The ADAPT's Submissions to the WMT20 Biomedical Translation Task

Prashanth Nayak, Rejwanul Haque and Andy Way
The ADAPT Centre, School of Computing
Dublin City University, Dublin, Ireland
firstname.lastname@adaptcentre.ie

Abstract

This paper describes the ADAPT Centre's submissions to the WMT20 Biomedical Translation Shared Task for English-to-Basque. We present the machine translation (MT) systems that were built to translate scientific abstracts and terms from biomedical terminologies, and using the state-of-the-art neural MT (NMT) model: Transformer. In order to improve our baseline NMT system, we employ a number of methods, e.g. "pseudo" parallel data selection, monolingual data selection for synthetic corpus creation, mining monolingual sentences for adapting our NMT systems to this task, hyperparameters search for Transformer in low-resource scenarios. Our experiments show that systematic addition of the aforementioned techniques to the baseline yields an excellent performance in the English-to-Basque translation task.

1 Introduction

The ADAPT Centre participated in the Biomedical Translation Shared Task of the Fifth Conference of Machine Translation (WMT20). This task is about evaluating systems on the translation of documents from the biomedical domain. The test data consists of biomedical abstracts and terminologies. The task addresses a number of language pairs, and we participated in the English-to-Basque translation task. To make the readers familiar with the biomedical translation task and to understand the challenges of this task, we show a couple of examples from the blind test set and two terminological expressions from terminology test set in Table 1.

For building our MT systems we used the Transformer model (Vaswani et al., 2017). Our strategies to build the competitive MT systems for the task roughly include (i) pseudo in-domain parallel and monolingual data selection, (ii) augmenting training data (Sennrich et al., 2016a; Zhang and Zong, 2016; Burlot and Yvon, 2018; Poncelas et al., 2018; Caswell et al., 2019; Chen et al., 2019), (iii) mining monolingual sentences to adapt our NMT systems to the task, and (iv) finding the optimal hyperparameter configuration for Transformer in this low-resource settings.

(1) No cardiovascular risk factor differences were found in terms of age.
(2) Congenital tumors show a different pattern than tumors in other pediatric ages.
(3) Open bite of thyroid gland, sequela
(4) poisoning by oxytocic drugs, undetermined, subsequent encounter

Table 1: Sentences ((1) and (2)) from the blind test set and sample terminological expressions ((3) and (4)).

The remainder of the paper is organized as follows. In Section 2, we present our methods, and Section 3 details of the data sets used. Section 4 presents the results and discussions, while Section 5 concludes our work with avenues for future work.

2 Our Approaches

2.1 Selecting pseudo In-domain Parallel Sentences

The shared task organisers released parallel training data with a limited number of in-domain examples (only 24,247). The organisers also provided the participants with moderate-sized three out-of-domain corpora (totalling to approximately 770K bitexts). In an attempt to improve the quality of our baseline MT systems, we extracted those sentence-pairs from the out-of-domain corpora that are similar to the styles and domain of the texts we aim to translate, and were used in system building.

2.1.1 Selection using Bilingual Cross-Entropy Difference

We followed the state-of-the-art sentence selection approach of Axelrod et al. (2011) that extracts pseudo in-domain sentences from out-of-domain

corpora using bilingual cross-entropy difference over each side of the corpus (source and target). The bilingual cross-entropy difference is computed by querying in- and out-of-domain (source and target) language models.

2.1.2 Selection using Terminology

Terms are usually indicators of the nature of a domain and plays a critical role in domain-specific MT (Haque et al., 2020). Sentences that contain domain terms are likely to be a domain text. However, a 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). We can see that 'cold' could have very different meanings depending on the context in which it appears. Moreover, a polysemous term (e.g. 'cold') could have many translation equivalents in a target language.

In our second sentence selection approach, we mine those sentences from large out-of-domain or general domain corpus that contain domain terms. As pointed out above, an extracted sentence that contain a domain term may not represent the desired domain; however, the training examples that include such extracted sentences may play crucial role in minimising lexical selection errors as far as terminology translation is concerned (Haque et al., 2020).

To this end, we exploit the approach 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 (Basque) corpora. 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. We select those sentence-pairs from the out-of-domain bilingual corpus whose source or target sides contain at least one domain term.

2.2 Training Data Augmentation

The data augmentation methods in NMT (Sennrich et al., 2016a; Zhang and Zong, 2016; Burlot and Yvon, 2018; Bogoychev and Sennrich, 2019; Caswell et al., 2019; Chen et al., 2019), which usually employ the unlabeled monolingual data in addition to limited bitexts, can positively impact translation quality and are very popular among the MT developers and researchers (Barrault et al., 2019). In other words, use of synthetic data to improve a NMT system is nowadays a common practice, especially in the under-resource scenarios.

The synthetic training data whose source-side sentences are original is more effective for domain adaptation. The learning method that uses such training data is called self-training (Ueffing et al., 2007). The synthetic training data whose target-side is original is more effective for domain text translation and generation of fluent translations (Sennrich et al., 2016a). Many studies (e.g. Chen et al. (2019); Bogoychev and Sennrich (2019)) have shown that self-training and back-translation can be complementary to each other.

In this task, in order to improve our baseline Transformer models, we augmented our training data with both the target- and source-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 source-to-target and target-to-source MT systems are already excellent in quality, those were used to translate the monolingual data.

As in Section 2.1, we extract those sentences from large monolingual data that are similar to the styles of texts we aim to translate. We used the extracted pseudo in-domain monolingual sentences to produce the source- and target-original synthetic bitexts. As for the NMT training, we believe that synthetic parallel data created from pseudo in-domain sentences could be the better alternatives than those selected randomly.

2.2.1 Selection using Language Model Perplexity

Sentences of a large monolingual corpus similar to the in-domain sentences when selected based on the perplexity according to an in-domain language model were found to be effective in MT (Gao et al., 2002; Yasuda et al., 2008; Foster et al., 2010; Axelrod et al., 2011; Toral, 2013). Accordingly, we select "pseudo" in-domain sentences from a large monolingual data based on their perplexity scores accoring to the in-domain language model, which

are then translated to form synthetic training data.

2.2.2 Selection using Terminology

We mine "pseudo" in-domain sentences from large monolingual corpora following the method described in Section 2.1.2. We select those sentences from the monolingual corpus that contain at least one domain term. For mining monolingual sentences we create an efficient Trie structure given the large monolingual data. The idea is to store indices of the sentences (i.e. we restrict this number to 50) for each *n*-gram (upto trigram) of the corpus. Given the domain terms of the in-domain text, we can instantly retrieve the sentences from corpus.

2.3 Mining Sentences for Fine-tuning

Chinea-Ríos et al. (2017) demonstrated that in case of specialised domains or low-resource scenarios 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 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. 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, 2018; Wuebker et al., 2018; Huck et al., 2019).

2.3.1 Mining Source Language Monolingual Sentences

Since English–Basque is a low-resource language-pair and have a little amount of bitexts pertaining to the targeted domain (biomedical), we followed Chinea-Ríos et al. (2017) in order to mine those sentences from large monolingual data that could be beneficial for fine-tuning the original NMT models. In other words, we followed the method described in Section 2.1.2 in order to extract sentences form large monolingual corpus. As above, we identify terms in the test set (i.e. scientific abstracts of Medline) to be translated. As for the sub-task where the task is to translate the domain terms from English to Basque, we observed that many terminological entries are in fact a part of full sentences (e.g. 'person on outside of car injured in collision with pedestrian or animal in traffic accident, initial encounter') and contain general domain tokens. Therefore, we treat the terminological entries as normal sentences and translate them similarly to the Medline abstracts.

In addition to following the standard terminology extraction methods of Haque et al. (2014, 2018) who used a large corpus which is generic in nature as a reference corpus, in a second setup, we used either side of the authentic training bitexts on which the NMT systems were trained as the reference corpus. The intuition is to extract those terminological expressions from the test set that do not occur or rarely occur in the training data and are more indicative of the test corpus. We merged the two sets of terms extracted following the two setups above. Given the resultant list of terms, we mine sentences from monolingual corpus. The source sentences that have been mined are translated with the MT system in order to form synthetic bitexts to be used for adaptation.

2.3.2 Mining Bitexts

Farajian et al. (2017, 2018) exploit the similarity between the source sentences of the training examples and each test sentence and update their generic NMT model on-the-fly on a set of most similar training examples. Like them, we mine training examples form the bilingual training corpus. However, unlike them, our extraction process is driven by the domain terms appearing in the test set which is to be translated. In sum, we follow the bilingual sentence-pair extraction method described in Section 2.1.2 given the test set. For extraction we considered both in-domain and out-of-domain parallel corpora. The extracted bitexts are merged with the generated synthetic segment-pairs above (cf. Section 2.3.1). As in Chinea-Ríos et al. (2017), the best NMT system is finally fine-tuned on the combined train data.

2.4 Tuning Hyperparameters for Transformer

The NMT systems are Transformer models (Vaswani et al., 2017). To build our NMT systems, we used the MarianNMT (Junczys-Dowmunt et al., 2018) toolkit. The tokens of the training, evaluation and validation sets are segmented into sub-word units using Byte-Pair Encoding (BPE) (Sennrich et al., 2016b). We found that performance of the Transformer model more-or-less similar whether BPE is applied individually or jointly on the source and target languages. We kept the former setup, i.e. BPE is applied individually on the source and target languages. Recently, Sennrich and Zhang (2019) demonstrated that commonly used hyperparameter configuration do not lead to the best results in

low-resource settings. Accordingly, we carried out a series of experiments in order to find the best hyperparameter configuration for Transformer in our low-resource setting. In particular, we played with some of the hyperparameters, and found that the following configuration lead 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, respectively, and (iii) learning-rate: 0.0003. The models are trained with the Adam optimizer (Kingma and Ba, 2014), reshuffling the training corpora for each epoch. As for the remaining hyperparameters, we followed the recommended best setup from (Vaswani et al., 2017). The early stopping criteria 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 12. We make our final NMT model with ensembles of 8 models that are sampled from the training run.

3 Data Used

This section presents the data sets which were used for system building. We used the bilingual data provided by the WMT20 Biomedical Shared Task organisers only. As for English monolingual corpus, we used all in-domain texts released by the organisers including the English side of the bilingual corpora of the language-pairs. As for Basque monolingual data, organisers provided us with a tiny set of in-domain sentences. Since the participants are allowed to use external data, we used the CommonCrawl[1] corpus for Basque. Table 2 presents the corpus statistics. The out-of-domain

Bilingual			
in-domain	sentences	words (EN)	words (EU)
train	24,247	201,583	205,334
development	2,000	16,324	16,667
out-of-domain	770,273	12,637,438	11,289,811
Monolingual (sentences)			
	in-domain	CommonCrawl	
Basque	41,151	12,583,122	
English	9,015,051		

Table 2: The Corpus statistics.

parallel corpora for the English-to-Basque task are from three different sources (i.e OPUS (Tiedemann, 2012), IWSLT 2018 (Jan et al., 2018) and WMT16 IT Shared task (Bojar et al., 2016)). We merged segment-pairs of all three data sources, and after applying cleaning scripts to the data we are left with 770K parallel segments (cf. fifth row of Table 2).

[1] https://commoncrawl.org/

Since the size of English in-domain monolingual corpus is reasonably big, we did not use any English out-of-domain data for system building. In order to perform tokenisation for English and Basque texts, we used the standard tool of the Moses toolkit. The development data released by the task organisers contains 2,000 sentences (cf. fourth row of Table 2), out of which 1,000 sentences are used as the test set. The remaining sentences of the development set are used for validation.

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).

4.1 The Baseline MT System

First, we build an English-to-Basque NMT system on the in-domain parallel corpus (cf. Table 2) only, and we refer the MT system as Base. Note that size of the original test set is 1,000 and its sentences were randomly sampled from development set released by the organisers (cf. Section 3). We evaluate Base on the original test set and report its BLEU score in Table 3. As far as the BLEU score on original test set is concerned, it is excessively high. When we looked at the translations, we saw that they are nearly perfect. We

	BLEU
Original test set (1,000)	91.12
test set (200)	47.14

Table 3: The BLEU scores of the basline NMT system (Base).

checked how similar the original test set sentences is to the in-domain training set sentences. For this, we apply fuzzy string matching with a threshold of 80%, and used SimString[2] algorithm (Okazaki and Tsujii, 2010) for search. We found that the number of the non-matching sentences of the test set is 200 (out of 1,000), and same of the development set is 194 (out of 1,000). This indicates that the test and development sets sentences are very similar to those of the training set. The scores on the original test and development sets could be misleading for the evaluation and validation of MT systems. Therefore, for fair evaluation we used the non-matching sentences as the test set (200).

[2] http://www.chokkan.org/software/simstring/

Note that the BLEU scores reported in this paper are on this test set. The BLEU scores of Base on the test set is reported in the last column of Table 3. Similarly, we used the non-matching sentences of the original development set as the development set (194).

4.2 The Improved MT Systems

We applied the pseudo in-domain bilingual sentence selection strategies described in Section 2.1 to the out-of-domain bilingual data (cf. Table 2). We first apply the bilingual cross-entropy difference (BCED) measure described in Section 2.1.1. The so-called pseudo in-domain parallel sentences that were extracted from the out-of-domain data were appended to the in-domain training data, and the BLEU scores of the NMT systems trained on the combined training data are shown in the top rows of Table 4. As can be seen from the table, when the size of pseudo in-domain data is 100K, the MT system (Base+BCED-100K) produces 50.68 BLEU on the test set (a 3.54 BLEU points corresponding to 7.5% relative gain over the Base).

	BLEU
Base+BCED-100K	50.68
Base+BCED-150K	49.02
Base+BCED-200K	47.38
Base+BiTerm	52.19
Base+BiTerm+BCED-100K	53.07

Table 4: The BLEU scores of the NMT systems trained on the in-domain added with the pseudo in-domain training data.

Next, we apply our second method (cf. Section 2.1.2), and the pseudo bilingual corpus extracted following this method contains 294,998 segment-pairs. As above, we append this data to the in-domain data. The BLEU score of MT system (Base+Term) built on the combined data is reported in Table 4. We see from the table that this strategy provides us a 5.05 BLEU points (corresponding to 10.7% relative) gain over the baseline.

When we merge these two pseudo in-domain parallel data with the real in-domain data and train the MT model on the combined data, we further achieved a moderate BLEU gain over the baseline (a 5.95 BLEU points corresponding to 12.6% relative gain). We used this MT system (Base+BiTerm+BCED-100K) for further experimentation, which, from now on, is referred to Base2.

	BLEU
Base2+BT1	52.72
Base2+BT2	53.65
Base2+BT3	53.70
Base2+FT1	52.02
Base2+FT2	51.45
Base2+BT3+FT1	52.76

Table 5: The BLEU scores of the NMT systems trained on augmented training data.

As pointed out above, we augment our bilingual training data with forward and back-translated synthetic data. The BLEU scores of the MT systems trained on the augmented training data are reported in Table 5.

First, we create a synthetic train data by back-translating the tiny monolingual in-domain training data, and the BLEU score of the MT system built on the training data that includes this synthetic data is shown in the second row of Table 5 (i.e. Base2+BT1). This data could not improve Base2.

We extract 275,125 sentences from Basque monolingual data following the method described in Section 2.2.2 (i.e. using the list of terminology extracted from in-domain corpus), and created synthetic bitexts as above. We further add these synthetic bitexts to the training data.[3] The BLEU score of the MT system trained on this data (Base2+BT2) is shown in Table 5. This MT system brings about a 0.58 BLEU points improvement over Base2, and this time, the improvement is not statistically significant.

We further select top 200K target sentences (Basque) based on perplexity scores following the method described in Section 2.2.1. Note that many extracted sentences overlap with those extracted using terminology. We obtained the similar BLEU score on the test set when the synthetic data that is created from this data is further appended to training data (i.e. Base2+BT3).

As mentioned above, we have large monolingual in-domain corpus for English (cf. Table 2). Therefore, we directly used the in-domain English sentences for self-learning. We carried out a number of experiments with adding the source-original synthetic sentences with the original training data, e.g. Base2+FT1 and Base2+FT2 refer to 200K and 1M synthetic segment-pairs. We started doing forward translation with the Medline text. The self-training strategy could not surpass the best-performing MT

[3] Note that this training data refers the one that corresponds to Base2+BT1.

system, i.e. Base2+BT3.

4.3 Fine-tuning the best NMT systems

This section presents the MT systems that were prepared by the adaptation technique described in Section 2.3. We select Base2+BT3 and Base2+BT3+FT1 for adaptation. Following the method described in Section 2.3.1 we mine the source monolingual sentences from the large English in-domain corpus given the list of terms extracted from the test set. Then, synthetic data is created by translating the source sentences by the source-to-target MT systems. We follow the method described in Section 2.3.2 and mine sentence-pairs from in- and out-of-domain bitexts given the list of terms extracted from the test set. The synthetic data and extracted sentence-pairs are merged to form training data for adaptation. Finally, the best MT systems were fine-tuned on this training data. The BLEU scores of the adapted MT systems on the test set are reported in Table 6. When we compare the original MT systems reported in Table 5 with the adapted MT systems, we see that (i) the adapted version of Base2+BT3 produces a 1.1 BLEU points (corresponding to 2.05% relative) improvement over Base2+BT3, and (ii) the same of Base2+BT3+FT1 produces a 1.51 BLEU points (corresponding to 2.87% relative) improvement over Base2+BT3+FT1. The improvements are statistically significant.

	BLEU
Base2+BT3	54.80
Base2+BT3+FT1	55.21

Table 6: The BLEU scores of the adapted MT systems.

As above, we create the adapted MT systems for the blind test set and terminology. Then, we translate the blind test set sentences and terminological entries with the adapted MT systems (Base2+BT3, Base2+BT3+FT1). For our third submission we chose a non-adapted MT system, Base+BiTerm+BCED-100K (cf. Table 4).

In Table 7, we show the BLEU scores of MT systems on the blind test sets. As for abstract translation, Base+BiTerm+BCED-100K is found to be the best system. This system earned us the third position in the task. For the evaluation of terminology translation, in addition to BLEU, the organisers used the accuracy metric which relies on strict matches between ground truth and predictions (cf. Table 7). Base2+BT3 and Base2+BT3+FT1 produce the best BLEU and accuracy scores, respectively, on the terminology test set. Our systems earned us the second position in the terminology translation task.

		BLEU
Base+BiTerm+BCED-100K		**8.67**
Base2+BT3 (adapted)		8.25
Base2+BT3+FT1 (adapted)		8.08
	Acc.	BLEU
Base+BiTerm+BCED-100K	0.73	70.83
Base2+BT3 (adapted)	0.75	**72.39**
Base2+BT3+FT1 (adapted)	**0.76**	71.79

Table 7: Performance of our submitted MT systems in the abstract (top 3 rows) and terminology (bottom 3 rows) translation tasks.

5 Conclusion

This paper presents the ADAPT system description for the WMT20 Biomedical Translation Shared Task. We participated in the English-to-Basque translation task. The task is to translate scientific abstracts and terms from biomedical terminologies. We aimed to build a competitive translation system for this task. For this, we applied various strategies, e.g. selecting monolingual and bilingual texts that are similar to the in-domain data, mining monolingual sentences, applying adaptation technique for adapting the neural MT models to the task, hyperparameters search. We found that our strategies to improve the baseline MT system were effective and yields excellent performance.

This paper demonstrated a novel adaptation approach for translating domain texts. This method is found to be effective in this translation task. In the future, we aim to test the on-the-fly adaptation method (Farajian et al., 2017, 2018) to translate domain texts.

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. This project has partially received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567, and the publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077.

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, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

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.

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.

M. Amin Farajian, Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Evaluation of terminology translation in instance-based neural mt adaptation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 149–158, Alicante, Spain.

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.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 451–459.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):3–33.

Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2020. Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation (in press)*, 34.

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.

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.

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.

Niehues Jan, Roldano Cattoni, Stüker Sebastian, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–6.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, Beijing, China.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA. ACL.

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.

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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. 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 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2214–2218, Istanbul, Turkey.

Antonio Toral. 2013. Hybrid selection of language model training data using linguistic information and perplexity. In *Proceedings of the second workshop on hybrid approaches to translation*, pages 8–12.

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. *arXiv preprint arXiv:1811.01990*.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

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.