第十四届机器翻译研讨会中科院自动化所技术报告

刘宇宸，闫璟辉，张家俊，宗成庆
（中国科学院自动化研究所， 北京 100190）

摘要：本文主要介绍中国科学院自动化研究所参加 CWMT2018 机器翻译系统评测的总体情况。在本次评测中，我们参加了汉英新闻领域机器翻译项目。报告将主要阐述本次参评系统采用的神经机器翻译系统框架、数据处理、译码策略等技术，以及它们在评测数据上的性能表现，同时对翻译结果进行了比较和分析。

关键词：神经机器翻译；汉英翻译；自注意力机制

CASIA Technical Report for the CWMT208

Yuchen Liu, Jinghui Yan, Jiajun Zhang, Chenqing Zong
(Institute of Automation, Chinese Academy of Science, Beijing 100190)

Abstract: This paper describes an overview of CASIA technical report for CWMT2018. In the evaluation of this year, CASIA participates in Chinese-to-English news domain translation task. The report mainly describes our neural machine translation framework, data processing and decoding strategies, and the performance in the evaluation data set. Additional, we conduct the detailed analysis and comparisons on the translation result.

Key words: neural machine translation; Chinese-to-English; self-attention

1 引言

CWMT2018 评测由翻译任务、多语言翻译任务和翻译质量评估任务三个任务组成。其中翻译任务评测包含汉英新闻、英汉新闻、蒙汉日常用语、藏汉政府文献、维汉新闻 5 个方向的评测项目。中国科学院自动化研究所在本届机器翻译评测中参加了汉英新闻领域机器翻译。本文将介绍我们在此次翻译评测中使用的翻译系统框架、数据处理、译码策略等技术，并对系统的性能表现和实验结果进行比较分析。

2 系统介绍

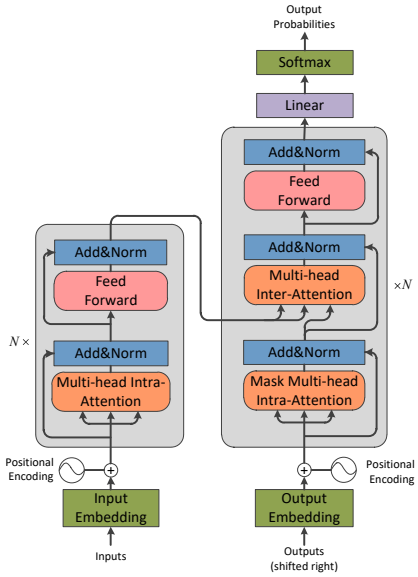
不同于往年[1][2],本次评测使用的系统是基于自注意力机制的 Transformer 模型。该模型由 Vaswani 提出，已在多个语言对上达到当前最优效果[3]。我们在此基础上进行了修改和调优，下面我们将对系统进行简单的介绍。

2.1 模型结构

Transformer 模型结构如图一所示。该模型完全基于注意力机制实现。这一模型在实现算法并行性、加快模型训练速度的同时，还能进一步提高翻译质量。具体来说，模型采用的是编码器-解码器的模型框架，编码器和解码器由 N 个层块堆叠而成。其中，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全连接前馈神经网络。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，在这两个模块之间另外加入了一个与编码器输出层相连的多头注意力模块，用于关注源端的信息。为了避免层数过多导致模型难以收敛的问题，编码器与解码器都使用了残差连接和层级正则技术。编码器和解码器的输入都加入了位置编码向量，使得模型更好得学习数据的序列信息。此外，在解码器的输入中加入了标志符，以避免解码器关注到还未产生的序列信息。

2.2 自注意力机制

自注意力函数的作用是将查询以及一个键-值对的集合映射到输出上。这里的查询、键、值和输出都是向量。输出实际上是值的加权和，权重是由查询和键采用点积的方式计算得到。具体来说，在编码器和解码器中，查询、键、值都来自上一层的输出，而在解码器与编码器之间的多头注意力模块中，查询来自解码器的上一模块的输出，键、值来自编码器最顶层的输出。这里的自注意力机制采用的是多头注意力机制。它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得关注到不同位置不同表示子空间的特征信息。



图一 神经机器翻译模型结构

3 数据处理

3.1 语料预处理

本次评测使用了 CWMT2018 提供的汉英新闻领域双语平行数据。我们对语料进行了一系列预处理操作，关键的预处理步骤如下：

- 全角转半角
- 转义字符处理
- 分词与 token
- 大小写转换
- 长句切分

其中，中文分词工具采用实验室开发的词法工具 Urheen[1]，英文 tokenize 使用的是开源系统 Moses 提供的脚本工具[2]。

众所周知，训练语料的质量对于最终的翻译模型的结果有很大影响。为减少平行语料中的噪声，缓解低质量的语料对翻译质量造成的影响，我们进行了较为细致的语料过滤操作，包括：

- 删除重复的句子
- 删除长度大于 120 个词和小于 3 个词的句子
- 删除源端与目标端的长度比小于 0.5 或大于 2 的句子
- 删除双语语料中汉语端包含非中文字符和其他特殊词组的句子
- 我们使用 fast_align[3]工具学习双语语料的词对齐分布，并删除了双语句子中词对齐比例低于 0.65 的句子

3.2 数据选择

除了 CWMT 提供的训练数据外，WMT 还提供了 UN Parallel Corpus V1.0 和 News Commentary v13 的双语平行数据。这部分数据量较大，受限于计算资源和时间限制，我们仅从中选取了一小部分语料用于后续的模型训练。同时我们认为并不是所有数据都对模型性能的提升起到同等的作用。Wees 等人比较了两种数据选择策略，分别是静态数据选择和动态数据选择[4]。其基本思想都是通过选择与评测领域更相关的数据作为训练数据，以达到减少数据规模并提升模型性能的目的。我们在此处借鉴了静态数据选择策略进行语料的选择。

首先，从双语平行语料中采样出汉语句子和英语句子分别训练了一个汉语的语言模型和一个英语的语言模型，作为通用领域的语言模型。同时使用开发集的单语句子训练了另外两个语言模型，作为评测领域的语言模型。对于训练语料中的每个双语句对，分别使用评测领域语言模型和通用领域语言模型计算困惑度。一般认为，困惑度越高的句子与训练语料的分布越不一致，噪声越大。我们将两个语言模型的困惑度相加进行排序，并从中选择困惑度得分较低的平行句对作为训练语料。最终得到的训练语料大小如表一所示。

3.3 单语数据利用

Zhang 等人和 Sennrich 等人提出了一种使用单语句子生成伪平行句对的数据增强技术，可以有效扩充训练语料，提升翻译质量[5][6]。我们采用了这一技术，并利用 CWMT 提供的大量的汉语和英语单语数据，构造了一批伪平行句对。

我们首先利用双语平行语料训练一个汉语到英语和一个英语到汉语的神经翻译模型，分别使用这两个模型将汉语单语句子和英语单语句子翻译到对应的语言，构成伪平行句对。为保证训练数据的质量，我们使用上文提到数据过滤方法对伪平行句对进行了进一步的过滤。最终使用的单语句子规模如表一所示。伪平行句对可以

[1] https://www.nlpr.ia.ac.cn/cip/software.html
[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
[3] http://github.com/clab/fast_align

与双语平行句对混合训练，也可以在训练好的模型基础上进行再训练，得到更好的翻译模型。我们分别采用了这两种方式训练模型，并将它们用到了后续的集成译码。

评测项目	CWMT 语料	UN 语料	汉语单语	英语单语
汉英新闻	8.9M	5.1M	4.1M	3.8M

表一 训练语料大小

3.4 bpe 切分与词表共享

为避免训练语料出现过多的集外词，Sennrich 等人提出了 BPE 算法，将词语划分为更小细粒度的子词表示，可以有效减少稀有词的数量[7]。我们对汉语句子和英语句子分别学习 BPE 模型[4]。不同的是我们在模型中使用了参数共享技术，我们按照词频从高到低的顺序分别生成中英文的词表，具有同样词频的中文子词和英文子词共享同一个词向量。最终，源语言的词向量、目标语言的词向量以及目标端的 softmax 层共享一套参数。这一方式可以有效减少参数规模，同时也可以提升翻译质量。

4 译码策略

4.1 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。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。因为模型训练采用的随机梯度下降法，每次仅对一个 mini-batch 的样本进行优化，造成模型参数更适应于这一批的样本。通过模型平均的方法可以减少这种不稳定性，使得模型参数更加鲁棒。

4.2 集成译码

在神经机器翻译中，模型集成可以将多个小模型集成到一个统一的大模型中进行译码操作。具体来说，它是在预测当前时刻目标端语言单词的时候，将多个小模型的输出单词的概率分布进行加权平均，来联合预测当前目标语言单词。

用于做集成译码的模型可以是同一模型在训练的不同时刻保存的模型，也可以是模型结构相同但是参数初始化方式不同的模型，甚至是模型结构和初始化方式均不同的模型。一般而言，结构和初始化方式均不同的模型通常更具有差异性，能够带来更大的提升[8]。

4.3 重打分

神经机器翻译通常采用从左到右的译码方式，面临着不平衡输出问题[9]。而且一旦译文前半段产生错误，后续很难再产生正确的结果。我们在开发集上发现了一个有趣的现象。我们首先利用柱搜索方法产生 n-best 的候选译文，然后对每个候选译文计算句子级别的 BLEU 得分，继而选择得分最高句子作为最终输出。这种方式得到的结果可以被看作是模型输出的上界。通过这种方式得到的最终结果，与直接选取柱搜索得分最高的译文相比，其 BLEU 得分高出超过 10 个点。因此我们认为，模型可以产生好的候选译文，但是没法被很好的选出来。

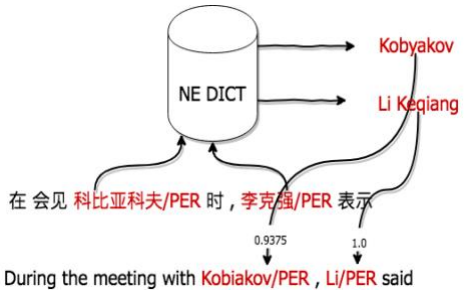
为此，我们尝试了多种重打分的方法。我们首先选择了一系列模型对候选译文进行打分，这些特征包括反向（英到汉）的模型、用于做集成的多个模型、目标端从左到右的模型、目标端的单语神经语言模型和 n-gram 语言模型，以及译文和输入的长度比、目标端和源端的词汇化翻译概率等。之后利用这些特征，我们使用了线性回归、序列排序和最小贝叶斯风险解码等方法对候选译文进行重打分[10][11]，最后选择得分最高的译文作为最终输出译文。

4.4 命名实体替换

新闻领域的语料经常伴随着大量的命名实体（人名，地名，组织名），这些命名实体一般出现的次数较少，尤其以人名为甚。神经网络翻译模型在训练过程中会将这些出现次数较少的词语标记为集外词，而出现漏翻或错翻的现象。由于训练集成模型较多，综合时间等因素，我们选择在后处理阶段对命名实体进行替换操作。首先我们使用Dong等人和Lample等人的方法[12][13]，分别训练了英语和汉语两套命名实体识别模型，并分别对训练数据进行命名实体标注，之后利用词对齐工具对标记标签的训练数据进行词对齐训练，得到一个源端和目标端对齐的命名实体对词典。这种从原训练数据得到的命名实体对词典具有很强的领域相关性，这对

[4] https://github.com/rsennrich/subword-nmt

于解决命名实体翻译的多译性起到了一定帮助。



图二 命名实体替换过程

我们对翻译结果中的命名实体采用“识别-替换”的方式进行处理，如图二所示。我们使用同样的方式对test数据以及翻译结果分别进行命名实体标注,并通过利用对齐训练语料库得到的命名实体对词典进行对比并替换。首先将test数据中所标记的命名实体在命名实体库中找到其对应的英文翻译，再将其与所对应的翻译结果中的命名实体依次对比相似度，相似度采用Li等人的方法进行计算[14]。对于标记好的训练数据，我们采用基于采样的词对齐工具 anymalign[5]生成其phrase-table[15]，再按以下规则抽取：源端与目标端都带有同种类命名实体标识；源端与目标端的翻译概率大于一定的阈值。除了上诉情况，测试集中还含有大量的数字命名实体，如日期，数字，百分比等，经观察，其格式并不统一，且无特定规律。我们使用了统一的格式来对这些数字命名实体进行表示。

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
primary-a	0.2528	0.2679	7.1472	0.5853	0.6432	0.4752	0.2581	0.2397	0.6660
contrast-b	0.2468	0.2634	7.0700	0.5817	0.6546	0.4839	0.2546	0.2379	0.6735

表二 汉英新闻领域机器翻译结果

5 实验与结果

5.1 参数设置

本次评测系统在开源项目tensor2tensor[6]上加以修改。参数设置如下，每个模型使用 3 块 GPU 核进行训练，每个 batch 大约含有 4096 个中文 token 和英文 token，模型训练 20 万 steps，每 30 分钟保存一次模型用于之后的模型平均。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。dropout 设为 0.3，我们使用 Adam 梯度优化算法，初始学习率为 0.1，warmup 设为 8000。训练语料汉英两端均采用 BPE 切分，选取中文词表大小为 42K，英文词表大小为 32K，两者的词表不共享但是共享词向量。

5.2 实验结果

表二是本次提交的汉英新闻领域测试集上评测结果，该结果采用的评测指标是大小写敏感的。其中 primary-a 是主系统的结果，使用了 4 个模型集成，beam size 大小等于 50，经过线性回归模型重打分的方法得到的最终结果。contrast-b 是对比系统的结果，使用了重打分方法中最小贝叶斯风险解码策略解码得到。可以看出不同的重打分策略对结果的影响有所不同，使用多特征融合的方法相比最小贝叶斯风险解码有更大的提升。

5.3 实验分析

在本小节中我们将针对上文提及的方法和策略对于翻译质量的影响分别加以分析。实验分析部分采用的评测指标是大小写不敏感的 BLEU，使用评测工具 multi-bleu.perl[7]。这里的实验结果为开发集上的结果，开发集选择的是 CWMT2017 汉英新闻领域测试集，共 1000 个句子。

5.3.1 基本分析

在开发集上的实验结果如表二所示。其中 baseline 是仅使用平行语料训练得到的单模型的结果，+synthetic 为使用平行语料加上伪平行语料后的结果，+average 是指经过模型平均后的结果，

[5] https://anymalign.limsi.fr/
[6] https://www.github.com/tensorflow/tensor2tensor
[7] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

+ensemble 是经过模型集成后的译码结果，+reranking 则是在集成译码基础上采用重打分策略后的翻译结果，+named entity 是指经过命名实体替换得到的结果。

系统	BLEU
baseline	25.52
+synthetic	28.94
+average	29.10
+ensemble	30.18
+reranking	31.23
+ named entity	31.41

表三 基本方法评测结果

从上述实验结果可以看出，伪平行数据使用、模型平均、集成译码、重打分策略和命名实体替换对翻译质量的提升均有一定的帮助。其中加入伪平行数据的方法对结果有显著提升，证明了该方法的有效性。

虽然重打分策略对结果有一定提升，但是与模型最优输出上界相比还有很大的差距。如何选择特征对参考译文进行表征并加以选择是一个值得研究的问题。

命名实体替换带来的 BLEU 值提升较小，这是由于经过 BPE 处理后的翻译模型本身对于命名实体已经具备不错的翻译能力，同时这与数据本身包含的命名实体数量也存在着一定的关系。但是，在对开发集结果进行观察后我们发现，一些关键的人名的翻译上，例如“马云”由 “Ma Yun”被合理地替换成了“Jack Ma”,这证明了我们从训练语料中提取出来的命名实体对词典确实具有很好的领域相关性并起到了一定的纠错作用。

5.3.2 长度惩罚分析

我们在实验中发现，模型产生的译文较为简短精炼，很多译文长度均短于参考译文。即使翻译质量不差，但受到长度惩罚因子的影响，最后 BLEU 得分较低。为此，我们尝试在柱搜索算法中，引入长度正则因子，使模型倾向于产生长度更长的句子。表四给出了在不同长度正则因子α下，单模型输出的长度惩罚因子和 BLEU 得分情况。可以看出随着长度正则因子的增大，译文长度在不断增长，与参考答案的长度比也不断提升。而译文质量的变化是先提升后下降，这说明过大的长度正则因子可能会导致柱搜索中的得分偏离而无法选择出正确的单词。根据实验结果，我们在系统中使用的长度正则因子是 1.3。

α	1.0	1.1	1.2	1.3	1.4	1.5
length	0.953	0.957	0.963	0.969	0.976	0.988
BLEU	28.79	28.85	28.88	28.94	28.73	28.03

表四 长度惩罚因子对翻译效果的影响

5.3.3 大小写转换方法分析

由于最终提交的系统采用大小写敏感的评测工具进行打分，而我们的模型输出文本均为小写。为此，我们尝试了两种方法进行后处理操作将小写结果转换为大小写混合的文本输出。我们使用 transformer 模型训练了一个小写转大小写混合的模型，使用 moses 提供的 recase[8]脚本训练另一个模型。此处使用 CWMT 提供的工具计算 BLEU4-SBP 得分。从结果可以看出，相比于 recase 训练的模型，我们的模型在大小写转换的正确性上有一定的提升。

System	Transformer	moses
BLEU4-SBP	29.97	29.44

表五 不同大小写转换方法

6 总结

本文介绍了中科院自动化研究所在本次 CWMT2018 汉英新闻领域评测任务上使用的主要技术和方法。总结来看，我们在模型上使用了基于 self-attention 的 transformer 的架构，使用了数据选择策略和伪平行数据方法，在译码策略上使用了模型平均、模型集成、重打分。实验结果证明，这些方法能够有效提高翻译的质量。

受限于时间和计算资源，我们还有许多方法没有尝试，最终翻译结果与其他系统相比也有一定差距。通过本次评测，我们发现了一些不足和问题，我们的翻译模型和系统仍存在很大提升空间。在今后的研究中我们期望能够学习各方先进技术，为提升我国的机器翻译水平贡献绵薄之力。

[8] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/recase.perl

致谢

在此次评测中，中科院自动化所模式识别国家重点实验室的很多老师和同学付出了艰辛的劳动，给予了很多工作上和精神上的支持。在此对他们表示衷心地感谢！

参考文献

[1] 周龙, 王亦宁, 赵阳, 张家俊, 宗成庆. 第十三届机器翻译研讨会中国科学院自动化研究所技术报告. 2017. 第十三届全国机器翻译研讨会论文集.

[2] Jiajun Zhang, Feifei Zhai, Yu Zhou, Kun Wang, Yufeng Chen, Mei Tu, Xiaoqing Li and Chengqing Zong. 2013. RoleTrans: Multilingual machine translation system in CASIA . In Proceedings of CWMT2013.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2017. Advances in Neural Information Processing Systems. 5998-6008.

[4] Van der Wees M, Bisazza A, Monz C. Dynamic Data Selection for Neural Machine Translation. 2017. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 1400-1410.

[5] Zhang J, Zong C. Exploiting source-side monolingual data in neural machine translation. 2016. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 1535-1545.

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

[7] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. 2016. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 1715-1725.

[8] Sennrich R, Birch A, Currey A, et al. The University of Edinburgh's Neural MT Systems for WMT17. 2017. arXiv preprint arXiv:1708.00726.

[9] Liu L, Utiyama M, Finch A, et al. Agreement on target-bidirectional neural machine translation. 2016 In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 411-416.

[10] Shen L, Sarkar A, Och F J. Discriminative reranking for machine translation. 2004. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004.

[11] Shu R, Nakayama H. Later-stage Minimum Bayes-Risk Decoding for Neural Machine Translation. 2017. arXiv preprint arXiv:1704.03169.

[12] Dong C, Zhang J, Zong C, et al. Character-based lstm-crf with radical-level features for chinese named entity recognition. 2016. Natural Language Understanding and Intelligent Applications. Springer, Cham, 239-250.

[13] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition. 2016. arXiv preprint arXiv:1603.01360.

[14] Xiaoqing, Jiajun Zhang, Chengqing Zong. Neural Name Translation Improves Neural Machine Translation. 2016. arXiv preprint arXiv:1607.01856.

[15] Lardilleux A, Lepage Y. Sampling-based multilingual alignment. 2009. In Recent Advances in Natural Language Processing. 214-218.