Huawei's Submissions to the WMT20 Biomedical Translation Task

Wei Peng[1], Jianfeng Liu[1], Minghan Wang[2], Liangyou Li[3], Xupeng Meng[1], Hao Yang[2], Qun Liu[3]
[1]Artificial Intelligence Application Research Center, Huawei Technologies
{peng.wei1,liujianfeng,mengxupeng}@huawei.com
[2]Huawei Translation Service Center, Huawei Technologies
{wangminghan,yanghao30}@huawei.com
[3]Noah's Ark Lab, Huawei Technologies
{liliangyou,qun.liu}@huawei.com

Abstract

This paper describes Huawei's submissions to the WMT20 biomedical translation shared task. Apart from experimenting with finetuning on domain-specific bitexts, we explore effects of in-domain dictionaries on enhancing cross-domain neural machine translation performance. We utilize a transfer learning strategy through pre-trained machine translation models and extensive scope of engineering endeavors. Four of our ten submissions achieve state-of-the-art performance according to the official automatic evaluation results, namely translation directions on English⇔French, English→German and English→Italian.

1 Introduction

Neural machine translation (NMT) models built upon the Transformer architecture (Vaswani et al., 2017) start to dominate the leader board of WMT biomedical shared tasks in recent years (Bawden et al., 2019). In-domain data (parallel and monolingual corpora) have been widely used in finetuning general domain NMT models. Despite ongoing improvements on the translation quality observed from recent biomedical shared tasks, domain adaptation remains an open problem. The in-domain data is hard to obtain and, as a consequence, greatly limits the cross-domain translation capability an NMT model can offer. Domain terminologies, on the other hand, are regarded as critical resources to improve the quality of machine translation by mitigating effects of scarce in-domain bitexts (Bawden et al., 2019). However, few research works leverage domain-specific terminologies (or dictionaries) in training cross-domain NMT systems.

In this paper, we present the system architecture and research approaches underpinning Huawei's submissions to the WMT20 biomedical translation task. We implement two NMT systems to maximize the performances of the shared task. The system I is an in-house NMT system built upon the transformer-big architecture (Vaswani et al., 2017) and trained using general domain data. We explore means to enhance cross-domain coverage of an NMT model by finetuning the NMT model with in-domain bitexts. We also investigate the effects of domain dictionaries in this domain adaptation process. Reusing pre-trained models has been regarded as an efficient way of transfer learning. Pre-trained NMT models (Ng et al., 2019) are adopted in the system II to this end.

All NMT systems are evaluated against the test set released in the WMT19 biomedical shared task. We submitted translated results for a total of ten language directions between English (EN) and other five languages including French (FR), German (DE), Italian (IT), Russian (RU) and Chinese (ZH). Four of the submissions achieve the best BLEU scores according to the official automatic evaluation results. Substantial increases in BLEU scores are recorded in translation directions of DE→EN (+3.9 BLEU), ZH→EN (+3.5 BLEU), and EN→DE (+2.8 BLEU) compared to our submissions last year (Peng et al., 2019). The improvements on EN⇔DE can be ascribed to strong pre-trained NMT baseline models and a series of optimization techniques, for example, in-domain data augmentation and a reranking method with strong language models. High-quality in-domain data and large-scale back-translation contribute to the improvements of the ZH→EN model.

2 The Data

Table 1 captures the number of sentences pairs used in this shared task. The system I is trained using in-house general domain data (OOD) and finetuned on the in-domain data (IND) provided by

Proceedings of the 5th Conference on Machine Translation (WMT), pages 857–861
Online, November 19–20, 2020.

Directions		Train				Dev.	Test	Vocab.
		OOD	IND	IND-Dict.	IND-Aug.			
System I	EN→FR	146M	4M	59K	-	4K	440	40K
	FR→EN	186M	4M	59K	-	4K	417	40K
	EN→IT	83M	219K	-	-	3.8K	400	40K
	IT→EN	150M	219K	-	-	3.8K	400	40K
	EN→ZH	164M	-	59K	-	5K	448	50K
	ZH→EN	200M	-	59K	55M	5K	115	50K
System II	EN→DE	-	40K	-	56K	435	-	42K
	DE→EN	-	40K	-	56K	373	-	42K
	EN→RU	-	54K	-	-	300	-	32K(EN)/31K(RU)
	RU→EN	-	54K	-	-	300	-	32K(EN)/31K(RU)

Table 1: Data used for training and finetuning systems I and II. Note that "IND-Dict." refers to the in-domain dictionary. "IND-Aug." is the augmented data derived from processing IND data. For the system I, "IND-Aug." is created from back-translating monolingual data. For the system II, "IND-Aug." is the pre-processed IND data in combination with the data selected from some OOD data based on the similarity to the Medline data. M is for "million," and K stands for "thousand".

WMT20.[1] The in-domain data consist of bitexts from EMEA (Tiedemann, 2012), UFAL,[2] Pubmed, and Medline.[3] The data is processed by methods in the next section. The test data for the system I are from the WMT19 shared task.

The data used for finetuning the system II are different from those for the system I. The system II only focuses on Medline as we discovered it is the most effective IND data for this shared task. The development (dev.) set for the system II is the OK-aligned test data from the WMT19 biomedical shared task.

A batch of monolingual Medline data in English dated before July 2018 has been extracted to provide a basis for data augmentation and noisy channel model reranking (Ng et al., 2019). It produces the augmented IND data for the ZH→EN translation direction via back-translation ("IND-Aug." in Table 1). Due to time and resource constraints, we could not fully explore this monolingual Medline data in other translation directions.

3 The Approaches

The proposed systems are finetuned and enhanced using the following methods. All models are trained on Tesla V100 GPUs. Systems I and II use batch sizes of 6,144 and 8,000 tokens respectively in the finetuning process.

3.1 In-domain Dictionary

Bilingual dictionaries have been studied in the machine translation community for various purposes. The lexicons are used to enhance the translation quality for rare and unknown words in the parallel corpus (Zhang and Zong, 2016). Research works in domain adaptation for NMT showed that incorporating domain-specific dictionaries is a viable solution (Hu et al., 2019; Thompson et al., 2019; Peng et al., 2020). Inspired by these studies, we apply domain-specific dictionaries derived from SNOMED-CT, [4] which is a collection of multilingual clinical terminology, to finetune general domain NMT models to boost cross-domain coverage. The dictionaries are treated as bitexts attached to the end of training data.

3.2 Reranking

Apart from adopting a data-driven approach mentioned above, we also apply a transfer learning approach by reusing the publicly available pre-trained NMT models provided at fairseq (Ott et al., 2019). [5] After finetuning the selected pre-trained NMT models on the in-domain data, we apply a noisy channel model reranking method (Ng et al., 2019). The weights λ in Equation 1 are learned with a

[1] http://www.statmt.org/wmt20/biomedical-translation-task.html

[2] https://ufal.mff.cuni.cz/ufal_medical_corpus

[3] https://github.com/biomedical-translation-corpora/corpora

[4] https://www.nlm.nih.gov/healthit/snomedct/index.html

[5] https://github.com/pytorch/fairseq

System I	EN→FR	FR→EN	EN→IT	IT→EN	EN→ZH	ZH→EN
baseline	38.98	38.31	30.85	35.73	36.22	34.37
+ ft BS, IND	-	-	31.04	35.93	-	-
+ ft IND, IND-Dict.	41.66	38.44	-	-	-	-
+ ft BS,IND-Dict.,IND-Aug.	-	-	-	-	35.90	35.66
WMT19 Submission	42.41	38.24	-	-	37.09	32.16
WMT20 Submission	**43.51**	**44.45**	**42.57**	**49.74**	**45.46**	**35.28**
WMT20 Best Official	**43.51**	**44.45**	**42.57**	**50.11**	**46.86**	**35.28**

Table 2: BLEU scores of the system I on all related submissions. The baseline models are finetuned (ft) in various configurations, including mixed finetuning on in-house OOD data (aka BS), IND bitexts, "IND-Dict." and the augmented IND data ("IND-Aug."). Note that the WMT20 best official score for ZH→EN excludes those results currently under investigation.

random search for the best performing candidate on the validation data.

$$\lambda_1 logP(y|x) + \lambda_2 logP(y) + \lambda_3 logP(x|y) \quad (1)$$

Due to time constraints, we did not implement the reranking approach on the system I.

3.3 Data Processing

A data processing pipeline is applied to enhance the quality of training data:

- Data cleaning is implemented to filter out noisy data. An important step is to handle misalignment in the parallel corpus. An alignment model trained by fast-align (Dyer et al., 2013) [6] is applied to this end (Lu et al., 2018). In addition, we remove bitexts with a source and target sentence length ratio exceeding a certain threshold (i.e., 2.5). A language detection tool [7] is used to filter out bitexts with abnormal language patterns, i.e., sentences with undesirable *langid*. Other noisy data, such as those with HTML tags and extra spaces, are removed.

- Scripts from Moses (Koehn et al., 2007) are used to perform punctuation normalization and tokenization. SentencePiece (Kudo and Richardson, 2018) segments words into subwords.

- We extract "in-domain" data which are close to Medline from general domain data by using TFIDF-based similarities. Similar data augmentation approaches can be identified in Wang et al. (2017) and Peng et al. (2020).

- Post-processing is performed after decoding to detokenize subwords and remove undesirable spaces between special characters and numbers, i.e., converting "23 - 25" into "23-25".

[6] https://github.com/clab/fast_align

[7] https://github.com/aboSamoor/polyglot

4 Experimental Results

The systems are trained with OOD data and finetuned using IND data to produce the submitted results. We benchmarked the submissions using WMT19 test data. The BLEU scores are calculated using the MTEVAL script from Moses (Koehn et al., 2007). Results are shown in Table 2 and Table 4. The final two rows demonstrate the scores of our submissions on this year's test sets and the best official records released by the organizers.

4.1 English ⇔ French

The system I is our in-house system equipped with an extensive data processing pipeline to handle noisy data, i.e., the application of sentence alignment and language detection tools. Our EN→FR and FR→EN submissions achieve the best official results in the WMT20 shared task. IND bitexts and "IND-Dict." have contributed to up to 2.7 BLEU in enhancing the baseline performance. We presume the improvement is due to the enhanced domain coverage the IND data brought forth. Note that even with much larger OOD bitexts than last year, the system produces similar benchmark scores. It appears an over-representation of OOD data is not helpful in cross-domain NMT. An analysis of domain coverage is performed to investigate the effect of IND information on cross-domain translation. We count the number of unique terms (1-2 grams)

Data	EN→FR		FR→EN	
	Unigrams	**Bigrams**	**Unigrams**	**Bigrams**
OOD	2,763	5,752	2,989	6,317
OOD + IND + IND-Dict.	2,773 (+10)	5,827 (+75)	2,997 (+8)	6,372 (+55)

Table 3: Domain coverage analysis for data used to train English⇔French.

System II	**EN→DE**	**DE→EN**	**EN→RU**	**RU→EN**
baseline	34.12	37.39	-	-
+ ft All Medline	35.58 (+1.46)	39.06 (+1.67)	-	-
+ ft Pre-proc. Medline	36.90 (+1.32)	40.98 (+1.92)	27.30	33.38
+ ft IND-Aug.	37.13 (+0.23)	41.79 (+0.81)	-	-
+ reranking	38.17 (+1.04)	42.74 (+0.95)	-	-
WMT19 Submission	35.39	38.84	-	-
WMT20 Submission	**36.89**	**41.46**	**34.64**	**43.03**
WMT20 Best Official	**36.89**	**41.65**	**39.36**	**43.31**

Table 4: BLEU scores of system II on English ⇔ German. "Pre-proc." stands for "pre-processed." Note that "IND-Aug." contains the pre-processed Medline data and the data derived from OOD via TFIDF selection. Numbers in the brackets depict the incremental increase from the baseline models.

at the intersection of a data source (i.e., the OOD training data) and the test data. Table 3 indicates that the increase of BLEU may be associated with a level of domain coverage enhancement. An increasing number of distinctive IND terms is recorded.

4.2 English ⇔ German

We perform ablation tests on pre-trained NMT models (the system II) in English ⇔ German under various conditions. As shown in Table 4, an EN→DE model finetuned on a preprocessed version of Medline outperforms that trained on the full version of Medline by 1.32 BLEU, indicating the effectiveness of the data preprocessing method. The EN→DE model finetuned on the "IND-Aug." data adds 0.23 to the BLEU score. The performance of the model can be boosted by 1.04 BLEU using the reranking method. Both EN→DE and DE→EN models outperform our last year's submissions significantly by 2.78 and 3.90 BLEU, respectively.

4.3 Other Translation Directions

The submissions for other translation directions are illustrated in Table 2 and Table 3. Note that we did not perform the experiments on the same level as those for English⇔German due to time constraints. It is observed that finetuning on IND data has contributed to improving the performance of baseline models in EN→IT, IT→EN, and ZH→EN directions. The result for EN→ZH is inconclusive, most likely due to potential issues during training.

5 Conclusion

This paper depicts Huawei's submissions to the WMT20 biomedical shared task. For all ten translation directions, we have explored the effects of using IND bitexts and dictionaries on enhancing the performances of cross-domain NMT. We have demonstrated the benefits of the transfer learning strategy of reusing pre-trained NMT models. Four of our ten submissions achieve the best records according to the released WMT20 official results.

Acknowledgments

We would like to show our gratitude to colleagues from Huawei Noah's Ark Lab, AARC, Huawei Translation Service Center, and HTRDC AIE for their support during this work.

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared*

Task Papers, Day 2), pages 29–53, Florence, Italy. Association for Computational Linguistics.

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, Atlanta, Georgia. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jun Lu, Xiaoyu Lv, Yangbin Shi, and Boxing Chen. 2018. Alibaba submission to the WMT18 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 917–922, Belgium, Brussels. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

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 NAACL-HLT 2019: Demonstrations*.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation.

Wei Peng, Jianfeng Liu, Liangyou Li, and Qun Liu. 2019. Huawei's NMT systems for the WMT 2019 biomedical translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 164–168, Florence, Italy. Association for Computational Linguistics.

Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. HABLex: Human annotated bilingual lexicons for experiments in 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 1382–1387, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Bridging neural machine translation and bilingual dictionaries. *CoRR*, abs/1610.07272.