Using Confidential Data for Domain Adaptation of Neural Machine Translation

Sohyung Kim **Arianna Bisazza** **Fatih Turkmen**
University of Groningen
s.kim22@student.rug.nl
{a.bisazza, f.turkmen}@rug.nl

Abstract

We study the problem of domain adaptation in Neural Machine Translation (NMT) when domain-specific data cannot be shared due to confidentiality or copyright issues. As a first step, we propose to fragment data into phrase pairs and use a random sample to fine-tune a generic NMT model instead of the full sentences. Despite the loss of long segments for the sake of confidentiality protection, we find that NMT quality can considerably benefit from this adaptation, and that further gains can be obtained with a simple tagging technique.

1 Introduction

The availability of in-domain data remains essential to ensure the quality of Neural Machine Translation (NMT), especially in technical domains (Koehn and Knowles, 2017). However, obtaining such data is often challenging, and in many real-world scenarios this is further aggravated by data confidentiality or copyright concerns. In fact, when data content is sensitive, the owner may simply deny providing its Translation Memories to the translation company it is hiring (Cancedda, 2012). This can lead to considerably worse MT quality, higher post-editing efforts, and subsequently higher translation costs for the data owners themselves.

When the complete data cannot be shared in its original form, releasing *fragmented* data can be considered as a compromise. The most well-known example of releasing fragmented data is *Google N-gram* (Michel et al., 2011). N-gram tables consisting of sequences of *n* words and their counts in a given corpus were routinely used to train count-based language models (Kneser and Ney, 1995; Brants et al., 2007) before the advent of neural methods. However, N-grams are not optimal for training state-of-the-art NLP models such as sequence-to-sequence LSTM (Bahdanau et al., 2015) or Transformers (Vaswani et al., 2017). In fact, one of the main strengths of these models is the ability of handling arbitrarily long contexts, which would be hindered by the use of fragmented data. In this paper, we take a pragmatic approach and ask: If the data owner can *only* release fragmented data due to confidentiality issues, can this still benefit downstream NMT quality in any way?

Motivated by the brittleness of NMT in out-of-domain settings (Koehn and Knowles, 2017) and the increasing availability of large pre-trained models (Ng et al., 2019), we focus on the task of adapting a strong-performing general-domain NMT system to various technical domains. We show that fine-tuning on *phrase pairs* can be a viable solution to exploit confidential data, but the scale of improvements varies strongly across target domains.

2 Background

To our knowledge, the use of confidential data in MT has not received much attention recently. Cancedda (2012) proposed an encryption-based (one-time pad) method for phrase-based statistical machine translation (PB-SMT). However, PB-SMT is nowadays clearly outperformed by NMT (Bentivogli et al., 2016), which function completely differently and therefore require new solutions to preserve data confidentiality.

In the broader context of NLP, secure multi-party computation (Feng et al., 2020) and homomorphic encryption (Al Badawi et al., 2020) have been used to provide strong privacy guarantees. Since these cryptographic methods incur high performance penalties (see (Riazi et al., 2019) for an overview of their performance in deep learning), more recent proposals have focused on the careful use of simpler cryptographic primitives while training a model over encrypted text due to confidentiality reasons. For instance, TextHide (Huang et al., 2020) allows to perform natural language understanding tasks while requiring the participants to complete an encryption step in a federated setting. The aforementioned studies mostly focus on

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 46–52
July 5–10, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.26615/978-954-452-056-4_006

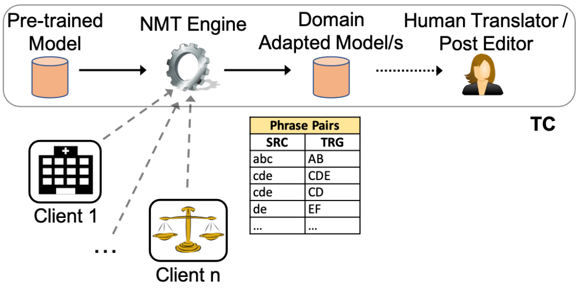

Figure 1: Motivating scenario: a Translation Company (TC) uses confidential data from its clients to adapt a pre-trained generic NMT system to different technical (e.g. medical, legal) domains.

preventing explicit/implicit leakage of partial information while training the models. By contrast, we explore the possibility of using fragmented data to improve state-of-the-art NMT applications.

Scenario As illustrated in Figure 1, we consider a common case where a translation company (TC) provides professional services based on a pipeline of NMT and human post-editing. TC wants to improve the quality of its NMT models by training or adapting them on the clients' previously translated data. Due to confidentiality concerns, the clients only provide their data in a fragmented form as a compromise. If this kind of data can be used to improve the NMT model, both the clients and the company will benefit by abating human post-editing costs. Thus, we want to study the possibility of sharing fragmented data for improving utility while preserving the confidentiality of data.

Threat Model We assume an honest but curious model in which the receiver of the partial data (e.g. the translation company) is untrusted or only partially trusted. The main threat we focus on is the **full reconstruction** of the original text from a list of given n-grams of phrases rather than the protection of partial information (e.g. key phrases (Hard et al., 2018), names, social security numbers). This setting is useful in various contexts where only partial data release is desired such as copyright protection. Examples of text where sensitive information is encoded in long sequences (sentences or paragraphs) include patent applications, as well as not (yet) publicly available product analysis reports or drug reaction reports.

3 Approach

Releasing fragmented data in the form of N-grams has a long tradition in NLP (Michel et al., 2011). However, fixed-size N-gram extraction is not directly applicable to parallel data because it breaks translation equivalence with the target side. As a solution, we propose to use *phrase pairs* (Koehn et al., 2003) as a text fragmentation method.

3.1 Phrase Pairs

Like N-grams, phrases are short sequences of consecutive words extracted from the input sentences. Unlike N-grams, phrases are always extracted in pairs from source-target sentence pairs in a way that is *consistent* with their word-level alignment. Formally, a phrase pair $(\bar{f}, \bar{e})$ is consistent with word alignment A if all source words $f_1, \cdots, f_n$ in $\bar{f}$ that have alignment points in A are connected with target words $e_1, \cdots, e_m$ in $\bar{e}$ and viceversa (Koehn et al., 2003; Koehn, 2009). As Figure 2 illustrates, the words of the target language (German) are first automatically aligned (grey connecting lines) with the words of the source language (English) by a statistical alignment model. Then, phrase pairs of various lengths (denoted by boxes) are extracted.

Phrase pairs and their statistics constitute the main component of PB-SMT systems, together with the target language model. In this work, however, we only use phrase extraction as a text fragmentation technique. After extraction, we shuffle the large set of phrase pairs extracted from the whole dataset and, finally, discard a random sample of phrase pairs (e.g. 50%) to preserve confidentiality. In the example of Figure 2, this would mean protecting the hypothetically sensitive connection between the drug name (*Abraxane*) and its reported side effect (*tiredness*).

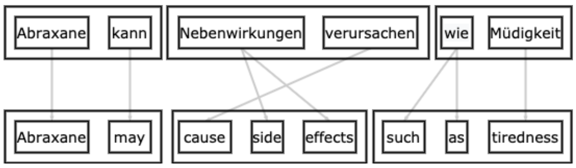

Figure 2: Example sentence pair from the EMEA corpus with extracted phrase pairs of maximum length 3 (every black box is a phrase). Grey lines denote word alignment. Shorter phrases imply more data protection.

3.2 Domain Adaptation

NMT models are trained on full sentences, and their ability to capture large context is one of their main strengths compared to classical SMT approaches. As a result, training NMT on fragmented data is likely to lead to a very poor performance. Nonetheless, we postulate that phrase pairs may still contain very valuable information for the *adaptation* of a general-domain system to a specific target domain. In fact, much of domain adaptation has to do with learning new words or short phrases, as well as new senses for known words and phrases (Irvine et al., 2013). As the domain adaptation technique, we choose fine-tuning (Luong et al., 2015; Sennrich et al., 2016b) which consists of continuing training a previously trained model on a, typically smaller, in-domain dataset.

We start by directly fine-tuning a general-domain NMT system on a random sample of phrase pairs (occurrences, not types) extracted from the in-domain dataset. Since this is expected to bias the model to produce shorter sentences, we also experiment with a simple phrase tagging technique (Sennrich et al., 2016a) so that the model may learn to represent the special nature of phrases and be less inclined to produce short outputs when translating full sentences in the test phase.

4 Experimental setup

We evaluate our approach on German-English in the domains of medicine descriptions, software manuals, and EU legislation. To simulate a realistic production setup, we start from a strong NMT system pre-trained on large amounts (28M sentences) of publicly available data.

Baseline NMT We use the Transformer-based system (Vaswani et al., 2017) pre-trained by Facebook for the WMT'19 news translation task (Ng et al., 2019)[1] and released as part of the Fairseq toolkit (Ott et al., 2019). This model was ranked first in the WMT'19 news competition (Barrault et al., 2019) with a BLEU score of 40.8.

Datasets We simulate confidential translation data by using publicly available datasets from three technical domains:[2] EMEA (medical), GNOME (software) and JRC-Acquis (legal) (Tiedemann, 2012; Steinberger et al., 2006).[3] Data statistics are shown in Table 1. Following the Fairseq model pipeline, we segment our data with FastBPE byte-pair encoding (Sennrich et al., 2016c).[4]

Type	Domain	#sent	#tok(DE)	#tok(EN)
Train	EMEA	10k	199k	209k
	GNOME		179k	194k
	JRC		279k	396k
Valid	EMEA	150	3k	3k
	GNOME		3k	3k
	JRC		4k	5k
Test	EMEA	2k	38k	42k
	GNOME		29k	30k
	JRC		53k	82k

Table 1: Size of datasets used in our fine-tuning experiments. The baseline NMT model was pre-trained on a separate corpus of 28M sentence pairs, not shown here.

Phrase extraction We first word-align the in-domain datasets using FastAlign (Dyer et al., 2013)[5] and compute the union of source-to-target and target-to-source word alignment links (known as union symmetrization heuristic) to obtain the alignment A. Then we use the phrase extraction utility from the Moses phrase-based SMT toolkit (Koehn et al., 2007)[6] to extract all phrases consistent with A. After the phrase extraction step, our dataset has been fragmented into a list of aligned phrases of various lengths. We experiment with a maximum source-side phrase length of either 4 or 7 words, and in both cases we randomly discard 50% of the extracted phrases (occurrences, not types).

Fine-Tuning During fine-tuning, we provide phrase pairs to the models as if they were sentence pairs. Note that this data is shuffled and has many duplicates. Additionally, we experiment with a simple tagging technique by adding <P> and </P> at the front and end of each phrase respectively, in both source and target side. During testing, full sentences with no tags are given to the model.

We apply the hyper-parameters described by Ng et al. (2019) with only a few adjustments inspired from previous work on fine-tuning regularization (Miceli Barone et al., 2017) and tuned on a small (full-sentence) validation set in each domain (150 sentences, see Table 1). Specifically, learning rate is divided by 4 (0.000175), weight decay rate is set to 0.0001 and dropout probability to 0.2. The same small validation set is used for early stopping.

[1] https://github.com/pytorch/fairseq/tree/master/examples/translation

[2] To simulate a professional translation scenario, we split the datasets by documents. We release the benchmarks at https://github.com/Sohyo/Using-Confidential-Data-for-NMT

[3] https://opus.nlpl.eu

[4] https://github.com/glample/fastBPE

[5] https://github.com/clab/fast_align

[6] http://www.statmt.org/moses

	Baseline	Fine-Tuning				
		Max length 4		Max length 7		*Original data*
	(No fine-tuning)	No tag	Tag	No tag	Tag	*(Full sentences)*
EMEA	35.5	39.1	40.5	41.5	37.2	*45.2*
GNOME	29.8	36.0	37.0	35.8	36.8	*38.9*
JRC	29.0	29.4	30.0	29.2	29.7	*54.7*

Table 2: BLEU scores of German-English NMT in three different domains: medical (EMEA), software (GNOME), and legal (JRC). The baseline is the pre-trained Fairseq WMT19 news system (Ng et al., 2019) based on Transformer (Vaswani et al., 2017) and ranked first in the WMT19 competition.

5 Results

We evaluate the quality of NMT models by BLEU (Papineni et al., 2002) computed with SACREBLEU (Post, 2018). The phrase-adapted models are compared to the non-adapted baseline (Ng et al., 2019), and to fine-tuning on the original (non fragmented) dataset in order to determine the maximum possible gains. Results are reported in Table 2.

Our main finding is that phrase pairs can indeed be used to fine-tune a NMT model without any changes to the architecture or the need of specific fine-tuning algorithms. The BLEU gains over the non-adapted baseline vary between +7.0 on EMEA and +1.0 on JRC. This is relevant for our scenario because even translation companies without significant in-house NMT expertise could easily apply our solution to their workflow. Our approach is also applicable in cases where TC uses NMT as an outsourced (cloud-based) service, by sending the provider phrase pairs instead of full sentences for model adaptation.

Effect of phrase tagging The addition of tags appear to improve NMT quality in most cases. Figure 3 shows that tagging yields slightly longer system outputs, suggesting the model indeed learned to associate the <P> tag with shorter training samples. While differences look small, they have a large impact on BLEU because of the Brevity Penalty (Papineni et al., 2002). As a notable exception to this positive trend, BLEU score decreases with tagging on EMEA (max length 7). We are currently investigating this result further.

Effect of phrase length We expected longer phrases to be considerably more useful for fine-tuning, at the expense of less confidentiality protection. By contrast, increasing the maximum length from 4 to 7 does not have a positive effect on BLEU but actually lowers it in the GNOME and JRC domains. This counter-intuitive result may be due to the fact that increasing the maximum length leads to a much larger number of extracted phrases that are redundant and overlapping. Previous work on lexicon-augmented NMT also reported negative results when fine-tuning on very large numbers of segments (Thompson et al., 2019b). In future work, we plan to experiment with minimum phrase length as a way to reduce the total number of phrase pairs.

Domain differences The benefits of fine-tuning on phrases appear to vary strongly across domains: on EMEA we obtain large gains but there is still space for improvement, on GNOME our approach nears the ceiling of fine-tuning on the original data, whereas on JRC gains are small and scores remain very far from the ceiling. To explain these results, we inspected our datasets and specifically looked for peculiarities of the JRC dataset. We find that JRC is rather different in terms of sentence length distribution, with much longer sentences on average. As shown in Figure 3, only fine-tuning on the original data leads to reasonably long outputs, whereas baseline and phrase-adapted systems all generate sentences that are, on average, about 10 words shorter than they should be. This suggests that our tagging technique is not sufficient to address the shorter-output bias in a robust way. Recent techniques to prevent overfitting during fine-tuning (Kirkpatrick et al., 2017; Thompson et al., 2019a) may overcome this problem in future work.

6 Conclusions

We have studied the problem of domain adaptation of NMT models when domain-specific data cannot be shared due to confidentiality or copyright concerns. Inspired by a common NLP practice of sharing confidential data in the form of N-grams (Michel et al., 2011), we propose to use phrase extraction (Koehn et al., 2003), shuffling and subsampling as a data fragmentation technique for translation data. Our experiments on three different

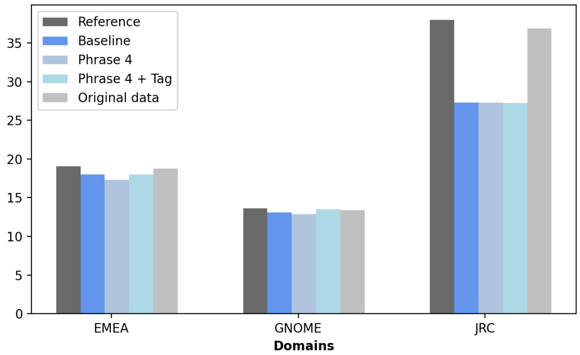

Figure 3: Average length (in tokens) of reference translations and outputs of different NMT systems, including a non-fine-tuned baseline and four differently fine-tuned systems.

domains show that this type of data can be used to fine-tune NMT models leading to considerable improvements on top of a strong baseline and further gains when a simple phrase tagging technique is used. We also find that the magnitude of these gains varies largely across domains, which we tentatively attribute to the different length profiles of our datasets (e.g. legal domain has much longer sentences than the other domains).

While our results show that text fragmentation is indeed compatible with modern machine translation systems adaptation, more work needs to be done before our method can be applied on actual sensitive data. To this end, we plan to determine metrics for the quantification of confidentiality protection (or violation) when an adversary tries to reconstruct the original documents. Our starting point for this direction would be Gallé and Tealdi (2015), who presented a technique for this purpose only in the context of (monolingual) N-grams.

Acknowledgements

We are grateful to Global Textware for giving us early input on the motivating scenario of this work. Arianna Bisazza was partly funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646.

References

Ahmad Al Badawi, Louie Hoang, Chan Fook Mun, Kim Laine, and Khin Mi Mi Aung. 2020. Privft: Private and fast text classification with homomorphic encryption. *IEEE Access*, 8:226544–226556.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Nicola Cancedda. 2012. Private access to phrase tables for statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 23–27, Jeju Island, Korea. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Qi Feng, Debiao He, Zhe Liu, Huaqun Wang, and Kim-Kwang Raymond Choo. 2020. Securenlp: A system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security*, 15:3709–3721.

Matthias Gallé and Matías Tealdi. 2015. Reconstructing textual documents from n-grams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 329–338.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. 2020. Texthide: Tackling data privacy for language understanding tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 1368–1382. Association for Computational Linguistics.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184 vol.1.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Minh-Thang Luong, Christopher D Manning, et al. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the international workshop on spoken language translation*, pages 76–79.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494, Copenhagen, Denmark. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

M Sadegh Riazi, Bita Darvish Rouani, and Farinaz Koushanfar. 2019. Deep learning on private data. *IEEE Security & Privacy*, 17(6):54–63.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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. 2016c. 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.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019a. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019b. HABLex: Human annotated bilingual lexicons for experiments in machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1382–1387, Hong Kong, China. Association for Computational Linguistics.

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

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.