。德语-英语的语料来自 IWSLT16。使用的维汉语料来自 CWMT19。土耳其语-英语的数据来源于 WMT17 的可用数据集 SETIMES2，测试集为 newstest2017，验证集为 newsdev2016。乌兹别克语-汉语语料来源于新疆大学多语种实验室自建的数据。实验数据量如表 2 所示：

表 2 实验数据统计结果

Tab2 Statistical results of experimental data

语言对	训练集	测试集	验证集
CS-EN	161838	2999	2656
ET-EN	226978	2000	2000
DE-EN	196884	2298	2052
UY-ZH	164316	1000	1000
UZ-ZH	176009	1000	1000
TR-EN	204936	3007	1001

对数据进行预处理是实验的关键步骤，因此本文针对不同的语言进行了仔细的处理，关键步

[2] https://github.com/clab/fast_align

骤如下：

(1)去除语料中的 HTML 标签。

(2)去除重复的句子，及源和目标相同的句子。

(3)使用 Moses tokenizer 工具对英语(EN)、德语(DE)、爱沙尼亚语(ET)、土耳其语（TR）、乌兹别克语（UZ）进行分词。

(4)使用新疆大学多语种实验室研发的编码转换工具对汉语进行全角半角转换、繁体简体转换、乱码过滤。

(5)利用哈尔滨工业大学开源的 pyltp 对中文语料进行分词处理，使用自主研发的维吾尔语分词工具对维吾尔语进行分词。

(6)最后，使用 Subword-nmt 对所有的句子进行 BPE 切分从而使用子词单元，由于不同的 BPE 切分对翻译性能影响较大，因此使用 BPE 切分时分别采用 8K 和 16K 的合并操作。

以 DE-EN 为例，展示在 BPE 合并操作次数分别为 8K、16K 时的实验数据，实验数据如表 3 所示：

表 3 不同合并操作次数下的 BPE 切分示例

Tab3 Example of BPE partitioning with different merge operations

Lang	BPE	示例
DE	8K	Der Wasser@@ spie@@ gel des Se@@ es begann zu sin@@ ken .
	16K	Der Wasser@@ spiegel des Se@@ es begann zu sinken .
EN	8K	The water level of the la@@ ke has started dr@@ ying up .
	16K	The water level of the lake has started dr@@ ying up .

从以上示例可以看出相对于 16k 的合并次数，使用 8K 的合并次数对单词进行切分可以得到更细粒度的子词单元。在德语的 8K 切分时，将单词“Wasserspiegel”切分成三部分“Wasser”、“spie”和“gel”。而在 16k 切分时，只切分成了两部分“Wasser”和“spiegel”。同样的在英语上，16K 的切分并没有将单词“lake”切分开，而 8K 切分成了两个子词单元“la”和“ke”。

3.2 实验参数设置

实验中使用用于机器翻译的平行语料训练子词嵌入，并将获得的子词嵌入初始化 LSTM 和 Transformer 翻译模型，比较子词嵌入初始化方法对不同翻译模型的影响。使用的翻译模型为 FairSeq[3]中集成的 LSTM 和 Transformer 模型。实验的基线是不使用任何嵌入方法进行初始化的翻译模型。使用的参数如下：

Transformer 模型: Transformer 中的编码器和解码器的层数为 6 层，每一层有 8 个注意力头，词向量维度为 512，全连接隐藏层状态维度为 2048，使用 adam 优化器更新模型参数，初始学习率为 0.0007，将在残差连接，注意力机制，前馈层使用的 dropout 设置为 0.3，使用 Label smoothing=0.1，使用 gleu 激活函数，一个批次中有接近 4096 个词。

LSTM 模型：LSTM 中采用 1 层双向编码器和一层的解码器，除了一些包含词向量维度、优化器、学习率，dropout 等的关键参数和 Transformer 保持一致，其余均使用默认参数。

对所有的模型迭代训练训练 40 轮，对最后 20 轮求最大值。采用 Beam Search 策略来进行预测，beam 大小均为 4，解码时长度惩罚因子为 0。

实验中，训练单语词嵌入的 FastText 以及 Word2Vec 采用相同的参数设置，词嵌入维度大小设置为 512 维，窗口大小设置为 10，训练迭代次数为 15。

实验中使用跨语言词嵌入工具 Vecmap 对单语词嵌入进行对齐。所使用的 Vecmap 工具均使用默认参数。对于监督的训练方法需要提供词典，该词典通过使用无监督词对齐工具 FastAlign 进行词典抽取，词典的大小设置为 16000。

3.3 实验结果与分析

3.3.1 预训练子词嵌入初始化不同的翻译模型

先前初始化翻译模型嵌入层的方法均是使用单词级别的嵌入，然而随着机器翻译的发展，人们常常使用 BPE 切分的子词单元作为机器翻译模型的输入。因此，实验以 DE-EN 为例，仅使用平行语料库训练不同 BPE 切分的源语言和目标语言子词嵌入，并在基于 LSTM 和 Transformer 的机器翻译模型上分别进行实验，以探讨子词嵌入初始化翻译模型嵌入层方法对不同翻译模型的影响。实验结果如表 4，表 5 所示：

表 4 BPE 合并次数为 8K 时的翻译结果

Tab4 The translation results of BPE merging times of 8K

DE-EN		LSTM	Transformer
Baseline		32.73	34.59
Skip-gram	FastText	**33.15**	33.85
	Word2Vec	**33.17**	33.99
CBOW	FastText	30.97	19.41
	Word2Vec	30.80	18.77

表 5 BPE 合并次数为 16K 时的翻译结果

Tab5 The translation results of BPE merging times of 16K

DE-EN		LSTM	Transformer
Baseline		31.26	32.77
Skip-gram	FastText	**31.73**	31.62
	Word2Vec	**31.96**	31.47
CBOW	FastText	30.57	19.32
	Word2Vec	30.70	20.12

[3] https://github.com/pytorch/fairseq

通过表 4 表 5 可以观察到 BPE 的合并操作次数对机器翻译模型的性能有较大的影响，并且对 Transformer 的影响更为显著。在 Transformer 上 8k 和 16k 上 Baseline 的 BLEU 值相差 2.18，在 LSTM 上相差 1.47。同时可以发现 LSTM 模型与 Transformer 模型相比在性能上确有一定的差距，在 BPE 的合并操作次数为 8K 时，LSTM 和 Transformer 的 BLEU 值相差 2.2。重要的是，无论 BPE 的合并操作次数是多少, 当使用 LSTM 模型时, Skip-gram 算法训练的嵌入总是能够提升机器翻译的性能，CBOW 算法训练的嵌入降低了翻译性能。但当使用预训练的子词嵌入初始化 Transformer 模型时，无论使用哪种子词嵌入的训练算法，均未对机器翻译模型带来性能上的提升。并且在使用 CBOW 训练的子词嵌入初始化时，对翻译性能有巨大的损害。这可能是因为语料规模较少，该方法无法得到良好的子词嵌入表示，而良好的嵌入表示对 Transformer 影响巨大。该实验表明在 LSTM 翻译模型上表现良好的初始化方法并没有在 Transformer 上带来有益的提升。

3.3.2 跨语言嵌入初始化 Transformer

由于直接使用预训练的子词嵌入初始化 Transformer 翻译模型嵌入层非但无法提升性能，反倒会有损模型的训练。提出使用跨语言嵌入的学习方法对预训练的源语言和目标语言子词嵌入进行对齐，试图在机器翻译的解码过程中提升对齐单词的注意力，从而提高机器翻译的性能。为探讨提出的方法对当前最有效的机器翻译模型的影响，该实验采用的机器翻译模型为 Transformer，使用的 BPE 合并操作次数为 8K，所选用的子词嵌入训练算法为 Skip-gram, 并在多个翻译任务上进行了实验以验证提出方法的有效性。实验结果如表 6、表 7 所示：

表 6 跨语言嵌入初始化 Transformer 结果

Tab6 The result of initializing Transformer with cross-language embedding

	CS-EN	DE-EN	ET-EN	UY-ZH	UZ-ZH
Baseline	17.27	34.59	15.10	28.39	36.52
FT_pre	13.72	33.85	13.19	27.72	34.06
W2V_pre	12.89	33.99	12.97	26.76	34.26
FT_un	17.57	35.13	15.91	29.16	37.08
W2V_un	17.72	35.1	15.53	**29.27**	**37.48**
FT_su	17.69	35.18	**15.95**	29.06	37.42
W2V_su	**18.46**	**35.42**	15.84	29.20	37.12
Δ	1.19	0.83	0.85	0.88	0.96

表 7 跨语言嵌入初始化 Transformer 结果

Tab7 The result of initializing Transformer with cross-language embedding

	EN-CS	EN-DE	EN-ET	ZH-UY	ZH-UZ
Baseline	16.13	30.18	13.95	30.39	46.18
FT_pre	12.19	29.12	11.46	28.35	39.48
W2V_pre	11.92	29.27	11.39	28.34	41.45
FT_un	16.37	30.64	**14.34**	30.84	46.18
W2V_un	13.93	30.65	13.95	**30.99**	46.76
FT_su	**16.64**	**30.92**	14.19	30.93	**47.03**
W2V_su	15.38	30.91	14.14	30.80	46.82
Δ	0.51	0.74	0.39	0.6	0.85

实验中用 FT 表示使用 FastText 训练的嵌入，W2V 表示 Word2Vec 训练的嵌入。*_pre 表示的是使用预训练子词嵌入直接进行初始化的结果。*_un 表示使用无监督方法对子词嵌入进行对齐后的初始化。*_su 表示使用监督方法对子词嵌入进行对齐后的初始化。最后一行表示相对于基线模型，使用初始化方法带来的收益。

以上在 5 个语言对 10 个方向上的实验结果一致表明，无论是使用融入了子词信息的 FastText 还是 Word2Vec,在平行语料库上训练的子词嵌入直接初始化 Transformer 翻译模型的嵌入层总是会降低翻译模型的性能。然而对齐后的子词嵌入相对于未对齐的子词嵌入对机器翻译模型带来的收益在 2-5 个 BLEU。并且对于基线模型仍有明显的提高，平均提高了 0.78 个 BLEU 值。最高可达 1.2 个 BLEU 值，在 EN-ET 上提升最小，提升了 0.39 个 BLEU 值。同时可以发现，对子词嵌入使用监督的对齐方式和无监督的对齐方式带来的差距并不大，这表明无监督方法在平行语料上进行子词嵌入对齐的鲁棒性。比较使用 FastText 与 Word2Vec 训练的子词嵌入进行初始化的结果，发现相比于 Word2Vec,融入了字符信息的 FastText 并没有显示出特别的优势，这可能是由于子词嵌入的训练是基于子词单元的。因此，实践中无论使用哪种工具训练子词嵌入都是可靠的。

为了验证提出的方法对 BLEU 值的提高是否显著，我们分别将表 6 和表 7 中四种场景下对齐后的向量进行翻译模型嵌入层初始化在不同语言上的实验数据与 Baseline 数据进行了显著性检验。结果如表 8 所示：

表 8 不同词嵌入初始化方法下的 p 值

Tab 8 P value under different word embedding initialization methods

	p
FT_un	0.00022394
W2V_un	0.3738

FT_su	5.86191e-06
W2V_su	0．00996846

通过对四种数据进行显著性检验，根据 p<0.05 可以判定显著性差异存在这一条件，我们可以发现除了使用 Word2Vec 训练的词嵌入对齐后初始化翻译模型嵌入层得到的 BLEU 值与 Baseline 没有显著性差异，其余的方法均符合显著性差异。

由于 3.3.1 中的实验表明无论是 Word2Vec 还是 FastText 词嵌入工具，直接使用 CBOW 算法训练的子词嵌入初始化翻译模型的嵌入层均会损害翻译模型的性能，因此，我们以 DE-EN 和 UY-ZH 翻译任务为例，使用 CBOW 算法训练子词嵌入并进行对齐来初始化翻译模型嵌入层，观察对翻译性能的影响。实验结果如表 9 所示：

表 9 CBOW 算法下跨语言子词嵌入对 Transformer 的影响

Tab 9 The impact of cross-language sub-word embedding on Transformer in the CBOW scenario

	DE-EN	UY-ZH
Baseline	34.59	28.39
FT_pre	19.41	12.75
W2V_pre	18.77	12.21
FT_un	35.32	**29.25**
W2V_un	35.24	28.71
FT_su	35.34	29.18
W2V_su	**35.41**	28.82
Δ	0.82	0.86

实验结果表明尽管 CBOW 算法训练的子词嵌入直接初始化翻译模型嵌入层严重降低了翻译性能。但当使用跨语言词嵌入的映射方法对齐之后再初始化翻译模型嵌入层的结果可以和对 Skip-gram 对齐后的结果相媲美。在 DE-EN 上，对齐后的 Skip-gram 算法和 CBOW 算法获得的最高 BLEU 分别为 35.42（见表 6）和 35.41。而在 UY-ZH 上的最高 BLEU 值分别 29.27 和 29.25。

为探讨使用 CBOW 算法训练的子词嵌入初始化翻译模型嵌入层性能低下，对齐后的 CBOW 嵌入初始化翻译模型性能显著提升的原因，我们以 DE-EN 为例，在 FastText 词嵌入工具上对 CBOW 算法和 Skip-gram 算法上训练的词嵌入以及无监督对齐前后的嵌入进行了可视化,我们对语料中词频最高的前 200 个进行了可视化。可视化的结果如图 2 所示。图中我们可以直观地看到，在对直接训练的词嵌入进行对齐之前，CBOW 算法训练的两种语言的词嵌入分布相差较大（a）而 Skip-gram 算法训练的两种语言的词嵌入分布差异较小（c）。这可能是导致 CBOW 算法训练的词嵌入初始化翻译模型性能低下的直接原因。图 2 中的 b 图将词嵌入空间明显的分为三部分，可以看到源和目标进行了良好的对齐。而图 d 相比于图 b 对齐效果相对较差，这也与我们的实验结果相对应。在 FastText 词嵌入工具下，使用 CBOW 词嵌入训练算法训练的词嵌入进行无监督对齐后（b）初始化翻译模型嵌入层的 BLEU 值为 35.32，使用 Skip-gram 词嵌入训练算法训练的词嵌入进行无监督对齐后（d）初始化翻译模型嵌入层的 BLEU 值为 35.13。这可能是由于使用 CBOW 算法中从周围词预测中心词的思想使得训练的词嵌入较为集中，而 Skip-gram 算法从中心词预测上下文词的策略使得训练的词嵌入在空间中较为分散。这使得 CBOW 算法更容易学习源语言和目标语言中对应词的映射关系。

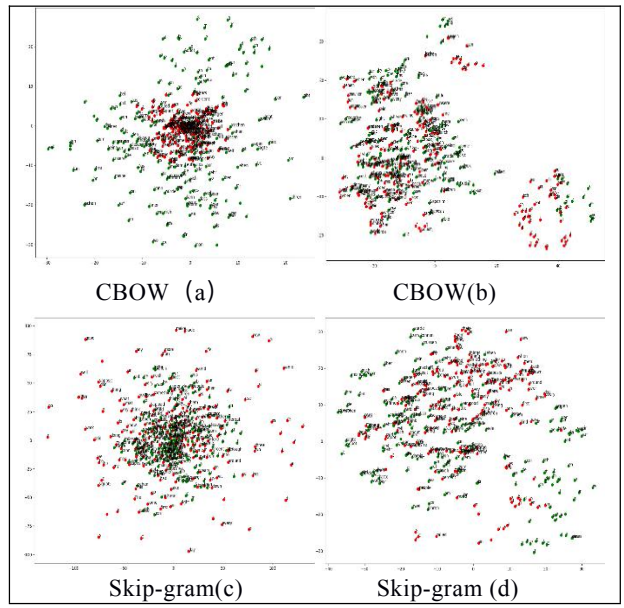

图 2 不同场景下的词嵌入可视化

Fig2 Visualization of word embedding in different scenarios (In the figure, the green dots represent German words, and the red dots represent English words. a and c show the word embedding before alignment, and b and d show the word embedding after alignment.)

在以上实验中，我们研究了将预训练的子词嵌入初始化翻译模型编码器和解码器嵌入层的结果。接下来我们继续以 DE-EN 和 UY-ZH 为例，探讨将解码器嵌入层和 Softmax 输出层参数共享时的情况，并与单独初始化翻译模型编码器嵌入层和解码器嵌入层的结果进行比较，试图对翻译模型的初始化方法进行全面的探究。实验中使用先前 FastText 以 Skip-gram 训练的词嵌入。实验结果如表 10 所示。

表 10 多种初始化翻译模型的方法结果

Tab10 Results of various methods of initializing translation models

	DE-EN	UY-ZH
Baseline	34.59	28.39
Src	33.75	27.09

Tgt	34.86	28.59
Src_Tgt	33.85	27.72
Src_Tgt_Share	34.43	28.62
Src_Tgt_Un_Aligned	35.13	29.16
Src_Tgt_Un_Aligned_Share	**35.89**	**30.05**

表中的 Src 表示只初始化编码器嵌入层，Tgt 表示只初始化解码器嵌入层。Src_Tgt 表示同时初始化编码器和解码器嵌入层。Src_Tgt_Share 表示在上一步的基础上将解码器嵌入层和 Softmax 输出层参数共享。Src_Tgt_Un_Aligned 表示使用无监督对齐后的向量初始化编解码器嵌入层。同样的，Src_Tgt_Un_Aligned_Share 表示在上一步基础上增加了解码器嵌入层和 Softmax 输出层参数的共享。

实验结果表明相比于仅初始化编码器嵌入层和同时初始化编解码器的嵌入层，仅用预训练的子词嵌入初始化解码器嵌入层能带来翻译性能上的更大收益。同时，将解码器嵌入层和 Softmax 输出层参数共享能带来进一步的性能上的显著提高。这一点，在对词嵌入对齐前后都有明显的体现。如在实验中对预训练的词嵌入对齐后初始化翻译模型嵌入层性能已经达到了 35.13 和 29.16 比 Baseline 高出 0.54 和 0.77 个 BLEU 值。但当共享之后,BLEU 分别提高了 1.3 和 1.66 个 BLEU。

3.3.3 跨语言嵌入初始化在多语言场景下的表现

跨语言子词嵌入的对齐对双语机器翻译系统带来了显著收益。对于多语言机器翻译系统，会有怎样的影响？为回答这个问题，以德语、爱沙尼亚语、土耳其语和英语为例进行实验，实验使用一个共享的编码器和解码器。为 DE、ET、TR 分别加上代表语言的标签并拼接在一起。使用该混合的语料训练基线模型以及子词嵌入。实验结果如表 11 所示：

表 11 多语言子词嵌入初始化结果

Tab11 Initialization results of multilingual subword embedding

	DE-EN	ET-EN	TR-EN
Baseline	**36.75**	21.57	22.22
FT_pre	35.75	20.82	21.01
FT_un	36.63	21.91	**22.53**
W2V_pre	35.79	20.66	21.24
W2V_un	36.73	**21.94**	22.23
Δ	-0.02	0.37	0.31

多语言机器翻译相比双语机器翻译，能够提供除了双语语料以外的信息资源，因此，多语言机器翻译的结果相比于双语机器翻译性能有显著的提升。以 ET-EN 为例，在双语机器翻译中,BLEU 值为 15.10，但是在多语机器翻译中，BLEU 值能够达到 21.57。在将词嵌入对齐的跨语言嵌入方法应用在多语言场景时，在 DE-EN 上观察到了 0.02 的下降，在 ET-EN 和 TR-EN 上分别观察到了 0.37.0.31 的提升。实验结果表明，提出的方法在其他语言作为重要资源的多语言机器翻译的场景下，仍能达到竞争性的结果，这进一步证明了提出的方法的有效性。

3.3.4 实验结果分析

以上实验已经证明提出的方法的有效性。为进一步验证我们的猜想：对预训练嵌入的对齐能够在机器翻译的过程中提高源语言和目标语言之间对齐单词的关注度，从而提高机器翻译质量。对实际产生的结果进行分析。从 CS-EN 中随机挑选两个较短的句对进行分析，实验产生的翻译结果如表 12 所示：

表 12 CS-EN 翻译结果示例

Tab12 Examples of CS-EN translation results

1	src	Další nej@@ bliž@@ ší v úrovni nad@@ šení ?
	ref	The next clos@@ est in enthusi@@ as@@ m ?
	Baseline	Another approximately in the above level ?
	Our_method	Another ne@@ a@@ rest in the level of enthusi@@ as@@ m ?
2	src	Ro@@ le , kterou hra@@ jete ve tvor@@ bě zpráv je velmi důležit@@ á .
	ref	The part you play in making the news is very important .
	Baseline	The collection you play in production is very important .
	Our_method	The role you play in creating messages is very important .

在示例 1 中，无论是 Baseline 还是跨语言子词嵌入初始化的方法，对照真实的翻译“The next”，系统均翻译成为了 Another。但对于真实的翻译“closest”，Baseline 将其翻译成为“approximately”，提出的方法将其翻译成为“nearest”。虽然翻译的结果与真实的翻译有所差别，但相比 Baseline 确有强相关性。猜测是由于数据量较小，或在语料中将大多数“nejbližší”翻译成了“nearest”而导致无法学习到该句对中“nejbližší”和“closest”的对应关系。同时，在源语言中的“nadšení”对应目标语言中的单词“enthusiasm”。而这在 Baseline 中却未被正确翻译，提出的方法准确的进行了翻译。

示例 2 中，通过比较 Baseline、提出的方法与

参考译文的不同，可以发现：Baseline 中将“part”翻译成了“collection”，“making the news”翻译成了“production”。而在提出的方法中分别翻译成了“role”和“creating message”。同时对于示例 2，图 3、图 4 展示了 Baseline 系统和跨语言嵌入对齐系统上 Transformer 翻译模型最后一层解码器与编码器在 8 个头上注意力的平均值。可以发现，对于源语言单词“zpráv”,在 Baseline 中对其关注最多的目标语言是单词“is”，这与参考译文中的“news”差距甚大。而在跨语言子词嵌入初始化的系统里是与“news”语义上相近的“message”。虽然整体来说，翻译结果并没有和参考译文保持一致，但同示例一一样表达了和参考译文之间的强相关性，从而进一步验证了我们的猜想。针对翻译结果与参考译文语义相近但不一致的问题，可以试图为机器翻译引入大规模的单语语料，训练子词嵌入，再进行跨语言子词嵌入的训练，最后初始化翻译模型编码器、解码器的嵌入层。

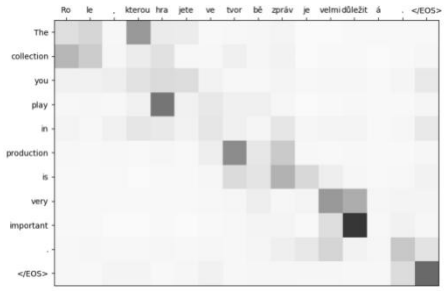

图 3 Baseline 注意力可视化

Fig3 Attention visualization of Baseline

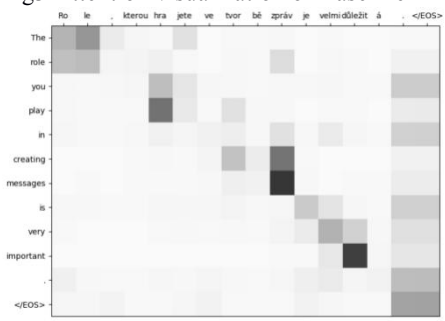

图 4 跨语言嵌入方法注意力可视化

Fig4 Attention visualization of cross-language embedding method

4 总结

使用预训练词嵌入初始化机器翻译模型的方法一度是提高机器翻译性能的实用技巧。然而，随着对单词的子词切分，以及机器翻译模型的迭代，以往的词嵌入初始化方法便不再适用。本文首先以德-英上的实验为例，证明了原始的预训练嵌入初始化方法并不适用于 Transformer 模型，相比于使用预训练的嵌入初始化方法，当前的最佳模型能够学习到更好的表示。基于当前最先进的模型，提出使用跨语言子词嵌入初始化翻译模型嵌入层的方法。该方法试图将不同语言上训练的子词嵌入映射到同一空间，从而缩短源嵌入和目标嵌入之间的距离，提高对齐单词之间的注意力得分。多个翻译任务上的实验证明了该方法的有效性。此外，我们通过实验表明将对齐后的词嵌入初始化翻译模型嵌入层，并将解码器嵌入层和 Softmax 输出层参数共享会进一步提升翻译性能。最后，在多语言机器翻译的场景下，该方法仍然有很大的竞争性。提出的方法虽有时无法翻译出和参考译文一致的结果，但相对于基线系统，能够翻译出与参考译文语义上更相近的结果，可将大规模的单语语料引入单语词嵌入的训练之中，缓解该问题。此外，可对不同的语言采用不同的 BPE 合并操作次数进行切分，从而获得每种语言的最佳切分方式，将前后缀，单词主干等准确切分有助于获得更高质量的对齐。此外，我们将使用最先进的 XLM 来为所有语言预训练大的语言模型，通过生成更好的跨语言表示来拓展这一初始化方法作为后续的研究工作。

参考文献

[1] Ramachandran P, Liu P J, Le Q V. Unsupervised pretraining for sequence to sequence learning[J]. arXiv preprint arXiv:1611.02683, 2016.

[2] Klementiev A, Titov I, Bhattarai B. Inducing crosslingual distributed representations of words[C]//Proceedings of COLING 2012. 2012: 1459-1474.

[3] Neishi M, Sakuma J, Tohda S, et al. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size[C]//Proceedings of the 4th Workshop on Asian Translation (WAT2017). 2017: 99-109.

[4] Di Gangi M A, Federico M. Monolingual embeddings for low resourced neural machine translation[C]//Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT’17). 2017: 97-104.

[5] Li Z, Specia L. A Comparison on Fine-grained Pre-trained Embeddings for the WMT19Chinese-English News Translation Task[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019: 249-256.

[6] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. arXiv preprint arXiv:1409.3215, 2014.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[8] Gouws S, Bengio Y, Corrado G. Bilbowa: Fast bilingual distributed representations without word alignments[C]//International Conference on Machine Learning. PMLR, 2015: 748-756.

[9] Søgaard A, Agić Ž, Alonso H M, et al. Inverted indexing for cross-lingual NLP[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1713-1722.

[10] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation[J]. arXiv preprint arXiv:1309.4168, 2013.

[11] Artetxe M, Labaka G, Agirre E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[12] Conneau A, Lample G, Ranzato M A, et al. Word translation without parallel data[J]. arXiv preprint arXiv:1710.04087, 2017.

[13] Artetxe M, Labaka G, Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings[J]. arXiv preprint arXiv:1805.06297, 2018.

[14] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.

[15] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In In EMNLP. Citeseer.

[16] Venugopalan S, Hendricks L A, Mooney R, et al. Improving lstm-based video description with linguistic knowledge mined from text[J]. arXiv preprint arXiv:1604.01729, 2016.

[17] Qi Y, Sachan D S, Felix M, et al. When and why are pre-trained word embeddings useful for neural machine translation?[J]. arXiv preprint arXiv:1804.06323, 2018.

[18] Mi H, Wang Z, Ittycheriah A. Supervised attentions for neural machine translation[J]. arXiv preprint arXiv:1608.00112, 2016.

[19] Liu L, Utiyama M, Finch A, et al. Neural machine translation with supervised attention[J]. arXiv preprint arXiv:1609.04186, 2016.

[20] K Kuang S, Li J, Branco A, et al. Attention focusing for neural machine translation by bridging source and target embeddings[J]. arXiv preprint arXiv:1711.05380, 2017.

Neural machine translation method of cross language embedding initialization under low-resource

HanYue[1,2],YI Nian[1,2] AISHAN Wumaier*[1,2],WANG lie jun[1,2],
LIU Sheng-quan[1,2],Turgun Ibrayim[1,2]

(1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046;

2. Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi 830046)

Abstract : Using pre-trained word embeddings to initialize the encoder and decoder embedding layers of the end-to-end model is a practical technique for neural machine translation under low-resource.The general method is to train the word embedding on large-scale monolingual data and use it to initialize the RNN neural machine translation model. However, the existing machine translation models usually take subwords based on BPE segmentation as input and use Transformer as translation models. Therefore, this paper first compares the performance of initialization methods based on subword embedding in different translation models. Aiming at the poor performance of the pre-training initialization method based on subword embedding in Transformer, this paper proposes an approach to initialize the translation model embedding layer using cross-language embedding. The proposed method improves the average BLEU value by 0.78 in 10 directions for 5 languages, with the highest increase of 1.19 and the lowest increase of 0.39, and also observed an increase of 0.37 BLEU in the multi-language scenario.

Keywords: Subword Embedding; Cross-language embedding; Machine translation; Pre-training; Transformer

韩越（1997-），女，陕西咸阳人，硕士研究生，研究方向为自然语言处理及机器翻译，E-mail: xjdxhydyx@163.com

宜年（1992-），男，河南焦作人，硕士研究生，研究方向为自然语言处理及机器翻译，E-mail: 15709918429@163.com

艾山·吾买尔（1981-），通讯作者，男，新疆乌鲁木齐人，博士，教授，博导，研究方向为自然语言处理、机器翻译等，E-mail: hasan1479@xju.edu.cn

汪烈军（1975-）男，汉族，中国共产党党员，博士，教授，院长，E-mail:wljxju@xju.edu.cn

刘胜全（1963-），硕士，教授，研究领域为智能信息处理，语义 web，计算机网络安全，E-mail：liu@xju.edu.cn

吐尔根·依布拉音(1958-),男，教授，博士生导师，研究方向为自然语言处理及计算机应用，E-mail：turgun@xju.edu.cn