The Path to Continuous Domain Adaptation Improvements by HW-TSC for the WMT23 Biomedical Translation Shared Task

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,weidaimeng,lizongyao,yuzhengzhe,lishaojun18,chenxiaoyu35,
shanghengchao,guojiaxin1,xieyuhao2,leilizhi,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper presents the domain adaptation methods adopted by Huawei Translation Service Center (HW-TSC) to train the neural machine translation (NMT) system on the English↔German (en↔de) WMT23 biomedical translation task. Our NMT system is built on deep Transformer with larger parameter sizes. Based on the biomedical NMT system trained last year, we leverage Curriculum Learning, Data Diversification, Forward translation, Back translation, and Transductive Ensemble Learning to further improve system performance. Overall, we believe our submission can achieve highly competitive result in the official final evaluation.

1 Introduction

Machine translation (MT) (Lopez, 2008) refers to the automatic translation of text from one language to another. The WMT23 biomedical translation task aims to evaluate the performance of MT systems in the biomedical domain. Due to the lack of sufficient in-domain data, domain adaptation (Chu and Wang, 2018; Wu et al., 2023) has naturally become the main research direction of this task.

This paper presents the domain adaptation methods adopted by HW-TSC to train the NMT (Bahdanau et al., 2015) system on en↔de language pair of the WMT23 biomedical translation task. Our method is mainly based on previous works (Wei et al., 2022, 2021; Yang et al., 2021). We try to train a domain classifier to select biomedical data from general data, then perform multi-step data cleaning on the selected in-domain data and keep only a high-quality subset for training. Based on the biomedical NMT system trained last year, we leverage Curriculum Learning (Zhang et al., 2019), Data Diversification (Nguyen et al., 2020), Forward Translation (Abdulmumin, 2021), Back Translation (Sennrich et al., 2016), and Transductive Ensemble Learning (Wang et al., 2020b) to further improve system performance.

Our system report includes four parts. Section 2 focuses on our data processing strategies while section 3 describes our training details. Section 4 explains our experiment settings and training processes, and section 5 presents the results.

2 Data

2.1 Data Volume

We obtain bilingual and monolingual data from various data sources, except medical database. Then, we use biomedical data and general data to train a domain classifier based on fasttext (Joulin et al., 2016) to select biomedical data from general data. Table 1 lists the final size of the training data.

language pairs	bitext data	monolingual data
en↔de	11.6M	en: 12.3M, de: 10.1M

Table 1: Bilingual and monolingual used for training.

2.2 Data Pre-processing

Our data processing procedure is basically the same as our method last year (Wu et al., 2022), including deduplication, XML content processing, langid (Lui and Baldwin, 2012) and fast-align (Dyer et al., 2013) filtering strategies, etc. As we use the same data pre-processing strategy as last year's, we will not go into details here.

2.3 Data Denoising

Since there may be some semantically dissimilar sentence pairs in bilingual data, we use LaBSE (Feng et al., 2022) to calculate the semantic similarity of each bilingual sentence pair, and exclude bilingual sentence pairs with a similarity score lower than 0.7 from the training corpus.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 271–274
December 6–7, 2023.

3 System Overview

3.1 Model

We continue using Transformer (Vaswani et al., 2017) as our neural machine translation (NMT) model architecture. As we did last year, we use a 25-6 deep model architecture. The parameters of the model are the same as Transformer-big. We just change the post-layer normalization to the pre-layer normalization, and set encoder layers to 25.

3.2 Curriculum Learning

A practical curriculum learning (CL) (Zhang et al., 2019) method should address two main questions: how to rank the training examples, and how to modify the sampling procedure based on this ranking. For ranking, we choose to estimate the difficulty of training samples according to their domain feature (Wang et al., 2020a). The calculation formula of domain feature is as follows, where θ_{in} represents an in-domain NMT model, and θ_{out} represents an out-of-domain NMT model.

$$q(x, y) = \frac{\log P(y|x; \theta_{in}) - \log P(y|x; \theta_{out})}{|y|} \quad (1)$$

For sampling, we adopt a probabilistic CL strategy[1] that takes advantage of the spirit of CL in a nondeterministic fashion without discarding the good practice of original standard training, like bucketing and mini-batching.

3.3 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a data augmentation method to boost NMT performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset on which the final NMT model is trained. DD is applicable to all NMT models. It does not require extra monolingual data, nor does it add more computations or parameters. To conserve training resources, we only use one forward model and one backward model when performing DD.

3.4 Forward Translation

Forward translation (FT) (Abdulmumin, 2021), also known as self-training, is one of the most commonly used data augmentation methods. FT has proven effective for improving NMT performance by augmenting model training with synthetic parallel data. Generally, FT is performed in three steps: (1) randomly sample a subset from the large-scale source-side monolingual data; (2) use a "teacher" NMT model to translate the subset data into the target language to construct the synthetic parallel data; (3) combine the synthetic and authentic parallel data to train a "student" NMT model.

3.5 Back Translation

An effective method to improve NMT with target monolingual data is back translation (BT) (Sennrich et al., 2016; Wei et al., 2023). There are many works broaden the understanding of BT and investigates a number of methods to generate synthetic source sentences. Edunov et al. (2018) find that back translations obtained via sampling or noised beam outputs are more effective than back translations generated by beam or greedy search in most scenarios. Caswell et al. (2019) show that the main role of such noised beam outputs is not to diversify the source side, but simply to tell the model that the given source is synthetic. Therefore, they propose a simpler alternative strategy: Tagged BT. This method uses an extra token to mark back translated source sentences, which generally outperforms noised BT. For better joint use with FT, we use sampling back translation (ST).

3.6 Transductive Ensemble Learning

Ensemble learning (Garmash and Monz, 2016), which aggregates multiple diverse models for inference, is a common practice to improve the performance of machine learning models. However, it has been observed that the conventional ensemble methods only bring marginal improvement for NMT when individual models are strong or there are a large number of individual models. Transductive Ensemble Learning (TEL) (Zhang et al., 2019) studies how to effectively aggregate multiple NMT models under the transductive setting where the source sentences of the test set are known. TEL uses all individual models to translate the source test set into the target language space and then fine-tune a strong model on the translated synthetic data, which significantly boosts strong individual models and benefits a lot from more individual models.

[1] https://github.com/kevinduh/sockeye-recipes/tree/master/egs/curriculum

4 Experiment Settings

We use the open-source fairseq (Ott et al., 2019) for training, then we use SacreBLEU (Post, 2018) and multi-eval tool [2] to measure system performances. The main parameters are as follows: each model is trained using 8 A100 GPUs, batch size is 6144, parameter update frequency is 1, and learning rate is 5e-4. The number of warmup steps is 4000, and model is saved every 1000 steps. The architecture we used is described in section 3.1. We adopt dropout (Srivastava et al., 2014), and the rate varies across different training phases. When the training data is higher than tens of millions, the dropout ratio is set to 0.1, otherwise it is set to 0.3.

5 Results

Regarding en↔de, we use Curriculum Learning (CL), Data Diversification (DD), Forward Translation (ft), Back Translation (BT), and Transductive Ensemble Learning (TEL). The evaluation results of en→de and de→en NMT system on WMT22 biomedical test set are shown in Tables 2.

We see that CL can stably bring 3 SacreBLEU and multi-eval improvement, while DD, FT & ST and TEL can further slightly improve SacreBLEU and multi-eval. Our final en→de and de→en submissions achieve 40.48 and 48.75 SacreBLEU, 41.22 and 49.91 multi-eval respectively.

	en→de		de→en	
	SacreBLEU	multi-eval	SacreBLEU	multi-eval
last year's baseline	37.11	37.80	44.45	45.50
+ CL	40.11	40.89	47.77	48.89
+ DD, FT & ST	40.23	41.00	48.60	49.76
+ TEL	**40.48**	**41.22**	**48.75**	**49.91**

Table 2: BLEU scores of en→de and de→en NMT system on WMT22 biomedical test set.

6 Conclusion

This paper presents the submission of HW-TSC to the WMT23 biomedical translation task. We participate in en↔de language pair and perform a series of domain adaptation experiments based on the biomedical NMT system trained last year. The effectiveness of each domain adaptation method is demonstrated. Our experiments show that domain adaptation methods are effective for model training.

[2] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/generic/mteval-v14.pl

References

Idris Abdulmumin. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers*, volume 1350, page 355. Springer Nature.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

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*, page 489. Association for Computational Linguistics.

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.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020a. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2020b. Transductive ensemble learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6291–6298.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7944–7959, Toronto, Canada.

Zhanglin Wu, Zongyao Li, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Xiaoyu Chen, Zhiqiang Rao, Zhengzhe Yu, Jinlong Yang, Shaojun Li, et al. 2023. Improving neural machine translation formality control with domain adaptation and reranking-based transductive learning. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 180–186.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022. Hw-tsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, et al. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.