低资源神经机器翻译的关键技术研究

张文博[1,2,3]，张新路[1,2,3]，杨雅婷[1,3*]，董瑞[1,3]

(1.中国科学院新疆理化技术研究所，乌鲁木齐,830011;2.中国科学院大学，北京,100049;3.新疆民族语音语言信息处理实验室，乌鲁木齐,830011;)

摘要:本文描述了中国科学院新疆理化技术研究所参加第16届全国机器翻译大会（CCMT2020）翻译评测任务总体情况以及采用的技术细节。在评测中，中国科学院新疆理化技术研究所提交了两个翻译任务，分别是蒙汉日常用语机器翻译和维汉新闻领域机器翻译；本文使用所提供的平行数据加上大量回译单语数据得到伪平行语料来训练最先进的神经机器翻译系统。然后，本文主要采用了目前已被证明最有效的技术，如微调和模型集成来改善翻译效果，最后在该数据集上进行了详细的对比与分析。

关键词：神经机器翻译；低资源语言；回译;

中图分类号：TP391　　**文献标识码**：A

1 介绍

本文介绍了中国科学院新疆理化技术研究所参加第16届全国机器翻译大会（CCMT2020）翻译评测任务的情况。本文提交了CCMT2020 中2个有关少数民族语言的翻译项目，分别是蒙汉翻译和维汉翻译，这两个翻译方向都属于低资源语言之间的翻译任务。

基于注意力机制的神经机器翻译[1-3]目前已经成为最常见的机器翻译方法。本文在此次评测采用Transformer[3]神经网络机器翻译架构作为基线系统，模型结构如图1所示。为了提高翻译效果，本文还采用了目前已被证明最有效的技术，例如回译、微调和集成翻译等方法。通过尝试多种BPE[4]融合数以及Dropout[5]参数来改善低资源神经机器翻译的效果。本文采用额外的单语数据来扩充训练数据，通过多阶段训练翻译模型来改善翻译效果。最后，我们使用模型平均和集成提升系统的鲁棒性，进一步提升翻译质量。

2 数据及预处理

使用CCMT2020提供的维汉、蒙汉平行语料以及汉语单语语料。其中维汉平行语料是新闻领域数据，蒙汉平行语料是日常用语数据，汉语单语语料是新闻领域数据。

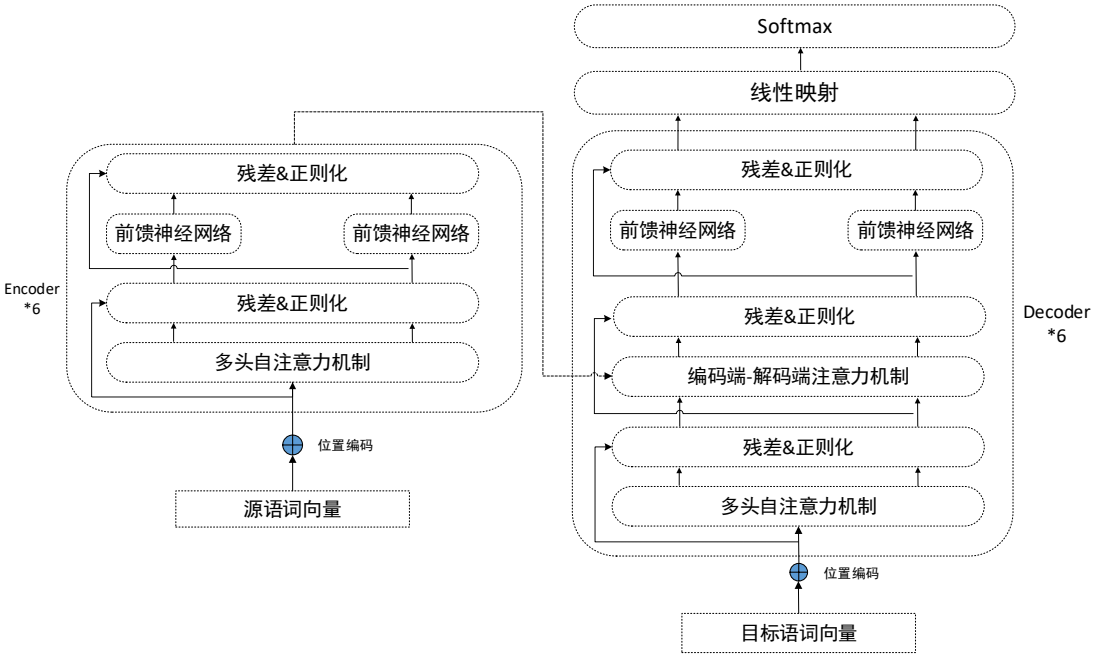

图 1 Transformer 模型结构

Fig.1 Architecture of the Transformer model

2.1 预处理

在该部分中本文过滤了平行语料中重复的句对，将语料中的字母和数字的全角形式转换成半角形式。在维汉翻译、蒙汉翻译两个任务中，本文分别使用 moses 脚本(https://github.com/moses-smt/mosesdecoder)将维吾尔语、蒙古语当做英文进行分词，使用 Jieba 分词工具(https://github.com/fxsjy/jieba)对汉语分词。

2.2 子词化处理

BPE[4]是目前缓解机器翻译任务中未登录词问题[6]的一种普遍采用的方法。本文使用 fastBPE 工具(https://github.com/glample/fastBPE)处理维汉和蒙汉这两个任务，融合数分别为 1000 和 5000。并且联合源语言和目标语言进行 BPE 切分处理，而不是分别处理源语言和目标语言。因此，在机器翻译模型中，翻译模型也在源语言和目标语言之间共享 embedding 矩阵。同时，本文都只使用平行语料作为 BPE 的词频统计语料，因此对汉语单语语料，分别采用维汉平行语料和蒙汉平行语料学习得到的模型进行汉语单语切分。

3 单语数据的使用

为了提升低资源语言对之间的翻译质量，回译[7]等能够有效地利用目标端单语数据的数据增强方法[7-13]已经被广泛采用，因此本文也同样使用回译利用汉语单语数据提升维汉和蒙

汉翻译质量。

3.1 单语语料筛选

和之前的一些研究[12]一样，本文根据单语句子中所有词在平行语料词典中出现的比例挑选和平行语料领域接近的单语数据。为了降低平行语料中低频词的干扰，在统计完平行语料的词典（BPE 之前）之后，删除词典中频率小于 3 的词。对维汉和蒙汉本文都挑选比例大于 0.9 的单语句子用来提升维汉和蒙汉的翻译质量。同时本文还将比例等于 1 的单语数据额外拿出一份作为高领域相似的单语数据，用于下面的分段式训练方法。

3.2 伪平行语料生成

首先训练一个目标端到源端的翻译模型，利用翻译模型将筛选到的目标端单语语料翻译成源端单语语料，翻译得到的源端单语语料和原来的目标端单语语料联合构成伪平行语料。在解码过程中，为了增强数据多样性，本文设置 beamsize 等于 1 并使用 sampling[13]随机采样的方式生成源端数据。迭代地回译[8]可以同时利用源端单数数据和目标端单语数据获得更好的效果，但是本次评测只提供汉语单语数据，因此本文中的伪平行语料是由在平行语料训练得到的反向翻译模型和汉语单语数据一次生成。为了过滤噪声数据，本文删除伪平行语料中长度（BPE 之后）小于 5 或大于 250 以及长度比大于 2 的句对。

3.3 分段式训练方法

回译常见的做法是将伪平行语料和平行语料合并作为新的平行语料用来直接训练翻译模型或者微调[7]。在本次评测中，由于单语语料的规模要远大于平行语料，所以直接合并训练，并不能有效地利用平行语料。过采样这样的方法虽然可以使平行语料和伪平行语料保持接近的比例，但是对平行语料采样次数过多，也容易使得模型对平行语料过拟合。因此本次评测中，本文使用两段式的训练方法。第一阶段本文只使用所有伪平行语料训练翻译模型；第二阶段本文将高领域相似的伪平行语料（由经过筛选得到的高领域相似的汉语单语生成）和平行语料结合，继续训练翻译模型，并且在训练过程中使用过采样使伪平行语料和平行语料保持 1:1 的比例。对维汉翻译任务，本文在第二阶段训练完成之后，进一步使用平行语料微调。

4 模型平均和集成

模型平均和集成都可以提升模型的鲁棒性，有助于进一步提升翻译质量。前者将同一个模型在训练阶段不同时刻保存的参数平均作为最后的模型参数。后者是使用多个模型同时解

码，在生成候选词概率表时，将多个模型生成的概率词表平均作为用于生成下一个词的概率表。本次评测本文多次训练翻译模型和微调，因此在每个训练阶段之后，根据在验证集的表现选择平均最后 5 个、10 个模型或者是在训练过程中的 best 模型作为该阶段的输出模型。对维汉和蒙汉本文分别使用三个不同的随机种子训练了三个模型，最后集成这三个模型对测试集进行解码。

5 实验

5.1 参数设置

使用开源工具 fairseq(https://github.com/pytorch/fairseq)在两块 32G V100 进行机器翻译模型的训练，本文采用 transformer_big 模型作为翻译模型，并且使用 GELU 激活函数。在训练中，使用 fairseq 的 update-freq 将每个 batch 的最大 token 设置为 64000。对维汉翻译任务，dropout[5]被设置为 0.3，activation-dropout 被设置为 0.3，attention-dropout 被设置为 0.2。对蒙汉翻译任务，dropout 被设置为 0.3，activation-dropout 和 attention-dropout 都设置为 0。所有模型都是 Adam 优化器，在训练过程中，维汉第一阶段使用 0.0005 的学习率，蒙汉第一阶段使用 0.0007 的学习率，warmup 设置为 4000。第二阶段，warmup 都被设置为 1000，学习率为 0.0003。对维汉翻译，最后使用固定学习率 0.0001 在平行语料进一步微调。本文所有系统在解码时，都是固定 beamsize 为 12。

5.2 实验结果与分析

对维汉翻译和蒙汉翻译，都分别提交了 3 个结果，其中主系统 primary-a 为使用单语语料分段式训练翻译模型，并使用 3 个不同随机种子训练得到的模型进行集成的结果。对比系统 contrast-c 为没有进行集成的单个模型的结果。对比系统 contrast-b 为只使用平行语料训练得到的单个模型的结果。本文使用 multi-bleu.perl 在验证集上计算不同系统基于字的 BLEU[14]值，结果如表 1 所示。

表 1 不同系统在验证集的测试结果

Tab.1 Test results of different systems in validation set

系统	uy2zh	mn2zh
contrast-b	41.20	65.65
contrast-c	47.26	71.51
primary-a	48.11	72.66

由表 1 可以看出使用汉语单语语料可以显著提升翻译质量，在维汉翻译和蒙汉翻译上

都提升了 6 个 BLEU 左右。最后使用模型集成可以在原有基础上再提升 1 个 BLEU 左右。

为了验证分段式训练方法的有效性，表 2 展现本文在本次评测中筛选之后得到数据规模。表 3 对比不同训练方式在验证集上的实验结果，为了简单起见，本文中所有实验结果中只有表 1 的数据为经过模型平均和集成的结果，其它均采用 best 模型进行对比。

表 2 不同类型语料的规模

Tab.2 The scale of different types of corpus

语料类型	uy2zh	mn2zh
平行语料规模	16.8 万	26.4 万
伪平行语料规模	954.2 万	772.9 万
高领域相似伪平行语料规模	370.8 万	282.5 万

表 3 不同训练策略在验证集的结果

Tab.3 The results of different training strategies in validation set

训练策略	uy-zh	mn-zh
只使用平行语料	39.09	64.75
只使用伪平行语料	44.04	66.01
伪平行语料混合平行语料	45.39	67.59
伪平行语料预训练 + 高领域相似伪平行语料混合经过上采样的平行语料微调	46.60	70.79

由表 3 可以看出语料规模对翻译结果有着至为重要的影响，只使用大量的伪平行语料就可以获得比只使用平行语料更好的性能。在伪平行语料的基础上加上平行语料虽然可以获得进一步的提升，但是提升幅度并不大。而分段式地训练可以在只使用伪平行语料的基础上获得更显著的提升。

5.3 重要参数对比

低资源神经机器翻译往往具有容易过拟合以及存在较多集外词[4]的问题，因此本文使用 transformer_big 模型在平行语料上，通过对 dropout[5,15]和 BPE 融合数这两个参数的调节来缓解这两个问题。表 4 是在维汉翻译和蒙汉翻译上使用 30000 融合数时，不同 dropout 的结果。表 5 是在维汉翻译和蒙汉翻译上使用表 4 中 dropout 最好的设置下，不同 BPE 融合数

的结果。

表 4 不同 dropout 参数在验证集的测试结果

Tab.4 Test results of different dropout parameters in validation set

dropout	0.1	0.3	0.3	0.3
activation-dropout	0	0	0.1	0.3
attention-dropout	0	0	0.1	0.2
uy2zh	30.19	37.91	38.19	**39.09**
mn2zh	52.83	63.07	**64.40**	62.73

表 5 不同 BPE 融合数参数在验证集的测试结果

Tab.5 Test results of different BPE parameters in validation set

BPE 融合数	uy2zh	mn2zh
1000	**40.10**	64.77
5000	39.97	**65.37**
10000	39.92	64.07
30000	39.09	64.40

从表 4 和 5 中可以看出，其中最好的设置和最差的设置之间相差巨大。该结果表明 dropout 和 BPE 融合数这两个超参数对维汉翻译和蒙汉翻译在低资源神经机器翻译中具有较大的影响。对低资源神经机器翻译，合适的 dropout 可以显著提升翻译质量，适当较小的融合数可能会表现更好。本文主实验在蒙汉翻译任务上使用的 dropout 超参数和这里最好的设置略有不同，主要为了更快地验证不同超参数的表现，本文在评测中使用 transformer_base 模型尝试不同的超参数的效果。

6 结论

本文主要通过调节 dropout 和 BPE 融合数两个参数缓解低资源神经机器翻译易过拟合以及存在较多低频词和集外词的问题，并且借助回译，通过分段式训练同时有效地利用单语语料和平行语料资源较大地提升了低资源情况下维汉翻译和蒙汉翻译的质量。

参考文献

[1] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [C] //Proceedings of ICLR 2015 . San Diego: International Conference on Learning Representations, 2015:

1409.

[2] LUONG M, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation [C] // Proceedings of EMNLP 2015 . Lisbon: Association for Computational Linguistics,2015 :1412-1421.

[3] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need [C] // Proceedings of NIPS 2017. Long Beach: Conference on Neural Information Processing Systems,2017:1706 .03762.

[4] 1.SENNRICH R, HADDOW B,BIRCH A. Neural machine translation of rare words with subword units [C] // Proceedings of ACL 2016 . Berlin: Association for Computational Linguistics,2016 :1715-1725.

[5] Hinton G E , Srivastava N , Krizhevsky A , et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer ence, 2012, 3(4): 212-223.

[6] Luong M T, Manning C D. Achieving open vocabulary neural machine translation with hybrid word-character models[J]. arXiv preprint arXiv:1604.00788, 2016.

[7] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[C] // Proceedings of ACL 2016 . Berlin: Association for Computational Linguistics,2016 :86-96 .

[8] Hoang V C D, Koehn P, Haffari G, et al. Iterative back-translation for neural machine translation[C]//Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. 2018: 18-24.

[9] Imamura K, Fujita A, Sumita E. Enhancement of encoder and attention using target monolingual corpora in neural machine translation[C]//Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. 2018: 55-63.

[10] Gulcehre C, Firat O, Xu K, et al. On using monolingual corpora in neural machine translation[J]. arXiv preprint arXiv:1503.03535, 2015.

[11] Graa M, Kim Y , Schamper J, et al. Generalizing Back-Translation in Neural Machine Translation[C] // Proceedings of ACL 2019. Florence: Association for Computational Linguistics,2019 : 45–52 .

[12] Jiajun Zhang, Chengqing Zong. Exploiting Source-side Monolingual Data in Neural Machine Translation[C]// Proceedings of EMNLP2016. Austin, Texas: Conference on Empirical Methods in Natural Language Processing. 2016：1535–1545.

[13] Edunov S , Ott M , Auli M , et al. Understanding Back-Translation at Scale [C] // Proceedings of ACL 2018 . Brussels: Association for Computational Linguistics，2018 : 489-500

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic eval-uation of machine translation. [C] // Proceedings of ACL 2002. Philadelphia: Association for Computational Linguistics, 2002: 311–318 .

[15] 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.

Research on key technologies of neural machine translation with low resources

ZHANG Wenbo[1,2,3],ZHANG Xinlu[1,2,3],YANG Yating[1,3*] ,DONG Rui[1,3]

(1.The Xinjiang Technical Institute of Physics & Chemistry.CAS,Wulumuqi,830011,China; 2. University of Chinese Academy of Sciences,Beijing,100049,China; 3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Wulumuqi,830011,China;)

Abstract： This paper describes the overall situation and technical details of translation evaluation tasks of XinJiang Technical Institute of Physics and Chemistry Chinese Academy of Sciences participating in the 16th China Conference on Machine Translation (CCMT2020). In the evaluation, XinJiang Technical Institute of Physics and Chemistry Chinese Academy of Sciences participated in two translation tasks, namely, Mongolian-Chinese daily language machine translation and Uyghur-Chinese news machine translation; We use the parallel data provided and a large number of monolingual data of reverse translation to train the most advanced neural machine translation system. Then, we have adopted the most effective techniques, such as fine tuning and model ensemble, to improve the translation effect. We made a detailed comparison and analysis on this data set.

Key words： Neural Machine Translation; Low-resource language; Back translation;