第47卷 第12期 计 算 机 工 程 2021年12月
Vol.47 No.12 Computer Engineering December 2021

·人工智能与模式识别· 文章编号：1000-3428(2021)12-0112-06 文献标志码：A 中图分类号：TP18

# 面向汉维机器翻译的BERT嵌入研究

陈 玺[1,2,3]，杨雅婷[1,2,3]，董 瑞[1,2,3]

（1.中国科学院新疆理化技术研究所，乌鲁木齐 830011；2.中国科学院大学，北京 100049；
3.新疆民族语音语言信息处理实验室，乌鲁木齐 830011）

**摘 要：**针对训练汉维机器翻译模型时汉语-维吾尔语平行语料数据稀疏的问题，将汉语预训练语言BERT模型嵌入到汉维神经机器翻译模型中，以提高汉维机器翻译质量。对比不同汉语BERT预训练模型编码信息的嵌入效果，讨论BERT不同隐藏层编码信息对汉维神经机器翻译效果的影响，并提出一种两段式微调BERT策略，通过对比实验总结出将BERT模型应用在汉维神经机器翻译中的最佳方法。在汉维公开数据集上的实验结果显示，通过该方法可使机器双语互译评估值（BLEU）提升1.64，有效提高汉维机器翻译系统的性能。

**关键词：**汉维翻译；神经机器翻译；预训练语言模型；BERT模型；两段式微调策略

**开放科学（资源服务）标志码（OSID）：**

**中文引用格式：**陈玺，杨雅婷，董瑞.面向汉维机器翻译的BERT嵌入研究[J].计算机工程，2021，47(12)：112-117.

**英文引用格式：**CHEN X，YANG Y T，DONG R.Research on BERT embedding for Chinese-Uyghur machine translation[J]. Computer Engineering，2021，47(12)：112-117.

# Research on BERT Embedding for Chinese-Uyghur Machine Translation

CHEN Xi[1,2,3]，YANG Yating[1,2,3]，DONG Rui[1,2,3]

（1. Xinjiang Technical Institute of Physics and Chemistry，Chinese Academy of Sciences，Urumqi 830011，China；
2. University of Chinese Academy of Sciences，Beijing 100049，China；
3. Xinjiang Laboratory of Minority Speech and Language Information Processing，Urumqi 830011，China）

**【Abstract】**The Chinese-Uyghur parallel corpus required for training Chinese-Uyghur machine translation models suffer from data sparsity.To address the problem，this paper embeds the Chinese pre-trained language BERT model into a Chinese-Uyghur neural machine translation model to improve the quality of translation. This research compares the embedding effects of coding information of different Chinese BERT pre-trained models，explores the influence of the coding information at different hidden layers of Chinese BERT on Chinese-Uyghur neural machine translation，and on this basis proposes a two-stage BERT fine-tuning strategy.By comparative experiments，this paper summarizes the best method of applying the BERT model to the Chinese-Uyghur neural machine translation.The experimental results on the Chinese-Uyghur public dataset show that the proposed model increases the BLEU value by 1.64，and significantly improves the performance of the Chinese-Uyghur machine translation system.

**【Key words】**Chinese-Uyghur translation；Neural Machine Translation（NMT）；pre-trained language model；BERT model；two-stage fine-tuning strategy

**DOI：**10. 19678/j. issn. 1000-3428. 0059863

## 0 概述

近年来，基于深度学习的神经机器翻译（Neural Machine Translation，NMT）技术取得了较大的进展，网络结构从循环神经网络[1-3]发展到卷积神经网络[4]，再到完全基于自注意力机制的网络[5]。在这些不同的网络结构中，基于自注意力机制而又高度并行化的Transformer[5]取得了非常好的效果。

目前的神经机器翻译模型在面对英法、英中等拥有大规模平行语料资源丰富的语言对时，取得了较好

**基金项目：**国家自然科学基金“融合复杂形态特征的多语言神经机器翻译研究”（U1703133）；国家重点研发计划“维吾尔语、哈萨克语到汉语的机器翻译研究”（2017YFC0822505-04）；新疆高层次引进人才项目（新人社函[2017]699号）；中国科学院“西部之光”人才培养计划A类项目“以和田墨玉为例的维汉翻译关键技术研究”（2017-XBQNXZ-A-005）。

**作者简介：**陈 玺（1995—），男，硕士研究生，主研方向为自然语言处理、机器翻译；杨雅婷，研究员、博士；董 瑞，副研究员、博士。

**收稿日期：**2020-10-28 **修回日期：**2020-12-02 **E-mail：**chenxi184@mails.ucas.ac.cn

的翻译效果。但是由于汉语-维吾尔语平行语料的缺乏且2种语言的差异性较大,其在汉维翻译方面效果并不如维吾尔语-汉语上翻译[6]。本文主要研究如何提升汉维神经机器翻译模型的翻译效果。

BERT[7]、Roberta[8]、GPT[9]等预训练语言模型在大规模的无标签单语语料上训练得来,在一系列自然语言理解任务(如文本分类[10]、阅读理解[11]等)上都取得了非常好的效果。BERT是一种多层的基于Transformers的双向编码表示模型,通过在大量的单语语料上以屏蔽语言模型建模任务(Masked Language Model, MLM)和下一句预测任务(Next Sentence Prediction,NSP)为训练目标得到。

尽管BERT在一系列自然语言理解任务上取得了不错的效果,但其在自然语言生成任务(如机器翻译、摘要生成[12]等)上的应用却鲜有人探索。文献[13]比较了在机器翻译翻译模型当中应用BERT的几种方式,包括将BERT作为NMT模型的输入嵌入层、利用BERT的参数初始化NMT模型的编码器层然后微调BERT、利用BERT的参数初始化NMT模型的编码器层然后冻结BERT参数。文献[14]将BERT应用于篇章级别的机器翻译,在法语-英语、汉语-英语、西班牙语-英语上取得了较好的翻译效果。文献[15]将BERT和机器翻译模型中的编码器模块和解码器模块分别进行注意力机制交互,然后进行特征融合来提升机器翻译的效果,在WMT语料和IWSLT语料上均取得了较好的效果。

本文借鉴文献[15]方法,设计一系列实验来探究如何在汉维机器翻译中更好地应用BERT。通过设计两段式微调BERT的方法,将BERT中的先验知识迁移到NMT模型中,同时根据对比实验总结出在汉维机器翻译中应用预训练BERT模型的最佳方法。

## 1 模型架构与嵌入策略

### 1.1 基于BERT嵌入的汉维神经机器翻译模型

本文采用文献[15]提出的基于注意力机制的BERT-fused模型,将源语言汉语输入BERT中,并固定BERT的参数,提取源语言经过BERT编码的预训练表示,然后借助于BERT编码器部分和BERT解码器部分的注意力机制模块,将得到的预训练表示分别与NMT模型编码器模块和解码器模块的每一层进行注意力交互,再将交互得到的结果与编码器模块和解码器模块每一层自身的自注意力特征进行融合。通过这样的方法,可以将BERT编码源语言的预训练特征表示融入到编码器模块和解码器模块的每一层当中,以充分利用预训练语言模型BERT,同时避免BERT模型和机器翻译模型在词切分技术上不同的问题。该模型结构如图1所示。

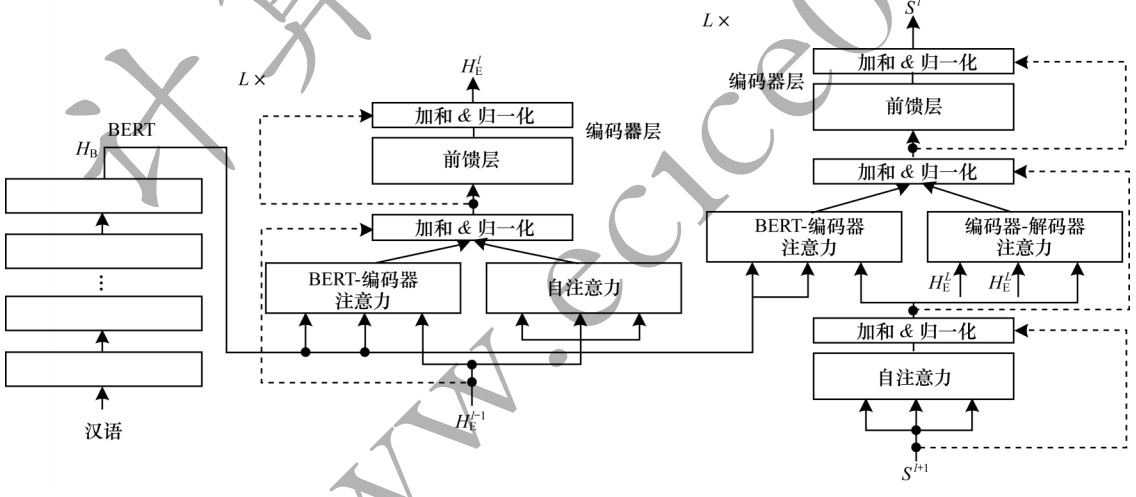


**图1 基于BERT的机器翻译模型架构**

**Fig.1 Architecture of machine translation model based on BERT**

在图1中,从左至右依次分别为BERT模块、编码器模块和解码器模块,虚线代表残差连接,$H_B$和$H_E^L$分别代表BERT模块和编码器模块最后一层的输出。与标准的基于Transformer的NMT模型相比,除了模型原有结构,还有2个额外的注意力机制模块,即BERT编码器注意力机制模块和BERT解码器注意力机制模块。源语言经过BERT模型的输出与NMT模块每一层的输入计算BERT编码器注意力、BERT解码器注意力,再与NMT模型自身的自注意力机制相融合。BERT编码器注意力机制模块与编码器的自注意力机制模块融合后的输出如式(1)所示:

$$\tilde{h}_i^l=\frac{1}{2}\left(\mathrm{attn}_s\left(h_i^{l-1},H_E^{l-1},H_E^{l-1}\right)+\mathrm{attn}_B\left(h_i^{l-1},H_B,H_B\right)\right)\quad \forall i\in\left[l_x\right] \tag{1}$$

给定源语言输入,BERT将语言输入编码为$H_B$。在式(1)中,$H_E^l$代$x$表编码器模块第$l$隐藏层的输出表示,$H_E^0$为源语言序列的词向量表示,$l_x$代表源语言句子中的第个子词,记$H_E^l$中的第$i$个子词为$h_i^l$,$\mathrm{attn}_s$和$\mathrm{attn}_B$为Transformer中的缩放点积注意力,它们拥

有不同的参数。然后，得到的每个$\tilde{h}_i^l$被输入前馈层，得到第$l$层的输出，如式(2)所示：

$$H_{\mathrm{E}}^l=\left(\mathrm{FFN}\left(\tilde{h}_1^l\right),\mathrm{FFN}\left(\tilde{h}_2^l\right),\cdots,\mathrm{FFN}\left(\tilde{h}_{l_x}^l\right)\right) \tag{2}$$

对于解码器端，以$S_{<t}^l$代表在时刻$t$之前位于解码器模块第$l$层的隐藏状态。在第$l$层可以得到解码器的自注意力模块和BERT-解码器注意力模块，分别如式(3)和式(4)所示：

$$\hat{s}_t^l=\mathrm{attn}_{\mathrm{S}}\left(s_t^{l-1},S_{<t+1}^{l-1},S_{<t+1}^{l-1}\right) \tag{3}$$

$$\tilde{s}_t^l=\frac{1}{2}\left(\mathrm{attn}_{\mathrm{B}}\left(\hat{s}_t^l,H_{\mathrm{B}},H_{\mathrm{B}}\right)+\mathrm{attn}_{\mathrm{E}}\left(\hat{s}_t^l,H_{\mathrm{E}}^l,H_{\mathrm{E}}^l\right)\right)$$

$$s_t^l=\mathrm{FFN}\left(\tilde{s}_t^l\right) \tag{4}$$

编码器-解码器注意力模块的输出进行融合后通过前向传播网络。在式(3)和式(4)中，$\mathrm{attn}_{\mathrm{S}}$、$\mathrm{attn}_{\mathrm{B}}$、$\mathrm{attn}_{\mathrm{E}}$分别代表解码器的自注意力模块、BERT-解码器注意力模块和编码器-解码器注意力模块。将式(3)和式(4)应用在解码器每一层中，最后一层解码器可以得到$s_t^L$，对$s_t^L$通过线性变换和softmax分类函数得到第$t$个被预测的单词$\hat{y}_t$，解码器一直进行解码直到输出句子结束符。

模型通过DropNet方法来将BERT-编码器注意力、BERT解码器注意力与NMT模型自身的自注意力机制相融合，从而将BERT的输出特征表示嵌入到NMT模型中。DropNet比率$P_{\mathrm{Net}}$取值范围在[0,1]之间，在每次训练迭代的过程中，对于每一层$l$，在均匀分布[0,1]上采样得到$U^l$，每一层$\tilde{h}_i^l$的计算公式如式(5)所示：

$$\begin{aligned}\tilde{h}_{i,\mathrm{Drop-Net}}^l=&I\left(U^l<\frac{p_{\mathrm{Net}}}{2}\right)\cdot\mathrm{attn}_{\mathrm{S}}\left(h_i^{l-1},H_{\mathrm{E}}^{l-1},H_{\mathrm{E}}^{l-1}\right)+\\&I\left(U^l>1-\frac{p_{\mathrm{Net}}}{2}\right)\cdot\mathrm{attn}_{\mathrm{B}}\left(h_i^{l-1},H_{\mathrm{B}},H_{\mathrm{B}}\right)+\\&\frac{1}{2}I\left(\frac{p_{\mathrm{Net}}}{2}\leqslant U^l\leqslant 1-\frac{p_{\mathrm{Net}}}{2}\right)\cdot\\&\left(\mathrm{attn}_{\mathrm{S}}\left(h_i^{l-1},H_{\mathrm{E}}^{l-1},H_{\mathrm{E}}^{l-1}\right)+\right.\\&\left.\mathrm{attn}_{\mathrm{B}}\left(h_i^{l-1},H_{\mathrm{B}},H_{\mathrm{B}}\right)\right)\end{aligned} \tag{5}$$

其中，$I(\cdot)$是指示函数。对于任意一层编码器模块，以$p_{\mathrm{Net}}/2$概率去选择BERT编码器注意力或自注意力，每次只选择其中一种。在推理阶段，每种注意力机制都会被用到，如式(1)所示，即推理时有：

$$E_{U\sim\mathrm{uniform}[0,1]}\left(\tilde{h}_{i,\mathrm{Drop-Net}}^l\right) \tag{6}$$

同理，对于解码器模块，有：

$$\begin{aligned}\tilde{s}_{t,\mathrm{Drop-Net}}^l=&I\left(U^l<\frac{p_{\mathrm{Net}}}{2}\right)\cdot\mathrm{attn}_{\mathrm{B}}\left(\hat{s}_t^l,H_{\mathrm{B}},H_{\mathrm{B}}\right)+\\&I\left(U^l>1-\frac{p_{\mathrm{Net}}}{2}\right)\cdot\mathrm{attn}_{\mathrm{E}}\left(\hat{s}_t^l,H_{\mathrm{E}}^l,H_{\mathrm{E}}^l\right)+\\&\frac{1}{2}I\left(\frac{p_{\mathrm{Net}}}{2}\leqslant U^l\leqslant 1-\frac{p_{\mathrm{Net}}}{2}\right)\cdot\\&\left(\mathrm{attn}_{\mathrm{B}}\left(\hat{s}_t^l,H_{\mathrm{B}},H_{\mathrm{B}}\right)+\mathrm{attn}_{\mathrm{E}}\left(\hat{s}_t^l,H_{\mathrm{E}}^l,H_{\mathrm{E}}^l\right)\right)\end{aligned} \tag{7}$$

在推理阶段，BERT解码器注意力和编码器-解码器注意力都会被用到。

### 1.2 不同汉语BERT对于翻译结果的影响

本文采用的BERT模型具体情况如表1所示，其中包括以下6种汉语BERT模型：

1)BERT-base-multilingual-uncased模型[16]，由12层Transformer组成，隐藏层的特征维数是768维，Transformer模块的多头注意力包含12个头，共包含110M参数。该模型是在102种语言的维基百科语料上训练得来的，其中包含汉语，但不包含维吾尔语，预训练时中文按字进行切分。

2)BERT-base-Chinese模型[17]，由12层Transformer组成，隐藏层的特征维数是768维，Transformer模块的多头注意力包含12个头，共包含110M参数。该模型是在中文维基百科语料上训练得来的，在预训练的过程中按字进行切分。

3)BERT-wwm-ext模型[18]，由12层Transformer组成，隐藏层的特征维数是768维，Transformer模块的多头注意力包含12个头，共包含110M参数。该模型是在中文维基百科语料和通用数据上训练得来的。同时，在预训练的过程中按词进行切分，使用了全词遮罩技术。Google发布的BERT-base-Chinese模型中文是以字为粒度进行切分，没有考虑到中文自然语言处理中的中文分词问题。文献[19]提出了基于全词遮罩(Whole Word Masking，WMM)技术的中文预训练模型BERT-wwm，将全词遮罩的方法应用在了中文中。将该模型在中文维基百科语料上进行训练，在许多任务上都取得了非常好的效果。本研究将基于全词遮罩的BERT模型应用到模型之中。

4)RoBERTa-wwm-large-ext模型[20]。RoBERTa[8]是BERT通用语义表示模型的一个优化版，它在BERT模型的基础上提出了动态遮罩方法，去除了下一个句子预测预训练目标，同时在更多的数据上采用更大的批处理大小训练更长的时间，在多个任务中取得了很好的效果。该模型由24层Transformer组成，隐藏层的特征维数是1 024维，Transformer模块的多头注意力包含16个头，共包含330M参数。该模型是在中文维基百科语料和通用数据上训练得来的。同时，在预训练的过程中按词进行切分，使用了全词遮罩技术。

5)RoBERTa-wwm-ext模型[21]，由12层Transformer组成，隐藏层的特征维数是768维，Transformer模块的多头注意力包含12个头，共包含110M参数。该模型是在中文维基百科语料和通用数据上训练得来的。同时，在预训练的过程中按词进行切分，使用了全词遮罩技术。

6)RBTL3模型[22]。该模型以RoBERTa-wwm-large-ext模型参数初始化前3层Transformer以及词向量层并在此基础上继续训练了1M步，在仅损失少量效果的情况下大幅减少参数量，得到了更轻量的模型。同时在预训练的过程中按词进行切分，使用了全词遮罩技术。

表1　不同BERT模型的比较

Table 1　Comparison of different BERT models

| 模型 | Transformer层数 | 预训练语料 | 隐藏层特征维数 | 多头注意力头数 | 参数个数/$10^6$ | 中文是否全词遮罩 |
|---|---|---|---|---|---|---|
| BERT-base-multilingual-uncased | 12 | 102种语言的维基百科 | 768 | 12 | 110 | 否 |
| BERT-base-Chinese | 12 | 中文维基百科 | 768 | 12 | 110 | 否 |
| BERT-wwm-ext | 12 | 中文维基和通用数据 | 768 | 12 | 110 | 是 |
| RoBERTa-wwm-large-ext | 24 | 中文维基和通用数据 | 1 024 | 16 | 330 | 是 |
| RoBERTa-wwm-ext | 12 | 中文维基和通用数据 | 768 | 12 | 110 | 是 |
| RBTL3 | 3 | 中文维基和通用数据 | 1 024 | 16 | 65 | 是 |

注：通用数据包括百科、新闻、问答等数据。

### 1.3　BERT不同隐藏层对翻译结果的影响

在模型中，BERT的输出作为一个额外的源语言序列表示，使用额外的注意力机制来将其引入到NMT模型当中。将BERT最后一层输出作为模型中额外注意力机制的输入，预训练模型的输出特征被引入到NMT模块的所有层中，以确保预训练模型的特征能够被完全利用。本文使用注意力机制将NMT模块和BERT预训练特征相结合，使NMT模块能够动态地决定从BERT中得到哪些特征。

文献[23]提出的BERT预训练语言模型学习到了一些结构化的语言信息，例如BERT的底层网络学习到了短语级别的信息表征，中层网络学习到了丰富的语言学特征，而高层网络则学习到了丰富的语义信息特征，将源语言用BERT编码后，底层、中层、高层分别有不同的语言信息表征。本文探索使用不同层次的BERT特征对于模型的翻译效果的影响。因为神经机器翻译模型是用编码器将源语言编码成语义特征，再送入解码器进行解码，所以猜想将BERT高层的语义特征引入到模型当中应该会取得较好的效果。单独将BERT的1、3、5、7、9、11隐藏层输出分别引入到模型当中，观察得到的模型翻译效果。同时，将BERT的1、3、5、7、9、11层编码的特征分别引入到编码器和解码器的1~6层中进行对比实验。

### 1.4　两段式BERT微调策略

灾难性遗忘是迁移学习中经常出现的一个问题，指模型在学习新知识的过程当中将原有预训练的知识遗忘[24]。当以较大的学习率微调BERT时会导致模型发生灾难性遗忘问题，而且直接微调BERT和整个模型的参数会使得模型的效果变差。本文探索如何微调BERT，提出一种两段式微调BERT的方法。首先固定BERT的参数，将BERT模型作为一个特征提取器，将提取到的预训练表示融入到NMT模型当中，只训练模型剩余部分的参数直到模型收敛，即训练BERT-fused模型BERT以外的部分直至收敛。然后微调模型中包括BERT在内的整个模型的参数。在微调的过程中，不改变其他训练参数，只改变学习率和预热更新步数（warmup updates）。在此基础上，通过实验对比不同的学习率和预热更新步数对模型翻译效果的影响。

## 2　实验与结果分析

### 2.1　实验数据集情况

本文采用2017年全国机器翻译研讨会（CWMT）公开的维吾尔语-汉语语料数据集进行实验。其中，训练集的数量为336 397，开发集的数量为700，测试集的数量为1 000。对维语语料按照词进行切分，对汉语语料按照字进行切分。所有的维吾尔语句子都通过字节对编码（Byte-Pair Encoding，BPE）技术[25]进行预处理，BPE融合数设置为10 000。实验评测指标为机器双语互译评估值（BLEU）。

### 2.2　训练参数设置

本文实验基于fairseq[26]，fairseq是Facebook开源的自然语言处理框架，基于pytorch开发，具有多卡训练性能好、支持混合精度训练等优点。在fairseq实现Transformer模型的基础上引入BERT编码器注意力和BERT解码器注意力，然后进行2种注意力的融合。模型使用6层Transformer作为编码器模块，使用6层Transformer作为解码器模块，词嵌入维度为512维，全连接层维度为1 024维，失活（dropout）率设置为0.3。drop-net比率$p_{\mathrm{Net}}$设置为1.0。使用BLEU[27]值来评估翻译质量，值越大翻译质量越好。

首先训练一个和BERT-fused模型中NMT部分同样架构同样参数的NMT模型直到收敛，然后利用这个已经得到的模型初始化图1所示模型的编码器和解码器，BERT和编码器之间的注意力模块参数与BERT和解码器之间的注意力参数随机进行初始化。使用的分批训练数据大小（max tokens）为8 000，使用Adam优化算法进行模型参数优化，初始学习率是0.000 5。在生成翻译结果的过程中，设置分批训练数据大小为128，设置集束搜索（beam search）的大小

为5,惩罚长度因子设置为1.0。

### 2.3 不同汉语BERT实验结果

不同汉语BERT的实验结果的对比如表2所示,其中基线是指完全基于6层Transformer不引入BERT特征的模型,加粗表示最优数据。对于所有的BERT模型,都将最后一层的输出特征融入到NMT模型中。由表2的实验结果可以看出,在所有BERT模型当中效果最好的是BERT-base-Chinese模型,相较于基线不引入BERT的模型BLEU值提高了1.02;拥有同样参数的基于全词遮罩的模型BERT-wwm-ext并没有表现出更好的效果,BLEU值仅提高了0.56;拥有同样参数的RoBERTa-wwm-ext全词遮罩模型并没有BERT-wwm-ext模型效果好,使BLEU提高了0.30;模型网络层数更深,参数更多的RoBERTa-wwm-large-ext模型效果较RoBERTa-wwm-ext要好,较基线BLEU值提高了0.93,但仍不及BERT-base-Chinese模型;更轻量的RBTL3模型和多语言版本BERT-base-multilingual-uncased模型得到的翻译效果甚至都没有完全基于6层Transformer不引入BERT的翻译效果好。

**表2 不同汉语BERT模型的实验结果**

**Table 2 Experimental results of different Chinese BERT models**

| 模型 | BLEU值 |
|---|---|
| 基线 | 30.77 |
| BERT-base-multilingual-uncased | 30.63 |
| BERT-base-Chinese | **31.79(+1.02)** |
| BERT-wwm-ext | 31.33(+0.56) |
| RoBERTa-wwm-large-ext | 31.70(+0.93) |
| RoBERTa-wwm-ext | 31.07(+0.30) |
| RBTL3 | 30.22 |

### 2.4 BERT不同隐藏层实验结果

表3中的基线为完全基于Transformer,不引入BERT特征的模型,基线右侧分别为BERT-fused模型融入BERT模型的1、3、5、7、9、11层输出的特征,奇数层指将BERT的1、3、5、7、9、11层编码的特征分别引入到编码器和解码器的1~6层中进行融合,实验使用bert-base-chinese模型,加粗表示最优数据。由表3的实验结果可以看出,将BERT的1、3层特征信息的引入对于模型起到了负面的影响,使模型的翻译效果出现了下降,而5、7、9、11层特征信息的引入让模型的翻译效果逐步上升。最后一层特征的引入效果最好,相较于基线提高了1.02 BLEU值,这验证了文献[23]得到的关于BERT的高层网络学习到了丰富的语义信息特征的结论。将BERT的1、3、5、7、9、11层依次融入到NMT的编码器和解码器1~6层当中效果并没有只融入最后一层的效果好。

**表3 BERT不同隐藏层的实验结果**

**Table 3 Experimental results of different hidden layers of BERT**

| 隐藏层 | BLEU值 | 隐藏层 | BLEU值 |
|---|---|---|---|
| 基线 | 30.77 | 第7层 | 31.28(+0.51) |
| 第1层 | 29.73 | 第9层 | 31.43(+0.66) |
| 第3层 | 30.49 | 第11层 | **31.79(+1.02)** |
| 第5层 | 30.78(+0.01) | 奇数层 | 31.19(+0.42) |

### 2.5 两段式BERT微调策略实验结果

不同微调BERT参数策略的实验结果如表4所示,其中不微调BERT是指训练BERT-fused模型,固定BERT参数不微调,加粗表示最优数据。直接微调BERT是指在训练BERT-fused模型的过程中直接微调BERT的参数。由表4的实验结果可以看出,在训练BERT-fused模型的过程中直接微调BERT的效果与不微调相比翻译效果会明显变差。在两段式微调的过程中,学习率过大会导致模型无法收敛;当预热更新步数为15 000时,模型取得了最高BLEU值32.41,相较于不微调BLEU值提高了0.62;当学习率固定为8e-5时,预热更新步数为15 000时模型的翻译效果最好。

**表4 不同微调BERT策略的实验结果**

**Table 4 Experimental results of different fine-tuning strategies of BERT**

| 微调方式 | 学习率 | 预热更新步数 | BLEU值 |
|---|---|---|---|
| 不微调BERT | 8e-5 | 15 000 | 31.79 |
| 直接微调BERT | 8e-5 | 15 000 | 30.82 |
| 两段式微调BERT（预热更新步数15 000） | 5e-4 | 15 000 | 无法收敛 |
| | 2e-4 | 15 000 | 31.33 |
| | 1.2e-4 | 15 000 | 32.22 |
| | 1e-4 | 15 000 | 31.78 |
| | 8e-5 | 15 000 | **32.41(+0.62)** |
| | 4e-5 | 15 000 | 32.12 |
| 两段式微调BERT（学习率8e-5） | 8e-5 | 4 000 | 31.95 |
| | 8e-5 | 8 000 | 31.95 |
| | 8e-5 | 10 000 | 31.88 |
| | 8e-5 | 12 500 | 32.20 |
| | 8e-5 | 15 000 | **32.41(+0.62)** |
| | 8e-5 | 17 000 | 31.66 |

## 3 结束语

本文针对汉语-维吾尔语平行语料资源匮乏的问题,将BERT-fused模型应用于汉维机器翻译,通过一系列对比实验总结得到在汉维机器翻译中应用预训练语言模型BERT的最佳方法。将本文提出的两段式微调BERT的方法在CMWT 2017评测语料上进行实验,结果表明,该方法能够显著提高汉维机器翻译的性能。后续将研究如何把预训练语言模型应用到维吾尔语-汉语的机器翻译任务中,进一步提高维汉机器翻译的效果。

## 参考文献


[1] SUTSKEVER I, VINYALS O, QUOC V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2014: 3104-3112.

[2] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. (2016-05-19)[2020-09-10]. https://arxiv.org/pdf/1409.0473.pdf.

[3] MENG F, ZHANG J. DTMT: a novel deep transition architecture for neural machine translation[C]//Proceedings of 2019 AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2019: 224-231.

[4] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning. New York, USA: ACM Press, 2017: 1243-1252.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 6000-6010.

[6] 哈里旦木·阿布都克里木,刘洋,孙茂松. 神经机器翻译系统在维吾尔语汉语翻译中的性能对比[J]. 清华大学学报(自然科学版),2017,57(8):878-883.
ABUDUKELIMU H, LIU Y, SUN M S. Performance comparison of neural machine translation systems in Uyghur-Chinese translation[J]. Journal of Tsinghua University(Science and Technology), 2017, 57(8): 878-883. (in Chinese)

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24) [2020-09-10]. https://arxiv.org/pdf/1810.04805.pdf.

[8] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2020-09-10]. https://arxiv.org/pdf/1907.11692v1.pdf.

[9] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[10] 李俊,吕学强. 融合BERT语义加权与网络图的关键词抽取方法[J]. 计算机工程,2020,46(9):89-94.
LI J, LÜ X Q. Keyword extraction method based on BERT semantic weighting and network graph[J]. Computer Engineering, 2020, 46(9): 89-94. (in Chinese)

[11] RAJPURKAR P, JIA R, LIANG P. Know what you don't know: unanswerable questions for SQuAD[EB/OL]. (2018-06-11)[2020-09-10]. https://arxiv.org/pdf/1806.03822.pdf.

[12] ZHANG H, XU J, WANG J. Pretraining-based natural language generation for text summarization[EB/OL]. (2019-02-25)[2020-09-10]. https://arxiv.org/pdf/1902.09243v2.pdf.

[13] CLINCHANT S, JUNG K W, NIKOULINA V. On the use of BERT for neural machine translation [EB/OL]. (2019-09-27)[2020-09-10]. https://arxiv.org/pdf/1909.12744.pdf.

[14] LI L, JIANG X, LIU Q. Pretrained language models for document-level neural machine translation [EB/OL]. (2019-11-08)[2020-09-10]. https://arxiv.org/pdf/1911.03110.pdf.

[15] ZHU J, XIA Y, WU L, et al. Incorporating BERT into neural machine translation[EB/OL]. (2020-02-17)[2020-09-10]. https://arxiv.org/pdf/2002.06823.pdf.

[16] BERT-base-multilingual-uncased model[EB/OL]. [2020-09-10]. https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip.

[17] BERT-base-Chinese model[EB/OL]. [2020-09-10]. https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip.

[18] BERT-wwm-ext model[EB/OL]. [2020-09-10]. https://drive.google.com/file/d/1iNeYFhCBJWeUsIlnW_2K6SMwXkM4gLb_/view.

[19] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[EB/OL]. (2020-02-17)[2020-09-10]. https://arxiv.org/pdf/1906.08101v2.pdf.

[20] RoBERTa-wwm-large-ext model[EB/OL]. [2020-09-10]. https://drive.google.com/open?id=1-2vEZfIFCdM1-vJ3GD6DlSyKT4eVXMKq.

[21] RoBERTa-wwm-ext model[EB/OL]. [2020-09-10]. https://drive.google.com/open?id=1eHM3l4fMo6DsQYGmey7UZGiTmQquHw25.

[22] RBTL3 model[EB/OL]. [2020-09-10]. https://drive.google.com/open?id=1qs5OasLXXjOnR2XuGUh12NanUl0pkjEv.

[23] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2019: 3651-3657.

[24] McCLOSKEY M, COHEN N J. Catastrophic interference in connectionist networks: the sequential learning problem[J]. Psychology of Learning and Motivation. 1989, 24: 109-165.

[25] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[EB/OL]. (2016-06-03)[2020-09-10]. https://arxiv.org/pdf/1508.07909v4.pdf.

[26] OTT M, EDUNOV S, BAEVSKI A, et al. fairseq: a fast, extensible toolkit for sequence modeling [EB/OL]. (2019-04-01)[2020-09-10]. https://arxiv.org/pdf/1904.01038.pdf.

[27] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2002: 311-318.


编辑 金胡考