新疆大学信息学院 CCMT2022 维-汉机器翻译系统评测技术报告

买日旦·吾守尔[1]，斯拉吉丁[2]，艾斯卡尔·艾木都拉[1]

[1]新疆大学信息科学与工程学院 [2]新疆三剑客信息科技有限公司

摘要：本文主要介绍新疆大学信息学院研究团队参与第十八届全国机器翻译大会机器翻译测评中的维汉翻译测评基本情况。在此次评测中，本研究团队参与了维-汉机器翻译任务，并在该任务上提交了评测系统。本技术报告主要介绍参赛的维汉神经网络机器翻译系统选用的不同模型对比分析、不同粒度选择、数据增强使用方法以及该系统在开发集上的性能。

关键词：CCMT2022 维-汉；神经机器翻译；音节；模型微调

中图分类号：TP391　　文献标识码：A

Abstract: This document details about the Uyghur-Chinese machine translation model developed by the NLP research team from the Xinjiang University as a participant to the CCMT2022 evaluation task. In this evaluation, the research team accepted the task of Uyghur-Chinese machine translation and submitted the translation result to the evaluation system. This document details the comparative analysis of different models selected by the varies approaches such as the selection of different granularities, the use of data enhancement methods, and the performance of the system on the development set.

Key Words: Uyghur-Chinese; NMT; Phoneme; Finetuning

1. 引言

第十八届全国机器翻译大会(CCMT 2022)由中国中文信息学会主办，其主要目的在于促进国内外科研单位、产业界相关单位之间的学术交流和联系，共同推动机器翻译研究和技术的发展。本次机器翻译大会的机器翻译测评主要包括：延续在线翻译评测，汉英、英汉新闻领域的翻译评测，维汉、蒙汉、藏汉的翻译评测，专利领域的日、汉、英多语言翻译评测，翻译质量估计评测以及新增评测任务（自动译后编辑任务和低资源机器翻译评测项目）。

在本届 CCMT2022 机器翻译测评中，新疆大学研究团队参与维汉机器翻译评测任务，并在该任务上提交了评测系统。本技术报告主要介绍参赛的维汉神经

网络机器翻译系统选用的开源平台对比分析、亚词单元选择、数据增强使用方法、知识蒸馏等及该系统在开发集上的性能。

2. 评测系统

近年来，随着深度学习方法的迅速发展，神经机器翻译在机器翻译领域取得了巨大的成功。相对于传统的统计机器翻译而言，基于深度学习的神经机器翻译在译文质量上更加准确，尤其是 2017 年提出的 Transformer[1]翻译模型为神经机器翻译带来了新的活力，使得机器翻译领域中模型性能与翻译质量有了大幅的提升。基于 Transformer 的翻译模型主要特点在于既不依赖于 RNN[3]，也不依赖于 CNN [错误!未找到引用源。]，而仅通过自注意力机制计算输入序列和输出序列的表示，实现端到端的神经机器翻译框架。Transformer 模型在机器翻译任务上的表现超过了 RNN 和 CNN，只用 encoder-decoder 和 attention 机制就能达到很好的效果，其最大的优点是可以高效地并行化。鉴于其优秀性能，到目前为止大部分工作都是在 Transformer 上进行的，其优越性也得到了广泛认可，无论是从 BLEU[4]值还是语言的流利度都提升了很多，因此已成为机器翻译领域的主流模型。

2.1 Transformer 模型

Transformer 模型由编码器和解码器两部分组成。编码器由一个 Multi-Head 网络和一个简单的全连接前馈神经网络组成，在这两个网络中间添加了一个残差连接，并进行层标准化操作。而解码器由一个 Masked Multi-Head Attention 网络、Multi-Head Attention 网络以及一个全连接前馈网络组成，同样也用了残差连接及层标准化操作。编码器用于将输入语料转化成特征向量，解码器输入为编码器的输出以及已经预测的结果，用于输出最后结果的条件概率。

编码器的输入先经过第一个 self-attention 层被编码成包含位置信息的向量。位置编码器负责捕获位置信息，首先输入向量上创建查询(Query，Q)和键值(Key-Value，KV)对向量，其次通过缩放点积注意力机制训练该三个向量并用 Softmax 对训练结果归一化，最后乘以值(V)获取注意力向量。计算数学公式如式（1）:

$$Attention(Q,K,V) = softmax(\frac{Qk^T}{\sqrt{d_k}})V \quad (1)$$

其中，$\sqrt{d_k}$是缩放因子。在常见的加法 Attention 和点积 Attention 中，当$\sqrt{d_k}$的值比较小的时候，这两个机制的性能相差相近，当$\sqrt{d_k}$的值比较大的时候，加法 Attention 比不带缩放的点积 Attention 性能更好。

多头注意力机制（Multi-head Attention）是 Transformer 模型中另一个重要的注意力机制，由多个缩放的点乘注意力(scaled dot-product attention)经过堆叠而成其使模型捕获不同子空间的信息。多头注意力机制把 Query 和 Key 映射到高维空间的不同子空间中去计算相似度，且其输入都是 Q、K、V 三个元素，只是取值不同。原来的输入长度除于头数 h 将输入向量分为 h 个子向量，经过不同的线性层后拼接输出得到原向量长度的矩阵，计算公式如下(2)和(3)所示：

$$MultiHead(Q,K,V) = Concat(head_1,\cdots,head_n)W^o \tag{2}$$

$$head_i = Attention（QW_i^Q, KW_I^K, VW_I^V） \tag{3}$$

其中，$W_i^Q \in R^{d_{model}\times d_k}$，$W_i^K \in R^{d_{model}\times d_k}$，$W_i^V \in R^{d_{model}\times d_u}$，$W_i^O \in R^{hd_u\times d_{model}}$

d_k=$d_u = d_{model} \div h = 64$

在 Decoder 中，与 Encoder 的不同之处在于 Decoder 多了一个 Encoder-Decoder 注意力模块，与 Self-Attention 分别用于计算源语言和目标语的权值。

2.2 Dynamic Convolutional 模型

通过对 RNN、CNN 和 self-attention 模型的使用，在序列建模方面已有许多进展。RNN 通过在每个时刻更新隐藏状态来汇集上下文信息，CNN 通过进行多层固定大小的卷积来汇总上下文信息，而 self-attention 则直接在每一层汇总所有上下文信息。self-attention 被定义为基于内容的表示，其中通过将当前时刻与上下文中的所有元素进行比较来计算注意力权重。在这种不受限制的上下文大小上计算比较的能力被视为 self-attention 的核心功能。然而，self-attention 在较长的上下文在计算上非常具有挑战性，此外，在实践中长序列需要引入层次结构。

针对上述问题，引入了具有深度可分离的、softmax 归一化并在通道维度上共享权重的轻量级卷积，其卷积的权重比标准不可分离卷积少几个数量级。与 self-attention 不同，轻量级卷积不管上下文长度如何都为上下文元素重用相同的权重。但是，轻量级卷积神经网络的计算预算较低，不仅限制了其卷积层的深度，还限制了其信道的宽度，从而导致其性能下降。为了解决这个问题，Chen 等人提出了动态卷积（Dynamic Convolution），这是一种在不增加网络深度或宽度的情况下增加模型复杂性的新设计。动态卷积不是每层只使用一个卷积核，而是根据注意力来动态聚合多个与输入相关的并行卷积核。

动态卷积基于轻量级卷积构建，通过在每个时刻预测不同的卷积核，卷积核仅是当前时刻的函数，而不是整个环境。动态卷积在每个位置上权重都发生变化，这些卷积核随各个时刻的学习函数而变化，且是由模型动态生成的，而不是经过训练后固定的。在大规模机器翻译任务中，在翻译质量上 Dynamic Convolutional 模型比强大的 self-attention 模型有所改进。

2.3 维吾尔语音节特点

维吾尔语是一种典型的黏着语，具有丰富的形态变化。维吾尔语中有 24 个辅音字母和 8 个元音，共有 32 个字母，包含，同时每个字母具有不同的形式，共计约有 130 种。在维吾尔语中，一个或多个单词组成句子，句子中以空格分开每个单词，其中单词由一个或多个音节组成。音节是表音语系中最小的语音结构，由单个元音音素和辅音音素组合发音的语音单位。在维吾尔语中，单词由一个或多个音节组成，而构成单词的音节具有一定的语义信息。维吾尔语的音节切分有一定的规则，音节结构固有（起音）＋领音＋（收音），其中音节中必须要有领音且必须是元音，而在起音和收音中可以有也可以没有。

2.4 带标记音节的机器翻译

维吾尔语语法和形态的复杂性、维汉平行语料的匮乏以及数据稀疏问题使维-汉机器翻译研究进展相对缓慢。将维吾尔语数据切分成具有一定语义信息的音节，汉语数据划分为单个字符，这样可以使翻译单元数量减少，出现频率增加。每一个翻译单元出现频率的增加使得网络模型学习能力增强。而翻译单元数量减少，不仅能缩小词表规模，降低模型的复杂运算，缩短模型的训练时间。同时还能有效地解决集外词（OOV）问题，缓解维汉神经机器翻译中的数据稀缺问题，使得翻译质量得到提升。

首先，通过维吾尔语分词与编码转换工具来对维吾尔语语料进行分词并编码转换，使用繁体简体及全角半角转换工具对汉语语料进行处理，使得语料进行统一编码。然后，通过基于规则的维吾尔语音节切分工具来对维吾尔语语料进行音节切分，对汉语语料进行字符级切分，同时对维汉语料进行 BME 标记。这种切分方法把数据切分成更小的单位，使词表规模更小。最后，将切分好的维吾尔语音节向量化以后作为神经机器翻译模型的输入单元，把汉字向量作为模型的输出单元，训练一种基于音节粒度神经机器翻译模型。

3. 数据及数据处理

3.1 数据来源介绍

本次评测模型训练所采用的语料来自于 CCMT2022 提供的 17 万维汉双语平行语料与 800 万中文单语数据。

3.2 数据预处理

语料数据预处理是机器翻译任务中的关键一步，需要细致，有效地处理策略和方法。在本次的评测系统数据预处理包括编码转换、全角半角转换、乱码过滤、分词、BPE 切分以及音节提取等方法。

首先，利用编码转换工具分别对维-汉料进行编码转换，包括基本扩展区转换、全角半角转换、繁体简体转换、乱码过滤与去重；其次，利用开源的清华大学中文 NLP 工具 THULAC 对中文语料进行分词处理，同时利用新疆大学多语种实验室小组研发的维语分词工具对维文语料进行分词处理；然后，利用 subword-nmt 开源工具对中文语料进行 BPE 切分处理；最后，利用新疆大学多语种实验室小组研发的音节切分工具对维语料音节拆分处理。

4. 实验与结果

在本次测评所有的模型采用 Face Book(Meta) AI 开源的 FairSeq 框架 Pytorch 版本训练。我们分别尝试了 Transformer 与 Dynamic Convolution 模型。除此之外，本次我们还使用了模型平均，反向翻译，模型微调，模型集成解码等技术。部分实验结果如表 4-1 所示。

表 4-1 实验结果

序号	系统	BLEU
1	Base Line（Transformer + Syllable）	39.32
2	Base Line（Dynamic Convo + Syllable）	39.01
3	+Back Translation	42.44
4	+Fine Tuning（Org+KD）	43.30
5	+Fine Tuning（Org）	44.01
6	+Avg	44.23
7	+Ensemble	45.52

从上述实验结果可以看出，模型平均，模型微调，反向翻译，模型集成等策略均能对翻译效果有所提升，其中加入不同粒度生成反向翻译数据以后效果提升最大，说明加入反向翻译语料可以使得模型学到更多的知识。在最后采用

了模型平均与集成解码以后，模型效果也有明显的提升，将近有 1.5 个 BLEU 点。

5. 总结与展望

在本次维汉机器翻译评测系统的研究中，对机器翻译系统性能有影响的开源系统、粒度选择、反向翻译与数据评价和微调等方面进行了比较详细的对比，同时使用了模型平均和集成等相关技术手段，实验证明这些方法都能够有效的提升评测系统的性能。

由于时间和计算资源受限，同时受疫情影响，还有许多方法没来得及尝试。通过本次评测，我们发现了自己的不足和问题，模型和系统仍存在很大提升空间。在后续的研究工作中我们继续学习各方先进技术，对翻译模型速度,翻译质量上面继续探索并改进。

参考文献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30, pp. 5998-6008.

[2] [1] Wu F， Fan A， Baevski A , et al. Pay Less Attention with Lightweight and Dynamic Convolutions[C]// 2019.

[3] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. 2014. p. 3104-3112.

[4] Gehring J, Auli M, Grangier D, et al. A convolutional encoder model for neural machine translation[J]. arXiv preprint arXiv:1611.02344, 2016.

[5] 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.

[6] 瓦依提·阿不力孜,加米拉·吾守尔,吐尔根·依布拉音,阿依佐克拉·瓦依提.现代维吾尔文音节自动切分方法及其实现[J].中国科技论文,2015,10(08):957-961.

[7] 艾山·吾买尔,斯拉吉艾合麦提·如则麦麦提,西热艾力·海热拉,刘文其,吐尔根·依布拉音,汪烈军,瓦依提·阿不力孜.带标记音节的双向维汉神经机器翻译方法[J].计算机工程与应用,2021,57(04):161-168.