doi:10.6043/j.issn.0438-0479.202110035

结合 BERT 数据增强的基于词切分的蒙汉神经机器翻译系统

何乌云,秀 芝,包晶晶,陈美兰,王斯日古楞*
(内蒙古师范大学计算机科学技术学院,内蒙古 呼和浩特 011500)

摘要:神经机器翻译是目前机器翻译领域主流研究方法,但是蒙汉平行语料的稀缺使得蒙汉神经机器翻译性能难以提升.本文针对基于 Transformer 的蒙汉神经机器翻译系统,利用深度学习模型对蒙古文词切分方法进行研究,分析了蒙古文部分切分、BPE 子词切分和 BiLSTM-CNN-CRF 神经网络切分方法对于蒙汉机器翻译模型的影响,并在此基础上利用基于 BERT(bidirectional encoder representations from Transformers)中文语义相似度计算的数据增强技术去扩充蒙汉机器翻译训练数据.在 CCMT2019 提供的数据集上进行对比实验,实验结果表明,数据增强方法的 BLEU 值相较于基线实验提升显著,且 BLEU4 值达到了 75.28%.

关键词:蒙汉神经机器翻译;Transformer 神经网络;BERT;语义相似度
中图分类号:TP 391 **文献标志码**:A **文章编号**:0438-0479(2022)04-0667-08

蒙汉机器翻译任务属于资源稀缺的少数民族语言翻译任务范畴,这些翻译任务中普遍存在资源匮乏导致的数据稀疏和构词多样性而造成语言特征学习困难的问题.蒙古文形态变化繁杂,而且蒙古文的语序也和汉语语序有所不同,这都成为了蒙汉机器翻译的难题.

自神经机器翻译被提出以来,它就利用大规模的高质量平行语料库来实现端到端的翻译.在数据量充足时,神经机器翻译可以得到良好的翻译结果.但是,在一些资源稀缺的少数民族语言翻译任务中,有时神经机器翻译的性能还不如基于统计的机器翻译.同时,由于训练数据量有限,会导致模型的泛化能力下降.为解决这个问题,一种最直接有效的方法就是通过数据增强方法提高训练数据的多样性,从而提高模型的鲁棒性,避免过拟合.另外,通过改变训练数据,可以降低模型对某些特征或属性的依赖,从而提高模型的泛化能力.

规模庞大的平行语料库是神经机器翻译方法能够获得良好效果的前提.在少数民族机器翻译任务中,平行语料的获取难度是极大的,但单语语料却十分容易获取.合理使用单语语料信息完善翻译模型的翻译能力也成了一种研究课题.其中,数据增强是一个非常重要的研究热点[1].近年来,研究者们尝试使用正向翻译[2]和反向翻译[3]构建伪平行语料库,从而利用相对容易获得的单语数据解决数据资源不足的问题[4].正向翻译的基本思想是训练从源语言到目标语言的翻译模型,将源语言的单语数据翻译成目标语言的句子,与源语言端句子组合成伪双语句对,和真实的双语句对混合作为训练数据用于进一步的训练.反向翻译技术使用目标单语数据来提高模型的翻译质量[5].将目标语言端句子反向翻译后得到源语言端句子,与目标语言端句子组合成伪双语句对,并和真实的双语句对混合作为模型的训练数据,目的是为了采样多样化的训练数据,提升模型的泛化能力.

Sennrich 等[6]在机器翻译任务中,最先提出了回译方法来进行数据增强,该方法操作简单、具有很高的实用性.Fadaee 等[7]和 Sugiyama 等[8]使用回译方法,利用对目标端的单语数据进行反向翻译,生成更

收稿日期:2021-10-27 **录用日期**:2022-03-30
基金项目:国家自然科学基金(61762072);内蒙古自治区科技计划项目(2021GG0139);内蒙古自治区高等学校科学研究项目(NJZY21546);内蒙古师范大学研究生科研创新基金资助项目(GXJJS20129)
* **通信作者**:siriguleng@imnu.edu.cn
引文格式:何乌云,秀芝,包晶晶,等.结合 BERT 数据增强的基于词切分的蒙汉神经机器翻译系统[J].厦门大学学报(自然科学版),2022,61(4):667-674.
Citation:HE W Y,X Z,BAO J J,et al. Mongolian-Chinese neural machine translation system based on word segmentation with BERT data enhancement[J]. J Xiamen Univ Nat Sci,2022,61(4):667-674. (in Chinese)

多的源语言增强版本. Poncelas 等[9]对真实数据、合成数据、混合数据以及回译数据的比例对于翻译效果的影响进行了实证分析,实验数据得出基于采样和加入噪声的束搜索合成数据对翻译效果提升较大. Edunov 等[10]通过实验也同样论证在合成数据中添加噪声数据可以有效提高翻译质量. 谷舒豪等[11]在不同领域的语料中利用了数据增强技术,并验证了其方法的有效性.

基于回译的方法可以不用修改神经机器翻译模型就能有效地利用单语数据;但是当平行语料规模较小时,通过回译生成的伪平行语料质量较差,而且大规模的伪数据和较小规模的平行语料混合使得真实的平行语料难以被有效地利用. 针对这些问题,本文提出一种针对较小规模平行语料的解决方案. 该方法在蒙古文进行词切分的基础上通过 BERT 数据增强的方法去扩充蒙汉机器翻译训练语料库. 该方法利用 BERT 模型可以通过词的上下文信息去获得更好的词向量表示,从而可以针对多语义词区分表示这一特点,进一步计算句子的语义相似度,提升伪平行语料的质量,有效扩充蒙汉机器翻译训练语料库.

1 数据预处理

训练语料来源是 CCMT2021 提供的蒙汉双语平行语料库,包含 26.16 万句对. 由于平行语料的质量会对翻译的效果产生很大影响,因此需要对平行语料进行预处理. 预处理中对传统蒙古文语料库中的非法符号、非 Unicode 字符及平行语料中含有乱码句对进行了过滤处理,主要包括源语言端蒙古文语料里包含的汉语句对以及目标语言端汉语语料里包含的蒙古文句对;其后将改正后的传统蒙古文语料用本实验室开发的转换工具转换成蒙古文拉丁形式.

对目标语言端汉语数据,首先将语料库中的所有字母和数字转换成半角形式. 然后利用中科院开源工具 ICTCLAS2019 对语料进行切分工作. 评测实验中对涉及到的所有语料,包括训练集、开发集、测试集都做了相同的预处理工作.

1.1 蒙古文词切分处理

蒙汉双语语料库非常稀缺,且蒙古文和汉语有着很大的语言差异化. 蒙古文是主宾谓结构,汉语是主谓宾结构,两者的语序有很大的不同. 在蒙汉机器翻译中,由于语法和句法结构的差异,源语言和目标语言的语序完全不同. 蒙古文是一种形式丰富的黏着性语言. 同词根名词和动词的屈折变化较多,表达的概念相似,变形附加成分也具有丰富的词性和时态信息,这些信息对分析机器翻译有着重要的作用. 词级别机器翻译无法捕捉到这些隐藏的含义,同时将同一个词根的不同变体训练成不同的单元,从而增加神经机器翻译词汇词典的规模. 因此,针对蒙古文形态丰富特征和神经机器翻译的有限词汇词典问题,本文中对机器翻译蒙古文语料进行了切分预处理工作. 通过使用部分切分方法、BPE(byte-pair encoding)子词切分方法、基于神经网络的词切分方法对蒙古文语料进行了多种粒度的切分工作.

1.1.1 蒙古文部分切分

对蒙古文语料进行部分切分处理指的是对蒙古文构形附加成分中的分写附加成分的切分. 分写附加成分中包括格附加成分和领属,以及一些名词的复数附加成分. 通过部分切分可以有效减少单频词的数量,会对缓解未登录词和罕见词有很大的帮助,从而可以缓解数据稀疏问题.

1.1.2 蒙古文 BPE 子词切分

BPE 编码技术是在 2016 年,Sennrich 等[12]提出的一种处理文本切分粒度的方法. BPE 算法在机器翻译任务中的应用有两种方法,一种是独立的 BPE,即建立两个独立的词典,分别为源语言词典和目标语言词典. 它的优点是处理后的语料非常的简洁. 另一种是联合 BPE,即源语言和目标语言共同生成一个词典. 它的优点是能够保证源语言和目标语言切分粒度的一致性. 在本研究中只对蒙古文语料进行 BPE 子词切分,所以采用了独立 BPE. BPE 子词粒度切分是将蒙古文句子切分成介于词和字符之间的粒度大小,这种切分方法在一定程度上既能保留蒙古文词的局部特征,又能有效地减小词粒度. 通过 BPE 子词切分方法可以彻底解决集外词问题,任何一个词都可以切分成子词序列.

1.1.3 蒙古文神经网络词切分

传统的蒙古文词切分方法已经相对成熟,但也存在需人工提取文本特征、特征提取方面不够充分,以及模型训练时间过长等一系列问题. 由于神经网络模型有较强的泛化能力,能帮助研究者们从人工编写特征工程中解放出来. 因此,本文构建了一种基于 BiLSTM-CNN-CRF 神经网络模型实现对蒙古文词的切分,并将其应用在蒙汉机器翻译任务当中. BiLSTM-CNN-CRF 神经网络词切分实验采用的数据为少数民族语言分词技术评测 MLWS2017(Minority Language Word Segmentation)活动提供的蒙古文训

练评测语料,规模大小为 50 017 句,共包含 851 765 个词. 数据集随机划分为训练集、开发集和测试集,按 8∶1∶1 的比例去进行划分.

词切分模型中通过 BiLSTM 模型[13]可以有效地模拟远距离依赖关系,提取输入字符序列的上下文信息. CNN 模型[14]一般由卷积层、池化层和全连接层组成. 卷积层主要进行卷积运算;池化层利用数据局部信息重复的特征,对局部信息用总体特征来代表整个区域的输出;全连接层是传统的前馈神经网络,主要是将多个输出拓展为一层全连接形式. 词切分模型中通过 CNN 模型提取蒙古文字符的局部特征信息. CRF 模型[15]是一种无向图模型,它兼备了最大熵模型和隐马尔可夫模型的特征,在序列标注任务中具有较高准确率. 词切分模型中通过 CRF 模型对整个序列进行评分,并加入一些约束条件,保证预测标签的合理性. BiLSTM-CNN-CRF 神经网络模型对蒙古文词进行切分,能够使单频词总数明显减少,有效解决了数据稀疏及未登录词的问题. 通过词切分可使训练集语料总词数减少,能有效降低神经网络机器翻译源语言词汇词典的规模.

1.2　过滤蒙古文语料中控制字符

蒙古文语料中除了代表词边界的天然空格以外,还存在蒙古文特殊控制字符. 现行蒙古文 Unicode 编码中使用控制字符表达字的不同变形. 因此,蒙古文词中会出现编码为“180E”“202F”“180D”等的控制字符,但是在 Transformer 训练工具抽取词汇表时不能正确处理蒙古文词中出现的控制字符,而是把带有控制字符的词分成多个字符串. 为此,本文中在切分工作基础上去掉了语料中出现次数较多的蒙古文 Unicode 编码为“180E”“202F”“180D”的控制字符. 其中,控制字符“180E”是蒙古文元音间隔符,用于词尾的分写元音字母 A/E 与它前面的辅音字母之间;控制字符“202F”是蒙古文窄宽度无间断空格,用于字和附加成分之间,表示有空隙,虽然有空隙但不是词的边界;控制字符“180D”是蒙古文自由变体选择符,用于区别在同一条件下出现的同一个字母的不同的自由变体[16].

2　基于 BERT 数据增强的蒙汉神经机器翻译方法

2.1　蒙汉神经机器翻译模型

神经机器翻译的研究正处于快速发展阶段,尤其是资源匮乏语种的神经机器翻译研究. 近年来,出现了许多不同类型的神经机器翻译方法. 例如,基于循环神经网络[17]、基于卷积神经网络[18]和基于 Transformer[19]的神经机器翻译方法等.

Transformer 翻译模型不再依赖循环神经网络和卷积神经网络,只需要注意力机制就能解决序列到序列的翻译问题,并且能够一步到位获取全局信息. Transformer 模型的整体架构仍然采用编码器-解码器结构,其最大的优点是可以实现高效地并行运算. 且基于 Transformer 的神经机器翻译模型是目前最主流的机器翻译方法. 因此,本研究中在基于 BERT 数据增强的基础上采用 Transformer 进行蒙汉机器翻译.

2.2　BERT 数据增强的蒙汉神经机器翻译系统实现

现有的蒙汉训练语料规模较小,这种情况会直接影响神经网络模型训练的结果. 为了增强蒙汉神经机器翻译中的训练数据规模,本研究在词素切分及去掉语料控制字符预处理方法的基础上采用 BERT[20]数据增强方法去扩充训练语料库. 具体的整体流程框架如图 1 所示.

第一步利用 LCQMC(large-scale Chinese question matching corpus)数据集提供的训练集去训练 BERT 预训练语言模型,得到 BERT 中文语义相似度计算模型,用于之后蒙汉机器翻译语料的中文句子的语义相似度计算;第二步通过蒙汉初始的 Transformer 模型去翻译 CCMT2021 提供的蒙汉机器翻译训练集中的蒙古文单语语料获得相应的中文译文;第三步使用第一步得到的 BERT 中文语义相似度计算模型去计算第二步得到的中文译文和蒙汉机器翻译训练集中的真实中文语料之间的语义相似度. 第四步将计算得到的语义相似度较高的中文译文句子及对应的蒙古文句子构成一对伪蒙汉数据集,并扩充到 CCMT2021 提供的蒙汉机器翻译训练集中,训练基于 BERT 数据增强的 Transformer 蒙汉神经机器翻译模型.

2.2.1　BERT 中文语义相似度计算模型

BERT 的内部结构其实就是多个 Transformer 的编码器,Transformer 编码器因为有自注意力机制,因此 BERT 自带双向功能. 本研究中为了使获得的伪平行语料具有更少的噪声,选用了 BERT 训练中文语义相似度计算模型. BERT 中文语义相似度计算模型的框架如图 2 所示.

从图 2 可以看出,模型主要分为输入层、特征提

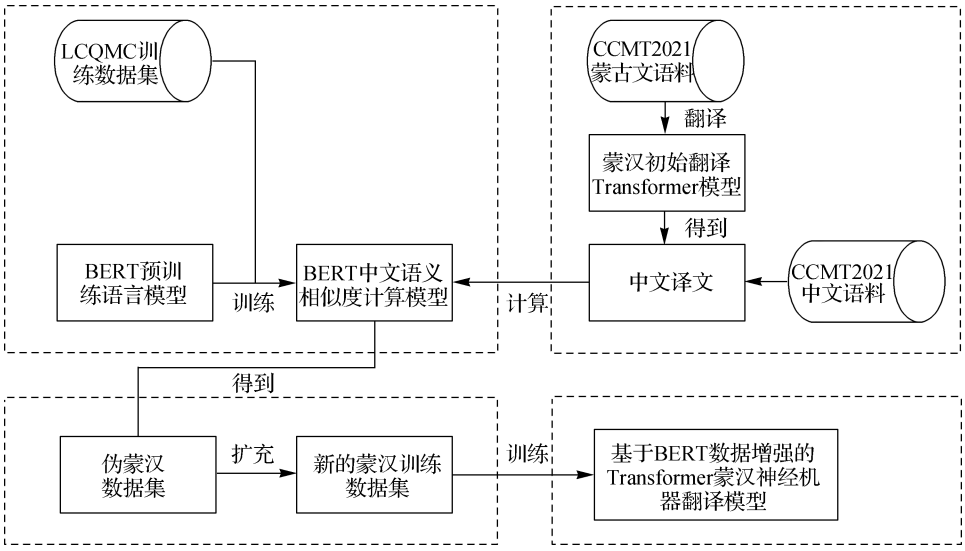


图 1 系统整体流程框架

Fig. 1 System overall process framework

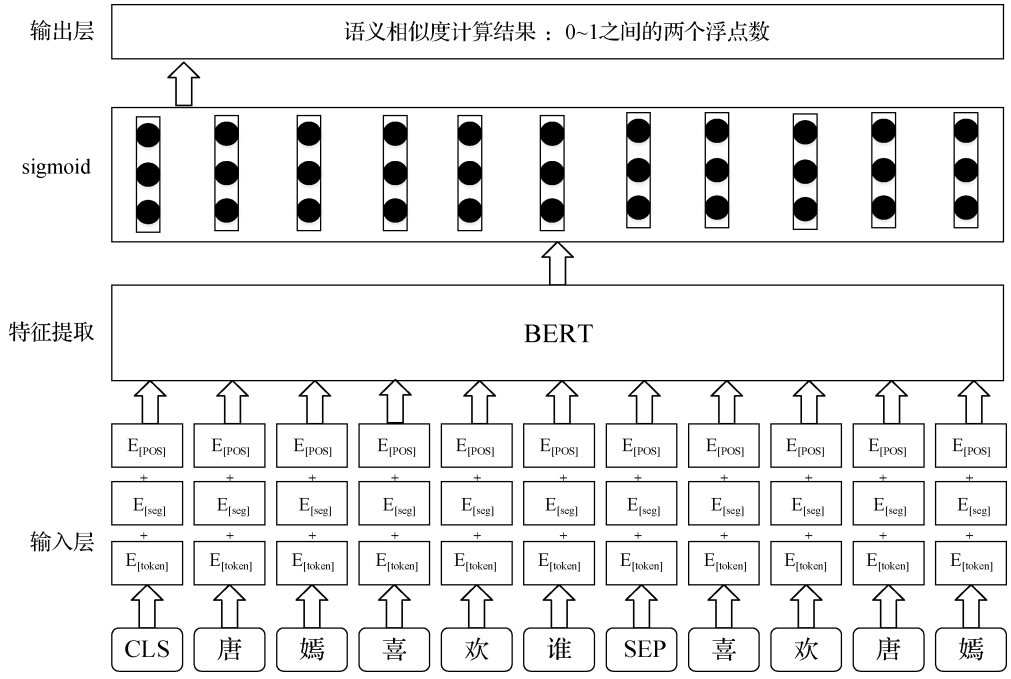


图 2 BERT 中文语义相似度计算模型结构图

Fig. 2 BERT Chinese semantic similarity calculation model framework

取层、sigmoid 函数层以及输出层. 模型的输入是一对中文句子. 句子中的[CLS]是一个特殊分隔符,表示每个样本的开头,[SEP]是样本中每个句子的结束标记符. 得到输入表示后,使用 BERT 模型的组成结构 Transformer 编码器去提取句子的特征,通过多层 Transformer 编码器逐层细化语义特征表示. 接着,取 BERT 的输出中[CLS]对应的输出向量作为整个句子的语义表示. 采用两个语义信息向量的余弦距离表征语义相似度,然后将其传递给 sigmoid 函数进行预测计算得到两个语义相似的概率. 输出结果是 0 到 1 之间的两个浮点数,接近于 1 表示两个句子的语义相似度高,接近 0 表示两个句子的语义相似度比较低. 余弦距离是从向量之间的角度的余弦得出的. 设句子 A 和 B 的语义信息向量分别为 $\boldsymbol{W}$ 和 $\boldsymbol{S}$:

$$\begin{cases}\boldsymbol{W}=[w_1,w_2,\cdots,w_n],\\ \boldsymbol{S}=[s_1,s_2,\cdots,s_n],\end{cases} \tag{1}$$

二者的夹角余弦值等于:

$$\cos(\boldsymbol{W},\boldsymbol{S}) = \frac{\boldsymbol{W}\boldsymbol{S}^{\mathrm{T}}}{|\boldsymbol{W}||\boldsymbol{S}|} = \frac{\sum_{i=1}^{n} w_i s_i}{\sqrt{\sum_{i=1}^{n} w_i w_i}\sqrt{\sum_{i=1}^{n} s_i s_i}}. \tag{2}$$

2.2.2　机器翻译蒙汉训练语料的扩充

利用 BERT 中文语义相似度计算模型进行相似度计算,最后会得到两个概率值相加为 1 的浮点数.最终实验得到的翻译结果中共有 36 636 条与原训练集不同的句子,因为当语义相似度为 0.5 以下时会使两个句子语义出现相反或不同的现象,所以本文中相似度取值范围选择 0.5,0.6,0.7,0.8 和 0.9.

当语义相似度大于 0.8 或 0.9 时,扩充语料库规模并没有明显的增加.当语义相似度大于 0.5 时,扩充语料库规模达到 23 092 句,说明可以有效扩充蒙汉机器翻译训练语料库的规模,从而缓解蒙汉对齐语料的数据稀疏问题.通过对语义相似度在大于 0.5,0.6,0.7,0.8,0.9 的取值下进行语料库扩充后的蒙汉机器翻译对比实验,发现语义相似度等于 0.5 时进行语料库扩充的实验效果最佳.所以根据不同语义相似度取值下扩充语料库对蒙汉机器翻译实验 BLEU 值的大小选择了相似度等于 0.5 概率值的句子.

3　蒙汉机器翻译实验及分析

3.1　实验数据

本文中 BERT 中文语义相似度计算模型的训练集为 LCQMC (large-scale Chinese question matching corpus) 数据集.LCQMC 是哈尔滨工业大学为 COLING2018 自然语言处理国际会议构建的语义匹配数据集.它的目的是确定两个句子是否具有相同的语义.数据集中训练集和开发集均为两个中文句子末尾带有 0 和 1 标签的数据.数据集中包括训练集(251 266 句对)、开发集(8 800 句对)和测试集(12 000 万句对).在本文实验中没有使用数据集中的测试集,而是选用了蒙汉机器翻译数据集中的中文训练集作为测试集,将中文译文和真实的中文句子使用空格作为间隔放到一行中并在句子末尾添加标签"1"作为测试集.

蒙汉机器翻译实验均采用了 CCMT2021 提供的训练集和开发集.测试集选用了 CCMT2019 提供的离线测试集,共 1 000 个蒙古文句子,每个句子对应 4 个汉语参考答案,且对测试集语料进行了校正编码工作.实验数据预处理方法见本文第 1 节.

3.2　实验参数设置

BERT 中文语义相似度计算模型的实验参数设置:损失函数选用交叉熵损失函数,Batch_size 为 32;Dropout 为 0.1,优化器使用 Adam 算法进行优化,激活函数使用 relu;迭代训练 2.2×10^4 轮.

蒙汉机器翻译实验参数设置:Transformer 采用 6 层编码器和 6 层解码器,每个 batch 的最大 token 设置为 2 048,学习率设置为 0.001,dropout 设置为0.3,优化器使用 Adam 算法进行优化.2.5×10^5 轮迭代训练,每 1 000 步保存一个翻译模型.在完成 2.5×10^5 轮迭代翻译训练后,将最新的 20 个翻译模型保存在检查站中.

运行环境:操作系统为 Ubuntu 16.04 版的 Linux 平台;CPU 均为 Intel(R) Core(TM) i7-8700k CPU @ 3.70 GHz;GPU 为 GTx1080Ti 独立显卡、i7 8700k 处理器和 256 GB 的硬盘;内存为 32 GB,运行平台是谷歌公司开发的 GPU 版本的 Tensorflow.

3.3　实验结果与分析

3.3.1　基于词切分的蒙汉神经机器翻译模型实验及分析

蒙汉机器翻译的基线实验是在 Transformer 的神经网络模型上不采用切分的实验.对比实验分别采用部分切分方法、BPE 切分方法及 BiLSTM-CNN-CRF 神经网络切分方法对蒙古文进行切分.基于不同粒度的词切分实例对比,如表 1 所示.

表 1　不同粒度词切分实例对比

Tab.1　Comparison of examples of word segmentation with different granularities

切分方法	ᠰᠠᠶᠢᠨ ᠠᠷᠢᠬᠢᠨ ᠳᠤ ᠰᠤᠷᠲᠠᠯᠴᠢᠯᠠᠭ᠎ᠠ ᠬᠡᠷᠡᠭ ᠦᠭᠡᠢ᠃
不切分	SAYIN ARIHIN-DV SVRTALCILAG _ A HEREG UGEI.
部分切分	SAYIN ARIHIN/-DV SVRTALCILAG _ A HEREG UGEI/.
BPE 切分	SAYIN ARIHI@@N-DV SVR@@TAL@@CILAG_A HEREG UGEI.
神经网络切分	SAYIN ARIHI/N/-DV SVRTALCILA/G_A HEREG UGEI/.

从表 1 中可以看出,拉丁转写不作任何切分处理,只以词切分为准时:分写构形附加成分与前面的词未切分;格附加成分"DV"与前面的词之间由"-"控

制字符连接着,看成一个词. 部分切分方法中格附加成分“DV”的控制字符“-”与前面的词之间用斜杠(/)切分,把格构形附加成分看成一个单独的词. BPE 切分方法把蒙古文单词“SVRTALCILAG_A”拆分成“SVR@@”“TAL”“CILAG_A”,其中“@@”标记是分隔符. BPE 切分方法主要是统计每一个连续字节对的出现频率,切分后的子词并不符合蒙古文的构词规则. 神经网络切分方法把“ARIHI/N/-DV”一词切分成了词干“ARIHI”和构形附加成分“N”和“-DV”.

机器翻译评测指标选用了 BLEU4[21]、BLEU4_SBP[22]和 NIST5[23]值,不同的切分方法在 CCMT2019 测试集上的实验结果如表 2 所示.

表 2 蒙古文词切分方法在 Transformer 机器翻译中的实验结果

Tab. 2 Experimental results of Mongolian word segmentation method in Transformer machine translation

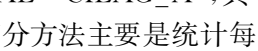
切分方法	BLEU4/%	BLEU4_SBP/%	NIST5/%
不切分	68.12	65.66	10.467 3
部分切分	69.87	67.21	11.272 9
BPE 切分	68.74	66.78	10.808 9
神经网络切分	69.12	66.83	10.982 8

从表 2 可以看出,对蒙古文语料做切分预处理工作与不做切分工作相比可以有效地提高机器翻译质量,这是因为蒙古文有着特殊的构词结构. 如果不进行词切分,基于词的蒙古文训练语料库中单频词的数量会高达总词数的一半,这是很严重的数据稀疏问题. 因此,从不同的粒度去切分蒙古文,可以有效缓解数据稀疏的问题.

通过对蒙古文语料进行部分切分、BPE 切分及基于神经网络的词干词缀切分实验对比发现,在 Transformer 神经机器翻译系统上基于部分切分的方法要高于其他不同粒度的切分方法,且 BLEU 值能够达到 69.87%. 通过分析得出,使用 BPE 进行切分时,因为 BPE 主要以词频统计为主要依据进行切分,所以会使蒙古文语料丢失掉原有的词性及词义. 采用 BiLSTM-CNN-CRF 神经网络切分后的蒙古文语料库进行分析,发现该词切分方法准确率很高,但在蒙汉机器翻译任务中会使蒙古文句子变长、蒙古文词的切分粒度过细且语料中出现大量单个的词素. 因此,后续实验将在 BLEU 值最高的部分切分方法的基础上进行.

3.3.2 蒙汉机器翻译对比实验及分析

在对蒙古文语料进行部分切分基础上,本文中对蒙古文语料的控制字符进行了过滤处理. 过滤掉控制字符后的蒙古语句子拉丁转写后的形式如下所示.

传统蒙古文句子: 

拉丁转写后的词素切分形式:BI TAN -TAI 0CIGAD -CV NEMERI ALAG_A.

过滤掉控制字符以后形式:BI TAN TAI 0CIGAD CV NEMERI ALAG A.

即将格附加成分“-TAI”的控制字符“-”和“-CV”的控制字符“-”以及名词复数附加成分“_A”的控制字符“_”替换为空格.

在对蒙古文进行部分切分且过滤掉控制字符的基础上,使用 BERT 模型进行中文语义相似度计算,从而扩充训练集平行语料. 对蒙古文进行过滤控制字符预处理方法和 BERT 数据增强方法在 Transformer 神经机器翻译系统上的实验结果,如表 3 所示.

表 3 蒙汉神经机器翻译结果

Tab. 3 Mongolian-Chinese neural machine translation results

实验方法	BLEU4/%	BLEU4_SBP/%	NIST5%
部分切分	69.87	67.21	11.272 9
+过滤蒙古文中控制字符	72.00	70.23	11.302 3
+BERT 数据增强	75.28	73.52	11.504 9

从表 3 实验结果可以看出,在蒙古文部分切分基础上对蒙古文控制字符的过滤有效地提高了机器翻译性能,与部分切分实验结果相比,BLEU4 值提高 2.13个百分点. 通过分析得出,不过滤控制字符的情况,在模型训练或标签还原时可能会与正常的词边界空格混淆. 通过过滤可以解决这一问题,使模型训练任务变得简单.

在过滤掉控制字符基础上使用 BERT 模型进行数据增强,会进一步提升机器翻译的质量,使蒙汉机器翻译 BLEU4 值达到 75.28%. 通过对实验结果分析得出,此方法可以有效扩充蒙汉双语训练语料库,进一步提高了蒙汉神经机器翻译模型的性能. 基于 BERT 的数据增强方法对低资源语言机器翻译任务质量提升至关重要.

4 总 结

本研究采用 Transformer 神经机器翻译模型作为蒙汉翻译的基线系统. 针对低资源语种上神经机器翻

译面临的训练语料匮乏问题,首先采用部分切分方法、BPE 切分方法及 BiLSTM-CNN-CRF 神经网络切分方法对蒙古文进行切分工作.通过实验结果发现部分切分的方法在基于 Transformer 神经机器翻译系统上性能更优.在此基础上,结合部分切分与过滤掉蒙古文语料控制字符方法,提出一种基于 BERT 数据增强的方法.该方法主要利用 BERT 模型计算得到语义相似度较高的中文句子去构建蒙汉伪平行语料以扩充训练语料.本文在蒙汉语种上的机器翻译实验结果表明,BERT 数据增强技术应对蒙汉机器翻译任务的平行双语数据稀疏问题,能有效提高神经机器翻译对低资源语言的泛化能力.

在蒙汉双语机器翻译任务中,目前最严峻的是双语数据稀疏问题,针对该问题后续可以使用迁移学习、译文后处理等方法进行研究.

参考文献:

[1] SU J S, ZHANG X W, LIN Q, et al. Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding[J]. Artificial Intelligence, 2019, 277: 103168.

[2] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation [C] // Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1412-1421.

[3] LAMPLE G, DENOYER L, RANZATO M. Unsupervised machine translation using monolingual corpora only[EB/OL]. [2021-11-01]. https: // www. arxiv-vanity. com/papers/1711. 00043/.

[4] NIU X, DENKOWSKI M, CARPUAT M. Bi-directional neural machine translation with synthetic parallel data[C] // Workshop on Neural Machine Translation and Generation. Stroudsburg: ACL, 2018: 84-91.

[5] WU L J, WANG Y R, XIA Y C, et al. Exploiting monolingual data at scale for neural machine translation [C] // Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2019: 4207-4216.

[6] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[EB/OL]. [2021-11-01]. https: // arxiv. org/abs/1511. 06709.

[7] FADAEE M, MONZ C. Back-translation sampling by targeting difficult words in neural machine translation [EB/OL]. [2021-11-01]. https: // arxiv. org/abs/1808. 09006.

[8] SUGIYAMA A, YOSHINAGA N. Data augmentation using back-translation for context-aware neural machine translation[C] // Workshop on Discourse in Machine Translation. Stroudsburg: ACL, 2019: 35-44.

[9] PONCELAS A, SHTERIONOV D, WAY A, et al. Investigating backtranslation in neural machine translation [C] // Annual Conference of the European Association for Machine Translation. Stroudsburg: ACL, 2018: 249-258.

[10] EDUNOV S, OTT M, AULI M, et al. Understanding back-translation at scale[C] // Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 489-500.

[11] 谷舒豪,单勇,谢婉莹,等.基于数据增强及领域适应的神经机器翻译技术[J].江西师范大学学报(自然科学版),2019,43(6):643-648.

[12] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C] // Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 1715-1725.

[13] YAO Y, HUANG Z. Bi-directional LSTM recurrent neural network for chinese word segmentation[C] // International Conference on Neural Information Processing. Cham: Springer, 2016: 345-353.

[14] WANG C Q, XU B. Convolutional neural network with word embeddings for Chinese word segmentation[C] // International Joint Conference on Natural Language Processing. Taiwan: Asian Federation of Natural Language Processing, 2017: 163-172

[15] JOHN L, ANDREW M, FERNANDO C N P. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C] // International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001: 282-289.

[16] 国家质量监督检验检疫总局,中国国家标准化管理委员会.信息技术 蒙古文变形显现字符集和控制字符使用规则: GB/T 26226—2010[S].北京:中国标准出版社,2011.

[17] 包乌格德勒,赵小兵.基于 RNN 和 CNN 的蒙汉神经机器翻译研究[J].中文信息学报,2018,32(8):60-67.

[18] 任众,侯宏旭,武静,等.基于统计和神经网络的蒙汉机器翻译研究[J].中文信息学报,2018,32(11):1-7.

[19] 申志鹏.基于注意力神经网络的蒙汉机器翻译系统的研究[D].呼和浩特:内蒙古大学,2017.

[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2018: 4171-4186.

[21] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2002: 311-318.

[22] CHIANG D, DENEEFE S, CHAN Y, et al. Decomposability of translation metrics for improved evaluation and efficient algorithms[C] // Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2008: 610-619.

[23] DODDINGTON G. Automatic evaluation of machine translation quality using *n*-gram co-occurence statistics [C] // International Conference on Human Language Technology Research. San Francisco: Morgan Kaufmann Publishers Inc, 2002: 138-145.

Mongolian-Chinese neural machine translation system based on word segmentation with BERT data enhancement

HE Wuyun, XIU Zhi, BAO Jingjing, CHEN Meilan, WANG Siriguleng*

(Collage of Computer Science and Technology, Inner Mongolia Normal University, Hohhot 011500, China)

Abstract: Although neural machine translation (NMT) is currently regarded as the mainstream research method in the field of machine translation, the scarcity of Mongolian-Chinese (M-C) parallel corpus retards improvements of M-C NMT performances. In this study, aiming at the M-C NMT systems based on Transformer, we use the deep-learning model to study the Mongolian word segmentation method. Mongolian partial segmentation, BPE subword segmentation and BiLSTM-CNN-CRF neural network segmentation method for the influence of the Mongolian-Chinese machine translation model are compared among one another. On this basis, the data enhancement technology based on bidirectional encoder representations from Transformers (BERT), Chinese semantic similarity calculation is used to expand the M-C machine translation training data. In comparison of experimental results on the dataset provided by CCMT2019, ours show that the BLEU value of the data enhancement method is significantly improved when compared with that of baseline experiments, and the BLEU4 value reaches 75.28%.

Keywords: Mongolian-Chinese neural machine translation; Transformer neural network; BERT; semantic similarity

(责任编辑:任滢滢)