HW-TSC's Submissions to the WMT23 Discourse-Level Literary Translation Shared Task

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

Abstract

This paper introduces HW-TSC's submission to the WMT23 Discourse-Level Literary Translation shared task. We use standard sentence-level transformer as a baseline, and perform domain adaptation and discourse modeling to enhance discourse-level capabilities. Regarding domain adaptation, we employ Back-Translation, Forward-Translation and Data Diversification. For discourse modeling, we apply strategies such as Multi-resolutional Document-to-Document Translation and TrAining Data Augmentation.

1 Introduction

Transformer architectures (Vaswani et al., 2017) have achieved outstanding performance on sentence-level machine translation tasks, but still have some shortcomings when it comes to discourse-level machine translation. Particularly, for machine translation scenarios that are highly discourse-dependent, such as novel translation and conversation translation, the performance is unsatisfactory.

This paper presents the submission of HW-TSC to the WMT23 Discourse-Level Literary Translation shared task. We utilize an effective data cleaning pipeline summarized in our previous works (Wei et al., 2022; Wu et al., 2022; Yang et al., 2021) to process the training data. We employ Regularized Dropout, Forward Translation, Back Translation, Data Diversification to train a strong baseline. On top of the baseline, we apply strategies including Multi-resolutional doc2doc Translation (MR-doc2doc), TrAining Data Augmentation (TADA) to enhance discourse-level translation capabilities.

The general translation model does not work well in novel translation. We found that the biggest factor affecting the quality of translation is domain adaptation; however, domain adaptation cannot solve the consistency of named entity such as names, addresses, and zero pronoun in novel translation. The consistency needs to be optimized by using strategies such as MR-doc2doc and TADA.

2 Data

2.1 Data Source

We use the same training data as that for the general MT shared task to train a sentence-level baseline. Then We use GuoFeng Webnovel Corpus[1] (Wang et al., 2023) and web-crawled novel data for domain adaptation and discourse-level capability enhancement. The data size is shown in Table 1.

	Bilingual	Source	Target
General MT	25M	50M	50M
GuoFeng Webnovel Corpus	1.9M	-	-
web-crawled novel data	10M	100M	400M

Table 1: Bilingual and monolingual data used for training.

2.2 Data Pre-processing

The data preprocessing pipeline follows our previous work (Wei et al., 2021), including deduplication, XML content processing, langid (Lui and Baldwin, 2012) and fast-align (Dyer et al., 2013) filtering strategies, etc. We will not repeat the details here.

3 System Overview

3.1 Sentence-level baseline

We directly employ the model we trained for the general MT shared task as the sentence-level baseline in this task. The following is the strategy we use to train the sentence-level baseline.

[1] http://www2.statmt.org/wmt23/literary-translation-task.html

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 302–306
December 6–7, 2023.

3.1.1 Regularized Dropout

Regularized Dropout (R-Drop) (Wu et al., 2021) improves performance over standard dropout, especially for recurrent neural networks on tasks with long input sequences. It ensures more consistent regularization while maintaining model uncertainty estimates. The consistent masking also improves training efficiency compared to standard dropout. Overall, Regularized Dropout is an enhanced dropout technique that often outperforms standard dropout.

3.1.2 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a simple but effective strategy to boost neural machine translation (NMT) (Bahdanau et al., 2015) 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. This method is more effective than knowledge distillation and dual learning.

3.1.3 Forward Translation

Forward translation (FT) (Abdulmumin, 2021) uses source-side monolingual data to improve model performance. The general procedure of FT involves three steps: (1) randomly sampling a subset from large-scale source monolingual data; (2) using a "teacher" NMT model to translate the subset into the target language, thereby constructing synthetic parallel data; and (3) combining the synthetic and authentic parallel data to train a "student" NMT model.

3.1.4 Back Translation

Augmenting parallel training data with back-translation (BT) (Sennrich et al., 2016; Wei et al., 2023) has been shown effective for improving NMT using target monolingual data. Numerous works have expanded the understanding of BT and investigated various approaches to generate synthetic source sentences. Edunov et al. found that back-translations obtained via sampling or noised beam outputs tend to be more effective than those via beam or greedy search in most scenarios. For optimal joint use with FT, we employ sampling back-translation (ST) (Edunov et al., 2018).

3.1.5 Alternated Training

While synthetic bilingual data has been shown effective for NMT, adding more synthetic data may deteriorate performance as synthetic data inevitably contains noise and errors. To address this issue, alternated training (AT) (Jiao et al., 2021) introduces authentic data as guidance to prevent model training from being disturbed by noisy synthetic data. AT views synthetic and authentic data as two types of different approximations for the authentic data distribution. The key idea is to iteratively alternate between synthetic and authentic data during training until convergence. Authentic data provides guidance to overcome noise in synthetic data. By alternating data types, AT ensures the usage of a large amount of synthetic data while prevents model deterioration from noisy data.

3.1.6 Curriculum Learning

A practical curriculum learning (CL) (Zhang et al., 2019) approach for NMT should address two key issues: ranking training examples by difficulty, and modifying the sampling procedure based on ranking. For ranking, we estimate example difficulty using domain features (Wang et al., 2020). The domain feature is calculated as:

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

Where θ_{in} is an in-domain NMT model, while θ_{out} is an out-of-domain model. The novel domain is treated as in-domain.

We fine-tune the model on the valid set to get the teacher model and select top 40% of the highest scoring data for finetuning.

3.2 Domain Adaptation

We found that the translation style of novel translation and general domain translation is completely different, so domain adaptation is very important. So we finetune the sentence-level baseline model with bilingual/monolingual novel data. For web-crawled novel data, we use 100M Chinese monolingual data and 400M English monolingual data to construct FT and ST corpus respectively, and use GuoFeng Webnovel Corpus bilingual 1.9M, web-crawled novel data bilingual 10M, finally mix the four parts data together and shuffle them.

3.3 Discourse Modeling

Although the translation quality has improved with domain adaptation, it still unable to solve document-level translation problems such as NE

consistency and zero pronoun translation. MR-doc2doc and TADA need to be used to solve the problem. It is expected to further improve the ability of discourse-level translation on the basis of section 3.2. We employ monolingual and bilingual novel data, and reconstruct them according to the method of discourse-level translation.

3.3.1 Multi-resolutional doc2doc

Multi-resolutional doc2doc (MR-doc2doc) (Sun et al., 2020) is a document-level neural machine translation approach that operates on different granularities of the document. It utilizes both sentence-level and document-level information during translation to improve context modeling and overall translation quality. Specifically, we split each document averagely into kparts multiple times and collect all the sequences together. For example, a document containing eight sentences will be split into two four-sentences segments, four two-sentences segments, and eight single sentence segments. Finally, fifteen sequences are all gathered and fed into sequence-to-sequence training. In this way, the model can acquire the ability to translate long documents since it is assisted by easier, shorter sentences and paragraphs. By doing so, the model can acquire discourse-level translation capabilities.

3.3.2 TrAining Data Augmentation

The key idea of TrAining Data Augmentation (Ailem et al., 2021) is to use tags to mark words or phrases that needs to be constrained in the source sentence during translation. When the model encounters a tagged token in the source, it is biased towards directly copying the expected lexical constraint following the tagged source word into the target output. This allows enforcing lexical constraints without changing the core NMT architecture, simply by using tags in the source. The model learns this copy behavior during training when exposed to tagged source sentences and the expected lexical constraints in the target. Thus, the approach can easily guide NMT to satisfy terminology constraints by just tagging the source sentence appropriately. It provides a simple and efficient way to constrain NMT output lexicons by merely adding tags on the source side. We use this method to ensure consistent translation of named entities (such as person names, location names, etc.) at both inference and training phases.

4 Experiments

4.1 Experiment Settings

We use SacreBLEU (Post, 2018) to measure system performances. The main parameters are as follows: the model is transformer-big with 25 encoder layers and 6 decoder layers. It trained using 8 A100 GPUs, batch size is 8192, parameter update frequency is 1, and learning rate is 5e-4. The number of warmup steps is 4000, and the model is saved every 1000 steps. We adopt dropout, and the rate varies across different training phases. R-Drop is used in model training, and we set λ to 5. We use fairseq (Ott et al., 2019) for training.

4.2 Testing Datasets

4.2.1 Simple Set

Simple Set[2] (Wang et al., 2023) contains unseen chapters in the same web novels as the training data.

4.2.2 Difficult Set

Difficult Set[3] (Wang et al., 2023) contains chapters in different web novels from the training data.

5 Results

As shown in Table 2, each step is fine-tuned based on the model from the previous step. In the Domain Adaptation stage (ST, ST & FT & AT & DD, CL), we observe significant s-BLEU improvement, while d-bleu is also improved. In the Discourse Modeling stage, MR-doc2doc can improve both s-bleu and d-bleu. TADA works well on NE consistency, but does not significantly improve d-BLEU and leads to a slight decrease in s-bleu.

As shown in Table 3, We extracted 75 NEs by W2NER (Li et al., 2022) from the test set, which occurred 1241 times in total. We count the word frequency of consecutive and identical NEs as an indicator to evaluate the consistency. We found that the TADA strategy can bring significant improvements in NE consistency.

6 Conclusion

It was believed that algorithm enhancement can make the model handle long inputs so that discourse-level translation would be improved. However, it can only achieve slight improvement.

[2] http://www2.statmt.org/wmt23/literary-translation-task.html

[3] http://www2.statmt.org/wmt23/literary-translation-task.html

	Simple		Diffcult	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU
sentence-level baseline + R-Drop	26.63	15.87	23.47	12.97
+ ST	29.36	22.63	26.02	17.74
+ ST & FT & AT & DD	29.49	26.52	25.97	21.54
+ CL	30.96	26.52	27.4	21.92
+ MR-doc2doc	30.71	26.99	27.27	22.16
+ TADA	30.58	**27.27**	27.12	**22.48**

Table 2: BLEU scores of zh→en NMT system on WMT23 web fiction test set.

models	NE consistency accuracy
sentence-level baseline	43.3%
MR-doc2doc	67.0%
TADA	71.8 %

Table 3: NE consistency accuracy of zh→en NMT system on WMT23 web fiction test set.

It is more important to achieve domain adaptation first for the sentence-level model. The discourse-level translation strategy can get the best performance based on domain adaptation.

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.

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. *Le Centre pour la Communication Scientifique Directe - HAL - Diderot,Le Centre pour la Communication Scientifique Directe - HAL - Diderot*.

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*.

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.

Rui Jiao, Zonghan Yang, Maosong Sun, and Yang Liu. 2021. Alternated training with synthetic and authentic data for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1828–1834.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 10965–10973.

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).

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2020. Capturing longer context for document-level neural machine translation: A multi-resolutional approach. *Cornell University - arXiv,Cornell University - arXiv*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*.

Longyue Wang, Zefeng Du, DongHuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Shuming Shi, and Zhaopeng Tu. 2023. Guofeng: A discourse-aware evaluation benchmark for language understanding, translation and generation.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020. 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.

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.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

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.