内蒙古师范大学 CCMT2021 蒙汉机器翻译系统评测技术报告

何乌云，包晶晶，萨出拉，陈美兰，特日格勒呼，王斯日古楞*

（内蒙古师范大学 计算机科学技术学院, 内蒙古自治区 呼和浩特市 011500）

摘要：本文介绍了内蒙古师范大学计算机科学技术学院蒙古文大数据研究基地参加 CCMT2021 机器翻译评测中的蒙汉日常用语机器翻译评测项目的情况。本文采用基于 Transformer 神经网络的蒙汉机器翻译系统, 在此基础上使用基于词素的方法实现了蒙汉神经机器翻的主系统。对比系统分别是过滤掉蒙古文语料的控制字符预处理方法和基于 BERT(Bidirectional Encoder Representations from Transformers)中文语义相似度计算的数据增强方法。在 CCMT2019 提供的数据集上进行了对比实验, 实验结果表明对比系统相比于主系统，BLUE_SBP 值分别提高 3.02%和 3.85%。

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：TP391　　**文献标志码：** A

1 引言

“第十七届全国机器翻译大会(China Conference on Machine Translation, CCMT 2021)”, 包括机器翻译双语翻译（汉英、英汉、维汉、藏汉和蒙汉），多语言翻译（汉、日、英），低资源语言翻译（俄中、泰中、越中），自动译后编辑（汉英）和翻译质量自动评估（汉英、英汉）等多个任务。我们参加了蒙汉机器翻译任务，并提交了一个主系统和两个对比系统翻译结果。下面给出数据预处理，实验方法介绍，实验结果及分析。

2 数据预处理

本次评测中，将源语言端的传统蒙古文形式的数据用本实验室开发的转换工具全部转换成内大拉丁形式，转换过程中会把格附加成分前面的控制字符转换为“-”，把元音间隔符转换为“_”。在机器翻译任务中，对语料的切分是非常重要的一步预处理工作。本评测中我们采用了蒙古文部分切分方法，即对蒙古文词的格附加成分进行切分，并将词与标点符号全部进行分割。目标语言端对汉语进行全角转半角操作，利用中科院开源工具 ICTCLAS2019 进行分词处理。评测中训练集、开发集、测试集都做了相同的处理工作。

3 基于 transformer 神经网络的蒙汉机器翻译系统

目前, 基于神经网络的方法在机器翻译领域占据着主导地位, 并出现了多种不同类型的神经机器翻译方法, 例如, 基于循环神经网络[1]、基于注意力机制[2][3]、基于序列到序列[4][5][6]和基于 Transformer[7]等的神经机器翻译方法等。本次评测中，我们使用了基于 Transformer

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）；内蒙古师范大学研究生科研创新基金资助项目（GXJJS20129）。

通信作者：E-mail：siriguleng@imnu.edu.cn

的神经网络机器翻译方法。

3.1 基于 transformer 神经网络的蒙汉机器翻译系统

Transformer 神经网络是“Sequence to Sequence”模型框架，由编码器和解码器两部分组成。其中编码器有 6 层，解码器也有 6 层，如图 1 所示。在编码器部分，低层的编码器是表层的词法信息，逐步向上进行抽象之后，上层的编码器将表示抽象语义信息。最上层编码器部分到每个解码器部分的连接线表示 Attention 的部分，解码器和编码器也有信息传递和交互的。

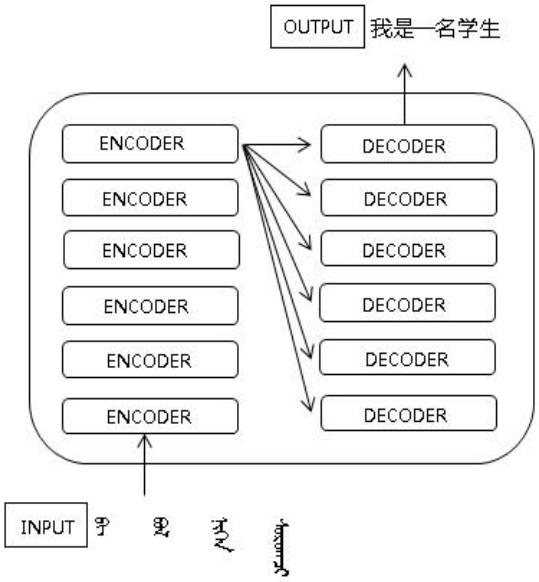

图 1 Transformer 蒙汉机器翻译模型实例

我们利用 Google 开源的基于 Transformer 神经网络模型搭建了蒙汉机器翻译平台，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 @ 3.70GHz；内存为 32GB，运行平台是谷歌公司开发的 GPU 版本的 Tensorflow。

3.2 基于词素的蒙汉神经机器翻译方法

在蒙汉机器翻译中一方面蒙汉双语语料库非常稀缺。另一方面蒙汉两种语言差异比较大，蒙古文是主宾谓结构，汉语是主谓宾结构，两者语序上存在很大的差异。在蒙汉机器翻译中，由于语法、句法结构等差异，导致源语言和目标语言语序完全不同，因此存在长距离调序问题。

另外蒙古文是黏着性语言，其形态丰富，存在词根相同的名词和动词的许多变形形式，表示相似的概念，而构形附加成分部分也具有丰富的词性和时态等信息，这些信息对于分析

机器翻译有重要的作用。但是，词级别机器翻译无法捕获这些隐藏的意义，同时，相同词根的不同变形被训练为不同的单元，从而增加开发神经机器翻译词汇词典的规模。因此针对蒙古文形态丰富的特征和神经机器翻译的有限词汇词典问题，我们开展了基于词素的蒙汉神经机器翻译研究。本文以此作为基线系统进行了蒙汉神经机器翻译实验。

3.3 过滤掉蒙古文语料的控制字符预处理方法

由于现行蒙古文 Unicode 编码中使用控制字符表达字的不同变形。因此，一个蒙古文的词中会出现蒙古文 Unicode 编码为“180E”、“202f”和“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，我们去掉了语料中出现次数较多的蒙古文 Unicode 编码为“180E” 、“202f”和“180D”的控制字符。

3.4 基于 BERT 数据增强的蒙汉神经机器翻译方法

现有的蒙汉训练语料规模较小，这种情况直接影响神经网络训练模型的结果。为了增强蒙汉神经机器翻译中的训练数据，本评测在去掉语料的控制字符预处理方法基础上采用基于 BERT 模型[8]去计算中文语义相似度扩充训练语料库的方法，有效扩充了蒙汉训练语料库。

3.4.1 BERT 中文语义相似度计算模型

为了解决中文译文和真实中文句子之间语义相似度计算问题，我们研究并且训练了 BERT 中文语义相似度计算模型。模型的框架如图 2 所示。

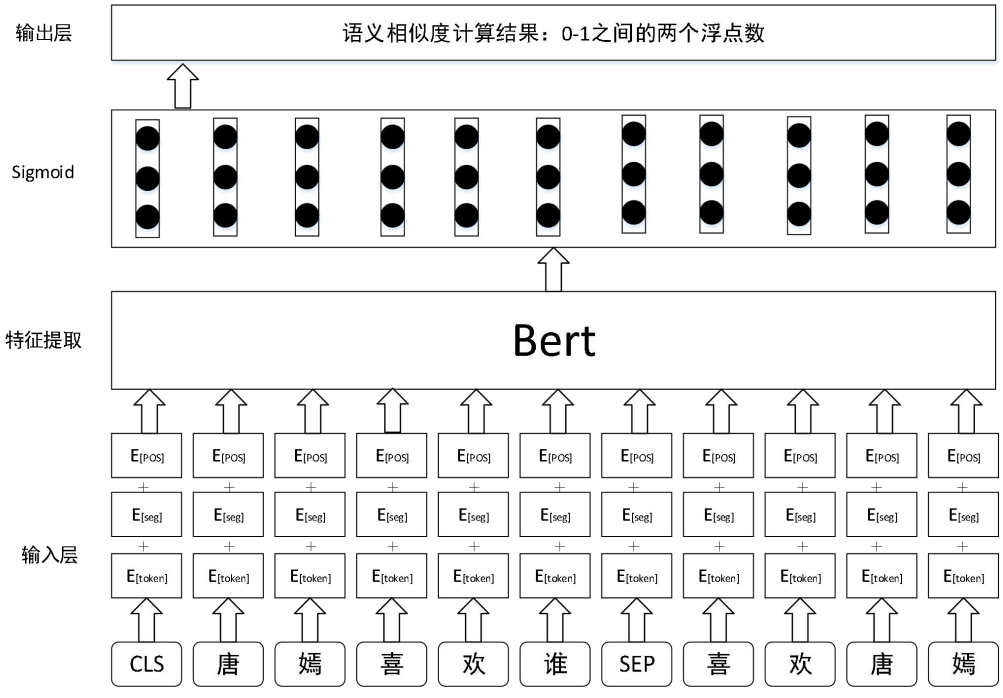

图 2 BERT 中文语义相似度计算模型框架

中文语义相似度训练模型时使用基于 LCQMC 的数据集，训练集为两个中文句子后面

带有 0 和 1 标签的数据。从图 2 可以看出，模型的输入是一对中文句子。模型首先在句子中自动添加两个符号，代表句子开头的[CLS]特殊符号和代表句子结尾的[SEP]特殊符号，然后通过词嵌入层得到该句对中每个字的字向量、代表位置的位置向量和区分两个句子的句子向量三者总和作为模型的输入。

获得输入表示后，使用 Transformer 编码器提取句子的特征，经过多层 Transformer 编码器逐层传递并细化特征表示。计算公式如式（1）所示：

$$R_l = Transformer_l(R_{l-1}) \tag{1}$$

式中，l 表示对应的层数，R_l 为经过对应层特征的上下文表示。每一层编码器结构相同，均包括多头自注意力层和前馈神经网络层并经过残差连接和层-归一化。

[CLS]符号与其他句子中的单词相比没有明确的语义信息，可以更准确地融合句子中每个单词的相关信息，因此，此符号相对应的输出向量用作整个句子的语义表示。最后模型的输出是包含全部语义信息的[CLS]向量，使用余弦距离算法计算语义相似度，然后传给 sigmoid 函数进行预测计算，输出结果是两个 0 到 1 之间的浮点数，靠近 1 表示两个句子语义相似度高，靠近 0 表示两个句子语义相似度低。余弦距离来源于向量之间夹角的余弦值。两个句子中，每个句子都有包含语义信息的[CLS]向量如式子（2）所示：

$$\begin{aligned} W &= (w_1, w_2, ..., w_n) \\ S &= (s_1, s_2, ..., s_n) \end{aligned} \tag{2}$$

那么二者的夹角余弦值等于：

$$cos = \frac{W*S^T}{|W|*|S|} = \frac{\sum_{i=1}^{n} w_i * s_i}{\sqrt[2]{\sum_{i=1}^{n} w_i * w_i} * \sqrt[2]{\sum_{i=1}^{n} s_i * s}} \tag{3}$$

其中 W 表示句子 A 的向量，S 表示句子 B 的向量。

3.4.2 实验数据及实验参数

数据来源为 LCQMC 带有 0 和 1 标签的数据。LCQMC 是哈尔滨工业大学为自然语言处理国际顶会 COLING2018 构建的语义匹配数据集，其目标是判断两个句子的语义是否相同，数据集包含训练集(23 万句对)、开发集(8 千句对)和测试集(1.2 万句对)。我们在实验中没

有使用数据集中的测试集，而是选用我们机器翻译的训练集，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 1 标签作为测试集。

实验参数设置：损失函数为交叉熵损失函数，Batch_size 为 32；Dropout 为 0.1，使用 Adam 算法进行优化，激活函数使用 RELU；迭代训练 22384 轮，每 1000 步保存一个模型，最后在 checkpoint 中保存最近的 5 个语义相似度计算模型。

3.4.3 蒙汉训练语料的扩充

基于 BERT 的中文语义相似度计算模型实验中，我们首先利用 CCMT2021 提供的训练集训练了蒙汉初始翻译模型，使用该模型翻译训练集中蒙古文获得中文译文，然后将中文译文和真实中文句子组成句对送到 BERT 中文语义相似度计算模型进行计算，最后得到两个概率值相加为 1 的浮点数。通过实验结果得出，翻译结果中共有 26636 条与原训练集不同的句子。得到句子的语义相似度计算结果后，选择大于等于 0.5 的概率值的句子，将筛选出来的 12092 条句子与对应的蒙古文扩充到机器翻译训练语料中进行实验。实验流程如图 3 所示。

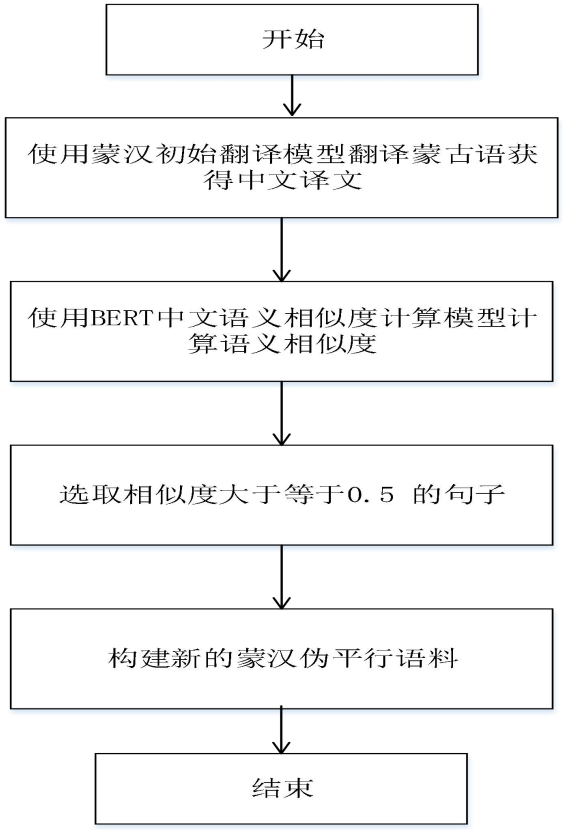

图 3 蒙汉训练语料库扩充流程图

4 评测

4.1 评测数据

本次机器翻译评测以及本文实验均采用了 CCMT2021 提供的训练集（26 万句对），开

发集（1001 句对）。利用 CCMT2019 提供的 1000 个句子的蒙古文语料库和对应的 4 个汉语参考答案作为测试集进行对模型的优化，最终对 CCMT2021 提供的 5970 个句子作为测试集进行了翻译实验。数据预处理方面，训练集、开发集和测试集做了相同的处理，包括：对中文进行全角转半角操作和采用中科院开源的 NLPIR(ICTCLAS2019)分词系统进行分词；对蒙古文进行传统蒙古文到内大拉丁形式的转写、部分切分，去掉控制字符等工作。

4.2 参数设置

Transformer 全连接层隐藏单元个数为 512；学习率设置为 0.001， Dropout 设置为 0.3，使用 Adam 算法进行优化，batch size 为 2048。迭代训练 250000 轮，每 1000 步保存一个翻译模型，直到完成 250000 轮的迭代翻译训练之后，在 checkpoint 中保存最近的 20 个翻译模型。完成 250000 轮的迭代翻译训练大约需要 12 小时的时间。

4.3 评测结果与分析

首先进行基于词，BPE 切分和词素切分方法在基于 Transformer 的神经网络模型框架下译文质量。通过实验结果，主系统采用基于词素的蒙汉神经机器翻译系统。对比系统 1 是过滤掉蒙古文语料中控制字符方法，对比系统 2 是利用 BERT 模型计算中文语义相似度进行扩充方法。我们在 CCMT2019 提供的 1000 个句子测试集上做了实验及结果分析。机器翻译评测指标选用了 BLUE5[8]、BLUE_SBP5[9]和 NIST5[10]值。

4.3.1 基于词的模型

基于词的蒙汉神经机器翻译模型对语料未进行任何切分处理，分写的构形附加成分跟着前面一个词。

词模型的语料形式如下实例所示：

VLVS-VN HUMUSVN AJILCIN-V HAVLI .

分写附加成分“VN”和“V”与前面的词用“-”符号连接着，看成一个词。

4.3.2 BPE 切分模型

BPE 编码技术是在 2016 年，Sennrich 等人[11]提出的一种处理文本切分粒度的方法，它有独立 BPE 和联合 BPE 两种应用方法。本实验中切分机器翻译语料时选用了独立 BPE。实验设置中将 BPE 操作数分别设置为 30000、35000、40000、45000、50000，将蒙古文词切分成子词粒度。最终得出，BPE 操作数 40000 的时候机器翻译指标 BLEU5 值最高，达到 68.74%。

BPE 切分模型的语料形式如下实例所示：

VLVS VN HUMU@@ SVN AJILCIN V HAVLI .

蒙古文单词“HUMUSVN”被拆分成了“HUMU@@”和“SVN”，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，并不符合蒙古文的语法规则。

4.3.3 词素切分模型

对蒙古文端进行词素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，还有部分名词的复数附加成分。通过词素切分可以有效减少单频词的数量，从而可以缓解数据稀疏问题。

词素切分模型的语料形式如下实例所示：

VLVS VN HUMUSVN AJILCIN V HAVLI .

分写附加成分“VN”和“V”与前面的词之间的连接符号“-”替换成空格，把分写的构形附加成分看成一个单独的词。

基于词的机器翻译系统，BPE 词切分机器翻译系统，主系统和对比系统在 CCMT2019 测试集上的评测结果，如表 1 所示。

表 1 1000 句子评测结果

实验	BLUE_SBP	BLUE	NIST
基于词的模型	0.6566	0.6812	11.0673
BPE 切分模型	0.6668	0.6874	10.8089
主系统	0.6721	0.6987	11.2729
对比系统 1	0.7023	0.7200	11.5023
对比系统 2	0.7408	0.7586	11.8261

从表 1 可以看出，使用基于词素切分的方法相比于基于词的方法，机器翻译结果 BLUE_SBP 值提高 1.55%，BLUE 值提高 1.75%。相比于 BPE（Byte Pair Encoding）切分方法，机器翻译结果 BLUE_SBP 值提高 0.53%，BLUE 值提高 1.13%。所以本评测中选用基于词素的蒙汉神经机器翻译系统为主系统。

我们的对比系统实验结果与主系统实验结果相比都有所提升，对比系统 1 的过滤掉蒙古文语料的控制字符预处理方法跟主系统的基于词素的基线实验相比，BLUE_SBP 值提高 3.02%，BLUE 值提高 2.13%。在对比系统 1 的基础上，通过 BERT 模型去计算中文语义相似度扩充训练语料会使得 BLUE_SBP 值进一步提高 3.85%，BLUE 值提高 3.86%。

5 总结

本文介绍了内蒙古师范大学计算机科学技术学院蒙古文大数据研究基地在 CCMT2021 蒙汉日常用语翻译评测的基于 Transformer 的蒙汉神经机器翻译系统搭建情况，且使用蒙古文拉丁转写、蒙古文部分切分、过滤掉蒙古文控制字符、基于 BERT 中文语义相似度去扩充蒙汉训练语料等多种方法。

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、无监督学习等方法进行研究。

参考文献

[1] Kalchbrenner N, Blunsom P. Recurrent Continuous Translation Models. [C]. EMNLP, 2013.

[2] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]. ICLR, 2015.

[3]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]. 内蒙古大学, 2017.

[4] Sutskever I, Vinyals O, Le Q V, et al. Sequence to Sequence Learning with Neural Networks[C]. NIPS, 2014.

[5] Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]. EMNLP, 2014.

[6] Siriguleng Wang,Wuyuntana.The Research on Morpheme-Based Mongolian-Chinese Neural Machine Translation[J].2019 IEEE 2nd International Conference on Information and Computer Technologies. 2019:138-142.

[7] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv, 2018.

[8] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. 311-318.

[9] Chiang D , Deneefe S , Chan Y , et al. Decomposability of translation metrics for improved evaluation and efficient algorithms[C]// 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008.

[10] Doddington G. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics[C]// Proceedings of the HLT Conference, San Diego, California, 2002.p138-145.

[11] Sennrich R， Haddow B， Birch A . Neural Machine Translation of Rare Words with Subword Units[J]. Computer Science, 2015.

IMNU CCMT2021 Mongolian and Chinese machine translation system evaluation technical report

He Wuyun, Bao Jingjing, Sachula, Chen Meilan, Terigelehu, Wang Siriguleng*

(Inner Mongolia Normal University, Inner Mongolia Autonomous Region Huhehaote 011500, China)

Abstract: This paper presents the participation of the Mongolian Big Data Research Base of the School of Computer Science and Technology of Inner Mongolia Normal University in the machine translation evaluation project of CCMT2021. This paper adopts a Mongolian-Chinese machine translation system based on Transformer neural network, on which the main system of Mongolian-Chinese neural machine translation is implemented using a lexical element-based approach. The contrasting systems are a control character pre-processing method that filters out the Mongolian corpus and a data enhancement method based on the BERT (Bidirectional Encoder Representations from Transformers) Chinese semantic similarity calculation, respectively. The comparison experiments were conducted on the dataset provided by CCMT2019, and the experimental results showed that the BLUE_SBP values of the comparison system were improved by 3.02% and 3.85% respectively compared with the main system.

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