文章编号:1000-5862(2020)02-0153-07

基于 CSGAN 的多模型融合蒙汉神经机器翻译研究

武子玉[1],侯宏旭[2*],白天罡[2],吉亚图[2],乌尼尔[2],郭紫月[2],王雪姣[2],孙 硕[2]
(1. 内蒙古大学计算机学院,内蒙古 呼和浩特 010021;2. 内蒙古自治区蒙古文信息处理技术重点实验室,内蒙古 呼和浩特 010021)

摘要:由于低资源语料稀少而导致的语义捕获不充分现象已成为影响机器翻译质量的主要因素. 为此,该文在预处理的基础上利用 CNN 和门控机制来改进 Transformer 模型,通过对抗训练的方式来引导模型参数的优化,同时通过加入命名实体识别来提高模型对实体的翻译性能. 此外,通过多模型融合的方式将来自多个机器翻译的输出经过改进、重组、合并转变为一个单一的改进的翻译结果. 通过 3 组对比实验表明,该方法优于基准方法.

关键词:蒙汉机器翻译;数据稀疏;系统融合;命名实体
中图分类号:TP 391.1 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2020.02.07

0 引言

蒙古语是一个广泛使用的跨多国、多地区的语言. 蒙汉翻译任务受众群体广、需求量大,所以对蒙汉机器翻译进行研究,以便于促进我国民族文化的交流与发展,这具有重要的意义. 但是这些翻译任务普遍存在资源稀少导致的数据稀疏和构词多样性导致的特征学习困难问题.

在本文中涉及 3 个系统:多个 CNN 融合系统、融合 CNN 门控机制的 Transformer 系统以及融合命名实体识别的多模型融合系统. 融合命名实体识别的多模型融合系统在 CCMT2019 蒙汉翻译任务中取得了本文所有系统的最好的实验效果($B_{\text{LEU5-SBP}}$ = 0.489 1),排第 2 和第 3 的系统分别为结合多个 CNN 模型的翻译系统以及融合了 CNN 门控机制的 Transformer 系统. 在蒙汉机器翻译任务中,常常会遇到平行语料稀缺导致的数据稀疏问题,通常会使用蒙古文校正技术和 Byte-Pair Encoding(BPE)切分技术,这 2 种关于语言方面的处理技术可以有效地缓解数据稀疏问题. 但在实验中发现,不是所有的切分粒度都有利于模型捕获语义关系. 在翻译任务中引入字符粒度切分,神经机器翻译模型就可以通过深层次的网络结构学习到更加深层次的语义信息,学习到更多的语言特征就会较大程度地提高翻译质量. 但是当将翻译粒度控制为字符级别时,句子长度会变为原先句子长度的 4 ~ 5 倍,此时需要学习更长距离下字符之间的依赖关系,因此在 Transformer 的基础上进行了修改,加入 CNN 用来学习长距离特征. 但是卷积层的输出有大量的无益信息,因此采用 GLU 门控单元对经过 CNN 卷积操作之后的输出进行无用信息的 2 次过滤,以得到更好的翻译结果. 针对翻译结果中出现的较多命名实体翻译不准确的情况,在数据的预处理过程中对语料进行了命名实体抽取以作为后期翻译结果的辅助词典对翻译结果进行矫正.

本文基于 CCMT2019 蒙汉测试集合,对 3 个系统进行了对比实验,并对实验结果进行分析.

1 数据预处理

1.1 基于 BPE 的双语切分

在机器翻译任务中对语料进行不同粒度的切分会产生不同的翻译效果. 因此,该步骤是语料预处理过程中非常重要的一步. BPE[1] 切分技术是采用迭代的方法,在每一轮的迭代中使用单个字节对出现次数最多的翻译单元对进行替换操作,而且这个作为替换的字节应该是不在待压缩的数据中的. 一句话包含的特征信息往往是由多个局部信息组成的,因此在对语料进行预处理的过程中,当切分得到的局部信息个数较少时,虽然可以对缓解数据稀疏问

收稿日期:2019-09-02
基金项目:内蒙古自然科学基金(2018MS06005)和内蒙古自治区科技成果转化"蒙古文机器翻译与辅助翻译云平台建设与推广"(2019CG028)资助项目.
通信作者:侯宏旭(1972-),男,山东济南人,教授,博士,主要从事自然语言处理、信息检索研究. E-mail:cshhx@imu.edu.cn

题有积极影响,但是对局部信息的获取就较为粗糙;相反地,当切分得到的局部信息个数较多时,就可以保留更为完整的局部特征,但是这样就会引起更加严重的数据稀疏问题. 因此,在双语平行语料特别稀缺的蒙古语-汉语机器翻译任务中,对句子切分粒度的把控十分重要.

关于 BPE 切分技术,就是将句子的组成部分(粒度)切分为介于词粒度和字粒度之间的单元,该方法在一定程度上可以在缓解数据稀疏问题的同时保留较多的句子局部信息,从而提高翻译模型的翻译质量. 目前的蒙古文文字的拉丁拼写方式与西欧地区以及世界各地使用拼音文字的国家和地区拼写方式类似[2]. 字节对编码切分方法是采用对拼音文字统计相邻字符共现次数的一种切分方法,共现次数高的连续字符通常可以当作是一个组合. 而针对蒙古文文字中的词根以及多个词缀的组合特性恰好就可以使用这样的切分方式进行切分,所以在蒙古语的预处理阶段对蒙古语进行 BPE 切分. 同时对汉语在已经分词的情况下进行 BPE 切分.

1.2 命名实体识别

本文利用人工方式对语料进行词级标注,并利用 CRF 模型进行命名实体识别训练[3],通过预先设定的特征模板,学习出符合特征模板要求的命名实体,当模型完成文档翻译时,将生成文档作为测试语料进行命名实体的识别,并将得到的命名实体和目标译文中的翻译单词进行替换得到最终的译文. 具体的方法如图 1 所示.

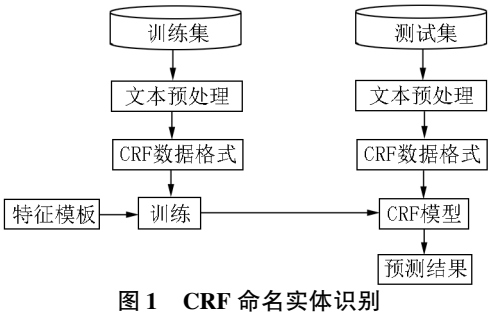

图 1 CRF 命名实体识别

采用 BMEWO 标注方式标注语料,将语料以字粒度为基础,B 表示实体的首部,M 表示实体中部,E 表示实体尾部,W 表示单个实体,O 表示非实体,同时对与每个实体又细分为时间命名实体 TIME、机构命名实体 ORGANIZATION、人名命名实体 NAME、地名命名实体 LOCATION 等,具体标注表示类别如表 1 所示.

表 1 具体标注表示类别

文字	类别	文字	类别	文字	类别
一	B_TIME	中	B_ORGANIZATION	中	B_LOCATION
九	M_TIME	共	M_ORGANIZATION	国	E_LOCATION
九	M_TIME	中	M_ORGANIZATION	合	O
八	M_TIME	央	E_ORGANIZATION	作	O
年	E_TIME	总	O	的	O
新	B_TIME	书	O	先	O
年	E_TIME	记	O	行	O

对于特征模板来说,主要采用 Unigram 形式特征模板,充分学习当前字符与其前后 n 个字符的关系,若符合特征模板条件要求则对当前所学内容的参数信息进行保存. 特征模板如图 2 所示.

```
#Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]/%x[-1,0]/%x[0,0]
U06:%x[-1,0]/%x[0,0]/%x[1,0]
U07:%x[0,0]/%x[1,0]/%x[2,0]
U08:%x[-1,0]/%x[0,0]
U09:%x[0,0]/%x[1,0]
```

图 2 特征模板图

本文暂时不考虑命名实体研究中的实体边界问题和识别混淆问题.

2 模型结构

2.1 CNN 系统

针对长句翻译中的长距离依赖问题,本文通过加入深层 CNN 来获取长距离依赖信息,并为每个卷积层配备一个注意力机制,但是卷积神经网络无法获得源语言句子的位置信息,因此在输入时针对输入词添加一个绝对位置信息,用以辅助 CNN 实现文本语义的抽取. 具体方式由 Fairseq 实现[4-5],这里关于 CNN 的输入设置为词向量及其对应的绝对位置向量的加权和.

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n),$$
$$\boldsymbol{w} = (w_1, w_2, \cdots, w_n),$$
$$\boldsymbol{p} = (p_1, p_2, \cdots, p_n),$$
$$\boldsymbol{e} = (w_1 + p_1, w_2 + p_2, \cdots, w_n + p_n),$$

其中 $\boldsymbol{x}$ 为输入序列,$\boldsymbol{w}$ 为输入序列对应的词向量,$\boldsymbol{p}$ 为位置向量,$\boldsymbol{e}$ 为 CNN 的输入向量. 解码器由多层卷积神经网络结构组成,为解码器的每个卷积层加入 1 个注意力机制[6].

$$\boldsymbol{d}_i^l = \boldsymbol{W}_d^l \boldsymbol{h}_i^l + \boldsymbol{b}_d^l + \boldsymbol{t}_i,$$
$$\boldsymbol{a}_{ij}^l = \exp(\boldsymbol{d}_i^l \boldsymbol{z}_j^u) / \sum_{t=1}^{n} \exp(\boldsymbol{d}_i^l \boldsymbol{z}_t^u),$$
$$\boldsymbol{c}_i^l = \sum_{j=1}^{n} \boldsymbol{a}_{ij}^l (\boldsymbol{z}_j^u + \boldsymbol{e}_j),$$

其中 $\boldsymbol{t}_i$ 表示生成的前一个目标词的词向量信息,$\boldsymbol{h}_i^l$ 表示解码器的第 l 层的第 i 个隐状态,$\boldsymbol{z}_j^u$ 表示编码器的最后一层的第 j 个输出,$\boldsymbol{a}_{ij}^l$ 表示第 l 层的第 i 个隐状态同源端句子中第 j 个单元的对齐信息,$\boldsymbol{e}_j$ 表示编码器的第 j 个输入,$\boldsymbol{c}_i^l$ 表示解码器的第 l 层的第 i 个注意力向量.

2.2　结合 CNN 门控机制的 Transformer 模型

Transformer 基线模型是 A. Vaswani 等[6] 提出的一种基于注意力机制的翻译模型. 解码器由多个包含自注意力模块的子层构成,用于获取目标语言单词之间的语义关系,并通过前馈网络层将整体的向量信息汇总,便于后续解码推理. 每个子层均由残差连接[7] 并进行层正规化操作[8]. 本层为全连接层,该层的激活函数为 ReLU 函数[9],本层的输出位对输入的信息进行了 2 次线性变换后的结果.

解码层是由多个 CNN 结构堆叠而成的,每个 CNN 结构由 3 层组合而成. 第 1 层为自注意力层,用以计算编码器中向量之间的自注意力权向量;第 2 层用以计算编码器的输出和解码器隐层状态的注意力向量;第 3 层为全连接前馈神经网络层,用以计算向量权重,该层结构与编码器中的全连接前馈神经网络结构一致,该层输出由下式得到:

$$F_{FN}(x) = \max(0, \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2,$$

其中 $\boldsymbol{x}$ 为编码器输入,$\boldsymbol{W}_1$ 是经过第 1 次线性变换的参数矩阵,$\boldsymbol{b}_1$ 是第 1 次线性变换的偏置向量,$\boldsymbol{W}_2$ 是经过第 2 次线性变换的参数矩阵,$\boldsymbol{b}_2$ 是第 2 次线性变换偏置向量.

在采用的融合 CNN 门控机制的 Transformer 模型中,将卷积网络和门控单元[10] 加入到模型的输入层和编码器中,抽取关键的语义信息以及过滤掉噪声信息,经过残差连接[11] 传递到编码器中进行其他编码的计算. 通过增加卷积层网络层数来获取深层语义信息以达到更好的效果.

每层输出的计算方式为 $O_l(\boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{W} + \boldsymbol{b}_w) \otimes \sigma(\boldsymbol{X}\boldsymbol{V} + \boldsymbol{b}_v)$,其中 d 为输入词向量的维度,k 为卷积核的宽度,n 为输出向量的维度,$\boldsymbol{X} \in \mathbf{R}^d$ 用来表示第 l 层卷积层的输入向量,$\boldsymbol{W} \in \mathbf{R}^{k \times d \times n}$ 是卷积核的参数矩阵,$\boldsymbol{b}_w \in \mathbf{R}^n$ 是卷积核的偏置,$\otimes$ 为逐元素相乘操作,σ 是一个 sigmoid 函数,$\boldsymbol{V} \in \mathbf{R}^{k \times d \times n}$ 是 GLU 的卷积核矩阵,$\boldsymbol{b}_v \in \mathbf{R}^n$ 是 GLU 卷积核的偏置,$O_l(\boldsymbol{X})$ 用来表示第 l 层的输出. 结合基于门控机制的 CNN 与 Transformer 融合模型的结构图如图 3 所示,将输入层的词向量传入到卷积层中,其中每个卷积层输出的一部分用来计算 GLU,剩下的部分作为卷积层的输出同 GLU 的输出进行逐元素相乘.

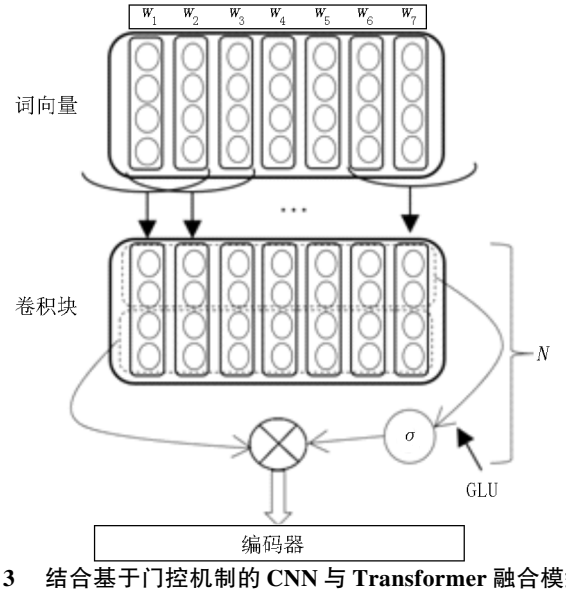

图 3　结合基于门控机制的 CNN 与 Transformer 融合模型的结构图

2.3　模型训练

本文使用基于 M. Mirza 等[12] 提出的条件约束生成对抗网络(Conditional Generative Adversarial Networks,CSGAN),针对蒙汉机器翻译任务的特点进行了改进. 由于蒙汉任务资源稀缺,模型接受的数据存储较为稀疏,这使得模型很难在少量的语料中发现有用的上下文信息. 对于这个问题,提出一种多粒度混合策略,在此基础上添加一个基于价值迭代的过滤器,用于帮助模型识别当前序列最合适的粒度. 结构图如图 4 所示.

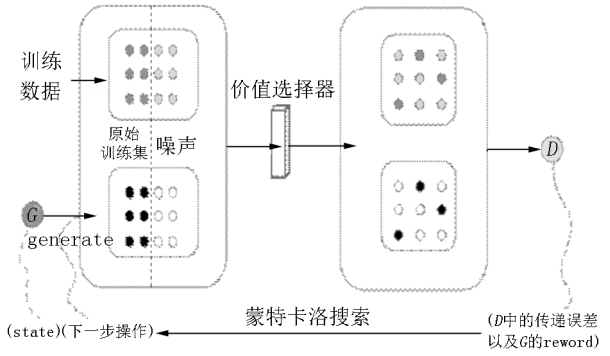

图 4 CSGAN(Transformer/LSTM) 结构图

所谓噪声,其实就是对原始训练语料经过不同粒度切分后产生的语料,这里可以理解为伪数据. 生成器 G 用于译文的产生,本文中的生成器为前面提到的几种模型架构;将生成的译文通过价值选择器,过滤掉部分无用信息传入判别器;判别器 D 用于区分生成器 G 生成的译文与真实译文,判别器 D 采用深度卷积结构. 生成器 G 采用策略梯度进行训练,计算公式为

$$J(\theta) = \sum_{\boldsymbol{Y}_{1:N}} G_\theta(\boldsymbol{Y}_{1:N} \mid \boldsymbol{X}) R_D^{G_\theta}(\boldsymbol{Y}_{1:N-1}, \boldsymbol{X}, \boldsymbol{y}_N, \boldsymbol{Y}^*),$$

G_θ 为生成器中源端句子为 $\boldsymbol{X}$ 的情况下目标端句子为 $\boldsymbol{Y}$ 的概率. 句子的奖励使用蒙特卡洛搜索进行计算,$R_D^{G_\theta}$ 表示最后一个词之前累积奖励加上当前奖励. $R_D^{G_\theta}$ 由下式计算得到:

$$R_D^{G_\theta} = D(\boldsymbol{X}, \boldsymbol{Y}_{1:N}) - b(\boldsymbol{X}, \boldsymbol{Y}_{1:N}),$$

D 由翻译译文与真实译文来进行训练,使用判别器作为奖励函数能够进一步让生成器迭代训练从而动态更新判别器,进而得到更加真实的生成序列. 随后再训练生成器,生成器内部参数更新公式为

$$\nabla J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{\boldsymbol{y}_n} R_D^{G_\theta}(\boldsymbol{Y}_{1:N-1}, \boldsymbol{X}, \boldsymbol{y}_N, \boldsymbol{Y}^*) \cdot$$

$$\nabla_\theta(G_\theta(\boldsymbol{Y}_{1:N} \mid \boldsymbol{X})) = \frac{1}{N} \sum_{n=1}^{N} E_{\boldsymbol{y}_n \in G_\theta}(R_D^{G_\theta}(\boldsymbol{Y}_{1:N-1}, \boldsymbol{X}, \boldsymbol{y}_N, \boldsymbol{Y}^*) \nabla_\theta \log p(\boldsymbol{y}_n \mid \boldsymbol{Y}_{1:n-1}, \boldsymbol{X})).$$

D 在训练中随着 G 同步更新.

2.2 多模型融合

本文使用的多模型融合系统是 Kenneth Heafied 等[13-14]开发的开源的基于词级别融合系统 Multi-Engine Machine Translation(MEMT). 对模型进行融合的目的是利用不同模型的优势对来自多个不同的翻译系统的翻译结果进行改进、重组以及合并以得到在 BLEU 值以及语言的流利度等多方面均有提升的翻译译文. MEMT 融合模型主要由 3 部分组成,它们分别是对齐方式、搜索空间以及多特征评分.

首先是对齐方式. 与混淆网络中使用的对齐方式[15]不一样,MEMT 使用的是将每个翻译系统的输出都和其他的翻译假设使用 METEOR[16]进行对齐计算,对齐计算后的结果包含更多的语义信息. 将该结果作为该融合系统的输入(见图 5).

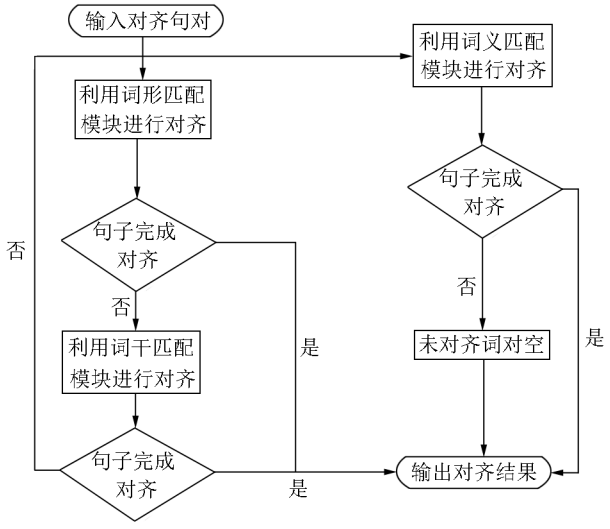

图 5 METEOR 的对齐方式

其次是搜索空间. 要翻译一句话,从第 1 个待翻译单元开始,搜索空间可以在这句话里面进行搜索也可以跳转到别的句子中进行搜索. 对这些翻译模型进行融合的目标是将多个系统的输出进行组合优化以达到更好的翻译译文,而想要实现这个目标必须要确保重组后的句子中各组成片段不重复,而且这些片段可以在不同的句子中进行转换,在搜索的过程中使用启发式方法.

最后是多特征评分. 句子的流畅度是保证搜索空间在句子之间灵活转换的必要条件. 因此,在这个融合系统中使用多个特征对句子进行打分、完善,从而保证句子的流畅度. 针对特征的选择问题,采用了句子长度约束,即翻译假设长度(类似 Moses[17]),弥补句子长度对其他特征的影响;语言模型,为了保证翻译结果的流利度采用 n-gram 语言模型;回退限制,在 n-gram 模型中能匹配到 N 元短语的平均长度可以对回退行为进行约束;匹配程度,参考译文中与翻译假设可以相配的 N 元短语的数量. 这些特征通过线性加权得到最终的特征函数,在实验的过程中,发现匹配特征的权重较大程度上依赖于底层的系统构建,所以在的融合系统中使用 Z-MERT 对权重进行调整. 在该系统的解码阶段采用柱搜索(beam search)算法,在对齐后的结果上使用柱搜索进行解

码,对解码后的高分译文进行命名实体的修正,得到最终的翻译译文(见图 6).

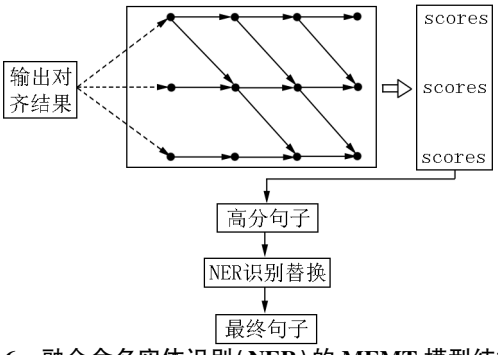

图 6　融合命名实体识别(NER)的 MEMT 模型结构

3　实验

3.1　实验数据

本文的实验数据均是由 CCMT2019 提供的蒙汉日常用语翻译评测任务的 26 万句对平行训练语料,1 000 句对验证集平行语料和 1 001 句对的测试集平行语料. 实验的实际训练数据包括训练集和开发集经过长度处理后的全部数据. 在数据的预处理阶段,蒙古文使用蒙古文校正技术进行校正,蒙汉双语均使用 BPE 切分技术进行切分.

3.2　实验配置

本文使用的基线 CNN 系统是 Facebook AI Research 提供的开源系统 fairseq. 其中编码器层数为 5 层,解码器层数为 9 层,解码器的结构在原来的基础上每一层配一个注意力机制,编码器和解码器的核宽度均为 3,词向量维度为 512,隐层单元个数设置为 512,batch size 为 32,训练过程中使用 Nesterov's accelerated gradient(NAG). Transformer 系统中编码器和解码器层数均为 6 层,batch size 为 2 048,其他配置与 CNN 系统相同. 结合基于门控机制的 CNN 与 Transformer 融合模型由 4 层 CNN 组成,卷积核宽度设置为 5,并在每一层的卷积层后附加一个线性门控单元(GLU),编码器层数为 6,解码器层数为 6. 词向量维度 512,隐层单元个数设置为 512,batch size 为 2 048.

3.3　实验结果与分析

表 2 为 3 个系统在 BLEU5-SEP、METEOR、ICT 以及 TER[18] 指标下的不同得分. 观察表 2 可以发现,融合命名实体识别的多模型融合系统(Primary a)取得的效果最好,其次是多个 CNN 融合系统(Contrast b),最后是融合 CNN 门控机制的 Transformer 系统(Contrast c).

表 2　实验结果

系统	评测指标			
	BLEU5-SBP	METEOR	ICT	TER
Primary a	0.489 1	0.675 9	0.538 9	0.337 4
Contrast b	0.476 3	0.658 6	0.495 8	0.355 6
Contrast c	0.420 2	0.606 2	0.465 0	0.427 2

根据不同句子长度,对 3 种模型的 BLEU 值进行了分析(见图 7).

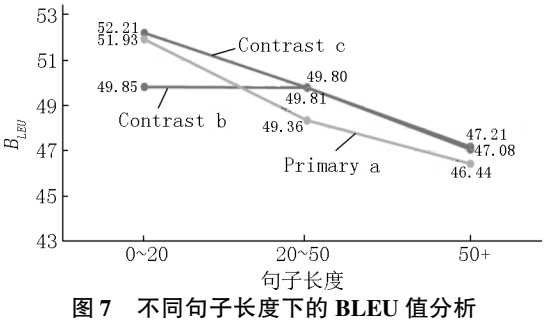

图 7　不同句子长度下的 BLEU 值分析

分析图 7 可发现,融合 CNN 门控机制的 Transformer 模型翻译效果并未达到预期要求,与在 CWMT2009 进行实验所得到的实验结果不符. 因此,对 CWMT2009 与 CCMT2019 的语料进行分析,2 者实验数据统计如表 3 所示,其中 Mo 和 Ch 2 项分别代表实验语料中蒙语和汉语句子的平均长度. 经过对实验数据规模的观察和比对,认为 Transformer 变体对蒙古文进行切字母处理会使得句子长度是切词的 4 ~5 倍. 这一操作本就增长了句子长度,而 CCMT2019 中又存在长句,加重了长距离依赖问题,因此导致融合 CNN 门控机制的 Transformer 系统的效果并不理想. 如表 4 所示,发现 Transformer 变体在部分短句的翻译上表现得更加细致,能够翻译出更多细节,表现比 CNN 更好.

表 3　CWMT2009 与 CCMT2019 实验数据规模对比

数据集		训练集		验证集		测试集	
		2009	2019	2009	2019	2009	2019
双语句对		66 808	261 643	1 000	1 001	1 000	1 001
规模		18.3 MB	112.0 MB	214 kB	782 kB	213 kB	316 kB
平均长度	蒙	40	71	23	82	27	60
	汉	74	123	37	286	39	115

表 4 Transformer 变体在长短句上的表现

句子		长句	短句
	原句	(ᠵᠢᠵᠢᠩᠭᠦᠯᠬᠦ ᠬᠡᠷᠡᠭ ᠪᠦᠲᠦᠶᠢᠯᠡᠬᠦ ᠵᠢᠨ ᠬᠡᠷᠡᠭ ᠬᠥᠳᠥᠯᠭᠡᠬᠦᠶᠢᠨ ᠬᠣᠯᠪᠣᠭᠳᠠᠯ ᠳᠤ ᠬᠡᠷᠡᠭ ᠬᠡᠷᠪᠡ ᠵᠢᠨ ᠬᠥᠳᠥᠯᠭᠡᠬᠦᠶᠢᠨ ᠬᠡᠷᠡᠭᠵᠢᠭᠡᠯᠡᠭᠡᠨ)	ᠲᠤᠰᠬᠠᠢ ᠮᠡᠳᠡᠭᠡᠯᠡᠯ ᠳᠤ ᠪᠠᠷᠢᠭᠠᠳᠠᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠶᠢᠨ᠎ᠠ ᠃᠃
	目标译文	(六)加强人民政协理论研究,以理论创新推动工作创新	电视里正在介绍西藏
译文	CNN	(六)加强人民政协理论研究,推动理论创新促进工作创新	电视刚刚介绍西藏
	融合 CNN 门控机制的 Transformer	(六)加强政协研究,推动理论创新推动创新	电视正在介绍西藏

4 总结

本文构建了 3 个基于不同神经网络架构的蒙汉翻译系统,利用 CNN 和门控机制改进 Transformer 模型,通过使用条件约束对抗训练的方式来优化参数,使得模型性能得到良好的提升. 语料的预处理对其翻译质量会有较大的影响,因此针对蒙古文形态多样复杂的特性,使用了蒙古文校正技术以及 BPE 切分技术对语料进行预处理. 针对低资源语言机器翻译模型中命名实体翻译效果不理想的情况,使用了命名实体抽取融合以及多模型融合的技术改善最终的译文质量. 通过实验发现大语种的机器翻译任务上使用的方法,并不完全适合蒙汉双语的翻译任务,因此运用合适的方法,或者将其他任务上的方法修改成更适合蒙汉翻译的方法显得更为重要. 在今后的工作中,也会致力于关于语料预处理部分的研究以及构建更适合蒙汉翻译这类翻译的机器翻译模型.

5 参考文献

[1] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [EB/OL]. [2019-03-16]. https://arxiv. org/abs/1508. 07909.

[2] 清格尔泰. 蒙古语语法 [M]. 呼和浩特:内蒙古人民出版社,1992.

[3] Le H Q, Nguyen T M, Vu S T, et al. D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine – tuned embeddings of various linguistic information [J]. Bioinformatics, 2018, 34(20): 32-36.

[4] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning [EB/OL]. [2019-03-19]. https://arxiv. org/abs/1705. 03122.

[5] Gehring J, Auli M, Grangier D, et al. A convolutional encoder model for neural machine translation [EB/OL]. [2019-03-19]. https://arxiv. org/abs/1611. 02344.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [EB/OL]. [2019-03-13]. https://arxiv. org/abs/1706. 03762.

[7] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [EB/OL]. [2019-04-13]. https://arxiv. org/abs/1512. 03385.

[8] Ba J L, Kiros J R, Hinton G E. Layer normalization [EB/OL]. [2019-04-18]. arXiv preprint arXiv: 1607. 06450, 2016.

[9] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks [EB/OL]. [2019-04-22]. https://www. researchgate. net/publication/215616967 _ Deep _ Sparse _ Rectifier_Neural_Networks.

[10] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated linear units [EB/OL]. [2019-04-25]. https://arxiv. org/abs/1612. 08083.

[11] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [EB/OL]. [2019-04-29]. https://arxiv. org/abs/1512. 03385.

[12] Mirza M, Osindero S. Conditional generative adversarial nets [EB/OL]. [2019-03-29]. https://arxiv. org/abs/1411. 1784.

[13] Bangalore S, Bordel G, Riccardi G. Computing consensus translation from multiple machine translation systems [EB/OL]. [2019-03-12]. http://citeseerx. ist. psu. edu/viewdoc/summary? doi = 10. 1. 1. 69. 4528.

[14] Heafield K, Lavie A. Combining machine translation output with open source: the carnegie mellon multi-engine machine translation scheme [J]. Prague Bulletin of Mathematical Linguistics, 2010, 93(1): 27-36.

[15] Rosti A V I, Matsoukas S, Schwartz R M. Improved word-

level system combination for machine translation [C]. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague: DBLP, 2007: 312-319.

[16] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments [EB/OL]. [2019-03-02]. http://www.wendangku.net/doc/ed5d2e83e53a580216fcfe46.html.

[17] Hieu Hoang, Alexandra Birch, Chris Callison-burch, et al. Moses: open source toolkit for statistical machine translation [EB/OL]. [2019-05-02]. https://dl.acm.org/citation.cfm? doid=1557769.1557821.

[18] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation [EB/OL]. [2019-05-05]. https://www.researchgate.net/publication/228668187_A_study_of_translation_edit_rate_with_targeted_human_annotation.

The Research on Multi-Model Fusion for Mongolian-Chinese Neural Machine Translation Based on CSGAN

WU Ziyu[1], HOU Hongxu[2*], BAI Tiangang[2], JI Yatu[2], WU Nier[2], GUO Ziyue[2], WANG Xuejiao[2], SUN Shuo[2]

(1. College of Computer Science, Inner Mongolia University, Hohhot Inner Mongolia 010021, China; 2. Inner Mongolia A. R. Key Laboratory of Mongolian Information Processing Technology, Inner Mongolia University, Hohhot Inner Mongolia 010021, China)

Abstract: The phenomenon of insufficient semantic capture due to the scarcity of low-resource corpus has become a major factor affecting the quality of machine translation. Therefore, based on the pretreatment, the paper improves the transformer model by using CNN and gating mechanism and guides the optimization of model parameters using confrontation training. At the same time, named entity recognition is added to improve the translation performance of model entities. Reorganize and merge the output from multiple machine translations into a single improved translation result through multi-model fusion. The Mongolian-Chinese translation experiments show that the proposed method is superior to the benchmark method.

Key words: Mongolian-Chinese machine translation; data sparsity; fusion system; named entity

(责任编辑:冉小晓)