doi:10.6043/j.issn.0438-0479.202312017

融合 BERT 预训练语言知识的神经机器翻译方法

谷雪鹏[1],郭军军[1,2*],余正涛[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:[**目的**]针对在神经机器翻译任务中仅使用微调的方法不能充分利用预训练语言知识的问题进行研究.[**方法**]提出一种双阶段交互融合预训练模型的神经机器翻译方法.首先提取 BERT 预训练模型的多层表征,利用多层表征构建掩码知识矩阵,将 BERT 包含的预训练知识作用于神经机器翻译模型编码端词嵌入层.其次,通过自适应融合模块提取 BERT 多层表征中的有益知识,并与神经机器翻译模型交互融合.[**结果**]实验结果表明,与 Transformer 基线模型相比,所提方法在多个神经机器翻译任务上 BLEU 评分获得了 1.41~4.20 的提升,相较于其他融合预训练知识的神经机器翻译方法,所提方法也有较为明显的模型性能提升.[**结论**]本文提出的双阶段交互融合预训练模型的神经机器翻译方法缓解了灾难性遗忘问题,缩小了预训练模型与神经机器翻译模型因训练目标不同而导致的差异,可以有效利用预训练语言知识来提升神经机器翻译模型性能.

关键词:机器翻译;预训练语言模型;注意力机制;Transformer 网络模型
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2024)06-1024-09

Neural machine translation method integrating BERT's pre-trained language knowledge

GU Xuepeng[1],GUO Junjun[1,2*],YU Zhengtao[1,2]
(1. Faculty 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 Technol,Kunming 650500,China)

Abstract:[**Objective**] To study the problem that only fine-tuning method can not make full use of pre-trained language knowledge in neural machine translation task. [**Methods**] A neural machine translation method based on two-stage interactive fusion of pre-trained models is proposed. First, the multi-layer representation of BERT pre-trained model is extracted, then the mask knowledge matrix is constructed by using the multi-layer representation, and the pre-training knowledge contained in BERT is applied to the encoding word embedding layer of neural machine translation model. Second, the beneficial knowledge obtained from BERT multilayer representation is extracted by adaptive fusion module and interactively fused with neural machine translation model. [**Results**] Experimental results show that, compared with Transformer baseline model, the proposed method achieves an improvement of BLEU score of 1.41~4.20 in multiple neural machine translation tasks. Compared with other neural machine translation methods that integrate pre-training knowledge, the proposed method also secures a significant model performance improvement. [**Conclusion**] The neural machine translation method proposed herein, which combines pre-trained models with two-stage interaction, resolves the problem of catastrophic forgetting, reduces the difference between pre-trained models and neural machine translation models due to different training objectives, and can effectively use pre-trained language knowledge to improve the performance of neural machine

收稿日期:2023-12-18 **录用日期**:2024-03-29
基金项目:国家重点研发计划(2020AAA0107904);国家自然科学基金(62366025);云南省自然科学基金(2019FB082,2019QY1801);云南省重大科技专项(202002AD080001,202103AA080015,202202AD080003)
* **通信作者**:guojjgb@163.com
引文格式:谷雪鹏,郭军军,余正涛.融合 BERT 预训练语言知识的神经机器翻译方法[J].厦门大学学报(自然科学版),2024,63(6):1024-1032.
Citation:GU X P,GUO J J,YU Z T. Neural machine translation method integrating BERT's pre-trained language knowledge[J]. J Xiamen Univ Nat Sci,2024,63(6):1024-1032. (in Chinese)

translation models.

Keywords: machine translation; pre-trained language model; attention mechanism; transformer network model

神经机器翻译(neural machine translation,NMT)是机器翻译的主流方法,经过不断创新与完善[1-2],NMT 模型表现出强大的翻译性能,然而 NMT 模型的性能仍依赖于大规模的双语数据,自然界中单语数据更易获取,且成本较低,研究如何利用现有大规模单语数据增强 NMT 模型翻译效果具有重要意义. 预训练语言模型依靠大规模单语数据训练,如 BERT[3]、ELMo[4]、XLM[5]等在自然语言处理领域受到广泛关注. 微调预训练语言模型在许多自然语言理解任务中取得了最好的效果,如文本分类、阅读理解等. 首先,以自监督的方式在大规模单语语料库上对预训练模型进行训练,然后使用任务特定的损失函数和数据集对下游任务进行模型微调. 在自然语言理解任务上对预训练模型进行微调非常简单,通常可以将预训练语言模型视为特征提取器,并在此基础上引入新的分类模块. 然而,对于序列到序列框架的 NMT 任务[6],由于单语预训练语言模型与 NMT 模型训练目标的不同,利用预训练语言模型初始化 NMT 参数并微调的方法收效甚微,Imamura 等[7]使用预训练语言模型替换 NMT 模型编码器,在训练过程中采用解码器训练和微调两阶段优化,取得了较好效果,而直接微调的方法性能低于基线模型. BERT 为目前最经典的预训练语言模型之一,研究如何利用 BERT 增强 NMT 模型性能具有重要意义. 许多研究在这方面做了探索,并提出了将 NMT 模型与 BERT 相结合的方法,但如何最大限度地在 NMT 中利用 BERT 知识仍然是一个悬而未决的问题. 首先,以往利用 BERT 增强 NMT 模型的研究大多只使用 BERT 的最后一层表示,没有充分利用中间层. 然而现有研究表明,BERT 的不同层编码了不同语言特征[8-11],BERT 的低层网络包含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 BERT 的高层网络则学习到了丰富的语义信息特征. 其次,Yang 等[12]的研究表明预训练语言模型在 NMT 编码端融合知识更高效,在解码端效果较差,如何在 NMT 编码端以有效的方式利用 BERT 知识还有很多值得研究的地方.

针对上述问题,本文提出了一种有效将预训练语言模型与 NMT 模型交互融合的方法,旨在充分利用预训练语言模型知识提升 NMT 模型效果. 将 BERT 的不同层信息通过掩码知识矩阵和自适应融合模块与 NMT 模型融合,提升 NMT 模型对语言的表征能力. 本文的主要贡献如下:

1) 提出一种利用 BERT 的多层表征分别与 NMT 模型的词嵌入层与编码端输出交互融合的方法,以提升 NMT 模型翻译性能.

2) 通过构造掩码知识矩阵,在 NMT 模型词嵌入层交互融合 BERT 的多层知识,减小单语预训练语言模型与 NMT 模型的语义差异. 通过自适应融合模块,将 BERT 的多层表征与 NMT 编码器输出交互融合,自适应筛选 BERT 多层表征中相匹配的知识.

3) 在多个神经机器翻译任务上评估了所提模型,实验结果表明,本文所提方法能够有效将 BERT 包含的预训练语言知识与 NMT 模型融合,并显著提升了基线模型的翻译性能.

1　相关工作

1.1　预训练语言模型

在自然语言处理领域,已经提出了许多从大规模单语数据中学习上下文信息的预训练语言模型. 预训练语言模型旨在通过自监督从大量文本语料库中学习强大的上下文语言表示. Peters 等[4]提出的 ELMo 模型以长短时记忆网络(long short-term memory,LSTM)为基础架构,构造了双向 LSTM 结构,能够提取输入序列的语义和句法信息,并成功地将其应用于问答系统、文本蕴涵等任务. 受其启发,Radford 等[13]提出的 GPT 模型使用基于自注意网络的语言模型代替双向 LSTM 结构,进一步提高了预训练模型的性能. 随后,Devlin 等[3]提出基于 Transformer 编码器的降噪自编码语言模型 BERT,其预训练目标包含掩码语言模型和下一句预测,掩码语言模型类似于完形填空,给定一个输入句子 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$,随机选择 $\boldsymbol{x}$ 中的一小部分(通常是 15%)标记并用一个特殊符号[MASK]替换,然后模型尝试基于序列中其他未被掩码的词来预测被掩码的词. 下一句预测任务的目的主要是更好地理解两个句子的关系,预测两个输入序列是否相邻. 之后,一些变种模型被相继提出,如 Song 等[14]提出序列到序列语言模型 MASS,能够对编码器和解码器联合训练来提高特征抽取和语言模型的表达能力. Xu 等[15]提出一种双语预训练语言模型 BIBERT,使源语言和目标语言数据能够丰富彼此的语义信息,从而更好地进行双向翻译. 但由于这些模型配备了大

规模的训练语料库,因此从头开始训练需要耗费大量的时间和资源. 在本文中,我们专注于在 NMT 模型中利用开源的预训练语言模型 BERT.

1.2 预训练语言模型在 NMT 中的应用

为了进一步提高 NMT 的性能,最近的一些研究试图将预训练语言模型,特别是 BERT,融合到 NMT 模型中,探索如何在 NMT 模型中更好地利用预训练语言模型输出的上下文信息. Lample 和 Conneau[5]利用预训练语言模型的参数初始化 NMT 模型编码器和解码器的参数,表明了参数初始化的方法有助于无监督神经机器翻译. Edunov 等[16]使用预训练语言模型输出的上下文向量初始化 NMT 模型的编码器或解码器的词嵌入层,使用此种方法可以提高 NMT 模型翻译性能,但在使用微调时几乎没有什么收获,而当 NMT 的可用语料增多时,获得的增益将减小. 这意味着用预训练语言模型输出初始化 NMT 参数的方法往往面临灾难性遗忘[17]问题. 为了解决灾难性遗忘问题,Yang 等[18]提出渐进蒸馏的方法,让 BERT 作为教师模型,指导 NMT 模型翻译. Weng 等[19]提出在 NMT 模型编码端采用动态融合,解码端采用知识蒸馏的方法,在词级和句子级两种粒度上提取 BERT 知识. Chen 等[20]提出条件掩码语言模型(C-MLM)对 BERT 进行微调,然后利用微调后的 BERT 作为教师模型来改进用于文本生成的序列到序列模型. Zhu 等[21]将 BERT 最后一层的输出通过交叉注意力机制分别融合到 NMT 模型编码器和解码器的每一层,有效利用了 BERT 中的语义信息. Guo 等[22]通过引入轻量级 adapter 模块,将来自源语言和目标语言域的两个预训练的 BERT 模型集成为一个序列到序列模型,该方法在训练时只微调 adapter 模块参数,有效解决了灾难性遗忘问题. Zhang 等[23]提出使用联合注意力机制将 BERT 输出融合到 NMT 模型中,并采用三阶段优化策略训练所提出的模型,逐步解冻不同组件,以克服微调期间的灾难性遗忘. 最近,ChatGPT 的出现给自然语言处理任务带来了显著的影响,Jiao 等[24]对 ChatGPT 用于 NMT 任务做了初步的评估,发现在高资源场景下,ChatGPT 的翻译效果与商用翻译系统相当,但在低资源场景下却明显落后.

2 双阶段交互融合预训练模型的 NMT 方法

在双语数据有限的情况下,NMT 模型很难生成合适的句子上下文表征,预训练语言模型可以为 NMT 模型提高有效的语言知识补充. 然而,利用预训练语言模型初始化 NMT 模型参数并微调的方法并不适合 NMT 任务. 因此,本文提出一种双阶段交互融合预训练模型的 NMT 方法. 在 NMT 模型编码端词嵌入阶段,构造掩码知识矩阵,在词嵌入层自适应地融合 BERT 表征中对 NMT 任务有益的语言知识;在 NMT 模型编码端输出阶段,通过自适应融合模块筛选 BERT 中的有效信息,增强编码端对源语言句子的上下文表征能力. 图 1 为本文所提模型的整体框架.

2.1 BERT 多层输出融合

为了充分发掘 BERT 不同层输出包含的预训练知识,本文设计了一种简单且有效的 BERT 多层输出融合方法,通过对 BERT 不同层输出融合,选择性地输出 BERT 不同层的有效信息. 具体实现过程如下:

$$\hat{\boldsymbol{B}} = \mathrm{Concat}(B_1, B_2, \cdots, B_n), \tag{1}$$

$$\boldsymbol{B} = \mathrm{Glu}(\hat{\boldsymbol{B}}\boldsymbol{W}), \tag{2}$$

其中,$B_1, B_2, \cdots, B_n$ 表示 BERT 的每一层词级表征输出,Concat(·)表示对 BERT 的多层输出拼接,$\hat{\boldsymbol{B}} \in \mathbb{R}^{b\times l\times 768\times 12}$,$b$ 表示训练批次,l 表示句子长度,Glu(·)为激活函数,$\boldsymbol{W} \in \mathbb{R}^{12\times 2}$ 为线性变化矩阵,$\boldsymbol{B} \in \mathbb{R}^{b\times l\times 768}$ 为 BERT 多层融合后的输出.

2.2 掩码知识矩阵

之前的研究表明直接用 BERT 输出作 NMT 模型的嵌入层,面临灾难性遗忘问题. 为了更好地利用 BERT 中的有效信息,本文通过构建掩码知识矩阵,将 BERT 知识在词嵌入阶段融入 NMT 模型. 首先,取 BERT 多层融合输出 $\boldsymbol{B}$ 的[CLS]表征 $\boldsymbol{B}_{\mathrm{CLS}}$,$\boldsymbol{B}_{\mathrm{CLS}}$ 包含了输入序列的全局信息,这使得它在执行下游任务时,对整体序列信息有很好的表示效果. 将 $\boldsymbol{B}_{\mathrm{CLS}}$ 与编码端词嵌入层 $\boldsymbol{E}$ 做相似度计算,得到相似度矩阵 $\boldsymbol{A}$,具体计算过程如下:

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{\boldsymbol{E}(\boldsymbol{B}_{\mathrm{CLS}}\boldsymbol{w}_1 + b)^{\mathrm{T}}}{\sqrt{d}}\right), \tag{3}$$

其中:$\boldsymbol{w}_1 \in \mathbb{R}^{768\times 512}$ 为可训练参数矩阵,使 $\boldsymbol{B}_{\mathrm{CLS}} \in \mathbb{R}^{b\times 768}$ 的最后一维维度转换为 512,b 为偏置,d 的值设置为 512,相似度矩阵 $\boldsymbol{A} \in \mathbb{R}^{b\times l}$.

再对相似度矩阵 $\boldsymbol{A}$ 中元素 a_i 二值化,得到掩码知识矩阵 $\boldsymbol{M}$,具体如下:

$$m_i = \begin{cases} 1, a_i < \theta, \\ 0, a_i \geqslant \theta, \end{cases} \tag{4}$$

其中:m_i 为 $\boldsymbol{M} \in \mathbb{R}^{b\times l}$ 掩码知识矩阵中的元素,θ 为阈值,本文设置为 0.03.

将掩码知识矩阵 $\boldsymbol{M}$ 作用于编码端词嵌入层得到

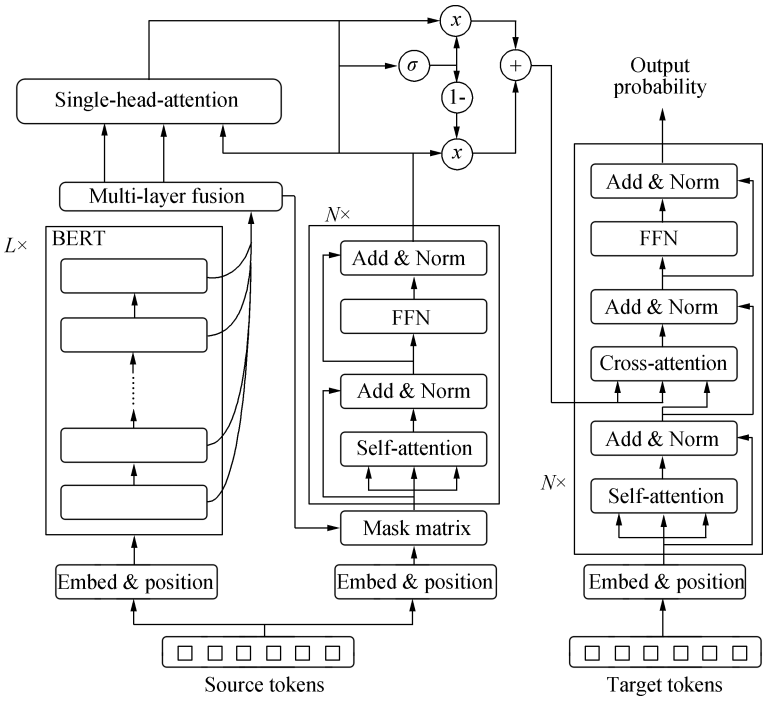

图 1　模型整体框架

Fig. 1　Overall framework of the model

$\widetilde{\boldsymbol{E}}$,具体地,首先将 $\boldsymbol{M}$ 矩阵的最后一维扩维得到 $\boldsymbol{M}'\in\mathbb{R}^{b\times l\times 1}$,将 $\boldsymbol{E}$ 中对应 $\boldsymbol{M}'$ 中元素值为 1 的词表征填充为一个很小的值,本文将填充值设置为 -1×10^{-9},减小词嵌入层中不相关信息的干扰,并通过门控选择 $\boldsymbol{B}$ 和 $\boldsymbol{E}$ 中的有效信息,与 $\widetilde{\boldsymbol{E}}$ 相加,作为新的编码端词嵌入表征 $\overline{\boldsymbol{E}}$,具体流程如图 2 所示.

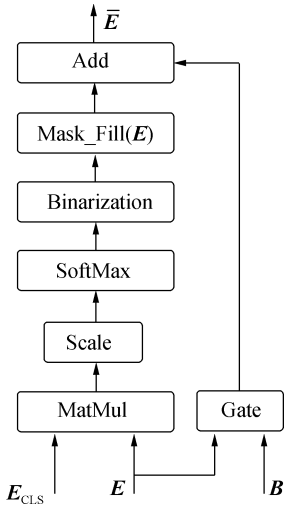

图 2　掩码知识矩阵流程

Fig. 2　Mask knowledge matrix

门控输出 $\boldsymbol{G}$ 的具体实现方法如下:

$$\boldsymbol{G}=\lambda\boldsymbol{B}+(1-\lambda)\boldsymbol{E},\tag{5}$$

$$\lambda=\mathrm{Relu}((\mathrm{Concat}(\boldsymbol{B}\boldsymbol{w}_3,\boldsymbol{E}))\boldsymbol{w}_2),\tag{6}$$

其中,$\boldsymbol{w}_2$、$\boldsymbol{w}_3$ 为可训练参数矩阵,Relu(·)为激活函数.

2.3　自适应融合模块

自适应融合模块包含两部分,分别是单头注意力机制和选择性融合模块,单头注意力机制利用 NMT 模型编码端输出信息关注 BERT 中的有效信息,减小 BERT 输出与 NMT 模型编码端输出差异性干扰. 选择性融合模块自适应融合经单头注意力机制输出的 BERT 信息与 NMT 模型编码端输出信息,并送入解码端.

单头注意力机制将 BERT 多层输出信息 $\boldsymbol{B}$ 与 NMT 模型编码端输出信息 $\boldsymbol{E}_N$ 关联起来,其中查询向量、键向量和值向量分别是 $\boldsymbol{E}_N$、$\boldsymbol{B}$ 和 $\boldsymbol{B}$,单头注意力机制的输出 $\boldsymbol{E}^{\boldsymbol{B}}_{\mathrm{attn}}$ 由以下方式计算:

$$\boldsymbol{E}^{\boldsymbol{B}}_{\mathrm{attn}}=\mathrm{Softmax}\left(\frac{\boldsymbol{E}_N\boldsymbol{B}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{B}.\tag{7}$$

选择性融合模块自适应的融合单头注意力机制输出的 BERT 信息 $\boldsymbol{E}^{\boldsymbol{B}}_{\mathrm{attn}}$ 和 NMT 模型编码端输出信息 $\boldsymbol{E}_N$,送入解码端. 融合后的信息 $\boldsymbol{E}^{\mathrm{out}}$ 可以被定义为:

$$\eta=\mathrm{Sigmoid}(\boldsymbol{U}\boldsymbol{E}_N+\boldsymbol{V}\boldsymbol{E}^{\boldsymbol{B}}_{\mathrm{attn}}),\tag{8}$$

$$\boldsymbol{E}^{\mathrm{out}}=(1-\eta)\boldsymbol{E}_N+\eta\boldsymbol{E}^{\boldsymbol{B}}_{\mathrm{attn}},\tag{9}$$

其中:$\boldsymbol{U}$、$\boldsymbol{V}$ 为可训练权重矩阵,Sigmoid(·)为激活函数,$\eta\in[0,1]$.

2.4 训练策略

本文在训练策略上参照 Zhu 等[21]提出的 warm up 方法,先训练一个 Transformer-base 模型至收敛,然后用得到的模型初始化本文所提模型对应的 NMT 模型编-解码器.

3 实验与分析

3.1 数据集和预处理

对于低资源的场景,本文在 IWSLT 的 4 个翻译任务上对所提出的模型进行了评估,分别是 IWSLT'14 英语-德语(En-De)、IWSLT'14 德语-英语(De-En)、IWSLT'15 英语-越南语(En-Vi)、IWSLT'17 英语-法语(En-Fr). 本文对数据集中的语料都进行了 BPE 分词[25]处理,合并语言对的源语言和目标语言构建联合词典,词典大小为 10K. 对 IWSLT'14 德-英任务,本文遵循 Edunov[26]的设置,小写所有单词,验证集是通过从训练集中抽取 7k 个句子构建的,测试集是通过拼接多个 TED Talk 文件构建的. 对于其他 IWSLT 任务,将相应年份的 TED Talk 文件用作验证/测试集.

对于高资源的场景,本文在 WMT 的 2 个翻译任务上对所提出的模型进行了评估,分别是 WMT'14 英语-德语(En-De)、和 WMT'14 英语-法语(En-Fr). 对数据集中的语料都进行了 BPE 分词[25]处理,表 1 给出了数据集的划分方式和大小.

表 1 数据集结构

Tab. 1 Dataset structure

翻译任务	训练集	验证集	测试集
IWSLT'14 En-De	160k	{从训练集划分 7k 个句子}	{dev2010,dev2012,tst2010,tst2011,tst2012}
IWSLT'15 En-Vi	113k	{tst2012}	{tst2013}
IWSLT'17 En-Fr	236k	{tst2011,tst2012,tst2013,tst2014,tst2015}	{tst2016,tst2017}
WMT'14 En-De	4.5M	{newstest2012,newstest2013}	{newstest2014}
WMT'14 En-Fr	36M	{newstest2012,newstest2013}	{newstest2014}

3.2 实验设置

本文使用预训练语言模型 BERT 提取源语言的上下文表征,在提取上下文表征时,BERT 的参数冻结,以减少训练的参数和时间,具体来说,我们使用 bert-base-uncased 提取英语的下文表征,使用 bert-base-german-uncased 提取德语的上下文表征. 本文所提的模型基于 Facebook 开源工具包 Fairseq[27]实现,采用 Transformer_iwslt_de_en 作为基础模型,NMT 模型编-解码器的层数均为 6,词向量维度为 512,前馈神经网络中间层维度为 1 024,注意力头数为 4. 模型训练时使用 Adam 作为优化器,参数 $\beta_1=0.9$,$\beta_2=0.98$. 权重衰减 weight-decay 为 0.000 5,warmup 步数为 4 000,学习率 $lr=0.000\ 1$. 损失函数采用交叉熵损失函数,编-解码器 dropout 设为 0.3,单头注意力机制的 dropout 设为 0.2. 根据句子长度划分为不同的 batch,max-token 参数设为 4 096. 本文所有实验均在一张 NVIDIA GeForce RTX 3090 显卡上实验.

3.3 对比模型

本文与相同任务的其他基线模型做了对比,包括利用预训练语言知识增强 NMT 的模型和其他 NMT 模型.

首先是利用预训练知识增强 NMT 的模型:

1) C-MLM[20]:提出条件掩码语言模型,以实现 BERT 的微调,然后利用微调后的 BERT 作为额外的监督来改进用于文本生成的传统序列到序列模型.

2) BERT-fused[21]:将 BERT 最后一层输出通过注意力机制分别与 NMT 模型编-解码器每一层融合,充分利用 BERT 知识.

3) AB-NET[22]:通过引入轻量级 adapter 模块,将来自源语言和目标语言域的两个预训练的 BERT 模型集成为一个序列到序列模型.

4) XLM-R_ENC[28]:将 XLM-R 跨语言预训练语言模型应用在 NMT 模型的源语言端,提高译文质量.

5) BERT-JAM[23]:提出分别使用联合注意力机制和交叉注意力机制将 BERT 输出与 NMT 编-解码端每一层融合,并采用三阶段优化策略训练所提出的模型,逐步解冻不同组件.

其次是在相同任务上的其他 NMT 模型:

1) DynamicConv[29]:提出轻量化的动态卷积神经网络来解决由于长输入序列所导致的注意力机制计算量过大的问题.

2) Tied-Transformers[30]:通过共享编-解码器权重改进 NMT 模型翻译效果.

3.4　实验结果

表 2 展示了本文所提模型与对比模型在 IWSLT'14 英语-德语(En-De)、IWSLT'14 德语-英语(De-En)、IWSLT'15 英语-越南语(En-Vi)、IWSLT'17 英语-法语(En-Fr)等翻译任务上的实验结果.

表 2　不同方法在 IWSLT 任务上的 BLEU 值

Tab. 2　BLEU of different methods on the IWSLT task

模型	En-De	De-En	En-Fr	En-Vi
C-MLM	—	35.60	—	31.51
BERT-fused	30.34	36.11	38.70	31.97
AB-NET	—	36.45	—	—
XLM-R_ENC	29.07	—	—	31.39
BERT-JAM	**31.20**	**38.66**	39.80	—
DynamicConv	—	35.20	—	—
Tied-Transformers	29.07	35.10	—	—
Transformer	28.57	34.64	35.90	30.76
Ours	30.70	36.05	**40.10**	**33.03**

表 3 展示了本文所提模型与对比模型在 WMT'14 英语-德语(En-De)和 WMT'14 英语-法语(En-Fr)翻译任务上的实验结果.

表 3　不同方法在 WMT 任务上的 BLEU 值

Tab. 3　BLEU of different methods on the WMT task

模型	En-De	En-Fr
BERT-fused	30.75	43.78
AB-NET	30.60	43.56
XLM-R_ENC	29.09	—
BERT-JAM	**31.59**	—
DynamicConv	29.70	43.20
Transformer	29.30	41.30
Ours	31.47	**43.96**

可以看到,本文所提的方法在表中所有翻译任务上都优于 Transformer 基线模型,相较于 Transformer 基线模型,BLEU 值提高了 1.41～4.20,表明本文方法可以有效在 NMT 模型中利用预训练语言知识.

与其他利用预训练知识提升 NMT 性能的模型相比,本文在 IWSLT17'En-Fr、IWSLT'15 En-Vi 和 WMT'14 En-Fr 任务上取得了最好的效果,在其他翻译任务上也取得了较好的效果,表明本文所提方法在低资源和高资源场景下,对 NMT 模型的翻译效果均有较好的提升. 与 BERT-fused 模型在编码端和解码端的每一层融合 BERT 知识的方法不同,本文仅在编码端融合 BERT 知识,减小了训练参数量的同时,在 IWSLT'14 En-De、IWSLT'15 En-Vi、IWSLT'17 En-Fr 翻译任务上,BLEU 值分别提高了 0.36、1.06、1.40,说明本文所提方法可以更高效的发掘 BERT 中的预训练语言知识,本文的方法在英语-德语互译任务上未超过 BERT-JAM 模型,本文认为可能的原因之一是 BERT-JAM 提出了三阶段优化的策略,在不同的训练阶段解冻不同的模块,以达到最好的效果,其次是提出的联合注意力机制对于德语-英语这类属于同一语系,且具有相似语言结构特征的翻译任务具有较好的学习能力,但在语言差异性较大的翻译任务上,如英语到法语、英语-越南语翻译任务,效果则没有那么出色. 本文的方法通过在词嵌入层引入掩码矩阵,可以通过预训练语言知识屏蔽词嵌入层中的不相关信息,减小 BERT 输出和 NMT 模型嵌入层的差异性,并在编码端输出阶段自适应选择预训练语言模型中的有效信息,增强 NMT 模型对语言的表征和学习能力,进而提升 NMT 模型的翻译性能.

与相同任务的其他 NMT 方法相比,本文的方法也展现出了较好的优势,表明预训练语言知识在 NMT 模型中的有效性.

总的来说,本文所提的方法能够有效利用预训练语言模型中的知识,通过掩码矩阵和自适应融合模块,特别是在语言差异性较大的翻译任务中取得了更好的效果.

3.5　消融实验

为了分析验证各个模块的有效性,本节在 IWSLT'14 En-De 任务上进行了消融研究. 将主模型的部分模块移除得到新模型,具体如下:

M_0:表示完整的主模型;

M_1:表示去掉 BERT 的多层融合模块,仅使用 BERT 最后一层的输出;

M_2:表示去掉掩码知识矩阵;

M_3:表示去掉自适应融合模块;

表 4 给出了消融实验的结果:

表 4 消融实验结果

Tab. 4 Ablation experiment results

模型	En-De
M_0	30.70
M_1	30.44
M_2	30.52
M_3	28.82

消融实验结果表明,本文所提模型的各个模块均有助于 NMT 模型更好的利用 BERT 预训练语言知识. M_1 表明 BERT 的多层知识有助于 NMT 模型更好的对句子表征,BERT 的不同层包含的不同语言特征,如短语级别信息表征、语言学信息表征、语义信息表征对 NMT 任务是有益的. M_2 表明本文提出的掩码知识矩阵可以生成更好的 NMT 编码端词嵌入层表征,在 NMT 模型词嵌入层交互融合 BERT 的多层知识,成功减小了单语预训练语言模型与 NMT 模型的语义差异. M_3 表明本文设计的自适应融合模块可以有效筛选 BERT 中的有益信息,促进 BERT 的多层知识与 NMT 编码端更好地融合,以提高机器翻译性能.

为了更好地理解不同掩码阈值对模型性能的影响,本节还在 IWSLT 任务上对掩码阈值 θ 进行了消融实验,实验结果如表 5 所示.

表 5 不同掩码阈值对实验结果的影响

Tab. 5 Influences of different mask thresholds on experimental results

θ	掩码阈值					
	0.00	0.01	0.02	0.03	0.04	0.05
En-De	30.52	30.60	30.63	30.70	30.53	30.49
De-En	35.60	35.67	36.03	36.05	35.77	35.63
En-Fr	38.89	40.01	40.06	40.10	39.93	38.95
En-vi	32.55	32.80	32.97	33.03	32.73	32.51

从表中数据可以看出,当掩码阈值 θ 取 0.03 时,所提方法获得了更好的翻译效果. 当 θ 取值从 0.00 增长到 0.03 时,模型性能也得到了相应的提升,表明掩码矩阵对 NMT 模型的词嵌入表征是有益的,分析其原因是掩码矩阵根据 BERT 多层输出的[CLS]表征 $\boldsymbol{B}_{\mathrm{CLS}}$ 与 NMT 模型编码端词嵌入层 $\boldsymbol{E}$ 计算相似度得到,可以对 $\boldsymbol{E}$ 中语义差异较大的特征掩码,增强模型对语言的学习与表征能力,进而提升 NMT 模型的翻译性能. 当 θ 取值从 0.03 增长到 0.05 时,模型的性能反而有所下降,分析其原因是较高的掩码阈值对 NMT 模型编码端词嵌入层的有益信息也做了掩码,影响了 NMT 模型的翻译性能.

3.6 实例分析

为了进一步验证本文方法的有效性,表 6 给出了在 IWSLT'14 De-En 任务上,Transformer 基线模型和本文方法生成的译文示例.

表 6 译文示例

Tab. 6 Translation example

源语言	manche menschen waren enttäuscht dass es keine poesie war.
参考译文	some people were disappointed there was not poetry.
Transformer	some people were a little disappointed we didn't use it for a poem.
Ours	some people were disappointed that it wasn't poetry.
源语言	ich denke, dass das geändert werden kann indem man rekombinante impfstoffe im vorhinein erstellt.
参考译文	i think that can be changed by building combinatorial vaccines in advance.
Transformer	i think this approach could start with a recombinant vaccine
Ours	i think that can be changed by creating recombinant vaccines in advance.

由表 6 的翻译示例可以看到,本文所提模型相较于 Transformer 基线模型,在语义的正确表征上有较为明显的提高,更加接近参考译文,表明本文所提模型有效利用了 BERT 中包含的大规模语义信息,提升了 NMT 的翻译性能.

4 结 论

在本文中,我们探索了如何在 NMT 模型编码端最大限度利用 BERT 预训练语言模型,在编码端词嵌入阶段,引入掩码知识矩阵,用 BERT 的多层预训练语言知识指导词嵌入生成,在编码端输出阶段,通过自适应融合模块,筛选预训练语言模型中对 NMT 任务的有效信息,增强编码端对句子的表征能力. 综合实验表明,本文所提方法在多个翻译任务上取得较好

的 BLEU 分数,表明本文所提方法能够有效在 NMT 任务中利用预训练语言知识.

参考文献:

[1] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2014-09-01) [2023-12-01]. http://arxiv.org/abs/1409.0473.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-01-12) [2023-12-01]. http://arxiv.org/abs/1706.03762.

[3] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11) [2023-12-01]. http://arxiv.org/abs/1810.04805.

[4] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 2227-2237.

[5] LAMPLE G, CONNEAU A. Cross-lingual language model pretraining[EB/OL]. (2019-01-22) [2023-12-01]. http://arxiv.org/abs/1901.07291.

[6] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 4: 3104-3112.

[7] IMAMURA K, SUMITA E. Recycling a pre-trained BERT encoder for neural machine translation[C]// Proceedings of the 3rd Workshop on Neural Generation and Translation. Stroudsburg: Association for Computational Linguistics, 2019: 23-31.

[8] 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. Florence: Association for Computational Linguistics, 2019: 02131630.

[9] BAPNA A, CHEN M A, FIRAT O, et al. Training deeper neural machine translation models with transparent attention[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 3028-3033.

[10] WANG Q, LI F, XIAO T, et al. Multi-layer representation fusion for neural machine translation [EB/OL]. (2020-02-16) [2023-12-01]. http://arxiv.org/abs/2002.06714v1.

[11] DOU Z Y, TU Z P, WANG X, et al. Exploiting deep representations for neural machine translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 4253-4262.

[12] YANG J C, WANG M X, ZHOU H, et al. Towards making the most of BERT in neural machine translation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9378-9385.

[13] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL]. [2023-12-01]. https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf. 2018.

[14] SONG K T, TAN X, QIN T, et al. MASS: masked sequence to sequence pre-training for language generation[EB/OL]. (2019-06-21)[2023-12-01]. http://arxiv.org/abs/1905.02450.

[15] XU H R, VAN DURME B, MURRAY K. BERT, mBERT, or BiBERT? A study on contextualized embeddings for neural machine translation[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6663-6675.

[16] EDUNOV S, BAEVSKI A, AULI M. Pre-trained language model representations for language generation [C]// Proceedings of the 2019 Conference of the North. Stroudsburg: Association for Computational Linguistics, 2019: 4052-4059.

[17] SERRÀ J, SURÍS D, MIRON M, et al. Overcoming catastrophic forgetting with hard attention to the task [EB/OL]. (2018-05-29) [2023-12-01]. http://arxiv.org/abs/1801.01423.

[18] YANG J C, WANG M X, ZHOU H, et al. Towards making the most of BERT in neural machine translation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9378-9385.

[19] WENG R X, YU H, HUANG S J, et al. Acquiring knowledge from pre-trained model to neural machine translation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9266-9273.

[20] CHEN Y C, GAN Z, CHENG Y, et al. Distilling knowledge learned in BERT for text generation[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7893-7905.

[21] ZHU J, XIA Y, WU L, et al. Incorporating BERT into neural machine translation [EB/OL]. (2020-02-17)

[2023-12-01]. http://arxiv.org/abs/2002.06823.

[22] GUO J L, ZHANG Z R, XU L L, et al. Incorporating BERT into parallel sequence decoding with adapters[J]. Advances in Neural Information Processing Systems, 2020, 33: 10843-10854.

[23] ZHANG Z B, WU S, JIANG D W, et al. BERT-JAM: Maximizing the utilization of BERT for neural machine translation[J]. Neurocomputing, 2021, 460: 84-94.

[24] JIAO W X, WANG W X, HUANG J T, et al. Is ChatGPT A good translator? yes with GPT-4 As the engine[EB/OL]. (2023-11-02) [2023-12-01]. http://arxiv.org/abs/2301.08745.

[25] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2016: 1715-1725.

[26] EDUNOV S, OTT M, AULI M, et al. Classical structured prediction losses for sequence to sequence learning[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 355-364.

[27] OTT M, EDUNOV S, BAEVSKI A, et al. Fairseq: a fast, extensible toolkit for sequence modeling[EB/OL]. (2019-04-01) [2023-12-01]. http://arxiv.org/abs/1904.01038.

[28] 王倩,李茂西,吴水秀,等. 基于跨语种预训练语言模型XLM-R的神经机器翻译方法[J]. 北京大学学报(自然科学版), 2022, 58(1): 29-36.

[29] WU F, FAN A, BAEVSKI A, et al. Pay less attention with lightweight and dynamic convolutions[EB/OL]. (2019-02-22) [2023-12-01]. http://arxiv.org/abs/1901.10430.

[30] XIA Y C, HE T Y, TAN X, et al. Tied transformers: neural machine translation with shared encoder and decoder[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 5466-5473.

(责任编辑:汪 军)