

文章编号: 1003-0077(2020)01-0058-05

基于 transformer 神经网络的汉蒙机构名翻译研究

安苏雅拉, 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 010022)

摘要: 机构名翻译是机器翻译的研究内容之一, 在机器翻译任务中机构名翻译的准确度, 直接影响着翻译性能。在很多任务上, 神经机器翻译性能优于传统的统计机器翻译性能, 该文中使用基于 transformer 神经网络模型与传统的基于短语的统计机器翻译模型和改进后的基于语块的机器翻译模型做了对比试验。实验结果表明, 在汉蒙机构名翻译任务上, 基于 transformer 神经网络的汉蒙机构名翻译系统优于传统的基于语块的汉蒙机构名翻译系统, BLEU4 值提高了 0.039。

关键词: 神经网络; 汉蒙机器翻译; 机构名

中图分类号: TP391

文献标识码: A

Chinese-Mongolian Organization Name Translation Based on Transformer

AN Suyala, WANG Siriguleng

(School of Computer Science and Technology, Inner Mongolia Normal University,
Hohhot, Inner Mongolia 010022, China)

Abstract: Organization name translation directly affects translation performance. In this study, a transformer-based neural network model is proposed for this task. Compared with a traditional phrase-based SMT model and an improved block-based MT model, the experimental results show that the transformer NMT increased by 0.039 in terms of BLEU 4 in the Chinese-Mongolian Organization name translation task.

Keywords: neural network; Chinese-Mongolian machine translation; organization name

0 引言

随着社会的发展, 机器翻译成为了我们生活中必不可少的技术。翻译技术方面, 从基于规则的方法到目前的基于神经网络的方法, 性能不断在提高。而在翻译对象方面, 命名实体、领域术语等有实际意义的对象, 成为了研究人员的关注点。命名实体(named entity, NE)是指具有特定意义的实体, 主要包含专有名词、人名、地名和机构名等, 其一般承载着主要的信息。因此, 机器翻译中命名实体的翻译准确度, 对译文的整体翻译质量具有重要影响。人名、地名可以通过查字典或音译的方法进行翻译, 而机构名本身结构复杂, 造成了翻译困难。由于基于

神经网络的方法在蒙汉机器翻译任务上取得了非常好的效果^[1], 所以在本研究中, 利用基于 transformer 神经网络的机器翻译模型实现了汉蒙机构名翻译。

1 相关工作

机器翻译从最早的基于规则的机器翻译到如今的神经机器翻译, 其性能在不断提高^[2-6]。神经机器翻译是一种新兴的机器翻译方法, 由 Kalchbrenner 和 Blunsom, Sutskever 等和 Cho 等提出^[7-9]。在此基础上, 2015 年 Bahdanau 等^[10]提出了注意力机制, 明显提高了神经机器翻译模型的性能。2017 年, 谷歌发表的论文^[11]中提出了一种新的网络架构, 称为 transformer, 其抛弃了传统的卷积神经网络(convolutional

收稿日期: 2019-08-15 定稿日期: 2019-10-09

基金项目: 国家自然科学基金(61762072); 内蒙古自然科学基金(2016MS0623)

neural networks, CNN)或者循环神经网络(recurrent neural network, RNN)的定式,超越了以往的算法,激起了工业界和学术界的广泛讨论,成为了研究热点。

随着机器翻译技术的发展,汉蒙机器翻译经历了从基于规则的机器翻译^[12]、基于实例的机器翻译^[13],到基于短语的统计机器翻译^[14]。但在汉蒙命名实体自动翻译方面的研究较少。杨萍^[15]运用了非对称的“汉文”新蒙古文命名实体对齐策略完成命名实体的翻译。藏丹^[16]利用基于语块的方法做了“汉文”到传统蒙古文机构名自动翻译研究。哈斯高娃^[17]利用 transformer 神经网络模型做了蒙汉机器翻译研究。本研究中,以基于短语的实验为基线实验,做了基于 transformer 神经网络的汉蒙机构名翻译对比实验。

2 基于语块的汉蒙机构名翻译系统

该研究首先利用 CRF 对“汉文”机构名做切分,对地名语块和机构标识语块做 XML 标记,然后分别对 XML 标记的语块做模块翻译,最后利用 Moses 解码器完成整个机构名的翻译。基于语块的汉蒙语机构名翻译系统的流程如图 1 所示。

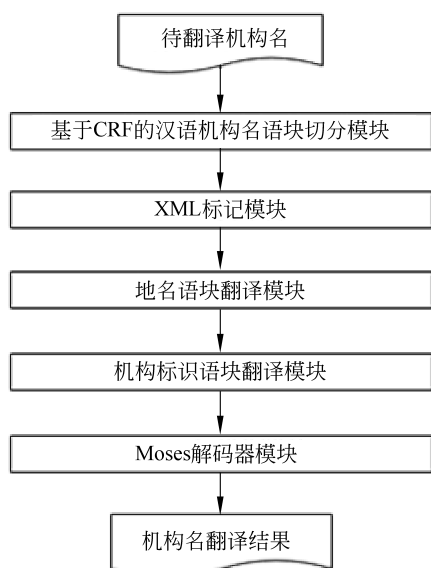


图 1 基于语块的汉蒙语机构名翻译系统的流程图

基于 CRF 的汉语机构名切分模块,利用开源的 CRF++-0.58 工具包实现了机构名切分,主要完成了以下几步工作:语料库标注、基于 CRF 的切分模块的特征模板制定、训练、测试及评测。其中语料标注是对 1 万个机构名人工做了 BIO 标注。

XML 标记是对已切分好的语料进行加工,对地名语块和机构标识语块做标注,以便地名语块翻

译模块和机构标识语块翻译模块能识别其需要翻译的词语。

地名翻译模块使用查词典的方法实现,为此建立了汉蒙地名词典。该词典中有 9 273 个地名词条,包括国家名、世界常用城市名、世界主要山川河海名、中国各省及主要城市名和内蒙古行政区划名。若要翻译的地名不在词典中,则利用音译和意义相结合的方法进行翻译。最后把翻译结果写到 XML 标记中。

机构标识语块翻译模块使用了音译的方法进行翻译。汉蒙翻译中机构标识语是一个商家或者企业的代号,每一个机构名都有各自的标识,用音译的方法把汉语机构标识翻译成蒙古语,再把翻译结果写到 XML 标记中。

通过 XML 标记可以把预处理的地名语块和机构名标记语块的翻译引入到 Moses 解码器中,不需要重新训练或更改翻译模型,直接对剩余的语块进行翻译,最后得到完整的机构名翻译结果。

3 基于 transformer 的汉蒙神经机器翻译系统

Transformer 模型的结构由编码器—解码器组成。编码是将输入序列 (x_1, x_2, \dots, x_n) 转换成一个连续的序列表示 $z = (z_1, z_2, \dots, z_n)$, 解码过程是根据前面的 z 将生成对应的输出序列 (y_1, y_2, \dots, y_m) 。其中编码器有 6 层,每一层包含两个子层,分别是多头注意力层(multi-head attention layer)和全连接前馈网络层(feed forward),每个子层之间用残差连接(residual connection)。解码器也有 6 层,每一层包含三个子层,分别是多头注意力层和编码器—解码器注意力层,最后是全连接前馈网络层。masking 的作用是防止在训练的时候使用未输出的单词。其模型结构如图 2 所示,其中左半部分是编码器,右半部分是解码器。

位置向量(positional encoding)是对输入序列中每个词的顺序进行编码,对每一个输入的词向量从 0 开始依次标记。位置向量是采用正弦函数和余弦函数,将位置索引为 pos 的词向量映射为一个 d_{model} 维的位置向量。如式(1)、式(2)所示, pos 是位置索引, i 是索引对应的向量值。

$$\text{PE}_{(\text{pos}, 2i)} = \sin(\text{pos}/1000^{2i/d_{\text{model}}}) \quad (1)$$

$$\text{PE}_{(\text{pos}, 2i+1)} = \cos(\text{pos}/1000^{2i/d_{\text{model}}}) \quad (2)$$

多头注意力机制(multi-head attention)是该模型的主要模块。注意力机制是对源语言中元素的

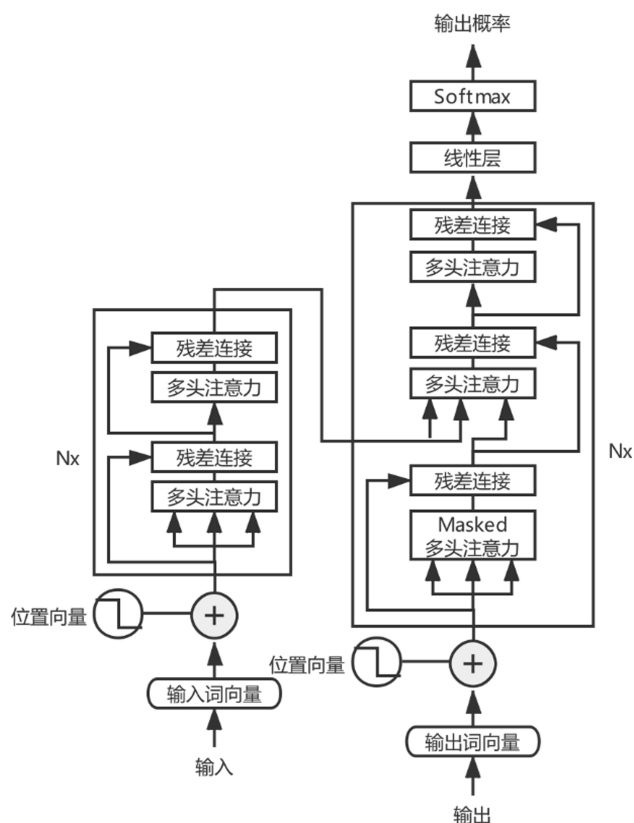


图2 transformer 模型结构

Value 值进行加权求和,而 Query 和 Key 用来计算对应 Value 的权重系数。Query、Key、Value 首先经过一个线性变换,然后输入到 Scaled Dot-Product Attention,做 h 次训练,一次算一个头,头之间参数不共享,每次 Q 、 K 、 V 进行线性变换的参数是不一样的。然后将 h 次的 Scaled-Dot-Product Attention 结果进行连接,再进行一次线性变换,将得到的值作为多头注意力的结果,如图 3 所示。

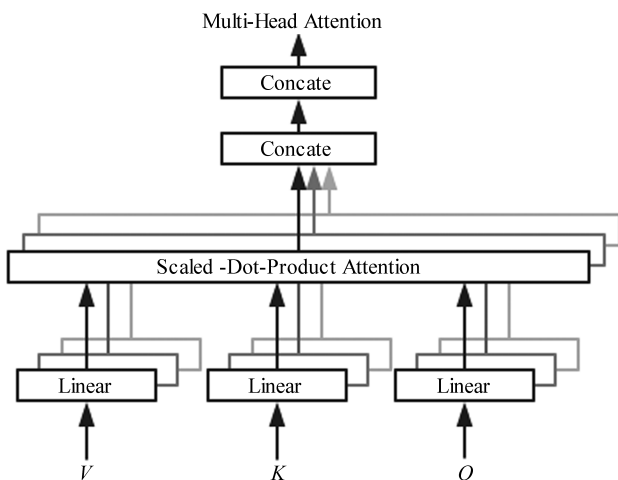


图3 多头注意力机制框架

基于 transformer 神经网络的汉蒙机构名翻译

系统,有三个模块,分别是数据生成模块、训练模块和翻译模块。

数据生成模块的主要功能是要生成在训练模型时用的语料。其中包括定义词汇表的大小,训练数据集、开发集和测试集的位置。首先生成汉蒙语料的词汇表,再把词频少于 \min_cnt 的词标成 $\langle unk \rangle$,最终生成数据集。

训练和翻译模块中,词向量训练是 transformer 模型自带的方法,首先将对齐语料的每个词表示成词向量(embedding),并加上每个词的位置向量,得到新的词向量,把新的词向量输入到编码器中,每次输入 $batch_size$ 大小的对齐语料。编码器中的输入通过自注意力层进行残差连接和正则化(layer normalization),再进入全连接前馈网络层进行残差处理以及正则化。最后输出的结果作为下一个编码器的输入,经过 6 次重复操作后,作为输入到解码器的部分。翻译模块是解码器收到编码器的输出,然后和位置向量相加。输出结果通过解码器端的自注意力层和全连接前馈网络层进行解码,通过 6 层解码器得到的向量会经过 linear 层和 softmax 层完成解码。

本实验每隔 1 000 步会自动保存生成的模型,模型只保留最后 20 个模型,且在训练模型期间可以使用已生成的模型将测试语料翻译出来。

4 实验及结果分析

基于 transformer 神经网络实验的硬件环境为: 32GB 内存、GTx1080Ti 独立显卡、i7 8700k 处理器和 256GB 硬盘。在 Ubuntu 16.04 版本 64 位操作系统环境下,搭建了 GPU 版本的 TensorFlow, Python 版本为 3.5。

4.1 实验数据

本研究中的训练语料包括汉蒙对齐语料库和汉蒙词典两类。汉蒙对齐语料库包括通用汉蒙句子对齐语料库和汉蒙机构名对齐语料库。汉蒙词典包括汉蒙短语词典、汉蒙地名词典和汉蒙机构名词典。规模如表 1 所示。

表1 训练语料规模

语料编号	训练语料名称	语料数量
1	汉蒙句子对齐语料库	166 324
2	汉蒙短语词典	212 779
3	汉蒙地名词典	9 273
4	汉蒙机构名词典	7 096
5	汉蒙机构名对齐语料库	19 110

实验使用了与文献[16]相同的语料,训练语料开发集为 500 条、测试集为 111 条,评测标准答案只有 1 个译文,在此数据集上我们做了基于 transformer 神经网络的汉蒙机构名翻译实验。由于文献[16]中采用的测试集数据较少,为了验证模型的实际泛化能力,我们从 500 个句子的开发集中抽出 189 个句子,扩充到测试集中,这样开发集为 311 个句子,测试集为 300 个句子,同时把评测标准答案做了 4 个不同的译文,在此基础上做了基于语块的汉蒙机构名翻译实验和基于 transformer 神经网络的汉蒙机构名翻译实验。

汉语语料使用中科院汉语分词器 NLPPIR2016 进行了分词处理,对一些由蒙古语音译过来的地名分词错误,根据汉蒙地名词典进行了自动修正。蒙古文语料做了 Unicode,转换成内大拉丁转写形式,并做了基于词素的部分切分,即名词格附加成分的分分。

基于 transformer 神经网络的汉蒙机构名翻译系统的参数: batch_size 为 32,学习率为 0.001,隐藏单元的个数为 512,min_cnt 为 20,dropout_rate 为 0.1,设置同文献[11]。

4.2 实验结果分析

基线实验 1 是使用基于短语的统计机器翻译方法做的汉蒙机构名翻译实验,最好的 BLEU4 值达到了 0.700 5。用基线实验的语料做了基于语块的实验 2^[16]和基于 transformer 神经网络的实验 3,结果如表 2 所示。从表 2 中可以看出,基于 transformer 神经网络的实验 3 的结果优于基线实验和基于语块的实验 2 的结果,BLEU4 值分别提高了 0.056 9 和 0.014 7。这表明基于 transformer 神经网络的汉蒙机构名翻译系统优于基于短语和基于语块的汉蒙机构名翻译系统。

实验 4 和实验 5 分别是对于开发集和测试集进行调整平衡后做的基于语块的实验和基于 transformer 的实验。从实验结果可以看出,基于 transformer 神经网络的实验明显优于基于语块的实验,BLEU4 值提高了 0.039。

表 2 基线实验、基于语块的实验和基于 transformer 的实验在不同测试集上的 BLEU4 值

实验编号	实验名称	测试集	评测结果
			BLEU4
1	基线实验	111 句	0.700 5

续表

实验编号	实验名称	测试集	评测结果
			BLEU4
2	基于语块	111 句	0.742 7
3	基于 transformer	111 句	0.757 4
4	基于语块	300 句	0.423 3
5	基于 transformer	300 句	0.462 3

我们对基于 transformer 实验的错误翻译结果进行了分析,翻译错误的主要原因有以下两个方面:一个是机构名中出现命名实体简称时翻译不准确,另一个是机构名太长时翻译不准确。

测试集中出现的命名实体简称总共有 46 个词,其中翻译错误的有 18 个词,包含重复出现的词。实际翻译例子如下所示。

源语言: 包钢白云鄂博铁矿区职工医院

transformer 译文: B0V BAI YUN 0B0G _ A AGVRHAI-YIN AJILTAN-V EMNELGE-YIN H0RIY_A ○

《 ᠪᠠᠭᠠᠨ ᠪᠠᠢ ᠶᠤᠨ ᠣᠪᠣᠭ ᠠ ᠠᠭᠦᠷᠬᠠᠢ-ᠶᠢᠨ ᠠᠵᠢᠯᠲᠠᠨ-ᠦ ᠡᠮᠨᠡᠯᠭᠡ-ᠶᠢᠨ ᠬᠣᠷᠢᠶ ᠠ ᠣ 》

参考答案: B0GVTV-YIN BVLVD-VN BAI YUN 0B0G _ A TEMUR-VN AGVRHAI-YIN AJILTAN-V EMNELGE-YIN H0RIY_A

《 ᠪᠠᠭᠠᠨ ᠪᠠᠢ ᠶᠤᠨ ᠣᠪᠣᠭ ᠠ ᠠᠭᠦᠷᠬᠠᠢ-ᠶᠢᠨ ᠠᠵᠢᠯᠲᠠᠨ-ᠦ ᠡᠮᠨᠡᠯᠭᠡ-ᠶᠢᠨ ᠬᠣᠷᠢᠶ ᠠ ᠠ ᠠᠶᠢᠷ ᠠ 》

例子中“包钢”是“包头钢铁”的缩写,因训练语料中没有出现“包钢”一词,所以 transformer 采用音译的方式对它进行了翻译,但是只翻译了第一个字,漏掉了对第二个字的翻译。

测试集中最长的机构名有 43 个字,最短的机构名有 4 个字,平均长度为 16 个字。我们发现机构名长度超过 20 个字时,其模型总是不能翻译后面几个字。在测试集中长度超过 20 个字的机构名有 50 个。例如:

源语言: 永鑫保险销售服务有限公司内蒙古分公司包头营业部

transformer: YUNG SIN B0R0LAGVLVLTA-YIN UILECILEGEN-U HIJAGARTV GUNGSe-YIN OBOR M0NGG0L-VN GUNGSe EJENGNELTE-YIN GAJAR

《 ᠶᠤᠩ ᠰᠢᠨ ᠪᠣᠷᠣᠯᠠᠭᠦᠯᠦᠯᠲᠠ-ᠶᠢᠨ ᠤᠢᠯᠡᠴᠢᠯᠡᠭᠡᠨ-ᠤ ᠬᠢᠵᠠᠭᠠᠷᠲᠤ ᠭᠤᠩᠭᠤᠰᠡ-ᠶᠢᠨ ᠣᠪᠣᠷ ᠮᠣᠩᠭᠣᠯ-ᠶᠢᠨ ᠭᠤᠩᠭᠤᠰᠡ ᠡᠵᠡᠩᠭᠡᠨᠡᠯᠲᠡ-ᠶᠢᠨ ᠭᠠᠵᠠᠷ 》

参 考 答 案: YUNG SIN DAGADHAL B0R0LAGVLHV UILECILEGEN-U HIJAGARTV

