FJWU participation for the WMT20 Biomedical Translation Task

Sumbal Naz[1], **Sadaf Abdul Rauf**[1,2], **Noor e Hira**[1], **Syeda Abida**[1] **and Sami Ul Haq**[3]
[1] Fatima Jinnah Women University, Pakistan
[2] Univ. Paris-Saclay, CNRS, LIMSI France
[3] National University of Sciences and Technology, Pakistan
{sadaf.abdulrauf,sumbalnaz01,noorehira94,sami.haq99}@gmail.com

Abstract

This paper reports system descriptions for FJWU-NRPU team for participation in the WMT20 Biomedical shared translation task. We focused our submission on exploring the effects of adding in-domain corpora extracted from various out-of-domain sources. Systems were built for French to English using in-domain corpora through fine tuning and selective data training. We further explored BERT based models specifically with focus on effect of domain adaptive subword units.

1 Introduction

In this paper, we present Neural Machine Translation (NMT) systems developed by Fatima Jinnah Women University for participation in WMT20, Biomedical shared Translation task. The systems are developed for translating English/French (EN/FR) in both directions for biomedical domain using fairseq (Ott et al., 2019) and BERT (Devlin et al., 2018). To tackle in-domain corpus shortage challenge, selective data training and fine tuning are explored. We focused our submission on investigating the effects of adding in-domain corpora extracted from out-of-domain sources of various domains, objective was to study the effect of domain non-relatedness in schemes involving data selection through information retrieval or any sentence selection method. We further explored BERT based models specifically with focus on effect of domain adaptive subword units.

Neural Machine Translation systems have shown substantial growth with the ongoing introduction of new tool kits and training techniques to support developers in training models (Bahdanau et al., 2014; Wu et al., 2016). But the availability and cleaning of domain related corpora to achieve terminology advantage and fluency is still a challenge for many researchers as the accessible corpora is relatively small in size and comparatively noisy. In order to improve in-domain NMT systems, out-of-domain data is used and the most common method is to fine tune pre trained NMT models on in-domain data and selective data training (Hira et al., 2019).

Our last years submission presented promising results using selective data training incorporating data retrieved from News Commentary corpus by building two layered RNN systems. We extend our framework to study the quality of retrieved sentences from 3 more parallel corpora. We did not restrict to parallel data for mining biomedical sentences, rather this year we included monolingual data in our framework and studied the effect of using Back Translations (BT) in our framework. For building NMT models we explored subword units and report the results on using pre-trained BERT fused embedding.

2 Data Selection Architecture

Improving translation quality is a challenging task especially for domains where enough in-domain parallel corpus is not available to train a good translation system. To overcome data scarcity problem, several data selection techniques have been proposed over the years including information retrieval (IR) (Rauf and Schwenk, 2011), edit distances (Wang et al., 2013), cross entropy measures (Axelrod et al., 2011) and several others. We used the approach of relative query sentences using information retrieval to retrieve matching sentences from general domain corpora.

French-English is not a resource scarce language pair and has numerous parallel corpora available for various domains. There exist sizable corpus for the Biomedical domain to train the initial systems, but the great difference of terminologies and language jargon in various sub domains makes it challenging as the results of previous years bio med-

Proceedings of the 5th Conference on Machine Translation (WMT), pages 849–856
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

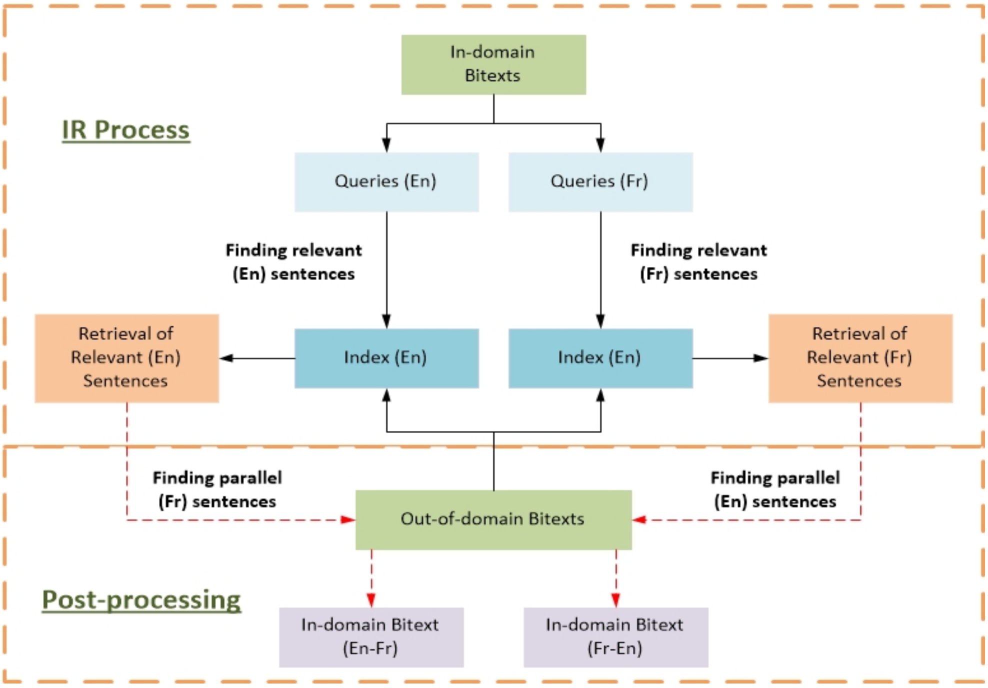
IR Process
In-domain Bitexts
Queries (En)
Queries (Fr)
Finding relevant (En) sentences
Finding relevant (Fr) sentences
Retrieval of Relevant (En) Sentences
Index (En)
Index (En)
Retrieval of Relevant (Fr) Sentences
Finding parallel (Fr) sentences
Finding parallel (En) sentences
Out-of-domain Bitexts
Post-processing
In-domain Bitext (En-Fr)
In-domain Bitext (Fr-En)

Figure 1: Data selection Architecture.

ical tasks indicate. Parallel corpora extracted from "other" easily available corpora, like comparable corpora and monolingual corpora do help improve MT performance (Abdul-Rauf and Schwenk, 2009; Abdul-Rauf et al., 2016). But, what is the effect of the domain of the corpus used to find the related sentences, is the question we focus on in our data selection design.

Our aim is to study the improvements achieved by using the sentences from different genre/domain of corpora. However, to be able to extract sizeable amount of biomedical sentences, the corpora should not be very unrelated, for example, the Europarl corpus (Koehn, 2005) which is composed of Parliament proceedings would not be a good choice[1]. Our intent was to do a comparative study of quality of extracted sentences from varied but yet not too far off domain corpora. Thus, for mining related sentences from general domain corpora we used Books[2], News Commentary[3] and WikiPedia[4] corpus obtained from Open Parallel Corpus (OPUS) (Tiedemann, 2012). French WikiPedia[5] (FrWikipediaMono) was also used which was available as monolingual corpus and was translated to English.

Corpus	Corpus Size	Retrieved Sentences	Unique Sentences
Books	127085	1235684	42827
News Commentary	209479	1244026	72011
WikiPedia	818302	1236092	105880
FrWikiPediaMono	8766978	938834	162743

Table 1: Number of sentences retrieved for each corpus for top-2 using French side of Medline tiltles as queries.

Our data selection strategy is graphically presented in figure 1. We followed the data selection approach based on IR as proposed by (Abdul-Rauf et al., 2016). The choice of corpus to use as queries was a critical one: queries should have maximum biomedical terminologies to enable targeting and choosing domain specific sentences from the general domain corpora. We chose Medline titles as queries hypothesising on the fact that the title essentially contains the specific domain terminology. We used English side of Medline titles as queries when retrieving similar sentences

[1] It must have some biomedical sentences from parliamentary debates on health issues, but the amount will be very little.

[2] http://opus.nlpl.eu/Books-v1.php

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

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

[5] https://www.dropbox.com/s/le4yxfijxt0uiia/frwiki-20181001-corpus.xml.bz2?dl=0

from English side of the corpora and French side of Medline titles as queries for IR from French side. We retrieved 10-best sentences and experimented with top-1, top-2 and top-3 sentences as shown in section 4. Table 1 shows the number of retrieved sentences per each corpus and the unique sentences chosen from these to build our models.

Corpus	**Sentences**
In-domain training data	
Ufal	2358164
Scielo	6827
EDP	2200
Medline Abstracts	51520
Medline Titles	567257
Selective IR training data	
News Commentary-IR1	40645
News Commentary-IR2	60671
News Commentary-IR3	75347
Books-IR1	27938
Books-IR2	39901
Books-IR3	48291
WikiPedia-IR1	46439
WikiPedia-IR2	74595
WikiPedia-IR3	97554
Monolingual	
FrWikipediaMono-IR1	81851
FrWikipediaMono-IR2	133259
FrWikipediaMono-IR3	177266
Development data	
Scielo	3606
EDP	295
Khresmoi	1452
Test Data	
Medline 18	231
Medline 19	442

Table 2: Sentence Pairs for Training, Development and Test sets. Sizes are given for cleaned corpora.

3 Corpora

In this section, we present details of corpora used to train our systems, pre-processing and training parameters. We used the in-domain corpora provided by the organizers along with our mined in-domain sentences from the general domain corpora. The in-domain corpora included were:

- Ufal medical corpus, where a subset of medical corpora were extracted including CESTA, ECDC, EMEA, Subtitles and patTR medical corpus. (Yepes et al., 2017)
- Scielo corpus that included scientific bio-domain articles. (Neves et al., 2016)
- EDP dataset containing documents from EDP database for scientific publications. (Névéol et al., 2018)
- Medline abstracts and titles from publications.(Bawden et al., 2019)

Books, News Commentary, WikiPedia and Fr-WikipediaMono corpora were used as the out-domain corpora to perform in data selective training experiments by extracting relevant in-domain sentences as explained in section 2. Development set included EDP, Scielo and Khresmoi (Dušek et al., 2017). Medline test corpora provided by WMT18 (Neves et al., 2018) and WMT19 (Bawden et al., 2019) were used as test sets.

3.1 Pre-processing

Our pre-processing pipeline includes data cleaning, punctuation normalization, tokenization, truecasing and subword segmentation.

Data cleaning was done to remove noisy data. Some of the provided corpora, including EDP, Scielo and Subtitles, were not completely aligned so we used Microsoft's bilingual sentence aligner[6] (Moore, 2002) for their complete alignment. Empty lines, hyperlinks, parenthesis, white spaces if present at the beginning of sentences were removed. Sentences having more than 120 tokens were dropped using Moses cleaning scripts (Koehn et al., 2007), punctuation and normalization was also applied. Table 2 shows our corpus sizes in terms of number of sentences (after cleaning).

For our French to English systems, we tokenized the corpora using Moses tokenizer[7]. Byte Pair Encoding (BPE) sub word units with a vocabulary of 32K units were computed on true cased data using subword-nmt (Sennrich et al., 2015) . BertTokenizer[8] was only used for our submitted English to French system.

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

[7] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[8] https://huggingface.co/transformers/model_doc/bert.html

ID	Train Set	Size	Test sets	
French to English		(No of sentences)	Medline18	Medline19
Baseline	WMT	2,985,968	29.6	34.7
S1	WMT + News Commentary-IR1	3,026,613	32.5	35.1
	WMT + News Commentary-IR2	3,046,639	**33.1**	**36.8**
	WMT + News Commentary-IR3	3,061,315	29.5	34.2
S2	WMT + Books-IR1	3,013,906	32.7	36.4
	WMT + Books-IR2	3,025,869	**32.9**	**37.0**
	WMT + Books-IR3	3,034,259	29.4	32.7
S3	WMT + WikiPedia-IR1	3,032,407	**33.1**	36.4
	WMT + WikiPedia-IR2	3,060,563	31.9	**37.2**
	WMT + WikiPedia-IR3	3,083,522	29.4	33.6
S4	WMT + FrWikipediaMono-IR1	3,067,819	32.3	36.1
	WMT + FrWikipediaMono-IR2	3,119,227	**32.5**	**36.9**
	WMT + FrWikipediaMono-IR3	3,163,234	31.4	35.5

Table 3: BLEU scores for (BERT-fused NMT) French to English Models trained with selective data training from out-of-domain corpus

3.2 Training and Parameters

We used Fairseq (Ott et al., 2019), an open-source toolkit for training simple transformer (Vaswani et al., 2017) model and Bert-nmt[9] for training BERT-fused NMT systems. Our experiments can be grouped in three categories depending upon the corpora used during training and their training approach. I) Models trained using all the in-domain corpora provided by WMT. II) Models trained on all the in-domain WMT corpora with addition of in-domain corpus retrieved from out-of-domain corpora using IR. III) Models fine tuned on Medline abstracts and titles (since test corpus is from Medline), from few models built in second category. We used transformer base (Vaswani et al., 2017) architecture provided by fairseq as `transformer_iwslt_de_en`. Adam optimizer and a batch size of 4K words was used in all the experiments. Training was done till complete convergence, models were checked for improvements on test data, and training was stopped if no further improvement in BLEU scores is calculated after 2-3 successive checkpoints. For BERT-fused NMT models same training parameters were used as for NMT models except that multilingual bert base was incorporated during training following the approach of (Zhu et al., 2020)

4 Experiments and Results

In this section we report the details of the experiments we performed for our participation in the WMT20 Biomedical task. We performed several different experiments to investigate the performance of NMT with different training approaches. Several different models were trained for French to English translation direction and one model was trained for English to French translation direction. The experiments were conducted as an extension of our last year's submission (Hira et al., 2019) with two different objectives. First, to investigate the performance of BERT-fused NMT over state-of-the art transformer model and the other to explore the effect of out-of-domain corpus used for selective data training. We evaluated our models on Medline 18 and Medline 19 test sets, scores were calculated using sacrebleu (Post, 2018).

4.1 Corpus Selection for Selective Data training

The significant gains in performance due to selective data training, as achieved in our WMT19 participation moved us to explore further to catego-

[9] https://github.com/bert-nmt/bert-nmt

ID	Approach	Training sets	Test sets	
French to English			**Medline 18**	**Medline 19**
M1	Transformer	WMT	33.2	36.3
M2	BERT-fused transformer (cased)	WMT	29.5	32.6
M3	BERT-fused transformer (uncased)	WMT	29.6	34.7
R1	BERT-fused transformer (SD)	WMT + all IR2	31.8	37.2
R2	R1 fine tuned	WMT + all IR2	35.1	**38.4**
R3	BERT-fused transformer (SD)	WMT + all IR3	29.7	34.0
R4	R3 fine tuned	WMT + all IR3	**47.5**	36.8
English to French				
M4	Transformer	WMT + Books + WikiPedia	**32.5**	**35.8**

Table 4: BLEU scores for BERT-fused NMT with IR incorporated French to English models.

rize which out-of-domain corpus is a better choice. We extended out-of-domain corpora to four different resources for selective data training. Alongwith News Commentary, which was also used in WMT19 participation, we extended the list with WikiPedia corpus, Books corpus and back translated FrWikipediaMono corpus. These were used to build four different sets of models from $S1$ to $S4$ as listed in Table 3. Adding the IR retrieved data has unanimously helped improve the scores to almost 3 BLEU points on both test sets.

These models were trained using WMT20 in-domain corpora with addition of selective `top-1`, `top-2` and `top-3` retrieved IR sentences. $S1$ represents models built using additional News commentary IR corpus. Best scores were obtained on `top-2` yielding 33.1 and 36.8 BLEU points on Medline 18 and Medline 19 test sets. $S2$ consists of models trained on additional Books IR corpus and best scores were again achieved on `top-2` giving 32.9 and 37.0 points on Medline 18 and Medline 19 test sets. $S3$ comprises of models trained using additional WikiPedia IR corpus that reveal change in trend by giving best points 33.1 on `top-1` for Medline 18 and 37.2 on `top-2` for Medline 19 test sets. Similarly, systems represented by $S4$ show the effect of adding back translated FrWikipedia-Mono IR corpus in training set, that followed the trend of $S1$ and $S2$ giving best points 32.5 and 36.9 on `top-2` for Medline 18 and Medline 19 respectively. We can safely conclude that `top-2` IR retrieved sentences give us the best score. As for the effect of domain/type of the corpus used for IR, we don't see any significant advantage of any corpus over the other. For example, News Commentary and Books are very different corpora, but still sentences from both the corpora yield more or less the same improvement. Same is the case with WikiPedia, whether parallel or monolingual. This is an expected outcome as the IR process retrieves the sentences most relevant to the query sentence (Medline titles in our case).

4.2 BERT-fused NMT

To target our second objective, investigation of BERT-fused NMT performance over transformer model, we trained three models using in-domain data provided by WMT20 Bio-medical translation task; $M1$, $M2$ and $M3$. And four models, $R1$ to $R4$, using additional IR data, for French to English translation direction. Whereas 1 model ($M4$) for English to French translation direction, as shown in table 4. $M1$ was trained with simple transformer architecture without BERT fusion and it scored 33.2 and 36.3 BLEU points on Medline 18 and Medline 19 test sets respectively. $M2$ was trained under BERT-fused NMT setting with cased multilingual BERT base fused in transformer architecture. This model yielded 29.5 BLEU score on Medline 18 and 32.6 BLEU score on Medline 19 test set. Unexpectedly $M2$, despite being trained in BERT-fused NMT setting, didn't show improvements in BLEU points over simple transformer model ($M1$). One reason of this unexpected decrease in the BLEU scores of $M2$ over $M1$ could be the use of BERT trained on general domain. It seems that BERT trained on much huge general domain corpus has suppressed the learned parameters from in-domain training corpus. $M3$ was trained as similar to $M2$ but with uncased BERT, to explore the difference in the performance of cased and uncased BERT model, and it showed little improvement than $M2$ on Medline 18 test data with a difference of only 0.1 BLEU points whereas an increase of 2.1 BLEU points on Medline 19 test set, as listed in Table 4. Based on this result, we selected uncased BERT model for our further experiments.

Further, we tried to evaluate the performance of selective data training in BERT-fused NMT setting, and trained four models for this investigation as shown in Table 4. $R1$ was trained over in-domain WMT20 corpus concatenated with `top-2` queried all IR data, since these proved to be most beneficial as shown by the results from section 4.1. $R1$ scored 31.8 and 37.2 BLEU points on Medline 18 and Medline 19 test sets respectively. Comparing $R1$ with $M2$ depicts that BERT-fused NMT also benefits from data selective training approach, as the results show considerable increase in BLEU points, increasing 2.3 and 2.5 BLEU scores on Medline 18 and Medline 19 respectively by adding only 0.3M (308426 sentences) IR data. Though the addition of IR data for training $R1$ improved scores compared to $M2$ but did not outperform $M1$ which initiated the need to verify our assumption that general domain BERT is suppressing the learned parameters from in-domain training data. So, for verification we fine tuned $R1$ on Medline abstracts and Medline titles data to train a new system $R2$. $R2$ showed improvements in scores as fine tuned on in-domain corpus (Medline abstracts and titles). It gave highest BLEU score points of 38.4 for medline 19 test set and also producing 35.1 BLEU points on Medline 18. This verify that our assumption about the unexpected results of BERT-fused NMT model was correct. Another model $R3$ was built to test the effect of queried IR data. $R3$ was trained over in-domain WMT20 corpus concatenated with `top-3` queried all IR data. It yielded 29.7 and 34.0 BLEU scores on Medline 18 and Medline 19 respectively. $R3$ is then fine tuned on Medline abstracts and Medline titles data to train a new system $R4$. $R4$ scored highest BLEU points of 47.5 for Medline 18 and gave 36.8 BLEU points for Medline 19 test set.

For English to French translation direction, we trained transformer model with hugging face BERT tokenizer instead of BERT-fused NMT ($M4$). The model was trained with transformer architecture on in-domain data and selective data from Books and WikiPedia corpus. This model ranked third in official results provided by WMT20 and scored 32.5 and 35.8 BLEU points on Medline 18 and Medline 19 test sets respectively.

5 Related Work

Numerous challenges arise when dealing with biomedical data used for translation due to limited size of corpus and unstructured alignments. Various approaches have been adopted by researchers in WMT biomedical translation. (Khan et al., 2018) submitted a NMT system that combined in-domain data set and used transfer learning approach to train the model along with ensemble learning. (Huck et al., 2018) trained by using transformer architecture using biomedical and news domain and employed cascaded word segmentation along with BPE. (Tubay and Costa-jussà, 2018) emphasize on using multi-source approach like Romance languages with in-domain data by implementing transformer architecture using OpenNMT in PyTorch. (Carrino et al., 2019) created terminology list for biomedical words using BabelNet API, inserted the information at a token level and trained NMT system using transformer model (Vaswani et al., 2017). (Hira et al., 2019) used selective learning for building additional corpus from out-of-domain data and incorporated transfer learning approach by using recurrent encoder decoder NN model for training of in-domain biomedical data.(Peng et al., 2019) trained their Transformer model on in-domain and out-of-domain data for six translations using transfer learning methods. The model used attention mechanism along with RELU activation function yielding better results for in-domain biomedical data. (Saunders et al., 2019) used transfer learning using Bayesian Interpolation for multi-domain data for ensemble weighting. (Soares and Krallinger, 2019) participated in WMT19 with four translation directions by creating concatenating corpora from UMLS, out-of-domain and in-domain data and trained the systems using Transformer model.

6 Conclusion

In this paper, we present our submission for WMT20 Biomedical tasks. Our model trained for English to French language direction ranked third in official scores provided by WMT20. We trained different models to investigate the performance of BERT-fused NMT over transformer model and to explore the effect of selective data training in BERT-fused NMT for French to English language direction. Results show decline in performance of BERT-fused NMT models over transformer architecture as general domain BERT suppressed the learned parameters from in-domain training corpus. BERT-fused models yielded better results when fine tuned on in-domain corpus and trained with IR data.

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 and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece. 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.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, et al. 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.

Casimiro Pio Carrino, Bardia Rafieian, Marta R. Costa-jussà, and Josà© A. R. Fonollosa. 2019. Terminology-aware segmentation and domain feature for the wmt19 biomedical translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 153–157, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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 158–165, Florence, Italy. Association for Computational Linguistics.

Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2018. Lmu munichâ^{TM}s neural machine translation systems at wmt 2018. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 659–665, Belgium, Brussels. Association for Computational Linguistics.

Abdul Khan, Subhadarshi Panda, Jia Xu, and Lampros Flokas. 2018. Hunter nmt system for wmt18 biomedical translation task: Transfer learning in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 666–672, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

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 ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Robert C Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144. Springer.

Aurélie Névéol, Antonio Jimeno Yepes, L Neves, and Karin Verspoor. 2018. Parallel corpora for the biomedical domain.

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.

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.

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. *arXiv preprint arXiv:1904.01038*.

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 166–170, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine translation*, 25(4):341–375.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2019. Ucam biomedical translation at wmt19: Transfer learning multi-domain ensembles. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 171–176, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Brian Tubay and Marta R. Costa-jussà. 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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Longyue Wang, Derek F Wong, Lidia S Chao, Junwen Xing, Yi Lu, and Isabel Trancoso. 2013. Edit distance: A new data selection criterion for domain adaptation in smt. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 727–732.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, et al. 2017. Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.