DUTNLP Machine Translation System for WMT21 Triangular Translation Task

Huan Liu **Junpeng Liu** **Kaiyu Huang*** **Degen Huang**
School of Computer Science, Dalian University of Technology
{liuhuan4221,liujunpeng_nlp, kaiyuhuang}@mail.dlut.edu.cn
{huangdg}@dlut.edu.cn

Abstract

This paper describes DUT-NLP Lab's submission to the WMT-21 triangular machine translation shared task. The participants are not allowed to use other data and the translation direction of this task is Russian-to-Chinese. In this task, we use the Transformer as our baseline model, and integrate several techniques to enhance the performance of the baseline, including data filtering, data selection, fine-tuning, and post-editing. Further, to make use of the English resources, such as Russian/English and Chinese/English parallel data, the relationship triangle is constructed by multilingual neural machine translation systems. As a result, our submission achieves a BLEU score of 21.9 in Russian-to-Chinese.

1 Introduction

The WMT2021 Shared Task on translating sentences from Russian into Chinese provides a challenging mixed-genre test for machine translation systems and a triangular relationship for researchers to evaluate new techniques. The task focuses on translation between non-English languages and optimally mixing direct and indirect parallel resources. In this task, the participants must use only the provided parallel training data and use of other data is not allowed. The provided data is shown in Table 1, including parallel data in three directions. Given the language pair (Russian-to-Chinese), the bulk of previous NMT work has pursued one of two strategies that are illustrated in Figure 1:

Direct: Collect parallel Russian-to-Chinese data from the public resource, and train a Russian-to-Chinese translator.

Pivot: Collect parallel Russian-to-English and English-to-Chinese data (usually larger than direct data), train two translators (Russian-to-English + English-to-Chinese), and make a cascade translator from Russian to Chinese.

Direction	# SENT
Russian/Chinese	33,388,455
Russian/English	69,155,404
English/Chinese	28,528,290

Table 1: The provided training data in the constrained data track.

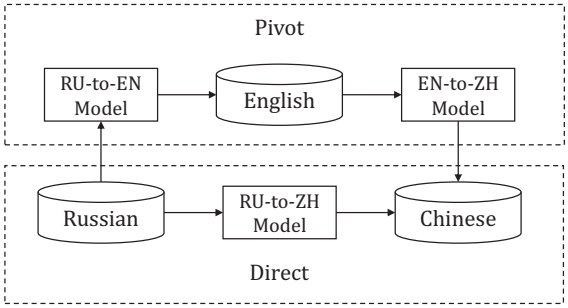

Figure 1: The illustration of two strategies for triangular machine translation.

The DUTNLP submission to the constrained data track is based on the mainstream architecture Transformer (Vaswani et al., 2017). According to the scale of datasets, this shared task should be considered as the high-resource translation direction. We use the Transformer-big setting for better performance, on the contrary, a low-resource translation task often utilizes the Transformer-base due to the limited parallel training data. Moreover, to enhance the baseline model and investigate the usage of triangular direction data, we utilize two pipelines for combining English related resources: 1) incorporates English-to-Chinese translator into direct translation process to normalize the translation of rare words. 2) adopts a multilingual training strategy to make use of English ↔ X parallel resources. Besides, some of the provided Russian ↔ Chinese parallel corpora are crawled from the

*Corresponding author

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 331–335
November 10–11, 2021.

web, which has many noise issues. We filter the training bilingual corpora with several techniques, including language model and constrained rules.

This paper is structured as follows: Section 2 describes variants of models we used in the shared task. In Section 3, we introduce the system overview using several techniques for model enhancement, including data pre-processing and filtering, triangular translation strategy and fine-tuning. In Section 4, this paper presents experimental settings, main results and analysis. Finally, in Section 5 we draw a brief conclusion of our work in the WMT2021 Triangular Translation Task.

2 Model

2.1 Transformer

Recent advances in Transformer (Vaswani et al., 2017) have led to significant improvement of Neural Machine Translation (NMT) and achieve human parity on Automatic Chinese to English News Translation (Hassan et al., 2018). The Transformer adopts a sequence-to-sequence structure, using stacked encoder and decoder layers of self-attention. Each encoder layer consists of a self-attention mechanism and a feed-forward network. Each decoder layer consists of a masked self-attention layer, a cross self-attention layer, and a feed-forward network layer. Moreover, the Transformer leverages positional embedding, residual connections and layer normalization for enhancement (Ba et al., 2016). In this paper, we adopt the Transformer-big as the baseline model, in which both the encoder and decoder have 6 layers, the hidden size is 1024, and the feed-forward inner size is 4096.

2.2 Multilingual Architecture

Multilingual neural machine translation (mNMT) handles the translation between multiple languages by joint training in a multi-task setup (Johnson et al., 2017), which greatly eases the model deployment. Previous works (Lakew et al., 2018; Tan et al., 2019) show that the mNMT model can facilitate cross-lingual knowledge transfer between languages. It also enables zero-shot translation between unseen language pairs (Johnson et al., 2017; Al-Shedivat and Parikh, 2019; Zhang et al., 2020). Following Johnson et al. (2017), we build our multilingual translation system based on the advanced Transformer model by adding a pretending language token to each source sentence, which indicates the language to be translated into. And the multilingual model is also fine-tuned on pre-trained mNMT models such as mBART(Liu et al., 2020) and mRASP(Lin et al., 2020).

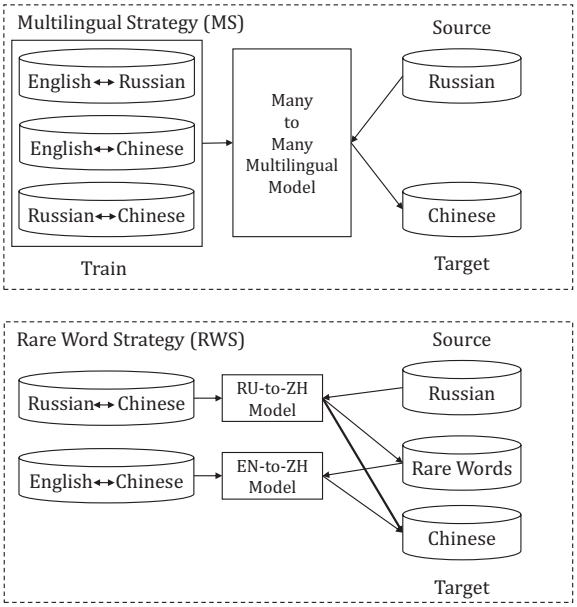

Figure 2: The illustration of two strategies for incorporating English resources into the baseline systems.

3 System Overview

3.1 Data Pre-processing and Filtering

To improve the quality of data, especially the Russian ↔ Chinese parallel data, we filter noisy data with several techniques. The flow of all training data pre-processing and filtering is set to step by step as follows:

- Punctuation normalization with Moses scripts (Koehn et al., 2007) for all language pairs.
- Chinese word segmentation using the open segmentation tool (Huang et al., 2020). Splitting the English and Russian words using clearly delimiter by Moses "tokenizer" script.
- True-casing. The uppercase letter may influence the generation of vocabulary dictionaries. We transfer the uppercase letter into lower case automatically by Moses scripts.
- Filtering out the sentence pairs longer than 256 or duplicated translation.
- Filtering out the sentences by the multilingual parallel data filter tools LASER [1].

[1]https://github.com/facebookresearch/LASER

- Filtering out the sentence according to their characteristics in terms of language identification and length ratios, in particular, the sentence pairs whose length ratio between the source and target are not in range of 1:2.5 and 2.5:1 are abandoned.

- The bilingual direct translator utilizes BPE to encode text into sub-word unit (Sennrich et al., 2016). In the multilingual translator, the system applies sub-word processing using SentencePiece tool (Kudo and Richardson, 2018).

3.2 Triangular Translation Strategy

In this task, we exploit two strategies to incorporate English resources into the baseline systems, which are shown in Figure 2.

Rare Word Strategy According to the official provided data, the performance of the direct translator is better than the pivot. And the provided parallel Russian ↔ Chinese training data can be considered as the high-resource. The baseline model can be trained successfully with the bilingual training data and can achieve competitive performance in most cases. However, the translations of rare words are always terrible by the direct translator, for example, most Russian names will be translated into English. We utilize an English ↔ Chinese translator to alleviate this issue. The external translator only works in the specific case to reduce the error propagation and redundant computational cost. This strategy is to complement the direct baseline model and improve the performance in an interpretable way.

Multilingual Strategy The mNMT models are effective in low resource settings due to knowledge transfer. To make use of all provided parallel data, we train a multilingual many-to-many models with 3 language pairs(i.e., 6 directed translation direction). However, it is expensive to train different parameter sets to get the best translation result in the target direction. According to Lin et al. (2020), mRASP can obtain more improvements with rich-resource language pairs than multilingual many-to-many models, so multilingual strategy (MS) is based on fine-tuning large-scale pre-training models. We fine-tune on two pre-trained models respectively: mBART(Liu et al., 2020) and mRASP(Lin et al., 2020). In particular, we use MS-mBART and MS-mRASP for the two methods in section 4. We use mBART to continue training on the filtered parallel data subset, and select the appropriate checkpoints according to the performance on the validation set. The mRASP model pre-trained on a dataset contains 32 English-centric language pairs, including English-Russian and English-Chinese, and Russian-Chinese are not the direct training objective of it. We stop the fine-tune process when the loss on validation does not decrease for 5 consecutive steps (measured every 50 updates). The experiment shows that fine-tuning on mRASP with filtered parallel data achieves anticipated improvements.

Models	VALID.	TEST.
Direct	20.2	17.0
Baseline	24.92	20.4
+RWS	26.40	21.8
+MS-mBART	19.47	-
+MS-mRASP	25.21	21.7

Table 2: The BLEU-4 scores in the constrained data track.

4 Experiment

4.1 Experimental Settings

The implementation of our models is based on Fairseq (Ott et al., 2019). All the models are carried out on 2 NVIDIA 3090 GPUs each of which has 24 GB of memory. The parameters of Transformer-Big, mBART, and mRASP are all followed by the architectures themselves. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The batch size is set to 4096 tokens and the "update-freq" parameter in Fairseq is set to 2. In particular, for pre-trained language models settings (i.e., mBART and mRASP), the batch size is 2048 and "update-freq" is 4. The initial learning rate is set to $5e^{-4}$ for training and $3e^{-5}$ for fine-tuning. The learning scheduler is inverse_sqrt and all the dropout probabilities are set to 0.1. We select the checkpoint with the average of the top 5 sacreBLEU scores on the development set as the final checkpoint in each training. We calculate the BLEU-4 score for all experiments, which is officially recommended.

4.2 Main Results

Table 2 shows the Triangular translation results on validation and test set. We train multiple models in each setting and report the best scores in Table

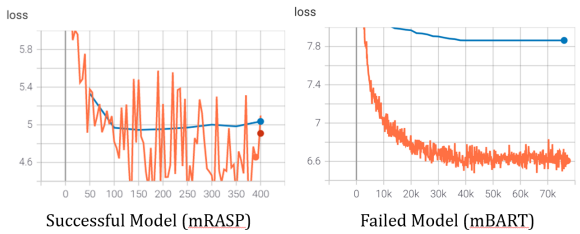

Figure 3: Loss curve of the multilingual models and the local minima.

Ref: 按照历史学家的说法，**阿尔巴津圣母圣像**是1492年尼科季姆修士所绘。
Baseline: 根据历史学家的推测，Albazinskaya是在1492年由尼克科姆修道院写的。
RWS: 根据历史学家的推测，**阿尔巴津圣母像**是在1492年由尼科季姆修道院写的。
MS-mRASP: 根据历史的推测，**阿尔巴津圣母神像**是在1492年由尼古迪姆修道院写成的。
Ref: **坦波夫地区**是农业地区，但该地区在种子生产领域也面临严重问题。
Baseline: Tambovshchyna是一个农业地区，但在种子生产方面也面临严重问题。
RWS: **坦波夫地区**是一个农业地区，但在种子生产方面也面临严重问题。
MS-mRASP: **坦布夫地区**是一个农业地区，但面临着严重的农业问题。

Figure 4: The results by different NMT systems. The rare words are bold.

2. In particular, the direct model is trained with all available data, and the Baseline utilizes the clean parallel data after filtering to train. It improves the direct method by 4.72 and 3.4 BLEU scores on validation and test set, respectively. Moreover, to make use of the English resources, we propose two triangular training strategies to investigate the effect of the triangle relationship. The rare word strategy (RWS) can effectively improve the baseline from 24.92 to 26.40 in terms of BLEU scores. And the multilingual strategy (MS) improves the baseline from 24.92 to 25.21 in terms of BLEU scores.

4.3 Triangular Translation Analysis

The official results do not open the reference on the test set. So we investigate the experiments further on the validation set.

Multilingual Models In this task, we utilize the multilingual training strategy which is fine-tuned on mBART and mRASP. It is surprising to find that the model trains failed on this dataset, which is fine-tuned on mBART, shown in Figure 3. However, the mRASP worked in this scenario. Follow by Lin et al. (2020), the mRASP is beneficial to fine-tune the high-resource language pairs. And this method (MS-mRASP) achieves the best BLEU-4 scores on the validation set. Although the training curve of MS-mBART is more stable, the overall loss has remained relatively high level and cannot be effectively declined.

Results Discussion Figure 4 shows the results from different models. The two multilingual translation strategies can effectively improve the performance of baseline model, especially for the rare words. In the baseline model, some rare words are translated into English, and most of the words are translated correctly depends on the direct training method. It is worth mentioning the two multilingual translation strategies can alleviate this issue effectively.

5 Conclusion

This paper presents the DUTNLP Translation systems for WMT2021 Russian-to-Chinese triangular translation tasks. We investigate various neural architectures and data filtering to build strong baseline systems. Then the two triangular translation strategies are used to improve the baselines. We also prove that in-domain finetuning is very effective for this translation task. Finally, we discuss the results carefully and analyze the influence of different triangular strategies for further improvement. A number of advanced technologies reported in this paper focus on alleviating the issue of triangle translation. As a result, our system outperforms the strong baseline by 1.48 and 1.4 BLEU scores on the validation set and test set, respectively. In the future, we will investigate the technologies in the low-resource scenario and continue to improve the performance of this task through post-evaluation submissions.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and suggestions to improve the quality of the paper. The authors gratefully acknowledge the financial support provided by the National Key Research and Development Program of China (2020AAA0108004) and the National Natural Science Foundation of China under(No.U1936109).

References

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,

pages 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Kaiyu Huang, Degen Huang, Zhuang Liu, and Fengran Mo. 2020. A joint multiple criteria model in transfer learning for cross-domain chinese word segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3873–3882.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.