FJWU participation for the WMT21 Biomedical Translation Task

Sumbal Naz[1], **Sadaf Abdul Rauf**[1,2] **and Sami ul Haque**[3]
[1] Fatima Jinnah Women University, Pakistan
[2] Univ. Paris-Saclay, LISN-CNRS, France
[3] National University of Science and Technology, Pakistan
{sadaf.abdulrauf,sumbalnaz01}@gmail.com

Abstract

In this paper we present the FJWU's system submitted to the biomedical shared task at WMT21. We prepared state-of-the-art multilingual neural machine translation systems for three languages (i.e. German, Spanish and French) with English as target language. Our NMT systems based on Transformer architecture, were trained on combination of in-domain and out-domain parallel corpora developed using Information Retrieval (IR) and domain adaptation techniques.

1 Introduction

Due to vast availability of multilingual information, Neural Machine Translation (NMT) systems have achieved remarkable growth over Statistical Machine Translation (SMT) systems. Although the amount of training resources has significantly increased in the past few years but availability of large in-domain parallel data is still a challenging task. Performance of NMT system may quickly degrade as soon as the application domain deviates from training domain. Domain adaptation (Koehn and Schroeder, 2007) is a promising active research topic to enhance the translation quality when faced with data scarcity issues. In domain adaptation, initially large amount of parallel out-of-domain corpora is utilized for training NMT models and then fine-tuning is performed on small in-domain data for adapting to novel domains (Freitag and Al-Onaizan, 2016; Luong and Manning, 2015). Fine-tuning does not require building system from scratch, instead it is fast and efficient method of integrating in-domain data. An NMT model already trained on general domain data is further fine-tuned on in-domain data with less time and effort (Chu et al., 2017; Hira et al., 2019). Training MT systems on back-translated data is a proven domain adaptation method (Abdul Rauf et al., 2020; Sennrich et al., 2015), where synthetic parallel data is combined with original data to generate large in-domain training corpus. In addition, information retrieval (IR) technique to extract relevant sentences from out-of-domain corpus has shown promising results to overcome data scarcity (Naz et al., 2020).

NMT system incorporating multiple languages into single model is known as multilingual NMT (MNMT) (Dabre et al., 2020). Multilingual NMT systems are gaining popularity due to effective use of available resources and boosting translation quality with Translation Knowledge Transfer (Pan and Yang, 2009).

In this paper, we present study on adapting MNMT systems (Multiway many-to-one) for translating English (EN) language from French (FR), German (DE) and Spanish (ES) using fairseq (Ott et al., 2019) implementation of Transformer model. Our main focus is to investigate the effect on EN translation in Biomedical domain using multilingual NMT systems. We have also explored the domain adaptation for fine-tuning of bilingual NMT models into multilingual NMT models using out-domain and in-domain corpora. Furthermore, we show the effectiveness of utilizing in-domain data generated through IR techniques (Naz et al., 2020) by training a NMT system on combined parallel in-domain data. We also compare in-domain multilingual and bilingual models.

The remainder of this paper is organized as follows. Section 2 introduces the literature review followed by corpus processing in Section 3. Section 4 presents experiments and results. In Section 5, we conclude the findings of our work.

2 Literature Review

MNMT models tend to acquire knowledge from more than one language which helps in generalization and in building systems for low resource languages. MNMT models may help in miti-

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 857–862
November 10–11, 2021.

Corpus	DE/EN	ES/EN	FR/EN
In-domain training data			
UFAL	2.6 M	631 K	2.6 M
Scielo Health	-	124 K	9 K
Scielo Biological	-	581 K	-
EDP	-	-	3 K
Medline Titles	-	285 K	612 K
Medline Abstracts	18 K	66 K	46 K
EMEA	1.10 M	1.09 M	1.09 M
In-domain IR training data			
News Commentary-IR2	-	-	65 K
WikiPedia-IR2	-	-	84 K
Out-domain training data			
UFAL	30.9 M	74.8 M	73.3 M
UFAL Dictionary	733 K	544 K	744 K
Scielo	-	433 K	-
UN	-	21.9 M	25.8 K
Development data			
Medline18	321	239	311
Medline19	439	437	400
Test data			
Medline20	409	466	479

Table 1: Sentence Pairs Used for Training, Development and Testing of MNMT models (K stands for "Thousand" and M stands for "Million")

gating the problem for resource poor languages (Dabre et al., 2020), where limited training data is available. Tubay and Costa-jussÃ (2018) submitted their NMT systems for English translation with multi-source similar languages including Portuguese, French and Spanish showing improvement of 6 BLEU points over single source NMT system. Soares and Krallinger (2019) also built NMT systems using two of the Romance languages, Spanish and Portuguese for translating into English language. For domain adaptation in NMT, fine-tuning models on in-domain parallel text is a common and effective approach (Peng et al., 2020). We assume that, the same can be used for training multilingual (many-to-one) NMT models. Chu and Dabre (2019) focused on fine-tuning MNMT models for domain adaptation, they initially trained different MNMT models using single domain and then further fine-tune on multi-domain corpora with mixed (combination of out-domain and in-domain) corpora.

3 Corpus Pre-processing

This section describes parallel corpora used in training and evaluation of our models. Statistics of train, development and test data are presented in Table 1. Main sources of data were provided by WMT21 Biomedical Translation Task. Data sources include:

- Medline abstracts and titles in-domain corpora consists of scientific publications (Bawden et al., 2019). We used datasets available for DE/EN, ES/EN and FR/EN provided by WMT. These datasets are aligned through Bilingual Sentence Aligner[1] (Moore, 2002).
- EDP are the in-domain texts of scientific publications available for FR/EN language pair only (Neves et al., 2018).
- EMEA provides in-domain biomedical parallel corpus of documents related to medicinal products (Tiedemann, 2012). We used corpora provided for DE/EN, ES/EN and FR/EN language pairs.
- Scielo in-domain corpus provided by WMT comprises of abstracts and titles in biological and health sciences domain (Neves et al., 2016). We used datasets provided for FR/EN and ES/EN language pairs.

  We also used Scielo full text corpus of scientific articles as general domain data set provided for ES/EN language pair (Soares et al., 2018).
- UFAL Medical Corpus provides various in-domain medical texts and out-domain corpus sources including dictionaries (Jimeno Yepes et al., 2017). We included corpora provided for DE/EN, ES/EN and FR/EN language pairs.
- United Nations (UN) parallel corpus comprises of official records in general domain (Ziemski et al., 2016) . We used sources provided for ES/EN and FR/EN language pairs.

News Commentary[2] and WikiPedia[3] in-domain IR corpora are used. These corpora are extracted using data selection based on IR approach (Abdul-Rauf et al., 2016) by using Medline titles as queries

[1] https://www.microsoft.com/en-us/download/details.aspx?id=52608

[2] http://opus.nlpl.eu/News-Commentary-v14.php

[3] http://opus.nlpl.eu/Wikipedia-v1.0.php

		ID	Train Set	DE → EN	ES → EN	FR → EN
Multilingual	System I	M1	In-domain	26.96[3]	39.12[3]	30.23
		M2	M1⇒Medline	27.22[2]	39.40[1]	33.79[1]
		M3	M2⇒Indomain+IR-2	27.38[1]	39.36[2]	32.07[2]
Multilingual	System II	M4	Out-domain	20.97	29.75	24.86
		M5	M4⇒In-domain	26.40	38.87	30.94[3]
		M6	M5⇒Medline	26.40	38.74	**34.73**
Bilingual	System III	M7	In-domain	**27.40**	**41.60**	30.20

Table 2: Bilingual and Multilingual DE/ES/FR → EN Transformer models and their BLEU scores for Medline20 test-sets for three language directions (DE→EN, ES→EN and FR→EN. (M here stands for 'Model'). Superscripts denote the runs submitted.

for retrieving related biomedical domain sentences. From experiments conducted by (Naz et al., 2020), corpora with top-2 best sentences gave good results in training NMT models for biomedical domain.

Medline18 and 19 testsets are used as development set. We used Medline20 testset provided by WMT20 (Bawden et al., 2020) as initial test sets to determine quality of our translation models. Preprocessing of data include tokenization and learning joint Byte Pair Encoding (BPE) (Sennrich et al., 2016) using sentencepiece[4] with a vocabulary size of 32K over in-domain corpus and encoding all available corpora with learned BPE.

4 Experiments and Results

In this section we present details of experimentation along with training configurations.

4.1 Training and Parameters

We employed Fairseq toolkit to train MNMT systems for (German, Spanish, French) → English translation. We used Transformer architecture and followed similar configuration parameters for out systems as reported in original paper (Vaswani et al., 2017). Batch size of 4K words and Adam optimizer was used in all experiments. Training was done till convergence and stopped if no improvement was noted in BLEU scores on development sets for 2-3 consecutive checkpoints. Fine-tuned models were trained for 150K steps unless early stopping is employed based on bleu score convergence.

[4] https://github.com/google/sentencepiece

4.2 NMT Models

We have categorized our experiments into 3 classes based on the corpora and training technique used. I) Multilingual models trained using all in-domain corpus and fine-tuned on Medline and IR. II) Multilingual models trained on all out-domain corpus and fine-tuned on all in-domain and Medline corpus. III) Bilingual models trained on all in-domain corpus. Results of all experiments are depicted in Table 2. BLEU score for all models is calculated using Sacrebleu (Post, 2018) on Medline20 test-set for German-English, Spanish-English and French-English.

For System I:

- $M1$: this is trained on all in-domain parallel corpus with a total size of 3.71M (DE-EN), 2.77M (ES-EN), 1.78M (FR-EN) sentences. Best BLEU score of 39.12 on Medline20 test-set was achieved for ES→EN as it has high rate of Medline sentences (66K) as compared to FR→EN (46K) and DE→EN (18K).
- $M2$: this model derived from $M1$ by further tuning it on Medline corpus for domain adaptation which resulted in significant increase in BLEU score of +3.56 for FR→EN as compared to previous model ($M1$). An increase of +0.26 BLEU for DE→EN and +0.28 BLEU for ES→EN is achieved.
- $M3$: $M2$ was further fine-tuned on IR corpus for FR→EN language pair but we observe no significant improvements in term of BLEU score on out test-set for all combinations of languages. IR corpus was extracted

from News commentary and Wikipedia parallel corpora. Apparently these corpora are far in language jargon from the traditional Medline tests, so we see no apparent gain. It is pertinent to note that IR data was only available for FR-EN thus the in-domain training corpus was used for other language pairs.

For System II:

- $M4$: this model is trained on all out-domain parallel corpus with a total size of 3.82M (DE-EN), 10.6M (ES-EN) 8.33M (FR-EN) sentences. Highest BLEU score of 29.75 is achieved with ES→EN test set. As the model is mainly trained on out-domain corpora, the huge difference in score is visible as compared to previous models. When compared with models trained on in-domain we see significant loss in BLEU scores for our current model.

- $M5$: previous model is fine tuned on in-domain corpus that shows substantial improvements over baseline model. A gain of +5.43 points for DE→EN, +9.12 points for ES→EN and +6.08 points for FR→EN was achieved. This clearly indicates that fine-tuning is an effective method for improving quality of multilingual NMT.

- $M6$: $M5$ is further fine tuned on Medline corpus yielding an improvement of +3.79 points for FR→EN giving best score of **34.73** among all models . No significant improvement in DE→EN and ES→EN is observed.

For System III:

- $M7$: Represents the bilingual models trained on all in-domain corpus. Comparing with the multilingual models; ES→EN achieved the best score of `41.60` BLEU points in bilingual mode. Bilingual DE→EN results are comparable to the multilingual systems whereas for FR→EN multilingual systems majorly outperformed the bilingual systems. Interestingly, ES→EN had more medline corpus as compared to other two. The three language pairs that we work on are not similar and thus do not have too much to gain from each other. Introducing other romance languages in the systems might lead to better performance for French and Spanish. The factor of training corpus imbalance is also playing it's part, we intend to employ better sampling strategies for multilingual systems in future.

5 Conclusion

In this paper we have described our system submissions at WMT21 biomedical shared translation task under FJWU's submission. For our submission we trained multilingual NMT systems for German, Spanish and French languages with English as target language. We focused on utilizing in-domain and out-domain parallel corpora and domain adaptation techniques for training multilingual NMT systems. We showed that, domain adaptation using fine-tuning of multilingual NMT model can be a reasonable alternative to achieve good translation quality for novel domains.

Acknowledgments

This study is funded by the National Research Program for Universities (NRPU) by Higher Education Commission of Pakistan (5469/Punjab/NRPU/R&D/HEC/2016).

References

Sadaf Abdul Rauf, José Carlos Rosales Núñez, Minh Quang Pham, and François Yvon. 2020. Limsi @ wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 803–812, Online. Association for Computational Linguistics.

Sadaf Abdul-Rauf, Holger Schwenk, Patrik Lambert, and Mohammad Nawaz. 2016. Empirical use of information retrieval to build synthetic data for smt domain adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):745–754.

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.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova.

2020. Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *arXiv preprint arXiv:1906.07978*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Noor-e Hira, Sadaf Abdul Rauf, Kiran Kiani, Ammara Zafar, and Raheel Nawaz. 2019. Exploring transfer learning and domain data selection for the biomedical translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 156–163, Florence, Italy. Association for Computational Linguistics.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. Association for Computational Linguistics.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation*, pages 224–227.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT, Da Nang, Vietnam.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proc. AMTA'02*, Lecture Notes in Computer Science 2499, pages 135–144, Tiburon, CA, USA. Springer Verlag.

Sumbal Naz, Sadaf Abdul Rauf, Noor-e Hira, and Sami Ul Haq. 2020. Fjwu participation for the wmt20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 849–856, Online. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Wei Peng, Jianfeng Liu, Minghan Wang, Liangyou Li, Xupeng Meng, Hao Yang, and Qun Liu. 2020. Huawei's submissions to the wmt20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 857–861, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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, Berlin, Germany.

Felipe Soares and Martin Krallinger. 2019. Bsc participation in the wmt translation of biomedical abstracts. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 177–180, Florence, Italy. Association for Computational Linguistics.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A large parallel corpus of full-text scientific

articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

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

Brian Tubay and Marta R. Costa-jussÃ. 2018. Neural machine translation with the transformer and multi-source romance languages for the biomedical wmt 2018 task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 678–681, Belgium, Brussels. Association for Computational Linguistics.

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.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).