Terminology-Aware Sentence Mining for NMT Domain Adaptation: ADAPT's Submission to the Adap-MT 2020 English-to-Hindi AI Translation Shared Task

Rejwanul Haque*, Yasmin Moslem and Andy Way
ADAPT Centre
*School of Computing, National College of Ireland
School of Computing, Dublin City University
Dublin, Ireland
firstname.lastname@adaptcentre.ie

Abstract

This paper describes the ADAPT Centre's submission to the Adap-MT 2020 AI Translation Shared Task for English-to-Hindi. The neural machine translation (NMT) systems that we built to translate AI domain texts are state-of-the-art Transformer models. In order to improve the translation quality of our NMT systems, we made use of both in-domain and out-of-domain data for training and employed different fine-tuning techniques for adapting our NMT systems to this task, e.g. mixed fine-tuning and on-the-fly self-training. For this, we mined parallel sentence pairs and monolingual sentences from large out-of-domain data, and the mining process was facilitated through automatic extraction of terminology from the in-domain data. This paper outlines the experiments we carried out for this task and reports the performance of our NMT systems on the evaluation test set.

1 Introduction

ADAPT Centre participated in the Adap-MT 2020 Translation Shared Task[1] of the 17th International Conference on Natural Language Processing (ICON 2020).[2] This task aims at evaluating machine translation (MT) systems on the translation of documents from two domains (AI and Chemistry) involving low-resource Indic languages. The task addresses a number of translation directions, and we participated in the English-to-Hindi translation task and focused on translating the AI texts only. To make the readers familiar with the AI translation task and to understand the challenges of this task, we show a couple of sentences from the blind test set in Table 1.

(1)	Machine learning (ML) is a branch of AI that allows chatbots to identify patterns in human language and learn from past conversations.
(2)	Approaches include statistical methods, computational intelligence, and traditional symbolic AI.

Table 1: Sentences from the AI blind test set.

Our MT systems are Transformer models (Vaswani et al., 2017) which were trained using the Marian-NMT toolkit.[3] In this work, we applied different data augmentation and domain adaptation techniques to train our models, such as using synthetic data from target-side monolingual data through the use of back-translation (Sennrich et al., 2016a; Poncelas et al., 2018), mixed fine-tuning (Chu et al., 2017) and on-the-fly model adaption (Chinea-Ríos et al., 2017). As for the latter two approaches, we mined sentences and sentence pairs from large out-of-domain monolingual and parallel corpora, respectively, based on domain terms appearing in the in-domain data. Note that the terms were extracted automatically from the in-domain data.

This remainder of the paper is organized as follows. Section 2 presents our approaches. We describe the resources we utilized for training in Section 3. Section 4 presents the results obtained, and Section 5 concludes our work with avenues for future work.

2 Our Approaches

2.1 Training Data Augmentation

The use of unlabeled monolingual data in addition to limited bitexts for NMT training (Sennrich et al.,

[1] https://ssmt.iiit.ac.in/machinetranslation.html
[2] https://www.iitp.ac.in/~ai-nlp-ml/icon2020/main_prog.html
[3] https://github.com/marian-nmt/marian

Proceedings of the 17th International Conference on Natural Language Processing: Adap-MT 2020 Shared Task, pages 17–23
Patna, India, December 18 - 21, 2020. ©2020 NLP Association of India (NLPAI)

2016a; Zhang and Zong, 2016; Burlot and Yvon, 2018; Poncelas et al., 2018; Caswell et al., 2019) is nowadays a common practice in MT development (Barrault et al., 2020). This has even more impact when applied to the specialised domains and many language pairs, for which obtaining parallel data is a challenge.

In this task, in order to improve our baseline English-to-Hindi Transformer model, we augmented our training data with target-original synthetic data. As in Caswell et al. (2019), in order to let the NMT model know that the given source is synthetic, we tag the source sentences of the synthetic data with the extra tokens. Iterative generation and training on synthetic data can yield increasingly better NMT systems, especially in low-resource scenarios (Hoang et al., 2018; Chen et al., 2019). Since our baseline target-to-source (Hindi-to-English) MT system is already good in quality, it was used to translate the Hindi monolingual data.

2.2 Mixed Fine-Tuning

As for adapting our baseline MT model to the AI domain, we implemented mixed fine-tuning of model parameters, where fine-tuning is conducted on the training data that consists of both in-domain and out-of-domain data as described in Chu et al. (2017). The shared task organisers released parallel training data of the AI domain with a limited number of in-domain examples (only 4,872 sentence pairs). The in-domain data was augmented by oversampling the AI training set several times, and an almost similar sized out-of-domain data set is mined from the parallel (out-of-domain) training corpus on which our baseline NMT system was trained. This strategy worked well for us when we translated business scene dialogue (Jooste et al., 2020) in the WAT 2020[4] (Nakazawa et al., 2020) document-level translation task. However, the adaptation method presented in this paper slightly differs from the conventional mixed fine-tuning (Chu et al., 2017; Jooste et al., 2020), and is described below.

Terms are usually indicators of the nature of a domain and play a critical role in domain-specific MT (Haque et al., 2019, 2020a). Sentences that contain in-domain terms are likely to be in-domain sentences. However, an ambiguous term could have more than one potential meaning. As an example of lexical ambiguity, 'cold' has several possible meanings in the Unified Medical Language System Metathesaurus (Humphreys et al., 1998) including 'common cold', 'cold sensation' and 'cold temperature' (Stevenson and Guo, 2010). Moreover, a polysemous term (e.g. 'cold') could have many translation equivalents in a target language. With this in mind, we mined those training examples (i.e. sentence pairs) from the large out-of-domain domain parallel corpus whose source or target sentences contain at least one domain term. As pointed out earlier, an extracted out-of-domain sentence that contain a domain term may not represent the desired domain; however, the training examples that include such sentences may play a crucial role in minimising lexical selection errors as far as terminology translation in NMT is concerned (Haque et al., 2019, 2020a).

To this end, we exploit the approaches of Rayson and Garside (2000) and Haque et al. (2014, 2018) in order to automatically identify terms in the in-domain texts. The idea is to identify those words which are most indicative (or characteristic) of the in-domain corpus compared to a reference corpus. Haque et al. (2014, 2018) used a large corpus which is generic in nature as a reference corpus. We adopted their approach and used a large generic corpus in order to identify terms in the in-domain source (English) and target (Hindi) corpora. In our setup, we also used the source and target sides of the out-of-domain training bitexts on which our baseline NMT system was trained as the reference corpora. The intuition is again the same, i.e. to extract those (terminological) expressions from the in-domain data that do not occur or rarely occur in the training data and are more indicative of the in-domain AI corpus. Given the lists of source and target terms, we mine sentences independently from the source and target sides of the out-of-domain bilingual corpus. As pointed out above, we select those sentence pairs from the out-of-domain bilingual corpus whose source or target sides contain at least one domain term. In Nayak et al. (2020b), we empirically showed that such "pseudo" in-domain sentences are more effective than those mined using bilingual cross-entropy difference according to the in-domain language model (Axelrod et al., 2011) for NMT model adaptation.

As in Kobus et al. (2017), in order to inform the NMT model about the domain during training and decoding, we add a (domain) tag at the begin-

[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/index.html

ning of the source sentences of the in-domain data, which allows us to control the output domain of the trained system. The NMT system is finally fine-tuned on the mixture of the in-domain and mined out-of-domain corpora.

2.3 Mining Sentences for Fine-tuning

Chinea-Ríos et al. (2017) demonstrated that in the case of specialised domains where parallel corpora are scarce, sentences of a large monolingual data that are more related to the test set sentences to be translated could be effective for fine-tuning the original general domain NMT model. They select those instances from a large monolingual corpus whose vector-space representation is similar to the representation of the test set instances. The selected sentences are then automatically translated by an NMT system built on a general domain data. Finally, the NMT system is fine-tuned with the resultant synthetic data. The synthetic training data whose source-side sentences are original could be more effective for domain adaptation, and the learning method that uses such training data is called 'self-training' (Ueffing et al., 2007). In a similar line of research, it has also been shown that an NMT system built on general domain data can be fine-tuned using just a few sentences (Farajian et al., 2017; Wuebker et al., 2018; Huck et al., 2019).

We followed Chinea-Ríos et al. (2017) in order to mine those sentences from large monolingual datasets that could be beneficial for fine-tuning the original NMT model. As in Jooste et al. (2020); Nayak et al. (2020b); Parthasarathy et al. (2020), we first identified terms in the AI test set to be translated, and given the list of extracted terms, English sentences which were mined from large monolingual data are similar in style to the AI test set sentences. To put it another way, we followed the method described in Section 2.2 in order to extract sentences form large monolingual corpus. The monolingual corpus that we used for this purpose contains 95,918,840 sentences which were sampled from CommonCrawl[5] and Wikipedia Dumps.[6] The English source sentences that have been mined were translated into Hindi using the best MT system (cf. through mixed fine-tuning strategy) to create synthetic data (i.e. source-side original synthetic corpus (SOSC)) to be used for fine-tuning the same NMT model.

3 Data Used and Training Setups

For building our baseline models (forward and backward), we used only the bilingual data provided by the task organisers. As for Hindi monolingual sentences for back-translation, we sampled them from AI4Bharat-IndicNLP Corpus (Kunchukuttan et al., 2020). The out-of-domain parallel data is compiled from a variety of existing sources, e.g. OPUS[7] (Tiedemann, 2012), and after applying standard cleaning procedures including applying a language identifier[8] we are left with just over 1.1 million parallel sentence pairs. Table 2 presents the corpus statistics. The development set

In-domain	sentences	words (EN)	words (HI)
Train	4,872	77,301	82,815
Development	400	7,031	7,064
Out-of-domain	1,102,511	22.4M	23.4M
Hindi Monolingual			
Setup 1	1M		18.8M
Setup 2	7.82M		142.9M

Table 2: The Corpus statistics.

(cf. Table 2) of the AI English-to-Hindi translation task consists only of 400 sentence pairs. For experimentation, we considered its first 200 sentence pairs as development set and the remainder as the evaluation test set. We used two different sized monolingual datasets for our back-translation experiments (cf. last rows of Table 2).

As pointed out earlier, our NMT systems are Transformer models. The tokens of the training, evaluation and validation sets are segmented into sub-word units using Byte-Pair Encoding (BPE) (Sennrich et al., 2016b), and BPE is applied individually on the source and target languages. From our experiences (Jooste et al., 2020; Haque et al., 2020b; Nayak et al., 2020b,a; Parthasarathy et al., 2020) in the participation in the recent shared translation tasks (Barrault et al., 2020; Mayhew et al., 2020; Nakazawa et al., 2020) involving low-resource language pairs and domains, we found that the following configuration usually leads to the best results in our low-resource translation settings: (i) the BPE vocabulary size: 6,000, (ii) the sizes of the encoder and decoder layers: 4 and 6,

[5] http://web-language-models.s3-website-us-east-1.amazonaws.com/wmt16/deduped/en-new.xz

[6] http://data.statmt.org/wmt20/translation-task/ps-km/wikipedia.en.lid_filtered.test_filtered.xz

[7] http://opus.lingfil.uu.se/

[8] https://pypi.org/project/pycld2/

respectively, and (iii) learning-rate: 0.0003. As for the remaining hyperparameters, we followed the recommended best setup from Vaswani et al. (2017). The early stopping criterion is based on cross-entropy; however, the final NMT system is selected as per the highest BLEU score on the validation set. The beam size for search is set to 6. We make our final NMT model with ensembles of 8 models that are sampled from the training run.

4 Experiments and Results

This section presents the performance of our MT systems in terms of the automatic evaluation metric BLEU (Papineni et al., 2002). Additionally, we performed statistical significance tests using bootstrap resampling methods (Koehn, 2004). We obtained the BLEU scores of our MT systems to evaluate them on the test set, and the scores are reported in Table 3. The first row of Table 3 represents our baseline English-to-Hindi MT system. The Hindi-to-English MT system which has been used to translate the Hindi monolingual sentences to English is of good quality (i.e. it produces 28.76 BLEU points on the test set). The BLEU scores of the MT systems (Base2 and Base3) trained on training data that consists of both authentic and synthetic parallel data are shown in the next two rows of Table 3 (cf. Section 2.1).

	BLEU
Base	28.97
Base2 (Base + 1M Syn)	30.80
Base3 (Base + 8M Syn)	29.97
Base2 + Mixed FT	42.02
Base3 + Mixed FT	43.03
Base2 + Mixed FT + ST	43.00
Base3 + Mixed FT + ST	43.51

Table 3: The BLEU scores of the English-to-Hindi NMT systems.

Source–target sentence pairs were mined from out-of-domain training bitexts for mixed fine-tuning (see Section 2.2). The number of sentence pairs that have been mined is 167,234. We also augmented the in-domain parallel corpus via over-sampling in-domain sentences, and by this, the size of the in-domain bitexts becomes 97,440. We finally fine-tuned Base2 and Base3 on the training data that is a mixture of (augmented) in-domain and (mined) out-of-domain data. The BLEU scores of the MT systems (Base2 + Mixed FT and Base3 + Mixed FT) which are the results of the fine-tuning process are presented in the fourth and fifth rows of Table 3. One of our three submission (*Run1*) is with Base3 + Mixed FT. We select Base2 + Mixed FT and Base3 + Mixed FT for further adaptation.

Following the method described in Section 2.3, we mined English sentences (a total of 27,644 sentences) from a large monolingual corpus (cf. Section 2.3) given the list of terms (a total of 356 terms) appearing in the test set. Then, SOSC was created by translating these mined English sentences into Hindi using the respective MT system. Finally, the best MT systems (Base2 + Mixed FT or Base3 + Mixed FT) were fine-tuned on the resultant SOSC. The BLEU scores of the adapted MT systems on the test set are shown in the last rows of Table 3. When we compare the original MT systems with the adapted MT systems, we see that (i) the adapted version of Base2 + Mixed FT, Base2 + Mixed FT + ST, produces a 0.98 BLEU point (corresponding to 2.33% relative) improvement over Base2 + Mixed FT, and (ii) the same of Base3 + Mixed FT, Base3 + Mixed FT + ST, produces a 0.48 BLEU point (corresponding to 1.1% relative) improvement over Base3 + Mixed FT. The former improvement is statistically significant but the latter is not.

As above, we created the adapted MT systems for the blind test set which consists of 401 sentences. Our terminology extraction model identified 1,599 AI terms in the blind test set. We mined 98,009 English sentences from the large monolingual data given the list of terms. We followed the approach described above for fine-tuning our best two models (Base2 + Mixed FT and Base3 + Mixed FT) in order to translate the blind test set sentences. The BLEU scores of our MT systems on the blind test set, which the task organisers published, are shown in Table 4.

MT systems	Submissions	BLEU
Base2 + Mixed FT	*Run1*	35.78
Base2 + Mixed FT + ST	*Run2*	36.71
Base3 + Mixed FT + ST	*Run3*	39.15

Table 4: The BLEU scores of the MT systems on the blind test set.

5 Conclusion

In this paper, we described our MT systems that were submitted to the Adap-MT 2020 AI translation shared task. We presented our results obtained at the time of development of our MT systems. In order to adapt our MT systems to translate texts of AI domains, we subsequently applied two existing fine-tuning techniques while using a term extraction model in the translation pipeline for mining sentences similar to the domain and style of those of the AI data. We showed that, in the case of limited in-domain training data, both out-of-domain data which are selected via term-based mining protocol and in-domain data are useful for fine-tuning model parameters, which essentially provides our best results in this translation task. Furthermore, making use of synthetic parallel data in training also greatly increased the performance of our MT systems. As for the shared task's system rankings, our three submissions *Run3*, *Run2* and *Run1* secured second, third and fourth positions, respectively.

In future, we aim to apply our strategy to other domains and language pairs.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. The publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077 and 18/CRT/6224.

References

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, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–54, Online. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2018. Using Monolingual Data in Neural Machine Translation: a Systematic Study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Belgium, Brussels. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Peng-Jen Chen, Jiajun Shen, Matthew Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott, and Marc'Aurelio Ranzato. 2019. Facebook AI's WAT19 Myanmar-English Translation Task Submission. In *Proceedings of the 6th Workshop on Asian Translation*, pages 112–122, Hong Kong, China. Association for Computational Linguistics.

Mara Chinea-Ríos, Álvaro Peris, and Francisco Casacuberta. 2017. Adapting Neural Machine Translation with Parallel Synthetic Data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark. Association for Computational Linguistics.

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.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-Domain Neural Machine Translation through Unsupervised Adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2019. Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 437–446, Varna, Bulgaria. INCOMA Ltd.

Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2020a. Analysing Terminology Translation Errors in Statistical and Neural Machine Translation. *Machine Translation (in press)*, 34.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020b. The ADAPT System Description for the STAPLE 2020 English-to-Portuguese Translation Task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 144–152, Online. Association for Computational Linguistics.

Rejwanul Haque, Sergio Penkale, and Andy Way. 2014. Bilingual Termbank Creation via Log-Likelihood

Comparison and Phrase-Based Statistical Machine Translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 42–51, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Rejwanul Haque, Sergio Penkale, and Andy Way. 2018. TermFinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation*, 52(2):365–400.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative Back-Translation for Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Matthias Huck, Viktor Hangya, and Alexander Fraser. 2019. Better OOV Translation with Bilingual Terminology Mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5809–5815, Florence, Italy. Association for Computational Linguistics.

Betsy L. Humphreys, Donald A. B. Lindberg, Harold M. Schoolman, and G. Octo Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11.

Wandri Jooste, Rejwanul Haque, and Andy Way. 2020. The ADAPT Centre's Neural MT Systems for the WAT 2020 Document-Level Translation Task. In *Proceedings of the the 7th Workshop on Asian Translation (WAT 2020), AACL-IJCNLP 2020*, page (in press), Suzhou, China.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages. *arXiv preprint arXiv:2005.00085*.

Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous Translation and Paraphrase for Language Education. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 232–243, Online. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th Workshop on Asian Translation. In *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. Association for Computational Linguistics.

Prashanth Nayak, Rejwanul Haque, and Andy Way. 2020a. The ADAPT Centre's Participation in WAT 2020 English-to-Odia Translation Task. In *Proceedings of the the 7th Workshop on Asian Translation (WAT 2020), AACL-IJCNLP 2020*, page (in press), Suzhou, China.

Prashanth Nayak, Rejwanul Haque, and Andy Way. 2020b. The ADAPT's submissions to the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation (Shared Task Papers (Biomedical)*, Punta Cana, Dominican Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Venkatesh Balavadhani Parthasarathy, Akshai Ramesh, Rejwanul Haque, and Andy Way. 2020. The ADAPT system description for the WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation (Shared Task Papers (News))*, Punta Cana, Dominican Republic.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating Backtranslation in Neural Machine Translation. In *Proceedings of The 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, pages 249–258, Alicante, Spain.

Paul Rayson and Roger Garside. 2000. Comparing Corpora using Frequency Profiling. In *The Workshop on Comparing Corpora*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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. Association for Computational Linguistics.

Mark Stevenson and Yikun Guo. 2010. Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus. *Journal of Biomedical Informatics*, 43(5):762–773.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA).

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32, Prague, Czech Republic. 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 6000–6010.

Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 881–886, Brussels, Belgium. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting Source-side Monolingual Data in Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.