Manifold's English-Chinese System at WMT22 General MT Task

Chang Jin
Soochow University*
cjin@stu.suda.edu.cn

Tingxun Shi
OPPO
shitingxun@oppo.com

Zhengshan Xue
OPPO
xuezhengshan@oppo.com

Xiaodong Lin
Rutgers University
lin@business.rutgers.edu

Abstract

Manifold's English-Chinese System at WMT22 is an ensemble of 4 models trained by different configurations with scheduled sampling-based fine-tuning. The four configurations are DeepBig (XenC), DeepLarger (XenC), DeepBig-TalkingHeads (XenC) and DeepBig (LaBSE). Concretely, DeepBig extends Transformer-Big to 24 encoder layers. DeepLarger has 20 encoder layers and its feed-forward network (FFN) dimension is 8192. TalkingHeads applies the talking-heads trick. For XenC configs, we selected monolingual and parallel data that is similar to the past newstest datasets using XenC, and for LaBSE, we cleaned the officially provided parallel data using LaBSE pretrained model. According to the officially released autonomic metrics leaderboard[1], our final constrained system ranked 1st among all others when evaluated by bleu-all, chrf-all and COMET-B, 2nd by COMET-A.

1 Introduction

This report describes Manifold's machine translation system submitted to WMT 22 English→Chinese general domain translation task. The general domain translation task of WMT is a new task set up this year, replaces the time-honored news translation task. However, as the newstest datasets released previously were created by professional translators manually, they were considered to have better quality compared with the web crawled dataset. Therefore, generally our strategy is selecting data from the provided datasets according to their similarity between past WMT test sets, and training big models with different architectures or training methods based on the selected data.

2 Data Preprocessing

For officially released bilingual data, we merged ParaCrawl v9, News Commentary v16, Wiki Titles v3, UN Parallel Corpus v1.0, CCMT corpus and WikiMatrix corpus together, then applied multiple rules to preprocess and filter the officially released bilingual data, including

- Normalizing punctuation.
- Removing unprintable characters.
- Tokenizing texts. jieba[2] was applied for Chinese texts, and moses tokenizer[3] was applied for English texts.
- Removing all the sentence pairs whose source text or target text tokens count exceeds 150, and the segment pairs with a source-target token ratio lower than 2/3 or higher than 3/2.
- Removing all the sentence pairs that belong to other languages. We applied the fasttext model (Joulin et al., 2016)[4] as our language identifier.
- Deduplicating the dataset.

3 Basic models

After applying the aforementioned preprocessing and filtering rules, 30M segment pairs were kept [5]. For the kept data, we adopted two different selection strategies:

1. Following NiuTrans' system submitted to WMT21 (Zhou et al., 2021), we selected 12M segment pairs from the kept data which are

*Work done during Jin's internship in OPPO.

[1] https://github.com/wmt-conference/wmt22-news-systems/blob/main/scores/automatic-scores.tsv

[2] https://github.com/fxsjy/jieba

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[4] https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin

[5] We did not make use of official back-translated corpus

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 275 - 279
December 7-8, 2022 ©2022 Association for Computational Linguistics

Config Item	Big	Deep	Deeper	DeepBig	DeepLarger
# Encoder Layer	6	30	40	24	20
# Attention Heads	16	8	8	16	16
Embedding Size	1024	512	512	1024	1024
FFN Size	4096	2048	2048	4096	8192
Pre-Norm	No	Yes	Yes	No	No

Table 1: Main configurations for different architectures we applied for basic models. Decoder layers numbers were fixed to 6 for all the architectures if not specially mentioned.

most similar to our validation set (i.e. newstest2020enzh data) using XenC (Rousseau, 2013). This part of data will hereinafter be referred as "XenC" for short.

2. We further cleaned the 30M segment pairs using LaBSE (Feng et al., 2022)[6], set the threshold to 0.7 according to our experience in filtering out un-aligned data. After this step, 24.3M segment pairs were kept. This part of data will hereinafter be referred as "LaBSE" for short.

Config Item	Value
dropout	0.3
learning rate	0.0005
max tokens	4096
warmup init lr	1e-7
warmup steps	4000
label smoothing	0.1
num max updates	300,000
update frequency	8

Table 2: Hyper-parameters for training models. All experiments were conducted on 4 or 8 Tesla V100 GPUs.

We applied BPE-subword (Sennrich et al., 2016)[7] to divide tokens into subwords. BPE codes were learned jointly with 32K merge operations but dictionaries for the source language and target language are generated separately.

Basic models were trained to generate synthetic, pseudo bilingual segment pairs for the next step. As is discovered by Zeng et al. (2021), sub-model diversity is a key factor to enhance the performance of ensemble model. Therefore, we trained various Transformer models (Vaswani et al., 2017) applying different architectures. For very deep Transformers, we followed the suggestion given by DLCL (Wang et al., 2019) to use Pre-Norm. Main configurations for the architectures we trained are listed in Table 1. Other hyper-parameters for training models are listed in Table 2 (same for all the experiments we took in the contest).

For each architecture X listed in Table 1, we also trained its talking-heads attention variant (Shazeer et al., 2020) to further increase model diversity. Such variants will be denoted as "X-th" in the following part of the report. All the models were developed and trained using fairseq (Ott et al., 2019) [8]

4 Data Augmentation

Previous studies show that adding synthetic data can help to boost the performance of machine translation systems (Edunov et al., 2018) (Hoang et al., 2018). We adopted four data augmentation methods during the contest, including back-translation (with sampling), forward-translation, sequence knowledge distillation and R2L translation. For R2L translation, we reversed the token sequences of inputs, e.g. converted "This is a book ." to ". book a is This", and left the target sentences unchanged.

We selected 12M sentences from officially released English and Chinese monolingual datasets respectively[9], also applying XenC algorithm on them. The selected monolingual data was preprocessed by the similar pipeline presented in section 2, with the difference on skipping the steps to filter unaligned bilingual data. The reason behind selecting 12M sentences is we expect that the data from each monolingual dataset has the same size compared with the bilingual training data, as Edunov et al. (2018) indicated.

The preprocessed Chinese monolingual data was

[6] We downloaded the pretrained model from transformers official website on October 11th., 2021. As the model was released before February 2022, the constrained system requirement is not violated.

[7] https://github.com/rsennrich/subword-nmt

[8] https://github.com/facebookresearch/fairseq

[9] For English, we combined News Crawl, News Discussions, News Commentary and Europarl v10 corpus together; for Chinese, we combined News Crawl, News Commentary and Common Crawl corpus.

No.	Method	Data Size	Newstest 2020	Newstest 2021
1	Deep baseline model	12M	42.7	32.4
2	1 + Forward-translation (ForT)	24M	44.7 (+2.0)	33.3 (+0.9)
3	1 + Top-p back-translation (topp BT)	24M	45.8 (+3.1)	32.9 (+0.5)
4	1 + R2L KD + R2L ForT	36M	44.2 (+1.5)	33.5 (+1.1)
5	Sequence KD	12M	42.7 (-)	32.5 (+0.1)
6	5 + ForT + topp BT	36M	45.3 (+2.6)	33.7 (+1.3)
7	4 - 1 + 6	60M	45.5 (+2.8)	33.8 (+1.4)

Table 3: Model performances when applying different methods. All the models are trained by Deep architecture depicted in Table 1. Our validation dataset is from newstest2020 (Barrault et al., 2020) and test dataset is from newstest2021 (Akhbardeh et al., 2021).

No.	Method	Newstest 2020	Newstest 2021
1	Fine-tuned Deep	46.2	34.6
2	DeepBig	47 (+0.8)	35 (+0.4)
3	DeepLarger	47.5 (+1.3)	35.7 (+1.1)
4	DeepBig-th	47.6 (+1.4)	35.4 (+0.8)
5	DeepBig (LaBSE)	47.6 (+1.4)	35.5 (+0.9)
6	Ensemble of 2, 3, 4 and 5	48.1 (+1.9)	36.2 (+1.6)

Table 4: Detailed performance information of the four bigger model and the final ensemble model. All the improvements are based on the baseline model (Deep model shown in the last line of Table 3, fine-tuned using past newsdev/test datasets, by applying decoder steps based schedule sampling as regularization). All sub-models (No. 2 to 5) are also fine-tuned by the same datasets applying the same regularization method.

then back-translated into English by an ensemble model, which is composed by a Big model, a Deep model and a Big-th model, all trained with the bilingual XenC Chinese→English corpus. Inspired by some ideas of Burchell et al. (2022), we performed top-p (nucleus) sampling (Holtzman et al., 2019) in the process of back-translation to import some noises, and set topp to 0.9. In this way, 12M Pseudo English-Chinese segment pairs were constructed.

The preprocessed English monolingual data was forward-translated into Chinese, leading to another 12M English-Pseudo Chinese dataset. The ensemble model used to generate data has the same architecture with the back-translation model introduced above, the only difference is it is trained by the parallel XenC English→Chinese corpus.

We further applied sequence-level knowledge distillation (Kim and Rush, 2016) to distill this ensemble model by translating English sentences of the parallel XenC corpus into pseudo Chinese. Furthermore, we also trained a Deep model with the reversed parallel XenC data to get a right-to-left (R2L) model, and generated forward-translation and the results of knowledge distillation using reversed monolingual and parallel English corpora. After having acquired these different synthetic datasets, we trained Deep models by various combinations of them. The concrete model performance is shown in Table 3.

5 Fine-tuning and Bigger Models

After having acquired extra data by back/forward-translation, knowledge distillation, and R2L data augmentation, we experimented on several other methods to further improve our system.

Fine-tuning. We fine-tuned our Deep model using the combination of newsdev2017, newstest2017, newstest2018, and newstest2019 datasets. As the dataset used for fine-tuning is quite small, we applied decoder steps based scheduled sampling (Liu et al., 2021) as a means of regularization. We set k to 0.99. With this step, a 0.7 BLEU gain has been brought on the validation set and 0.8 BLEU gain on the test set.

Ensemble of bigger models. We trained three bigger models on the 60M XenC Dataset (configuration No. 7 in Table 3), including a DeepBig model, a DeepLarger model and a DeepBig-th model. Furthermore, we replaced the distilled data in the XenC Dataset (12M) with the LaBSE data (24M), and got a dataset containing 72M segment pairs which is used to train another DeepBig model. We fine-tuned these four models using past newsdev/test datasets and applied decoder steps based schedule sampling, then made an ensemble model comprised of them. Table 4 lists the detailed performance of the bigger models and their ensemble model.

Post-processing. We converted punctuation from half-width symbols to full-width symbols for the generated results.

6 Conclusion

In this report, we describe our Manifold English→Chinese system submitted to WMT 22 general translation task. The core idea of our system is to train various big Transformer models utilizing in-domain (actually news domain) data, based on which an ensemble model is created. Although most training data belongs to a special domain, we still achieved compelling results in the final submission, i.e. our final system ranked first among the constrained systems, evaluated by BLEU score based on the two references.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussà, Cristina España Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79, Hybrid. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd workshop on neural machine translation and generation*, pages 18–24.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Scheduled sampling based on decoding steps for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3296.

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.

Anthony Rousseau. 2013. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100(1):73.

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.

Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. 2020. Talking-heads attention. *arXiv preprint arXiv:2003.02436*.

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.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. Wechat neural machine translation systems for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 243–254.

Shuhan Zhou, Tao Zhou, Binghao Wei, Yingfeng Luo, Yongyu Mu, Zefan Zhou, Chenglong Wang, Xuanjun Zhou, Chuanhao Lv, Yi Jing, et al. 2021. The niutrans machine translation systems for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 265–272.