Tencent AI Lab Machine Translation Systems for the WMT21 Biomedical Translation Task

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

Abstract

This paper describes the Tencent AI Lab submission of the WMT2021 shared task on biomedical translation in eight language directions: English-German, English-French, English-Spanish and English-Russian. We utilized different Transformer architectures, pre-training and back-translation strategies to improve the translation quality. Concretely, we explore mBART (Liu et al., 2020) to demonstrate the effectiveness of the pre-training strategy. Our submissions (**T**encent AI Lab **M**achine **T**ranslation, TMT) in German/French/Spanish⇒English are ranked 1st respectively according to the official evaluation results in terms of BLEU scores.

1 Introduction

This paper describes the Tencent AI Lab submission of the WMT2021 shared task on biomedical translation. Last year, we participated in three translation tasks: News (Wu et al., 2020), Chat (Wang et al., 2020a), and Biomedical (Wang et al., 2020b). In biomedical translation, we adopt DEEP TRANSFORMER (Dou et al., 2018, 2019), HYBRID TRANSFORMER (Hao et al., 2019) and DATA REJUVENATION[1] (Jiao et al., 2020). This year, we participated in eight language directions: English-German (En-De), English-French (En-Fr), English-Spanish (En-Es) and English-Russian (En-Ru) in the biomedical translation.

In this paper, we also apply the pre-train and fine-tune paradigm for the biomedical translation task. The pre-train model is first trained on the the large-scale monolingual data in a self-supervised manner, then is fine-tuned on downstream bilingual data. Specifically, we adopt the encoder-decoder pre-trained model mBART (Liu et al., 2020) to implement the pre-training strategy.

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

The rest of this paper is organized as below. Section 2 presents our system: Transformer and pre-trained model mBART. Section 3 describes the training and validation data used in our system. Section 4 reports experimental results in the participated eight language directions. Finally, we conclude our work in Section 5.

2 System

Our systems are implemented with Transformer (Vaswani et al., 2017) and the pre-trained model mBART. The training details of theses models are described in Section 4.

2.1 Transformer

We adopt the BIG and LARGE Transformer models used in the previous year (Wang et al., 2020b) as the basic Transformer models. BIG and LARGE Transformer models contain 6-layer and 20-layer encoders with TRANSFORMER-BIG setting (Vaswani et al., 2017), respectively.

2.2 Pre-train Model

For the sequence-to-sequence pre-training, we adopt mBART25 (Liu et al., 2020) as the pre-train model for our experiments, which consists of 12 encoder and decoder layers with the default size of hidden state is 1024. The model is pre-trained with the denoising objective on the large-scale monolingual data and is fine-tuned on the downstream tasks. mBART has achieved significant improvements on many low resource language paris.

3 Data

In this section, we present the training and validation data used in our system.

Besides the in-domain data provided by organisers, we collect the out-of-domain bilingual data from WMT news translation shared task.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 874–878

	En-De	En-Fr	En-Es	En-Ru
Out-of-domain	37.8M	28.0M	30.3M	92.0M
In-domain	2.5M	3.5M	1.6M	43.0K
Validation set	9.8K	1.5K	1.5K	4.0K

Table 1: The detailed statistics of training and validation data used in our system.

- En-De: Europarl-v10[2], Common Crawl corpus[3], ParaCrawl[4], News Commentary-v15[5] and Wiki Titles-v2[6].
- En-Fr: Europarl-v7[7], Common Crawl corpus, News Commentary[8], English-French Giga Corpus[9].
- En-Es: Europarl-v7[10], Common Crawl corpus, News Commentary[11], ParaCrawl[12].
- En-Ru: Common Crawl corpus, News Commentary[13], ParaCrawl[14], Yandex Corpus[15], Wiki Titles-v2, Back-translated news[16].

For the validation data, we use the Khresmoi development data[17] (En-De, En-Fr, En-Es) as the validation sets. We also use the HimL test sets 2015 and 2017[18] to enlarge the En-De validation set. For En-Ru, we randomly sample 4000 examples from the training data as the validation set.

[2] http://www.statmt.org/europarl/v10/

[3] www.statmt.org/wmt13/training-parallel-commoncrawl.tgz

[4] https://s3.amazonaws.com/web-language-models/paracrawl/release8/en-de.txt.gz

[5] http://data.statmt.org/news-commentary/v15/

[6] http://data.statmt.org/wikititles/v2/

[7] http://www.statmt.org/wmt13/training-parallel-europarl-v7.tgz

[8] http://www.statmt.org/wmt15/training-parallel-nc-v10.tgz

[9] http://www.statmt.org/wmt10/training-giga-fren.tar

[10] training-parallel-europarl-v7.tgz

[11] http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz

[12] https://s3.amazonaws.com/web-language-models/paracrawl/release8/en-es.txt.gz

[13] http://data.statmt.org/news-commentary/v16

[14] http://paracrawl.eu/download.html

[15] https://translate.yandex.ru/corpus?lang=en

[16] http://data.statmt.org/wmt20/translation-task/back-translation/

[17] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122

[18] https://www.himl.eu/test-sets

The statistics of the in-domain and out-of-domain training data and the validation data are listed in Table 1.

To enlarge the in-domain bilingual corpus, we follow Wang et al. (2020b) to adopt back-translation method to generate synthetic bilingual sentence pairs. For English-X pair, we train a English-X LARGE model on the combination of in-domain and out-of-domain data, and use the model to generate synthetic bilingual data. We also collect the En-Ru bilingual biomedical data (about 1.0 M sentence pairs) from Internet as the in-domain data.

In this work, all corpora are tokenized by sentence-piece (Kudo and Richardson, 2018) model[19] without any pre-processing procedures.

4 Experiments

For the corpus filtering, we follow Wang et al. (2020b) to filter duplicate sentence pairs (Khayrallah and Koehn, 2018), sentence pairs with wrong language (Khayrallah and Koehn, 2018) or length problem (Ott et al., 2018).

For the synthetic bilingual data generation, we adopt iterative knowledge distillation (Li et al., 2019) to improve the translation quality. Our iterative knowledge distillation is performed with 3 BIG Transformer teachers and 3 iterations. We also try to use the Right-to-Left (R2L) training (Wu et al., 2020) but fail in achieving significant improvements on the test sets.

We follow Wang et al. (2020b) to train the BIG and LARGE Transformer models. Specifically, we first use the combination of the out-of-domain data and the in-domain data to train the teacher model. Then we use the teacher model to generate the synthetic bilingual data. Finally, we train the student model on the combination of the synthetic and real bilingual data (Jiao et al., 2021). 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.

[19] https://github.com/google/sentencepiece

System	De		Fr		Es		Ru
	2019	2020	2019	2020	2019	2020	2020
Best Official 19 (Bawden et al., 2019)	38.84	–	38.24	–	48.33	–	–
Best Official 20 (Bawden et al., 2020)	–	41.65	–	44.45	–	50.75	43.31
Transformer-Big	38.66	39.15	37.32	41.92	50.63	48.22	30.89
Transformer-Large	39.41	39.64	38.12	42.77	52.58	49.26	31.92

Table 2: BLEU scores on the German/French/Spanish/Russian⇒English biomedical test sets. Only the correctly aligned sentences are used in the test set.

System	De		Fr		Es		Ru
	2019	2020	2019	2020	2019	2020	2020
Best Official 19 (Bawden et al., 2019)	35.39	–	42.41	–	48.96	–	–
Best Official 20 (Bawden et al., 2020)	–	36.89	–	43.51	–	46.72	39.36
mBART	29.96	28.47	40.13	44.04	44.79	42.92	32.23
Transformer-Big	30.43	29.56	40.33	43.58	44.23	42.87	31.96
Transformer-Large	31.60	30.89	41.04	44.01	44.68	43.05	31.79

Table 3: BLEU scores on the English⇒German/French/Spanish/Russian biomedical test sets. Only the correctly aligned sentences are used in the test set.

System	2019
Baseline	37.72
+ In-domain Data	38.14
+ Data Rejuvenation	38.47
+ Back-translation	38.66
+ Ensemble	39.14

Table 4: BLEU scores of the Transformer-Big model on the German⇒English WMT2019 biomedical test set. Only the correctly aligned sentences are used in the test set.

For the pre-train model, we adopt the publicly available mBART25[20] model and fine-tune the mBART25 on the in-domain data. In the fine-tuning phase, we minimize the label smoothed cross entropy with the smoothing factor of 0.2. We use the Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e{-}6$. The learning rate is scheduled to increase from 0 to the maximum value in the warm-up phase and decreases linearly to 0 in the remaining steps. The dropout rate is 0.3 for each residual connection and 0.1 for attention matrices.

We carry out ablation study on De→En translation task. The results are shown in Table 4. The in-domain data improves the baseline Transformer-Big model with 0.42 BLEU point. We then apply the Data Rejuvenation, Back-translation and model ensemble strategies and achieve the further improvement.

We adopt the In-domain Data, Data Rejuvenation, Back-translation as the default setting and apply the setting to Transformer-Big and Transformer-Large models on the eight language directions. We train 5 BIG and 5 LARGE Transformer models with different random seeds initialization. With the trained models, we employ the model ensemble strategy with the greedy based ensemble (Li et al., 2019; Wu et al., 2020) to get the final translation outputs. For model inference, the length penalty is set to 0.6 and the beam size is set to 4.

Translation results are reported in term of BLEU score in Table 2 and Table 3. From the tables, we find that 1) utilizing different Transformer architectures, pretraining and back-translation strategies achieve strong performance on the De→En, En↔Fr and Es→En translation tasks. 2) the lack of the large-scale in-domain data makes our En-Ru NMT system significantly lower than the state-of-the-art systems, demonstrating that the in-domain data plays a critical role in the development of NMT system.

[20] https://github.com/pytorch/fairseq/tree/master/examples/mbart

System	En-De		En-Fr		En-Es		En-Ru	
Direction	←	→	←	→	←	→	←	→
Best Official	38.16	27.76	48.05	44.65	52.99	48.52	36.23	30.78
TMT	38.16	23.32	48.05	43.90	52.99	41.57	29.98	25.43

Table 5: Official BLEU scores of our submissions for WMT21 biomedical task.

Post-process We find that several long sentences exist in the 2021 test sets, which pose a great challenge for our NMT system. Take the following two sentences for example:

Sentence 6 in doc73 in medline_fr2en_fr.txt: "Nous avons constaté que: (i) malgré le fardeau de plus en plus lourd des maladies non transmissibles, nombre de pays à faible et moyen revenu ne possédaient pas les fonds suffisants pour assurer des services de prévention; (ii) les professionnels de santé au sein des communautés manquaient fréquemment de ressources, de soutien et de formation; (iii) les frais non remboursables dépassaient 40% des dépenses de santé dans la moitié des pays étudiés, ce qui entraîne des inégalités; et enfin, (iv) les régimes d'assurance maladie étaient entravés par la fragmentation des systèmes publics et privés, le sous-financement, la corruption et la piètre mobilisation des travailleurs informels."

Sentence 3 in doc27 in medline_es2en_es.txt: "Este artículo tiene como objeto el análisis de los ensayos clínicos que permitieron dicha autorización, así como la revisión de nuevas terapias para el tratamiento del carcinoma urotelial localmente avanzado o metastásico. MÉTODO: Búsqueda bibliográfica realizada en Pub-Med y ClinicalTrials.gov mediante la combinación de las palabras clave, en español e inglés: "carcinoma urotelial", "cáncer de vejiga", "localmente avanzado","metastásico", "inmunoterapia", "CTLA-4", "PD1", "PDL-1", "atezolizumab", "nivolumab", "ipilimubab", "pembrolizumab","avelumab", "durvalumab", "tremelimumab", "terapia antiangiogénica", "terapia molecular dirigida" e "inhibidores VEGF"."

To address the problem, we manually split the long sentences into multiple sentences, and use the splitted ones as the system input to perform the translation.

We also find our system may generate wrong translations for the very short input sentences, e.g., "RéSUMé: " (Sentence 1 in doc92 in medline_fr2en_fr.txt), "(" (Sentence 4 in doc11 in medline_es2en_es.txt). To overcome the problem, we extract the target translation from the SMT phrase table and use it as the final translation output, as the NMT and SMT models are identical in modeling the bilingual knowledge (He et al., 2020).

5 Official Results

The official automatic evaluation results of our submissions for WMT 2021 biomedical translation task are shown in Table 5. Our final systems in German/French/Spanish⇒English are ranked 1st respectively, in terms of BLEU score.

6 Conclusion

In this paper, we present Tencent AI Lab machine translation systems for the WMT21 biomedical translation shared task. we participated in eight language directions: English-German (En-De), English-French (En-Fr), English-Spanish (En-Es) and English-Russian (En-Ru). Our systems German/French/Spanish⇒English are ranked 1st according to the official evaluation results in terms of BLEU scores.

It is worth mentioning that most advanced technologies reported in this paper are also adapted to our systems for news translation task (Wang et al., 2021), which achieve the 1st rank in Chinese⇒English task.

In the future, we plan to explore Non-Autoregressive machine Translation (NAT) models to improve the system performance (Zhou et al., 2020; Ding et al., 2020; Hao et al., 2021) and will integrate these advanced techniques in our Tencent TranSmart System (Huang et al., 2021)[21].

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, et al. 2019. Findings of the wmt 2019 biomedical translation shared task:

[21] https://transmart.qq.com/index

Evaluation for medline abstracts and biomedical terminologies. In *WMT*.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurelie Neveol, Mariana Neves, Maite Oronoz, et al. 2020. Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *WMT*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2020. Understanding and improving lexical choice in non-autoregressive translation. In *ICLR*.

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. 2019. Towards better modeling hierarchical structure for self-attention with ordered neurons. In *EMNLP-IJCNLP*, pages 1336–1341.

Yongchang Hao, Shilin He, Wenxiang Jiao, Zhaopeng Tu, Michael Lyu, and Xing Wang. 2021. Multi-task learning with shared encoder for non-autoregressive machine translation. In *NAACL*.

Shilin He, Xing Wang, Shuming Shi, Michael R Lyu, and Zhaopeng Tu. 2020. Assessing the bilingual knowledge learned by neural machine translation models. *arXiv*.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: a practical interactive machine translation system. *arXiv*.

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*.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *ACL*.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. of EMNLP*.

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*.

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. *TACL*.

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

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, Mu Li, Fangxu Liu, Shuming Shi, Zhaopeng Tu, Xing Wang, Shuangzhi Wu, Jiali Zeng, and Wen Zhang. 2021. Tencent translation system for the WMT21 news translation task. In *WMT*.

Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020a. Tencent AI lab machine translation systems for WMT20 chat translation task. In *WMT*.

Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2020b. Tencent AI lab machine translation systems for the WMT20 biomedical translation task. In *WMT*.

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 *WMT*.

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Improving autoregressive NMT with non-autoregressive model. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*.