Effectively Aligning and Filtering Parallel Corpora under Sparse Data Conditions

Steinþór Steingrímsson
Department of Computer Science
Reykjavik University
Iceland
steinthor18@ru.is

Hrafn Loftsson
Department of Computer Science
Reykjavik University
Iceland
hrafn@ru.is

Andy Way
School of Computing
ADAPT Centre
Dublin City University
Ireland
andy.way@adaptcentre.ie

Abstract

Parallel corpora are key to developing good machine translation systems. However, abundant parallel data are hard to come by, especially for languages with a low number of speakers. When rich morphology exacerbates the data sparsity problem, it is imperative to have accurate alignment and filtering methods that can help make the most of what is available by maximising the number of correctly translated segments in a corpus and minimising noise by removing incorrect translations and segments containing extraneous data. This paper sets out a research plan for improving alignment and filtering methods for parallel texts in low-resource settings. We propose an effective unsupervised alignment method to tackle the alignment problem. Moreover, we propose a strategy to supplement state-of-the-art models with automatically extracted information using basic NLP tools to effectively handle rich morphology.

1 Introduction

Machine translation (MT) quality has improved substantially with the advent of neural machine translation systems (NMT). However, while the quality gains over statistical machine translation (SMT) systems can be large, in low-resource and domain mismatch settings they are significantly reduced (Koehn and Knowles, 2017). In recent years, unsupervised NMT trained only on monolingual corpora has attracted considerable attention, and has been proposed for scenarios where there is a lack of bilingual data (Artetxe et al., 2018b; Lample et al., 2018). These methods have been shown to perform well for related language pairs (e.g. Wu et al. (2019)), but as the languages differ more the unsupervised methods become less effective (Leng et al., 2019). Kim et al. (2020) show that supervised and semi-supervised baselines with only a small parallel corpus of 50K bilingual sentences consistently outperform the best unsupervised systems for a range of languages, similar and distant. They also show that unsupervised NMT is very sensitive to domain mismatch, which poses a problem to low-resource language pairs where it can be difficult to match the data domain on both sides. Thus, it is evident that to achieve high quality MT, sentence aligned-texts in two or more languages are required.

NMT systems have been shown to be sensitive to noise in the training data (Khayrallah and Koehn, 2018), where noise is defined as segments that decrease output quality of systems trained on the data. It is, therefore, important to be able to accurately align multilingual texts and precisely filter out misalignments and bad translations that adversely affect performance. In the study, conducted on the impact of various types of noise on MT quality, untranslated and misaligned segments had the most detrimental effect. Misaligned segments were by far the most prevalent type of noise in the ParaCrawl[1] parallel corpus they used, twice as common as accepted segments. However, misalignments vary; a segment can have one extraneous word, it can have twice the content its counterpart has, or anything in between. It can be very useful to understand the intricacies of the effects different types and levels of noise have, why it is important not to have noise and whether some kinds of noise are more acceptable than others. This leads us to our first research question:

RQ1: How do different kinds of misalignments in a parallel corpus affect translation quality of an MT (SMT or NMT) system trained on that corpus?

If we can measure the effects of various misalignments, it could help us construct more effective methods to filter parallel corpora for MT.

[1] https://paracrawl.eu/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 182–190

As the usefulness of parallel corpora for MT was first becoming apparent, Harris (1988) pointed out that aligning such texts was a serious problem. Moreover, collecting multilingual texts is expensive and time-consuming, and for some languages it can be hard to obtain access to even small amount of texts. Thus, we need to be able to make the most out of what is available.

We describe a method using Bleualign (Sennrich and Volk, 2011) and Monoses (Artetxe et al., 2018b), an unsupervised SMT system, to align parallel corpora using only monolingual texts for training. The proposed method is language pair-independent and only assumes unaligned bitexts and monolingual corpora for both languages. It is the first step towards answering our second research question:

RQ2: How can we best build useful parallel corpora from bilingual texts, having no other resources but monolingual corpora?

In morphological typology, languages can be classified as analytic or synthetic (see e.g. Haspelmath and Sims (2013), Steinbergs (1996)). Analytic languages primarily rely on word order and auxiliary words to convey meaning, while synthetic languages use inflection. "Morphologically rich" languages are synthetic languages which commonly have a large number of different surface forms for any given lexeme. This can lead to a high rate of out-of-vocabulary (OOV) words, a data sparsity problem that machine learning algorithms struggle with.

Icelandic is a synthetic language with relatively few native speakers (approx. 350,000) where data sparsity problems are prevalent in most NLP tasks. In our work, we will focus on building a parallel corpus for the English-Icelandic language pair and confronting the issues that arise when working with a less-resourced and morphologically rich language.

When doing sentence alignment and filtering noise from parallel corpora, the sparsity problem caused by rich morphology leads to lower confidence scores for segment pairs resulting in lower classification accuracy, and thus smaller or less accurate parallel corpora. When ParIce (Barkarson and Steingrímsson, 2019), an English-Icelandic parallel corpus was compiled, the filtering process resulted in an estimated 20% reduction in corpus size. Out of what remained, about 5% was faulty (see Section 3). We will work with the same data with the goal of minimising these numbers. This leads us to the third and last research question this research proposal centres around:

RQ3: How can we filter parallel corpora to minimize noise, and still lose little or no useful data from the original texts?

Our approach to try to answer these questions is to experiment with common and recent methods from the alignment and filtering literature. We will build a toolset that can employ various known methods and compare and contrast them. We will investigate how word embeddings, a lemmatizer, a part-of-speech (PoS) tagger or a parser can help tackle the data sparsity problem, and which known methods benefit most from them. Evaluation data sets will be created for the purposes of the project and the methods evaluated according to a set of evaluation metrics. Finally, we will train and evaluate our system on a different language pair with comparable issues.

2 Related Work

Filtering parallel data is the task of removing incorrect translations, noise and otherwise faulty data from a set of two (or more) aligned texts. Alignment is the task of finding target segments with a corresponding meaning to that of source segments in multilingual texts. While these may seem to be different tasks, the same methods may apply partly to both problems. Filtering is often done by scoring sentences and removing the lowest-scoring ones, whereas in alignment the highest-scoring sentences can be used as anchors: elements in the data that can reliably be aligned and thus direct further processing. In the next subsections, we describe alignment and filtering methods used in prior work.

2.1 Alignment

The first approaches to automatic sentence alignment were length-based. Gale and Church (1991) found that "the correlation between the length of a paragraph in characters and the length of its translation was extremely high". Motivated by that, they describe a method for aligning sentences based on a simple statistical model of character lengths. Brown et al. (1991) also describe a length-based method, but use tokens instead of characters. In addition, they use signals in the markup as anchor points to segment the corpus into smaller chunks.

Kay and Röscheisen (1993) used bilingual lexicons induced from the corpus being aligned.

Haruno and Yamazaki (1996) show that combining an induced lexicon with an external dictionary yields better results. Papageorgiou et al. (1994) use part-of-speech, commonly preserved in translation, by computing the optimum alignment based on the PoS-tags. Tschorn and Lüdeling (2003) use a morphological analyzer to improve a dictionary-based distance measure, and Ma (2006) increases the robustness of a lexicon-based aligner by assigning greater weights to less frequent translated words.

Sennrich and Volk (2010) use machine translations and BLEU (Papineni et al., 2002) as a similarity score to find reliable alignments to use as anchor points. The gaps between the anchor points are filled using BLEU-based and length-based heuristics.

Thompson and Koehn (2019) describe a method based on bilingual sentence embeddings, using the similarity between the embeddings as the scoring function for alignment.

2.2 Filtering

Recently, neural networks have been used to find anchor points and detect misalignments. Many of these methods have been devised to extract parallel sentences from comparable corpora, by training classifiers to determine if source and target sentences are parallel.

Earlier work includes employing the IBM models (Brown et al., 1993) for word alignment. Khadivi and Ney (2005) filter out the noisy part of a corpus based on IBM models 1 and 4 and length-based models, and score the alignments on a linear combination of these. Taghipour et al. (2011) do outlier detection and show that their filtered corpus results in improved translation quality, even though sentences have been removed. Sarikaya et al. (2009) use context extrapolation to boost the sentence pair coverage, checking whether the distance of the sentences from an anchor point is the same, and whether the sentences have the highest similarity score compared to other pairs within a window, despite being below a defined threshold.

Crosslingual word embeddings have been used to calculate distance between equivalences in different languages (Luong et al., 2015; Artetxe et al., 2016). Defauw et al. (2019) treat filtering as a supervised regression problem and show that Levenshtein distance (Levenshtein, 1966) between the target and MT-translated source, as well as cosine distance between sentence embeddings of the source and target, are important features. While they use InferSent (Conneau et al., 2017), BERT (Devlin et al., 2019) has recently been employed for calculating crosslingual semantic textual similarity to detect misalignment with good results (Lo and Simard, 2019).

Zipporah (Xu and Koehn, 2017) uses a logistic regression model trained to classify sentence pairs. Noisy data is synthesized and used as negative samples in training. BiCleaner (Sánchez-Cartagena et al., 2018) uses a set of handcrafted hard rules to detect flawed sentences and then proceeds to use a random forest classifier based on lexical translations and several shallow features such as respective length, matching numbers and punctuation. Finally, it scores sentences based on fluency using 5-gram language models.

In 2019, at the fourth Conference on Machine Translation, WMT, the shared task on parallel corpora filtering focused on low-resource conditions. The method central to the best-performing submission was the use of crosslingual sentence embeddings, trained from parallel sentence pairs (Chaudhary et al., 2019). Artetxe and Schwenk (2019a) devised a similar method. Both papers tackle the inconsistencies of cosine similarity by investigating the neighbourhood of a given sentence pair, outperforming systems using only cosine similarity.

3 Experimental Framework

The continuum of morphologically rich languages is quite diverse with the one end of the continuum being agglutinative languages, that primarily rely on discrete particles for inflection, and the other being fusional languages, which tend to use a single inflectional morpheme to denote multiple features. While it may be worthwhile to investigate if the same unsupervised methods work across different language categories, it can be expected that if further processing is needed, different approaches have to be taken. Decompounding (Alfonseca et al., 2008) may be more useful for agglutinative languages to tackle the OOV problem, and for many fusional languages internal change and suppletion call for different approaches. In our study we focus on fusional languages. English is primarily an analytic language and Icelandic a fusional language with moderately rich morphology. We will be using the English-Icelandic language pair as a test case.

3.1 Data

ParIce, an English-Icelandic parallel corpus, was compiled from data consisting of 4.3 million translation segments. It was aligned with LF Aligner, which uses Hunalign (Varga et al., 2005), and then filtered using a sentence-scoring algorithm based on a bilingual lexicon bag-of-words method and a comparison between the original segment and an MT-generated translation. The filtering process resulted in 3.5 million translated segments. Manual evaluation of approximately 2000 sample pairs from the corpus indicate that approximately 5% are faulty, while over 50% of the deleted segments are estimated to be faulty using automatic methods.

From these numbers we can deduce that in the raw 4.3 million segment ParIce corpus, there are approx. 3.7 million good segments and around 600K faulty ones. Many of the faulty segments in the corpus are due to misalignment. We will be working with the raw data that made up the 4.3 million segment ParIce corpus. In order to compile a better corpus, we need improved alignment methods to reduce the number of faulty alignments, and we need a classifier that is able to identify the quality of the segments with high precision and recall in order to build as big a corpus as possible with as few faulty segments as possible.

3.2 Evaluation

We are building three evaluation sets, for alignment, filtering, and MT, all sub-sampled and extracted from the ParIce corpus. The MT evaluation set will contain 3000 manually aligned and error-free segments. The alignment evaluation set will have 2000 manually aligned sentences and the filtering set 2000 automatically aligned segments, each assigned one of four classes: correct, partially misaligned, partially incorrect translation, incorrect.

To evaluate the usefulness of our methods for MT, we will use our aligned and filtered corpora to train SMT and NMT systems and compare the results to a baseline where the raw ParIce corpus is used for training.

3.3 Tools and Models

In Section 4, we will discuss some of the methods we will be experimenting with. These include applying a variety of available tools and models as well as developing our own. ABLTagger (Steingrímsson et al., 2019) will be used for PoS-tagging Icelandic texts. The tagger employs biLSTMs and an external morphological lexicon (Bjarnadóttir et al., 2019). Lemmatising will be carried out using Nefnir (Ingólfsdóttir et al., 2019). For all English processing we will use tools available in the NLTK toolkit (Bird et al., 2009) or SpaCy.[2]

We will focus on the most common word embedding models: word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017) and ELMo (Peters et al., 2018). As using bilingual sentence embeddings with BERT has been shown to be effective for filtering (Lo and Simard, 2019), we want to experiment with different contextualized embedding models. The main hindrance with these models is the massive computational resources needed to train, which may limit our possibilites.

For alignment and filtering we experiment with Bleualign, Hunalign and vecalign for sentence alignment, Giza++ (Och and Ney, 2003) for word alignments, and Zipporah, BiCleaner and LASER (Artetxe and Schwenk, 2019b) for filtering, and possibly to help with anchoring the parallel texts for more effective alignment.

Moses (Koehn et al., 2007) will be employed for phrase-based SMT and our NMT system uses the reference implementation of Vaswani et al. (2017) of the transformer-base architecture that is part of the Tensor2Tensor package (Vaswani et al., 2018).

4 Research Plan

Our first goal is to set up an unsupervised pipeline for aligning parallel texts. While this is the first step in tackling RQ2, it is also necessary to devise a method to answer RQ1. We will outline how we seek to answer these questions, as well as RQ3. A secondary goal is to investigate methods to improve upon the unsupervised pipeline by exploring how basic NLP tools can help us deal with the data sparsity problem inherent to many morphologically rich languages. In the following subsections we describe how we intend to research these questions.

4.1 Unsupervised Alignment

Our initial pipeline for aligning parallel texts is trained only on monolingual corpora. While this is a starting point for language pairs lacking pre-existing parallel corpora or glossaries to use with alignment, it also serves as a baseline to compare to when additional processing modules are added, such as a lemmatizer or other NLP tools.

[2] https://spacy.io/

	LF Aligner			Bleualign + Monoses		
	Regulatory texts	Literary texts	Total	Regulatory texts	Literary texts	Total
Aligned pairs	184	69	253	166	61	227
- of which correct	143	57	79.1%	154	54	91.6%
- of which faulty	41	12	20.9%	12	7	8.4%
Aligned words	2470/2485	1652/1652	4122/4137	2427/2485	1539/1652	3966/4137
- of which correct	1980	1337	80.5%	2110	1539	92.0%

Table 1: Alignment results for both systems and number of source language words in the alignments. When no alignment was found the segments were discarded.

As stated in Section 1, we initially employ Bleualign for unsupervised alignment, but instead of bootstrapping an initial training set with length-based methods like Sennrich and Volk (2011), we train Monoses and use that to provide Bleualign with machine translations of the sentences being aligned. Monoses is trained by building cross-lingual word embeddings from monolingual corpora using word2vec and Vecmap (Artetxe et al., 2018a), inducing a phrase table. An SMT system is then trained on this data and used to translate the monolingual corpus in one of the two languages. The translated data is then used to train a standard SMT system in the opposite direction. A new phrase table is built and the process iterated three times for a final model.

To investigate the feasibility of our method we aligned two parallel texts, selected randomly from the ParIce data. We compared the results to LF Aligner, which employs Hunalign. To be able to evaluate the alignment methods accurately, evaluation sets are being compiled (see Section 3.2). Here, we present preliminary results acquired by manually evaluating the alignments. Results, given in Table 1, show that the Bleualign + Monoses method gives better results as measured by accuracy of the aligned pairs, with a total of 91.6% of the resulting pairs correctly aligned, vs. only 79.1% of the alignments by LF Aligner. Although our method yields 10% fewer aligned pairs, it results in a parallel corpus which has substantially more correct alignments both in terms of absolute numbers and percentage of alignments, regardless of whether we are looking at aligned pairs or aligned words.

There are a variety of ways to improve upon the unsupervised method. By training larger word embedding models we can increase the vocabulary. By investigating common *n*-grams within word embedding models we may be able to better pinpoint phrases or multi-word expressions. By extending the iteration process to the bitexts by selecting the highest-scoring sentence pairs after training and alignment, and add them to the training set of the SMT system, we would have more accurate training data, and probably derive better translations after each iteration. That in turn would likely raise the confidence for selecting the best alignments.

4.2 Investigating Misalignments

After setting up alignment pipelines and creating evaluation sets, we will initiate the filtering process using methods and strategies that have previously given good results for other language pairs.

One aspect of the filtering process is to decide which noise is most important to filter out. While Khayrallah and Koehn (2018) highlight the importance of filtering out certain types of noise in parallel corpora, we want more fine-grained results. We will conduct a similar study but investigate different classes of misalignments especially. This will help us decide whether to treat all misalignments the same or if some are worse than others.

We will do this by using available tools (see Section 3.3) to aggressively filter out possible faulty alignments to have as clean a corpus as possible. We will then systematically change the alignments to introduce different types of misalignments in the corpus. The effects of these variations will be investigated by training both SMT and NMT systems, and comparing the effect on changes in resulting translations. This method is intended to give us insight into the problem we pose in RQ1. We will use the results to help us make decisions on how to best set up a filtering system.

4.3 Filtering

We then start the filtering process again, with information about which type of faulty sentences

are likely to have the worst effect on MT systems trained by the data. To try to answer RQ3 we will investigate the practicality of applying different mechanisms to scoring sentences. We will look at features such as sentence length; word similarity based on dictionary lookup, both using an external dictionary and an induced one from raw parallel data; word similarity from word embeddings; distance between a machine-translated source sentence and the target sentence; and sentence similarity scores based on bilingual sentence embeddings.

4.4 Language Independence

After studying the effects of misalignments on MT systems and finding a good balance between the different mechanisms used for scoring the aligned segments, we will investigate the extent of this balance being language pair-dependent by running the same process for other language pairs. These could be English-Irish, Danish-Faroese or others that have some of the same characteristics the English-Icelandic pair has, e.g, at least one morphologically rich language and data sparsity. This will give us further insight to answer the three research questions posed in Section 1.

4.5 Aligning Morphologically Rich Languages

While the first goal is to create a completely unsupervised pipeline for building parallel corpora, applicable to any language pair, we also want to investigate the case of morphologically rich languages specifically by extracting latent information in the data that can help us tackle the data sparsity problem. This includes lemmas derived from the word forms, PoS-tags or constituent structures as additional features for sentence-pair scoring, and by training embedding models, both to help with the morphology and with semantics for unknown words. For this we use available tools such as a PoS-tagger and lemmatizer to try to outperform the unsupervised method alone. For many languages these tools are not available, as they usually rely on training data which may not exist for low-resource languages. Pursuing our second goal we will thus consider the case of a low- to medium-resource language which is morphologically rich and for which basic NLP tools are available. For the language pair selected as our test case, English-Icelandic, all necessary NLP tools are available, so successful methods can subsequently be tested on other language pairs. Furthermore, the only parallel corpus available for Icelandic is rather small and quite noisy and there is a pressing need to improve on it. For proof-of-concept we want our methods to achieve that goal.

No machine-readable English-Icelandic dictionary is available, and if we want to try to use semi-supervised methods for the language pair we will thus need to induce a lexicon from the parallel data, monolingual data or both. Other methods for building a glossary may include using external data such as Wiktionary or Wikipedia, and using available dictionaries in different languages for pivoting.

One of the products of this research will be a toolset to produce parallel corpora from multilingual texts. The software should: align bilingual parallel texts; filter bilingual parallel corpora; be modular; be language-pair independent – although optional language-specific features can be used; use external tools for linguistic annotation: PoS-tagging, parsing, lemmatising, machine translation or other methods that may be beneficial; offer a variety of strategies for aligning and filtering, depending on available resources; and it should aim at accuracy at the cost of speed.

5 Summary

We have given an overview of the literature on sentence alignment and parallel corpus filtering. We outlined challenges associated with implementing these methods for low-resource and morphologically rich languages and proposed initial experiments to tackle these challenges. The motivation for this research is to improve the quality of machine translations by making better use of and increasing the quality of parallel training data, especially in regard to sparse data scenarios. An unsupervised method that effectively aligns bilingual texts will lower the barrier for building high-quality MT systems for low-resource languages and our first results suggest that it may also play a role in improving MT for morphologically rich languages.

Acknowledgements

This work is supported by the Language Technology Programme for Icelandic 2019-2023, funded by the Icelandic government, and by the ADAPT Centre for Digital Content Technology which 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.

References

Enrique Alfonseca, Slaven Bilac, and Stefan Pharies. 2008. Decompounding query keywords from compounding languages. In *Proceedings of ACL-08: HLT, Short Papers*, Columbus, Ohio.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised Statistical Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Mikel Artetxe and Holger Schwenk. 2019b. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, Berkeley, California.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, Florence, Italy.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Arne Defauw, Sara Szoc, Anna Bardadym, Joris Brabers, Frederic Everaert, Roko Mijic, Kim Scholte, Tom Vanallemeersch, Koen Van Winckel, and Joachim Van den Bogaert. 2019. Misalignment Detection for Web-Scraped Corpora: A Supervised Regression Approach. *Informatics*, 6(3):35.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, Minneapolis, Minnesota.

William A. Gale and Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, Berkeley, California.

Brannon C. Harris. 1988. Bi-text, a New Concept in Translation Theory. *Language Monthly*, 54:8–10.

Masahiko Haruno and Takefumi Yamazaki. 1996. High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, Santa Cruz, California.

Martin Haspelmath and Andrea D. Sims. 2013. *Understanding Morphology*. Taylor & Francis, New York.

Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason, and Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland.

Martin Kay and Martin Röscheisen. 1993. Text-Translation Alignment. *Computational Linguistics*, 19(1):121–142.

Shahram Khadivi and Hermann Ney. 2005. Automatic Filtering of Bilingual Corpora for Statistical Machine Translation. In *Natural Language Processing and Information Systems*, pages 263–274, Berlin. Springer Berlin Heidelberg.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? *ArXiv*, abs/2004.10581.

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*, Prague, Czech Republic.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, Canada.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Chi-kiu Lo and Michel Simard. 2019. Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *NAACL Workshop on Vector Space Modeling for NLP*, Denver, Colorado.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Harris Papageorgiou, Lambros Cranias, and Stelios Piperidis. 1994. Automatic Alignment in Parallel Corpora. In *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico.

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*, Philadelphia, PA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistic*, New Orleans, Louisiana.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels.

Ruhi Sarikaya, Sameer Maskey, R. Zhang, Ea-Ee Jan, D. Wang, Bhuvana Ramabhadran, and Salim Roukos. 2009. Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In *Proceedings of INTERSPEECH 2009*, Brighton, United Kingdom.

Rico Sennrich and Martin Volk. 2010. MT-based Sentence Alignment for OCR-generated Parallel Texts. In *The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based Sentence Alignment of Parallel Texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*, Riga, Latvia.

Aleksandra Steinbergs. 1996. The classification of languages. In William O'Grady, Michael Dobrovolsky, and Francis Katamba, editors, *Contemporary Linguistics*, 3rd edition, chapter 9, pages 372–415. Longman, Harlow, UK.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel Corpus Refinement as an Outlier Detection Algorithm. In *Proceedings of MT Summit XIII*, Xiamen, China.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. 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)*, Hong Kong, China.

Patrick Tschorn and Anke Lüdeling. 2003. Morphological knowledge and alignment of English-German parallel corpora. In *Proceedings of the Corpus Linguistics 2003 conference*, Lancaster, UK.

Dániel Varga, Péter Halácsy, András Kornai, Nagy Viktor, Nagy László, Németh László, and Tron Viktor. 2005. Parallel corpora for medium density languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, Boston, Massachusetts.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, Long Beach, California.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Extract and Edit: An Alternative to Back-Translation for Unsupervised 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)*, Minneapolis, Minnesota.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.