doi:10.6043/j.issn.0438-0479.201908030

多策略切分粒度的藏汉双向神经机器翻译研究

沙　九[1],冯　冲[1*],张天夫[1],郭宇航[1],刘　芳[2]

(1.北京理工大学计算机学院,北京市海量语言信息处理与云计算应用工程技术研究中心,北京 100081;
2.北京理工大学外国语学院,工信部语言工程与认知计算重点实验室,北京 100081)

摘要:现有的机器翻译模型通常在词粒度切分的数据集上进行训练,然而不同的切分粒度蕴含着不同的语法、语义的特征和信息,仅考虑词粒度将制约神经机器翻译系统的高效训练.这对于藏语相关翻译因其语言特点而显得尤为突出.为此提出针对藏汉双向机器翻译的具有音节、词语以及音词融合的多粒度训练方法,并基于现有的注意力机制神经机器翻译框架,在解码器中融入自注意力机制以捕获更多的目标端信息,提出了一种新的神经机器翻译模型.在 CWMT2018 藏汉双语数据集上的实验结果表明,多粒度训练方法的翻译效果明显优于其余切分粒度的基线系统,同时解码器中引入自注意力机制的神经机器翻译模型能够显著提升翻译效果.此外在 WMT2017 德英双语数据集上的实验结果进一步证明了该方法在其他语种方向上的适用性.

关键词:音词融合;藏汉双向;神经机器翻译

中图分类号:TP 391　　**文献标志码**:A　　**文章编号**:0438-0479(2020)02-0213-07

藏语是我国古老的语言文字之一,属辅音文字型,具有丰富的表达功能和严密的语法结构,是具有典型逻辑格语法体系的复杂拼音文字.它最小的语言单位是字母,其次是音节,其音节至少由一个藏文字母构成,且每个音节的前后都有一个音节点“·”.音节是藏语最基本的形式构词单元和语言单位.藏语音节分辅音字母、元音符号和标点符号 3 部分,含 30 个辅音字母和 4 个元音符号.

与其他语言相比,藏语音节之间的分隔符存在省略现象.例如:虚词中的格助词“ར”、具格助词“ས”、属格助词“འི”、终结词“ཨ”、饰集词“འང”和离合词“འམ”与其前一音节不加分字分隔符组成一个音节,从而形成紧缩词或者黏着词,如“མགའི”是由音节“མགའ”和“ཀི”结合而成;其次词与词之间没有明显的间隔标记,如“མི་དང་ཆེན་པོ།”和“མི་དང་ ཆེན་པོ།”,不同的切分方式导致其表达的语义有所不同.藏语中的虚词和实词的界限也很难区分.虚词具有语法单位和词汇单位的两面性,并且具有层级性,不同虚词语法化和抽象化的程度不同.在一个句子中词和字携带不同的语义信息,计算机正确识别紧缩词或者黏着词对文本的歧义消解和句法、句型、语义处理有着十分重要的意义.

藏语的上述特点为机器翻译等自然语言处理技术提出了挑战.近年来随着编码器-解码器下模型结构的演化[1],基于端到端的网络模型在神经机器翻译(neural machine translation,NMT)任务中已取得了良好的进展[2].NMT 把源语言编码为词嵌入,再对这些词嵌入通过解码器生成目标语言.其中,在神经网络训练中,通过长短时记忆(long short-term memory,LSTM)等网络和注意力机制的正确连接,可以很好地解决长距离依赖问题[3].而传统的编码器仅对字或词进行编码,这样在训练过程中丢失了源语言所携带的语义信息,降低了 NMT 的忠实度.对于资源稀缺的语料难以构建高效的翻译模型.

目前对藏语的相关翻译还存在很多技术难题.首先,NMT 系统需要具备高质量大规模的平行语料才能训练出高效的翻译模型[4],而获取高质量大规模的平行语料比较困难,尤其是对资源稀缺的语料.其次,NMT 使用神经网络直接将源语言文本映射成目标语

收稿日期:2019-08-28　　**录用日期**:2019-12-16

基金项目:国家重点研发计划(2016YFB0801200,2018YFC0832104);国家自然科学基金(U1636203)

* **通信作者**:fengchong@bit.edu.cn

引文格式:沙九,冯冲,张天夫,等.多策略切分粒度的藏汉双向神经机器翻译研究[J].厦门大学学报(自然科学版),2020,59(2):213-219.

Citation:SHA J,FENG C,ZHANG T F,et al. Multi-strategic granularity of segmentation on Tibetan-Chinese bidirectional neural machine translation[J]. J Xiamen Univ Nat Sci,2020,59(2):213-219. (in Chinese)

言文本时,一般以句子为单位进行层次化处理,在句子级别上用不同的粒度,如以单词、子单词或字符为依据进行处理;但不同粒度在句法和语义层面所携带的特征对训练 NMT 模型在学习和获取特征时的影响有所不同,很难确定哪种粒度更适合特定的翻译任务. 单词通常被认为是语言交际的基本单位,早期的 NMT 系统要求句子用以词语为单位粒度的序列表示. 后来,Sennrich 等[5]建议将单词分割成更小的单元并像单词级别翻译. Britz 等[6]利用混合词元模型建立开放词汇的 NMT,主要在词层次上进行翻译,并为罕见的词查询字符提供组件. Murtay 等[7]提出了一种利用字符信息的循环神经网络(recurrent neural network, RNN)语言模型,从字符的 n 元语法(n-gram)嵌入构造单词嵌入,并与普通单词嵌入融合进行模型训练. Yan 等[8]专门研究电子商务领域中通过应用词级权重来提高 NMT 的效果.

为此,本文针对资源稀缺的藏汉双向翻译任务,通过不同粒度来改进 NMT 模型:首先提出了多策略切分不同粒度的翻译模型;其次,在 RNN 的解码器中引入自注意力机制[9],以修改网络结构获取不同粒度上所携带的语义特征信息,从而更好地保留原文中所携带的特征信息;最后通过多策略融合方式从不同粒度上获取更高效的译文.

1 NMT 模型

1.1 RNNSEARCH

大多数 NMT 系统都遵循 Bahdanau 等[4]提出的注意力机制的编码器-解码器框架. 目标是将一种序列转换成另一种序列. 基本思想为使用两个网络来处理翻译任务,分别为编码器和解码器,编码器将输入的序列转换成一个固定长度的内部表示向量,解码器则将该向量作为输入用以预测输出的序列. 两个网络之间由内部表示向量连接. 假设输入句子 $X=\{x_1,x_2,\cdots,x_{T_x}\}$,目标端输出句子为 $Y=\{y_1,y_2,\cdots,y_{T_y}\}$. 翻译过程实际上就是一个概率求解过程,具体计算形式为式(1):

$$P(Y\mid X;\theta)=\prod_{j=1}^{T_y}P(y_{<j},X;\theta), \tag{1}$$

其中,θ 为参数,$y_{<j}$ 为目标端生成词之前的所有词. 因为 X 与 Y 并不等长,所以一般构造 RNNSEARCH 模型,即由两个 RNN 组成的编码器-解码器结构. 编码器-解码器网络对源句的语义进行建模,并将源句转换为上下文向量表示形式,解码器从上下文向量表示形式中逐词生成目标词. NMT 的一个重要特征是将词汇表中的每个单词映射成一个低维实值向量,连续表示法的使用让 NMT 能够学习潜在的双语映射,以进行准确翻译,并探索单词间的统计相似性.

1.1.1 编码器

通常的 RNN 计算式(1)时会从 x_1 依次读取到 x_{T_x} 完成序列 X 的输入. 但是,本研究希望每个单词的注释不仅总结前面的单词,而且总结后面的单词. 因此编码器使用 m 个堆叠的 LSTM 层生成,隐层向量 $\boldsymbol{h}_j^k(k=1,2,\cdots,m)$,具体计算如下所示:

$$\boldsymbol{h}_j^k=\mathrm{LSTM}(\boldsymbol{h}_{j-1}^k,\boldsymbol{h}_j^{k-1}), \tag{2}$$

其中如果 $k=1$,则 $\boldsymbol{h}_j^{k-1}=\boldsymbol{x}_j$,$\boldsymbol{x}_j$ 为词 x_j 的向量表示.

1.1.2 解码器

解码器根据 y_j 上的上下文向量 $\boldsymbol{c}_j$ 计算输出概率 $p(y_j\mid y_{<j},X;\theta)$,$\theta$ 在不同的时间步长使用不同的上下文向量 $\boldsymbol{c}_j$,最终 y_j 的输出概率为

$$P(y_j\mid y_{<j},X;\theta)=\mathrm{softmax}\left(\begin{bmatrix}\boldsymbol{t}_{j-1}\\ \boldsymbol{d}_j\\ \boldsymbol{c}_j\end{bmatrix}\right), \tag{3}$$

其中,$\boldsymbol{t}_{j-1}$ 是 $j-1$ 时刻目标词的嵌入,$\boldsymbol{d}_j$ 为 j 时刻编码器端的隐藏状态. 其中 $\boldsymbol{d}_j$ 的计算公式为

$$\boldsymbol{d}_j=\mathrm{LSTM}(\boldsymbol{d}_{j-1},\begin{bmatrix}\boldsymbol{t}_{j-1}\\ \boldsymbol{c}_j\end{bmatrix};\theta_{j-1}). \tag{4}$$

注意力机制将上下文向量 $\boldsymbol{c}_j$ 计算为源注释的加权和:

$$\boldsymbol{c}_j=\sum_{i=0}^{l}\alpha_{ji}\boldsymbol{h}_j. \tag{5}$$

其中,$\boldsymbol{h}_j=[\boldsymbol{h}_j^1,\boldsymbol{h}_j^2,\cdots,\boldsymbol{h}_j^m]$,注意力机制权重的计算式如下:

$$\alpha_{ji}=\frac{\exp(e_{ji})}{\sum_{i=1}^{l}\exp(e_{ji})}, \tag{6}$$

$$\boldsymbol{e}_j=\boldsymbol{v}_a^{\mathrm{T}}\tanh(\boldsymbol{W}_a\boldsymbol{d}_{j-1}+\boldsymbol{U}_a\boldsymbol{h}_j^k), \tag{7}$$

其中,$\boldsymbol{v}_a$、$\boldsymbol{W}_a$ 和 $\boldsymbol{U}_a$ 是注意力机制的权重矩阵,$\boldsymbol{e}_j$ 在注意力机制模型中能够平衡 $\boldsymbol{d}_{j-1}$ 和 $\boldsymbol{h}_j^k$. 使用这种策略,解码器可以处理在给定时间内最相关的源语句.

1.2 Transformer

Transformer 模型[10]仅依赖于注意力机制,也采用了编码器-解码器架构,但其结构相比于注意力更加复杂,其中编码端由 6 个编码器堆叠在一起,解码端也一样. 每个编码器包含两层:一个自注意力层和一个前馈神经网络,自注意力能帮助当前节点不仅仅只关注当前的词,从而能获得上下文的语义. 每个解码器也包含编码器提到的两层网络,且在这两层中间还有一层注意力层,帮助当前节点获得当前需要关注的重点内容.

1.3　本文 RNN 模型

在通用序列建模中,自注意力是一种强有力的机制. 本文提出的模型(RNN * Self-Attention)在 RNNSEARCH 的解码端引入 Transformer 中解码器的自注意力. 解码器完全使用注意力机制不仅能够在不同的神经层之间传递信息,实现一个多层注意力机制的神经网络翻译模型,而且能更好地处理序列过长问题;并且自注意力能够计算每一个词之间的注意力. 捕获更多的原文信息. 如图 1 所示:

混合多策略架构是一个层次结构序列到序列的模型,通过在不同的粒度级别即音节、词语、音词融合上进行训练作为对比实验.

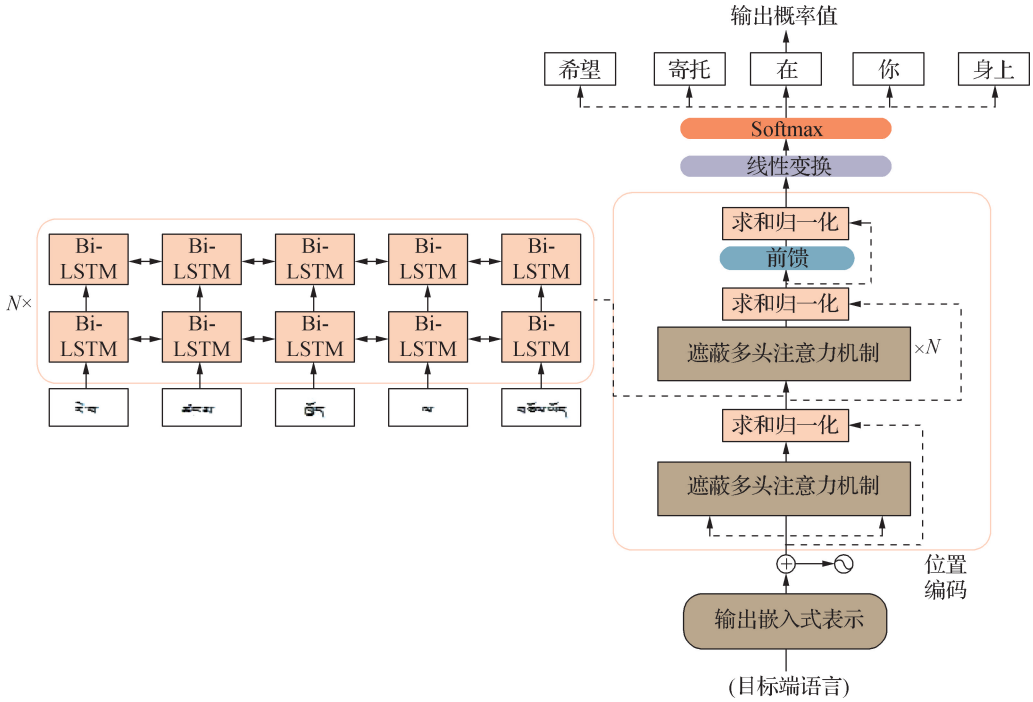

图 1　RNN * Self-Attention 的编码器-解码器网络模型

Fig. 1　Encoder-decoder network model of RNN * Self-Attention

2　不同粒度的切分策略

通过不同的切分粒度训练一个高效的翻译模型,使用字符集信息增强(子)字级别表示;自注意力机制的解码器,则使用不用粒度的表示协同控制翻译. 本文对音节、词语、子词和音词融合 4 种切分策略进行了实验,其中子词切分策略采用字节对编码(BPE),其余 3 种切分策略介绍如下.

2.1　以音节为粒度的切分策略

在 NMT 中广泛采用的字符级建模[11]具有多种优势:1) 字符是语言的基本单位,2) 字符级别能够更好地为数据稀疏提供支撑作用,3) 字符级建模的消失梯度. 考虑到藏语的特殊性和不断出现新词术语的情况,词汇库不可能涵盖所有的语言词汇. 词汇量不足将直接影响 NMT 的性能. 使用更大的词汇量虽然能提高译文质量,但是训练过程将变得更加复杂,而且词汇库中往往具有许多相似的词汇[3],它们共享一个词汇位,但具有不同的词形,几乎不可能利用所有可能的候选词汇来获得良好的性能. 为此,本文以藏语中的分割符“·”为切分依据,对藏语以音节为粒度进行切分,这样不仅缩小了词汇表,同时将完整地切分出每个独立的音节.

2.2　以词语为粒度的切分策略

目前大量的 NMT 系统都是以词语级别为单位进行训练,音节级别的 NMT 系统可能会遇到丢失语义或数据信息分散等问题. 在藏语中词与词之间没有明确的分割符,导致机器学习对齐和翻译更加困难. 为了缓解神经元使其在 NMT 中能够很好地获取词语级别上的信息特征,本文利用西北民族大学所开源的藏文分词工具 TIP-LAS[12],利用人工标记的大量数据进行训练,其次利用该分词工具对本文中使用的所有数据进行分词,最后通过人工校正分词所带来的歧义. 这样很好地避免了由于没有公开高效的藏语分词

工具而为藏语翻译系统带来的困难.

2.3 音词融合的切分策略

在资源匮乏的环境中,大型词汇表导致低频(子)词在训练时被表示为子字单元,并且模型还需学习这些高维表示的能力. Sennrich 等[13]提出了子字单元的最小频率阈值,并将频率较低的子字段分成较小的单元或字符,以减少仔细调整词汇量大小到数据集的需要,从而对较小数据集进行更准确地分割. 这种方法在分布表示上可以捕获多个混合的候选词,并在增强数据中有足够多的变化,是一种有效解决低资源的 NMT 技术. 合并来自细粒度的数据可以看作是一种很好的数据扩充方式. 尽管如此,简单的融合训练仍然存在歧义,例如"ཞབས་ཞེག་དང་སྒྲིལ་གསལ་སོགས་ཉེད་པའི་ལམ་ལུགས་འཇུག་ཚང་དུ་བཏང་།"(要完善保证和利补等制度),按照音节切分为"ཞབས་ ཞེག་ དང་ སྒྲིལ་ གསལ་ སོགས་ ཉེད་ པའི་ ལམ་ ལུགས་ འཇུག་ ཚང་ དུ་ བཏང་ །",按照词语切分为"ཞབས་ཞེག་ དང་ སྒྲིལ་གསལ་ སོགས་ ཉེད་པ་ ཡི་ ལམ་ལུགས་ འཇུག་ཚང་ དུ་ བཏང་ །",按照音词融合切分为"ཞབས་ཞེག་ དང་ སྒྲིལ་གསལ་ སོགས་ ཉེད་པའི་ ལམ་ལུགས་ འཇུག་ ཚང་ དུ་ བཏང་ །". 有许多词,甚至可能有许多句法结构不会在高度相关的语法之间共享. 其次藏语中具有大量的黏着语和紧缩词,例如"དེ་རིང་སྨན་པང་དུ་ཁོས་གཞེགས་བཅོས་བྱས།"和"དེ་རིང་སྨན་པང་དུ་ཁོར་གཞེགས་བཅོས་བྱས།"中除了虚词"ས"和"ར"外其他词序列与成分都完全一样,但是句意完全不同:第一句中"ཁོ"指"医生",第二句中"ཁོ"指"患者". 因此不能单纯地把藏语跟汉语同等地采用 BPE[5]. 为此,本文在分词的基础上对藏语按照藏文紧缩格识别方法[14],融合规则、统计和还原等方式进行切分. 具体步骤如下:在分词的基础上首先提取携带拟紧缩音节的词,对其中所携带的音节利用规则进行识别. 若规则不能识别,则用还原法识别;如果还原法不能识别,则利用最大熵模型识别,最终达到音词融合的效果. 该融合方法的一个特性是通过改变词汇表的大小可以控制混合基于音节和词语的融合模型.

3 实验与结果分析

3.1 实验设置

本文将 RNN * Self-Attention 与厦门大学开发的 XMUNMT 开源翻译框架中的 RNNSEARCH 系统(下文简称为 RNNSEARCH)、Tensor2Tensor 开源框架中的 Transformer 系统(下文简称为 Transformer)进行了对比实验. 考虑客观性,对比的其他两种系统的所有参数都保持了文献[10]中最优的参数设置.

为了能够跟以往的研究实验进行对比,所有的实验数据使用公开的数据集. 藏汉平行语料为 CWMT 2018 提供的数据集,经过预处理后最终使用的数据如表 1 所示. 在 CWMT 测试集上进行测试,测试数据如表 2 所示,其中 2017_dev、2017_test、2018_test 分别表示 CWMT2017 的开发集、测试集和 CWMT2018 的测试集,另外为了进一步说明在 RNNSEARCH 的解码端通过引入自注意力机制来进一步提升翻译效果,还在德英和英德翻译上进行实验,使用的数据集为 WMT2017 德英.

表 1 不同粒度下的训练语料

Tab. 1 Training corpus at different granularities

不同粒度	训练集		开发集	
	句对数目	平均句长	句对数目	平均句长
音节	147 434	81.58	1 000	46.21
词语	147 434	81.65	1 000	44.97
音词融合	147 434	99.69	1 000	57.35

表 2 测试语料

Tab. 2 Test corpus

测试集	句对数目	平均句长
2017_dev	650	143.23
2017_test	729	89.66
2018_test	1 000	57.35

3.2 结果分析

本文在每个模型下按不同粒度进行训练,对表 2 所示测试语料进行测试,表 3 展示了不同模型在不同粒度下的测试结果,使用的评分工具为 Moses 中的机器双语互译评估(bilingual evaluation understudy, BLEU)值.

3.2.1 藏汉翻译

在藏汉翻译任务中,本文对比了第 1 节中 3 种不同的 NMT 模型,采用 4 种不同切分策略分别训练,在表 2 的 3 种测试集上进行测试,结果表明:

1) 以音节为切分粒度训练时,3 种模型的翻译效果最差,BLEU 值均低于 15%. 这是由于在藏语中不加其他任何规则或限制条件按照分割符"·"进行切分时,失去了其语义信息,无法保留原文所表达的信息,同时引入不少的噪声. 同样地,汉语按照字符进行切分也会失去原文所携带的语义信息,无法对应藏语中相应的译文,导致双语平行语料无效,所以 NMT 无

表 3　藏汉双向翻译模型测试结果

Tab. 3　Test results of Tibetan-Chinese bidirectional translation models

任务	模型	粒度	BLEU/%		
			2017_dev	2017_test	2018_test
藏汉	RNNSEARCH	音节	14.08	13.95	13.15
		词语	45.50	35.56	29.58
		子词	48.69	45.85	36.78
		音词融合	50.17	47.85	37.22
	Transformer	音节	12.68	11.27	10.03
		词语	45.60	36.74	30.43
		子词	47.98	43.31	33.11
		音词融合	**51.74**	48.30	**38.07**
	RNN * Self-Attention	音节	14.70	14.21	13.68
		词语	46.09	35.56	30.05
		子词	48.12	47.35	34.28
		音词融合	50.20	**48.57**	37.27
汉藏	RNNSEARCH	音节	8.52	10.96	9.77
		词语	45.13	62.71	33.31
		子词	46.12	62.93	34.78
		音词融合	46.75	63.13	35.38
	Transformer	音节	6.17	8.68	9.01
		词语	43.17	60.45	29.07
		子词	45.56	61.96	32.50
		音词融合	47.19	63.40	35.59
	RNN * Self-Attention	音节	8.95	11.19	9.80
		词语	45.60	62.81	33.39
		子词	46.43	63.01	34.33
		音词融合	**47.29**	**63.88**	**35.75**

法在平行语料中学习语义特征,导致训练不佳.

2) 以词语为切分粒度进行训练的效果明显优于以音节为切分粒度的.因为在词语级别上能够相对稳定地保留原文的语义信息,同时能够较好地在双语句对中找到对应的平行词语,有助于 NMT 获取更多的信息以及保留更多的语义特征.本文提出的 RNN * Self-Attention 模型在 3 个测试集上采用词语级别训练时 BLEU 值均有明显提升,与 RNNSEARCH 模型在 2017_dev 和 2018_test 测试集上的 BLEU 值相比,分别提升了 0.59 和 0.47 个百分点.

3) 当把 2 种语言分词后按照普通的子词切分方式处理训练时,同样也能提升翻译质量,3 种模型在不同的测试集上都有所提升.

4) 按音词融合切分策略进行训练时,相比于词语级别和子词级别 3 个模型的 BLEU 值均有明显提升.如与子词级别相比,本文提出的模型采用音词融合切分策略在 3 种测试集上的 BLEU 值分别提升了 2.08,1.22 和2.99个百分点.可见本文提出的音词融合方法能够有效提升译文质量.

在 3 个测试集中,同一种粒度下训练模型在2017_dev 测试集上的 BLEU 值最高.经分析后发现,2017_dev 测试集和 CCMT2018 的训练集均为新闻、政府领域,可见领域适应问题对于机器翻译仍然十分重要.

3.2.2　汉藏翻译

在汉藏翻译任务中,本文同样利用 3 种不同的模型系统,采用 4 种不同切分策略进行训练,并在 3 种测

试集上进行测试,实验结果表明:

1) 以音节为切分粒度训练模型依然很难达到预期效果,在 3 个测试集上音节级别测试结果的 BLEU 值均低于 15%.

2) 与藏汉翻译相同,在词语级别上的 BLEU 值比音节级别上的高很多;与藏汉翻译相比,汉藏翻译采用不同切分策略训练的模型效果更好;本文提出的 RNN * Self-Attention 模型相比于 RNNSEARCH 模型在词语级别上具有显著提升,例如在 2017_dev、2017_test、2018_test 测试集上 BLEU 值分别提高了 0.47,0.10 和0.08个百分点.

3) 不加其他规则按照通用的 BPE 方式双语处理后进行实验,同样能够提升译文质量,可见藏语也能满足子词级别的切分,并且以子词切分的翻译效果比词语级别的更好.

4) 按音词融合切分策略进行训练时,在汉藏翻译任务中更具优势,因为本文在音词融合时,中文使用 BPE 方式进行切分,藏语在分词的基础上利用藏语本身具有的语言特色规则、统计及还原的方式进行切分,后期经人工处理之后达到更好的音词融合效果. 首先在译文中几乎没有未登录词,其次更好地将双语中稀缺或频率较少的词语切换成更小粒度的词,使藏语更好地遵循语法信息,并照顾到黏着语格助词和紧缩词的问题,这将极大地提升译文质量. RNNSEARCH、Transformer 和 RNN * Self-Attention 模型在音词融合上的训练效果相比于 BPE 有显著提升,其中 RNN * Self-Attention 模型在 2017_dev、2017_test 和 2018_test 测试集上的 BLEU 值分别提升了 0.86,0.87 和 1.42 个百分点. 同样地,RNN * Self-Attention 模型相比于 RNNSEARCH 和 Transformer 模型均有所提升,证明本文提出的方法在汉藏翻译上的效果得到有效提升.

3.2.3 其他翻译

为了进一步证明 RNN * Self-Attention 模型能够更好地获取目标端信息,本文使用 WMT2017 德英数据集,分别在德英和英德翻译上进行实验,其实验过程中的所有参数跟藏汉和汉藏翻译中的参数保持一致,实验结果如表 4 所示. 虽然在德英和英德翻译中效果最好的为 Transformer 模型,但是,RNN * Self-Attention 模型相比于 Transformer 模型具有复杂度低、训练效率高等优势,能够更快地训练出一个比较稳定的翻译模型,如表 4 最后一列显示,在同样的数据集上训练翻译模型所耗费的平均时间具有很大差距,RNN * Self-Attention 模型效率高于 Transformer 模型. 与 RNNSEARCH 模型相比,RNN * Self-Attention 模型在德英和英德翻译任务上的 BLEU 值分别提升了0.18和 1.20 个百分点. 引入自注意力构建的网络模型不仅提高翻译效果,同时没有增大复杂度,能够有效地解决长序列中词与词之间的依赖关系,更好地学习到一个句子内部的结构. 从而证明本文方法具有较强的通用性.

表 4 德英和英德翻译的测试结果

Tab. 4 Test results in German-English and English-German

任务	模型	BLEU/%	t/s
德英	RNNSEARCH	35.93	0.19
	Transformer	36.79	0.21
	RNN * Self-Attention	36.11	0.19
英德	RNNSEARCH	33.76	0.20
	Transformer	35.93	0.24
	RNN * Self-Attention	34.96	0.21

注:t 表示模型迭代一次所耗费的平均时间.

4 结 论

本文针对资源稀缺的藏汉双向翻译研究了多切分粒度的训练方法,在音节、词语、音词融合 3 种切分策略下,有效解决了汉语句子中普遍出现的一些介词和连词在藏语中没有对应词的翻译问题,避免词汇和句法级别的翻译困难. 同时,把低频词切分成相对高频的子字片段,缓解数据稀疏问题,使多种模型的翻译效果得到显著提升,通过缩小词典大小,显著缩短训练周期,正确传达源语言的意义,符合目标语言的语法. 从同语料不同模型和同模型不同语料两个角度进行实验验证,在 CWMT2018 藏汉和 WMT 的德英机器双向翻译任务上的实验结果表明,本文方法优于基线系统.

在下一步的工作中,希望结合不同粒度,改进编码器在音节和词语级别上同时编码,以有效获取更多的特征属性,避免中间繁琐的过程. 同时希望能够引入先验知识库来缓解资源稀缺问题.

参考文献:

[1] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL].

[2019-12-16]. https://arxiv.org/pdf/1406.1078.pdf.
[2] COSTA-JUSSÀ M R, ALDÓN D, FONOLLOSA J A R. Chinese-Spanish neural machine translation enhanced with character and word bitmap fonts [J]. Machine Translation, 2017, 31(1/2): 35-47.
[3] SUTSKEVER I, VINYALS O, LE V Q. Sequence to sequence learning with neural networks[EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1409.3215.pdf.
[4] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointlyl earning to align and translate[EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1409.0473.pdf.
[5] SENNRICH R, HADDOW R, BIRCH A. Neural machine translation of rare words with subword units[EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1508.07909.pdf.
[6] BRITZ D, LE Q, PRYZANT R. Effective domain mixing for neural machine translation[C]// Proceedings of the Second Conference on Machine Translation. Copenhagen: ACL, 2017: 118-126.
[7] MURRAY K, CHIANG D. Auto-sizing neural networks: with applications to *n*-gram language models [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015: 908-916.
[8] YAN S, DAHLMANN L, PETRUSHKOV P, et al. Word-based domain adaptation for neural machine translation[EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1906.03129.pdf
[9] LIN Z H, FENG M W, DOS SANTOS C N, et al. A structured self-attentive sentence embedding [EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1703.03130.pdf.
[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1706.03762.pdf.
[11] ZHAO S J, ZHANG Z H. An Efficient character-level neural machine translation [EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1608.04738.pdf.
[12] 李亚超,江静,加羊吉,等. TIP-LAS:一个开源的藏文分词词性标注系统[J]. 中文信息学报, 2015, 29(6): 203-207.
[13] SENNRICH R, BIRCH A, CURREY A, et al. The University of Edinburgh's neural MT systems for WMT17[EB/OL]. [2019-12-16]. https://arxiv.org/pdf/1708.00726.pdf.
[14] 拉玛扎西,才智杰,扎西吉. 藏文紧缩格识别方法[J]. 计算机应用研究, 2019, 36(4): 1080-1083.

Multi-strategic granularity of segmentation on Tibetan-Chinese bidirectional neural machine translation

SHA Jiu[1], FENG Chong[1*], ZHANG Tianfu[1], GUO Yuhang[1], LIU Fang[2]

(1. Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China; 2. Key Laboratory of Language Engineering and Cognitive Computing, Ministry of Industry and Information Technology, School of Foreign Languages, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Existing machine translation models are usually trained on word-granularity data sets. However, different segmentations contain different grammatical, semantic features. Segmenting word granularity merely will interfere efficient training of neural machine translation (NMT) models, and is particularly prominent for Tibetan-related translation due to Tibetan linguistic features. Hence, for bidirectional Tibetan-Chinese NMT, we propose a multi-granularity training method focusing on syllables, words and phonetic fusion. We also propose a novel NMT model within the attention-based NMT framework, where a self-attention mechanism is incorporated into the decoder to capture more target-side information. Experimental results on CWMT2018 Tibetan-Chinese bilingual dataset show that the translation performance of the phonetic word fusion segmentation granularity significantly outperforms other segmentation granularity, and that integrating self-attention mechanism into the decoder can improve the translation quality greatly. In this paper, we also use the additional WMT2017 German-English bilingual dataset to demonstrate the universality of the proposed method across different languages.

Keywords: syllable words fusion; Tibetan-Chinese bidirectional; neural machine translation