同方知网 CCMT2021 评测技术报告

刘嘉[1]，李楠[2]*

（同方知网（北京）技术有限公司全球学术文献大数据中心，北京 100192）

摘要：本文描述了同方知网参加第十七届全国机器翻译大会（CCMT'2021）中英新闻领域翻译评测任务的总体情况。本次评测任务采用了谷歌 Transformer 神经网络机器翻译模型架构，同时为了提高翻译效果，还应用了数据过滤、数据增强、模型微调，以及模型集成等技术。报告将主要阐述本次参评系统采用的技术方案，同时给出了不同设置下系统在评测数据上的性能表现，并进行对比和分析。

关键字：机器翻译；深度神经网络；注意力机制

中图分类号：TP391　　**文献标志码：** A

1. 引言

本文描述了同方知网参加第十七届全国机器翻译大会（CCMT'2021）中英新闻领域的翻译评测任务的总体情况。本次评测采用谷歌 Transformer 神经网络机器翻译架构[1]。在本次测评中，我们主要使用了数据过滤、数据增强、模型微调，以及模型集成等方法提升翻译质量，其中在数据增强阶段，采用了回译[4]方法构建伪平行语料，同时应用知识蒸馏方法[11]进一步补充神经机器翻译模型的训练集。在模型微调阶段，采用了基于领域内训练集数据微调以及验证集数据微调的两阶段微调方法，同时应用了迭代集成学习微调技术。在实验中，对比了系统在各任务上不同设置下的表现，并对实验结果进行了分析。

2. 系统介绍

本节主要描述了参加本次测评系统所使用的技术和方法。

2.1 模型

本次测评采用了基于自注意力机制的 transformer 模型，该模型是由编码器和解码器组成的 sequence-to-sequence 结构，其中编码器和解码器由 N 个层块堆叠而成。编码器的每层均包括两个子网络层，分别是多头自注意力层和前馈神经网络层。本次测评使用了基于开源的 opennmt-py 系统，基线模型采用 transformer-big 结构，每个模型使用 2 块 GPU 核进行训练，每个 batch 大小为 4096。词向量的维度为 1024，隐层状态维度为 4096，编码器与解码器均为 6 层，多头自注意力机制使用 16 个头。

2.2 数据预处理

由于平行语料的质量会对翻译的效果产生很大影响，因此需要对平行语料进行预处理。本次测评采用的预处理过程主要包括文本预处理和语料筛选两部分，文本预处理包括：

- 替换转义字符，删除控制字符；

* **通信作者：**ln6562@cnki.net

- 全角字符转半角字符，繁体字符转简体字符；
- 统一标点符号；
- 分词（tokenization）：中文使用 jieba，英文使用 Moses 进行分词。

在完成文本预处理后，还要对平行语料进行过滤，删除掉其中质量较低的语料，包括：

- 删除重复语料；
- 删除长度大于 160 个词的语料；
- 使用 fasttext 检测平行句对的语种，删除语种识别错误的数据；
- 删除长度比过大的句对（对于每个数据集删除其长度比前 5%的双语例句）；
- 使用 fast_align 对平行语料的对齐度按中->英和英->中方向分别进行打分，对齐度按分词后的词对齐度进行打分，并计算每个句对中->英和英->中方向对齐度的平均分，删除每个数据集得分最差的 5%的双语例句；
- 对于平行语料数据，在完成基线模型的训练后，使用基线模型生成将平行语料的中文部分翻译成英文，并计算其与原英文句子的 bleu 值，删除每个数据集得分最差的 10%的双语例句。

2.3 数据增强

本次测评采用了回译和知识蒸馏这两种数据增强方法。

回译（Back Translation）：为了有效利用目标端单语数据提升翻译质量，本次测评采用了回译的方法扩大中英平行语料的规模。回译采用的单语数据包括 News Crawl、News commentary 以及 XinHua news 等数据集。由于到这些语料集数据量较大，我们使用了语言模型对单语数据进行了筛选。具体过程如下：

1．首先基于 News Commentary 平行语料分别构建中英语言模型；

2．使用上述语言模型对中英单语数据分别进行排序，并各保留前 1000 万单语数据作为待回译的语料；

3．使用基线模型对保留语料进行目标端到原端的翻译，并将生成的回译数据加入到平行语料库作为训练数据。

在构建回译数据的过程中我们使用了 beamsearch 的解码方法，而没有采用 sampling 方法，也未加入噪音数据。因为经实验发现，在我们的测试系统中，采用后两种方法并未观察到能有效提升翻译效果。

知识蒸馏（Knowledge Distillation， KD）：为了进一步充分利用平行语料资源和原端的单语语料库，本系统使用了两种知识蒸馏方法将集成模型的知识迁移到单个模型，包括：

1． 集成知识蒸馏（Ensemble KD）：使用多个基线模型对平行语料原语言端文本进行翻译，将翻译的合成语料经过 2.2 节所描述的过滤步骤后加入到训练数据中；

2． 单语知识蒸馏（Monolingual KD）：使用多个基线模型对单语数据进行翻译，将翻译的合成语料经过 2.2 节所描述的过滤步骤后加入到训练数据中。

考虑在知识蒸馏训练过程中质量较差的合成语料会对单个模型造成伤害，我们在合成语料的筛选过程中提升了 bleu 值和 fast_align 对齐分数的阀值，删除了 bleu 值较低的 50%的合成语料。

2.4 模型微调

使用领域内（in-domain）的数据在已经训练好的模型上进行微调是一种有效地提升模型翻译效

果的方法。为了充分利用平行语料训练数据和验证集数据，本系统采用了两阶段的模型微调方法，即先从训练数据中筛选一部分语料进行初步微调，进而在数据量较小的验证集数据进一步进行精细的微调，包括混合测试集原语言端语料实现集成学习微调，具体过程如下：

训练集数据微调：使用 N-grams 语言模型选择平行语料和回译生成的合成语料中与领域内语料最相似的数据进行微调。首先，使用历年 WMT 发布的验证集和测试集数据分别构建中/英领域内语料；然后，使用 kenlm 训练相应的 tri-grams 语言模型作为领域内语言模型；进一步，使用平行语料分别训练中英 tri-grams 语言模型，作为领域外（out-of-domain）语言模型。根据文献[13]，我们使用如下的公式计算平行语料中每一个句对的交叉熵之差作为筛选分数：

$$CE(H_{I-SRC}, H_{O-SRC}) + CE(H_{I-tgt}, H_{O-tgt})$$

其中，O代表领域外语料，I代表领域内语料，H_{I-SRC}代表源语言端领域内语言模型，$\mathrm{H_{I-tgt}}$代表目标端领域内语言模型，$\mathrm{H_{O-SRC}}$代表源语言端领域外语言模型，$\mathrm{H_{O-tgt}}$代表目标语言端领域外语言模型，CE代表交叉熵函数。

对于平行语料数据和由单语语料构建的合成语料数据，使用上述公式对每一个句对进行打分，并且分别根据得分进行排序。从平行语料数据和合成语料数据中各抽取排序最高的 50 万句对，共计 100 万句对进行模型微调。由于训练集微调数据较多，不易发生过拟合问题，因此在微调时采用了和模型训练时相同的配置，并未降低学习率等参数。

验证集数据微调：在训练集数据微调的基础上，我们进一步使用了数据量较小的验证集数据进行更精细的微调。我们采用上述的 WMT 验证集和测试集作为微调数据。根据文献[14]的实验结果，我们将微调数据按源语言端数据来源分为两个部分，即根据 XML 文件中的原始语言属性，将验证集划分为来自于原始中文和原始英文这两个部分。对于中英方向翻译，我们使用 CWMT2008, CWMT2009 以及 newsdev2017，newstest2017，newstest2018，newstest2019 中原始语言为中文的语料作为验证集微调数据；而对于英中方向，我们使用 newsdev2017， newstest2017，newstest2018，,newstest2019 中原始语言为英文的语料作为验证集微调数据。由于验证集微调数据较少，为防止发生过拟合现象，在微调时降低学习率参数至 0.00001。

迭代集成学习微调：在前两步模型微调的基础上，我们还实验了文献中[7]提出的迭代集成学习微调方法，该方法在微调数据集中混合了测试集原语言端语料，利用知识蒸馏的思想，实现翻译模型的自我学习（self-learning）。具体过程如下：1）使用模型集成方法利用验证集和测试集源语言端数据构建合成语料，并使用该合成语料对模型进行微调；2）使用验证集对模型进行进一步微调；3）仅使用测试集合成数据实现模型对领域内知识的自学习。上述步骤可以迭代执行多次，直至在验证集上的得分不再增长。由于迭代集成学习微调中已包括验证集数据微调过程，因此在实际训练中，我们直接利用 1，3 方法进行两阶段微调。

2.5 模型集成

集成学习是指使用多种兼容的学习方法或模型来执行单个机器学习任务的方法。在机器翻译领域应用模型集成的方法可以有效地整合多个模型预测的概率分布，以达到整体提升翻译效果的目的。在本系统中，使用了如下的模型集成方法：

模型平均：模型平均是指将同一模型在训练不同时刻保存的参数进行平均以期得到更加鲁棒的

模型参数。使用模型平均可以减少模型参数的不稳定性，进一步提升模型鲁棒性。一般地，保存的参数通常是模型基本收敛时对应的最后 N 个时刻的参数。在本次测评中采用的模型平均策略是，从第 10 万步起每个新保存的模型与之前保存的两个模型进行平均。

解码器集成：解码器集成就是将目标端的训练看成词序列预测任务，结合多个模型预测概率来预测下一个词的翻译。在本次测评中，我们使用了不同随机种子和随机打乱训练语料等不同方法来构建多个模型。

3. 实验

3.1 实验设置

本次测评所使用的模型训练系统是基于开源的 opennmt-py 2.0，我们使用 transformer-big 作为基线系统。在训练阶段，使用 Adam 作为优化器，其参数配置为$\beta_1 = 0.9$，$\beta_2 = 0.98$，$\epsilon = 10^{-8}$，系统初始学习率为 0.0001，在验证集数据微调阶段，学习率降至 0.00001 以防止发生过拟合问题。其他训练参数为：Batchsize=4096，dropout=0.1，每次训练在两块 GPU 上迭代 50 万步，warmup 步数为 8000。

在数据预处理完成预处理后还需要对其进行子词化（sub-word），以降低模型词表的大小，缓解集外词问题，减少模型参数加快模型训练速度。针对中英语料的不同特点，我们使用了不同的子词化方法：对于英文数据，使用 BPE 进行切分，词表大小设置为 32000。对于中文数据，采用了二次分词的方法，具体过程为：基于结巴分词的结果按词频统计排序构建大小为 N 的二次分词词表，本次测评中 N 的取值与 BPE 词表大小相同，均为 32000。然后使用最大长度匹配法对已用结巴分词的中文文本进行二次切分。在二次分词的过程中，若中文文本中出现英文单词，则使用英文数据对应的 BPE 词表进行切分。

3.2英中方向测评结果

表 1 英中翻译评测在开发集上的结果

Table 1. The results of the English-Chinese translation evaluation on the development set

模型	news2019	news2020
基线模型	32.7	38.06
+数据过滤	34.75	39.74
+数据增强	33.83	40.62
+模型微调		
+训练集微调	36.45	43.12
+验证集微调	-	45.22
+模型集成	-	45.67

在英中方向任务中，使用的平行语料包括 CCMT 数据，News Commentary v16，ParaCrawl v7.1，UN Parallel Corpus V1.0，WikiMatrix，Wiki Titles v3 等数据集，单语数据使用了 News Commentary，News crawl 以及 Xinhua news 语料，同时使用了部分 WMT 提供的 Back-translated news 数据。训练得到的翻译模型在 WMT news 2019 和 2020 测试集上的结果如表 1 所示，其中使用数据过滤分别增长 2.05 和 1.68 分，使用数据增强技术在 news2020 上增长 0.88 分，但在 news2019 上反而下降 0.92。而模型微调对这两个验证集均有明显的提升效果，本次测评使用了训练集微调和验证集微调两种方法，其中验证集微调同时使用了迭代集成学习微调技术，在 news2019 上使用训练集微调提升了 2.62 分，由于验证集微调使用了 news2019 的数据，因此只在 news2020 上测试了提升效果，两个微调阶段总计提升了 4.6 分，结果显示两阶段模型微调对翻译效果有明显提升作用。最后，用两个模型应用集成技术使分数小幅提升了 0.45 分。

3.3 中英方向测评结果

对于中英方向翻译任务，除了在单语语料中没有使用 Xinhua news 数据，其他语料与英中方向任务所使用的语料相同，中英方向测评结果如表 1 所示，测评显示，数据过滤、数据增强、模型微调，以及模型集成方法对翻译效果均有不同程度的提升，并且与英中方向类似，模型微调仍然是对翻译效果提升最显著的优化技术。

表 2 中英翻译评测在开发集上的结果

Table 2. The results of the Chinese-English translation evaluation on the development set

模型	news2019	news2020
基线模型	28.79	30.08
+数据过滤	29.01	30.45
+数据增强	29.56	31.13
+模型微调		
+训练集微调	34.47	32.38
+验证集微调	-	34.88
+模型集成	-	35.27

4. 总结

本文描述了同方知网参加第十七届全国机器翻译大会机器翻译评测（CCMT'2021）中英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过滤、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。

参考文献：

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In Proceedings of NIPS 2017.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword

Units. In Proceedings of ACL 2016.
[3] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In Proceedings of NIPS 2014.
[4] Sennrich, R., Haddow, B., & Birch, A. (2016, August). Improving Neural Machine Translation Models with Monolingual Data. InProceedings of the 54th AnnualMeeting of the Association for Computational Linguistics.Volume 1: Long Papers.ACL 2016.pp. 86-96.
[5] 刘洋. 神经机器翻译前沿进展[J]. 计算机研究与发展，2017,54(06):1144-1149.
[6] Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013, August). Scalable modified Kneser-Ney language model estimation. InProceedings of the 51st Annual Meeting of the Association for Computational Linguistics ACL 2013.Volume 2: Short Papers.pp. 690-696.
[7] Wang, Y., Wu, L., Xia, Y., Qin, T., Zhai, C., Liu, T.Y.: Transductive ensemble learning for neural machine translation. In: AAAI (2020
[8] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In Proceedings of EMNLP 2018.
[9] Marzieh Fadaee and Christof Monz. 2018. Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation. In Proceedings of EMNLP 2018.
[10] Marlies van der Wees, Arianna Bisazza and Christof Monz. 2017. Dynamic Data Selection for Neural Machine Translation. In Proceedings of EMNLP 2017.
[11] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of EMNLP 2016.
[12] Hoang, V. C. D., Koehn, P., Haffari, G., & Cohn, T. (2018, July). Iterative back-translation for neural machine translation. In Proceedings of the 2nd Workshop onNeural Machine Translation and Generation NMT 2018.pp. 18-24.
[13] Tanfang Chen, Weiwei Wang, Wenyang Wei, Xing Shi, Xiangang Li, Jieping Ye and Kevin Knight.2020. in Proceedings of the Fifth Conference on Machine Translation. 2020
[14] Chen G , Wang S , Huang X , et al. Tsinghua University Neural Machine Translation Systems for CCMT 2020[C]// 2020.

CNKI Evaluation Technical Report for CCMT 2021

Liu Jia，Li Nan *

（Tongfang Knowledge Network Technology Co., Ltd.(Beijing)，Beijing 100192，China）

Abstract: This paper presents an overview and the technical details of the Chinese-English news translation evaluation task in the 17th China Conference on Machine Translation (CCMT'2021) attended by CNKI. The evaluation task adopts the framework of Google's Transformer neural network model. In order to improve the translation effect, data filtering, data augmentation, model fine-tuning, and ensemble learning are also applied. The report will mainly describe the technical framework of the system, give the performance of the system on the evaluation dataset under different settings and conduct a comparison and analysis

Keywords: *Machine translation; Deep neural network; Attention mechanism*