IOL Research Machine Translation Systems for WMT23 General Machine Translation Shared Task

Wenbo Zhang, Zeyu Yan, Qiaobo Deng, Jie Cai, and **Hongbao Mao**
Transn IOL Research, Wuhan, China

Abstract

This paper describes the IOL Research team's submission systems for the WMT23 general machine translation shared task. We participated in two language translation directions, including English→Chinese and Chinese→English. Our final primary submissions belong to constrained systems, which means for both translation directions we only use officially provided monolingual and bilingual data to train the translation systems. Our systems are based on Transformer architecture with pre-norm or deep-norm, which has been proven to be helpful for training deeper models. We employ methods such as back-translation, data diversification, domain fine-tuning and model ensemble to build our translation systems. An important aspect worth mentioning is our careful data cleaning process and the utilization of a substantial amount of monolingual data for data augmentation. Compared with the baseline system, our submissions have a large improvement in BLEU score.

1 Introduction

This paper describes our submissions to the WMT23 General Machine Translation shared task. We participated in two language translations: English-to-Chinese and Chinese-to-English. For both tasks, we built our system in a constrained scenario, using only official training data. Our systems are based on Transformer(Vaswani et al., 2017) architecture with pre-norm or deep-norm(Wang et al., 2022), which has been proven to be helpful for training deeper models. We used rule-based methods, language models, and alignment models to clean bilingual and monolingual data, and then used back-translation(Sennrich et al., 2016), data diversification(Nguyen et al., 2020), and model ensemble(Garmash and Monz, 2016) to leverage large-scale monolingual data to construct our translation systems. We also tried domain fine-tuning and found that this approach still helped in improving the BLEU(Papineni et al., 2002) scores on the WMT23 test set.

The design of the subsequent paper is as follows. We introduce the data source and processing strategy in Section2; Section 3 describes the details of our training procedure; Section 4 presents the experimental settings and results.

2 Data

2.1 Data Source

Bilingual corpus We used all provided bilingual data, including: ParaCrawl v9(Bañón et al., 2020), News Commentary v18.1, Wiki Titles v3, UN Parallel Corpus v1.0(Ziemski et al., 2016), CCMT Corpus, WikiMatrix(Schwenk et al., 2019), and Back-translated news.

English monolingual corpus The used English monolingual data including: News crawl, News discussions, Europarl v10, News Commentary, Common Crawl, Leipzig Corpora(Goldhahn et al., 2012), and English part of other bilingual data for WMT general task.

Chinese monolingual corpus The used Chinese monolingual data including: News crawl, News Commentary, Common Crawl, Leipzig Corpora, and Extended Common Crawl.

2.2 Data Preprocessing

For bilingual data we first filter out noisy sentences according to the rules, the filtering rules are as follows:

- Remove invisible characters.
- Remove sentences containing too more than 300 words or more than 1000 characters or less than 3 characters.
- Remove English sentences containing words exceeding than 40 characters.
- Remove Chinese sentences with a low rate of Chinese characters(less than 0.2).

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 187–191
December 6–7, 2023.

- Remove sentences that contain too many punctuation marks.
- Remove sentences that contain repeated substrings, which refers to a string composed of a single character that repeats more than 10 times, or two or more character that repeat more than 5 times.
- Remove sentences that contain HTML tags.
- Convert full-width characters to half-width characters, Traditional Chinese to Simplified Chinese.
- Remove duplicated sentence pairs.

Then we use fast-align(Dyer et al., 2013) to filter out sentence pairs with low alignment scores (less than 13) or low bilingual alignment ratio (less than 0.6), and use forward and reverse translation models to calculate the perplexity of sentence pairs, removing sentence pairs with high perplexity. For monolingual data we perform filtering using similar rules to bilingual data. At the same time, The KenLM(Heafield, 2011)[1] tool is used to train an n-gram language model to filter sentences with high perplexity scores (more than 10 000). The original parallel data totaled about 64 million sentences, and after cleaning, 46.06 million sentences were retained. Through data cleaning, we obtained 1.4 billion sentences Chinese monolingual data, and 1.2 billion sentences English monolingual data.

We used the Sentencepiece(Kudo and Richardson, 2018) tool to train the unigram model for subword segmentation, and vocabulary sizes for both Chinese and English were set to 36 000.

3 System Overview

We chose Transformer(Vaswani et al., 2017) as our base translation model and used both pre-norm and deep-norm(Wang et al., 2022) variants to help us train deeper models. To improve the quality of translation models, we first pre-trained the translation models from scratch on the synthesized datasets generated by back-translation, then continue training on the datasets generated by data diversification, and finally used domain data for fine-tuning. We also iteratively performed two rounds of data augmentation to improve the quality of the synthetic data. The final synthetic data is generated by the model after training on data diversification data of the first round. We only used domain fine-tuning in the final submission. This method we adopt is a commonly used method in the field of machine translation and has been proven to be effective. In the following sections, We show the specifics of how we use these methods.

3.1 Back-translation

Back-translation(Sennrich et al., 2016) is almost the most well-known data augmentation method in the field of machine translation, which can effectively utilize target monolingual data to improve translation quality, even in high resource situations. We used top-k sampling strategy to generate back-translation data with top-k=10, and used the method in section 2 to filter the generated data. To further increase the diversity of synthetic data, we also employed different back-translation models, such as the R2L model and the L2R model, and models with different structures to perform the back-translation method. Since this task is oriented to a general domain, we only use the cleaned monolingual data to generate synthetic data and do not select according to the domain. Because our systems are first pre-trained on back-translation data, unlike the original approach(Sennrich et al., 2016), the method back-translation in this paper refers to using only back-translation data and does not including the non-augmented corpora.

3.2 Data Diversification

Data diversification(Nguyen et al., 2020) is a data augmentation method by performing back-translation and forward-translation multiple times on the target-side and source-side data of the parallel corpus, respectively. Following this approach, we used different models to generate synthetic data by beam search. However, we not only use parallel data as source language for synthetic data, but also monolingual data. The ratio between monolingual and parallel data is 1:1.

3.3 Model Ensemble

Model ensemble can effectively improve the overall system performance by combining the strengths of multiple individual models. The larger the difference between multiple single models, the larger the improvement the ensemble model can receive. We mainly increase the diversity between single models by using different monolingual data, including different monolingual data in the back-translation

[1] https://github.com/kpu/kenlm

stage and different monolingual data in the data diversification stage.

3.4 Domain Fine-tuning

Although the WMT23 test set contains sentences from multiple domains and the WMT21 test set mainly consists of sentences from the news domain, we found that fine-tuning on the WMT21 test set can still improve the WMT23 test set. Therefore, we still attempted to fine-tune our model using newtest2021 as in-domain data.

4 Experiments

4.1 Experiment Settings

All of our translation models were implemented based on fairseq(Ott et al., 2019) and trained on 8 NVIDIA A100 GPUs. During training, we used the Adam(Kingma and Ba, 2014) optimizer with $\beta 1 = 0.9, \beta 2 = 0.98$, the learning rate scheduling strategy of inverse sqrt, the number of warmup step set to 4000, the maximum learning rate set to 0.0005 and FP16 to accelerate the training process .

We used a 24-encoder, 6-decoder transformer with pre-norm as baseline and the embedding size was set to 1024. It was trained only on a real parallel corpus, with a batch size set at 240,000 tokens. For the data augmentation models, we increased the dimension of the embedding size to 1536 and adjusted the number of the encoder and decoder layers, using equal encoder and decoder layers, or deep encoder layers and shallow decoder layers to increase the model parameter size to approximately 1 billion. The training process for these models used a batch size of 640,000 tokens. Maintaining the diversity of different models is a useful trick for model ensembles, so we trained multiple different models by adjusting the number of layers of different models, using pre-norm or deep-norm, using different synthetic data, with or without domain fine-tuning to improve diversity. Finally, we trained 4 models from Chinese to English and 5 models from English to Chinese for model ensemble.

4.2 Results

All experiments were evaluated using the sacrebleu(Post, 2018) tool to calculate BLEU(Papineni et al., 2002) scores on the WMT21, FLoRes(Goyal et al., 2021), and NTREX-128 test sets(Federmann et al., 2022). We used beam search with beam size=5 to decode all models and converted punctuation to Chinese characters in English-to-Chinese direction. Regarding the final results we submitted, we also used regular expressions for n-gram repetition detection. For translations containing repeated substrings, we set a repetition penalty of 1.5 to retranslate the source sentences. The results of Zh→En and En→Zh are shown in Table 1 and Table 2.

Based on Table 1, we can clearly see that the use of Back-Translation and Data Diversification shows significant improvements on multiple test sets. Compared to the baseline, using both data augmentation methods achieves more than 2 BLEU improvements on each test set. More than 0.5 BLEU improvement is also achieved on each test set with the model ensemble. In the end, we achieved BLEU improvements of +4.4, +3.7 and +2.8 on the three test sets of FLoRes, NTREX-128 and WMT21 respectively. The inclusion of domain fine-tuned models can further improve the WMT 23 test set compared to the model ensemble without domain fine-tuning.

From Table 2, we can see that there is a significant improvement using Back-translation on each test set. After using Data Diversification, only further improvement is achieved on the FLoRes test set, while there is varying degree of decrease on the other two test sets. Due to the decrease in diversity caused by fine-tuning multiple models with similar synthetic data generated by Data Diversification, and Data Diversification did not lead to a consistent improvement on the English to Chinese test set, in the model ensemble stage, 4 out of 5 models were trained on only Back-translation data. Finally, on the three test sets of FLoRes, NTREX-128, and WMT21, we achieve improvements of +6.5, +5.9, and +3.6 BLEUs compared to the baseline, respectively, with the model ensemble contributing the largest improvement. Similar to the results from Chinese to English, further improvements are obtained on the WMT23 test set after adding domain fine-tuning.

5 Conclusion

In this paper, we described IOL Research's submissions to the WMT2023 General Translation shared task. We participated in the English from and to Chinese translation. Our system aims to leverage as much monolingual data as possible to improve the quality of machine translation. Experimental

System	FLoRes	NTREX-128	WMT21	WMT23
Baseline	31.4	30.4	27.6	-
+Back-translation	34.2	33.2	28.4	-
+Data Diversification	35.2	33.2	29.7	-
+Ensemble	35.8	34.1	30.4	26.4
+Fine-tuning	-	-	-	27.2

Table 1: Zh→En BLEU scores on FLoRes, NTREX-128, WMT21, and WMT23 test sets. Due to the limited number of submissions, we only report part results of WMT23.

System	FLoRes	NTREX-128	WMT21	WMT23
Baseline	41.8	33.5	31.9	-
+Back-translation	44.6	37.4	33.9	-
+Data Diversification	45.2	34.5	32.8	-
+Ensemble	48.3	39.4	35.5	56.3
+Fine-tuning	-	-	-	56.9

Table 2: En→Zh BLEU scores on FLoRes, NTREX-128, WMT21, and WMT23 test sets. Due to the limited number of submissions, we only report part results of WMT23.

results show that by increasing the scale of monolingual data in the system through data augmentation and model ensemble, we have achieved substantial improvements on multiple test sets.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and NoahA. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. *North American Chapter of the Association for Computational Linguistics,North American Chapter of the Association for Computational Linguistics*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. *International Conference on Computational Linguistics,International Conference on Computational Linguistics*.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

DiederikP. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning,arXiv: Learning*.

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*.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

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 NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. 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. *Advances in neural information processing systems*, 30.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. Deepnet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.