HW-TSC's Submissions to the WMT21 Biomedical Translation Task

Hao Yang[1], **Zhanglin Wu**[1], **Zhengzhe Yu**[1], **Xiaoyu Chen**[1], **Daimeng Wei**[1], **Zongyao Li**[1], **Hengchao Shang**[1], **Jiaxin Guo**[1], **Minghan Wang**[1], **Lizhi Lei**[1], **Chuanfei Xu**[1], **Min Zhang**[1], **Ying Qin**[1],
[1]Huawei Translation Service Center, Beijing, China
{yanghao30,wuzhanglin2,yuzhengzhe,chenxiaoyu35,weidaimeng,
lizongyao,shanghengchao,guojiaxin1,wangminghan,
leilizhi,xuchuanfei,zhangmin186,qinying}@huawei.com

Abstract

This paper describes the submission of Huawei Translation Service Center (HW-TSC) to WMT21 biomedical translation task in two language pairs: Chinese↔English and German↔English (Our registered team name is HuaweiTSC). Technical details are introduced in this paper, including model framework, data pre-processing method and model enhancement strategies. In addition, using the wmt20 OK-aligned biomedical test set, we compare and analyze system performances under different strategies. On WMT21 biomedical translation task, Our systems in English→Chinese and English→German directions get the highest BLEU scores among all submissions according to the official evaluation results.

1 Introduction

We have witnessed great progress made by neural machine translations (Bahdanau et al., 2015; Vaswani et al., 2017) in recent years. However, domain adaptation remains to be a tough issue. As noted by Koehn and Knowles (Koehn and Knowles, 2017), translations by NMT systems in out-of-domain scenarios are relatively poor, and high-quality data in specific domains are difficult to obtain, which pose great challenges to certain translation tasks (e.g. biomedical translation). To address the domain adaptation issue, on one hand, we leverage data diversification (Nguyen et al., 2020), forward translation (Wu et al., 2019) and back translation (Sennrich et al., 2016a; Edunov et al., 2018) to generate synthetic in-domain corpora. On the other hand, fine-tuning (Sun et al., 2019) and ensemble (Freitag et al., 2017; Li et al., 2019) are used to further enhance system performances in the biomedical domain.

We introduce our data strategy in section 2, and model architectures as well as model enhancement techniques in section 3. Section 4 presents experimental results of both language pairs on the wmt20 OK-aligned biomedical test set. Section 5 is a conclusion of our work.

2 Dataset

2.1 Data Source

Our baseline model is trained with out-of-domain WMT21 news data. The sizes of bilingual and monolingual data for Chinese↔English and German↔English language pairs are shown in Table 1.

With regard to in-domain data, we use both the bilingual data and monolingual data provided by the WMT21 Biomedical Translation Shared task. For German↔English task, we select Biomedical Translation and UFAL Medical Corpus as in-domain training data. Besides, 21.43M in-house English monolingual data are used. For Chinese↔English task, the used in-house data includes: 1.35M parallel data, 21.43M English monolingual data, and 36.11M Chinese monolingual data. Table 2 shows the details of data in the biomedical domain for German↔English and Chinese↔English tasks.

2.2 Data Pre-processing

Our data pre-processing methods include:

- Filter out repeated sentences (Khayrallah and Koehn, 2018; Ott et al., 2018).
- Normalize punctuations using Moses (Koehn et al., 2007).
- Filter out sentences with repeated fragments.
- Filter out sentences with mismatched parentheses and quotation marks.
- Filter out sentences of which punctuation percentage exceeds 0.3.
- Filter out sentences with the character-to-word ratio greater than 12 or less than 1.5.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 879–884
November 10–11, 2021.

Language	corpus WMT21 News Shared Task's Corpus	Mono English	German	Chinese
German↔English	96.6M	150M	150M	-
Chinese↔English	16.5M	150M	-	150M

Table 1: Out-domain data size of WMT21 Biomedical Translation Task

Language	corpus Biomedical Translation && UFAL	In-house Corpus	Mono English	German	Chinese
German↔English	3.06M	-	21.43M	-	-
Chinese↔English	-	1.35M	21.43M	-	36.11M

Table 2: In-domain data size of WMT21 Biomedical Translation Task

- Filter out sentences with more than 120 words.
- Apply langid (Joulin et al., 2017, 2016) to filter sentences in other languages.
- Use fast-align (Dyer et al., 2013) to filter sentence pairs with poor alignment.

It should be noted that for the German↔English translation task, we employ joint SentencePiece model(SPM) (Kudo and Richardson, 2018; Kudo, 2018) for word segmentation, with the size of the vocabulary set to 32k. As for the Chinese↔English translation task, Jieba tokenizer is used for Chinese word segmentation while Moses tokenizer for English word segmentation. Byte Pair Encoding (BPE) (Sennrich et al., 2016b) is adopted for Chinese and English sub-word segmentation. We train BPE models with 32,000 merge operations for both the source and target sides.

3 System overview

3.1 Model

Our system uses Transformer (Vaswani et al., 2017) model architecture, which adopts full self-attention mechanism to realize algorithm parallelism, accelerate model training speed, and improve translation quality. Two Transformer deep-large model architectures are used in our experiments:

- Deep 25-6 (Wang et al., 2018; Li et al., 2019): Based on the Transformer-base model architecture, the deep 25-6 model features 25-layer encoder, 6-layer decoder, 1024 dimensions of word vector, 4096-hidden-state, 16-head self-attention and layer normalization.
- Deep 35-6 (Wu et al., 2020; Sun et al., 2019): Based on the Transformer-base model architecture, the deep 35-6 model features 35-layer encoder, 6-layer decoder, 788 dimensions of word vector, 3072-hidden-staten, 16-head self-attention and layer normalization.

We use the open-source Fairseq (Ott et al., 2019) for training. The main parameters are as follows: Each model is trained using 8 GPUs. The size of each batch is set as 2048, parameter update frequency as 32, learning rate as 5e-4 (Vaswani et al., 2017) and label smoothing as 0.1 (Szegedy et al., 2016). The number of warmup steps is 4000, and the dropout is 0.1. We also use the Adam optimizer (Kingma and Ba, 2015) with $\beta 1 = 0.9$, $\beta 2 = 0.98$. In the inference phase, The beam-size is 8, The length penalties for Chinese→English and German→English are set to 0.5, and the length penalties for the other two directions are set to 1.5.

3.2 Data augmentation

Given the small size of in-domain bilingual data, how to generate more training data becomes a crucial issue for model performance enhancement in the biomedical field. We adopt three data augmentation methods:

- Data diversification (Nguyen et al., 2020): Data diversification is a simple but effective strategy to enhance the performance of NMT. It uses predictions from multiple forward and backward models and then combines the results with raw data to train the final NMT model. The method does not require additional monolingual data and is suitable for all types of NMT models. It is more efficient than knowledge distillation and dual learning, and exhibits strong correlation with model integration. In our Chinese↔English and German↔English systems, we use only the forward model and the backward model to

create synthetic data and add the data to the original parallel corpora.

- Forward translation (Wu et al., 2019): Forward translation usually refers to using source language monolinguals to generate synthetic data through beam search decoding, and then add synthetic data to the training data so as to increase the training data size. Although merely using forward translation may not work well, forward translation can be used in conjunction with a back translation strategy, which also works better than using back translation alone. We do not use forward translation for the German→English system task due to the lack of high-quality in-domain German monolinguals. We then give up forward translation for the English→German direction because forward translation and back translation cannot be used jointly for better effects. Ultimately, we only adopt forward translation for our Chinese↔English systems.

- Back translation (Edunov et al., 2018): Back translation translates target side monolingual data back to the source language so as to increase the training data size, which has been proved to be an effective method to improve neural machine translation performances. There are many methods for generating synthetic corpus through back translation. In a non-extremely low-resource scenario, sampling or noisy beam search decoding method is more effective than beam search or greedy search, and the synthetic data generated by sampling or noisy beam search decoding method may introduce more diversity to training data. In our experiment, sampling decoding is adopted. We use back translation for all directions expect English→German, due to the lack of high-quality in-domain German monolinguals.

3.3 Training strategy

We first use in-domain training data to conduct incremental training with baseline models trained by WMT21 news data for domain transfer. Then, we use three monolingual enhancement strategies, data diversity, forward translation and back translation, to create synthetic data and add them to the in-domain training data to further expand the scale of the training data, and then perform incremental training again. In addition, we fine-tune our models with test sets from previous years of the same task in hope of further improving in-domain performances. Specifically, we ensemble multiple models to forward translate the source side of test sets to increase the size of the training data, and then add noise (Meng et al., 2020) to the target side of the training data to achieve a better fine-tuning effect. Finally, multiple models are ensembled to achieve better performance.

Algorithm 1: Strategies for selecting ensemble models

Input : The list of all NMT models to be selected $M := [M^1, ..., M^n]$, n is the Number of M, and the test Set T;
Output : The optimal model combination $B := [M^i, ..., M^j]$;

```
Initialize the test set T's maximum BLEU score maxbleu := 0;
Initialize the optimal model combination B := [];
for num ∈ range(1, n) do
    Generate a list of model combination numlist, which is all possible combination of num models in M;
    for current model combination subnumlist ∈ numlist do
        Calculate the current BLEU score curbleu of the current combined model on the test set T.;
        if curbleu > maxbleu: then
            B := subnumlist
            maxbleu := curbleu
        end
    end
end
return B
```

3.4 Ensemble

For each translation task, we randomize two sets of training data and train four models using the two model architectures mentioned above. In the course of our experiments, we find that directly ensemble all models does not necessarily perform better on test set than a single model. To achieve a better ensemble effect, we design an algorithm, as shown in the algorithm 1. The core idea is to traverse all combinations of models and find the best one in the

System	English→Chinese	Chinese→English
baseline	40.0	28.3
+ biomedical corpus	44.5 (+4.5)	31.4 (+3.1)
+ data diversification	44.9 (+0.4)	32.3 (+0.9)
+ forward translation & back translation	45.8 (+0.9)	34.6 (+2.3)
+ fine-tuning	47.6 (+1.8)	35.9 (+1.3)
+ ensemble	**47.7** (+0.1)	**36.4** (+0.5)
WMT20 best official	46.9	35.3

Table 3: Chinese↔English BLEU scores on the WMT20 OK-aligned biomedical test set.

test set. The experiment results show that ensemble with the best combination found by the traverse strategy is much better than simply ensemble all models. In our experiment, the model combination that performs best on the wmt20 OK-aligned biomedical test set is used as the final submission.

4 Experimental result

We train baseline models using WMT21 news data, then incrementally train them using medical bilingual corpora and synthetic data generated by data augmentation techniques, fine-tune models with previous years' test sets, and finally ensemble multiple models to produce submitted results. We benchmark our submissions using the WMT20 OK-align test set. BLEU scores are calculated using the MTEVAL script from Moses (Koehn et al., 2007). The results are shown in Table 3 and Table 4. Our models outperform last year's official best results in three language directions. The tables mainly show the results of deep 35-6 models. Only in the last ensemble phase, multiple model architectures are used. we compare our results with the best official results from last year. We notice that our baseline models trained by WMT news data may also perform quite well in the biomedical field. For example, in German→English, Our baseline model is only 2.2 BLEU below last year's best result.

4.1 Chinese↔English

For Chinese↔English task, we first train the baseline model on WMT21 news data. Then, incremental training is conducted with in-domain bilingual and synthetic data. Finally, models are fine-tuned with the previous test sets, and multiple models are ensembled to produce the final result. The experimental results of Chinese↔English are shown in Table 3. Compared with the baseline model, the final systems achieve improvements of 8.1 BLEU and 7.7 BLEU on Chinese→English and English→Chinese directions, respectively. Incremental training alone leads to increases of 3.1 BLEU and 4.5 BLEU on Chinese→English and English→Chinese respectively. Besides, the combination of data diversity, forward translation, and back translation also lead to significant improvements (3.2 BLEU increase for Chinese→English and 1.3 BLEU for the opposite direction). Fine-tuning on previous test sets further improves the model quality by 1.3 BLEU for Chinese→English and 1.8 BLEU for English→Chinese. Notably, no further improvements is achieved by ensemble all models, while ensemble the model combinations found through the ergodic approach further improves translation quality by 0.5 BLEU and 0.1 BLEU on Chinese→English and English→Chinese, respectively. Ultimately, on Chinese↔English task, our results outperform last year's official best results.

4.2 German↔English

For German↔English task, the model training strategy used is similar to that for Chinese↔English task, except data augmentation techniques. As mentioned above, due to the lack of in-domain German monolingual data, we use data diversity and back translation strategies for German→English direction and only data diversity for English→German direction. The German↔English experiment results are shown in Table 4. Data augmentation results in significant performance improvements, with 1.1 BLEU and 1.7 BLEU on German→English and English→German respectively. Fine-tuning with previous years' test sets has also improved the quality of in-domain translations. On German→English, we fine-tune the model with wmt18 and wmt19 test sets and see an improvement of 1.1 BLEU. On English→German, fine-tuning leads to an increase of 0.4 BLEU.

System	English→German	German→English
baseline	33.8	39.5
+ biomedical corpus	34.9 (+1.1)	39.8 (+0.3)
+ data diversification	35.5 (+0.6)	40.4 (+0.6)
+ back translation	-	40.6 (+0.2)
+ fine-tuning	35.9 (+0.4)	41.7 (+1.1)
+ ensemble	36.5 (+0.6)	**42.4** (+0.7)
WMT20 best official	**36.9**	41.7

Table 4: German↔English BLEU scores on the WMT20 OK-aligned biomedical test set.

Ensemble the model combinations found through the ergodic approach contribute to 0.7 BLEU increase for German→English and 0.6 BLEU for English→German. Ultimately, due to the lack of effective in-domain German monolingual data, we only surpass last year's official best results on German→English direction.

5 Conclusion

This paper presents the submissions of HW-TSC to the WMT21 Biomedical Translation Task. We perform experiments with a series of pre-processing and training strategies. The effectiveness of each strategy is demonstrated by our experiment results. Combining with data augmentation strategies, incremental training with in-domain data on the basis of a baseline model from new domain can effectively improve in-domain translation quality. Our systems in English→Chinese and English→German directions get the highest BLEU scores among all submissions according to the official evaluation results.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

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.

Markus Freitag, Y. Al-Onaizan, and B. Sankaran. 2017. Ensemble distillation for neural machine translation. *ArXiv*, abs/1702.01802.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jegou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models.

Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *ACL (1)*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP (Demonstration)*.

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 *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266.

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, et al. 2020. Wechat neural machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 239–247.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation.

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. 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. 2016a. Improving neural machine translation models with monolingual data. In *ACL (1)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 374–381.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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.

Qiang Wang, Bei Li, Jiqiang Liu, Bojian Jiang, Zheyang Zhang, Yinqiao Li, Ye Lin, Tong Xiao, and Jingbo Zhu. 2018. The niutrans machine translation system for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 528–534.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. 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 4207–4216.

Liwei Wu, Xiao Pan, Zehui Lin, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2020. The volctrans machine translation system for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 305–312.