HW-TSC Translation Systems for the WMT22 Biomedical Translation Task

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, Yuanchang Luo, Yuhao Xie, Miaomiao Ma, Ting Zhu, Lizhi Lei, Song Peng, Hao Yang, Ying Qin
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,yangjinlong7,raozhiqiang,yuzhengzhe,weidaimeng,chenxiaoyu35,
lizongyao,shanghengchao,lishaojun18,zhuming47,luoyuanchang,xieyuhao2,
mamiaomiao,zhuting20,leilizhi,pengsong2,yanghao30,qinying}@huawei.com

Abstract

This paper describes the translation systems trained by Huawei translation services center (HW-TSC) for the WMT22 biomedical translation task in five language pairs: English↔German (en↔de), English↔French (en↔fr), English↔Chinese (en↔zh), English↔Russian (en↔ru) and Spanish→English (es→en). Our primary systems are built on deep Transformer with a large filter size. We also utilize R-Drop, data diversification, forward translation, back translation, data selection, finetuning and ensemble to improve the system performance. According to the official evaluation results in OCELoT[1] or CodaLab[2], our unconstrained systems in en→de, de→en, en→fr, fr→en, en→zh and es→en (clinical terminology sub-track) get the highest BLEU scores among all submissions for the WMT22 biomedical translation task.

1 Introduction

Machine translation (MT) refers to the automatic translation of text from one language to another, and the biomedical translation task aims to evaluate the performance of MT systems in the biomedical domain. In this year's biomedical translation task, our team (HW-TSC) participates in five language pairs, including en↔de, en↔fr, en↔zh, en↔ru and es→en (clinical terminology sub-track).

Since the size of in-domain (ID) data is limited, we first use a large amount of out-of-domain (OOD) data to train our baseline neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) system, which is a deep transformer model (Dou et al., 2018; Li et al., 2019) leveraging R-Drop (Wu et al., 2021) training strategy. We then use the collected ID data (except the data from medical database) to further train the NMT model for domain transfer. To better use the limited ID training data, we employ data selection to extract ID data from OOD data, in addition to basic data augmentation strategies including data diversity, forward translation and back translation. Finally, we use finetuning (Dakwale and Monz, 2017) and model ensemble (Wang et al., 2020b) to further improve model performance in the biomedical domain.

This paper is structured as follows: we describe data size and data pre-processing methods in section 2; the model structure and training methods in section 3; final results in section 4; and conclusion in section 5.

2 Dataset

2.1 Data volume

The data size for each language pair for the WMT22 biomedical translation task is shown in Table 1. The OOD bilingual data, used to train our baseline model, comes from the WMT general MT task and our internal corpus; while the ID bilingual and monolingual data, used for transferring the domain (Yang et al., 2021), come from Biomedical Translation, UFAL Medical Corpus and our internal corpus. As there is no ID monolingual data, we use the OOD monolingual instead.

2.2 Data Pre-processing

The data pre-processing process is as follows:

- Remove duplicate sentences (Khayrallah and Koehn, 2018; Ott et al., 2018).
- Remove sentences with mismatched parentheses and quotation marks.
- Filter out sentences of which punctuation percentage exceeds 0.4.
- Filter out sentences with the character-to-word ratio greater than 12 or less than 1.5.

[1] https://ocelot-wmt22.mteval.org
[2] https://codalab.lisn.upsaclay.fr/competitions/6696#results

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 936 - 942
December 7-8, 2022 ©2022 Association for Computational Linguistics

	bilingual					monolingual				
	en↔de	**en↔fr**	**en↔zh**	**en↔ru**	**es→en**	**en**	**de**	**fr**	**zh**	**ru**
OOD	200M	600M	200M	200M	200M	-	10M	-	-	40M
ID	2.75M	6.05M	10.87M	0.24M	8.1M	46M	-	2M	92M	-

Table 1: The data size for each language pair in the WMT22 Biomedical Translation Task

- Filter out sentences with more than 150 words.
- Apply langid (Joulin et al., 2017, 2016) to filter sentences in other languages.
- Use fast-align (Dyer et al., 2013) to filter out sentence pairs that are poorly aligned.

It should be noted that for en↔de, en↔fr, en↔ru and es→en translation task, we adopt joint SentencePiece model (SPM) (Kudo and Richardson, 2018; Kudo, 2018) for word segmentation, with a vocabulary of 32k. As for en↔zh translation task, we use Jieba tokenizer[3] to pre-segment Chinese sentences, and Moses tokenizer (Koehn et al., 2007) to pre-segment English sentences. Then we use joint Byte Pair Encoding (BPE) (Sennrich et al., 2016) to perform subword segmentation on Chinese and English sentences. The vocabulary size of BPE is also set to 32k.

3 System Overview

3.1 Model

Transformer (Vaswani et al., 2017), as the current mainstream architecture for NMT, adopts a fully self-attention mechanism, which can realize algorithm parallelism, speed up model training, and improve model performance. Deep Transformer, as an improvement of Transformer, increases the number of encoder layers and uses pre-layer-normalization to further improve model performance. Therefore, for all four language pairs, we adopt the Deep Transformer (Wei et al., 2021) model architecture: Based on the Transformer-big model architecture, our Deep Transformer model features pre-layer-normalization, 25-layer encoder, 6-layer decoder, 16-head self-attention, 1024-dimension word embedding and 4096-dimension hidden state.

3.2 R-Drop

Dropout (Srivastava et al., 2014) is a powerful and widely used technique for regularizing deep neural networks. Though it can help improve training effectiveness, the randomness introduced by dropouts may lead to inconsistencies between training and inference. R-Drop (Wu et al., 2021) forces the output distributions of different sub-models generated by dropout be consistent with each other. Therefore, we use R-Drop to augment the baseline model for each task and reduce inconsistencies between training and inference.

3.3 Data Diversification

Data diversification (Nguyen et al., 2020) is a simple and effective strategy to improve the performance of NMT. It uses predictions from multiple forward and backward models, and combines the results with the original data to train the final NMT model. The method does not require additional monolingual data and is applicable to all NMT models. It is more efficient than knowledge distillation (Wang et al., 2021) and dual learning (He et al., 2016). In our en↔de, en↔fr, en↔zh and en↔ru translation tasks, we use only a forward model and a backward model to generate synthetic data, and then mix the synthetic data with the bilingual data for NMT model training.

3.4 Forward Translation

Forward translation (Wu et al., 2019), also known as self-training (Imamura and Sumita, 2018), refers to using a forward NMT model to translate source-side monolingual data to generate synthetic bilingual data, which is then used to expand the training data size. Forward translation usually relies on beam search (Freitag and Al-Onaizan, 2017) decoding to generate synthetic data. Therefore, we adopt the forward translation method based on beam search decoding.

3.5 Back Translation

Back translation (Sennrich et al., 2015; Edunov et al., 2018) refers to translating the target monolingual data back to the source language, and then using the synthetic data to increase the training data size. This method has been proven effective in improving the NMT model performance. There are many back translation methods, among which sampling (Graça et al., 2019), noise (Edunov et al.,

[3] https://github.com/fxsjy/jieba

2018) or tagged (Caswell et al.) back-translation methods work better. In the scenario where forward translation and back translation are used in combination (Wu et al., 2019), the improvement effect brought by sampling back translation is more significant. In our translation task, we adopt sampling back translation method.

3.6 Data Selection

Data selection (van der Wees et al., 2017) is a data augmentation method that we use to select ID bilingual data from OOD bilingual data. Inspired by the domain feature calculation in curriculum learning (Wang et al., 2020a), we use an ID NMT model and an OOD NMT model to calculate the decoding probability of OOD bilingual data. The bilingual data of which ID decoding probability is higher than OOD decoding probability can be selected as additional ID data. The data selection process is also shown in Algorithm 1:

Algorithm 1: Data selection process

Input : ID NMT model θ_I, OOD NMT model θ_O and OOD bilingual data set D_O.

Output : ID bilingual data set D_I.

for *each sentence pair* $(x, y) \in D_O$ **do**
  // x is the source sentence, y is the target sentence.
  $score = \frac{\log P(y|x;\theta_I) - \log P(y|x;\theta_O)}{|y|}$
  if $score > 0$ **then**
    add (x, y) to D_I
  end
end

3.7 Finetuning

Finetuning (Dakwale and Monz, 2017) is a way to achieve domain transfer. In our translation task, we adopt a two-stage finetuning strategy. In the first stage, we use ID bilingual data to continue training the OOD NMT model, and then use the data augmentation strategy mentioned above to improve the model performance. In the second stage, we use the development set and synthetic data generated from the source-side text in the test set to finetune the ID model for more fine-grained domain transfer.

3.8 Ensemble

Ensemble (Wang et al., 2020b) is a widely used method to integrate different models for better performance. It is worth noting that when using ensemble, increasing the number of models does not always lead to better performance, and sometimes even causes performance deterioration. Therefore, for each track, we train four models on the same data, and go through all combinations of models to choose the one that performs best on the development set. This is also the model selection strategy (Yang et al., 2021) we use in the WMT21 biomedical translation task.

4 Experimental Result

During the training phase, we use Pytorch-based Fairseq[4] (Ott et al., 2019) open-source framework, and use deep Transformer model architecture as our benchmark system. Each model is trained using 8 GPUs with a batch size of 2048. The update frequency is 4 and the learning rate is 5e-4, the label smoothing rate (Szegedy et al., 2016) is 0.1, the warm-up steps is 4000, and the dropout is 0.3. Adam optimizer (Kingma and Ba, 2015) with $\beta 1$=0.9 and $\beta 2$=0.98 is also used. Furthermore, we use reg_label_smoothed_cross_entropy as the loss function and set reg-alpha to 5 when applying R-Drop (Wu et al., 2021) training strategy. In the evaluation phase, we use Marian[5] (Junczys-Dowmunt et al., 2018) for decoding and then calculate the sacrebleu[6] (Post, 2018) on the WMT21 OK-aligned biomedical test set to measure the performance of each model.

4.1 en↔de

For en↔de track, Table 2 shows the results of using the methods mentioned above to improve the model performance. The results show that continuing training with ID bilingual data on the basis of an OOD baseline improves en→de translation performance by 1.6 BLEU, but has little effect on the de→en track, with an increase of only 0.1 BLEU. Data selection significantly improves en↔de translation performance by 0.9-1.2 BLEU. In addition, other training strategies also bring small performance improvements.

4.2 en↔fr

Table 3 shows the results of en↔fr model. The results show that data diversity brings the greatest improvement to translation of both directions

[4] https://github.com/facebookresearch/fairseq
[5] https://github.com/marian-nmt/marian
[6] https://github.com/mjpost/sacrebleu

System	en→de	de→en
OOD R-Drop baseline	27.3	39.7
+ ID bilingual data continue training	28.9	39.8
+ data diversification	29.0	40.1
+ forward translation & back translation	29.4	41.3
+ data selection	30.3	42.5
+ dev set & synthetic test set finetuning	30.8	42.9
+ ensemble	**31.0**	**43.2**

Table 2: BLEU scores of en↔de on the WMT21 OK-aligned biomedical test set.

System	en→fr	fr→en
OOD R-Drop baseline	44.8	46.1
+ ID bilingual data continue training	45.3	46.3
+ data diversification	46.0	47.6
+ forward translation & back translation	46.3	47.7
+ data selection	-	47.8
+ dev set & synthetic test set finetuning	46.8	48.4
+ ensemble	**46.9**	**48.6**

Table 3: BLEU scores of en↔fr on the WMT21 OK-aligned biomedical test set.

System	en→zh	zh→en
OOD R-Drop baseline	38.5	32.1
+ ID bilingual data continue training	41.4	35.0
+ data diversification	42.5	36.4
+ forward translation & back translation	42.7	37.3
+ data selection	42.8	-
+ dev set & synthetic test set finetuning	43.0	38.7
+ ensemble	**43.1**	**39.3**

Table 4: BLEU scores of en↔zh on the WMT21 OK-aligned biomedical test set.

System	en→ru	ru→en
OOD R-Drop baseline	35.4	46.8
+ ID bilingual data continue training	41.0	48.9
+ data diversification	41.7	50.3
+ forward translation & back translation	42.3	50.4
+ data selection	-	-
+ dev set & synthetic test set finetuning	42.4	50.9
+ ensemble	**42.5**	**51.1**

Table 5: BLEU scores of en↔ru on the WMT21 OK-aligned biomedical test set.

(0.7 BLEU and 1.3 BLEU respectively). However, data selection has little impact on fr→en translation, and even no impact on en→fr translation. We assume this is because not much ID bilingual data is selected from the OOD data.

4.3 en↔zh

For en↔zh track, continuing training with ID bilingual data on the basis of an ODD baseline, as well as data diversity, bring the greatest impact on the model performance, while data selection has the least impact. In addition, the methods such as forward translation & back translation, dev set & synthetic test set finetuning and ensemble have little improvement on en→zh translation, but have a great improvement on zh→en translation. The detailed results of en↔zh translation are shown in Table 4.

4.4 en↔ru

As shown in Table 5, for the en↔ru track, the results are similar to en↔zh translation task. Continuing training with ID bilingual data and data diversity have the greatest impact on model performance, while data selection does not lead to performance improvement. In addition, the performance improvements brought by other methods are also relatively limited.

4.5 es→en

We also participate in the es→en clinical terminology sub-track (ClinSpEn-CT) this year. The sample set contains 7,000 terms that are extracted from medical literature and clinical records, with a particular focus on diseases, symptoms, findings, etc. The translations are generated and revised by professional medical translators. We extract 1000 sentences from the sample set as the dev set.

The results are shown in Table 6. All chrF and BLEU scores are calculated on this dev set. Unlike other experiments above, for es→en clinical terminology sub-task, we abandon forward translation method for the sake of maintaining terminology accuracy. Instead, we perform two rounds of back translation using monolingual English ID data. Finally, we finetune the model with 6000 bilingual terms, which results in a significant improvement on the dev set.

4.6 Results In OCELoT Or CodaLab

The BLEU scores of our submissions to the WMT22 Biomedical Translation Task on OCELoT and CodaLab (ClinSpEn-CT) are shown in Table

System	chrF	BLEU
OOD R-Drop baseline	0.76	49.5
+ ID bilingual data continue training	0.77	50.7
+ back translation	0.79	53.4
+ 2nd round back translation	0.79	54.1
+ 6000 bilingual terms finetuning	0.82	56.7
+ ensemble	**0.82**	**57.2**

Table 6: chrF (Popović, 2015) and BLEU scores of es→en on the WMT22 biomedical ClinSpEn-CT 1000 sample set.

	en→de	de→en	en→fr	fr→en	en→zh	zh→en	en→ru	ru→en	es→en
our submission system	**38.7**	**45.6**	**38.8**	**48.6**	**49.9**	43.0	43.3	50.3	**41.57**

Table 7: BLEU scores of our submission systems on WMT22 Biomedical Translation Task on OCELoT or CodaLab, where the highest BLEU scores among all submissions are bolded.

7, where our submitted systems achieve the highest BLEU scores in six language directions of the WMT22 biomedical translation task. In conclusion, from the results on the WMT21 OK-aligned biomedical test set, continuing training with ID bilingual data, data diversity, forward translation and back translation have great impacts on the NMT model performance. When the OOD bilingual data contains a certain amount of ID bilingual, the data selection method can also achieve a good boost effect. In addition, dev set & synthetic test set finetuning and ensemble can lead to further performance gains.

5 Conclusion

This paper presents our translation system for the WMT22 en↔de, en↔fr, en↔zh, en↔ru and es →en biomedical translation task. During the experiment, we use R-Drop and ID bilingual data finetuning methods to build our ID translation system, and then use data diversity, forward translation, back translation and data selection methods to expand the size of training data for training a better system. We also adopt finetuning and ensemble to further improve the system performance. According to the official evaluation results in OCELoT or CodaLab, our submitted systems achieve the highest BLEU scores in six language directions of the WMT22 biomedical translation task.

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*.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. *WMT 2019*, page 53.

P Dakwale and C Monz. 2017. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data.

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang. 2018. Exploiting deep representations for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4253–4262.

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 and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 820–828.

Kenji Imamura and Eiichiro Sumita. 2018. Nict self-training approach to neural machine translation at nmt-2018. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 110–115.

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018.

Marian: Fast neural machine translation in c++. In *ACL (4)*.

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 Lei Ba. 2015. Adam: A method for stochastic optimization.

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.

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.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

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.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

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

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410.

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.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6456–6466.

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.

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.

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.

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.