中国科学技术信息研究所 CCMT'2020 评测技术报告

刘文斌，魏家泽，吴振峰，潘优，何彦青[1]*

（中国科学技术信息研究所情报理论与方法研究中心，北京 100038）

摘要：本文详细介绍了中国科学技术信息研究所（ISTIC）参加第十六届全国机器翻译大会机器翻译评测（CCMT'2020）的总体情况和采用的技术细节。在本次评测中，ISTIC 共参加了六个任务，分别是汉英新闻领域的翻译评测、蒙汉日常用语领域的翻译评测、藏汉政府文献领域的翻译评测、维汉新闻领域的翻译评测、专利领域的日、汉、英多语言翻译评测以及汉英平行语料过滤任务。报告将主要阐述本次参评系统采用的模型框架、数据预处理方法以及译码策略。报告最后给出了不同设置下评测系统在评测数据上的性能表现，并进行了对比和分析。

关键词：机器翻译；CCMT'2020；神经网络；自注意力机制

中图分类号：G355　　**文献标志码：** A

1 引言

本文详细介绍了中国科学技术信息研究所（ISTIC）参加第十六届全国机器翻译大会机器翻译评测（CCMT'2020）翻译评测任务的总体情况。ISTIC 共参加了 CCMT'2020 中的六个任务，分别是汉英新闻领域的翻译评测、蒙汉日常用语领域的翻译评测、藏汉政府文献领域的翻译评测、维汉新闻领域的翻译评测、专利领域的日、汉、英多语言翻译评测以及汉英平行语料过滤任务。

本次评测采用谷歌 Transformer[1]神经网络机器翻译架构。在数据预处理方面，针对评测方发布的数据，采取多种不同语料过滤方法减少语料噪声以提高训练语料的质量。同时为了进一步利用评测方发布的单语数据，在蒙汉日常用语领域的翻译评测、藏汉政府文献领域的翻译评测、维汉新闻领域的翻译评测任务上，使用了回译方法来构建伪平行语料，补充神经机器翻译模型的训练集。在译文输出过程中，采用了模型平均[2]和集成解码的策略，最后利用译文重排序策略给出最终的译文。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

图 1 给出了 ISTIC 在本次评测中翻译评测的整体流程图。本次评测使用的基线系统是基于自注意力机制的 Transformer 模型，其模型结构包含编码器和解码器两个部分[3]，如图 2 所示。模型采用完全自注意力机制，能够在实现算法并行性、加快模型训练速度的同时，进一步缓解长距离依赖，并提高翻译质量[4]。编码器和解码器由 N 个层块堆叠而成。编码器的每个层块包含两个子模块，分别是多头自注意力模块和一个前馈神经网络模块，其中多头自注意力模块将隐状态的维度划分为多个部分，每个部分分别使用自注意力函数计算得到，然后将这些输出向量拼接起来。多头的作用是使模型能够更大程度的关注到不同位置不同表示子空间的特征信息。多头注意力方法包括两个步骤：1)点积注意力计算；2)多头注意力计算。点积注意力的计算方式为:

[1] **基金项目**：中国科学技术信息研究所重点工作项目“面向垂直领域应用场景的机器翻译研究”（ZD2020-18）；中国科学技术信息研究所重点工作项目“俄汉跨语言知识发现与服务研究”（ZD2020-10）

* **通信作者**：heyq@istic.ac.cn.

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}})\mathrm{V}$$

其中 Q 为查询向量，K 为键向量，V 为值向量，d_k 为隐层状态的维度。在点积注意力的基础上，多头注意力机制的计算方式为：

$$\mathrm{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)\mathrm{W}^{\mathrm{O}}$$

W^O 为矩阵参数。每个头的注意力值为：

$$\mathrm{head} = \mathrm{Attention}(\mathrm{QW}_{\mathrm{i}}^{\mathrm{Q}},\mathrm{KW}_{\mathrm{i}}^{\mathrm{K}},\mathrm{VW}_{\mathrm{i}}^{\mathrm{V}})$$

解码器的每个层块由三个子模块构成，除了编码器中的两个模块外，还加入了一个解码器-编码器的注意力模块，这一模块的作用是用于在解码单词的时候关注源语言端信息。为了避免层数过多导致模型难以收敛，编码器与解码器都使用了残差连接和层级正则技术。为了使模型更好获得输入序列的位置信息，在编码器和解码器的输入层都加入了额外位置编码向量。在解码器得到隐层状态后，Transformer 模型将该隐层状态输入 softmax 层并与候选词表进行打分，得到最终的译文结果。

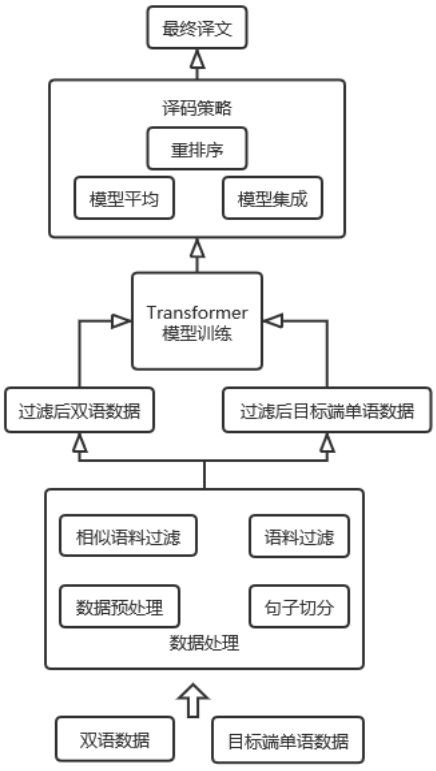

图 1 评测整体流程图

Fig. 1 Evaluation flow chart

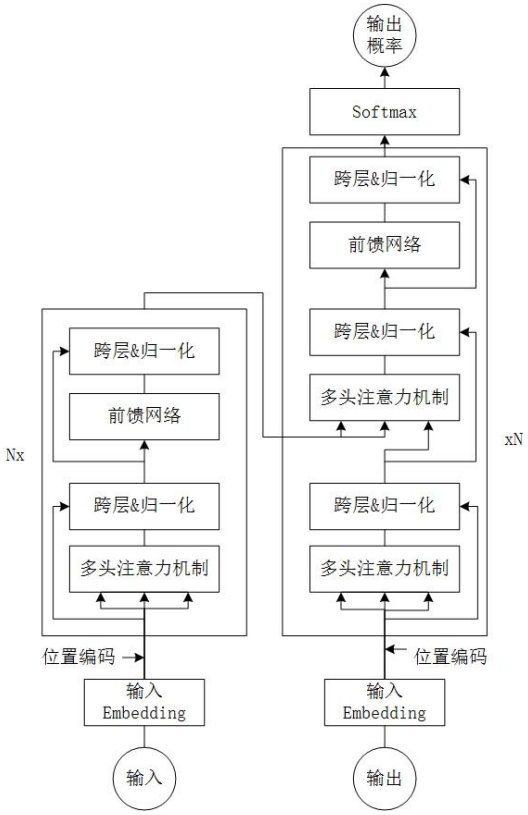

图 2 基于自注意力机制的 transformer 模型结构

Fig. 2 Transformer model structure based on self-attention mechanism

3 方法介绍

本次评测中，ISTIC 共参加了 CCMT'2020 中的汉英翻译评测、蒙汉、藏汉、维汉翻译评测、日汉英多语言翻译评测以及汉英平行语料过滤六个任务。以下分别介绍每个任务使用的方法，其中蒙汉、藏汉、维汉三个少数民族语言评测任务合并介绍。

3.1 汉英翻译评测

汉英翻译评测在采用基于自注意力机制的 Transformer 模型的基础上，使用了模型平均、集成、重排序方法。

3.1.1 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得到更加鲁棒的模型参数，以此减少模型参数的不稳定性，从而提升模型鲁棒性。这里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，作为一种增强模型鲁棒性的方法，模型平均通过整合多个时刻的训练结果，可以有效的减小训练过程中各参数以及最后预测结果的方差与偏置，比单一时刻得到的模型达到更好的效果，可以进一步提升模型的性能。

本报告在评测中对不同的评测任务采取不同的平均策略。在汉英新闻领域的翻译评测和专利领域的日、汉、英多语言翻译评测任务中，对训练器中指定 max-update 参数后训练得到的最后三个更新轮次 checkpoints 进行平均；在蒙汉日常用语领域的翻译评测、藏汉政府文献领域的翻译评测、维汉新闻领域的翻译评测任务中，从两个维度对最后三个更新轮次的模型参数进行平均，分别是训练器中指定 max-update 和 max-epoch 参数后训练分别得到的最后三个更新轮次即 update-checkpoints 和 epoch-checkpoints 进行平均。此外，模型平均得到的更为稳固鲁棒的单模型也将用于模型集成环节联合进行概率分布预测。

3.1.2 模型集成

集成学习是一种联合多个学习器进行协调决策的机器学习方法，应用在机器翻译任务的推断过程中可以有效整合多个模型预测的概率分布，达到提升翻译系统准确性的目的。神经机器翻译集成通常是在解码时，由多个模型同时预测当前时刻目标端词语的概率分布，最终将多个模型预测的概率分布进行加权平均，从而共同决定最终输出。现有的神经机器翻译模型集成方法主要有两种，一种是融合源端上下文信息的模型集成，另一种是基于解码器的模型集成。本报告中采用第二种方法，即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。

本报告在模型集成中根据不同任务采用了同模型不同随机种子和不同模型不同随机种子两种方法。在蒙汉常用语领域的翻译评测、藏汉政府文献领域的翻译评测、维汉新闻领域的翻译评测任务中，使用了同模型不同随机种子的方法，集成了模型平均环节从 update 和 epoch 两个维度各自得到的单模型以及主系统训练生成的 checkpoint_best 模型和 checkpoint_last 模型，生成最终译文。在汉英新闻领域的翻译评测和专利领域的日、汉、英多语言翻译评测任务中，使用了不同模型不同随机种子的方法，模型一由基于 transformer 在 wmt 数据集英德语言对上训练得到默认参数的基础上，设置参数学习率为 0.001，dropout 为 0.2 等进行训练得到。模型二由基于 transformer 在 iwslt 数据集德英语言对上训练得到默认参数的基础上，设置参数学习率为 0.0007，dropout 为 0.3 等进行训练得到，最后集成两个模型解码生成最终译文。

3.1.3 重排序

解码时的柱搜索策略可以提供多个候选结果，选取打分模型对其进行质量评估可以进一步提升翻译的准确性。打分模型具有多种策略：正向模型、反向模型、目标端译文从右到左、语言模型、候选译文与源端输入的长度比、正反向翻译的对齐概率与翻译覆盖率、候选译文之间的贝叶斯风险

概率等。本评测中，基于评测数据训练了多个打分模型来对候选译文进行质量评估。最后根据学习到的特征权重，对测试集中的候选译文进行打分并排序[5]，选择得分最高的译文作为最终的输出译文。

3.2 蒙藏维翻译评测

本评测中，由于评测方发布的蒙汉、藏汉、维汉平行语料规模小，所以使用了回译方法来构建伪平行语料，利用评测方发布的蒙汉、藏汉、维汉平行语料训练汉蒙、汉藏、汉维翻译模型，将过滤处理后的单语数据经过上述模型翻译得到蒙汉、藏汉、维汉伪平行语料来补充神经机器翻译模型的训练集。翻译模型在采用基于自注意力机制的 Transformer 模型的基础上，同样用到了模型平均、集成方法。与汉英和多语言评测任务不同的是，本评测在模型平均时对两种不同维度下各个时刻的参数进行平均，在模型集成时对模型平均环节得到的单模型及主系统训练得到的各模型混合集成，有效整合多个模型进行概率预测分布，解码得到最终译文。

3.3 多语言翻译评测

多语言机器翻译任务试图挖掘不同语言之间的对应和转换关系，只提供了专利领域的汉英、汉日平行语料，没有日英平行语料，任务目标是评价日英翻译方向的翻译性能。ISTIC 提出了两种思路训练日英翻译系统，整体框架图如图 3 所示。

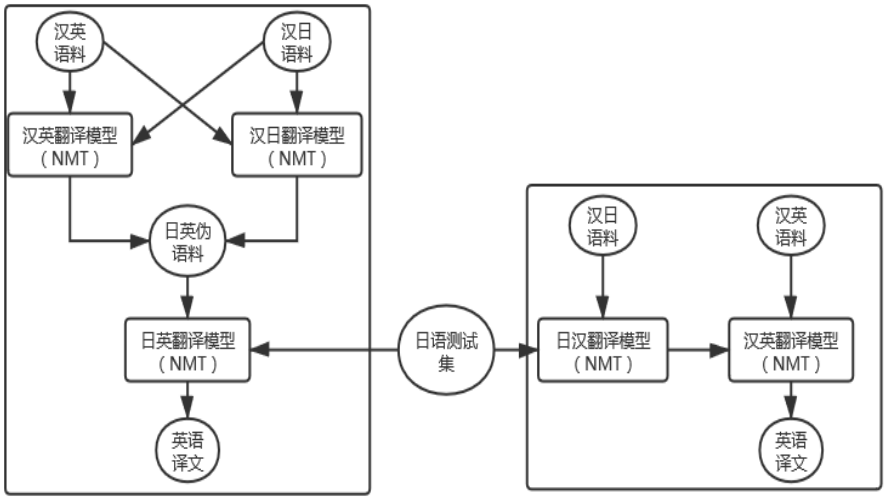

图 3 多语言翻译系统框架图

Fig. 3 Framework of Multilingual Translation System

第一种思路为非桥接方式，即利用汉日平行语料训练汉日翻译模型，利用汉英平行语料训练汉英翻译模型。通过汉日翻译模型将汉英平行语料中的汉语翻译为日语，并与汉英平行语料中的英文语料构建日英伪平行语料，通过汉英翻译模型将汉日平行语料中的汉语翻译为英语，并与汉日平行语料中的日语语料构建日英伪平行语料，最后合并日英伪平行语料训练得到日英翻译模型，最后通过日英翻译模型翻译日语测试集得到英语译文。第二种为思路为桥接方式，即首先利用汉日平行语料训练日汉翻译模型，利用汉英平行语料训练汉英翻译模型，然后将待翻译的日语测试集通过日汉翻译模型翻译成汉语，再将得到的汉语译文通过汉英翻译模型翻译得到最终英语译文。翻译模型在采用基于自注意力机制的 Transformer 模型的基础上，同样使用到了上文介绍的模型平均、集成、重排序方法，实验中涉及到的过渡翻译及最终的英语译文生成均基于主系统进行翻译，即基线+平均+

集成+重排序。

3.4 汉英平行语料过滤

汉英平行语料过滤任务需要将评测方发布的汉英平行语料训练集按某种排序方法由高到低排序，排序越高意味着质量越高。处理流程如图 4 所示。

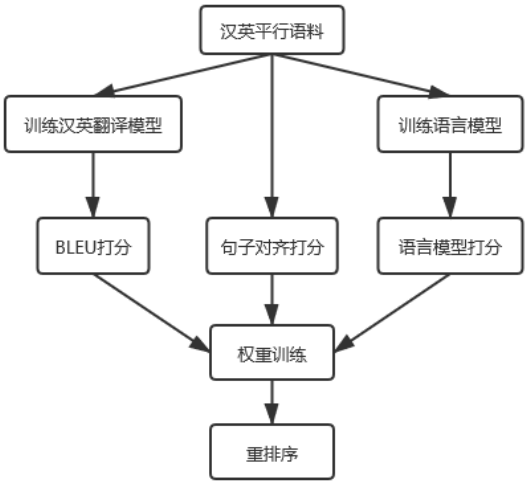

图 4 汉英平行语料过滤流程图

Fig. 4 Chinese-English parallel corpus filtering flowchart

该任务利用评测方发布的汉英平行语料分别训练汉英神经机器翻译模型和语言模型，对每对汉英平行句对：（1）将汉语做测试集通过训练好的汉英翻译模型翻译得到译文英文并于原英文之间进行 BLEU 打分，以句对的翻译质量作为第一个评价指标；（2）通过 Champillion[6]句子对齐方法对原汉英句对进行对齐打分以句对的对齐程度作为第二个评价指标；（3）将原汉英句对通过训练好的语言模型打分，以句对的平滑流畅度作为第三个评价指标。根据计算公式：

$$Score(S_i, T_i) = \lambda_{1} * BLEU(T_i, T'_i) + \lambda_{2} * CHAMPILLION(S_i, T_i) + \lambda_{3} * ML(S_i, T_i)$$

其中 λ_1 λ_2 λ_3 为权重训练得到的参数，S_i 为汉英平行语料中的源端语言（汉语），T_i 为汉英平行语料中的目标端语言（英语），T'_i 为汉英平行语料中的源端语言（汉语）经汉英翻译模型翻译得到的译文（英语）。对三个指标采用 Z-MERT[7]工具，以 CCMT2019 数据集为开发集，以 Bleu 得分为参考，进行权重融合训练。最后根据得到的权重计算对齐总分，按照由高到低顺序对汉英平行语料库进行重排序得到最终结果。

4 实验

本评测使用的系统为开源项 tensor2tensor[8]。主要参数设置如下，每个模型使用 1-3 块 GPU 核进行训练，每个 batch 大小为 2048。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。本次评测采用了 dropout[9]机制，dropout 设为 0.1。训练语料均采用 BPE[10]切分，其中源语言及目标语言的词表共享设定为 30K。初始学习率为 0.2，warmup 步数为 8000。

4.1 数据处理

4.1.1 语料预处理

本次评测每个任务都只使用了 CCMT'2020 提供的平行语料和单语数据。其中汉英平行语料为新闻领域，蒙汉平行语料为日常用语领域，藏汉平行语料为政府文献领域，维汉平行语料为新闻领域，日汉英多语言翻译平行语料为专利领域，汉语单语数据为新闻领域。几个评测任务的语言对的特点既有相似又有不同，为此对每项评测任务的数据均采用通用预处理和特定预处理两阶段处理法。通用处理阶段的的预处理操作为：汉语繁体到简体转化、邻接相似句的过滤、语言 token 占比过滤、特殊字符过滤、句子长度及句长比过滤。其中，邻接相似句的过滤指的是对语料库中相邻位置的句子判断其相似度，前后相邻两句的 Dice 相似度超过 0.9 时，将会删除后一句，以确保语料库的高质量。由于语料库中存在一定比例的句对质量较差，因此采用语言 token 占比[11]方法来过滤，将句子所有词中的语言 token 所占比例的阈值均设置为 0.1，小于该阈值的句对进行剔除。特定预处理阶段，汉英翻译评测、日汉英多语言翻译评测、蒙汉藏汉维汉翻译评测目标端语言（汉语）的分词均采用词法工具 Urheen[2]来实现，句子长度过滤区间设置为[1, 50]，句长比区间为[0.2, 5]，将源语言端与目标语言端句子的句长或句长比不在设置区间的句对进行剔除。蒙汉、藏汉、维汉翻译评测源语言端（蒙语、藏语、维语）的分词均采用词法工具 polyglot[3]来实现。句子长度过滤区间设置为[1, 100]，句长比区间为[0.1, 10]，将源语言端与目标语言端句子的句长或句长比不在设置区间的句对进行剔除。各项评测数据预处理前后句对数量变化如表 1 所示：

表 1 数据预处理结果

Tab.1 Data preprocessing results

评测任务	原始句数	过滤后
英汉	902w	861w
日汉英多语言	300w	288w
蒙汉	269462	261257
藏汉	162069	158972
维汉	170061	162454

4.1.2 单语数据

CCMT'2020发布的单语语料库共有662904篇新闻文章，大约1100万词汇，由于汉英翻译评测、日汉英翻译评测任务的训练集规模已经够大，所以本评测只在训练集规模较小的蒙汉、藏汉、维汉翻译评测任务上使用了单语数据增强训练集。由于用同样的单语语料对三种语言对的机器翻译增强并不能达到最好的语料增强效果，因此本评测根据三种语言对的特点分别对单语语料进行过滤和筛选，

[2] https://www.nlpr.ia.ac.cn/cip/software.html
[3] https://github.com/aboSamoor/polyglot

对提供的汉语新闻篇章，通过标点符号拆分成句，然后根据不同句长比例进行筛选过滤并相继采用通用预处理和特定预处理两阶段处理法，蒙藏汉最终获得的用于回译的汉语单语数据数量如表2所示：

表 2 单语数据数量

Tab.2 Number of monolingual data

语言对	数量
蒙汉	4000000
藏汉	2761421
维汉	5000000

对于筛选得到的单语数据，本评测采取了回译策略构造伪平行语料来增强机器翻译结果。根据CCMT'2020提供的蒙汉、藏汉、维汉平行语料构建汉蒙、汉藏、汉维神经机器翻译模型，继而通过该模型将筛选得到的汉语单语语料翻译成对应的少数民族语言。最终将回译得到的伪平行语料与CCMT'2020提供的预处理后的高质量双语平行语料进行混合训练，以提高蒙汉、藏汉、维汉的机器翻译质量。

4.2 实验结果

表 3-6 为本次评测提交的实验系统在英汉翻译评测、蒙藏维少数民族语言翻译评测、日汉英多语言翻译评测以及汉英平行语料过滤任务在开发集上的评测结果。其中，英汉翻译评测任务中基线是未利用单语数据，也未采用其他策略的系统；基线+平均是未利用单语数据，但采用了模型平均策略的系统；主系统是未利用单语数据，但采用了模型平均和集成策略的系统。蒙汉、藏汉、维汉翻译评测任务中基线是未利用单语数据，也未采用其他策略的系统；基线+回译+平均是利用了单语数据且采用了模型平均策略的系统；主系统是利用了单语数据且采用了模型平均和集成策略的系统。日汉英多语言翻译评测任务中基线是未利用单语数据，也未采用其他策略的系统；基线+平均是未利用单语数据，但采用了模型平均策略的系统；基线+平均+集成是未利用单语数据，但采用了模型平均和模型集成策略的系统；主系统是未利用单语数据，但采用了模型平均、模型集成和重排序策略的系统。汉英平行语料过滤任务中，使用评测方提供筛选脚本分别将排序后语料库中的前 1 亿个词以及前 5 亿个词对应的句对筛选出来（词数统计均以英文端分词后为标准），训练机器翻译模型在 CCMT2019 数据集上进行 bleu 打分。各任务结果如下表所示：

表 3 英汉翻译评测在开发集上的结果

Tab.3 The results of the English-Chinese translation evaluation on the development set

模型	BLEU
基线	23.94
基线+平均	24.25
主系统	25.86

表 4 蒙藏维少数民族语言翻译评测在开发集上的结果

Tab.4 The results of the Mongolian, Tibetan, and Uyghur minority language translation evaluation on the development set

模型 语言	基线	基线+回译+平均	主系统
蒙汉	41.11	45.73	47.09
藏汉	17.82	26.93	27.23
维汉	22.24	28.33	28.86

表 5 日汉英多语言翻译评测在开发集上的结果

Tab.5 The results of the Japanese-Chinese-English multilingual translation evaluation on the development set

模型	BLEU
基线	37.5
基线+平均	38.10
基线+平均+集成	38.85
主系统	39.13

表 6 汉英平行语料过滤任务在开发集上的结果

Tab.6 The results of the Chinese-English parallel corpus filtering task on the development set

词规模	BLEU-4
1 亿	18.4
5 亿	23.7

分析实验结果可知，英汉翻译评测任务中，在采用模型平均和模型集成策略后，翻译质量较基线系统提升 1.92 BLEU 值；蒙藏维少数民族语言翻译评测任务中，在采用了单语数据及模型平均和模型集成后，蒙汉方向 BLEU 值高出基线系统 5.98，藏汉方向 BLEU 值高出基线系统 9.41，维汉方向 BLEU 值高出基线系统 6.62；日汉英多语言翻译评测任务中，在采用模型平均、模型集成和重排序策略后，翻译质量比基线系统高 1.63 BLEU 值。因此可以得出结论：（1）模型平均、模型集成、重排序对翻译质量提升有一定的帮助；（2）单语数据回译构造伪平行语料有利于翻译质量的提升；（3）数据的预处理精度对翻译质量有很大的影响；（4）多维度多相似度融合的方法有助于过滤语料库，筛选更高质量的平行句对。

5 总结

本文详细介绍了中国科学技术信息研究所在 CCMT'2020 中各任务上使用的主要技术和方法。总结来说，本次评测在模型上使用了基于自注意力机制的 transformer 的架构。在数据预处理方面，探索了多种语料过滤方法。在译文输出过程中，采用了模型平均、集成、译文重排序的策略，在语料过滤中，采用了 bleu 相似度、语言模型打分、句子对齐打分相结合的方法。实验结果证明，这些方法能够有效提高翻译及过滤质量。

由于时间有限，本次评测中还有许多方法没有尝试，在评测过程中发现了一些问题和不足，采

用的翻译模型仍存在很大提升空间。在今后的研究中期望能够学习更多先进技术，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. 2017: 5998-6008.

[2] 李北，王强，肖桐，等. 面向神经机器翻译的集成学习方法分析[J]. 中文信息学报，2019,33(03):42-51.

[3] 王星，熊德意，张民. 神经机器翻译[EB/OL]. (2016-11- 04)[2019-07-14]. http://www.cipsc.org.cn/qngw/?p=953.

[4] 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.

[5] Zaidan O. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems[J]. The Prague Bulletin of Mathematical Linguistics, 2009, 91(1): 79-88.

[6] Ma X. Champollion: A Robust Parallel Text Sentence Aligner[C]//LREC. 2006: 489-492.

[7] Zaidan O. Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems[J]. The Prague Bulletin of Mathematical Linguistics, 2009, 91(1): 79-88.

[8] Vaswani A, Bengio S, Brevdo E, et al. Tensor2tensor for neural machine translation[C]//Proceedings of the 13th Conference of the Association for Machine Translation in the Americas,2018:193-199.

[9] Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks[C]//Advances in neural information processing systems. 2016: 1019-1027.

[10] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//Proceedings of the ACL, 2016:1715-1725.

[11] Lu J, Lv X, Shi Y, et al. Alibaba submission to the WMT18 parallel corpus filtering task[C]//Proceedings of the Third Conference on Machine Translation: Shared Task Papers. 2018: 917-922.

ISTIC Evaluation Technical Report for CCMT' 2020

Liu Wenbin, Wei Jiaze, Wu Zhenfeng, Pan You, He Yanqing*

(Research centor of information theory and methodology, Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: This paper describes an overview and the technical details adopted by Institute of Scientific and Technical Information of China (ISTIC) to participate in the 16th China Conference on Machine Translation (CCMT'2020) evaluation tasks. In the evaluation, ISTIC participated in six evaluation tasks, including Chinese-English news machine translation, Mongolian-Chinese daily language machine translation, Tibetan-Chinese government literature machine translation, and Uyghur-Chinese news machine translation, Japanese, Chinese, and English multilingual patents machine translation and Chinese-English parallel corpus filtering tasks. Thise report describes the model framework, datasets pre-processing methods and decoding strategies. The report gives the performance of the system on the evaluation dataset under different settings and conducts a comparative analysis.

Keywords: *Machine translation; CCMT'2020; Neural network; Self-attention*