Finetuning a Kalaallisut-English machine translation system using web-crawled data

Alex Jones
Dartmouth College
alexander.g.jones.23@dartmouth.edu

Abstract

West Greenlandic, known by native speakers as Kalaallisut, is an extremely low-resource polysynthetic language spoken by around 56,000 people in Greenland. Here, we attempt to finetune a pretrained Kalaallisut-to-English neural machine translation (NMT) system using web-crawled "pseudoparallel" sentences from around 30 multilingual websites. We compile a corpus of over 93,000 Kalaallisut sentences and over 140,000 Danish sentences, then use cross-lingual sentence embeddings and approximate nearest-neighbors search in an attempt to mine near-translations from these corpora. Finally, we translate the Danish sentences to English to obtain a synthetic Kalaallisut-English aligned corpus. Although the resulting dataset is too small and noisy to improve the pretrained MT model, we believe that with additional resources, we could construct a better pseudoparallel corpus and achieve more promising results on MT. We also note other possible uses of the monolingual Kalaallisut data and discuss directions for future work. We make the code and data for our experiments publicly available. [1]

1 Introduction

Low-resource machine translation—translation involving languages with very few linguistic resources, especially parallel data—is a massive challenge in NLP. The lack of data makes it incredibly difficult to train and evaluate NMT systems and, to a lesser extent, rule-based and statistical MT systems. Several works have summarized the manifold challenges that come with attempting low-resource MT (Haddow et al., 2021; Ranathunga et al., 2021).

In this paper, we tackle MT involving Kalaallisut, a polysynthetic Eskimo-Aleut language spoken by around 56,000 people in Greenland. Only a couple works have attempted Kalaallisut-English MT specifically (de Mol, 2020; Kelly, 2020), both of which see poor performance due to noisy and insufficient data. For this reason, we attempt different methods than them for aligning crawled data.

A small number of works have looked at MT and morphological segmentation for other polysynthetic Eskimo-Aleut languages, including Inuktitut and Yupik (Roest et al., 2020; Joanis et al., 2020; Ngoc Le and Sadat, 2020; Micher, 2018b; Le and Sadat, 2020; Micher, 2018a). Among these efforts was the creation of an Inuktitut-English parallel corpus, the Nunavut Hansard corpus.

In this paper, we first web-crawl monlingual texts in Kalaallisut and Danish from multilingual Greenlandic websites. We then attempt to find "pseudoparallel" sentence pairs, or sentences that are near-translations of each other, between these two sets of monolingual corpora. To this end, we deploy cross-lingual sentence embeddings and a highly optimized library for approximate nearest-neighbors search. We then try to finetune pretrained open-source MT model from HuggingFace (Wolf et al., 2020), namely the Helsinki-NLP/OPUS-MT model[2] (Tiedemann and Thottingal, 2020) trained using the MarianMT toolkit (Junczys-Dowmunt et al., 2018).

2 Methodology

2.1 Web crawling

The first step in our process was to gather monolingual sentences from data sources containing "similar" texts. In our case, we chose to crawl texts from multilingual websites that contain the same content in multiple languages. Originally, we planned to crawl texts in Kalaallisut, Danish, and English. However, due to the scarcity of English text we were able to obtain, we ended up not using the English sentences. The resulting set of Kalaallisut and

[1] https://github.com/AlexJonesNLP/KALComp

[2] https://huggingface.co/Helsinki-NLP/opus-mt-kl-en

Danish sentences is a “comparable corpus": a collection of similar texts (i.e. about the same topics or events) in two or more languages. Intuitively, one might assume that such corpora sometimes contain sentences that are (near-)translations of each other, which is indeed the case (Schwenk et al., 2021a,b; Rapp et al., 2021).

We compile a list of 29 Greenlandic websites with content in Kalaallisut and Danish. This included a diverse mix of news, government, corporate, scientific, and travel/hospitality websites. Next, we used the Python `requests`[3] library in tandem with `BeautifulSoup`[4] to obtain and parse HTML from these sites. In order to maximize the amount of text we crawl, we scrape text from webpages recursively. That is, we scrape text from a webpage, and then we retrieve links on that page that direct to other pages on the website. This allows us to scrape large portions of certain websites. After scraping, we clean the text using the `cleantext`[5] library. We end up with 93,093 Kalaallisut sentences and 140,802 Danish sentences, after removing duplicates.

2.2 Bitext mining

The bitext mining task consists of trying to find sentence pairs that are translations or near-translations of each other between two sets of monolingual sentences in different languages.

A large number of approaches have been attempted for this task over the past two decades (Zhao and Vogel, 2002; Resnik and Smith, 2003; Munteanu et al., 2004; Fung and Cheung, 2004; Munteanu and Marcu, 2006; Azpeitia et al., 2017, 2018; Bouamor and Sajjad, 2018; Hangya et al., 2018; Schwenk, 2018; Ramesh and Sankaranarayanan, 2018; Artetxe and Schwenk, 2019a,b; Hangya and Fraser, 2019; Schwenk et al., 2019a,b; Wu et al., 2019; Keung et al., 2020; Tran et al., 2020; Kvapilíková et al., 2020). However, current state-of-the-art techniques generally involve embedding sentences in both languages using a single cross-lingual encoder and then performing approximate nearest-neighbors search with some similarity metric. We hew closely to the method introduced in Artetxe and Schwenk (2019a).

[3] `https://docs.python-requests.org/en/latest/`

[4] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`

[5] `https://pypi.org/project/clean-text/`

2.2.1 Cross-lingual sentence embeddings

In order to create vector representations of sentences in different languages that are useful for the bitext mining task, we employ the cross-lingual sentence encoder LaBSE (Language-Agnostic BERT Sentence Embedding) (Feng et al., 2020). Cross-lingual encoders aim to create language-agnostic representations of sentences. In theory, two semantically equivalent sentences in different languages (i.e. translations) should map to the exact same vector. This makes them the ideal tool for this task.

LaBSE combines pretraining on the masked language modeling (MLM) and translation language modeling (TLM) (Conneau and Lample, 2019) tasks with the dual encoder translation ranking and additive margin softmax objectives. The model is trained on data in 109 languages, all of which have at least some parallel data aligned with English. LaBSE achieves state-of-the-art results when used for the bitext mining task (Feng et al., 2020; Reimers and Gurevych, 2020). We use the implementation of LaBSE available on the `Sentence Transformers` library.[6] (Reimers and Gurevych, 2019). The embeddings are 768-dimensional, and the embedding is done with a Quadro RTX 6000 GPU.

2.2.2 Dictionary translation

The issue with using LaBSE for embedding Kalaallisut sentences is that it was not trained on any Kalaallisut data. Furthermore, LaBSE doesn't even have training data for languages in the Eskimo-Aleut family that Kalaallisut is a part of, meaning it cannot leverage related-language data for zero-shot transfer as it could in less resource-scarce scenarios. Because of this, it is desirable to translate all or part of the Kalaallisut text to a higher-resource language before attempting similarity search with LaBSE.

We end up using the rather rudimentary method of translating parts of the Kalaallisut text using Kalaallisut-English and Kalaallisut-Danish dictionaries[7], resulting in a type of “code-switched" corpus. However, these dictionaries were highly incomplete, so we were only able to translate $\approx 16\%$ of the Kalaallisut words to English or Danish.

[6] `https://www.sbert.net`

[7] `https://www.mobileread.com/forums/showthread.php?t=20480&page=11`

2.2.3 BPE

Kalaallisut is a highly polysynthetic language, meaning it has many morphemes per word on average and a huge (effectively limitless) vocabulary size. This poses a significant challenge for a host of NLP problems, including cross-lingual similarity search and MT.

de Mol (2020) explored various methods for morphological segmentation of Kalaallisut, as well as their downstream consequences on MT. Although Conditional Random Fields performed best on a gold-standard morphological segmentation evaluation in their papere, we elect to use the simpler byte-pair encoding (BPE) algorithm (Sennrich et al., 2016), a language-agnostic and non-linguistic tokenization scheme that produces subword units based on statistical analysis of a training corpus. BPE has proven to be useful in MT training and has become a staple tokenization method in NLP. Segmentation has also been shown to help bitext mining performance between polysynthetic and analytic languages, such as Inuktitut and English (Jones et al., 2021).

We use the BPE implementation from the Python package `youtokentome`[8], with a vocabulary size of 10,000 and otherwise default settings. We train the BPE model on all the monolingual Kalaallisut data.

2.2.4 Approximate KNN search

Similarity search We use the `Faiss` (Facebook AI Similarity Search) library (Johnson et al., 2017) to perform similarity search between LaBSE embeddings of Danish and code-switched Kalaallisut sentences. `Faiss` is an aggressively optimized package that allows for extremely fast k-nearest neighbors search on GPUs. We use a Python implementation[9] of `Faiss`, although the original package[10] is written in C++. We use $k = 4$ neighbors and a batch size of 100 for our hyperparameter settings, and run the search on a Quadro RTX 6000 GPU. We also run both forward and backward searches, i.e. $\forall x \in \mathcal{X}$ we find the most similar sentence $y \in \mathcal{Y}$ and vice-versa. We then take the union of both searches to give us a larger pool of candidate sentence pairs to choose from (see Artetxe and Schwenk (2019a) for more details).

Margin criterion Many works use the cosine similarity metric to quantify similarity between vector representations, which is simply the dot product of two normalized vectors. However, we use the slightly more sophisticated margin criterion from Artetxe and Schwenk (2019a) instead, which has been shown to give better performance on the bitext mining task compared to simple cosine similarity. The margin score between two sentence embeddings $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is given by:

$$\text{score}_{\text{margin}}(x, y) = \frac{2k \cos(x, y)}{\sum_{z \in NN_k(x)} \cos(x, z) + \sum_{z \in NN_k(y)} \cos(y, z)}$$

where $NN_k(v)$ denotes the set of k-nearest neighbors of vector $x \in \mathbb{R}^n$ and $\cos(x, y)$ is the cosine similarity between x and y, i.e. $\frac{x \cdot y}{\|x\| \|y\|}$. The margin score measures the ratio between the cosine similarity of x and y and the average cosine similarity of x and its k-nearest neighbors and y and its k-nearest neighbors. The vector pair (x, y) that maximizes the margin criterion is the one that "stands out" the most in terms of similarity among its neighbors.

We use a margin score threshold of 1.04 to extract our final set of pseudoparallel sentence pairs, identical to Schwenk et al. (2021a). There is obviously a precision-recall tradeoff in selecting a margin threshold, and slightly higher thresholds such as 1.06 have also shown success (Schwenk et al., 2021b). Since our monolingual datasets are so small already, we opt for a threshold that will give us somewhat more sentence pairs at the expense of quality. We end up with 6,393 sentence pairs.

2.2.5 Post-translation

Recall that we performed similarity search between embeddings of Danish sentences and code-switched Kalaallisut sentences. After mining and filtering with a margin threshold, we simply map the code-switched Kalaallisut sentences back to the "pure" Kalaallisut sentences they came from. This resulted in a pseudoparallel Kalaallisut-Danish corpus. However, the model we wished to finetune is a Kalaallisut-English MT system. So we translate the Danish sentences to English using the Helsinki-NLP/OPUS-MT Da-En model[11]. We run this model on a NVIDIA TITAN V GPU, translating in batches of 32 sentences. The resulting

[8] https://pypi.org/project/youtokentome/

[9] https://pypi.org/project/faiss-gpu/

[10] https://github.com/facebookresearch/faiss

[11] https://huggingface.co/Helsinki-NLP/opus-mt-da-en

training set is a synthetic Kalaallisut-English pseudoparallel corpus.

2.3 NMT finetuning

The next step was to finetune the pretrained Kl-En MT model on the pseudoparallel data we had gathered. We finetune the model for only one epoch with a batch size of 8, for a total of 720 optimization steps. We use a learning rate of 2e-5 and a weight decay parameter of 0.01. Unfortunately, due to technical difficulties we were forced to run our code in a Google Colab instance, which meant using a Tesla K80 GPU instead of a more powerful GPU. Finetuning took around 23 minutes.

3 Evaluation

We evaluate using BLEU score (Papineni et al., 2002), the most popular metric in the MT literature. The test set we use is the same set[12] the pretrained model was evaluated on, which consists of texts from the JW300 parallel corpus (Agić and Vulić, 2019). Disappointingly, the finetuned model achieves a BLEU score of only 14.67, compared to a score of 27.75 for the pretrained model. There are multiple factors that may explain this. For one, the pseudoparallel data we extracted is incredibly noisy, despite attempts to bootstrap similarity search using dictionary translation. A glance at the dataset reveals that many, if not most, of the sentence pairs are not translations or even near-translations. Second, the pretrained model is almost certainly overfit to the JW300 data, since this corpus is also the source of the training data. Earlier in the experiment, when we tried to translate the crawled Kalaallisut to English, the pretrained model generated gibberish sentences that seemed to be replicating what it saw in the training corpus, with apparently no connection to the input. So although the data we mined is undoubtedly noisy, we do not put much stock in this evaluation due to the problem of overfitting to the test set's underlying data distribution. Future efforts should seek to use an unbiased test set, which we were unfortunately unable to find.

4 Discussion

It is unfortunate that we were unable to mine a sufficient number of high-quality sentence pairs to improve a pretrained MT system. However, we would like to take stock of the contributions of this project, as well as discuss avenues for potential improvement.

Although our efforts did not result in a high-quality aligned corpus, we still managed to construct a Kalaallisut-Danish *comparable* corpus purely from web-crawled data. We plan to release this data publicly for future research endeavors. Furthermore, the nearly 100K Kalaallisut sentences we extracted could be used for unsupervised tasks such as language modeling.

We believe we could improve the bitext mining process for Kalaallisut using larger dictionaries with better coverage, and/or a decent pretrained MT system. A larger Kalaallisut-English dictionary appears to be in the works[13], but we couldn't locate any text files associated with it. There is a rule-based Kalaallisut-Danish MT system[14] hosted by Oqaasileriffik, the language secretariat of Greenland. Cursory experiments with this system yielded promising results, but we were unable to co-opt the system for batch translation (e.g. in a Python environment) due to the fact that this tool is currently just an on-demand web API and UI. In the future, we can contact the makers of this tool to see if we can use it for large-scale translation.

Our crawling pipeline was also suboptimal due to its uniformity (and other reasons, such as time constraints and Python recursion depth limits), and we could likely mine considerably more monolingual sentences using custom parsers. This should lead to improved downstream performance as well.

5 Conclusions

In this paper, we attempt to finetune a pretrained Kalaallisut-English MT system using web-crawled data. We scrape monolingual text from roughly 30 multilingual websites in Kalaallisut and Danish and then try to mine pseudoparallel sentences using LaBSE and `Faiss`, after which we translate the Danish to English to create a synthetic, pseudoparallel Kalaallisut-English corpus. Although finetuning results are disappointing, we offer the Kalaallisut-Danish comparable corpus as a novel and valuable resource, and suggest multiple directions for possible improvement.

[12] https://object.pouta.csc.fi/OPUS-MT-models/kl-en/opus-2020-01-09.test.txt

[13] https://scholarspace.manoa.hawaii.edu/handle/10125/26190

[14] https://nutserut.gl/en

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.

Adoni Azpeitia, Thierry Etchegoyhen, and Eva Martinez 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. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Barbera de Mol. 2020. *A Comparison of Data-Driven Morphological Segmenters for Low-Resource Polysynthetic Languages: A Case Study of Greenlandic*. Ph.D. thesis, University of Groningen.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. *arXiv e-prints*, page arXiv:2007.01852.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, Geneva, Switzerland. COLING.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2021. Survey of Low-Resource Machine Translation. *arXiv e-prints*, page arXiv:2109.00486.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 7–13, Bruges, Belgium.

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. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale Similarity Search with GPUs. *arXiv e-prints*, page arXiv:1702.08734.

Alexander Jones, William Yang Wang, and Kyle Mahowald. 2021. A massively multilingual analysis of cross-linguality in shared embedding space. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Kevin Kelly. 2020. *An Evaluation of Parallel Text Extraction and Sentence Alignment for Low-Resource Polysynthetic Languages*. Ph.D. thesis, University of Groningen.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2020. Unsupervised bitext mining and translation via self-trained contextual embeddings. *Transactions of the Association for Computational Linguistics*, 8:828–841.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus

mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Ngoc Tan Le and Fatiha Sadat. 2020. Addressing challenges of indigenous languages through neural machine translation: The case of inuktitut-english.

Jeffrey Micher. 2018a. Using the Nunavut Hansard data for experiments in morphological analysis and machine translation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 65–72, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jeffrey C Micher. 2018b. Addressing challenges of machine translation of inuit languages.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 265–272, Boston, Massachusetts, USA. Association for Computational Linguistics.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia. Association for Computational Linguistics.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online). International Committee on 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.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural Machine Translation for Low-Resource Languages: A Survey. *arXiv e-prints*, page arXiv:2106.15115.

Reinhard Rapp, Serge Sharoff, and Pierre Zweigenbaum, editors. 2021. *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*. INCOMA Ltd., Online (Virtual Mode).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv e-prints*, page arXiv:1907.05791.

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, and Armand Joulin. 2019b. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. *arXiv e-prints*, page arXiv:1911.04944.

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.

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. European Association for Machine Translation.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual Retrieval for Iterative Self-Supervised Training. *arXiv e-prints*, page arXiv:2006.09526.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lijun Wu, Jinhua Zhu, Di He, Fei Gao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Machine translation with weakly paired documents. 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 4375–4384, Hong Kong, China. Association for Computational Linguistics.

Bing Zhao and Stephan Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, page 745, USA. IEEE Computer Society.