7 新疆大学 CCMT2022 英汉机器翻译评测任务技术报告

宜年[1]，艾山·吾买尔*[1]，汪烈军[1]

(1.新疆大学信息科学与工程学院，新疆 乌鲁木齐 830046)

摘要:本文主要介绍新疆大学信息科学与工程学院在第十八届全国机器翻译大会机器翻译评测项目中参赛的基本情况。在本次机器翻译评测中，参加了英汉新闻领域机器翻译评测项目。本文主要阐述本次参赛的英汉神经机器翻译系统采用的模型框架、数据预处理过程、数据增强以及模型微调和集成等方法。最后给出不同方法和模型在测试集上的性能表现，并进行对比和分析。

关键词: 英汉神经机器翻译；自注意力机制；模型微调；模型集成

中图分类号： TP302.1 **文献标志码：**A

1. 引言

本文介绍了新疆大学新疆多语种信息技术重点实验室所参加第十八届全国机器翻译大会（CCMT 2022）（China Conference on Machine Translation，简称 CCMT）的英汉机器翻译技术评测的主要情况。本次评测采用的模型结构为 Google 提出的 Transformer 神经机器翻译模型[1]和基于动态卷积的神经机器翻译模型[2]。在英汉新闻领域机器翻译任务上，使用的数据为 CCMT2022 和 WMT2022 提供的英汉平行数据，以及 CCMT2022 提供的汉语单语数据。在数据预处理部分，考虑到汉英数据量较大的原因，本文仅仅采用了传统的句子长度和长度比方法对平行数据进行过滤。除此之外，并没有采用其他的过于复杂的筛选方法。同时，为了提高模型的性能，本次测评采用回译[3]、模型平均、微调[4]和集成[5]等方式去提升模型的翻译性能。

在本次 CCMT2022 双语翻译评测任务中，新疆大学自然语言处理团队提交了英汉新闻领域机器翻译的评测系统。该测试报告中详细描述了新疆大学维汉机器翻译系统的网络架构、数据预处理、单语数据的使用、微调策略、模型集成等相关技术，并进行对比分析。

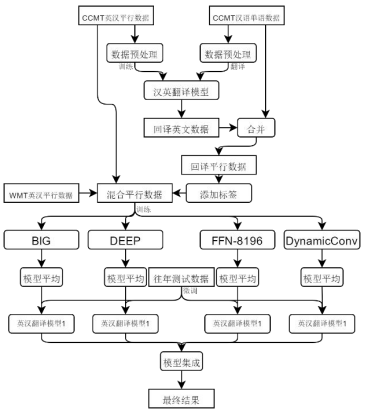

图 1 系统训练流程

Fig.1 Process of system training

基金项目：国家重点研发计划项目（2018YFB1403202）；国家自然科学基金资助项目(62166044)；

* **通讯作者:** hasan1479@xju.edu.cn；

2. 系统介绍

本次测评所使用的系统的整体训练训练流程如图 1 所示。首先对所使用的平行数据和单语数据进行数据预处理，之后会使用 CCMT 平行数据训练的汉英模型翻译汉语单语句子已得到与汉语单语句子对应的英文句子。英文句子与其对应的汉语单语句子会组成回译平行数据。之后会对回译平行数据进行长度比筛选并添加标签，图一中省略了该操作。之后将 CCMT 平行数据、回译平行数据以及 WMT 平行数据进行混合训练四种结构不同的模型。最终通过模型平均、微调和模型集成来得到最终的译文。下面将从数据预处理、模型结构、模型微调、平行和集成等方面进行详细介绍。

2.1 数据预处理

对于本次英汉翻译任务所使用的英汉平行数据来源于 CCMT 和 WMT，本文对平行数据进行了一下的预处理操作：

1. 全角转半角、转义、控制等特殊字符处理；
2. 分词：使用 spacy 工具[2]对英文进行分词，汉语分词则为哈工大分词系统[3]；
3. 删除重复的语对；
4. 根据 Wu 等[6]中预处理时筛选英汉平行数据所采用的句子长度，本文删除了句子长度小于 3 且大于 150 的句对；
5. 同样根据 Wu 等[6]中预处理时采用的比例，本文删除了平行数据中长度比大于 1.3 小于 0.76 的句对；

在本次测评中，本文在使用回译方法[3]时，仅仅使用了 CCMT 提供的单语数据，并对单语数据进行了同平行数据预处理相同的 1 到 4 步处理，由于单语数据仅仅使用了汉语数据，因此在第 2 步操作时，仅仅进行了汉语分词操作。本文在利用汉英模型翻译进行回译操作后，把翻译后的英文数据与汉语单语数据进行对应得到回译平行数据。之后对回译平行数据采用与双语数据相同的处理方式再处理一次。所有经过上述操作完成之后，CCMT 提供的平行数据保留了 7 百万条平行句子，CCMT 提供的单语数据保留了 4.5 百万条句子，WMT 提供的平行数据则保留了 2 千 1 百万条平行句子。本次测评将 CWMT2008，2009 和 2011 的验证数据、HTRDP2003 到 2005 的验证数据、NJU2017 测试和验证数据、SSMT2007 验证数据合并作为验证集，将 CCMT2019 的中英和英中验证集作为测试集，其中有多个翻译候选时，仅使用第一个候选与源语言句子作为平行数据。最终对训练、验证和测试数据集进行亚词切分处理，以减少低频词的影响。

2.2 模型介绍

在本次测评中，主要选用 FACEBOOK 团队研发的 Fairseq 开源系统[4]进行模型训练。本次测评按照模型参数设定和结构将评测中所使用的模型分为四种：BIG、DEEP，FFN-8196 和 DynamicConv。其描述如下：

BIG: Transformer 的编码器和解码器完全基于自注意力机制，能够节省大量的训练时间并显著的提升翻译质量，因此，本次测评将主要使用 Transformer 作为翻译模型。由于 Transformer 模型参数主要分为 transformer-base 和 transformer-big，而英汉翻译任务数据量比较丰富，因此本次测评将能容纳更多信息的 transformer-big 模型作为基线模型并称作 BIG，BIG 模型的参数与 Vaswani 论文[1]中设定的 Transformer-big 参数相同。

---

[2] https://spacy.io/

[3] http://ltp.ai/

[4] https://github.com/pytorch/fairseq

DEEP: 该模型同样是基于自注意了机制，主体结构与 transformer 结构相同。整体参数根据 Sun 等[7]所提供参数进行设置，主要是通过增加编码器层数来学习更好的源语言特征以得到更好的翻译结果，因此该模型除了将编码器层数由 6 层变为 15 层外，其他设置与 BIG 模型相同。

FFN-8196: Sun 等[7]认为前馈网络的维度对于 Transformer 的性能有很大的影响，因此增加了前馈网络的维度来提升模型的性能。该模型便根据 Sun 等[7]的方法扩大了前馈网络的维度，其他参数与 BIG 的参数相同。该模型的前馈网络维度设置与 Sun 等人[7]将维度设置为 15000 不同，因为所使用显卡的显存限制，本次测评将原始 Transformer-big 中前馈神经网络的维度由 4096 变为了 8196。

DynamicConv: 该模型与 Transformer 不同，主要使用动态卷积提取文本特征，并在某些情况下模型性能优于 Transformer。由于动态卷积与自注意力机制对于特征的提取存在一定的区别，本次测评将该模型作为次要模型，用于和 Transformer 模型进行集成以得到更好的效果，其结构与 Wu 等[2]一致。也可以在该模型参数的基础之上更改模型编码器的层数和前馈网络的维度，只不过由于时间的原因，本次评测过程中仅仅使用了最基础的模型参数。

2.3 数据增强

为了提升模型的性能，本次测评使用了回译方法[3]，该方法通过翻译模型翻译单语言数据来生成大量的数据来提高模型性能的方法。本文使用 CCMT 提供的平行数据来训练回译方法所需的汉英方向翻译模型，而单语数据则为 CCMT 提供的汉语单语，并使用 beam search[4]的解码方式来得到译文。回译方法得到数据的质量会影响到模型的性能，因此为了得到质量更好的翻译，本文将使用 CCMT 数据训练的 BIG 模型和 DynamicConv 模型进行集成，用于将汉语单语数据翻译为英文，考虑到回译方法得到的数据与真实数据存在一定的区别而且也无法确定回译得到的数据与平行数据领域是否相似，同时由于数据量过大，如果进行领域筛选需要花费大量时间，因此在训练过程中会使用加标签的方法[8]对回译得到数据进行处理，该方法与添加噪声的方法可以起到相同的作用，用于帮助模型区分平行数据和回译数据来提升模型的翻译效果。除此之外，为了进一步提升回译数据的质量，会对译文和汉语单语进行长度比过滤，处理方式与 2.2 小节中平行数据处理方法相同。鉴于 WMT 提供的平行数据中绝大部分数据为 2017 年和 2018 年爱丁堡大学系统回译的数据新闻，同时也不分确保 WMT 提供的数据与 CCMT 数据之间是否存在区别，因此我们将 WMT 提供的双语数据作为数据增强的数据来使用。

2.4 模型平均

模型平均是指将训练过程中不同时刻的模型参数进行平均从而减少模型参数的不稳定来提升模型的鲁棒性。本次测评使用了模型平均，会根据模型训练过程中的损失变化来确定模型收敛的范围，并将收敛范围内的 N 个时刻保存模型参数进行平行，这里由于本次测评训练模型时会将数据迭代 12 轮，并且每 4000 步保存一次模型。模型会在第 11 次迭代时收敛，但多训练几轮可能得到更好的结果，因此，本次测评在采用模型平均时会平均最后一轮和其前几轮的 10 个模型来得到性能更好的模型参数。

2.3 模型微调与集成

模型微调可以是模型参数更适用于验证集，由于本次测评中使用的数据虽然都是新闻数据，但是其领域与质量无法确定，因此使用了模型微调技术[4]。其中使用 CCMT 提供的 CWMT2008，2009 和 2011 的验证数据、HTRDP2003 到 2005 的验证数据、NJU2017 测试和验证数据、SSMT2007 验证数据以及 WMT 中往年的验证集和测试集作为数据进行微调，其中有多个翻译候选时，将每个候选句子与源语言句子作为平行数据。并对微调过后的模型进行平均。在实验中发现，基于 Transformer 架构的模型基本上微调 1 到 2 次迭代便可以达到

最好的结果，但是 DynamicConv 模型往往要微调 5 次迭代以上才能得到最好的结果，而微调之后，DynamicConv 模型的结果与 BIG 模型和 FFN-8196 模型的结果基本相同，没微调时却低于 BIG 模型和 FFN-8196 模型的结果。

模型集成可以通过提升模型的鲁棒性来进一步提升模型的性能。本次测评同样使用了模型集成[5]，考虑到集成模型的差异性越多，模型的提升越大。本次测评在上述四种不同的模型中选取微调后的四个模型以及微调后使用模型平均的四个模型来进行集成。

3. 实验

3.1 实验环境及模型参数

训练模型所用服务器的操作系统为 ubuntu21.04，CPU 为 E5-2640，内存为 256 GB，使用 4 块显存为 16G 的 V100 显卡。BIG 模型的参数是基于 Transformer Big Model，其中 batch size 为 2048 个 tokens，每次训练迭代 12 epochs，每 4000 步保存一次模型，并平均保存模型里面的最后 10 个或者 20 个模型的参数，dropout[9]设为 0.3，使用 Adam 优化模型参数，Adam[10]中β1 = 0.90，β2 = 0.98，初始学习率设为 0.001，warmup 为 4 000。DEEP 模型参数与 BIG 参数相同，仅仅将 BIG 模型的编码器层数改为 15 层，解码器层数不变。FFN-8196 的参数与 BIG 参数相同，只是将编码器和解码器的前馈神经网络的维度变为了 8192。DynamicConv 模型参数为：编码器层数为 7 层，解码器为 6 层，attention dropout 为 0.1,weight dropout 为 0.1 编码器和解码器中激活函数为 glu[13]，其他参数与 BIG 相同。在模型预测结果时，beam size 为 12，这里之所以将 beam size 设为 12 是参考 Sun 等人[7]论文里面的设置。数据预处理所采用的 BPE[11]迭代次数为 32000，翻译结果使用 Moses 中提供的评价脚本 multi-bleu.perl 计算得到词级的 BLEU 值[12]。

3.2 平均不同模型个数

表 1 平均不同个数模型参数对模型性能的影响

Tab.1 The effect of averaging different numbers of model parameters on model performance

t

平均模型个数	BIG	DEEP	FFN-8196	DynamicConv
1	29.93	31.01	30.07	29.81
5	30.59	31.22	30.48	29.87
10	30.61	31.39	30.48	29.89
15	30.47	31.34	30.36	29.80
20	30.47	31.31	30.43	29.77

为了验证模型平均对翻译质量的影响，本文对比不同个数模型参数以及单个模型的结果。由于数据集大小、GPU 个数以及每一个 batch 下 token 的个数会对每一 epoch 训练模型的步数产生影响。数据集越大、GPU 个数越少，每个 batch 中 token 个数越少，每一 epoch 中的步数越大。为了更科学的对比结果，本文中所有模型使用的这两个参数和数据大小相同，因此每一 epoch 中的步数基本相同。并且每个模型都训练 12epoch，且每 4000 步保存一次模型，所以最终四个中结构都会保存 48 个模型。由于本文在训练模型时设置的 epoch 和步数较大，这四个模型基本会在 8 到 12epoch 时收敛速度变缓，在 11 和 12epoch 收敛，所以在第 8epoch 内保存的模型在验证集上的结果与最终收敛的结果仅相差 0.2 个 BLEU。因此本文

对比从第 8epoch 保存的模型到第 12epoch 保存的模型进行不同个数模型平均的对比。这中间保存有 20 个模型，因此本文按照平均模型个数为 1、5、10、15、20 的结果进行对比。

实验结果如表 1 所示，基本上本文所使用的四种结构不同的模型在平均 10 个模型参数时结果最好，但与平均其他个数模型的结果差别没有那么明显。对于 DEEP 和 DynamicConv 模型，平均 10 个模型参数的结果对比平均 5 和 15 个的基本没有差别。从总体上看，随着平均模型个数的增加，平均模型参数后的结果会先上升后下降，在平均 10 个模型时到达顶点。因此，本文在进行模型平均时，为了方便都平均 10 个模型来得到最终的结果。出现这种结果的原因可能是因为模型在训练过程中会慢慢收敛，如果平均收敛范围左右的模型参数，且每个参数的模型结果相差不大，那么平均后的结果可能是最佳的。如果增加模型个数，则会加入未收敛或者过拟合的模型，这是模型的性能相差较大，平均后的结果反而相比个数少的结果差。在实验过程中，在第 8 和第 9epoch 内保存的模型结果变化幅度比较大，因此区间内保存的模型可能并没有完全收敛。如果继续增加模型个数，反而降低了结果。

3.2 英中方向翻译结果

表 2 英中翻译任务在测试集上的结果

Tab.2 The results of the English-Chinese translation task on the test set

模型	BIG	DEEP	FFN-8196	DynamicConv
基线	28.48	-	-	28.02
+数据增强	29.93	31.01	30.07	29.81
+模型平均	30.61	31.39	30.48	29.89
+微调	31.96	33.35	32.17	31.84
+模型平均	31.96	33.40	32.34	32.03
集成		34.04		

不同方法在英汉翻译任务上的结果如表 2 所示，其中基线为只使用 CCMT 双语数据训练的模型，由于时间原因，仅仅在 BIG 和 DynamicConv 上进行了实验，并没有使用 CCMT 双语训练 DEEP 和 FFN-8192 模型。数据增强则是使用 CCMT 双语数据、CCMT 单语数据回译的数据以及 WMT 双语数据进行的实验。所以模型结果的提升比较大。参考 CCMT 和 WMT 往年评测报告，本文发现在资源丰富的翻译任务上，如果通过数据增强来得到比较可观的提升，则需要投入大量的时间进行回译操作。因此，本次评测没有深入研究数据增强方法。在 BIG、DEEP 和 FFN-8196 模型结构上，在数据增强方法上进行模型平均都有比较大的提升。微调所使用的数据则是在第 2.3 小节所提到的数据进行的实验。而在微调之上进行模型平均得到的提升并不明显。从数据增强实验结果看，DEEP 模型的性能比 Big 提升了 0.78 个 BLEU，而 DynamicConv 和 FFN-8196 则比 Big 模型要低，但经过微调之后，DynamicConv 和 FFN-8196 则要比 Big 的结果要好。而 DEEP 的结果无论是在数据增强和微调情况下都取得了做好的结果。其中 DEEP 模型微调的结果比 Big 基线提升了 4.87 个 BLEU。将四种模型集成的结果也比四个不同模型微调的结果要好，其中相比 DEEP 微调的结果高了 0.69 个 BLEU。实验证明，通过增加编码器的层数确实可以有效的提升模型的性能，而增加前馈神经网络的维度的 FFN-8196 在微调时有一定的提升，但是不太明显。这种结果可能与前馈神经网络的维度的设置存在一定联系，可能维度大到一定程度才能起到作用例如将维度设为

15000。更深的层数也许对于小数据微调更敏感，因为 DEEP 和 DynamicConv 微调的提升明显要好于其他两个，这里 DEEP 的编码器的层数为 15，而 DynamicConv 的层数则是 7。

4. 总结

本文描述了新疆大学信息学院参加第十八届全国机器翻译大会机器翻译评测英中新闻领域的翻译评测任务的总体情况。在本次测评中，我们主要使用了数据增强、模型微调，以及模型集成等方法提升翻译质量。实验结果证明，这些方法能够有效提高翻译质量，其中模型微调对翻译效果的提升最为显著。由于时间和机器的原因，本次评测中还有一些比较有效的方法没有使用，当然在实验的过程中也不断的发现了些问题和不足。这些都将是在今后研究中主要需要解决的，期望之后的研究中能够学习更多有效的方法，并对机器翻译研究有所贡献。

参考文献

[1]. VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Massachusetts:MIT, 2017: 5998-6008.

[2]. Wu F, Fan A, Baevski A, et al. Pay Less Attention with Lightweight and Dynamic Convolutions[C]//International Conference on Learning Representations. 2018.

[3]. Sennrich R, Haddow B, Birch A. Improving Neural Machine Translation Models with Monolingual Data[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 86-96.

[4]. Chu C, Dabre R, Kurohashi S. An empirical comparison of domain adaptation methods for neural machine translation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017: 385-391.

[5]. Wang Y, Wu L, Xia Y, et al. Transductive ensemble learning for neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 6291-6298.

[6]. Wu S, Wang X, Wang L, et al. Tencent neural machine translation systems for the WMT20 news translation task[C]//Proceedings of the Fifth Conference on Machine Translation. 2020: 313-319.

[7]. Sun M, Jiang B, Xiong H, et al. Baidu neural machine translation systems for WMT19[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019: 374-381.

[8]. Caswell I, Chelba C, Grangier D. Tagged Back-Translation[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers). 2019: 53-63.

[9]. SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[10]. KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[11]. SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]// Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 1715-1725.

[12]. 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.

[13]. Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks[C]//International conference on machine learning. PMLR, 2017: 933-941.

English-Chinese Machine Translation Evaluation Task Technical Report

YI Nian[1], AISHAN Wumaier[1], Liejun Wang[1]
(1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046)

Abstract: This article mainly introduces the basic situation of the School of Information Science and Engineering of Xinjiang University participating in the machine translation evaluation project of the 18th National Machine Translation Conference. In this machine translation evaluation, I participated in the machine translation evaluation project in the English-Chinese news field. This paper mainly describes the model framework, data preprocessing process, data enhancement, and model fine-tuning and integration methods adopted by the English-Chinese neural machine translation system in this competition. The paper concludes with the performance of the submitted system on the test set.

Keyword: English-Chinese neural machine translation; self-attention mechanism; model fine-tuning; model ensemble