基于自注意力机制预训练跨蒙汉语言模型

苏依拉 高 芬 牛向华 仁庆道尔吉
(内蒙古工业大学信息工程学院 内蒙古 呼和浩特 010080)

摘 要 针对蒙汉机器翻译中平行语料资源稀缺的问题,提出利用单语语料库对蒙汉机器翻译进行研究。由于利用单语语料库进行机器翻译的效果较差,故将基于自注意力机制预训练跨蒙汉语言模型应用于基于单语语料库训练的蒙汉机器翻译系统中。实验结果表明,基于自注意力机制预训练跨蒙汉语言模型的方法极大改善了蒙汉机器翻译系统的性能。

关键词 蒙汉机器翻译 单语训练 自注意力机制 预训练 语言模型

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.02.028

PRE-TRAINING CROSS MONGOLIAN-CHINESE LANGUAGE MODEL BASED ON SELF-ATTENTION MECHANISM

Su Yila Gao Fen Niu Xianghua Ren Qingdaoerji
(*College of Information Engineering, Inner Mongolia University of Technology, Hohhot 010080, Inner Mongolia, China*)

Abstract Aiming at the scarcity of parallel corpus resources in Mongolian-Chinese machine translation, this paper proposes to use monolingual corpus to study Mongolian and Chinese machine translation. The translation effect of using the monolingual corpus for machine translation is relatively ineffective. Therefore, the pre-training cross Mongolian and Chinese language model based on self-attention mechanism was applied to the Mongolian-Chinese machine translation system based on monolingual corpus training. Through experimental comparison, the method of pre-training cross Mongolian-Chinese language model based on self-attention mechanism greatly improves the performance of the Mongolian-Chinese machine translation system.

Keywords Mongolian-Chinese machine translation Monolingual corpus training Self-attention mechanism Pre-training Language model

0 引 言

机器翻译是人工智能领域的重要研究课题之一,主要目标是研究如何使用计算机实现一种自然语言到另一种自然语言的自动转换[1]。针对蒙汉机器翻译中平行语料资源稀缺的问题,提出利用单语语料[2-8]训练来提高蒙汉机器翻译的质量。而实现基于单语语料库训练的蒙汉机器翻译系统最重要的基础就是需要一个翻译性能良好的初始化翻译模型,而初始化的翻译模型是基于字典的翻译系统和语言模型相结合才能实现的,故预训练一个表现较好的语言模型就显得尤为重要。

语言模型这个概念最早出现在统计机器翻译中,是统计机器翻译技术中较为重要的概率模型之一。通俗而言,语言模型就是一串词序列或者一句话的概率分布,其作用就是可以为一个长度为 n 的句子计算其存在的可能性的概率分布 p。在统计机器翻译中常用的是 N-gram 语言模型,N 表示 n 元模型,当 n 选取较大时,就会产生数据稀疏问题,从而导致估算结果出现较大的偏差。因此,实际中最常用的是三元模型,也就是 Tri-gram 模型。为了缓解 N-gram 语言模型导致的

收稿日期:2019-05-24。国家自然科学基金项目(61363052,61502255);内蒙古自治区自然科学基金项目(2016MS0605);内蒙古自治区民族事务委员会基金项目(MW-2017-MGYWXXH-03)。**苏依拉**,教授,主研领域:人工智能与模式识别,自然语言处理。**高芬**,硕士生。**牛向华**,硕士生。**仁庆道尔吉**,教授。

数据稀疏的问题,也由于计算机技术的快速发展,现在有很多研究者专注于预训练神经网络语言模型的研究。实际上,语言模型就是依据上下文信息去预测下一个词,因为不需要人工标注的数据,所以预训练语言模型,其实就是从大量的单语语料中,学习到比较丰富的语义知识的表示。使用神经网络训练时基本上都是基于后向传播算法(Back Propagation,BP),通过对网络模型的参数进行随机的初始化,再通过一些优化算法去优化初始的模型参数。预训练的原理是:对神经网络模型的参数不进行随机的初始化,而是有目的地训练一套网络模型参数,之后将这套参数用来做初始化,然后再训练。在迁移学习中也大量地使用了这样的思想。

近几年,在自然语言处理(NLP)领域,在多项 NLP 任务上使用预训练语言模型的方法皆获得了令人可喜的成果,因此预训练语言模型也受到人们广泛的关注。例如,2018 年在 NTM 领域引起热门讨论的几篇文章都和预训练语言模型有关,其中三个比较具有代表性的模型为 ELMo[9]、OpenAI GPT[10]和 BERT[11],它们都是神经网络语言模型。因神经网络语言模型在当前研究中表现最佳,且使用神经网络的方法可以结合上下文信息学到一些语义信息有利于翻译系统性能的提升,所以本文也采用了该方法预训练语言模型。

1 基于多头自注意力机制的研究

虽然基于 LSTM[13]神经网络的方法在一定程度上能够缓解长距离依赖问题,但是因为不能进行并行运算,所以计算速度相对比较缓慢。而且,基于双向 LSTM[14]神经网络预训练语言模型的方法看似是双向的,实则是两个单向运算结果的拼接。为了实现真正意义上的双向运算,本文使用基于多头自注意力机制(Multi-head Self-attention Mechanism)的框架预训练跨蒙汉语言模型,此框架依然是编码器-解码器的结构,只不过编码器由多层编码器层堆叠而成,在每一层编码器层中都包含前馈神经网络层和使用了基于多头注意力机制的 Transformer 的注意力层[15]。

注意力机制来自人类的视觉注意力机制,当人类使用视觉感知外界事物的时候,一般不会将场景中的事物都扫描一遍,而是有选择地去观察那些对自己有用的信息。而注意力机制也是基于这样的原理,注意力函数可以被形容为一个查询(Query)到一系列键值对(Key-Value)的映射,如图 1 所示,其计算公式如下:

$$Attention(Query,Source)=\sum_{i=1}^{t_x}Similarity(Query,Key_i)\cdot Value_i \tag{1}$$

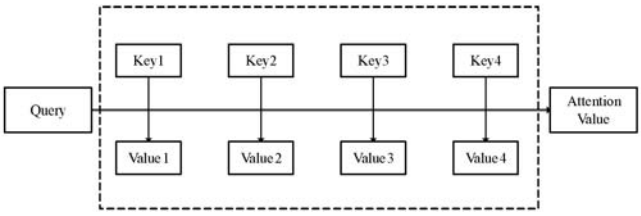

图 1 注意力机制原理

注意力函数值的计算步骤如下:首先是计算 *Query* 和每一个 *Key* 之间的相似度权重,相似度如果是通过点积再除以维度 K 进行计算的,则称为放缩点积注意(Scaled Dot-product Attention);然后通过 Softmax 将计算得到的权重进行归一化操作;最后将归一化之后的相似度权重和 *Value* 做加权求和就得到了注意力值。

自注意力机制就是运算过程中 *Key*、*Value*、*Query* 是相等的。多头注意力就是多个自注意力。多头注意力运算过程是 *Query*、*Key*、*Value* 首先进行一次线性变换,然后再执行多次的放缩点积注意力运算。而且每次 *Query*、*Key*、*Value* 进行线性变换的参数 W 是不一样的。然后将多次的放缩点积注意力结果进行拼接,再进行一次线性变换得到的值作为多头注意力机制的结果。多头注意力可以允许模型在不同的表示子空间里学习到相关的信息。在蒙汉语言模型的训练过程中,指的就是当输入一句话时,句子中的每个词都得和其他的词做注意力运算,这样就能学习到句子中每个词之间的相互依赖关系,即能更进一步地解决长期依赖的问题。

而 Transformer 的整体构成其实也是一个编码器-解码器的结构,只是整个编码器是由很多个子编码器组成的,每一个子编码器中包含一个多头注意力机制和一个前馈神经网络。它的解码器部分则和编码器部分类似,但是为了更好地优化网络,又添加了一个多头注意力层,并使其与其他层进行了残差连接。

1.1 框架构建

基于多头自注意力机制的框架最大的特点就是使用一个多层双向的 Transformer 编码器,因为 Transformer 中包含自注意力机制,所以该框架中所有的网络层都是直接相互连接的,这也是其优于双向 LSTM 神经网络模型的地方,因此它实现了真正意义上的双向和全局。基于多头自注意力机制框架的组成结构如图 2 所示。

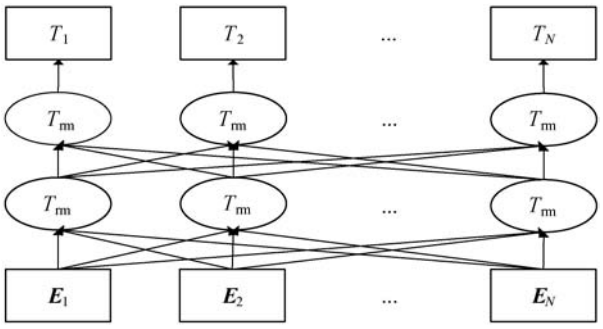

图 2　多头自注意力机制

其中:E_N表示加入了位置信息和句子切分信息的词向量,T_{rm}表示训练模型 Transformer,T_N表示翻译译文。在基于多头自注意力机制预训练语言模型的框架中,其输入是加入了位置信息和句子切分信息的词向量E_N,接着在多层 Transformer 上预测任务进行训练,称之为遮挡式语言模型(Mask Language Model,MLM)。此任务中,在输入词向量序列时,随机地使用标记 MASK 遮挡 15% 的词,然后通过训练来预测这些被遮挡的词是什么。

1.2　实验设置

首先对语料库进行预处理,使用词级粒度的方法对中文语料进行切分,再使用 BPE 进行子词级处理,学习到一个共享的蒙汉词汇表,将蒙汉语料库中共有的标记作为锚点来对齐两种语言的词向量空间,如数字和英文字母等,训练好一个跨语言词向量空间。接着借鉴 MLM 的方法进行语言模型建模,从 BPE 处理之后的文本中随机抽取 15% 的 BPE 标记,训练中 80% 的时间用 MASK 标签代替,10% 的时间用随机的标记替换,剩余 10% 的时间保持正确的 BPE 标记不变。MLM 预训练语言模型的过程如图 3 所示。

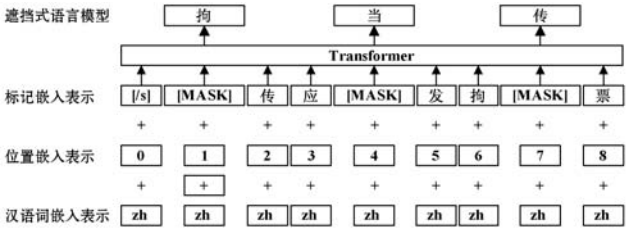

图 3　MLM 预训练语言模型的过程

实验环境配置为:Python 3.6.0,PyTorch 1.0.1,NumPy,fastBPE 和 Moses。

实验参数设置为:BPE codes 数量设置为 60 000,词向量的维度设置为 512,Transformer 的神经网络层数设置为 6 层,多头注意力机制设置为 8 头,droupout 设置为 0.1,注意力机制的 droupout 设置为 0,激活函数使用 GELU,batch_size 设置为 16,句子的最大长度设置为 256,使用 Adam 优化算法,学习率 lr 设置为 0.000 1,学习率衰减率设置为 0,一阶矩估计的指数衰减率 beta1 设置为 0.9,二阶矩估计的指数衰减率 beta2 设置为 0.98,epoch_size 设置为 300 000,batch_size 设置为 16。

本文主要以内蒙古工业大学构建的 123 万句蒙汉对齐语料库中的蒙古语作为源语言端单语语料库,以全球 AI 挑战赛(AI Changer)中给出的 1 000 万句英汉对齐语料库中的汉语作为目标语言端单语语料库为研究数据。训练集均为 123 万句对蒙汉单语语料,验证集均为 3 000 句对蒙汉平行语料,测试集均为 1 000 句对蒙汉平行语料。

1.3　实验结果

实验的评测使用的是语言模型评测中常用的两个指标:语义困惑度(Perplexity,ppl)和准确率(Accuracy,acc)来判断的。ppl 越小表示语言模型越好,而 acc 越高表示语言模型越好。

在整体预训练过程中,跨蒙汉语言模型的 ppl 和 acc 在测试上的变化趋势分别如图 4 和图 5 所示。

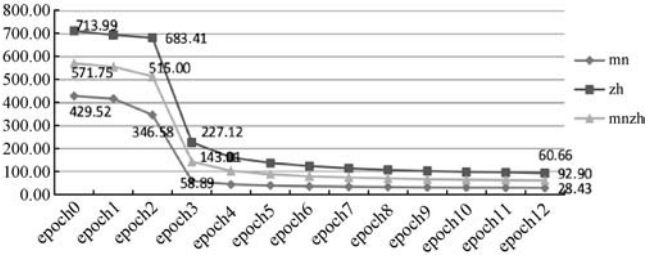

图 4　语言模型的 ppl 变化趋势

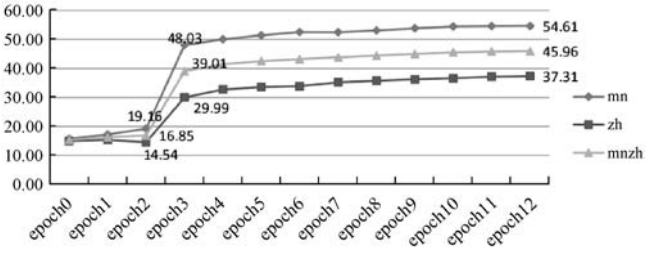

图 5　语言模型的 acc 变化趋势

可以看出在 epoch 3 时训练结果有了很大的提升,而当到达 epoch 10 左右结果的变化开始趋于平缓。这说明基于自注意力机制的预训练语言模型的方法,训练速度还是比较快的。由于结果可以很快收敛,所以短时间内也可以预训练一个比较好的跨蒙汉语言模型。

2　基于预训练语言翻译模型的构建

本文总体技术路线图如图 6 所示,图中虚线框表示的部分即为语言模型预训练部分。将其添加到蒙汉

语言模型之前对语言模型的参数进行初始化。

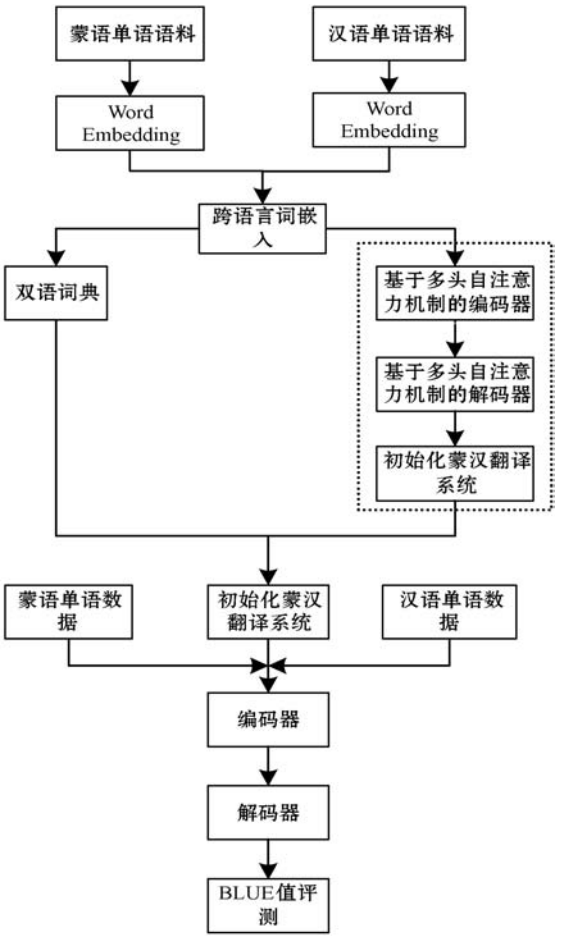

图 6 预训练语言模型的蒙汉机器翻译总体技术路线图

2.1 实验设置

实验环境为 Ubuntu 16.04 的 Linux 系统,实验架构的基础为 Pytorch,基于 Facebook AI Research 开源的框架 XLM 实现。参数设置如下:Transformer 的神经网络层数设置为 6 层,多头注意力机制设置为 8 头,droupout 设置为 0.1,注意力机制的 droupout 设置为 0,激活函数使用 GELU,batch_size 设置为 16,句子的最大长度设置为 256,优化函数使用 Adam 优化算法,学习率 lr 设置为 0.000 1,学习率衰减率设置为 0,一阶矩估计的指数衰减率 beta1 设置为 0.9,二阶矩估计的指数衰减率 beta2 设置为 0.98,epoch_size 设置为 300 000,batch_size 设置为 16,batch_size 的最大值限定为 64,实验终止条件设置为模型在验证集上 10 个 epoch 内不再发生改变。

2.2 实验结果

预训练语言模型的蒙汉机器翻译模型的 BLEU 值的变化趋势如图 7 所示。mn-zh 表示蒙汉翻译模型的结果;zh-mn 表示汉蒙翻译模型的结果;Test 表示在测试集上的结果;Valid 表示在验证集上的结果。

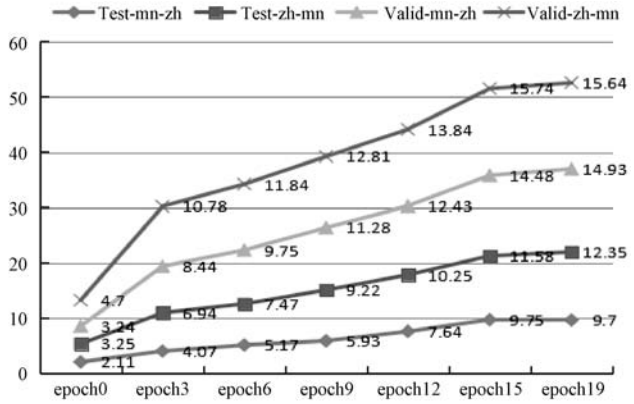

图 7 预训练之后的 BLEU 值

如图 7 所示,预训练跨蒙汉语言模型之后,基于单语语料库训练的蒙汉机器翻译模型的 BLEU 值在 20 个 epoch 上呈现增长趋势,在测试集上蒙汉翻译性能于 epoch17 时表现最好,此时 BLEU 值为 10.23,在验证集上蒙汉翻译性能于 epoch19 时表现最好,此时 BLEU 值为 14.93。

3 平行语料改进预训练语言模型

本节考虑使用少量平行语料来改进蒙汉翻译模型。因为跨蒙汉语言模型和汉语语言模型在验证集中的 ppl 都很高,所以提出应用翻译语言模型(Translation Language Model,TLM)来改进跨蒙汉预训练。方法是在基于多头自注意力机制预训练语言模型的框架中加入另一个预测任务,此任务是预测下一个句子。在此任务中,一方面输入同属于一句话中的两个句子,将其视为正例;另一方面则输入两句不属于同一句话中的句子,将其视为反例,两方面的任务各占一半。那么整个语言模型预训练的目标函数就是将这两个任务求和之后再使用极大似然函数求值。为了预测被遮挡的蒙语,该模型可以同时处理蒙语句子和与它对应的汉语句子,并通过重置汉语句子中的位置嵌入表示以使得蒙语和汉语词嵌入表示进一步对齐。TLM 预训练语言模型的过程如图 8 所示。实验设置同第 2 节一致,只是在数据集中添加了 3 万句对的蒙汉平行语料库作为训练集,平行语料的预处理方式同验证集和测试集一致。

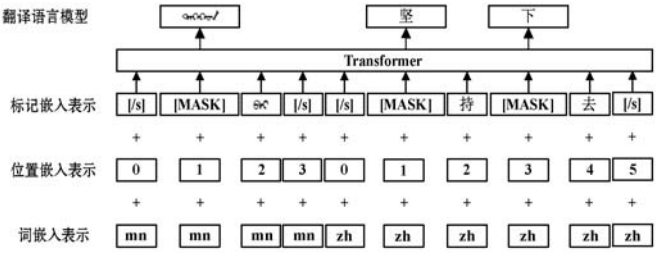

图 8 TLM 预训练语言模型的过程

加入 TML 预训练语言模型的蒙汉机器翻译模型的 BLEU 值的变化趋势如图 9 所示。

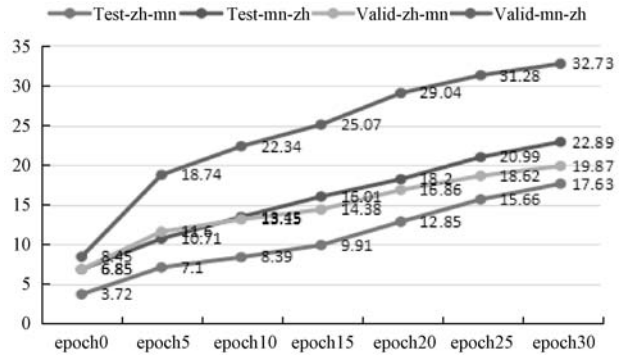

图 9　添加 TML 后 BLEU 值的变化趋势

可以看出,添加 TML 之后,蒙汉机器翻译模型的 BLUE 值产生了较大幅度的提升,蒙译汉于测试集上在 epoch 28 表现最好,此时 BLEU 值为 23.05,于验证集上也是在 epoch 28 表现最好,此时 BLEU 值为 32.743。这证明加入平行语料库进一步促进了蒙汉词嵌入表示的对齐程度,提升了基于单语语料训练的蒙汉机器翻译系统的整体翻译性能。

4　实验结果分析

本文叙述了基于自注意力机制预训练跨蒙汉语言模型的方法以及添加少量蒙汉平行语料来进一步优化预训练语言模型的方法,并基于这两种方法构建了蒙汉翻译系统进行实验。将实验结果和基于单语语料库训练方法进行了对比,统计到三组实验在 20 个 epoch 上测试集的 ppl 和 BLEU 如表 1 所示,其相应的变化趋势如图 10 所示。

表 1　三组实验结果

epoch	BLEU（单语）	BLEU（单语 + 预训练）	BLEU（单语 + 双语 + 预训练）	ppl（单语 + 双语 + 预训练）
epoch0	1.12	2.11	6.85	114.15
epoch1	2.07	3.08	8.65	88.53
epoch2	2.78	3.79	9.87	73.76
epoch3	3.08	4.07	11.03	66.67
epoch4	3.44	4.45	11.97	51.77
epoch5	3.5	4.56	10.71	48.27
epoch6	4.16	5.17	11.30	46.84
epoch7	4.52	5.53	11.62	44.71
epoch8	4.51	5.51	12.13	42.25
epoch9	4.82	5.93	13.43	41.25
epoch10	5.17	6.56	13.45	40.09

续表 1

epoch	BLEU（单语）	BLEU（单语 + 预训练）	BLEU（单语 + 双语 + 预训练）	ppl（单语 + 双语 + 预训练）
epoch11	5.81	6.84	14.18	38.17
epoch12	6.63	7.64	14.84	35.80
epoch13	7.27	8.26	15.24	34.38
epoch14	8.06	9.07	15.65	33.15
epoch15	8.75	9.75	16.01	32.89
epoch16	8.86	9.84	16.31	31.38
epoch17	**9.18**	**10.23**	16.85	30.21
epoch18	8.96	9.95	16.74	29.51
epoch19	**8.70**	**9.70**	**17.35**	**29.02**

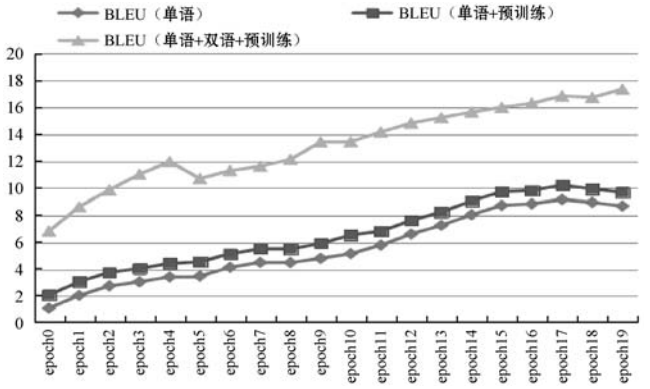

图 10　三组实验 BLEU 值变化趋势

可以明显地看出,在语料库相同的情况下,预训练跨蒙汉语言模型的方法结果表现较好,说明语言模型对于基于单语语料库训练的蒙汉机器翻译性能有很大的影响,预训练跨蒙汉语言模型能够在一定程度上改善其翻译质量。通过对比实验证实了融合单语和双语语料预训练语言模型提升基于单语语料库训练的蒙汉机器翻译的可行性。三组实验训练出的最优 BLEU 值如图 11 所示。

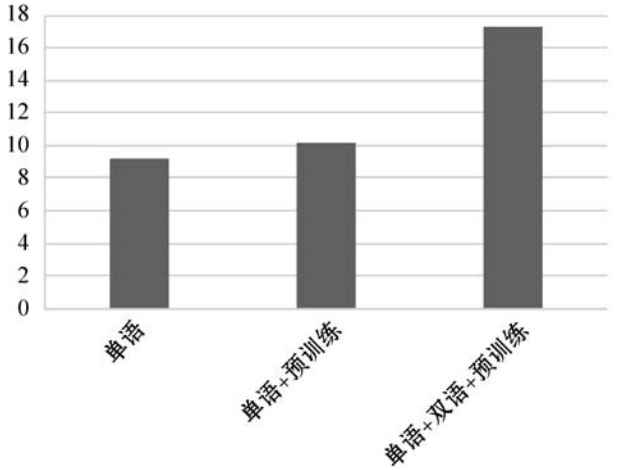

图 11　最优 BLEU 值

基于单语语料库训练的蒙汉机器翻译系统在不同的 epoch 上翻译的测试语句如表 2 和表 3 所示。

表 2　短句翻译模型测试

原文	[illegible]
参考译文	明天 是 雨天 。
epoch 0 译文	明天 有 风 。
epoch 19 译文	明天 要 下雨 。
epoch 28 译文	明天 是 雨天 。

表 3　长句翻译模型测试

原文	[illegible]
参考译文	找到 固定 工作 不 容易 ，可 找到 一些 临时 工作 倒 还有 可能 。
epoch 0 译文	工作 不 是 很 容易 ，但 是 可以 做 的 。
epoch 19 译文	找到 固定 工作 不 容易 ，可 找到 临时 工作 却 是 可以 找到 的 。
epoch 28 译文	找到 固定 工作 不 容易 ，可 找到 一些 临时 工作 倒 还 有 可能 。

如表 2 和表 3 所示，模型通过不断地学习最终可以将原句子表达的意思准确地翻译出来，而且在没有准确翻译出来时也没有出现 UNK，而是使用了和原文中词语意思较为接近的词来代替。

5　结　语

本文主要介绍了预训练思想和基于自注意力机制预训练跨蒙汉语言模型的方式以及基于此方法设置的实验，并将实验所得的跨蒙汉语言模型应用于基于单语语料库训练的蒙汉机器翻译系统中。最后，通过和基于单语训练的蒙汉翻译模型的实验结果做对比。基于自注意力机制预训练跨蒙汉语言模型的方法极大地改善了蒙汉机器翻译系统的性能。究其原因，其一是自注意力机制能够学习整个句子中每一个词之间的关系，可以更好地学习到上下文的信息；其二是预训练的方法可以对参数进行有目的的初始化；其三是使用了少量的蒙汉平行语料库进一步提升了语言模型的准确率以及蒙汉词嵌入表示的对齐程度。

参　考　文　献

[1] 李业刚，黄河燕，史树敏，等. 多策略机器翻译研究综述[J]. 中文信息学报，2015，29(2)：1 - 9.

[2] He D, Xia Y, Qin T, et al. Dual learning for machine translation[C]//Barcelona, Spain: Advances in Neural Information Processing Systems. 2016: 820 - 828.

[3] Cheng Y, Xu W, He Z, et al. Semi-supervised learning for neural machine translation[EB]. arXiv preprint arXiv:1606.04596, 2016.

[4] Gulcehre C, Firat O, Xu K, et al. On using monolingual corpora in neural machine translation[EB]. arXiv preprint arXiv:1503.03535, 2015.

[5] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data[EB]. arXiv preprint arXiv:1511.06709, 2015.

[6] Nakayama H, Nishida N. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot[J]. Machine Translation, 2017, 31(1 - 2): 49 - 64.

[7] Lample G, Conneau A, Denoyer L, et al. Unsupervised machine translation using monolingual corpora only[EB]. arXiv preprint arXiv:1711.00043, 2017.

[8] Artetxe M, Labaka G, Agirre E, et al. Unsupervised neural machine translation[EB]. arXiv preprint arXiv:1710.11041, 2017.

[9] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[EB]. arXiv preprint arXiv:1802.05365, 2018.

[10] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB]. 2018.

[11] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB]. arXiv:1810.04805, 2018.

[12] Lin Z, Feng M, Santos C N D, et al. A structured self-attentive sentence embedding[EB]. arXiv:1703.03130, 2017.

[13] Gers F A, Schmidhuber E. LSTM recurrent networks learn simple context-free and context-sensitive languages[J]. IEEE Transactions on Neural Networks, 2001, 12(6): 1333 - 1340.

[14] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5 - 6): 602 - 610.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[EB]. arXiv:1706.03762, 2017.

(上接第 109 页)

[14] 王日升，谢红薇，安建成. 基于分类精度和相关性的随机森林算法改进[J]. 科学技术与工程，2017，17(20)：67 - 72.

[15] 方匡南，吴见彬，朱建平，等. 随机森林方法研究综述[J]. 统计与信息论坛，2011，26(3)：32 - 38.

[16] 詹鹏伟，谢小姣. 几种降维技术在分类问题中的效果评估[J]. 科技创新与应用，2018，241(21)：22 - 23，26.

[17] 孙平安，王备战. 机器学习中的 PCA 降维方法研究及其应用[J]. 湖南工业大学学报，2019，33(1)：73 - 78.

[18] 和湘，刘晟，姜吉国. 基于机器学习的入侵检测方法对比研究[J]. 信息网络安全，2018，209(5)：1 - 11.