昆明理工大学 CCMT2021 评测报告

李佳佳 [1,2]，王家琪 [1,2]，朱俊国 [*1,2]，余正涛 [1,2]

（1. 昆明理工大学信息工程与自动化学院，云南省昆明市，650500；

2. 昆明理工大学云南省人工智能重点实验室，云南省昆明市，650500）

摘要：本文介绍了昆明理工大学云南省人工智能重点实验室参加 2021 年全国机器翻译评测（CCMT2021）的评测系统。本次评测我们提交了两个翻译任务，分别是蒙-汉和汉-英翻译任务。本文主要介绍了我们参加该评测任务所用到的 Transformer 模型以及模型在评测数据上的性能表现。

关键字：机器翻译，神经网络，评测

1 引言

昆明理工大学云南省人工智能重点实验室多年来一直积极地开展低资源机器翻译的相关研究工作。在 CCMT2021 的机器翻译评测中，我们提交了两个翻译任务，分别是蒙-汉机器翻译的评测任务和汉-英机器翻译的评测任务。针对这两个任务，我们提交了两组翻译结果。其中一组是蒙-汉翻译结果，另一组是汉-英翻译结果。其中，蒙-汉的翻译结果又包含两组结果，蒙-汉翻译任务的第一组翻译结果是利用真实平行数据进行蒙-中模型训练，第二组翻译结果是先利用真实平行数据训练中-蒙模型，再将处理过的汉语单语语料放进中-蒙模型生成蒙语，与之前的处理好的汉语单语语料构成伪平行语料，最后利用伪平行语料训练蒙-中模型。汉-英的翻译结果只有一组，这组翻译结果是利用真实平行数据训练出汉-英模型。最终，蒙-汉翻译任务第一组数据在验证集上达到 28.9 的 bleu 分数，在测试集上达到 26.1 的 bleu 分数，第一组的在翻译任务上的质量明显比第二组好。而汉-英翻译任务最终翻译结果在测试集上达到 17.34 的 bleu 分数。本报告将主要介绍我们的翻译系统的具体方法及实现。

2 模型概述

本次参加的机器翻译评测任务，我们采用的是基于 Transformer[1]的序列到序列[2]的神经机器翻译模型。Transformer 模型是由 Google 于 2017 年在 Attention is All You Need 一文中提出。近年来，注意力机制在序列建模中被广泛使用，用于解决 RNN 在建模长序列时

* 通讯作者：zhujunguohit@gmail.com

资助项目：云南省科技厅基础研究计划面上项目(202001AT070167)

对于较早编码的部分依赖不足的问题。Transformer 摒弃了使用循环递归结构来编码序列的基本范式，而是基于全局的注意力机制来计算序列的隐状态。因而比 RNN 能够更好地建模长序列中的依赖关系。从运行效率来看，RNN 模型需要按照序列既定的顺序依次计算每个位置的隐状态，而 Transformer 模型所采用注意力机制在训练阶段能够并行地计算整个序列的隐状态，具有更高的并行度。

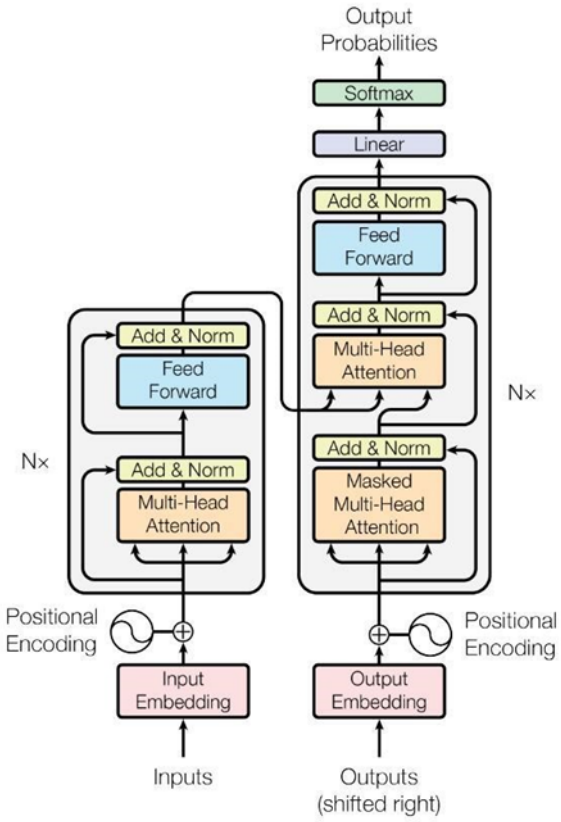

图 1 Transformer 模型结构

Fig. 1 Model Structure of Transformer

如图，模型由编码器和解码器以及注意力机制组成。编码器由 N 层组成，其中 N=6。每一层编码器由多头注意力机制、残差连接和归一化操作组成。其中自注意力机制可由以下公式表示：

$$Attention(Q,K,V) = Vsoft\max(\frac{K^T Q}{\sqrt{d}}) \qquad (1)$$

解码器与编码器类似，但解码器比编码器多了一个自注意力的子层。解码器的输入是前一个位置的解码器的输出和额外来自编码器的输出 H^{Le} 作为注意力机制的 K 和 V。Transformer 解码器的输出为序列对应的隐状态矩阵 $S^{Ld} \in R^{d*(M+1)}$，其中 L_d 为解码器的网络层数，最后经过线性映射和归一化得到每一个位置在词表 V 上的分布。

3 数据处理

3.1 蒙-汉数据处理

蒙汉模型的训练使用了评测官方提供的受限蒙-汉双语训练集，原始训练数据包括大约 23 万句。我们对平行语料进行了预处理。首先将语料的标签去掉并将语料中的字母和数字的全角全部转化成半角形式。然后对语料进行分词操作，蒙语部分使用了 moses 脚本的 tokenizer 来进行分词，中文部分使用了 jieba 分词进行精确分词。对于分词之后的语料，我们用了 moses 脚本的 clean 进行过滤。过滤之后剩余大概 22 万句对，挑选的是长度在 1-250 之间以及长度比小于 3 的句对。对于过滤后的句子，采用了 BPE 的方法分别将蒙语和汉语句子切分成了子词模型，BPE 处理是联合源语言和目标语言进行的，而不是分别对源语言和目标语言进行切分。同时，我们只使用了训练的平行语料作为 BPE 的词频统计语料，因此，验证集和测试集都是根据训练的蒙汉平行语料学习到的模型进行切分。

对于单语数据，首先进行去标签以及全角转半角、排序去重复和 jieba 分词操作，然后挑选比例大于 0.9 的单语句子（该比例是指单语句子单词在训练平行数据词典的比例）。其余步骤与处理训练的平行句对相似。

第一组用真实平行数据训练的模型的验证集是采用的官方提供的约 1000 个句对，而第二组用伪平行数据训练的模型的验证集是从伪平行数据中抽取的。

3.2 汉-英数据处理

汉-英平行数据采用的是 NEU2017 语料库，NEU2017 语料库是由中国东北大学的 NLP 实验室提供的。该语料库包含从网络上自动收集的 200 万个句子对，包括新闻、技术文档等。句子层面的对齐精度约为 90%。

由于英语中的标点符号和单词是连在一起的，所以如果只对单词进行分词，book 和 book, 就会占据词典中的两个位置，这是不合理的；对于句子中的英文单词，还存在同一个单词不同大小写的问题，也需要学习最适合它们的大小写形式。我们采用了 mosesdecoder 来对数据进行了以上的预处理。

首先对中英语料库中的标点符号进行标准化，其中还需要对中文进行分词处理，因为中文词语不像英文单词那样天然具有明显的单词边界。有很多比较有名的中文分词工具，像 jieba、pkuseg[3]等等，本文采用了 jieba 来进行分词处理。然后对进行了上述处理的双语文件进行标记化处理，这一步可以将英文单词与标点符号用空格分开，还可以将多个连续空格简化为一个空格，同时还能将很多符号替换为转义字符，如：把”替换成为"、把 don’t 替换成为 don 't 等。

对于英文语料，还需要进行大小写转换处理，尤其是句首单词，在数据中学习最适合它们的大小写形式。

然后对双语语料进行子词处理，会用到 BPE 算法。使用处理后的子词作为基本单位训练神经机器翻译模型。BPE 算法会将训练语料以字符为单位进行拆分，按照字符对进行组合，并对所有组合的结果根据出现的频率进行排序，出现的频率越高排名越靠前，排在最前面的便是出现频率最高的子词。

最后对经过处理的双语语料进行过滤，可以过滤最小长度和最大长度之间的句对，也可以过滤长度比不合理的句对，这同时也可以有效过滤掉空白行。

最后将双语语料按 98：1：1 的比例划分为训练集、验证集和测试集。

4 实验

4.1 实验环境

本次测评使用的服务器配置如表 1 所示：

表 1 服务器配置

GPU	NVIDIA GTX 3090
CPU	Gen Intel(R) Core(TM) i9-11900F @ 2.50GHz
内存	64G
操作系统	Ubuntu 20.04.2 LTS

4.2 实验数据

本次评测的实验使用的实验数据都是由官方提供的受限数据，具体信息如表 2 所示：

表 2 实验数据

语种对	训练集	验证集	测试集
蒙汉	222981	1000	1000
汉英	约 200 万	20000	20000
汉语单语	约 700 万		

其中汉语单语数据约 700 万，在经过去标签以及筛选比例大于 0.9 的句子之后仅剩约 100 万。

4.3 参数设置

本次评测我们用到的是 Facebook 实验室的开源代码 fairseq[4]来进行处理，后续也是用 fairseq 来进行训练和解码。通过双语数据我们分别构建了两种语言的词汇列表，并用 fairseq 内置的 Transformer 架构来训练我们的模型。

对于蒙-汉任务，Transformer 采用 6 层编码器和解码器，每个 batch 的最大 token 设置为 4096，学习率设置为 0. 0007，dropout 设置为 0. 3，activation−dropout 以及 attention−dropout 被设置为 0，优化器用的是 adam 优化器。解码阶段，beam size 设置为 5。第一组仅用真实

平行数据训练的模型在验证集上达到 28.9 的分数，第二组用部分伪平行数据训练的模型仅达到 18.7 的分数。

对于汉-英任务，Transformer 采用 6 层编码器和解码器，其余参数设置是学习率为 0.0007，warmup updates 为 16000，dropout 为 0.3，采用 adam 优化器来优化我们的模型。汉-英模型在我们的测试集上翻译评测最终的 bleu 分数为 17.34 分。

5 总结

本文主要介绍了昆明理工大学云南省人工智能重点实验室参加 CCMT2021 评测的情况，用 mosesdecoder 和 jieba 工具对数据进行预处理后，使用开源工具 fairseq，应用基于自注意力机制制的 Transformer 架构来训练我们的模型。其中蒙-汉翻译任务通过生成伪平行数据来提高训练数据的质量和规模，最终翻译质量未有提升。其原因是计算资源有限，我们没能尝试更多参数。今后，我们将深入研究利用伪平行数据训练模型，尝试更多参数，使得翻译任务质量得到提升。

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar,et al. 2017. Attention Is All You Need. In Proc. of NIPS.

[2] Myle Ott and Sergey Edunov and Alexei Baevski et al. 2019. fairseq: AFast, Extensible Toolkit for Sequence Modeling.Proceedings of NAACL-HLT 2019: Demonstrations.

[3] Luo, R., Xu, J., Zhang, Y., Ren, X., & Sun, X. (2019). PKUSEG: A Toolkit for multi-domain Chinese word segmentation. arXiv preprint arXiv:1906.11455.

[4] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., ... & Auli, M. (2019, June). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (pp. 48-53).

KUST Technical Report of Machine Translation for the 2021 China Conference on Machine Translation

Jiajia Li[1,2], Jiaqi Wang[1,2], Junguo Zhu*[1,2], Zhengtao Yu[1,2]

(1.School of Information Engineering and Automation, Kunming University of science and technology, Kunming,650500, China ;

2. Yunnan Key Laboratory of Artificial Intelligence, Kunming, 650500, China)

Abstract: This report introduces the machine translation system for the 2021 China Conference on Machine Translation (CCMT2021), which submitted by the Yunnan Key Laboratory of Artificial Intelligence of Kunming University of Science and Technology. In this task, we submitted two

translation tasks: Mongolian-Chinese and Chinese-English translation task. The paper mainly introduces the Transformer model and the performance in the evaluation data set of the machine translation system.
Keywords： machine translation, neural network, evaluation