CCMT2021 低资源俄汉机器翻译评测报告

刘欢，刘俊鹏，黄锴宇，黄德根*

（大连理工大学计算机科学与技术学院，辽宁 大连 116024）

摘要：介绍了大连理工大学参加第十七届全国机器翻译大会（CCMT2021）俄汉方向上的低资源语言评测任务的总体情况。本次参评系统通过扩充领域内数据和微调的方法完成旅游口语领域的俄汉机器翻译。首先利用多种语料过滤方法得到不同规模的通用领域俄汉语料，然后在不同参数配置的 Transformer 模型上进行预训练和微调的探索，得到最优模型，最后在译码阶段采用模型平均技术生成翻译结果。实验结果表明，该方法可以有效提升低资源场景下的俄汉机器翻译效果。

关键词：神经机器翻译；CCMT2021；低资源语言；mRASP

中图分类号：TP 391　　**文献标志码：** A

1 引言

本次评测采用基于注意力机制的 Transformer[1]神经网络机器翻译架构。针对低资源语言的机器翻译问题，主要利用通用领域的大规模俄汉平行语料[1]进行数据增强以及微调，以改善旅游口语领域俄汉机器翻译性能。在数据预处理方面，针对评测方发布的俄汉口语领域数据，主要采用长度比筛选，针对大规模通用领域数据，采用多种语料过滤方法如长度比、词占比、对齐模型等，来筛选出接近领域内的数据。在模型参数方面，对大规模通用领域语料筛选出来的不同规模的训练集，用不同参数配置的 Transformer 模型训练 NMT 模型，并在双语 NMT 模型和 mRASP[2]预训练模型上进行了微调，根据验证集上的表现选择最优模型。在译文输出过程中，对长度惩罚因子等参数进行调优，最后进行了模型平均。

2 数据处理

2. 1 预处理

首先对 CCMT2021 提供的旅游领域口语俄汉语料及大规模通用领域俄汉语料均进行了预处理，处理过程如下：（1）去除多余空格及非打印字符；（2）全角字符转换为半角字符；（3）对俄语和汉语语料分别进行分词处理，其中俄语使用 Moses 工具，汉语使用 Jieba 分词。数据处理后，为了缩减词表和解决集外词（OOV）的问题，采用 BPE 算法[3]分别将俄汉词语切分成更小粒度的子词，将通用语料和口语语料混合，生成统一的俄语词表和汉语词表。对所有数据进行子词长度及长度比过滤：最大子词长度比设置为 2，最大子词长度设置为 250 个 token。

基金项目：国家科技创新 2030“新一代人工智能”重大项目（2020AAA0108004）；自然科学基金（U1936109）

* **通信作者：**huangdg@dlut.edu.cn

[1] http://statmt.org/wmt21/triangular-mt-task.html

2. 2 数据增强

由于旅游口语领域数据匮乏，直接训练 NMT 模型的翻译效果十分有限，所以数据增强十分有必要。本系统主要从大规模域外数据筛选出靠近域内的数据来进行数据增强。

为了筛选出靠近旅游口语领域的平行句对，对于大规模通用领域俄汉语料，组合地使用以下方法进行数据筛选：（a）Fast-align[4][2]对齐分数过滤：选取 Fast-align 双向对齐分数大于 -400 的句子；（b）语言模型过滤：使用 KenLM[5][3]在俄语测试集上训练语言模型，选取困惑度小于 6300 的句子；（c）词占比[6]过滤：对口语语料的汉语分词后，取字长大于 1、词频大于 2 的词构建目标词典，对通用语料的汉语进行分词，选取所有词均在目标词典中，即目标词占比为 100%的句子。

利用上述方法进行语料过滤的结果如表 1 所示，其中，①为长度比过滤后的口语领域训练集，②～④为利用不同的方法对通用领域俄汉语料进行筛选后得到通用领域训练集：

表 1 不同过滤条件下获得的句对数量

Tab.1 Number of sentence pairs obtained under different filtering conditions

训练集	过滤条件	句对数
口语训练集①	-	44,619
通用训练集②	（a）+（b）+（c）	1,126,717
通用训练集③	（a）+（c）	2,077,391
通用训练集④	（a）+（b）	8,558,405

3 方法介绍

3. 1 模型微调

模型微调在许多自然语言处理任务上都取得了良好的效果。其通常做法是首先在领域相近的大规模语料上进行预训练，而后利用特定领域的语料对预训练模型的参数进行微调。再机器翻译任务上，模型微调是改善低资源语言对翻译的一种有效手段。然而预训练与微调两个阶段的所使用语料的领域适应也是一个不容忽视的问题。本文研究了在通用领域数据训练模型上进行跨领域微调的方法。为了缩小领域差距，表 1 筛选出了不同大小的训练集，其中通用训练集②与口语训练集①在所含词汇和语言模型分数上最为相似。在预训练的双语 NMT 模型上，分别进行：（1）口语训练集①微调；（2）口语训练集①和通用训练集②混合微调；（3）通用训练集②微调。

mRASP[2]是一个在包含 32 个语言对的大规模多语言语料库上得到的多语言翻译预训练模型。mRASP 利用随机对齐替换的方式拉近多语言之间的语义表示，并进一步对特定语言对进行微调。由于大规模预训练模型包含较多语义知识，因此我们使用 mRASP 预训练模型

[2] https://github.com/clab/fast_align
[3] https://github.com/kpu/kenlm

进行微调，将 2.1 节中的俄汉词表与 mRASP 的多语言联合词表进行合并，扩大俄语和汉语的词表占比，从而减少本任务中的集外词数量。

3. 2 模型平均

模型平均[7][8]可以提升模型鲁棒性，有助于提高翻译质量。模型平均是将同一个模型在训练的不同时刻保存的模型进行参数平均，通常选择模型基本收敛时对应的最后 N 个时刻或是训练过程中得分最高的 N 个时刻的参数，以同等权重对参数进行平均，得到鲁棒性更强的模型。本文选择训练阶段中最好的 5 个检查点进行模型平均。

3. 3 长度惩罚

如表 2 所示，对口语语料和通用语料统计句子长度，对俄语统计以空格分隔的单词和标点符号数，对汉语统计 Jieba 分词后以空格分隔的的词和标点符号数。由于口语语料的平均句子长度小于通用语料的平均句子长度，在预训练阶段，通用领域数据训练出来的模型更倾向生成长句子，为了更好地适应口语领域，在译码时对长度惩罚因子进行了调整。

表 2 旅游口语和通用领域数据集的平均句子长度

Tab.2 Average sentence length for travel spoken and general domain datasets

	口语训练集	口语验证集	通用训练集②
RU	5.04	9.06	16.68
ZH	5.45	10.38	20.82

4 实验

4. 1 参数设置

对于 Transformer 模型，在训练数据充足的情况下，增加隐层表示维度，即 Transformer big，能有效地提高翻译性能，而对于低资源的训练数据，适当减少隐层表示维度、编码器解码器堆叠层数反而更有益。对于口语领域的俄汉训练集，使用 Flores[9]参数设置，对于通用领域俄汉数据，使用了 base 和 big 两种参数设置，如表 3 所示。

预训练阶段学习率设置为 5e-4，warm-up 为 4000。所有模型均使用 Adam 优化器，β_1 为 0.9，β_2 为 0.98。

表 3 模型参数设置

Tab.3 Model parameter setting

参数	Flores	Transformer base	Transformer big
词嵌入维度	512	512	1024
编码器层数	5	6	6
解码器层数	5	6	6
注意力头数	2	8	16
FFN 大小	2048	2048	4096
attention dropout	0.2	0	0
residual dropout	0.4	0.3	0.3
relu dropout	0.2	0	0

4. 2 实验结果与分析

(1)长度惩罚因子

为了适应口语领域短句较多的情况，对长度惩罚因子α进行了分析，在通用训练集④上用 Transformer big 模型训练至基本收敛时，将束搜索大小固定为 4，调整长度惩罚因子的值来进行实验，使用 multi-bleu.perl[4]在验证集上计算不同系统基于字的 BLEU-4[10]值，结果如表 4 所示。

表 4 长度惩罚因子对 BLEU 值的影响

Tab.4 Influence of length penalty α on BLEU score

α	0.4	0.5	0.6	0.8	1.0	1.2	1.5
BLEU	26.87	26.91	27.06	27.12	**27.13**	26.47	22.17

由表 4 可知，将长度惩罚因子α调整到合适范围会对 BLEU 值产生正面影响，对于句子长度偏短的口语领域验证集，过大的长度惩罚因子可能会导致束搜索无法选择正确的翻译结果。

（2）模型平均

本文选择训练阶段中最好的 5 个检查点进行模型平均，结果如表 5 所示。

[4] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

表 5 不同系统进行模型平均的验证集测试结果

Tab.5 Results of different systems with model averaging on validation set

系统	训练方法	参数设置	BLEU（w/o model-averaging）	BLEU（w/ model-averaging）
baseline	train on ①	Flores	9.23	9.90
system-a	train on ②	Flores	14.85	15.55
system-b	train on ③	Base	17.80	18.79
system-c	train on ④	Big	27.13	**27.48**

（3）模型微调

本文对俄汉双语 NMT 模型和 mRASP 预训练模型进行微调，微调结果如表 6 所示。

表 6 不同系统进行微调的验证集测试结果

Tab.6 Results of different systems with fine-tuning on validation set

系统	参数设置	BLEU
baseline	Flores	9.90
system-a	Flores	15.55
system-b	Base	18.79
system-c	Big	**27.48**
+ fine-tune on ①		19.15
+ mix-fine-tune on ①②		19.19
+ fine-tune on ②		23.59
mRASP	Big	0.32
+ fine-tune on ①		**29.70**
+ fine-tune on ①②		24.62
+ fine-tune on ②		25.55

由表 6 可知，使用过滤后的通用领域语料能显著提升翻译质量，system-a 在相同的模型参数配置下，使用 112 万通用语料 system-a 的 BLEU 值提升了 5.65，使用 207 万通用语料 system-b 的 BLEU 值提升了 8.89，system-c 使用更宽的模型和 855 万筛选后的通用语料将 BLEU 值提升了 17.85。

在 system-c 的基础上，进行微调训练。直接用跨领域数据进行微调，模型性能出现明显下降。将 1126717 句对的通用语料和过滤后的 44619 句对的口语语料进行混合作为微调数据，system-c + mix-fine-tune 收敛更快，但是 BLEU 值仍然下降到 19.19。为了分析性能下降的原因，system-d 继续使用通用语料以混合微调相同学习率继续训练，BLEU 值下降到 23.59，可知更靠近口语领域训练集的数据对模型是有伤害的。这可能是由于口语领域的训练集和验证集差距导致的。

mRASP 预训练模型在口语训练集上微调获得了最好的效果，相比 baseline 结果明显提升。对比 system-c 与 mRASP+口语训练集①微调模型，可以看到更大规模的预训练模型对跨领域数据的容纳能力更强，少量的域内数据微调就能有效地改善特定领域的翻译效果。

4. 3 译文分析

表 7 展示了测试集源句及本系统生成译文的例子，其中参考译文来自谷歌翻译。对比 baseline 的译文，system-c 和 mRASP+口语训练集微调模型在 BLEU 值提高的同时，译文质量均有所改善。

低资源神经机器翻译往往具有集外词影响翻译性能的问题，使用 BPE 算法进行较小粒度的翻译在一定程度上能缓解集外词影响。然而对于俄汉机器翻译，仅使用子词切分并不能有效解决具有组合词义的集外词现象，如表 7 源句中的“на урок”组合出现时翻译为“上课”，而“урок”单独出现时翻译为“教训”。system-c 使用的训练数据中“урок”是一个高频词，而“на урок”组合并未出现过，因此产生了错译。mRASP 微调模型也未能很好地翻译这个组合。使用双语词典集和替换技术能缓解组合词义现象导致的错译和漏译问题[11][12]，由于资源和时间有限，本系统未能处理好集外词问题，在此仅作分析和参考。

表 7 源句及系统生成译文示例

Tab.7 Examples of source sentences and the system generated translation

	例子 1	例子 2
源句	Нашему（我们的） офису（办公室） нужен（需要） кондиционер（空调）.	Я（我） снова（又） опоздал（迟到了） на урок（上课）.
参考译文	我们 办公室 需要 空调 。	我 上课 又 迟到 了 。
baseline 生成译文	办公设备 。	我 迟到 了 。
system-c 生成译文	我们 办公室 需要 空调 。	我 再次 错过 了 教训 。
mRASP+ fine-tune on ①生成译文	我们 办公室 需要 空调。	我 再次 拖延 了。

5 总结

本文介绍了 CCMT2021 低资源语言评测项目俄汉方向上使用的主要技术和方法。主要借助通用领域的大规模俄汉平行语料来提升旅游口语领域俄汉机器翻译性能。融合了词占比、Fast-align、语言模型等数据筛选方法获取接近口语领域的域外数据，在 mRASP 预训练模型上进行微调获得了明显提升，译码时使用合适的长度惩罚因子和模型平均技术。实验结果显示，这些方法能够明显提高低资源情况下旅游口语俄汉翻译的质量。

由于时间有限，本次评测中还有许多方法有待尝试。在评测过程中发现了一些问题和不足，如跨领域微调的性能下降和集外词问题，有待于今后的进一步研究。

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] LIN Z, PAN X, WANG M, et al. Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[3] 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.

[4] DYER C , CHAHUNEAU V , SMITH N A . A Simple, Fast, and Effective Reparameterization of IBM Model 2[J]. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 644–48 .

[5] HEAFIELD K , POUZYREVSKY I , CLARK J H , et al. Scalable Modified Kneser-Ney Language Model Estimation[J]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013: 690-696

[6] 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.

[7] SENNRICH R, HADDOW B, BIRCH A. Edinburgh neural machine translation systems for wmt 16[C]//Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, 2016: 371 – 376.

[8] 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.

[9] FRANCISCO G, CHEN P J , OTT M , et al. The FLoRes Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 6097 – 6110.

[10] KISHORE P, SALIM R, TODD W, et al. Bleu: a method for automatic eval-uation of machine translation. [C] // Proceedings of ACL 2002. Philadelphia: Association for Computational Linguistics, 2002: 311 – 318 .

[11] Li X, ZHANG J, ZONG C. Towards zero unknown word in neural machine translation. International Joint Conference on Artificial Intelligence, 2016: 2852-2858.

[12] CHE W , YU Z , YU Z , et al. Towards Integrated Classification Lexicon for Handling Unknown Words in Chinese-Vietnamese Neural Machine Translation[J]. ACM Transactions on Asian and Low-Resource Language

Information Processing, 2020, 19(3):1-17.

Evaluation Technical Report for CCMT2021 Low Resource Russian-Chinese Translation task

LIU Huan, LIU Junpeng, HUANG Kaiyu, HUANG Degen*

(College of Computer Science and Technology, Dalian University of Technology, Liaoning Dalian 116024, China)

Abstract: In this paper, we describe the DUTNLP participation in the 17th National Machine Translation Conference (CCMT 2021) low-resource translation task for Russian-Chinese. We explore the strategy of data enrichment, by selecting general domain data that is close to the in-domain data from large-scale general Russian-Chinese parallel corpus. Firstly, multiple data filtering methods are used to obtain different training set in general domain. Then, pre-training and fine-tuning are carried out on Transformer models with different parameter configurations to obtain the optimal model. Finally, the translation results are generated by model averaging in the decoding stage. The experiment results show that our system can achieve significant improvements on Russian-Chinese translation in the low-resource scenario compared with baseline system.

Keywords: neural machine translation; CCMT2021; low resource language; mRASP