Exploring Transfer Learning and Domain Data Selection for the Bio-medical translation

Noor-e-Hira[1], **Sadaf Abdul Rauf**[1,2], **Kiran kiani**[1], **Ammara Zafar**[1] **and Raheel Nawaz**[3]
[1] Fatima Jinnah Women University, Pakistan
[2] LIMSI-CNRS, France
[3] Manchester Metropolitan University, UK
sadaf.abdulrauf@limsi.fr
{noorehira94,kianithe1,ammarazafar11}@gmail.com

Abstract

Transfer Learning and Selective data training are two of the many approaches being extensively investigated to improve the quality of Neural Machine Translation systems. This paper presents a series of experiments by applying transfer learning and selective data training for participation in the Bio-medical shared task of WMT19. We have used Information Retrieval to selectively choose related sentences from out-of-domain data and used them as additional training data using transfer learning. We also report the effect of tokenization on translation model performance.

1 Introduction

This paper describes the first system submission by Fatima Jinnah Women University under the NRPU project (NRPU-FJ) for the Bio-medical task. We have built our systems using the paradigm of Neural Machine Translation. We worked on translation between French and English (in both directions) and incorporated domain adaption by using selective data training utilizing information retrieval to retrieve domain related sentences from out-of-domain corpus.

Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014), is the current state-of-the-art in Machine Translation. Since its arrival, active research is being done to investigate the field and exploit its benefits to produce quality translations. These efforts have resulted in state of the art translation architectures (Vaswani et al., 2017; Gehring et al., 2017). Despite the winning results of NMT over it's counter part Statistical Machine Translation (SMT) for large training corpora; the quality of NMT systems for low resource languages and smaller corpora is still a challenge (Koehn and Knowles, 2017).

To overcome this challenge various studies explore numerous techniques to improve NMT quality especially in low resource settings. Domain adaptation (Freitag and Al-Onaizan, 2016), transfer learning (Zoph et al., 2016; Khan et al., 2018), fine tuning (Dakwale and Monz, 2017; Huck et al., 2018) and data selective training (van der Wees et al., 2017); are few terms being interchangeably used for such techniques as reported in the literature.

As is common in machine learning approaches, the quality of the system being built depends on the data used to train the system. This was true for SMT systems and still holds significance for NMT based systems (Sajjad et al., 2017; Chu et al., 2017). The domain of the training data is crucial to get quality translations. MT performance quickly degrades when the testing domain is different from the training domain. The reason for this degradation is that the learning models closely approximate the empirical distributions of the training data (Lambert et al., 2011). An MT system trained on parallel data from the news domain may not give appropriate translations when used to translate articles from the medical domain.

The availability of language resources has increased over the last decade, previously this was mainly true only for monolingual corpora, whereas parallel corpora were a limited resource for most domains. Most of the parallel data available to the research community was limited to texts produced by international organizations, parliamentary debates or legal texts (proceedings of the Canadian or European Parliament (Koehn, 2006), or of the United Nations,[1] MultiUN.[2] These only covered specific languages and domains which posed a challenge for the porta-

[1] https://cms.unov.org/UNCorpus/

[2] http://www.euromatrixplus.net/multi-un

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 158–165
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

bility of MT systems across different application domains and also its adaptability with respect to language within the same application domain.

Translation quality of medical texts also suffers due to fewer resources available to train a quality NMT system. Though, medical domain is a growing domain with respect to availability of parallel corpora like scielo (Neves et al., 2016), EMEA (Tiedemann, 2012), Medline (Yepes et al., 2017) and others in making are being made available to the research community.

In this paper we present an approach which aims at increasing the training corpus by mining similar in domain (Bio Med) sentences from out of domain data. We have developed NMT system for English-French language pair, for translation in both directions. Data selective training over cascaded transfer learning, approach has been used to train the model for English to French translation direction; whereas for French to English translation, data selective training approach was used over the whole corpus.

The systems were built with tokenized and untokenized data to study the affect of tokenization in NMT. Tokenization is an important prepossessing step to build MT system. It benefits the MT system by splitting the words into sub-word units, removing punctuations and any other unnecessary tags from the corpus; thus decreasing the vocabulary and helping to translate the unknown words. Tokenization, where, improves the MT system quality it also raises a challenge of developing good quality tokenizers for each language. Studies are performed to investigate tokenization for SMT systems (Zalmout and Habash, 2017; Chung and Gildea, 2009), the question arises how important tokenization is for NMT? Could tokenization be ignored in NMT? (Domingo et al., 2018) investigate tokenization in NMT, to explore the impact of tokenization scheme selected for building NMT, but do not report that, if tokenization is not done, how much will it affect the quality of the NMT system. We present an answer to this question along with other explorations.

The rest of the paper structured as follows; Section 2 provides a brief overview of the related work and background. Section 3 discusses the experimental setup. Results for the different systems are presented in section 3.3. The paper concludes with a brief conclusion.

2 Related Work

This section reports a brief review of the existing literature for machine translation in bio-medical domain. The literature for neural machine translation with the focus of bio-medical domain data is not in abundance. Few studies which we found are discussed followed by a brief overview of transfer learning, domain adaptation and data selective training methods

The system by (Huck et al., 2017) ranked highest in human evaluation in WMT17. They used linguistically informed cascaded word segmentation at the target language side using suffix and compound splitting and BPE. The system was built using attention based gated recurrent units (GRUs).

The techniques used to improve machine translation quality also include selection of best translation among various candidate translations from different translation models. (Grozea, 2018) focuses the mentioned dimension for bio-medical domain NMT system for English Romanian language pair. Percentages were computed for source words which have correspondence in the translation, to select the quality translation. The resultant BLEU scores did not improve more than 0.5.

Khan et al. (2018) trained three NMT systems with different corpus grouping. One experiment included only in-domain corpus, whereas two experiments were performed to train in-domain corpus by initializing the training parameters from general domain system. Learning rate was adjusted to 0.25 and dropout to 0.2, for all the training experiments. The study reveals that training in-domain corpus by transfer learning from general domain corpus increase the MT system quality. The study reports a gain of 4.02 BLEU points over the baseline through transfer learning.

2.1 Transfer Learning

Transfer learning is a process of training a model by utilizing the learned parameters of an already trained model. Learned knowledge of one model is transferred to initiate the training process of a new model for some related models. (Zoph et al., 2016) has defined the process in terms of parent and child model training. The model which is first trained then used to initialize the parameters of a new training process is considered as parent model and the new model which has utilized the knowledge of parent model for its training is considered as child model.

Jointly training both source-target and target-source models minimizes reconstruction errors of monolingual sentences as proposed in the dual learning framework by (He et al., 2016) where two translation models teach each other through a reinforcement learning process. (Wang et al., 2018) also proposed dual transfer learning by sampling several most likely source sentences (target-to-source) to avoid enumerating all source sentences, thus transferring the knowledge from the dual model to boost the training of the primal source-to-target translation.

2.2 Domain adaptation using selective data training

Adaptation using existing parallel texts has shown to be beneficial for translation quality by distributing the probability mass associated with the existing translation phrases. Our method also mostly distributes the probability mass of existing translation phrases and has shown improved results in the paradigm of SMT systems(Abdul-Rauf et al., 2016, 2017). In this study we show the effectiveness of the method in NMT systems. Information retrieval has been previously used in the context of translation model adaptation by (Hildebrand et al., 2005), who use IR to find sentences similar to the test set from the parallel training data. They use the source side of the test data to find related sentences from the parallel corpora. (Lu et al., 2008) use a similar technique of using IR to select and weight portions of existing training data to improve translation performance.

3 Experiments

We have studied two approaches being used to improve NMT in low resource settings. A detailed description of our experiments is provided in this section.

3.1 Corpus

We used in-domain and general domain corpora to train our systems. News-Commentary (Tiedemann, 2012) was used as general domain corpus to perform Information Retrieval for selective data selection. The books corpus was used as the main out-domain corpus. For in-domain corpus we used Medline abstracts training corpus, subset of scielo corpus (Neves et al., 2016), EMEA corpus (Tiedemann, 2012), Medline titles training corpus provided by WMT17 (Yepes et al., 2017), UFAL Medical corpus and Khresmoi corpus. Medline titles corpus was used as test set. Table 1 summarizes the details of our training, development and test corpora.

Corpus	English	French
In-domain:		
EMEA	12.3M	14.5M
Scielo	0.09M	0.1M
UFAL	1.4M	1.5M
Medline Abstracts	1.4M	1.7M
Medline Titles	6.0M	6.7M
Out-domain:		
Books	2.71M	2.76M
News Commentary (nc)	4.9M	5.9M
NC English IR, top-1 (ncSDE)	1.2M	1.5M
NC French IR, top-2 (ncSDF)	2.1M	2.5M
Development set	1.1M	1.2M
Test set	9.2K	10.9K

Table 1: Train, Development and Test set details in terms of number of words (tokenized).

3.2 Data Selection Procedure

We adopted the technique reported in (Abdul-Rauf and Schwenk, 2011) for our data selection procedure. In-domain Medline titles corpus were used as queries to retrieve related sentences from News Commentary corpus. We had a total of 627,576 queries for data selection. Top n ($1 < n < 10$) relevant sentences were ranked against each query. We used just the unique samples to train the systems.

The data selection process was done on both French and English. For the English News Commentary corpus, English side of Medline titles were used as queries and correspondingly for French News Commentary Corpus as Index using French part of Medline titles as queries. Two separate data selection pipelines were executed to investigate the effect of language used for data selection, inspired by the previous results on choice of translation direction reported in (Abdul-Rauf et al., 2016).

ID	Train Set	Detail	Test	
	English to French		**Un-tokenized**	**tokenized**
	Adding in-domain data to *In-domain* baseline:			
M1	em+sc+medAbs+uf	Baseline-in-domain	12.68	15.65
M2	em+sc+medAbs+uf+ncSDF	M1 ⇒ M2	14.57	19.56
M3	em+sc+medAbs+uf+ncSDE	M1 ⇒ M3	14.71	19.76
M4	em+sc+medAbs+uf+NewsComentary	M1 ⇒ M4	14.54	17.75
	Adding in-domain data to *Out-domain* baseline:			
M5	books	Baseline-outdomain	-	4.53
M6	books+em+sc+medAbs+uf	M5 ⇒ M6	-	14.48
M7	books+em+sc+medAbs+uf+ncSDF	M6 ⇒ M7	-	16.12
M8	books+em+sc+medAbs+uf-ncSDF+ncSDE+medTitle	M5 ⇒ M8	-	21.97
	French to English		**Un-tokenized**	**tokenized**
FE	em+sc+medAbs+uf+ncSDF+ncSDE+medTitle		-	15.94

Table 2: BLEU scores for English to French Models and English to French. ⇒ shows the direction of transfer learning while building the models. The best model from IR was chosen which was top2 for French IR and top1 for English IR.

3.3 Training Parameters

We used OpenNMT-py (Klein et al., 2017) to train the models. For English to French translation direction we adopted transfer learning approach along with selective data training. A series of experiments were performed to train a two layer RNN (Recurrent Neural Network) encoder decoder model, having 500 LSTM (Long Short Term Memory) units in each layer. Training was optimized via Adam optimizer and 0.001 learning rate fixed for all the experiments. Whereas for initial experiments we kept the batch size to 64 samples, and afterwards we increased the batch size to 128 samples. Validation was applied after every 10000 steps.

For training NMT system for French to English direction, we followed simple training process. The training model architecture and training parameters were same as for English to French experiments, except that the batch size was set to 128 through out the training process.

4 Results

This section describes the procedure and results of all experiments done by using tokenized and untokenized corpora in the training pipeline. Table 2 and Figure 1 show our results in values as well as graphically for English to French. The section is further sub-divided in two sub-sections, Adding in-domain data to *In-domain* baseline (section 4.1) and Adding in-domain data to *Out-domain* baseline (section 4.2), in which we discuss the results the results on tokenized data following the general MT convention. Effect of tokenization is discussed in the corresponding section 4.3. Experiments were performed with the aim to answer the following research questions:

- How important is the decision for selection of parent model for transfer learning.
- What is the effect of transfer learning when selective data training is initialized from an already trained in-domain model.
- Does selective data training has any benefit over simple training with out-domain corpus.
- How the source or target side data selection affects the translation performance.
- How the performance of a system is affected, if the corpus is not tokenized before starting the training pipeline.

4.1 Adding in-domain data to *In-domain* baseline

Table 1 summarizes the corpora used in our experiments. We have used the general domain News Commentary NC corpus having 4.9M English and 5.9M French tokens to do IR to select medical domain related sentences. We retrieved $top - 10$

sentences from both English and French NC corpus (section 3.2) and built NMT systems to choose the best system. The results of these experiments are graphically depicted in Figure 1. As is evident, selected data training always outperforms the baseline as well as the system built by adding the whole NC corpus to the baseline (row 4 Table 2). We then selected the best systems from both IR pipelines, which were $top-1$ (ncSDE) for English IR yielding 1.2M and 1.5M English and French tokens respectively. For the IR in French direction the best system was $top-2$ (ncSDF) having 2.1M and 2.5M English and French tokens respectively.

Table 2, summarizes the results of all the experiments. We have used short representations to name the corpora used for training. We represented $EMEA$ as em, $Scielo$ as sc, $Medline$ abstracts as $medAbs$, $UFAL$ as uf, selected data of News Commentary using French queries as $ncSDF$, selected data of News Commentary using English queries as $ncSDE$. Right arrow is used to show the application of transfer learning.

For our experiments on transfer learning and selective data training, we first trained a baseline system (M1), by concatenating in-domain EMEA corpus, $Medline$ abstracts corpus, $Scielo$ corpus and $UFAL$ corpus. We didn't add Medline titles corpus in our baseline training pipeline, to get a clear picture of the results of data selective training over transfer learning (as Medline titles were used as a key to select the data from general domain). The BLEU scores of the baseline system, calculated over test set from Medline titles corpus were 15.65.

In the second experiment we applied transfer learning to initialize the selective data training over sentences found by IR from News Commentary French corpus ($ncSDF$) from baseline model. For this experiment data selection was done using the French queries which is the target language in our case. Transfer learning over selective data training improved the system (M2) performance by 4 BLEU points from the baseline. Which is a significant improvement.

The third experiment was done by applying transfer learning to initialize the selective data training from baseline model, but this time data selection was performed using English queries ($ncSDE$). The resulting model (M3) performed better than the baseline with gain of 4.11 BLEU points. Comparing the resulting BLEU scores of

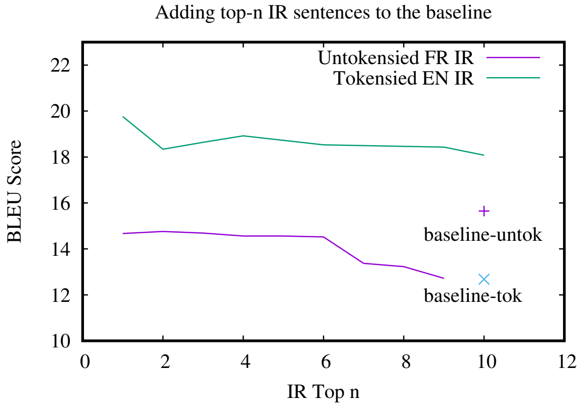

Figure 1: Effect of adding our IR selected data to baseline. In this figure we also show the difference between the the use of tokenized vs untokenized data

both the selective data training experiments, no obvious difference is observed by the change in language to select data.

To explore if selective data training gives any benefit over simple training using whole NC corpus; we built M5. In this experiment we continue to train the baseline system with full News Commentary corpus, which was used for finding domain related sentences for selective data training in above mentioned experiments (M2 and M3). Note that NC corpus is more than double the size of $ncSDE$ and $ncSDF$ (see table 1). The resulting model (M5) only achieved an improvement of 2.1 BLEU points. This clearly demonstrates the efficiency and performance of IR based data selection method.

4.2 Adding in-domain data to *Out-domain* baseline

Table 2 shows the detailed results of our experiments on building NMT systems for English to French translation focusing on the above stated research questions. We first trained the out-domain baseline system (M5) on 2.7M French words of $books$ corpus and getting a baseline score of 4.53 BLEU. We applied transfer learning to initialize the training of 17.2M French words of in-domain $em + sc + medAbs + uf$ corpus to train a new model (M6).

We see that starting from an out-domain baseline $books$ corpus, the addition of in-domain data drastically improves system performance, giving a total gain of around 10 BLEU points (Table 2 row 6). We did not observe this scale of improvement in previous experiments (section 4.1) when in-domain IR selected data was added to in-domain medical corpora.

Then, we evaluate the performance of cascaded transfer learning over selective data training. We applied transfer learning over the model which was first trained by transferring the parameters of baseline-out-domain to train over major in-domain corpus (M6). However, here we see a similar trend when we apply selective data training using $ncSDF$ (M6 ⇒ M7) and resulting improvement is of 1.64 BLEU points. Here, domain data selection exhibited the same trend as we observed in previous section.

In the last experiment we concatenated all the in-domain corpus and trained a model (M8) initiating from out-domain $books$ corpus. Interestingly, this is the best result achieved, giving a total improvement of 17.44 BLEU points from the out-domain $books$ corpus baseline (4.53 ⇒ 21.97). The improvement of 1.64 BLEU points (M6 ⇒ M7) achieved with a rather stronger baseline as in the previous section, strengthens our claim of efficiency of our IR based data selection method using selective data training.

4.3 Effect of Tokenization on translation quality

To study the effect of tokenization on performance of NMT systems, we built four models (from M1 to M4) with both tokenized and untokenized corpora. For the experiments done with tokenized corpora we used $MossesTokenizer$ (Koehn et al., 2007), which is reported to yield best results as compared to other tokenizers (Domingo et al., 2018).

Table 2 lists the results of our findings. All the models built using tokenized corpora significantly out-performed their corresponding counterparts built using untokenized corpora.

M1 dropped in performance, by 2.97 BLEU points when trained with untokenized corpora, than its corresponding system trained with tokenized corpora (12.68 ⇔ 15.65). Same trend is observed in M2 which showed a decline of 4.99 BLEU points by using untokenized corpora during training, in comparison to its training using tokenized corpora (14.57 ⇔ 19.56). The decline in performance of M3, when trained with untokenized corpora is highest. Its performance decreased by 5.05 BLEU points as compared to training using tokenized corpora. M4 maintained the trend of decrease in performance when trained with untokenized corpora. It lost 3.21 BLEU points with respect to its corresponding system trained using tokenized corpora. The trend of decline in performance for untokenized corpora can also be observed from Figure 1.

On average the decline in performance of the systems is around 4 BLEU points, which reveals the importance of tokenization of corpora in NMT. This concludes that tokenization of corpora is an important pre-processing step when building NMT systems.

It must be noted here that the selective data training maintained its trend to perform better than the baseline as well as the system built by adding $ncSDE$, $ncSDF$ and the whole NC corpus to the baseline for the systems built using untokenized corpora. Our IR based data selection method still holds it's efficiency claim here (see $M2$ and $M3$ vs $M4$). This adds to the efficacy of the data selective training approach we adapted to build our systems for domain adaptation.

4.4 French to English

To train the system for French to English direction, we followed simple training pipeline with selective data training using both source and target language as selection queries. We concatenated all the in-domain corpus and trained the system with selective data training from News Commentary. This model (FE) gave 15.94 BLEU score on the test set. The reported BLEU scores from WMT official results are 0.1972 and 0.2105 for all and OK sentences respectively.

5 Conclusion

In this paper, we have described our submission to the Bio Medical task based on the sequence-to-sequence NMT architecture for the WMT2019 shared task. In the Bio-medical task we worked on translation between French and English (in both directions). We used transfer learning approach to train our systems along with selective data training using information retrieval techniques.

We performed a series of experiments to investigate a few important research questions. Data selective training, though done with selected corpus smaller in size, yields better results than using the whole out-domain corpus in training for domain adaptation. Our study also adds to the previous results, that tokenization is an important pre-processing step for NMT and it helps significantly improve the system performance.

Over all our system achieved an improvement of 17.44 BLEU points from the out-domain *books* corpus baseline (4.53 ⇒ 21.97) by adding all in-domain data. The improvement of 4.11 and 1.64 BLEU points in selective data training from in-domain to in-domain and out-domain to in-domain respectively show the efficiency of our IR based data selection method using selective data training methods.

Acknowledgments

This study is funded by the projects: National Research Program for Universities (NRPU) by Higher Education Commission of Pakistan (5469/Punjab/NRPU/R&D/HEC/2016) and the project Alector (ANR-16-CE28-0005) by Agence National de la Recherche, France.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, pages 1–35.

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 (TASLP)*, 24(4):745–754.

Sadaf Abdul-Rauf, Holger Schwenk, and Mohammad Nawaz. 2017. Parallel fragments: Measuring their impact on translation performance. *Computer Speech & Language*, 43:56–69.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *arXiv preprint arXiv:1701.03214*.

Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 718–726. Association for Computational Linguistics.

Praveen Dakwale and Christof Monz. 2017. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. *Proceedings of the XVI Machine Translation Summit*, page 117.

Miguel Domingo, Mercedes Garcıa-Martınez, Alexandre Helle, and Francisco Casacuberta. 2018. How much does tokenization affect in neural machine translation? *arXiv preprint arXiv:1812.08621*.

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

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. JMLR. org.

Cristian Grozea. 2018. Ensemble of translators with automatic selection of the best translation–the submission of fokus to the wmt 18 biomedical translation task–. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 644–647.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the Meeting of the European Association for Machine Translation (EAMT)*, Budapest, Hungary.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017. Lmu munichs neural machine translation systems for news articles and health information texts. In *Proceedings of the Second Conference on Machine Translation*, pages 315–322.

Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2018. Lmu munichs neural machine translation systems at wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 648–654.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

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: Shared Task Papers*, pages 655–661.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn. 2006. Europarl: A parallel corpus for statistical machine translation.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland. Association for Computational Linguistics.

Yajuan Lu, Jin Huang, and Qun Liu. 2008. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Mariana Neves, Antonio Jimeno Yepes, and Aurlie Nvol. 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 2016)*, Paris, France. European Language Resources Association (ELRA).

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural machine translation training in a multi-domain scenario. *arXiv preprint arXiv:1708.08712*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Jrg 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).

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.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and T Liu. 2018. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712*.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondrej 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.

Nasser Zalmout and Nizar Habash. 2017. Optimizing tokenization choice for machine translation across multiple target languages. *The Prague Bulletin of Mathematical Linguistics*, 108(1):257–269.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.