Tencent AI Lab Machine Translation Systems for the WMT20 Biomedical Translation Task

Xing Wang, Zhaopeng Tu, Longyue Wang, Shuming Shi
Tencent AI Lab, Shenzhen, China
{brightxwang,zptu,vinnylywang,shumingshi}@tencent.com

Abstract

This paper describes the Tencent AI Lab submission of the WMT2020 shared task on biomedical translation in four language directions: German⇒English, English⇒German, Chinese⇒English and English⇒Chinese. We implement our system with model ensemble technique on different transformer architectures (DEEP, HYBRID, BIG, LARGE Transformers). To enlarge the in-domain bilingual corpus, we use back-translation of monolingual in-domain data in the English language as additional in-domain training data. Our systems in German⇒English and English⇒German are ranked 1st and 3rd respectively according to the official evaluation results in terms of BLEU scores.[1]

1 Introduction

Neural machine translation (Bahdanau et al., 2015; Vaswani et al., 2017, NMT) has achieved great progress in recent years. However, as Koehn and Knowles (2017) pointed out, NMT systems suffer from poor translation performance in out-of-domain scenarios, which poses a great challenge for the biomedical translation task.

In this paper, we present our submission to the WMT20 shared task on biomedical translation task. We participated in two language directions: German-English and Chinese-English. To address the domain problem, on one hand, we adopt model ensemble technique (Liu et al., 2018) with different transformer architectures to build a more robust model. On the other hand, we enlarge the in-domain bilingual corpus with back-translation approach (Sennrich et al., 2016a).

Our contributions are as follows:

- We adopt the model ensemble technique and the back-translation approach to achieve the state-of-the-art performance on WMT19 biomedical translation task test sets.
- To promote further studies, we release some pre-trained models and the in-domain synthetic Chinese-English bilingual data for the community.

The rest of this paper is organized as follows. Section 2 presents our system with four different transformer architectures: DEEP, HYBRID, BIG, LARGE Transformers. Section 3 describes the training data used in our system, including bilingual data, monolingual data and synthetic bilingual data. Section 4 reports experimental results in two language directions. Finally, we conclude our work in Section 5.

2 System

In our systems, we adopt four different model architectures with TRANSFORMER (Vaswani et al., 2017):

- **DEEP** TRANSFORMER (Dou et al., 2018; Wang et al., 2019a; Dou et al., 2019) is the TRANSFORMER-BASE model with the 40-layer encoder.
- **HYBRID** TRANSFORMER (Hao et al., 2019b) is the TRANSFORMER-BASE model with 40-layer hybrid encoder. The 40-layer hybrid encoder stacks 35-layer self-attention-based encoder on top of 5-layer bi-directional ON-LSTM (Shen et al., 2019) encoder.
- **BIG** TRANSFORMER is the TRANSFORMER-BIG model as used by Vaswani et al. (2017).
- **LARGE** TRANSFORMER is similar to TRANSFORMER-BIG model except that it uses a 20-layer encoder.

[1]Details of our systems are introduced in https://github.com/hsing-wang/WMT2020_BioMedical

Proceedings of the 5th Conference on Machine Translation (WMT), pages 881–886
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

	DEEP	HYBRID	BIG	LARGE
Encoder Layer	40	40	6	20
Decoder Layer	6	6	6	6
Attention Heads	8	8	16	16
Embedding Size	512	512	1024	1024
FFN Size	2048	2048	4096	4096

Table 1: Hyper-parameters of different Transformer models used in our system.

The main differences between these models are presented in Table 1. Pre-Norm (Wang et al., 2019a) is adopted in above four models. All models are implemented on top of the open-source toolkit Fairseq[2]. Model ensemble is used through ensemble decoding with different model architectures.

3 Data

The data used to train our system consists of three parts: bilingual data, monolingual data and synthetic bilingual data.

3.1 Bilingual Data

In-domain bilingual data The in-domain bilingual data is provided by WMT20 biomedical translation shared task. For German-English, we choose Biomedical Translation[3] and UFAL Medical Corpus[4] to use as the in-domain training data. For Chinese-English out-of-domain (OOD) data, we adopt data selection (Axelrod et al., 2011; Liu et al., 2014) to select the in-house data (8.5M sentence pairs) as the in-domain training data.

General-domain bilingual data To alleviate the data scarce problem, we collect general-domain bilingual data from WMT20 news translation shared task[5]. For German-English, we use Europarl-v10[6], ParaCrawl-v5.1[7], News Commentary-v15[8] and Wiki Titles-v2[9]. For Chinese-English, we use CCMT Corpus[10], UN Parallel Corpus v1.0[11], News Commentary-v15[12].

3.2 Monolingual Data

As WMT20 biomedical translation shared task provides in-domain bilingual data in other language pairs, we gather in-domain monolingual data from bilingual data in other language pair. Specifically, we collect the English side of the bilingual sentence pairs from Biomedical Translation and UFAL Medical Corpus.

The statistics of the in-domain bilingual and monolingual data is listed in Table 2.

3.3 Synthetic Bilingual Data

To enlarge the in-domain bilingual corpus, we adopt back-translation method (Sennrich et al., 2016a) to generate synthetic bilingual sentence pairs. For Chinese-English, as we lack of sufficient in-domain bilingual data, we use an online translation system TranSmart[13] to translate the in-domain monolingual English back to Chinese. For German-English, we train a English-German LARGE model on the combination of in-domain and general-domain bilingual data, and use the model to generate synthetic bilingual data.

4 Experiment

We report experimental results in four language pairs: German-English (de/en), English-German (en/de), Chinese-English (zh/en) and English-Chinese (en/zh).

4.1 Experimental Setup

Data Pre-Processing We follow previous work (Saunders et al., 2019; Peng et al., 2019) to

[2] https://github.com/pytorch/fairseq (Ott et al., 2019)
[3] https://github.com/biomedical-translation-corpora/corpora
[4] https://ufal.mff.cuni.cz/ufal_medical_corpus
[5] http://www.statmt.org/wmt18/translation-task.html
[6] http://www.statmt.org/europarl/v10/
[7] https://www.paracrawl.eu/index.php
[8] http://data.statmt.org/wikititles/v2/
[9] http://data.statmt.org/news-commentary/v15/
[10] http://mteval.cipsc.org.cn:81/agreement/description
[11] https://conferences.unite.un.org/UNCorpus/
[12] http://data.statmt.org/wikititles/v2/
[13] transmart.qq.com

Corpus	File	Zh/En	De/En	En
Biomedical Translation	wmt18training/es-en	n/a	n/a	287,811
	wmt18training/fr-en	n/a	n/a	627,576
	wmt18training/pt-en	n/a	n/a	74,645
	wmt19training/de-en	n/a	40,398	40,398
	wmt19training/fr-en	n/a	n/a	75,049
	wmt19training/es-en	n/a	n/a	100,257
	wmt19training/pt-en	n/a	n/a	49,918
	wmt20training/it-en	n/a	n/a	14,756
	wmt20training/ru-en	n/a	n/a	46,782
UFAL Medical Corpus	shuffled.de-en	n/a	37,814,533	37,814,533
	shuffled.cs-en	n/a	n/a	48,243,170
	shuffled.es-en	n/a	n/a	92,999,169
	shuffled.fr-en	n/a	n/a	88,526,658
	shuffled.hu-en	n/a	n/a	48,783,611
	shuffled.pl-en	n/a	n/a	39,442,076
	shuffled.ro-en	n/a	n/a	62,034,179
	shuffled.sv-en	n/a	n/a	23,142,661

Table 2: The detailed statistics of in-domain training data used in our system. "Zh/En" and "De/En" denote the Chinese-English and German-English bilingual data, respectively. "En" denotes the monolingual English data.

use Moses scripts[14] to preprocess[15] the data and filter the bilingual data with following heuristics rules:

- Filter out duplicate sentence pairs (Khayrallah and Koehn, 2018; Ott et al., 2018).
- Filter out sentence pairs with wrong language (Khayrallah and Koehn, 2018).
- Filter out sentences pairs containing more than 120 tokens or fewer than 3.
- Filter out sentence pairs with source/target length ratio exceeding 1.5 (Ott et al., 2018).

4.2 Evaluation

For German-English, we use the Khresmoi development data as the development set, and use the sentence pairs with the correct alignment in WMT19 biomedical translation task test set as our test set. For Chinese-English, we use the in-house bilingual test set (1,000 sentence pairs) and the sentence pairs with the correct alignment in WMT19 biomedical translation task test set as development set and test set, respectively.

Follow Bawden et al. (2019), we use multi-bleu.perl from Moses[16] to compute BLEU scores and report case-sensitive BLEU scores on development and test sets.

Data Pre-processing For each language pair, we perform byte-pair encoding[17] (BPE) (Sennrich et al., 2016b) processing on the combination of in-domain bilingual data and general-domain bilingual data, and set the number of BPE merge operations to 50,000 for source and target sides, respectively.

Model Training The learning rate is set to 0.0007. All models are trained for 600K steps on 8 Tesla V100 GPUs where each is allocated with a batch size of 8192 tokens.

4.3 German-English Results

For German-English task, we first train the models on the general-domain data. Then we combine the general-domain data and the in-domain data and train the models from scratch. Finally, we introduce the synthetic bilingual data to the combination data and use all data to train the models. The model

[14] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

[15] normalize-punctuation.perl, tokenizer.perl, remove-non-printing-char.perl

[16] https://github.com/moses-smt/mosesdecoder/

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

Dataset	Size	DEEP	HYBRID	BIG	LARGE	ENSEMBLE
General Domain	37.8M	37.62	37.81	38.03	38.27	38.95
+ In-domain Data	2.5M	38.18	38.12	38.65	39.56	40.22
+ BT In-Domain Data	5.4M	38.55	38.74	38.85	40.16	40.68

Table 3: BLEU scores on the WMT19 German⇒English biomedical test set. Only the correctly aligned sentences are used in the test set.

Dataset	Size	DEEP	HYBRID	BIG	LARGE	ENSEMBLE
General Domain	19.1M	20.31	19.56	19.41	20.52	21.26
+ BT In-Domain Data	5.4M	28.52	28.83	29.32	29.80	31.34
+ OOD In-house Data	8.5M	29.92	30.07	30.66	32.05	33.23

Table 4: BLEU scores on the WMT19 Chinese⇒English biomedical test set. Only the correctly aligned sentences are used in the test set.

with best validation loss throughout the training process is selected as the final model for the testing. For model inference, the length penalty is set to 0.6 and the beam size is set to 4.

The German-English results are listed in Table 3. Our observations are:

- Due to the largest model capacity, LARGE model obtains the best translation performance among the four model variants.
- Ensemble decoding with different transformer architectures (ENSEMBLE in Table 3) achieves best translation performance.
- Leveraging in-domain bilingual data ("+In-domain") and synthetic bilingual data ("+BT In-domain") achieves significant translation improvement.

Data rejuvenation[18] (Jiao et al., 2020) is an approach which exploits the inactive training examples for neural machine translation on large-scale datasets. We adopt the data rejuvenation approach to German⇒English translation task. Experimental results are presented in Tale 7 and the data rejuvenation approach achieves significant improvement over the baseline LARGE model.

4.4 Chinese-English Results

For Chinese-English task, we gradually add the general-domain data, the synthetic bilingual data and OOD in-house data to the training data and train the models from scratch. Since the development set and test set have different data distribution, we save checkpoints every epoch and average the last 5 checkpoints rather than choose the model with best validation loss. For model inference, the length penalty is set to 2.0 and the beam size is set to 8.

Similar phenomena are observed in Chinese-English translation task. Table 4 shows Chinese-English translation results. Finally, our systems obtain 32.24 BLEU points and 33.23 BLEU points on the development and test sets, respectively.

4.5 Main Results

Main results are reported in Table 5. Our submissions (**T**encent AI Lab **M**achine **T**ranslation, TMT) with model ensemble technique achieve strong performances in WMT19 German⇔English and Chinese⇔English biomedical test sets.

5 Official Results

The official automatic evaluation results of our submissions for WMT 2020 are presented in Table 6. Our final systems rank the 1st and 3rd places on German-English and English–German, respectively, in terms of BLEU score.

6 Conclusion

In this paper, we present Tencent AI Lab machine translation systems for the WMT20 biomedical translation shared task and release the pre-trained models as well as the in-domain synthetic Chinese-English bilingual data for the research commu-

[18] https://github.com/wxjiao/Data-Rejuvenation

System	De-En	En-De	Zh-En	En-Zh
ARC (Peng et al., 2019)	38.84	35.39	32.16	37.09
UCAM (Saunders et al., 2019)	38.07	34.69	n/a	n/a
Our System	40.68	35.53	33.23	37.85

Table 5: Evaluation of translation performance on the WMT19 German⇔English and Chinese⇔English biomedical test sets. Only the correctly aligned sentences are used in the test sets.

System	De-En	En-De	Zh-En	En-Zh
Best Official	41.65	36.89	35.28	46.86
TMT Primary Run	41.65	35.24	30.48	39.43

Table 6: Official BLEU scores of our submissions for WMT20 biomedical task. Only the correctly aligned sentences are used in the test sets.

	Dev	Bio19
LARGE	52.37	39.56
+*data rejuvenation*	52.69	40.31

Table 7: Effect of data rejuvenation strategy. BLEU scores on the WMT19 German⇒English biomedical test set. Only the correctly aligned sentences are used in the test set.

nity. Our systems in German-English and English-German are ranked 1st and the 3rd respectively according to the official evaluation results in terms of BLEU scores. We also participate in the news translation (Wu et al., 2020) and the chat translation tasks (Wang et al., 2020).

In the future, we plan to explore domain adaptation (Peng et al., 2019; Saunders et al., 2019; Chu and Wang, 2018; Wang et al., 2017a), phrase modeling (Wang et al., 2017b,c; Hao et al., 2019a), structural modeling (Hao et al., 2019c; Wang et al., 2019b) strategies to improve the system performance.

7 Acknowledgments

We thank Yongchang Hao for his implementation of Hybrid TRANSFORMER model, Wenxiang Jiao for his implementation of DATA REJUVENATION, and the anonymous reviewers for their insightful suggestions.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *EMNLP*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurélie Névéol, Mariana Neves, Felipe Soares, et al. 2019. Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies. In *WMT*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *COLING*.

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang. 2018. Exploiting deep representations for neural machine translation. In *EMNLP*.

Zi-Yi Dou, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2019. Exploiting deep representations for natural language processing. *Neurocomputing*.

Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019a. Multi-granularity self-attention for neural machine translation. In *EMNLP-IJCNLP*, pages 886–896.

Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019b. Towards better modeling hierarchical structure for self-attention with ordered neurons. In *EMNLP-IJCNLP*, pages 1336–1341.

Jie Hao, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. 2019c. Modeling recurrence for transformer. In *NAACL*.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In *EMNLP*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *WMT*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *WMT*.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, et al. 2019. The niutrans machine translation systems for wmt19. In *WMT*.

Le Liu, Yu Hong, Hao Liu, Xing Wang, and Jianmin Yao. 2014. Effective selection of translation model training data. In *ACL*.

Yuchen Liu, Long Zhou, Yining Wang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. A comparable study on model averaging, ensembling and reranking in nmt. In *CCF International Conference on Natural Language Processing and Chinese Computing*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *ICML*.

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. *NAACL*.

Wei Peng, Jianfeng Liu, PRC Shenzhen, Liangyou Li, and Qun Liu. 2019. Huawei's nmt systems for the wmt 2019 biomedical translation task. *WMT*.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2019. Ucam biomedical translation at wmt19: Transfer learning multi-domain ensembles. In *WMT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2019. Ordered neurons: Integrating tree structures into recurrent neural networks. In *ICLR*.

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 *NIPS*.

Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020. Tencent AI Lab machine translation systems for the WMT20 chat translation task. In *Proceedings of the Fifth Conference on Machine Translation*.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019a. Learning deep transformer models for machine translation. In *ACL*.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017a. Instance weighting for neural machine translation domain adaptation. In *EMNLP*.

Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. 2017b. Neural machine translation advised by statistical machine translation. In *AAAI*.

Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2019b. Self-attention with structural position representations. In *EMNLP-IJCNLP*.

Xing Wang, Zhaopeng Tu, Deyi Xiong, and Min Zhang. 2017c. Translating phrases in neural machine translation. In *EMNLP*.

Shuangzhi Wu, Xing Wang, Longyue Wang, Fangxu Liu, Jun Xie, Zhaopeng Tu, Shuming Shi, and Mu Li. 2020. Tencent neural machine translation systems for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*.