第十三届机器翻译研讨会中科院自动化所技术报告

周龙，王亦宁，赵阳，张家俊，宗成庆

（中国科学院自动化研究所，北京 100190）

{long.zhou, yining.wang, yang.zhao, jjzhang, cqzong}@nlpr.ia.ac.cn

摘要：本文将主要介绍中国科学院自动化研究所参加 CWMT2017 机器翻译系统评测的总体情况。在本次评测中，我们参加了 6 个评测项目中的 2 个子项–汉英新闻领域机器翻译和日汉专利领域机器翻译。报告主要阐述本次参评系统采用的神经机器翻译系统框架，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；日汉翻译

CASIA Technical Report for the CWMT2017

Long Zhou, Yining Wang, Yang Zhao, Jiajun Zhang, Chengqing Zong

(Institute of Automation, Chinese Academy of Science, Beijing 100190)

Abstract: This Paper describes an overview of CASIA technical report for CWMT2017. In the evaluation, CASIA takes part in two out of six translation tasks, including translation on Chinese-to-English news domain and Japanese-to-Chinese patents domain. The report mainly describes our neural machine translation (NMT) framework and the performance in the evaluation data set. Additionally, we conduct detailed analysis and comparisons.

Key words: neural machine translation; Chinese-to-English; Japanese-to-Chinese

1 引言

CWMT2017 一共包含了 6 个评测方向：汉英新闻领域机器翻译（CE）、英汉新闻领域机器翻译（EC）、蒙汉日常用语机器翻译（MC）、藏汉政府文献机器翻译（TC）、维汉新闻领域机器翻译（UC）以及日汉专利领域机器翻译（JC）。不同于往年[1][2]，中国科学院自动化研究所在本届机器翻译评测中参加了两个项目：汉英新闻领域机器翻译和日汉专利领域机器翻译。

本文将主要介绍我们采用的神经网络机器翻译系统框架、主要技术以及系统在各个翻译任务上的性能表现。最后我们将对实验结果进行比较和分析。

2 系统简介

本次机器翻译评测中我们使用的是基于神经网络的机器翻译系统。系统采用 C++/CUDA 编写，下面我们将对系统进行简单的介绍。

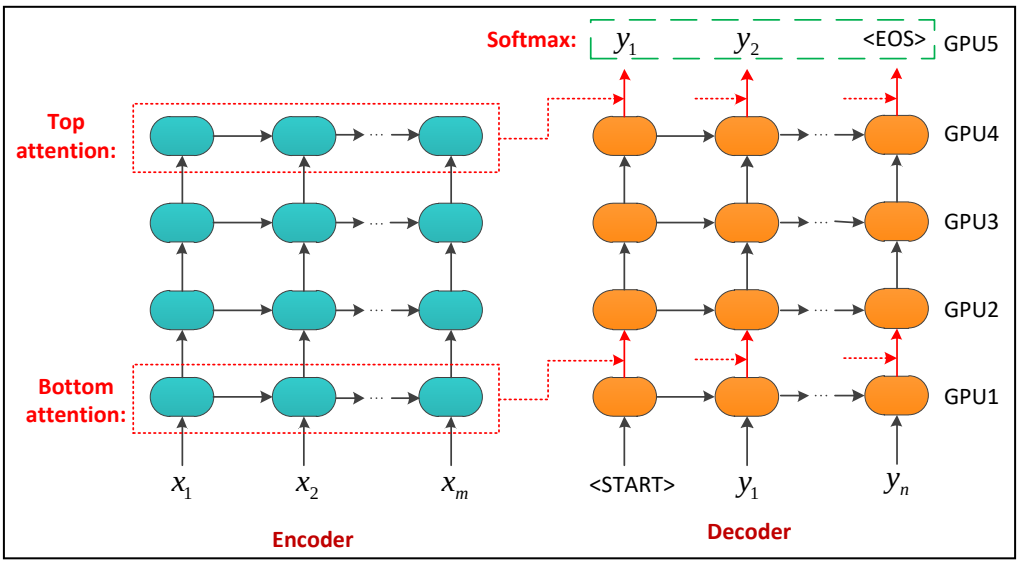
2．1 模型结构

我们的模型为基于注意力机制的端到端神经网络机器翻译模型。神经机器翻译模型（NMT）由两个长短时记忆网络（LSTM）连接而成。它包含三个部分：一个编码器，一个译码器，和两个注意力机制网络。编码器将变长的源端句子转化为一个上下文矢量表示。根据编码得到的矢量表示，译码器每个时刻产生一个目标单词，直到产生结束符为止。两个注意力机制模型一个连接着编码器与译码器的最顶层，一个连接着编码器和译码器的最底层。注意力模型让译码器解码时在最顶层和最底层分别关注不同的源语言单词。模型结构如图一所示。

2．2 平行注意力机制模型

注意力机制模型对目前的神经网络机器翻译起着至关重要的作用，它能使模型在翻译目标语言单词时更注重源端对应的单词。为了提高注意力机制模型的有效性和鲁棒性，我们提出了平行注意力机制模型。

系统包含两个注意力机制模型，一个接着编码器的顶层和译码器的顶层，注意力计算方式采用点积[3]。另一个注意力机制模型连接着编码器的最底层和译码器的最底层。这里以底层注意力机制为例进行介绍，首先

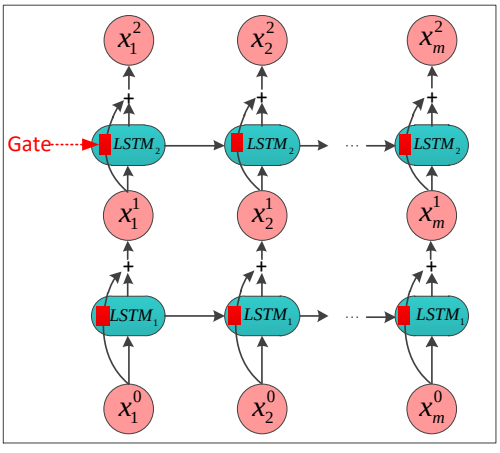


图一 神经机器翻译模型结构

由目标端最底层当前时刻的隐层状态与源端最底层的隐层状态进行注意力操作得到注意力权重，然后将源端隐层状态加权求和得到注意力矢量表示，最后将这个矢量表示和目标端当前时刻的矢量表示送入一个前馈神经网络，输出即下一层 LSTM 的输入。

2．3 门限残差网络

神经网络模型的深度对其效果有着显著的影响，然而由于梯度消失和爆炸等问题，导致深层的机器翻译模型无法达到理想的翻译质量。我们在残差网络的基础上，提出了一个门限残差网络，有效地增强了模型梯度流的更新。



图二 门限残差网络

模型示例如图二所示，红色的方块表示在残差网络中新加入的门（Gate），用于对跨层连接时的输入进行控制。在模型中，第一层 LSTM 的输入首先与一个门（Gate）按位相乘，其值按位相加到第一层 LSTM 的输出中，相加的和即第二层 LSTM 的输入。这里的控制门类似于 LSTM 的遗忘门，参数通过模型训练更新而来。

3 数据处理

3．1 语料预处理

本次评测中，汉英方向未使用 WMT2017 提供的数据。我们对 CWMT2017 提供的汉英新闻领域和日汉专利领域的所有语料进行了一定的预处理，关键预处理步骤如下：

- 全角转半角
- 转义字符处理
- 长句切分
- 双语数据过滤
- 单语数据过滤
- 分词与 token

其中汉语、日语分词工具采用实验室开发的词法分析工具 Urheen[1]。英语使用 Moses 中的 token 脚本工具[2]。

原始语料中噪声较多，部分平行语料对质量较差，为了缓解低质量的语料对翻译质量造成的影响，我们分别对句长较长、句长比例过于悬殊、词对齐比例过低、未登录词占比过大的句子进行替换和过滤操作。最终的训练语料如表一所示。

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

评测项目	平行语料	源端单语语料	目标端单语语料
汉英新闻	8.9M	5.1M	4.8M
日汉专利	3M	无	7.1M

表一 训练数据

3．2 命名实体识别对齐与翻译

汉英新闻语料中包含大量的人名，地名，机构名。其中，较大部分命名实体出现次数较少，在模型训练中会被当做未登录词进行处理。我们采用类别标签替换的方式，来解决命名实体翻译问题。

在训练阶段，参照 Dong 等人[4]和 Lample 等人[5]的方法[3]，我们分别训练英语和汉语两套命名实体识别模型，并使用该模型对训练语料进行标注。在替换过程中，为了尽可能降低识别错误产生的影响，我们利用实验室标注的大规模命名实体库，当且仅当汉英两端的识别结果是实体库中实体翻译对，才将其进行替换。与此同时，我们使用识别出来的命名实体翻译对训练一个基于字的命名实体翻译系统。

在翻译过程中，我们首先使用命名实体识别模型对测试集进行标注，用相应的标签进行替换，并保留其中的对应关系。翻译结束后，根据对应关系，使用训练集中提取出的命名实体翻译对进行替换，对于未能替换的命名实体，我们使用之前训练完成的基于字的命名实体翻译模型进行翻译，并用翻译好的命名实体替换掉相应的类别标签。

3．3 单语数据利用

CWMT 在汉英翻译方向和日汉翻译方向提供了大量的单语数据。Zhang 等人[6]和 Sennrich 等人[7]提出了一种利用单语数据生成伪训练数据的方法，有效地扩充了训练语料，增强了翻译质量。

经过预处理后的平行语料和单语数据如表一所示。我们首先在平行语料上训练一个 NMT 或者 SMT 翻译系统，然后使用训练好的翻译系统将单语数据翻译成对应语言，即可得到伪平行数据。伪平行数据可以和平行数据混合，或在平行数据训练的模型基础上，进行再训练。得到的模型可用作集成译码（4.2 节）。

3．4 混合字词切分

集外词处理是神经机器翻译的一大难点。Li 等人提出了一种“替换-翻译-替换”的集外词处理方法[8]，有效地保持了句子的语义结构，并明显地改善了模型的训练效果。 Sennrich 等人针对 NMT 的稀有词和集外词问题，提出了一个 BPE 算法[9]，它通过使用更小的细粒度来表示词语从而提高模型处理稀有词和未登录词的能力。

我们这里使用的是 Wu 等人提出的混合字词模型[10]。首先保持一个固定大小的词表，然后将集外词转换为连续字符的序列，并且在每个字符前加上特殊的标签。假设单词“核糖核酸”是未登录词，那么它将被切分为：<B>核 <M>糖 <M>核 <E>酸。

这里的<B>，<M>和<E>分别表示单词的开始、中间和结束。在翻译完成后，只需经过简单的后处理即可将这些特殊字符恢复为单词。

4 译码策略

4．1 网络参数动态调优

受限于训练数据的规模和模型的复杂度，一组固定的神经机器翻译参数不可能覆盖所有的翻译知识，因而无法对所有的测试句子都做到最优。针对这一问题，Li 等人提出了一种动态的句子敏感的参数更新方法[11]。

采用网络参数动态调优方法，我们针对每个测试句子（或每类测试句子），通过从双语语料中实时地检索出相似的子集，然后采用动态调优策略对解码当前句子（或当前类别句子）所需的翻译知识进行实时更新，从而得到其专属的网络参数。

评测中，汉英新闻方向测试句子分为政治、体育、经济等 10 个子类。日汉专利方向针对每个测试句子实时更新网络参数，输出 beam search 大小（这里为 12）个译文，其可作为重打分策略（4.2 节）的候选译文。

4．2 集成译码与重打分

集成译码在目标端生成单词前，将多个模型的概率分布进行平均，去预测当前的目

[3] https://github.com/glample/tagger

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
primary-a	0.2415	0.2532	6.9717	0.5962	0.6493	0.4586	0.238	0.2144	0.6924
contrast-b	0.2137	0.2220	6.6022	0.5766	0.6706	0.4783	0.2296	0.1997	0.7144

表二 汉英新闻领域机器翻译结果

System	BLEU5-SBP	BLEU5	BLEU6	GTM	mWER	mPER	ICT	METEOR	TER
primary-a	0.428	0.44	0.3869	0.7741	0.3897	0.258	0.3772	0.667	0.3823
contrast-b	0.4093	0.4262	0.3733	0.7612	0.4138	0.2781	0.3639	0.6487	0.4001

表三 日汉专利领域机器翻译结果

标语言单词。集成译码已经被证明在神经机器翻译中有着显著的效果[12][3][13]。

在神经网络机器翻译译码中，面临着不平衡输出问题[14]，导致模型译文不能准确表达源语言的意思。针对汉英方向和日汉方向，我们分别训练了两个反向的NMT模型（英汉和汉日）。使用训练得到的反向模型来对正向模型的 N-best 输出进行重打分，选择得分最高的译文作为最终的目标译文。

在本次评测中，我们使用了 8 个模型来进行集成译码，beam search 设置为 50，即对于每个句子可得到 50 个候选译文。另外，使用 4.1 节中的网络参数动态调优方法，每个句子可得到 12 个候选译文。然后，采用 4 个反向模型对候选译文进行重打分，这里的分值等于正向模型的平均分值加上反向模型的平均分值，选择得分最高的作为系统的最终输出译文。

4．3 小规模词汇表译码

译码速度是评价机器翻译系统的重要指标之一。而在实际系统中，巨大的词汇表将耗费大量的计算时间和内存资源。参照 Mi 等人的工作[15]，我们在评测系统中使用了小规模词汇表译码。

我们首先使用 fast_align 对齐工具[4]，对双语训练语料进行词对齐训练，统计得到每个源语言单词对应的N个最高概率的目标语言单词。测试时，针对每一句源语言，我们将源语言单词对应的所有目标语言单词合并起来得到一个集合，将这个集合与前 M 个高频词相加，即得到当前源语言句子的目标语言词汇表。

评测实验中，我们使用 N=50，M=200。对于每一个源语言句子，目标语言词汇表规模将从几万个减少为几千个，在译文质量不下降的情况下，极大提高解码速度。我们在单语数据回翻（3.3 节）、汉英方向、日汉方向译码中都使用了该方法。

5 实验与结果

5．1 实验设置

本次评测系统采用 C++/CUDA 编写，支持多 GPU 并行训练，部分代码借鉴 ZOPH_RNN[5]。如图一所示，我们采用了 4 层 LSTM 结构，每一个隐藏层和 Softmax 层可各在一个 GPU 上运行。模型采用了第 2 章介绍的平行注意力机制模型和门限残差网络来改善翻译质量。

具体地，经过混合字词切分后，汉英方向的词表大小分别为 91K 和 94K。日汉专利方向的源端和目标端词表大小分别 60K 和 56K。词向量的维度和隐层状态维度都设置为 1000。我们使用随机梯度下降法（SGD）来更新模型参数，初始学习率设为 0.1，当开发集上的困惑度不在下降时，学习率减半，模型的最大训练轮数为 15 轮。另外，每一个 mini-batch 有 128 个样本，对每一个层 LSTM 隐层状态进行 dropout，其大小设为 0.2。

5．2 实验结果

表二和表三是本次评测提交的汉英新闻领域和日汉专利领域评测结果。其中 contrast-b 是指只使用平行语料，单个模型，beam size 大小等于 12，得到的翻译结果。primary-a 是指使用了第 4 章中的译码

[4] https://github.com/clab/fast_align

[5] https://github.com/isi-nlp/Zoph_RNN

策略，最终的译文结果。

可以看出译码策略在汉英方向和日汉方向都有 2 到 3 个点的提升。但汉英新闻领域提升大于日汉专利领域，其原因可能是日汉专利领域为限定领域，基本模型就能实现较好的效果。

5．3 实验分析

本小节我们将分析上文提及的方法和策略在翻译质量和速度上的改进。实验分析部分采用的评测指标是大小写不敏感的 BLEU 评测工具 *multi-blue.perl*[6]。日汉方向采用基于词的评价方式。

5.3.1 基本分析

我们从汉英方向和日汉方向开发集随机各选出 1000 句作基本实验分析。实验结果如表四所示，其中 baseline 为在平行语料上训练单模型结果，+synthetic 为加上伪平行数据后单模型结果，+ensemble 是加上集成译码后翻译结果，+reranking 是在集成译码基础上采用 4.2 节中的重打分策略后的翻译结果。

系统	汉英	日汉
baseline	21.94	41.03
+synthetic	23.76	40.58
+ensemble	26.20	41.86
+reranking	27.33	42.55

表四 基本方法评测结果

实验结果表明，伪平行数据、集成译码和重打分策略均有利于翻译质量的提升。但是在日汉专利领域中，加入伪数据时，翻译质量出现下降，我们猜测，日汉方向领域较窄，在平行语料下能取得较理想的翻译质量，加入含有噪声的未平行数据，影响了翻译效果。

5.3.2 参数调优分析

我们对 4.1 节中的网络参数动态调优方法进行分析，Li 等的实验[11]显示，该方法在相似度高的语料上效果优于相似度低的语料。因此，我们在日汉专利领域的开发集和测试集进行了测试。

表五给出了日汉方向开发集和测试集上的参数调优结果。我们发现开发集上的 BLEU 的绝对值大于测试集，并且网络参数动态调优方法在开发集上的提高大于测试集。经过相似度分析，我们认为是因为开发集与训练语料的相似度高于测试集，并且训练语料中有少量开发集句子。

系统	开发集	测试集
Baseline	41.03	37.65
+fine-tune	42.25	38.17

表五 网络参数动态调优

5.3.3 译码速度分析

最后我们对 4.3 节中的小规模词汇表译码进行译码质量和译码速度的分析，实验在 1000 句日汉开发集上进行，表六给出了在单个 GPU 上译码的实验结果。

系统	BLEU	时间(s)	速度(w/s)
Baseline	41.03	472.7	62.8
+Voc.	41.11	157.6	188.5

表六 小规模词汇表译码

表中+Voc.即为使用小规模词汇表译码策略的结果。可以发现，采用该策略后，译码质量出现小量的提升，译码速度是原来模型的 3 倍。在本次评测模型（4 层 LSTM、平行注意力机制、目标端词表 56K、单 GPU 等）下，实现了每秒 188.5 个词的译码速度。

6 总结

本文主要介绍了中国科学院自动化研究所参加 CWMT2017 评测的总体情况，由于时间及计算资源的限制，我们参加了汉英新闻领域和日汉专利领域的评测，取得了比较理想的成绩。在神经网络翻译模型上，我们提出了平行注意力机制模型和门限残差网络。在译码策略上，我们采用了网络参数动态更新、集成译码与重打分、小规模词汇表译码等技术方法。实验结果表明，这些方法极大提高了机器翻译质量。

通过本次评测，我们也发现了一些不足和问题。同时，我们也意识到，我们的翻译模型和系统还有很大的提升空间。我们希望在以后的研究中与国内外同行多交流学习，不断提高我们现有的模型和系统，也为不断提升我国的机器翻译水平贡献绵薄之力。

[6] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.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！并特别感谢实验室周玉老师和向露老师给予的大力帮助和支持。

参考文献

[1] Yu Zhou, Feifei Zhai, Jiajun Zhang, Mei Tu, Yufeng Chen and Chengqing Zong. 2011. Multi-lingual machine translation system-CASIA technical report for CWMT2011 evaluation. In Proceedings of CWMT2011.

[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] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Proceedings of EMNLP2015.

[4] Chuaihai Dong, Jiajun Zhang, and Chengqing Zong. 2016. Character -based LSTM-CRF with radical-level feature for Chinese named entity recognition. In Proceedings of NLPCC2016.

[5] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In Proceedings of NAACL2016.

[6] Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In Proceedings of EMNLP2016.

[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In Proceedings of ACL2016.

[8] Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In Proceedings of IJCAI 2016.

[9] Rico Sennrich，Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In Proceedings or ACL2016.

[10] Yonghui Wu and Mike Schuster and Zhifeng Chen and Quoc V. Le and Mohammad Norouzi, et al. 2016. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXix:1609.08144.

[11] Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. One sentence one model for neural machine translation. arXiv preprint arXiv:1609.06490.

[12] Sebastien Jean, Kyunghyun Cho, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In Proceeding of ACL2015.

[13] Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. arXiv preprint arXiv:1704.06393.

[14] Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Summita. 2016. Agreement on target-bidirectional neural machine translation. In Proceedings of NAACL-HLT2016.

[15] Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Vocabulary manipulation for neural machine transaltion. In Proceedings of ACL2016.