Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches

Steinþór Steingrímsson[1], Pintu Lohar[2], Hrafn Loftsson[1], and Andy Way[2]
[1]Department of Computer Science, Reykjavik University, Iceland;
[1]ADAPT Centre, School of Computing, Dublin City University, Ireland
steinthor18@ru.is, pintu.lohar@adaptcentre.ie,
hrafn@ru.is, andy.way@adaptcentre.ie

Abstract

Parallel sentences extracted from comparable corpora can be useful to supplement parallel corpora when training machine translation (MT) systems. This is even more prominent in low-resource scenarios, where parallel corpora are scarce. In this paper, we present a system which uses three very different measures to identify and score parallel sentences from comparable corpora. We measure the accuracy of our methods in low-resource settings by comparing the results against manually curated test data for English–Icelandic, and by evaluating an MT system trained on the concatenation of the parallel data extracted by our approach and an existing data set. We show that the system is capable of extracting useful parallel sentences with high accuracy, and that the extracted pairs substantially increase translation quality of an MT system trained on the data, as measured by automatic evaluation metrics.

1 Introduction

High quality MT systems rely on the availability of parallel data. In low-resource settings, where parallel data is scarce, unsupervised methods have been proposed, where only monolingual corpora are used for training (Artetxe et al., 2018; Lample et al., 2018). Kim et al. (2020) show that supervised and semi-supervised approaches with only a small parallel corpus of 50K bilingual sentences consistently outperform the best unsupervised systems for a range of languages. However, there is a scarcity of parallel data, especially for languages with a low number of speakers. When parallel corpora are scarce, comparable corpora, which are far more common, can be used to supplement it. We will be working with the English–Icelandic language pair, for which no statistical or neural MT work had been published until last year (Jónsson et al., 2020).

When parallel sentences are extracted from comparable corpora, potential parallel sentence candidates can usually come from anywhere in two comparable documents. This means that a potential parallel counterpart of one sentence in the source-language document can be any sentence in the target-language document. If the average number of sentences in a comparable document is n, the number of potential sentence pairs that have to be evaluated are n^2. This quickly becomes overwhelming (as n increases) and so it is imperative to reduce the search space. Reducing the search space should ideally result in a list of a maximum of kxn candidates, where k is a constant number of allowed candidates for each sentence in the comparable documents. To retrieve useful sentence pairs from this list, the pairs have to be scored and filtered.

Our approach divides the problem into two main steps. We start by extracting parallel sentence candidates using an inverted index-based crosslingual information retrieval (CLIR) tool called *FaDA* (Lohar et al., 2016), that requires a collection of documents in two languages and only a bilingual lexicon without the need of any MT system. In the second step, we score the sentence candidates using two different scores, one based on contextualized embeddings and the other on high-precision word alignments. A binary classifier selects sentence pairs based on these scores.

We test our approach in three different ways. We use two different test sets to measure precision, recall and F1-scores, and we also use our approach to extract parallel sentences from Wikipedia and use the resulting data as supplemental data for training NMT systems. The systems are then evaluated in terms of BLEU scores (Papineni et al., 2002) and compared to a baseline in order to give an indication of the usefulness of the supplemental data for NMT training.

Proceedings of 14th Workshop on Building and Using Comparable Corpora in conjunction with RANLP-2021,
pages 8–17, held online, 6 September 2021.
https://doi.org/10.26615/978-954-452-076-2_003

Our main contributions are fourfold.

- We show that the combination of three different measures – CLIR, and scores based on contextualized embeddings and high precision word alignments – can effectively extract parallel sentence pairs from comparable corpora.
- We introduce WAScore, a score based on high precision word alignments and show its usefulness in filtering parallel sentence pairs.
- We publish two different test sets for measuring the effectiveness of parallel sentence extraction from comparable corpora for the English–Icelandic language pair.
- We publish a set of parallel sentences extracted from Wikipedia, shown to be useful for MT training.

2 Related Work

Comparable corpora have been shown to be a useful source for mining parallel segments that can help improve MT quality (Wolk et al., 2016; Hangya and Fraser, 2019). Afli et al. (2015) extract parallel data from a multimodal comparable corpus from the Euronews[1] and TED[2] web sites. Chu et al. (2015) extract parallel texts from the Chinese and Japanese Wikipedia and Ling et al. (2014) employ a crowdsourcing approach to extract parallel text from Twitter data in order to find the translations in tweets. The work of Karimi et al. (2018) describes the approach of extracting parallel sentences from English–Persian document-aligned Wikipedia entries. They use two MT systems to translate from Persian to English and the reverse and then use an IR system to measure the similarity of the translated sentences. Multilingual sentence embeddings have also been applied to the problem, obtaining state-of-the-art performance (Schwenk, 2018; Artetxe and Schwenk, 2019b). Recently, Ramesh et al. (2021) describe the collection of parallel corpora for 11 Indic languages from diverse comparable corpora using LaBSE embeddings (Feng et al., 2020), a language-agnostic BERT sentence embedding model trained and optimized to produce similar representations for bilingual sentence pairs that are translations of each other.

Word alignments have previously been used for parallel sentence extraction. Zariņa et al. (2015) identify parallel sentences using word alignments, experimenting with five different alignment based scores. They presume that if a pair of sentences are equivalent in two languages, there should be many word alignments between the sentences, and non-parallel sentences should have few or no word alignments. Stymne et al. (2013) use alignment based heuristics to filter out sentence pairs. Lu et al. (2020) use a word alignment based translation score as a part of their scoring ensemble for filtering a noisy parallel corpus. Their translation score is a simplified version of the translation score introduced by Khadivi and Ney (2005). Azpeitia et al. (2017) and Andoni Azpeitia and Garcia (2018) describe a method using CLIR and lexical translations obtained using word alignments, with a simple overlap metric. They obtained the highest results for the BUCC 2017 and BUCC 2018 shared tasks.

Our method uses an IR system to create a list of alignment candidates, thus reducing the search space. It then takes advantage of both LaBSE embeddings and word alignments. Our word alignment score is calculated by a simpler formula than most of the previous work, but relies on high precision alignments. It has been shown that they can be achieved by an ensemble method using *CombAlign* (Steingrímsson et al., 2021). A binary classifier is finally used to select acceptable sentence pairs.

3 Data

For the language pair we are working with, English–Icelandic, no test sets have previously been made available for parallel sentence extraction from comparable corpora. Therefore, we have to build test sets in order to be able to evaluate our approach. We prepare the following data sets for our experiments:

- *CompNews*: development and test sets using available news data,
- *CompWiki*: a manually curated small test set for Wikipedia data,
- *CompTrain*: training data for our logistic regression classifier, and
- *CompLex*: an English–Icelandic lexicon for word translation in an IR system.

[1] https://www.euronews.com/

[2] https://www.ted.com/

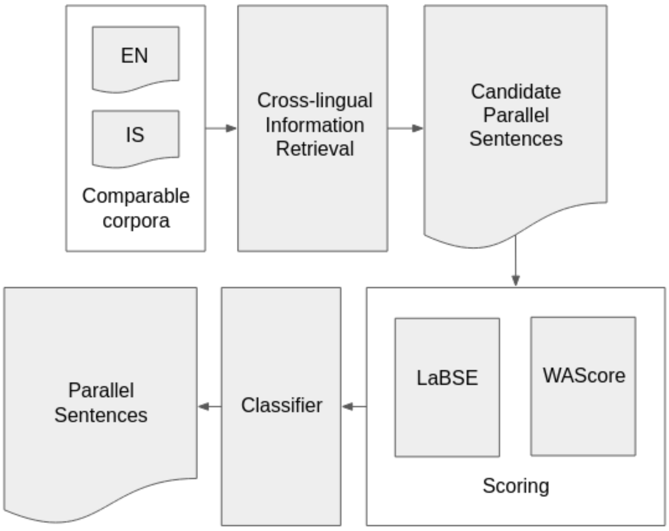

Figure 1: The system setup. English and Icelandic monolingual data are aligned by the CLIR system which outputs candidate pairs which are scored and a classifier outputs parallel sentence pairs.

All the data sets are published with open licenses on GitHub and in a CLARIN repository.

3.1 CompNews

We built development and test sets for identifying parallel sentences in news corpora, in similar style to the test sets compiled for the BUCC 2017 shared task on parallel sentence identification (Zweigenbaum et al., 2016), i.e. consisting of a small set of known parallel sentences, as well as a larger list of randomly sampled sentences from monolingual corpora in the same domain, but with no known parallel pairs. The parallel sentences used are the 2000 English-Icelandic sentence pairs made available as development data for the news translation task in WMT 2021.[3] The dev set for WMT 2021 contains 1000 sentences in each direction. The non-parallel sentences were randomly selected from Newscrawl 2018, and 2018 news texts sampled from the Icelandic Gigaword Corpus (Steingrímsson et al., 2018).

The texts were split into sentences. This resulted in two lists of $100,000$ sentences, English and Icelandic, with 2% of sentences in each list known to have a corresponding sentence in the other language.

We made a $40/60$ split, taking care that the true parallel sentence pairs were equally distributed between the splits. The smaller part was used as a development set and the larger part as a test set.

3.2 CompWiki

We randomly selected 15 Wikipedia articles available in both Icelandic and English. The texts were split into sentences and the CLIR tool (see Section 4.1) used to obtain translation candidates for each sentence. These sentence pairs were manually evaluated and marked as parallel, partially parallel or non-parallel. Out of a total of $10,098$ sentences, 86 were marked parallel and 421 as partially parallel.

3.3 CompTrain

In order to gain some information on the kind of scores the two scoring methods give to non-parallel data, on the one hand, and parallel data, on the other hand, we compiled a dataset with $50,000$ randomly sampled pairs from the two monolingual corpora used for CompNews and added parallel sentences from the English–Icelandic ParIce corpus (Barkarson and Steingrímsson, 2019). We selected $2,500$ random sentence pairs from a development set published with the corpus and filtered all sentences that have a minimum length of six tokens. This resulted in $1,743$ sentence pairs, marked as positive data for a classifier. The resulting $51,743$ sentence pairs are scored in the same way we score the parallel sentence candidates (see Section 4.2) and used to train the classifier.

[3] Available at: http://statmt.org/wmt21/translation-task.html

3.4 CompLex

FaDA, the cross-lingual information retrieval tool we use to obtain parallel pair candidates, requires a bilingual lexicon with lexical translation probabilities of words. It uses the lexicon to translate the query terms in the source language and searches these translated terms in the target-language index to retrieve the equivalent candidate sentences in the target language. It is described in more details in Section 4.1. As such a lexicon did not exist, we compiled it using a combination of approaches. We collected data that was available online, an English–Icelandic dictionary from Apertium (Brandt et al., 2011), Wiktionary entries and Wikipedia article titles. We obtained permission to use the bilingual ISLEX-dictionaries (Úlfarsdóttir, 2014), which go from Icelandic to five Nordic languages (Danish, Faroese, Finnish, Norwegian and Swedish) and used these to pivot to English using the aforementioned open dictionaries. We created word lists using word alignments to extract pairs from the ParIce corpus after lemmatizing both languages using SpaCy[4] for English and Nefnir (Ingólfsdóttir et al., 2019) and DIM (Bjarnadóttir et al., 2019) for Icelandic. We selected the most likely English equivalents for a list of Icelandic words using crosslingual word embeddings models based on Vecalign[5] (Thompson and Koehn, 2019). In addition, we translated both Icelandic words and words from the Nordic ISLEX-dictionaries, using models from OPUS-MT (Tiedemann and Thottingal, 2020). This resulted in a long list of word translation candidates which we then filtered using a threshold that required that each candidate was suggested by multiple sources. For each source word, we counted how many sources suggested that candidate and used the count to assign likelihood scores to the translations. This resulted in two files, an English–Icelandic lexicon with $140K$ entries and an Icelandic–English lexicon with $152K$ entries.

4 System Description

4.1 Sentence Alignment Using CLIR

We make use of an open source CLIR-based bilingual document alignment tool called *FaDA* (Lohar et al., 2016) in the first step of the alignment process. This tool is capable of aligning bilingual documents without the help of any MT system. In contrast, the MT-based alignment systems need additional time for translating all the source-language sentences into the target language. Therefore, *FaDA* reduces the computational overhead by skipping the translation process. As *FaDA* performs alignments at the document level, we consider each sentence separately and store it in a single document. Each document in our corpus therefore contains a single line of text. We then use the following functionalities of *FaDA* in our experiment.

(i) **Indexing**: First, we index both the source-language and the target-language documents,

(ii) ***Pseudo-query* construction**: Secondly, we construct a pseudo-query[6] from each source-language document using the terms selection procedure as shown in Equation (1).

$$\tau(t,d) = \lambda \frac{tf(t,d)}{len(d)} + (1-\lambda)\log(\frac{N}{df(t)}) \quad (1)$$

$tf(t,d)$ refers to the term frequency of a term t in a document d. $len(d)$ denotes the length of d, and N and $df(t)$ represents the total number of documents and the number of documents in which t occurs, respectively. $\tau(t,d)$ denotes the term-selection score which is a linear combination of the normalised term frequency of a term t in d, and the inverse document frequency (idf) of the term. The parameter λ controls the relative importance of tf and idf. We recommend the work of Lohar et al. (2016) for more details on *pseudo-query* construction.

(iii) **Word translation**: We then translate all the pseudo-query terms into the target-language with an English–Icelandic dictionary and search the translated query terms in the target-language index,

(iv) **Document retrieval**: Finally, we retrieve the *top-n*[7] target-language documents that are semantically equivalent to the source-language documents according to the IR-based retrieval.

4.2 Sentence Scoring

In the first step, the application of *FaDA* provides 10 (default value) target-language sentence candidates for each source-language sentence. This is

[4] https://spacy.io

[5] https://github.com/thompsonb/vecalign

[6] A *pseudo-query* is the modified version of the original query to improve the ranking of document retrieval. The terms in a pseudo-query are considered to be suitably representative of a document

[7] Note that $n = 10$ is the default value of n in *FaDA*. This means that the tool retrieves the top 10 candidate target-language documents by default.

done in both translation directions. We assume that most truly parallel sentences would be found in either direction and thus we create a subset of the *FaDA* outputs that contains an intersection of the candidate list for both directions. In order to test this hypotheses, we also create a union of both outputs when working with one of the test sets, *CompNews*.

We score our candidate lists using two methods, *LaBSE* (Feng et al., 2020), and *WAScore*, a word alignment-based score of our own device. Feng et al. (2020) show that *LaBSE* gives good results on the BUCC mining task when working with high-resource languages. However, the accuracy is reduced when working with less-resourced languages. In order to increase the accuracy of our extraction method, we use it together with another scoring mechanism that uses a very different approach. *WAScore* is calculated by collecting high precision word alignments using *CombAlign* (Steingrímsson et al., 2021). *CombAlign* uses a set of word alignment tools to perform the alignment and it has settings to aim for high precision or high recall, taking advantage of the fact that different alignment tools tend to make different guesses unless the alignment probabilities are high. We aim for high precision, thus removing most alignments that are not very likely to be correct. As this can be achieved by *CombAlign*, it makes WAScore an effective mechanism for measuring parallelism. *CombAlign* uses the following tools in our experiment; (i) *AWESoME* (Dou and Neubig, 2021), (ii) *eflomal* (Östling and Tiedemann, 2016), and (iii) *fast_align* (Dyer et al., 2013). WAScore is calculated for each sentence using Equation (2):

$$(s_a/s) * (t_a/t) \qquad (2)$$

where s is the number of words in the source sentence and s_a is the number of source words that are aligned to some word in the target sentence, t is the number of words in the target sentence, and t_a is the number of target words that are aligned to some word in the source sentence.

With a set of highly likely alignments for each sentence pair, the WAScore tends to favour sentences of similar length as a much longer sentence on one side usually has proportionately few alignment edges on that side which lowers the score substantially. In contrast, if a shorter sentence on one side has all tokens aligned to a longer sentence on the other side, it can result in a reasonable score. Such pairs are often partially parallel and using the *CompWiki* test set (Section 5.2) we see that our approach is suitable for extracting partially parallel pairs as well as truly parallel ones.

CompNews			
Set	Pr.	Rc.	F_1
Intersection	0.95	0.80	0.87
Union	0.92	0.86	0.86

Table 1: Precision, Recall F_1-measure and number of extracted sentences for a union and intersection of the *FaDA* output.

Finally, we use logistic regression to classify whether a sentence is parallel or not. All sentences accepted by the classifier are labelled as parallel sentences. The classifier is trained on the *CompTrain* training set, detailed earlier in Section 3.3.

5 Evaluation

We evaluate our system by calculating precision, recall and F1-scores using our (i) *CompNews* test set and (ii) *CompWiki* test set; and (iii) by training, testing and calculating BLEU scores for NMT systems, both with and without parallel sentences extracted from all Wikipedia articles that are available in both English and Icelandic.

5.1 Testing on News Data

The first experiment is on the *CompNews* test data, with the simple goal of extracting as many parallel sentence pairs as can be found from the two lists of $100K$ sentences in English and Icelandic. After running *FaDA* we obtain 10 candidates for each of the $100K$ sentences in each language. We create two different candidate sets, one by taking an intersection of both directions, en→is and is→en, and the other by taking a union of the two directions.

The intersection set contains $135K$ sentence pairs and an inspection of the set revealed that it

CompWiki			
Set	Pr.	Rc.	F_1
Parallel	0.39	0.90	0.54
+partially	0.84	0.33	0.47

Table 2: Precision, Recall and F_1-measure as measured when only looking at the sentence pairs marked as parallel in the test data, and when the partially parallel have been added to the desired output.

Wikipedia Training					
Training Data	Supplemental Sentences	TestEEA	TestEMA	TestOS	Combined
ParIce50K	0	9.0	9.0	1.6	8.1
ParIce50K+WikiMatrix	$313,875$	5.6	5.2	2.3	5.1
ParIce50K+Our approach	$55,744$	13.9	15.9	7.0	13.7

Table 3: BLEU scores for MT systems trained on parallel data and sentences extracted from comparable corpora.

included $1,693$ of the total $2,000$ known parallel sentence pairs in the data. The union set on the other hand had a total of 1.86 million pairs and $1,871$ of the $2,000$ correct sentence pairs.

We calculate LaBSE scores and WAScore for each of the candidates and apply our logistic regression classifier on the scores. The F-scores for both approaches were similar, but using the union data set obtains higher recall while using the intersection data obtains better precision. Table 1 shows the final results for the *CompNews* test set.

5.2 Testing on Wiki Data

The preparation of *CompWiki* was described in Section 3.2. It contains texts from 15 Wikipedia article pairs with a total of $10,098$ sentence pairs. We score the sentences in the same way as discussed before, using LaBSE and WAScore, and run our classifier on the scores. 200 sentence pairs are deemed parallel by our classifier. 77 of them are marked parallel in the test set, 90 are marked partially parallel and 33 are marked non-parallel. As can be seen in Table 2, our method achieves high recall on the sentences marked parallel, and 84% of our systems output is either marked parallel or partially parallel.

5.3 Parallel Sentence Extraction and MT Training

We collect all texts from Wikipedia articles that are linked and available both in English and Icelandic. The collection contains 412,442 Icelandic sentences and 4,259,150 English sentences from 35,690 article pairs. In our setup, *FaDA* searches for the parallel candidates in the paired documents. The candidate pairs are then scored as before and classified as parallel or non-parallel. Our system yields 55,744 sentence pairs that are classified as parallel sentences.

There have been previous efforts in extracting parallel sentence from the Wikipedia corpus. One of the largest such efforts is the WikiMatrix project (Schwenk et al., 2021) that mined parallel sentences in $1,620$ language pairs. When we compare the en–is language pair in WikiMatrix to the output of our system, the first obvious difference is that the WikiMatrix dataset has a lot more data, $314K$ sentence pairs compared to our $56K$. To compare the usefulness of the datasets, we trained an NMT system using Marian MT (Junczys-Dowmunt et al., 2018) in one direction, is→en, on $50K$ sentence pairs randomly sampled from the ParIce corpus and compared it to a system where WikiMatrix was added as supplemental data, and to a system where the results of our approach was used to supplement the ParIce data, using the same hyperparameters.

We compare BLEU scores for the different setups on a combination of three test sets (Barkarson and Steingrímsson, 2020), as well as on each of the test sets individually: TestEEA - containing sentence pairs from European Economic Area regulatory documents; TestEMA - containing sentence pairs from EMA drug descriptions; and TestOS - containing sentence pairs from OpenSubtitles. TestEEA and TestEMA are extracted from rather specialized texts, and generally have long sentences, while TestOS is from a rather open domain and tends to have shorter sentences. The test sets are used as filtered by Jónsson et al. (2020). All the sentence pairs in the test sets have been manually checked for correctness.

The fact that each of these three test sets are domain specific and that our NMT systems are not trained specifically on data from these domains, together with how small the training data sets are, results in low BLEU scores. But while the BLEU scores are quite low, the effect of our approach is evident.

We can see from Table 3 that when the WikiMatrix data is added to the $50K$ parallel sentences, the translation system trained on this augmented data set produces significantly lower BLEU scores as compared to the other two systems for the two test sets (TestEEA and TestEMA). However, it obtains higher BLEU scores than the baseline system (i.e, the system which is trained with only the $50K$ data) for the third test set (TestOS). In contrast, the system trained on the concatenation of the $50K$ sentence pairs and the data obtained from our approach significantly improves the BLEU scores for all the test sets, even though the number of sentence pairs in our data is less than 20% of the number of sentence pairs in WikiMatrix. This is most likely due to noise in WikiMatrix, as it has been shown that NMT is sensitive to noise in the training data (Khayrallah and Koehn, 2018).

Upon manual inspection of our data we see that our classifier accepted some sentence pairs even though they have a very low WAScore. We therefore train a number of NMT models using our data but apply thresholds for WAScore. As seen in Figure 2, the BLEU score rises when a low threshold is set, and then fluctuates when the threshold is raised, reaching the highest BLEU score for our combined test sets at a WAScore threshold of around 0.14. A WAScore of 0.14 means that if we have a pair of sentences containing ten tokens each, three tokens in one sentences align with four tokens in the other. If there are fewer alignments the sentence pair will not be accepted. At this threshold level we extract $34K$ parallel pairs to use for training. With further threshold filtering, we lose more beneficial data than detrimental data, and the BLEU score starts slipping down. This is an indicator of the usefulness of this scoring mechanism for MT training, showing that the score correlates with sentence pair parallelism, raising the BLEU score when it is used for filtering, and keeping it raised even though supplemental training data is reduced.

All of our data sets, for training and testing are available on Github, as well as a description of MarianMT training setup[8].

6 Conclusions and Future work

We have shown that our method, combining cross-lingual information extraction, contextualized embeddings and word alignments, is efficient at finding parallel segments in comparable corpora. Furthermore we introduce WAScore, a metric of translational equivalence based on high-precision word alignments, and show that as well as being a useful part of a binary classifier, it can be used effectively to filter out detrimental segments from parallel corpora. Finally, we publish two new test sets for extracting parallel sentences from comparable corpora, an automatically generated English–Icelandic lexicon with probability scores and a set of automatically extracted parallel segments that we show are useful for training MT systems.

When testing on the *CompWiki* test set we saw that while our method is efficient in finding parallel segments in comparable corpora, it also selects partially parallel segments. Although these segments seem to have information useful for training MT

[8] https://github.com/steinst/bucc2021-en-is

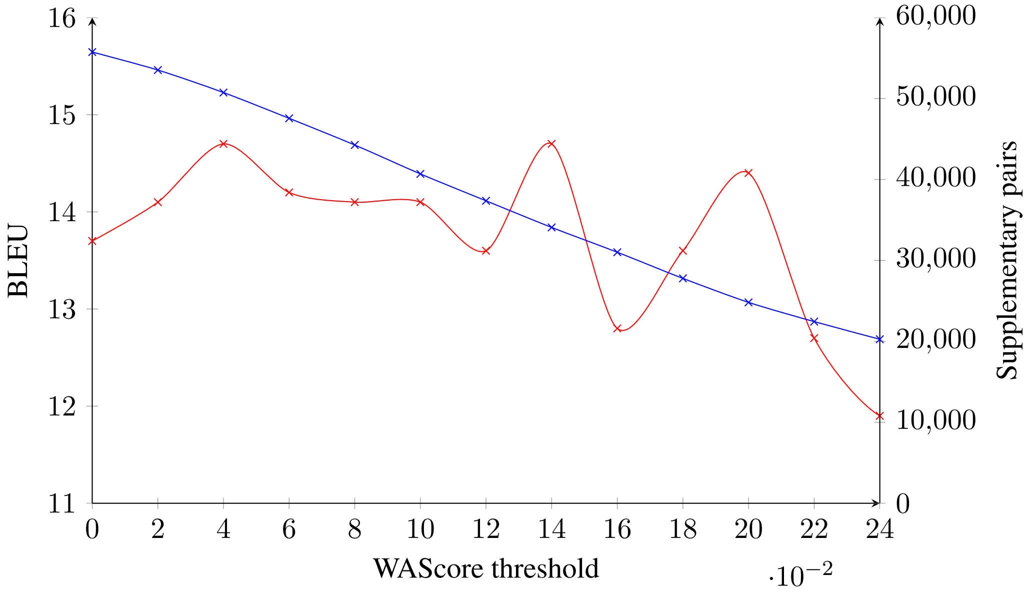

Figure 2: BLEU score for MarianMT models training with supplementary data, with different WAScore thresholds over the combined test sets.

systems, it is difficult to know to what extent they are useful and when they may become detrimental. For this reason, we plan to study these kinds of data further and investigate how they affect translation quality of an NMT system trained on it. Based on that, we want to explore more sophisticated ways to segment or concatenate alignment candidates in order to be able to build a data set that only contains segment pairs that are useful for training MT systems. There is previous work on parallel fragment extraction using word alignments (Yeong et al., 2019), and we will use their approach as a baseline to proceed further.

While the combination of the two scores used to measure the quality of the sentence pairs resulted in a list of sentence pairs that we show are useful for MT training, it still contains pairs that are detrimental, as shown by the simple filtering based on WAScore threshold. Other parallel sentence pairs may also remain to be found in the Wikipedia data. In order to improve our approach, more scores could be added to our classifier. While we opted to use raw LaBSE cosine similarity scores, shown by (Feng et al., 2020) to be more accurate than cosine similarity scores from other models, the margin-based ratio score proposed by Artetxe and Schwenk (2019a) has also been shown to be very effective for this task. Other scores to consider could include BLEU or ChrF (Popović, 2015), although they need reasonably good MT systems to be useful, margin-based cosine distance (Artetxe and Schwenk, 2019a), or Mahalanobis distance (Mahalanobis, 1936) as described in Littell et al. (2018). Doing an ablation study on the scores could help determine which are the most useful. Working with these scores, a comparison of applying different classifiers while using the same scoring mechanisms may be helpful. It is also to be noted that we extracted only 10 target-language candidate pairs in the first step, which is the default value used in *FaDA* as it gave optimal performance in their work. It also has the benefit of reducing the computational complexity in the next steps. However, we also plan to explore other higher values of candidate extraction in future and to investigate how it affects the overall system performance. Finally, we plan to conduct our experiments on other language pairs.

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

Haithem Afli, Loïc Barrault, and Holger Schwenk. 2015. Building and using multimodal comparable corpora for machine translation. *Natural Language Engineering*, 22(4):603 – 625.

Thierry Etchegoyhen Andoni Azpeitia and Eva Martínez Garcia. 2018. Extracting parallel sentences from comparable corpora with stacc variants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised Statistical Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, 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*, pages 3197–3203, 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.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45, Vancouver, Canada.

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*, pages 140–145, Turku, Finland.

Starkaður Barkarson and Steinþór Steingrímsson. 2020. ParIce dev/test/train splits 20.05. CLARIN-IS.

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*, pages 146–154, Turku, Finland.

Martha Dís Brandt, Hrafn Loftsson, Hlynur Sigurþórsson, and Francis M. Tyers. 2011. Apertium-IceNLP: A rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*, pages 217–224, Leuven, Belgium.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese–Japanese Wikipedia. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(2).

Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online.

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.

Fangxiaoyu Feng, Yin-Fei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *ArXiv*, abs/2007.01852.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised Parallel Sentence Extraction with Parallel Segment Detection Helps Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy.

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*, pages 310–315, Turku, Finland.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with different machine translation models in medium-resource settings. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 95–103.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective High-Quality Neural Machine Translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135, Melbourne, Australia.

Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2018. Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3477–3482, Miyazaki, Japan.

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, Heidelberg. 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*, pages 74–83, Melbourne, Australia.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and Why is Unsupervised Neural Machine Translation Useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal.

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*, pages 5039–5049, Brussels, Belgium.

Wang Ling, Luís Marujo, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2014. Crowdsourcing High-Quality Parallel Data Extraction from Twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 426–436, Baltimore, Maryland, USA.

Patrick Littell, Samuel Larkin, Darlene Stewart, Michel Simard, Cyril Goutte, and Chi-kiu Lo. 2018. Measuring sentence parallelism using mahalanobis distances: The NRC unsupervised submissions to the WMT18 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 900–907, Belgium, Brussels.

Pintu Lohar, Debasis Ganguly, Haithem Afli, Andy Way, and Gareth J. F. Jones. 2016. Fada: Fast document aligner using word embedding. *The Prague Bulletin of Mathematical Linguistics*, 106:169–179.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba Submission to the WMT20 Parallel Corpus Filtering Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 979–984, Online.

Prasanta Chandra Mahalanobis. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, AK Raghavan, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, J. Mahalakshmi, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and M. Khapra. 2021. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *ArXiv*, abs/2104.05596.

Holger Schwenk. 2018. Filtering and Mining Parallel Data in a Joint Multilingual Space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4361–4366, Miyazaki, Japan.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2021. CombAlign: a Tool for Obtaining High-Quality Word Alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Reykjavik, Iceland (Online).

Sara Stymne, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Tunable Distortion Limits and Corpus Cleaning for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 225–231, Sofia, Bulgaria.

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)*, pages 1342–1348, Hong Kong, China.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal.

Þórdís Úlfarsdóttir. 2014. ISLEX — a Multilingual Web Dictionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2820–2825, Reykjavik, Iceland.

Krzysztof Wolk, Emilia Rejmund, and Krzysztof Marasek. 2016. Multi-domain machine translation enhancements by parallel data extraction from comparable corpora. In Ewa Gruszczyńska and Agnieszka Leńko-Szymańska, editors, *Polish-Language Parallel Corpora*, pages 157–179. Instytut Lingwistyki Stosowanej, Warsaw, Poland.

Yin-Lai Yeong, Tien-Ping Tan, and Keng Hoon Gan. 2019. A Hybrid of Sentence-Level Approach and Fragment-Level Approach of Parallel Text Extraction from Comparable Text. *Procedia Computer Science*, 161:406–414.

Ieva Zariņa, Pēteris Ņikiforovs, and Raivis Skadiņš. 2015. Word Alignment Based Parallel Corpora Evaluation and Cleaning Using Machine Learning Techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 185–192, Antalya, Turkey.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2016. Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora. In *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*, pages 38–43, Portorož, Slovenia. ELDA.