Findings of the WMT 2023 Shared Task on Parallel Data Curation

Steve Sloto
Microsoft
ssloto@microsoft.com

Brian Thompson
AWS AI Labs
brianjt@amazon.com

Huda Khayrallah
Microsoft
hkhayrallah@microsoft.com

Tobias Domhan
Amazon
domhant@amazon.de

Thamme Gowda
Microsoft
thammegowda@microsoft.com

Philipp Koehn
Johns Hopkins University
phi@jhu.edu

Abstract

Building upon prior WMT shared tasks in document alignment and sentence filtering, we posed the open-ended shared task of finding the best subset of possible training data from a collection of Estonian-Lithuanian web data. Participants could focus on any portion of the end-to-end data curation pipeline, including alignment and filtering. We evaluated results based on downstream machine translation quality. We release processed Common Crawl data, along with various intermediate states from a strong baseline system, which we believe will enable future research on this topic.

1 Introduction

A machine translation (MT) system is only as good as the data it is trained on. However, the academic research community often overlooks the details of this task, using pre-curated corpora.

To promote research in this area, this shared task[1] focuses on finding pairs of sentences or documents that are translations of each other based on a collection of web crawled data. MT models are trained by the organizers on the data found by participants, and performance is then judged using automatic metrics. This shared task builds on prior shared tasks on document alignment (Buck and Koehn, 2016a) and sentence filtering (Koehn et al., 2018, 2019, 2020). However, this task is intentionally open-ended, and designed to allow participants to improve on various different parts of the data curation pipeline.

We chose the Estonian-Lithuanian language pair for several reasons. The amount of data we extracted in that language pair was enough to train a reasonable MT model, while being small enough that the task was still accessible to academic participants with limited hardware resources. We avoided English, as many toolkits are developed/optimized on English data, and results on English may not generalize well. And finally, we avoided languages which were closely related, as this could favor methods which do not generalize well.

To lower the barrier to entry and allow participants to focus their research and compute resources, we release intermediate stages of a strong baseline data curation system. We encourage future work to build upon resources provided in this shared task.

This paper gives an overview of the task, presents its results, and provides some analysis.

2 Related work

Parallel data has been required for training machine translation systems ever since the field transitioned to statistical machine translation (Brown et al., 1990). To train that first statistical system, Brown et al. aligned English-French sentences from the proceedings of the Canadian Parliament, often referred to as Hansards, using a very simple system to segment each side into sentences and then align them using only sentence length (Brown et al., 1991). The field of parallel data curation has come a long way since then, with modern methods extracting billions of sentence pairs in hundreds of languages, as opposed to the few million enabled by Hansards.

Currently, there are two main approaches to parallel data curation: (1) document and sentence alignment, and (2) comparable corpora methods.

Document & Sentence alignment The first approach is very similar in spirit to that used on Handards: Parallel documents are identified and then document pairs are aligned at the sentence level to produce sentence-level translation pairs. These steps are referred to as document alignment and sentence alignment, respectively. The web has become the default source of documents (Resnik, 1998), where businesses, governments, and individuals regularly release documents and translations of

[1] http://www2.statmt.org/wmt23/data-task.html

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 95–102

those documents–for example a user manual that is published in several languages. A very simple and computationally inexpensive approach to finding parallel documents is to locate URLs which differ in no more than a language code (Resnik and Smith, 2003). However, more accurate (and computationally expensive) methods have also been developed which look for documents which appear to contain similar information, for example by translating all documents into one language and then finding pairs via TF-IDF similarity (Buck and Koehn, 2016b). More recent approaches to document alignment have relied on finding similar vectors after converting documents into multilingual vectors, created via combining sentence embeddings (Thompson and Koehn, 2020) or by embedding entire documents (Guo et al., 2019). A WMT shared task on document alignment was held in 2016 (Buck and Koehn, 2016a).

Once parallel documents have been located, they are sentence aligned. Sentence alignment consists of finding a bipartite graph which matches minimal groups of sentences that are translations of each other. This is necessary because content may have been inserted or deleted in the translation process, and sentences may have been combined or split in the translation process. Additionally, sentence segmentation errors may cause sentences to be split or combined. An example of an early sentence alignment algorithm is Gale-Church (Gale and Church, 1993), which like the original IBM system uses only the length of each sentence, making it very computationally efficient but not particularly accurate. Bleualign (Sennrich and Volk, 2010, 2011) used an MT system to convert one text into the language of the other and then performed n-gram matching, similar to the BLEU MT metric (Papineni et al., 2002). A more recent sentence aligner is Vecalign (Thompson and Koehn, 2019), which uses multilingual sentence embeddings and a dynamic programming approximation (Salvador and Chan, 2007) which makes the algorithm linear with respect to the number of sentences being aligned. Widely used datasets created via document and sentence alignment include Paracrawl (Bañón et al., 2020) and CCAlign (El-Kishky et al., 2020).

Comparable Corpora A recent alternative to document and sentence alignment is to discard document information and simply create a collection of sentences in each language, and then find translation pairs by looking for sentences which are nearby by in a multilingual embedding space. LASER (Artetxe and Schwenk, 2019) was proposed for this task. The authors additionally proposed a margin-based score which gives preference to sentence pairs which are more similar to one another than other potential matches by at least a minimum margin. Approximate nearest neighbor search (Johnson et al., 2019) is used to make the search for sentence pairs tractable. Examples of widely-used datasets created via the comparable corpora method include Wikimatrix (Schwenk et al., 2021a) and CCMatrix (Schwenk et al., 2021b).

2.1 Parallel Corpora Filtering

Once data has been aligned, it is customary—especially for data coming from the web—to perform data filtering to remove low quality translation pairs before using the data for training, as unfiltered web-crawled data harms translation performance (Khayrallah and Koehn, 2018). There have been three prior shared tasks on bitext filtering at WMT (Koehn et al., 2018, 2019, 2020).

Popular approaches to data filtering include LASER margin filtering (Chaudhary et al., 2019), using an approach similar to the comparable corpora method described above, and dual conditional cross entropy (Junczys-Dowmunt, 2018), which trains NMT models on held-out clean data in both the forward and reverse directions and uses them to compute cross-entropy scores for the data being filtered. Sentence pairs with divergent or poor cross-entropies are down-weighted.

3 Shared Task Definition

This shared task presented the open-ended problem of finding the best possible subset of aligned sentence pairs from unaligned documents sourced from the internet. Participants were evaluated on downstream machine translation system performance.

Parallel data curation from web can be computationally demanding due to the sheer scale of webcrawled data. For this reason, in addition to our documents, we also released pre-computed intermediate steps from a baseline, so participants can choose to focus on one aspect of the task (e.g. sentence filtering.)

For this shared task, the organizers provided:

- Web-crawled data, as unique sentences or unique documents

- LASER2 sentence embeddings
- K-nearest neighbors by cosine similarity from our baseline
- End-to-end scripts for MT training and evaluation

End-to-end scripts enabled participants to supply a set of sentence ids and train and evaluate a Sockeye MT model (Hieber et al., 2022). Alongside the scripts, we provided a simple baseline based on 1-best cosine similarity.

Participants were allowed to use only pre-trained models and datasets publicly released with a research-friendly license on or before May 1, 2023.

3.1 Dataset

All of our inputs were derived from the 2023-06 snapshot of Common Crawl. We extracted the plain text from HTML using the *trafilatura* library (Barbaresi, 2021), and ran the first 2,000 characters through the 176-language fasttext language id model (Joulin et al., 2016a,b).

We kept all documents classified as Estonian or Lithuanian, unless their hostnames were included in the following lists from the blocklist project:[2] abuse, basic, crypto, drugs, fraud, gambling, malware, phishing, piracy, porn, ransomware, redirect, scam, torrent. No further data filtering was performed.

We split documents into paragraphs at line breaks, and segmented resulting paragraphs into sentences using the Media Cloud sentence splitter.[3]

Each unique sentence was given a Globally Unique IDentifier (GUID) and tagged with a language id based on *fastText*.

3.1.1 Dataset Statistics

Our dataset includes documents taken from 402,920 hosts. Only 24,319 of these hosts included documents in both languages. Table 1 includes overall counts on a per language basis.

3.1.2 Intermediate Outputs From Baselines

We provide participants with intermediate outputs from our baseline systems as additional resources, such that prospective participants could be able to access sentence embedding or sentence pair similarity information without needing computational resources to create these themselves.

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

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

	Estonian	Lithuanian
# Hosts	199,813	227,426
# Documents	3,449,211	4,571,947
# Sentences	53,234,425	63,488,253
# Sents w/ LangId	36,870,945	46,969,824

Table 1: Counts of unique hosts, documents, sentences, and sentences identified as the correct language in our dataset

We provide outputs of embedding each sentence with the LASER 2 model (Heffernan et al., 2022). We also release a smaller version of the embeddings, projected down to 128 dimensions via PCA and converted to float16.

To create baseline sentence pair alignments, we removed sentences detected as non-Estonian or non-Lithuanian, and used the FAISS library (Johnson et al., 2019) to index our LASER2 embeddings for fast retrieval. We applied L2 normalization to the embeddings, and added them to a flat inner product index, so that the resulting scores were equivalent to cosine similarity. We queried each index with embeddings in the other language, and returned the top eight results. These raw cosine similarity scores are shared with participants as a potential resource, and serve as the basis for our baseline submissions.

4 Evaluation

We evaluated submissions by using the curated data to train machine translation systems.

For preprocessing, we split sentences into subwords by applying Byte-Pair Encoding (BPE) (Sennrich et al., 2016) using 32,000 merge operations. The BPE vocabulary is learned jointly for the source and target language. We apply a minimum vocabulary frequency of 100 per language.

We use Sockeye (Hieber et al., 2022) to train Transformer (Vaswani et al., 2017) translation models with 512 hidden units, 8 attention heads, 6 layers and feed-forward layers of size 2048. For training we use an effective batch size of 400k target tokens. We use 4096 target tokens per GPU, and gradient accumulation to obtain 400k target tokens regardless of the number of GPUs.

We use the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_1 = 0.98$, an initial learning rate of 0.06325, a linear warmup for 4000 updates and an inverse square root learning rate

# Sents	Min Margin Score	EMEA	EUbookshop	Europarl	JRC-Acquis	average
1.6M	1.048	21.1	23.2	20.3	17.9	20.6
3.2M	1.027	**21.9**	**23.6**	**20.8**	**18.5**	**21.2**
4.8M	1.019	21.7	**23.6**	**20.8**	18.4	21.1
6.4M	1.013	21.6	23.4	**20.8**	18.3	21.0
8.0M	0.900	21.3	23.3	20.6	18.1	20.8

Table 2: Comparison of different training data sizes and margin score cutoffs on development set BLEU.

decay. Checkpoints are written every 500 updates and training is stopped once validation perplexity does not improve for 12 checkpoints. The checkpoint with the lowest validation perplexity is used as the final checkpoint.

All systems are trained on nodes with 8 V100 GPUs. We use BLEU (Papineni et al., 2002) and chrF (Popović, 2015) as quality metrics. Evaluation metrics are computed using Sacrebleu (Post, 2018).

We considered data from four domains for evaluation: EMEA,[4] EUbookshop,[5] Europarl,[6] JRC-Acquis,[7] and EUconst.[8] All data is released by OPUS (Tiedemann, 2012). From each domain, we created a dev, test, and held-out-test set. We use up to 10,000 lines for each. If less data is available, it is split between the three sets. We also kept EUconst as a held-out domain.

5 Systems

We report the results of four different systems: the baseline, two participant systems, and a contrastive system.

5.1 Baseline

The naive baseline was designed to give participants a simple end-to-end system, so they could focus on any part of the pipeline to improve upon. While participants were not required to build upon the baseline, doing so lowered the barrier to entry.

As described in Section 3.1.2, we used the LASER 2 model to embed all Estonian and Lithuanian sentences, indexed them with FAISS, and computed the eight nearest neighbors' cosine similarities for each sentence in each language. We provided these cosine similarity scores as an additional resource for participants.

Our naive baseline was created by taking all sentence pairs whose cosine similarities whose 1-best neighbor exceeded or matched the threshold of 0.9 in the Estonian → Lithuanian direction, meaning that multiple target sentences could be aligned to the same source.

This naive baseline was designed to be an end-to-end solution to allow participants to improve on any of the individual parts (filtering, alignment, margin scoring, etc).

5.2 Steingrímsson

Steingrímsson (2023b) first perform document alignment and sentence alignment, and then use matches from the provided top1-cosine data for sentences which were not aligned via document/sentence alignment.

They perform sentence alignment of all document pairs within each web domain and score the alignments to locate document pairs, similar to Thompson and Koehn (2020), to find high-quality document pairs. They use the recently proposed SentAlign[9] (Steingrímsson, 2023a; Steingrímsson et al., 2023b) sentence aligner, which in turn uses LaBSE (Feng et al., 2022) sentence embeddings.

They also perform extensive bitext filtering, using several different language ID tools and the filtering method proposed in Steingrímsson et al. (2023a) which uses perplexities of a GPT-2 model (Radford et al., 2019), LAESR embeddings (Chaudhary et al., 2019), NMTScore (Vamvas and Sennrich, 2022) using Prism (Thompson and Post, 2020a,b), and WAScore (Steingrímsson et al., 2021), as well as Bicleaner AI (Zaragoza-Bernabeu et al., 2022).

5.3 Nguyen-Hoang et al.

Nguyen-Hoang et al. (2023) focus on using the phrase based dictionary to distill the high-quality sentences and making a pipeline to re-ranking the top-K cosine similarity.

They begin with the released data, and an MGiza-based (Gao and Vogel, 2008) dictionary. They then extract sentence pairs using the a top-1 cosine score

[4] https://opus.nlpl.eu/EMEA.php
[5] https://opus.nlpl.eu/EUbookshop.php
[6] https://opus.nlpl.eu/Europarl.php
[7] https://opus.nlpl.eu/JRC-Acquis.php
[8] https://opus.nlpl.eu/EUconst.php
[9] https://github.com/steinst/SentAlign

	BLEU				ChrF			
Test	EMEA	EUbooks	Europarl	JRC-Acquis	EMEA	EUbooks	Europarl	JRC-Acquis
Top1_cosine	18.1	20.1	18.4	25.7	49.4	**53.0**	52.1	**55.7**
Nguyen-Hoang et al.	18.5	**20.4**	**19.1**	**25.8**	48.9	52.5	**52.5**	55.5
Steingrímsson	**20.4**	20.2	18.7	25.4	**51.4**	52.8	52.0	54.9
MarginScore 3.2M	21.5	22.4	20.2	27.9	52.5	54.7	53.4	57.8

Table 3: Test set BLEU and ChrF scores. Top1_cosine is the baseline, and Marginscore 3.2M is the contrastive system.

	BLEU					ChrF				
Held-out	EMEA	EUbooks	Europarl	JRC-A	EUconst	EMEA	EUbooks	Europarl	JRC-A	EUconst
Top1_cosine	18.7	14.0	18.2	22.9	23.8	49.8	**47.6**	52.4	54.0	**58.5**
Nguyen-Hoang et al.	19.3	**14.4**	**19.1**	**23.5**	**25.1**	49.7	47.4	**52.9**	**54.2**	58.3
Steingrímsson	**21.0**	**14.5**	18.7	23.1	23.2	**52.1**	**47.6**	52.3	53.6	57.8
MarginScore 3.2M	21.9	16.1	20.5	25.4	27.6	52.9	48.9	53.8	56.2	60.9

Table 4: Held-out test BLEU and ChrF scores. Top1_cosine is the baseline, and Marginscore 3.2M is the contrastive system.

and a threshold. From there, the dictionary is used to translate the source sentences. These dictionary-translated sentences are then compared with the translation from the baseline data. The translation from the baseline data is filtered based on the edit distance. Then a NMT model is trained, and the final threshold is set based on NMT model performance.

Nguyen-Hoang et al. (2023) also perform an analysis on the cosine score threshold, demonstrating how varying this value impacts both corpus size and translation quality.

5.4 Contrastive System

The participants in this task both performed data filtering on top of the the top-1 cosine baseline.

Since no participants experimented with using margin scoring, which Schwenk et al. (2021b) found significant for improving the quality of LASER-based mining, the organizers created a stronger contrastive system that did so.

We calculated margin scores for our four nearest neighbors in both directions. We performed competitive linking,[10] such that each sentence appeared only once in our contrastive submission. Although we computed cosine similarities for the eight nearest neighbors, no appreciable difference was found in MT quality by using k=8 instead of k=4 when computing margin scores.

We sorted our data by margin score and compared different data sizes, as shown in Table 2. We used a minimum margin score of 1.027 and data size of 3.2 million lines since it scored the highest on all development sets and had the highest average score.

6 Results

Table 3 and Table 4 show the BLEU and ChrF results of the naive top-1 cosine baseline, participant submissions, and the contrastive margin score system. Of the baseline and two participant systems, we bold the best and systems within 0.1 of the best. Overall, both participants improved over the naive baseline. On the held-out test sets, Steingrímsson had higher BLEU on EMEA and EUbookshop, while Nguyen-Hoang et al. had higher BLEU on Europarl, JRC-Acquis, and the held-out domain of EUconst.

We see that the contrastive margin score system outperforms the naive top-1 cosine baseline. This confirms the finding of Schwenk et al. (2021b) that margin scoring outperforms raw cosine similarity. The contrastive margin score system also outperforms the participant submissions that directly build and improve upon the naive top-1 cosine baseline.

Data filtering and alignment tend to be complimentary, so the filtering methods proposed by the

[10] Referred to as the "max strategy" by Schwenk et al. (2021b).

participants would likely improve upon the contrastive margin score system if they were applied on top of it.

7 Conclusion

While data curation is the first step in the training of any MT (or machine learning) model, this tends to be a less-published-upon topic in academic research.

In this shared task, we have released the processed webcrawled data, and a baseline system with intermediate outputs. We hope this task lowers the barrier of entry and allow participants to focus on any aspect of the data curation pipeline (document alignment, sentence alignment, filtering, etc.) We have trained and evaluated MT systems on the datasets curated by participating teams. We have presented results for two participant submissions, in addition to two more systems built by the shared task organizers.

We hope this work serves as a building block for future research on this topic.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176.

Christian Buck and Philipp Koehn. 2016a. 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. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016b. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 64–72, Florence, Italy. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, et al. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv preprint arXiv:2207.05851*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Minh-Cong Nguyen-Hoang, Van Vinh Nguyen, and Le-Minh Nguyen. 2023. A fast method to filter noisy parallel data WMT2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Philip Resnik. 1998. Parallel strands: a preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 72–82, Langhorne, PA, USA. Springer.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. 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. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

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

Steinþór Steingrímsson. 2023b. A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

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. University of Tartu Library.

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. 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). INCOMA Ltd.

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. Association for Computational Linguistics.

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.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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