A Sentence Alignment Approach to Document Alignment and Multi-faceted Filtering for Curating Parallel Sentence Pairs from Web-crawled Data

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavik, Iceland
steinthor.steingrimsson@arnastofnun.is

Abstract

This paper describes the AST submission to the WMT23 Shared Task on Parallel Data Curation. We experiment with two approaches for curating data from the provided web-scraped texts. We use sentence alignment to identify document alignments in the data and extract parallel sentence pairs from the aligned documents. All other sentences, not aligned in that step, are paired based on cosine similarity before we apply various different filters. For filtering, we use language detection, fluency classification, word alignments, cosine distance as calculated by multilingual sentence embedding models, and Bicleaner AI. Our best model outperforms the baseline by 1.9 BLEU points on average over the four provided evaluation sets.

1 Introduction

The aim of the Shared Task on Parallel Data Curation at the Eighth Conference on Machine Translation (WMT23) is to evaluate parallel data curation methods (Sloto et al., 2023). The goal is to find the best machine translation (MT) training data within a provided pile of web-crawled data.

The language pair chosen for the task is Estonian-Lithuanian. The provided data is extracted from a single snapshot of CommonCrawl,[1] which according to the task organizers should contain enough training data to train a reasonable Estonian → Lithuanian MT model, even with limited compute. As well as providing the data, the organizers release pre-computed intermediate steps from a baseline, so participants can choose whether to focus on one or more aspects of the task. We describe the provided data and the baseline in Section 3.

In our submission we experiment on two aspects of parallel data curation. Initially we try to to identify parallel documents in the two languages. We then align sentences in the documents using our own sentence alignment tool, SentAlign[2] (Steingrímsson, 2023; Steingrímsson et al., 2023b), and train an MT system on the resulting sentence pairs. SentAlign is a sentence aligner that uses LaBSE (Feng et al., 2022) to score all possible alignment combinations for a document pair, selects the highest scoring one, but then re-evaluates the results by looking at each individual alignment and their closest neighbours to see if localized scores can be raised. This is to counteract an effect of dynamic programming with cosine similarity, which often favours many-to-many alignments over 1-to-1 alignments (see e.g. Thompson and Koehn (2019). Steingrímsson et al. (2023b) show that this approach outperforms other aligners on two evaluation sets, as well as on a downstream task. The other aligners include aligners such as the length based Gale-Church (Gale and Church, 1991), MT-based Bleualign (Sennrich and Volk, 2010) and Vecalign (Thompson and Koehn, 2019) which is the most similar to SimAlign, using LASER embeddings (Artetxe and Schwenk, 2019b) to calculate cosine similarity of alignment candidates, and a recursive approximation to reduce the search space, as opposed to evaluating all possibilities as SentAlign does. We describe our approach to document alignment in Section 4.1. Subsequently, we try to identify parallel sentence pairs in all the other provided data and run a number of different filters to remove sentence pair candidates that we deem likely to be detrimental or useless for MT training. Our filtering approaches are described in Section 4.2

2 Related Work

Khayrallah and Koehn (2018) show that incorrect translations, untranslated target text, misalignments, and other noisy segments in a parallel corpus have a detrimental effect on the output quality of neural machine translation (NMT) systems trained on that corpus, as measured by using BLEU (Papineni et al., 2002). They specify five general

[1] https://commoncrawl.org/
[2] https://github.com/steinst/SentAlign

classes of noise commonly found in a German-English version of the ParaCrawl corpus: misaligned sentences, disfluent text, wrong language, short segments, and untranslated sentences. They find this distinction to be useful to give a general idea of which types of errors seem to have the least impact on MT systems (short segments, untranslated source sentences and wrong source language) and which have the most dramatic effect (untranslated target sentence). Misalignments, misordered words, and wrong language, in source or target texts, are also shown to be harmful, but not as harmful.

The Conference on Machine Translation, WMT, hosted three annual shared tasks on parallel corpus filtering (Koehn et al., 2018, 2019, 2020), focusing on filtering noisy web-crawled corpora. Submitted systems include the ones by Chaudhary et al. (2019) and Artetxe and Schwenk (2019a), who introduce approaches based on cross-lingual sentence embeddings trained from parallel sentences. Both papers use cosine similarity and consider the margin between a given sentence pair and its closest candidates to normalize the similarity scores.

Bicleaner (Sánchez-Cartagena et al., 2018; Ramírez-Sánchez et al., 2020) uses a set of handcrafted 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. It also scores sentences based on fluency using 5-gram language models. The tool ranked highly on the first two parallel corpus filtering tasks at WMT. Bicleaner AI (Zaragoza-Bernabeu et al., 2022) is a fork of Bicleaner using a neural classifier. It has been shown to give significant improvements in translation quality as measured by BLEU when used for filtering training data for multiple language pairs, as compared to the previous version of the tool.

In contrast to tools that apply a single method for parallel corpus filtering, Aulamo et al. (2020) implement multiple different filters in the OpusFilter toolbox. These include heuristic based filters, language identification, character-based language models and word alignment tools. The toolbox can furthermore be extended with custom filters.

Herold et al. (2022) revisit the noise classes specified by Khayrallah and Koehn (2018) to investigate how accurately two strong filtering approaches, cross entropy (Rossenbach et al., 2018) and LASER (Artetxe and Schwenk, 2019b) can filter out different noise classes. They find that for a common language pair, German→English, most types of noise can be detected with over 90% accuracy, although misalignments and poor synthetic translation can only be detected with an accuracy of less than 70%. For a less common language pair, Khmer–English, the performance of the filtering system degraded significantly and the accuracy of identifying noise was lowered by 8–19%, depending on the type of noise, indicating that results can vary dramatically depending on the languages.

3 Data and Baseline

The provided data is retrieved from the 2023-06 snapshot of Common Crawl. The organizers have extracted plain text from HTML documents and used the Fasttext (Joulin et al., 2017) language identification model to remove documents not classified as Estonian or Lithuanian by the model, based on the first 2,000 characters of the document. Unsafe and offensive content has been removed. Documents from host names in the following lists in the blocklist project[3] where removed: *abuse, basic, crypto, drugs, fraud, gambling, malware, phishing, piracy, porn, ransomware, redirect, scam, torrent*. Documents were split into paragraphs based on line breaks, and then into sentences using Mediacloud Sentence Splitter.[4] Each sentence was assigned a unique sentence id and classified using the Fasttext language identification model. The data is provided in TSV format.

The task organizers provide LASER2 sentence embeddings (Heffernan et al., 2022) for all sentences in the correct language, as classified by the Fasttext model. They index the embeddings and query each index to retrieve the top-8 results for each sentence, based on cosine similarity. We use these results as a starting point for our filtering approaches, as described in Section 4.2. The baseline simply takes top-1, i.e. the highest scoring sentence pair for each sentence, provided the score exceeds a threshold of 0.9. This results in a set of 2,654,090 parallel pairs for training a baseline model.

A script for training the baseline model using Sockeye (Hieber et al., 2022) was provided. We use the script and Sockeye for training all models on a single Nvidia GeForce RTX 3090 GPU card.

[3] https://github.com/blocklistproject/Lists

[4] https://github.com/mediacloud/sentence-splitter

4 System Architecture

In this section we describe our approaches to the parallel data curation problem. First, we try to identify parallel documents in the two languages and align them on the sentence level. Second, we use the provided sentence pair candidates, eight for each sentence in each language, and filter using a number of different approaches to remove possibly detrimental pairs from our training set. The sentence pairs from the aligned documents and the filtered sentence pairs are combined to compile our final dataset.

4.1 Document identification by Sentence Alignment

Bilingual document alignment is a matching task that considers documents in two languages and estimates the likelihood of the documents being a translation of each other. In the Bilingual Document Alignment Shared Task at WMT 2016 (Buck and Koehn, 2016), the submitted systems used a variety of approaches. Some of these include Gomes and Pereira Lopes (2016), who used a phrase table from a phrase-based statistical machine translation (SMT) system to compute coverage scores. Dara and Lin (2016) use MT to find corresponding documents based on n-gram matches, assisted by document length ratio, and Mahata et al. (2016) use text matching based on sentence alignment and word dictionaries. Thompson and Koehn (2020) present a document alignment method that uses information on sentence order both when generating candidates and when re-scoring the candidates. For re-scoring the candidate pairs they perform sentence alignment and score the alignment based on semantic similarity of the resulting sentence pairs.

In this paper, we use sentence alignment and average cosine distance as measured by LaBSE (Feng et al., 2022) to determine whether documents can be aligned. The provided dataset contains sentences scraped from the web, information on the web domain and an order of sentences within documents on the websites. We recreate documents, most likely to have a corresponding translation in the other language, using this information. In order to reduce the need for compute we only consider texts from the same domain to be possible candidates for document alignment.

Our approach is the following:

1. We start by collecting a list of all web domains common to both languages.
2. From these domains, we recreate all documents that contain more than five sentences. The recreated documents have one sentence in each line.
3. Using SentAlign, for each domain we align the recreated documents in Estonian to all the recreated documents in Lithuanian, and vice versa. SentAlign outputs all aligned sentence pairs, as well as the LaBSE score for the pair.
4. If more than half of the sentences in either language does not get an alignment, the document pair is discarded.
5. If the average LaBSE score for all sentence alignments for a given document pair is below a threshold of 0.6, the document pair is discarded.
6. We calculate an alignment ratio using Equation 1:

$$\frac{1}{2}\left(\frac{et_{aligned}}{et_{total}} + \frac{lt_{aligned}}{lt_{total}}\right) \quad (1)$$

   Where $et_{aligned}$ is the number of Estonian sentences that obtain an alignment to a Lithuanian sentence, and et_{total} is the total number of Estonian sentences in the document. $lt_{aligned}$ and lt_{total} are the corresponding numbers for Lithuanian.

   From a pool of documents for each web domain, a greedy algorithm selects the document pair with the highest alignment ratio, and if multiple pairs have the highest ratio, one of those with the highest average LaBSE score. The selected documents are then removed from the pool and the process repeated until all acceptable pairs have been collected for that domain.

The sentence alignment approach to identifying aligned documents in (Thompson and Koehn, 2020) uses Vecalign (Thompson and Koehn, 2019) and LASER embeddings to perform sentence alignment and judge sentence similarity. While we use a different aligner and embeddings, our approach follows the same general strategy, with the main difference being that language identification is part of their scoring function while we simply require over half the sentences in each document to obtain an alignment. We can do this as the provided data set we work with has been selected based on

language identification, so we can assume the sentences we work with are generally in the correct language.

Our process results in 4,372 document pairs, containing 160,787 sentence pairs after deduplication. We remove all sentence pairs that have less than three tokens in either language, disregarding all numbers and other non-alphabetical symbols. Furthermore, we remove all sentence pairs that obtain a LaBSE score lower than 0.4. While we do not have any statistics on what the ideal LaBSE threshold should be for this language pair, Steingrímsson et al. (2023a) show that for Icelandic-English over half the sentence pairs are acceptable when the LaBSE score exceeds 0.4, and we base our threshold on that. Our approach results in a set of 120,756 sentence pairs obtained from parallel documents, with 114,301 of those used for training after we have removed sentences that may overlap with test and development datasets.

4.2 Filtering Sentence Pair Candidates

Having extracted sentence pairs from aligned documents, we have yet to consider most of the data in the provided dataset. We experiment with various filtering filtering approaches and as a starting point we simply use the sentence pair candidates provided by the tasks organizers, eight Lithuanian sentences for each Estonian sentence and eight Estonian sentences for each Lithuanian sentence, as described in Section 3. To extract the best sentence pairs, we apply a number of diverse filtering approaches to these sentence pair candidates.

We start by filtering the sets of Estonian and Lithuanian sentences separately:

1. To start with, we have 142,516,521 sentences in Estonian and 210,914,146 sentences in Lithuanian. We deduplicate these sets, giving us 53,228,455 Estonian sentences and 63,536,939 Lithuanian sentences.

2. Although the Fasttext language detection model has been applied to the data, it still contains sentences that are in different languages. In order to remove these we run two additional language detection tools, lingua[5] and langdetect (Shuyo, 2010). From both of these tools we acquire a language classification for each sentence. We then remove all sentences that do not obtain the expected classification by at least two of the three classifiers that have been applied. This leaves us with 33,500,758 Estonian sentences and 43,173,412 lithuanian sentences.

3. In order to remove sentences that may be disfluent we use two pre-trained GPT-2 (Radford et al., 2019) models, one for each language,[6] to classify the sentences. For that we use the approach described in (Steingrímsson et al., 2023a): We collect two sets of sentences for each language, one containing sentences that are presumably fluent and the other one containing sentences that are likely to be disfluent. To train the classifiers, we selected 15,000 sentences randomly for each language from the Leipzig Corpora Collection (Goldhahn et al., 2012) for the fluent examples and 15,000 random sentences from the provided data we had already discarded in the previous step. The classifier uses the GPT-2 model to calculate perplexity for the sentences, and chooses potential thresholds as the average between two adjacent perplexity values. It then uses a maximization function to decide on a threshold that yields the most accurate prediction based on the training set. After classifying the remaining sentences, and removing the approximately 120 thousand sentences included in the document alignment data previously acquired, we are left with 31,298,451 Estonian sentences and 29,498,886 Lithuanian sentences.

Next, we consider the provided sentence pair candidates as calculated using LASER2. We have two candidate lists, one with eight candidates for each Estonian sentence and another with eight candidates for each Lithuanian sentence. We remove all pairs containing sentences not in our filtered sentence lists. We then take an intersection of the resulting sets. The intersection thus only contains sentence pairs where the Lithuanian sentence is one of the top 8 candidates for the Estonian sentence, and vice versa. This gives us a list of 36,250,870 sentence pairs, 35,720,955 after we have excluded all pairs containing sentences that overlap with sentences in the evaluation or development data sets. It should be noted that at this stage some sentences

[5] https://pemistahl.github.io/lingua-py

[6] Lithuanian model: https://huggingface.co/DeividasM/gpt2_lithuanian_small; Estonian model: https://huggingface.co/tartuNLP/gpt-4-est-base

are found in multiple sentence pairs. We proceed to filter this set of sentence pairs:

4. For each Estonian sentence we select only the Lithuanian sentence that gives the highest LASER2 score, and for each Lithuanian sentence we likewise select only the Estonian sentence with the highest score. This reduces the candidate list to 24,735,722 sentence pairs.

5. The sentences comprising the pairs are tokenized. We then run fast-align (Dyer et al., 2013) to obtain word alignments for each sentence pair. These word alignments are used to calculate a word alignment score, WAScore, a word alignment-based score devised to measure word-level parallelism, introduced in Steingrímsson et al. (2021). Steingrímsson et al. (2023a) show that when WAScore is low, very few sentences are good mutual translations. We remove all sentence pairs that have a WAScore lower than 0.15, indicating that 40% or fewer tokens in either sentence obtained an alignment on average. After that our candidate list contains 21,387,140 sentence pairs.

6. We calculate a LaBSE score for all the pairs. If the LaBSE score is higher than 0.9, we accept the sentence pair for our final training set without further processing. These are 891,313 sentence pairs. We also set a minimum threshold of 0.6, as suggested by Feng et al. (2022). This gives us 13,289,869 sentence pairs to processed further, and the rest is discarded.

7. Next, we train Bicleaner AI (Zaragoza-Bernabeu et al., 2022) to classify the Estonian-Lithuanian language pair. For training Bicleaner we need monolingual corpora and parallel corpora. For monolingual data we collected 5 million sentences in each language from the Leipzig Corpora Collection and used 100 thousand parallel pairs randomly selected from the set of sentence pairs extracted from the document alignment step described in Section 4.1. Our Bicleaner AI model gives low scores and we accept sentence pairs with scores over the threshold of 0.05. We run the model on all unfiltered sentences, removing over 20 million and leaving us with 14,988,586 sentence pairs, as shown in Table 1.[7] We later take an intersection of this set and the set obtained by applying other filters, as shown in Table 2.

8. Finally, we use the LASER2 scores, LaBSE scores, WAScore and NMTScore (Vamvas and Sennrich, 2022) with a classifier to predict whether a sentence pair contains a mutual translation. NMTScore is based on translation cross-likelihood, the likelihood that a translation of segment *A* into some language, could also be a translation of segment *B* into the same language. We used OPUS-MT models to translate the segments. We use a logistic regression (Cox, 1958) classifier trained on the same data as the GPT-2 classifiers described above. The classifier accepts as valid mutual translations, 2,967,348 sentence pairs out of the 13,289,869 marked for further processing in (6). When these are added to the set of previously accepted sentences from the aligned documents and the ones having very high LaBSE scores, we have 3,902,740 in our final training set, before applying the Bicleaner AI filter, as shown in Table 2.

5 Results

In addition to the baseline models described in Section 3, we trained eleven MT models using data sets at different stages of the compilation process and evaluated on the provided test sets, using BLEU[8] and chrF[9]. Table 1 shows the results after each filtering step until the logistic regression filter, and Table 2 shows the final sets after filtering and an ablation study on the effects of combining the sets acquired using different approaches. Our best model (**K**) was trained on a combination of sentence pairs from the aligned document pairs (**G**), sentence pairs with a LaBSE score over 0.9 (**H**) and the sentence pairs accepted by our logistic regression filter (**I**).[10]

[7] Our model is available at Github: `https://github.com/steinst/BicleanerAI-models`

[8] Sacrebleu signature: BLEU+nrefs.1+case.mixed+eff.no+tok.3a+smooth.exp+version.2.3.1

[9] Sacrebleu signature: chrF2+nrefs.1+case.mixed+eff.yes+nc.6+nw.0+space:no+version.2.3.1

[10] We submitted dataset *L* to the shared task, which has somewhat lower scores than dataset *K* and was the dataset that was used to train our second best model. This was due to an error in our training script used for selecting a dataset to submit. The script did not remove sentences overlapping with evaluation data, giving us incorrect results. This error has been rectified in all results given in this paper and when we talk about our best model we are always referring to the model trained on dataset *K*.

Data Filters	No. sent.	Bleu				ChrF			
		EMEA	EUB	EP	JRC	EMEA	EUB	EP	JRC
A. Unfiltered	35,720,955	16.2	14.8	15.1	18.2	45.0	43.3	45.9	45.8
B. A ∩ Bicleaner AI	14,988,586	18.7	17.4	17.3	21.8	49.2	48.0	49.5	50.2
C. A ∩ Best LASER2	24,735,722	15.1	15.1	14.7	18.2	45.9	45.3	46.3	48.3
D. C ∩ WAScore filter	21,387,140	19.4	18.9	17.3	23.9	49.3	48.7	49.0	52.0
E. D ∩ LaBSE > 0.6	13,958,582	19.9	19.0	18.3	23.3	50.3	50.6	50.3	52.4
F. B ∩ E	7,193,830	20.5	19.4	18.5	24.2	51.2	51.6	51.4	53.5

Table 1: Scores for the models trained on datasets compiled by applying different filters. We evaluate on the four provided test sets, with data from EMEA, EUBookshop (EUB), Europarl (EP) and JRC-Acquis. The table shows the number of sentences, BLEU and ChrF scores after different filters have been applied.

Data Filters	No. sent.	Bleu				ChrF			
		EMEA	EUB	EP	JRC	EMEA	EUB	EP	JRC
Baseline	2,654,090	18.2	19.1	17.8	24.3	49.5	52.2	51.5	54.8
G. Aligned Docs	114,301	8.0	10.9	9.3	16.2	33.8	41.6	40.3	44.5
H. LaBSE > 0.9	868,039	18.9	17.2	16.3	22.6	50.1	50.3	49.9	52.6
I. Logistic Regression	2,925,549	15.4	14.1	13.7	18.2	45.7	46.2	46.5	48.0
J. H ∪ I	3,788,511	20.2	19.5	18.3	24.8	51.2	52.1	51.7	54.4
K. G ∪ H ∪ I	3,902,740	20.4	20.7	19.1	26.6	51.4	53.3	52.2	56.1
L. K ∩ B	2,684,931	20.4	19.7	18.4	25.1	51.4	52.5	51.8	54.9

Table 2: Datasets created using different approaches and an ablation study for investigating the effect of each dataset on MT quality as measured by BLEU and ChrF. The logistic regression dataset is created by applying our logistic regression classifier on dataset E in Table 1. We evaluate on the four provided test sets.

Our best model outperforms the baseline by approximately 1.9 BLEU on average. We find that the sentence pairs from the aligned documents, only 114,301 pairs, improve the BLEU on average by 1.0 BLEU. This indicates that these sentence pairs are useful and that identifying document alignments in web-scraped data is worth the effort. We also find that the sentence pairs having high LaBSE scores, over 0.9, give much better results on their own than over three times more sentence pairs with LaBSE scores in the range 0.6 to 0.9, even though they have been filtered further using additional methods. As shown in Table 2, combining these two sets raises the scores substantially. Furthermore, while the Bicleaner AI model we trained seemed to give decent results in earlier stages, using it to filter the dataset we acquired using other approaches actually decreased the scores. This indicates that the Bicleaner AI model is rejecting too many useful sentence pairs. It could be useful to try to investigate further which of these rejected sentences are useful for MT training and which are truly detrimental, but we leave that for future work.

6 Conclusions and Future Work

Our alignment and filtering approach resulted in an improvement over the baseline in terms of BLEU score for the four evaluation sets. We identified 4,372 document pairs in the provided dataset, which we aligned on sentence level and used the resulting data set for training. We then combined a number of filtering approaches for determining which sentence pair candidates from a provided candidate list would be likely to be useful, these included an ensemble approach for language detection, using three different tools, a GPT-2 based classifier to determine whether sentences are fluent or disfluent, a logistic regression classifier based on word alignment scores and two sentence embedding based scores, LaBSE and LASER2, and finally a Bicleaner AI classifier.

Working in a similar vein, many different paths could be taken for future work on this problem. Steingrímsson et al. (2023a) show that it can be beneficial to inspect how different filters suit a given translation direction. A filtering method giving an optimal results for $\text{lang}_a \rightarrow \text{lang}_b$ is not necessarily the optimal filtering approach for $\text{lang}_b \rightarrow \text{lang}_a$.

In this work we did not try to evaluate the filtering approaches with regards to translation direction. For translating only from Estonian to Lithuanian, removing incoherent and ungrammatical Estonian sentences may not be as important as removing such sentences in Lithuanian, as it is more important that the target language data contains coherent and well written examples. Different levels of filtering for the different languages could thus be useful in order to add more useful examples.

The aim of our filters is to remove sentences likely to be detrimental in MT training. While we do know about some of the qualities that reduce translation quality, as discussed in Section 2, more fine-grained classifications may be useful. For example, we could designate different levels of misalignments, which include partial alignments defined as sentence pairs where a part of one or both sentences is not represented in the other sentence. Steingrímsson et al. (2023c) argue that extracting mutual translations from such pairs, while discarding the extraneous data, may improve the quality of MT models trained on the data, and show that for one parallel corpus. If that holds in general, it could be useful when working with web-scraped data to identify when misalignments become detrimental and when they can be useful, as well as helping to come up with effective ways to refine such sentence pairs.

Table 2 shows that the datasets compiled from the aligned documents and the one comprising sentence pairs with very high LaBSE scores are very useful as additional training data. We presume that this is an indication of these sets containing higher-quality data. While not suitable for the shared task, it would be an interesting experiment to use a curriculum learning approach for training models on web-scraped corpora such as the one we are using by training a model first on a large set of possibly useful sentences and then fine-tuning the model on the higher-quality data.

Finally, it should be noted that the training times for these models varied considerably. While our best model reached the optimal checkpoint in approximately 20 hours and the second best in 12 hours, the models trained on the larger datasets listed in Table 1 took between 50 and 80 hours of training, using the same settings, while still resulting in lower quality models. It shows that careful curation of training data for MT is not only important for improving model quality in terms of better translations, it also allows for much faster training resulting in a lower carbon footprint.

References

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.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A Configurable Parallel Corpus Filtering Toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany.

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)*, pages 261–266, Florence, Italy.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Aswarth Abhilash Dara and Yiu-Chang Lin. 2016. YODA system for WMT16 shared task: Bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 679–684, Berlin, Germany.

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, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland.

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.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702, Berlin, Germany.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *CoRR*, abs/2205.12654.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting Various Types of Noise for Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain.

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.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels.

Sainik Mahata, Dipankar Das, and Santanu Pal. 2016. WMT2016: A hybrid approach to bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 724–727, Berlin, Germany.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and Bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal.

Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney. 2018. The RWTH Aachen University Filtering System for the WMT 2018 Parallel Corpus Filtering Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 946–954, Belgium, Brussels.

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*, pages 955–962, Belgium, Brussels.

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.

Nakatani Shuyo. 2010. Language detection library for java.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2021. Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. In *Proceedings of the*

14th Workshop on Building and Using Comparable Corpora (BUCC 2021), pages 8–17, Online (Virtual Mode).

Steinþór Steingrímsson. 2023. *Effectively compiling parallel corpora for machine translation in resource-scarce conditions*. Ph.D. thesis, Reykjavik University.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023a. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023b. Sentalign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2023c. Do not discard – extracting useful fragments from low-quality parallel data to improve machine translation. In *Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation*, pages 1–13, Macau, 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.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, Abu Dhabi, United Arab Emirates.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner Goes Neural. In *Proceedings of the Language Resources and Evaluation Conference*, pages 824–831, Marseille, France.