CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web

Holger Schwenk and **Guillaume Wenzek** and **Sergey Edunov**
Edouard Grave and **Armand Joulin** and **Angela Fan**
Facebook AI
{schwenk,guw,edunov,egrave,ajoulin,angelafan}@fb.com

Abstract

We show that margin-based bitext mining in a multilingual sentence space can be successfully scaled to operate on monolingual corpora of billions of sentences. We use 32 Common Crawl snapshots (Wenzek et al., 2019), totalling 71 billion unique sentences. Using one unified approach for 90 languages, we were able to mine 10.8 billion parallel sentences, out of which only 2.9 billions are aligned with English. We illustrate the capability of our scalable mining system to create high quality training sets from one language to any other by training hundreds of different machine translation models and evaluating them on the many-to-many TED benchmark. Further, we evaluate on competitive translation benchmarks such as WMT and WAT. Using only mined bitext, we set a new state of the art for a single system on the WMT'19 test set for English-German/Russian/Chinese. In particular, our English/German and English/Russian systems outperform the best single ones by over 4 BLEU points and are on par with best WMT'19 systems, which train on the WMT training data and augment it with backtranslation. We also achieve excellent results for distant languages pairs like Russian/Japanese, outperforming the best submission at the 2020 WAT workshop. All of the mined bitext will be freely available.

1 Introduction

Parallel data, i.e. sentences in two languages which are mutual translations, are a crucial resource for many multilingual natural language processing tasks. Traditionally, high quality parallel texts are obtained from the publications of international organizations like the the United Nations (Ziemski et al., 2016) or the European Parliament (Koehn, 2005). These are professional human translations, but they are in a more formal language and tend to be limited to political topics. Another direction is to rely on volunteers to provide translations for public texts, such as the TED corpus (Qi et al., 2018), news commentary (Tiedemann, 2012) or OpenSubtitles (Lison and Tiedemann, 2016), but this approach lacks scalability.

There is also a large body of works which aims in mining bitexts by *comparing* huge collections of monolingual data. Our aim is to mine at massive scale, both in number of possible languages and in quantity of mined parallel sentences. Most existing large scale bitext mining techniques use a hierarchical approach. First, a subset of texts that may contain parallel sentences are selected at the document level. Subsequently, sentences within these aligned documents are compared to identify parallel ones. This local mining is potentially fast since only a few thousand sentences need to be compared for each document pair. However, sentences not present in these pre-selected documents cannot be aligned, which vastly limits the quantity of mineable bitext. A first system to globally compare all sentences in monolingual collections for many language pairs was presented in Schwenk et al. (2019), but was limited to only Wikipedia.

In this paper, we show that this type of global mining scales to extremely huge corpora: 71 billion sentences, about 120x larger than the work of Schwenk et al. (2019). Our contributions are:

- development of a new highly efficient and parallelized processing pipeline to confront the substantial computational challenge;
- unprecedented size: 10.8 billion mined parallel sentences in 90 different languages;
- all these resources are freely available;
- we demonstrate the quality of our mined data on a variety of machine translation benchmarks, such as TED, WMT, and WAT, achieving highly competitive results.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6490–6500
August 1–6, 2021.

2 Related work

Much previous work has explored the automatic creation of parallel data from monolingual resources. In this section, we detail various approaches and illustrate the differences of our algorithmic approach and the scale of our mining.

Mining Methodology At the start, various approaches used alignment on information beyond text itself, such as with document metadata (Resnik, 1999; Resnik and Smith, 2003). Later, work aligned based on text with techniques such as Jaccard similarity (Etchegoyhen and Azpeitia, 2016; Azpeitia et al., 2017, 2018), crosslingual document retrieval (Utiyama and Isahara, 2003; Munteanu and Marcu, 2005), language models (Buck and Koehn, 2016), translation (Abdul-Rauf and Schwenk, 2009; Bouamor and Sajjad, 2018), or bag-of-words (Buck and Koehn, 2016). In contrast, we use massively multilingual sentence embeddings trained on almost 100 languages, and then conduct margin-based mining in the multilingual embedding space (Schwenk, 2018; Artetxe and Schwenk, 2018a,b; Kvapilíková et al., 2020). Previous work such as España-Bonet et al. (2017); Hassan et al. (2018); Guo et al. (2018); Yang et al. (2019) used bilingual embeddings, which is not scalable for mining many different languages.

Compared to work such as Schwenk (2018), we drastically increase the scale of our mining and produce two orders of magnitude more data — this is possible by the increased efficiency and scalability of our improved mining methods. A few mining approaches were applied to large quantities of language pairs. For example, the ParaCrawl project[1] mined data for all European languages. Bitextor (Esplà-Gomis and Forcada, 2010) was applied to many languages, but took an approach that required identifying parallel documents first and then extracting aligned sentences. This is similar to the ccAligned project (El-Kishky et al., 2020). In contrast to these, we mine much larger quantities of parallel data due to the global margin-based mining approach that we take.

Data used to Mine Many previous methods for data mining focused on Wikipedia. Otero and López (2010) and Patry and Langlais (2011), for instance, aligned entire parallel documents. For example, Adafre and de Rijke (2006) and Mohammadi and GhasemAghaee (2010) used machine translation systems to compare Dutch and Persian Wikipedias to English, to identify aligned sentences. Various other worked used similarities in mentioned entities to align text, such as Gottschalk and Demidova (2017) and Tsai and Roth (2016). Work such as Smith et al. (2010); Tufis et al. (2013); Aghaebrahimian (2018) used Wikipedia to mine parallel sentences, but focused on fewer languages, often high resource. In contrast, our system mines not in Wikipedia but in CommonCrawl, a much larger source of data — and is applied to a much larger quantity of languages.

Work has extended mining beyond Wikipedia. For example, ParaCrawl[1] has been heavily used (e.g. in WMT), which is based on several noisy multilingual crawls (Koehn et al., 2018, 2019). El-Kishky et al. (2019) focused on mining documents in Common Crawl rather than sentences. Our work continues this line of scalable mining on the web, but pushes to large-scale mining to produce billions of aligned sentences.

3 Distance-based mining approach

We leverage massively multilingual sentence embeddings and a margin-based criterion to mine parallel sentences. The core idea is to learn a multilingual sentence embedding, or an embedding space in which semantically similar sentences are close, independent of the language they are written in. This means that distance in the embedding space can be used to determine if two sentences are mutual translations or not. We use the open source LASER (Artetxe and Schwenk, 2018b) embeddings as they cover over 90 different languages.[2] Another recent multilingual sentence embedding is LaBSE (Feng et al., 2020).

3.1 Margin criterion

Given two sentence embeddings, how can we decide if they are mutual translations? Using an absolute threshold on the cosine distance was shown to achieve competitive results (Schwenk, 2018), but is globally inconsistent (Guo et al., 2018). Therefore, we use margin-based mining (Artetxe and Schwenk, 2018a). The margin $\mathrm{M}(x, y)$ between two sentence embeddings x and y is defined as the ratio between the cosine distance between x and y, and the average cosine similarity of its nearest

[1] http://www.paracrawl.eu/

[2] https://github.com/facebookresearch/LASER

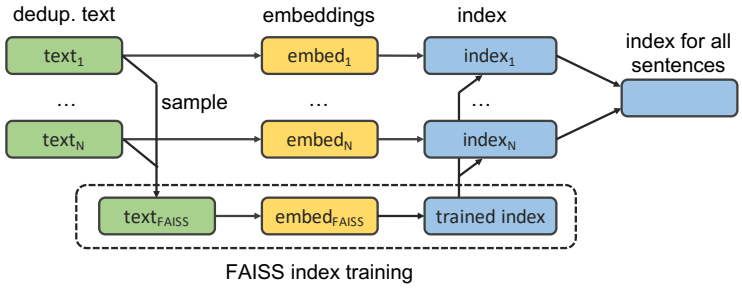

Figure 1: **Parallelized processing flow to create an FAISS index for each language**.

neighbors in both directions:

$$\mathrm{M}(x,y) = \frac{\cos(x,y)}{\sum_{z \in \mathrm{NN}_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in \mathrm{NN}_k(y)} \frac{\cos(y,z)}{2k}}$$

where $\mathrm{NN}_k(x)$ denotes the k unique nearest neighbors of x in the other language, and analogously for $\mathrm{NN}_k(y)$. We set k to 16.

Artetxe and Schwenk (2018a) describe the *max-strategy* as one of the best performing ones: the margin is calculated in both directions for all sentences in languages L_1 and L_2. Then, the union of forward and backward candidates is built, candidates are sorted, and pairs with source or target sentences which were already used are omitted. Finally, a threshold is applied to the margin score to decide if two sentences are mutual translations. This strategy was motivated by evaluation on the BUCC corpus (Zweigenbaum et al., 2018), where the reference alignments are known to be strictly 1:1. Our aim is to mine at the billion-scale, and at this size, the probability of finding multiple perfect translations increases. Therefore, we take the union of the best forward and backward alignments, excluding duplicate bitexts.

3.2 Scaling to billions of sentences

In this work, we mine billions of parallel sentences from the Web by using the data released in Common Crawl.[3] We preprocess the raw text following the pipeline used to create the CCNet dataset (Wenzek et al., 2019). We use 32 crawls spanning the period from December 2017 to February 2020.

Our CCNet corpus is about 120 times larger than Wikipedia: 71 billion compared to 595 million unique sentences (Schwenk et al., 2019). The largest corpora are English (14.3 billion), then German, French, and Spanish (more than 5.2 billion sentences). For 17 different languages, CCNet contains over one billion unique sentences (see Table 1). This requires a carefully designed mining approach in order to tackle the substantially computational complexity and successfully scale. We developed a multi-step mining procedure that is structured into three distinct tasks:

1. text extraction and processing including sentence splitting and language identification;
2. creation of a FAISS index for each language;
3. mining parallel data for each language pair using the sentence embeddings and indices.

Each step is parallelized as much as possible by splitting the data into several blocks.

Text extraction. The first task, text extraction and processing, consists of three steps: 1) extract text from the JSON data of CCNet and split the *paragraphs* into sentences; 2) mark duplicate sentences; and 3) perform language identification (LID) and exclude sentences not in the expected language. Each of these three steps processes blocks in parallel. At the final step, we merge all the block-wise deduplicated sentences and create one set of globally unique sentences for each language. We used a Python library[4] to detect sentence boundaries. If specific rules for a language are not available, we fall-back to a linguistically similar languages, e.g. using Spanish rules for Gallican, and default to English otherwise. Most of the Asian languages are handled by regular expressions. We exclude sentences with more than 500 characters. A major challenge of web data is noise. This particularly manifests in text that has the wrong language label. As noise in this stage will affect our mining process, we perform strict filtering using two LID systems on each sentence, fastText (Grave et al.,

[3] https://commoncrawl.org/

[4] https://pypi.org/project/sentence-splitter/

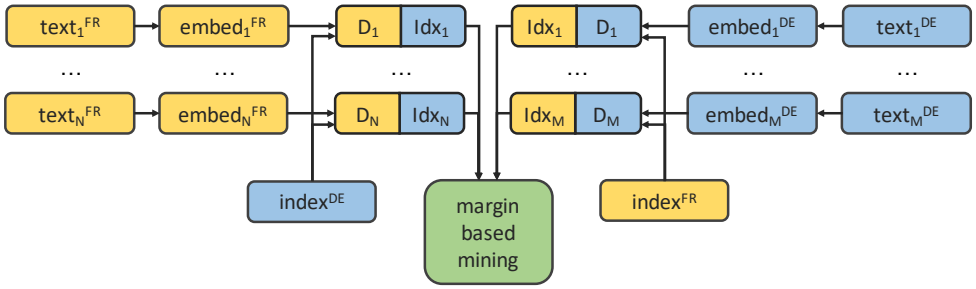

Figure 2: **Parallelized processing flow to mine parallel sentences**. **Left**: forward distances; **Right**: backward distances. **Middle**: both distances are combined according to Equation 3.1 and the extracted bitext.

2018) and LangID (Lui and Baldwin, 2011), and discard the data if the two disagree or have low confidence. This processing yields a corpus of N_i unique sentences for each language L_i. These texts are the basis for index creation and mining (see column *size* in Table 1).

Index creation. We follow Schwenk et al. (2019) and use the highly optimized FAISS library (Johnson et al., 2017)[5] to create compact indices of the sentence embedding. LASER's sentence representations are 1024-dimensional, which means that the embeddings of all sentences would require $71 \cdot 10^9 \times 1024 \times 4 \approx 290$ TB to store. To practically handle this scale, we use an aggressive vector compression based on a 64-bit product-quantizer (Jégou et al., 2011), and 64k cells to partition the search space. This corresponds to the index type `OPQ64,IVF65536,PQ64` in FAISS.

Exhaustive search in huge indices is tractable only if performed on GPU. FAISS supports sharding of a single index on multiple GPUs - this is most efficient if the GPUs are in the same machine and communicate very quickly. Our index type, using eight GPUs with 32GB of memory each, allows us to handle an index size of 3.2 billion sentences. Seven languages exceed this threshold, so we proceed to create multiple indices (English, German, French, Spanish, Russian, Chinese, and Japanese).

The processing pipeline to train and create the indices is summarized in Figure 1. We train an index on 40 million sampled sentences of the whole corpus. Once the index is trained, the data in each block is independently added to this common index, which can be performed in parallel. The individual indices are subsequently merged into one index per language. The largest indices have a size of around 210GB, making 90 indices total almost 4TB.

[5] https://github.com/facebookresearch/faiss/wiki/Faiss-indexes

Mining. After indices for all languages are created, we begin the mining process for each language pair. To illustrate the process, we describe it concretely with the example of two high resource languages, Italian and Portuguese, which have 2.5 billion sentences each. This requires $2.5\,10^9 \times 2.5\,10^9 = 6.25\,10^{18}$ distance calculations. Performing this on a single node with 8 GPUs would require more than 6 months. Instead, we tackle this computational challenge by decoupling the distance calculations of the forward and backward direction and the margin calculation, and processing these in parallel. This processing pipeline is illustrated in Figure 2.

For all language pairs, we compute both forward and backward distances, even for languages with multiple indices, such as English, French and German. All available alignments for one pair are merged, excluding duplicate sentence pairs.

In the current CCMatrix corpus, we have mined data for a diverse set of 90 languages, covering a variety of different language families and scripts (full list in the Appendix). As the mining process is computationally intensive, we focus on many commonly spoken languages to support existing translation systems, as well as mine several mid to low resource languages to provide parallel data for directions with limited to no public training data. We organized all languages into twelve groups which mostly correspond to well established linguistic language families, but we have also performed some geographic groupings, in particular for small language families or isolated languages. In addition, we have identified major languages in each group and use them as *"bridge languages"*. We mine for all bitexts among these 27 bridge languages. The motivation for this bridge language approach is to connect the languages of the various groups, but sill avoid mining the full matrix. Additional details are given in the Appendix.

3.3 Choosing the margin threshold

The margin threshold used to mine parallel sentences impacts the quality of mined bitexts. A higher threshold leads to better aligned sentences, and thus higher quality bitexts, but also to smaller datasets. Thus, there is a trade-off between size and quality. Exploratory experiments based on training different NMT models showed that a threshold around 1.06 gave good results. We display a representative example on Hungarian-Danish in Fig. 3.

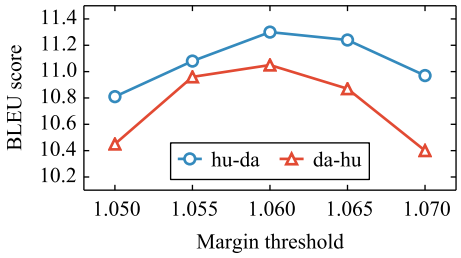

Figure 3: **BLEU scores on Hu-Da** TED dev set for various margin threshold values.

4 Quantity of Mined Data

4.1 Total Quantity

We mine a total of 10.8 billion parallel sentences out of which only 2913 million are aligned with English, considering a margin threshold of 1.06 for all language pairs. Table 1 gives a summary for the 54 largest languages. The full list of supported languages is given in the Appendix. In contrast to other works, such as the European ParaCrawl project,[1] we do not limit to alignments with English, but provide alignments for 1197 language pairs.

This yielded unprecedented amounts of bitexts of non-English language pairs, for example 286M for Spanish-French, 24M for Arabic-French and Spanish-Chinese, and a total of 326M bitexts with Norwegian (which is not present in Europarl). Further, a variety of different Asian languages were mined, producing 7.2M pairs for Japanese-Korean, 7.8M for Indonesian-Malay, and 1.3M for Bengali-Hindi. To the best of our knowledge, this makes CCMatrix the largest collection of high-quality mined parallel texts, with coverage over a wide variety of languages. Providing multiple aligned bitexts for many languages also opens the possibility of improved training of massively multilingual NMT systems (Fan et al., 2020), as this substantially increases the amount of bitexts for low resource languages. As an example, Nepali has less than 1M bitexts with English, but 17M bitexts with multiple languages (see last column of Table 1).

4.2 Analysis of mined bitexts

Table 1 gives the amount of mined bitexts for various language pairs. The general tendency is of course that mining in large monolingual corpora leads to larger extracted bitexts. This is however not systematically true. Let us consider for example Danish, a Germanic language. When aligned with Norwegian, also a Germanic language, we obtain 17.7M bitexts. The pair Danish-Italian, however, has only 14.7M bitexts although Italian has almost six times more sentences than Norwegian. One one hand, a possible explanation could be that LASER alignments are more reliable for languages which are very similar, i.e. in the same language family. On the other hand, it may also be that people which live in nearby countries have similar interests which increases the chance to find translations on the Web. Additional analysis and examples are provided in the Appendix.

5 Evaluation on Translation Benchmarks

To assess our mined bitext, we train NMT systems only on our mined data and evaluate on several public benchmarks. We do not use any of the training data provided with these corpora, so do not use any available human translated data, and have no guarantee our bitext covers the same domain as the test sets. Nevertheless, we show on the many to many TED corpus that our mined data produces high quality translation systems, even through distant language pairs not aligned through English and low resource languages. Finally, we demonstrate that models trained on CCMatrix can surpass state of the art systems in WMT'19 and WAT'20.

5.1 TED Evaluation

We examine the quality of our mined bitext across a diverse set of languages, focusing on performance of bitext pairs not aligned through English. Following Gottschalk and Demidova (2017), we evaluate on the test sets of the TED corpus (Qi et al., 2018), which contains parallel TED talk transcripts in 58 languages. This corpus is tokenized, so we detokenize using Moses, with the exception of pairs

ISO	Name	Size	ko	vi	zh	af	da	de	en	is	nl	no	sv	ca	es	fr	gl	it	pt	ro	be	bg	cs	hr	pl	ru	sk	sl	sr	uk	el	et	fi	hu	lt	lv	eu	sq	tr	ar	he	fa	sw	bn	hi	mr	ne	si	ur	ta	id	mg	ms	ml	tl	Total
ja	**Japanese**	**3929**	7.2	3.4	12.4	-	4.1	14.5	40.9	-	12.4	4.0	12.2	-	15.4	15.0	-	8.5	8.8	4.1	-	4.1	10.9	-	11.6	9.5	4.1	3.5	3.4	2.0	3.9	-	4.0	4.2	2.8	-	-	-	13.0	3.8	7.1	2.3	0.9	1.7	1.8	-	-	-	-	0.7	7.7	-	1.7	1.1	0.3	**268.7**
ko	**Korean**	**1492**	-	3.9	5.1	-	2.5	9.8	19.4	-	8.4	2.7	9.3	-	10.3	10.1	-	4.6	9.4	-	-	2.4	6.7	-	7.6	7.1	3.0	2.2	2.0	1.0	2.1	1.7	2.4	2.4	1.7	1.2	-	-	6.3	2.5	4.3	1.6	0.6	0.9	1.5	-	-	-	-	0.5	2.8	-	-	-	-	**167.1**
vi	**Vietnamese**	**2003**		-	8.0	-	4.0	12.8	50.1	-	12.2	4.4	7.4	-	34.0	16.0	-	8.8	8.2	4.8	0.2	4.7	9.9	-	5.8	8.4	5.5	3.8	3.5	1.9	8.4	2.6	7.6	8.0	5.5	1.9	-	-	7.3	8.8	6.0	5.6	2.5	2.8	3.3	-	-	-	-	1.3	9.3	-	4.2	-	1.6	**308.3**
zh	**Chinese**	**4363**			-	-	4.9	18.6	71.4	-	8.2	4.7	7.5	-	24.1	21.4	-	10.4	12.2	5.1	-	5.3	6.6	2.6	7.4	14.0	4.8	3.7	3.7	2.3	5.0	3.2	4.6	4.8	3.4	2.4	-	-	7.2	6.6	5.2	4.9	1.1	1.9	2.3	-	-	-	-	1.1	6.2	0.5	2.0	1.2	0.5	**328.3**
af	Afrikaans	15				-	1.1	3.6	8.7	0.3	2.3	1.4	1.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**21.8**
da	Danish	453					-	28.3	52.3	1.3	16.6	17.7	23.1	2.2	24.1	24.3	-	14.7	14.0	7.7	0.2	7.6	5.3	2.9	5.9	8.5	7.9	6.4	4.1	2.5	7.6	5.4	9.5	7.7	5.8	4.2	1.0	2.3	5.5	4.5	3.2	2.2	-	-	-	-	-	-	-	0.5	5.6	-	-	-	-	**358.9**
de	**German**	**5726**						-	247	3.4	68.8	23.8	43.9	-	112	137	-	71.6	53.7	21.7	0.7	20.6	32.9	8.9	44.6	45.9	25.0	18.0	12.0	7.5	19.7	12.5	20.0	23.4	13.1	9.6	-	8.6	20.5	16.0	9.2	7.4	2.4	3.7	5.7	-	-	-	-	2.3	17.0	0.7	4.0	2.4	1.1	**1286.4**
en	**English**	**14333**							-	8.7	107	47.8	77.0	21.3	409	329	13.2	146	174	55.6	1.9	44.6	56.3	18.8	74.1	140	38.1	27.4	26.5	20.2	49.3	22.0	36.0	36.4	23.3	16.7	7.8	22.4	47.0	49.7	25.2	24.6	5.8	10.1	15.1	2.9	0.7	6.3	6.1	7.3	70.5	1.7	10.7	6.8	3.1	**2833.5**
is	Islandic	32								-	1.7	1.3	1.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**18.5**
nl	**Dutch**	**1712**									-	14.0	23.8	-	48.6	38.1	-	29.3	25.8	11.7	0.3	10.5	25.5	-	30.1	25.7	11.8	8.9	5.9	3.3	11.0	6.9	11.4	11.6	7.4	5.5	-	-	18.9	7.7	9.5	7.4	1.2	1.8	2.8	-	-	-	-	1.0	9.4	-	-	-	-	**663.9**
no	Norwegian	436										-	23.3	2.0	20.6	20.7	-	12.7	12.0	6.3	0.2	6.1	8.0	2.5	5.0	7.5	6.7	5.5	4.5	2.5	5.9	4.2	8.1	6.2	4.6	3.4	0.9	2.3	6.1	4.6	3.3	2.3	-	-	-	-	-	-	-	0.7	6.1	-	-	-	-	**326.4**
sv	**Swedish**	**1733**											-	-	38.0	36.6	-	21.1	31.1	9.2	0.3	8.9	22.6	-	27.0	21.4	12.1	9.4	5.7	2.9	9.1	6.2	18.1	10.0	6.5	4.8	-	3.8	19.4	6.6	8.8	3.3	0.9	1.6	2.6	-	-	-	-	0.8	7.9	-	-	-	-	**587.2**
ca	Catalan	162												-	65.4	24.4	3.0	11.4	12.7	5.8	-	2.0	-	-	-	-	2.0	1.7	1.4	1.1	2.5	1.2	1.8	2.0	1.4	1.0	1.1	1.2	-	1.6	1.3	1.1	0.3	0.6	0.6	-	-	-	-	0.4	1.8	-	-	-	-	**176.4**
es	**Spanish**	**5268**													-	286	32.2	148	198	57.1	-	9.6	33.4	-	43.6	55.9	22.5	16.3	12.6	8.6	45.2	11.4	34.1	37.8	23.7	9.2	6.6	10.6	24.3	35.4	22.0	19.2	6.9	3.9	11.8	-	-	-	-	9.1	39.9	-	-	-	-	**2077.5**
fr	**French**	**5684**														-	17.2	148	128	58.6	-	10.3	34.2	9.2	44.0	48.3	22.7	16.6	12.3	7.6	44.4	11.8	36.3	38.7	24.6	9.3	-	-	24.2	24.4	23.2	8.3	8.1	10.1	11.4	-	-	-	-	8.2	21.9	-	4.4	2.3	1.1	**1826.6**
gl	Galician	43															-	4.0	9.9	1.8	-	-	-	-	-	-	-	-	-	-	4.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**85.8**
it	Italian	2441																-	69.6	29.7	0.4	14.0	10.0	-	13.0	18.6	13.9	10.7	7.0	2.7	16.0	7.1	11.4	13.5	8.3	5.8	1.9	5.9	11.1	11.0	-	4.5	1.2	1.8	3.2	-	-	-	-	1.1	10.8	-	-	-	-	**943.3**
pt	**Portuguese**	**2548**																	-	18.6	-	6.7	25.6	-	33.9	27.1	13.0	9.6	7.0	4.7	23.3	6.7	10.8	12.6	8.0	5.5	2.0	-	20.9	18.3	11.3	9.8	3.6	5.2	6.4	-	-	-	-	4.6	20.8	-	-	-	-	**1083.5**
ro	Romanian	566																		-	0.3	10.5	10.4	3.7	12.9	-	8.8	6.7	5.1	3.4	9.6	5.0	6.4	9.2	6.2	4.4	-	3.4	5.5	5.2	3.2	2.7	-	1.2	1.8	-	-	-	-	0.7	6.0	-	-	-	-	**430.1**
be	Belarusian	25																			-	0.9	1.5	0.1	2.3	4.0	0.9	0.6	0.5	0.9	0.2	0.2	0.2	0.4	0.2	0.1	-	0.1	2.0	-	-	-	0.0	0.1	0.1	-	-	-	-	-	-	-	-	-	-	**19.6**
bg	Bulgarian	271																				-	11.1	4.1	19.5	32.3	13.3	9.9	8.7	8.1	9.5	5.6	6.5	8.3	7.1	4.9	0.9	3.2	5.4	2.7	3.3	-	0.6	1.2	1.9	-	-	-	-	0.6	3.5	-	-	-	-	**340.9**
cs	**Czech**	**986**																					-	4.5	31.8	28.0	37.4	9.9	6.1	4.0	9.1	6.7	8.9	11.3	7.5	5.3	-	3.2	14.4	5.5	3.8	2.9	0.7	1.4	2.0	-	-	-	-	0.7	6.2	-	-	-	-	**522.0**
hr	Croatian	81																						-	5.2	7.7	3.8	4.2	4.9	1.8	-	2.4	2.5	3.3	2.9	2.0	-	1.6	-	1.8	-	1.2	-	-	-	-	-	-	-	-	0.6	-	-	-	-	**103.1**
pl	**Polish**	**1598**																							-	36.6	26.0	16.4	11.6	5.8	10.4	7.0	10.0	12.6	9.2	6.0	-	4.0	17.3	7.2	8.8	7.1	1.0	1.6	2.4	-	-	-	-	0.9	13.7	-	-	-	-	**640.9**
ru	**Russian**	**4255**																								-	21.7	15.8	14.2	26.5	12.8	10.1	10.3	12.4	14.8	10.1	-	5.8	11.0	11.9	9.5	5.5	1.4	2.6	3.5	0.7	0.2	1.4	1.4	1.5	14.1	0.4	2.5	1.6	0.6	**770.9**
sk	Slovak	395																									-	14.4	8.6	6.1	7.6	5.8	7.4	10.8	7.0	5.1	1.1	3.1	6.6	5.0	3.2	2.2	0.6	1.1	1.8	-	-	-	-	0.5	5.0	-	-	-	-	**408.0**
sl	Slovenian	186																										-	7.8	4.5	6.1	5.3	6.0	7.4	6.2	5.1	0.9	2.5	4.7	3.7	2.4	1.8	0.5	0.9	1.7	-	-	-	-	0.5	4.1	-	-	-	-	**293.4**
sr	Serbian	165																											-	4.6	4.6	3.4	3.6	4.3	3.9	3.3	0.8	2.9	4.1	3.3	2.4	1.9	0.5	0.9	1.8	-	-	-	-	0.5	4.2	-	-	-	-	**229.9**
uk	Ukrainian	481																												-	2.7	2.0	2.3	2.6	3.0	1.9	-	1.5	2.4	2.1	1.8	1.1	0.3	0.6	0.9	-	-	-	-	0.3	2.2	-	-	-	-	**164.1**
el	**Greek**	**578**																													-	4.7	6.3	7.6	5.4	4.1	2.5	3.2	10.8	3.9	6.3	2.6	1.5	3.0	3.3	-	-	-	-	1.8	9.0	-	-	-	-	**401.7**
et	Estonian	92																														-	6.0	5.5	5.6	4.1	1.5	1.4	3.3	2.7	2.0	1.5	0.8	0.8	1.4	-	-	-	-	0.4	3.4	-	-	-	-	**201.3**
fi	**Finnish**	**477**																															-	7.8	5.4	3.9	2.4	2.0	5.3	4.0	5.6	2.0	1.4	2.8	3.1	-	-	-	-	1.8	8.1	-	-	-	-	**348.3**
hu	**Hungarian**	**579**																																-	6.3	4.4	2.5	2.5	6.2	4.3	5.7	2.1	1.4	3.0	3.2	-	-	-	-	1.8	4.5	-	-	-	-	**370.6**
lt	**Lithuanian**	**159**																																	-	5.3	1.6	1.6	7.4	3.2	3.9	1.6	0.9	2.1	2.5	-	-	-	-	1.4	3.7	0.5	-	-	-	**266.1**
lv	Latvian	85																																		-	1.4	1.2	2.6	2.2	1.6	1.0	0.8	0.6	1.2	-	-	-	-	0.3	2.5	-	-	-	-	**165.7**
eu	Basque	32																																			-	0.9	3.8	2.0	1.8	1.8	-	-	-	-	-	-	-	-	-	-	-	-	-	**47.0**
sq	Albanian	84																																				-	7.5	2.6	1.5	2.0	-	-	-	-	-	-	-	-	-	-	-	-	-	**114.6**
tr	**Turkish**	**2157**																																					-	5.7	7.6	4.0	3.4	5.1	5.5	1.8	4.5	3.7	3.2	3.3	6.6	-	4.1	3.4	1.5	**422.4**
ar	**Arabic**	**786**																																						-	3.7	4.5	1.8	3.1	3.6	1.1	2.2	2.0	1.7	1.9	9.6	-	2.7	2.0	0.9	**320.9**
he	**Hebrew**	**263**																																							-	4.1	1.0	2.3	2.8	1.0	1.2	1.9	0.9	1.9	6.1	0.5	1.9	1.9	0.8	**246.0**
fa	**Farsi**	**778**																																								-	1.6	2.6	2.8	0.9	1.7	1.6	1.9	1.8	6.4	-	2.2	1.7	0.7	**184.1**
sw	**Swahili**	**9**																																									-	0.9	1.0	-	-	-	-	1.4	2.4	-	-	-	-	**61.3**
bn	**Bengali**	**136**																																										-	1.3	0.2	0.0	0.4	0.5	0.3	3.3	0.4	0.5	0.5	0.1	**91.4**
hi	**Hindi**	**0**																																											-	0.5	0.1	0.4	0.8	0.4	4.0	0.4	0.6	0.5	0.1	**125.7**
mr	Marathi	23																																													-	0.0	0.1	0.1	0.1	1.3	-	-	-	-	**10.5**
ne	Nepali	19																																														-	0.0	0.0	0.0	6.3	-	-	-	-	**17.0**
si	Sinhala	32																																															-	0.3	0.2	2.3	-	-	-	-	**20.6**
ur	Urdu	35																																																-	0.3	2.0	-	-	-	-	**19.3**
ta	**Tamil**	**154**																																																	-	2.3	0.2	0.3	0.3	0.2	**68.7**
id	**Indonesian**	**1014**																																																		-	1.2	7.8	2.3	1.7	**393.6**
mg	Malagasy	1																																																			-	0.1	0.1	0.1	**6.8**
ms	Malay	45																																																				-	0.3	0.2	**50.1**
ml	Malayam	59																																																					-	0.2	**28.5**
tl	Tagalog	42																																																						-	**14.9**

Table 1: **CCMatrix**: number of extracted parallel sentences for each language pair (**all numbers in millions**) for a margin threshold of 1.06. See text for details.

	ar	bg	cs	da	de	el	en	es	et	fa	fi	fr	he	hu	id	it	ja	ko	lt	nl	no	pl	pt	ru	sv	tr	uk	vi	zh
ar	-	16.6	14.4	19.3	18.4	20.0	33.3	24.6	12.6	8.7	10.0	25.5	14.8	12.7	19.2	21.4	13.0	4.0	12.1	19.1	24.6	12.4	23.8	15.5	22.6	9.5	13.4	21.2	16.7
bg	7.2	-	22.3	30.9	26.1	27.7	41.4	28.6	20.0	-	14.8	28.7	17.5	18.0	22.1	27.8	14.4	4.8	17.4	26.3	30.0	17.0	26.8	20.6	27.9	9.8	19.0	25.0	18.0
cs	8.6	24.9	-	22.0	23.7	22.9	33.4	26.0	19.0	8.3	15.4	27.5	14.3	16.4	21.8	21.2	13.3	-	16.8	-	26.8	18.1	25.9	19.3	24.6	10.0	17.2	22.3	17.5
da	8.6	30.5	18.5	-	29.6	28.6	48.9	31.5	19.3	9.4	16.6	33.0	18.7	17.9	26.4	29.3	14.7	5.5	17.5	31.3	36.7	16.5	31.8	19.3	36.2	12.5	19.1	23.4	17.9
de	10.2	27.3	22.4	33.5	-	24.7	39.8	30.0	19.8	9.3	15.5	31.3	17.5	18.3	24.2	26.9	-	4.8	16.8	28.8	25.4	17.3	29.9	19.3	28.6	12.7	17.4	24.3	19.0
el	11.0	28.3	20.7	30.6	24.6	-	41.0	31.0	18.1	9.4	14.3	31.9	17.7	17.8	24.6	27.7	-	4.9	17.3	-	31.8	16.3	30.0	20.2	27.2	12.1	17.2	24.5	18.7
en	16.8	38.5	26.9	43.5	33.9	36.3	-	43.1	22.5	13.5	19.8	42.8	26.9	22.0	35.0	37.3	16.3	6.4	21.3	35.3	46.6	20.6	43.1	23.3	39.6	17.5	23.2	30.7	24.7
es	13.6	27.0	21.3	32.1	26.8	28.0	44.3	-	19.9	9.9	15.9	34.6	18.2	18.3	26.0	32.0	15.5	5.0	16.6	27.9	29.2	17.6	34.1	21.0	28.5	12.5	18.0	26.1	20.5
et	6.9	22.9	18.2	22.0	20.7	20.2	26.9	23.3	-	10.4	15.2	24.4	12.3	17.4	21.4	19.9	-	-	14.3	21.2	15.4	12.4	21.6	17.3	21.6	8.4	15.3	17.9	17.5
fa	8.3	-	13.0	18.1	16.3	17.4	30.9	20.1	12.8	-	7.4	22.0	10.7	11.3	18.2	18.8	-	3.7	8.6	16.5	16.2	10.0	19.7	14.0	17.1	9.1	10.9	18.4	15.2
fi	6.5	18.4	16.1	21.0	17.6	17.1	25.4	20.8	16.1	4.9	-	20.5	11.5	14.4	15.8	18.5	-	4.0	12.7	19.8	18.2	13.5	19.2	14.0	19.7	9.3	12.3	18.3	13.6
fr	11.9	26.2	21.4	32.1	26.6	28.0	42.9	33.6	19.4	9.7	15.6	-	17.8	18.0	26.6	29.7	15.0	4.9	17.8	27.9	30.0	17.6	33.2	21.4	28.7	12.6	19.1	26.4	17.1
he	11.3	25.5	18.4	27.1	23.1	23.1	39.3	27.3	16.5	8.3	12.3	28.2	-	14.6	22.6	-	-	4.5	0.0	21.9	24.5	14.3	27.4	17.2	25.1	10.3	14.9	22.1	14.7
hu	9.3	21.5	17.0	22.6	20.4	20.5	29.1	23.0	18.1	7.7	13.8	23.6	13.0	-	20.0	21.5	-	-	14.6	20.9	19.1	14.8	23.2	16.2	21.2	10.5	13.9	20.1	17.3
id	8.6	20.6	16.9	25.7	21.6	21.7	35.3	26.3	15.5	8.7	11.7	27.0	14.5	14.9	-	23.6	13.5	5.2	12.6	24.1	26.0	14.6	26.3	16.6	22.8	11.4	15.2	24.7	18.2
it	11.8	28.8	19.0	30.3	26.0	27.6	41.1	33.9	18.2	10.1	15.8	34.2	-	17.7	25.6	-	13.8	5.1	16.8	26.8	31.7	15.9	33.0	19.2	29.4	13.4	11.5	25.6	19.3
ja	4.3	9.6	6.5	9.8	-	9.3	15.5	12.2	-	4.5	-	11.5	4.6	-	10.0	9.2	-	-	-	9.1	9.5	6.6	10.5	8.5	8.6	4.5	5.5	13.1	13.9
ko	5.3	12.8	-	13.6	11.2	12.4	19.6	15.0	-	5.9	7.8	15.3	6.9	-	15.4	13.8	-	-	-	12.1	-	8.6	14.6	9.7	12.4	6.9	7.5	16.9	16.6
lt	8.2	21.7	17.5	22.2	20.8	19.4	29.8	23.1	14.4	5.9	12.9	24.8	0.0	14.7	17.7	21.2	-	-	-	20.8	21.9	15.3	23.0	19.3	23.9	9.8	16.7	20.7	16.0
nl	10.3	26.3	-	32.1	27.3	-	39.3	30.8	19.3	9.0	15.7	31.0	16.3	17.4	25.3	27.2	14.1	4.6	17.0	-	21.0	17.6	30.0	18.9	27.5	11.2	16.5	23.6	17.3
no	13.5	29.5	24.2	37.1	25.4	29.9	51.2	31.1	11.6	6.3	15.3	32.2	18.7	16.5	27.9	33.3	15.0	4.8	15.7	20.9	-	12.4	31.7	18.6	34.6	13.7	19.3	19.6	-
pl	7.6	20.2	18.1	19.9	19.0	18.2	25.3	21.9	12.9	6.6	12.5	23.4	12.0	13.9	19.0	18.4	12.1	4.0	13.9	19.7	14.5	-	21.4	16.8	19.4	8.8	15.0	19.9	15.6
pt	12.5	26.9	22.2	34.0	27.8	29.4	47.9	36.5	18.3	10.2	15.5	35.8	19.3	19.0	28.3	33.3	14.1	5.0	18.1	29.5	29.0	18.2	-	21.5	29.3	13.3	18.9	25.5	17.2
ru	8.4	21.3	17.1	19.1	19.1	20.0	26.9	23.2	16.0	8.1	11.4	23.8	12.6	14.2	18.7	19.8	14.1	3.6	15.7	19.5	19.4	15.2	22.4	-	20.6	6.8	20.9	20.5	17.3
sv	12.7	28.2	21.5	36.0	28.4	27.0	44.7	30.8	20.8	9.1	16.2	32.4	18.3	17.8	25.6	30.0	14.1	5.0	19.8	27.8	34.9	16.5	30.6	20.2	-	12.7	17.7	26.8	18.0
tr	8.9	16.6	14.3	19.5	19.0	18.6	29.2	22.1	13.7	8.4	11.1	22.4	12.2	14.0	19.5	20.9	12.8	4.6	12.1	18.5	19.1	12.0	21.4	11.7	19.4	-	6.6	20.8	16.7
uk	8.4	22.4	17.6	23.9	20.2	20.9	30.3	24.0	14.9	7.2	11.5	25.1	12.2	14.2	19.9	15.2	11.3	3.8	16.1	19.8	20.9	16.0	23.4	23.8	19.8	4.8	-	20.8	15.7
vi	8.5	19.9	13.0	20.1	18.5	18.7	29.0	21.4	12.7	7.8	11.5	23.6	11.8	12.1	21.8	20.6	12.6	4.3	11.8	17.3	15.9	12.6	20.4	15.4	20.5	9.2	13.4	-	16.2
zh	6.0	14.3	10.8	13.4	13.4	14.4	21.6	17.3	9.9	6.0	7.9	16.7	8.3	10.4	15.8	-	13.2	3.8	8.2	14.3	15.1	9.7	15.6	12.5	14.8	6.9	9.5	18.8	-

Table 2: **BLEU scores on the TED test set.** NMT systems were trained on bitexts mined in CCMatrix only.

involving Chinese, Japanese and Korean as it creates artifacts.

We consider 29 different languages, resulting in 778 NMT systems to train. We apply the same preprocessing and training procedure for all language pairs. We train a SentencePiece Model (Kudo and Richardson, 2018) with a vocabulary of size 50k. The bitext were not filtered to remove sentences which may appear in the TED dev or test sets. Also, we did not try to optimize the architecture of the NMT models to to size of the bitexts for each language pair. Instead, for all the pairs, we use the same architecture, a Transformer model with six layers for both the encoder and decoder. We use a dimension of 512 and 4096 for the feed-forward. We train each model for 50 epochs with an initial learning rate of 0.001. We keep the model with the best BLEU on the TED validation set.

In Table 2, we report tokenized BLEU on the test sets. When translating into Chinese, we scored with `sacrebleu -tok zh`, and Kytea[6] was used to tokenize Japanese, respectively. The average BLEU over all pairs is 18.8 and 33.0 for pairs with English. There are 86 pairs out of 778 with BLEU above 30, compared to 10 out of 1620 language pairs for WikiMatrix. The best WikiMatrix pair reached 37.3 BLEU (for Brazilian Portuguese to English), while here 25 pairs are over 37.3, the best pair reaching 51.2 BLEU (Norwegian to English). These results show the quality of the mined bitexts and suggest that our mining strategy is robust to the noise and domain differences existing in large corpora like Common Crawl. However, since we did not optimize the NMT systems for each language pair, these BLEU score should not be considered as the best possible ones based on the CCMatrix bitexts. In particular, we anticipate that better results can be obtained when using models with more parameters for the high-resource language pairs.

Further, our mined data provides a starting point for those interested in training translation systems directly between languages that currently have no available bitext training data. In particular CCMatrix bitexts have been used to train a massively multilingual NMT systems for 100×100 languages (Fan et al., 2020).

5.2 WMT'19 Evaluation

Next, we focus on arguably the most competitive translation benchmark, the WMT news translation task, to compare our mined data to the best existing systems. We only consider the high resource directions, as they constitute the largest challenge — existing systems perform strongly, and previous work incorporating mined data from Paracrawl (Ott et al., 2018) only found marginal gains.

We follow Ng et al. (2019) and trained systems on en-de, en-ru, en-zh, and de-fr. We used the Transformer Big architecture with FFN size 8192,

[6] http://www.phontron.com/kytea/

	System	de-en	en-de	en-ru	ru-en	zh-en	en-zh	de-fr	fr-de
Single systems	NT'18 WMT bitext	46.2	45.9	33.5	33.4	25.8	39.2	-	-
	NT'18 CCMatrix	**49.9**	**50.3**	**35.7**	**36.9**	**30.2**	**40.8**	-	-
	NT'19 WMT bitext	41.0	40.4	31.4	38.1	-	-	-	-
	NT'19 CCMatrix	**43.3**	**44.5**	**35.5**	**41.8**	34.8	35.6	**37.9**	33.5
	NT'20 WMT bitext	**40.3**	31.9	24.0	35.5	-	-	-	-
	NT'20 CCMatrix	39.2	**35.1**	**25.5**	**37.1**	35.0	38.8	33.8	33.8
Ensembles + BT + Reranking	NT'19 best	42.8	44.9	36.3	40.2	39.9	44.6	37.3	35.0

Table 3: **BLEU scores on the Newstest'18, Newstest'19 and Newstest'20 test sets**. Newstest'18 WMT bitext, Newstest'19 WMT bitext and Newstest'20 WMT bitext are the results for single models trained on parallel WMT'19 data, En-De and En-Ru using the setup from Ng et al. (2019), and En-Zh results from Sun et al. (2019). Newstest'19 best are the best BLEU scores from ensembles trained on parallel and back-translated WMT'19 data, according to http://matrix.statmt.org/.

Language pair	src	tgt
En-De	3.9%	2.2%
En-Ru	4.2%	2.5%
En-Zh	3.0%	0.7%
De-Fr	3.6%	3.1%

Table 4: **8-gram test data overlap**. Percentage of 8-gram BPE tokens from Newstest 2019 that are also found in CCMatrix training data.

embedding size 2048, with 9 encoder/decoder layers, with LayerDrop (Fan et al., 2019). We trained for 400k updates on 8 GPUs. Given the large amounts of mined bitext (see Table 1), we train only on data with a margin threshold at least 1.07, and perform some additional filtering, resulting in 146M for en-de, 78M for en-ru, 82M for de-fr and 31M for en-zh. For each direction, we learn joint source-target BPE (Sennrich et al., 2016) and share input/output embeddings. We tune training parameters on WMT'12-13 when available and on the WMT'19 dev set for de-fr.

In Table 3 we demonstrate that the performance of a single model trained on mined data is better than the performance of the best published single models trained on WMT bitext, this can be seen as a clear indicator of the quality of the mined data.

Because CCMatrix data is mined from the Web, we want to make sure there is no significant leakage of the test sets that might be available online into the training data. While there are no exact matches of test and train samples, partial overlap is still possible. Following Radford et al. (2019) and Shoeybi et al. (2019) in Table 4 we report the percentage of 8-gram BPE tokens from the test data that are also found in CCMatrix training data. Finally, in Table 3 we also report performance on Newstest'20 tests sets that were not available at the time of mining the data.

We further investigate the impact of training on a combination of human translated and mined data. We examine En-De and include the WMT'19 training data. We found that this system outperforms the system trained on CCMatrix data only on average by only 0.6 BLEU, achieving BLEU score 50.9 on newstest2018 and 45.1 on newstest2019.

5.3 WAT'20 Evaluation

Finally, we examine the quality of our mined data on low resource, distant language pairs. We focus on Russian-Japanese, a language direction in the 2020 Workshop on Asian Translation (WAT) (Nakazawa et al., 2020).The organizers provide a tiny amount of parallel data from the Global Voices domain for training (12k sentences), and a development (486 sentences) and test set (600 sentences) from the News Commentary domain, respectively.[7]

We trained an NMT system on CCMatrix Japanese-Russian mined data only, without using other resources or texts aligned with English. We applied a threshold of 1.06 on the margin which yielded 9.5 million parallel sentences. We filtered the mined bitexts to exclude all sentences which

[7] https://github.com/aizhanti/JaRuNC

System	ja-ru	ru-ja
CCMatrix dev	13.68	20.38
CCMatrix test	**14.77**	**19.60**
WAT'20 test best	14.36	18.48

Table 5: **BLEU scores on WAT'20**.

appear in the WAT dev or test set.

We use the same NMT architecture as in Section 5.1. We report tokenized BLEU in Table 5. When translating from Russian into Japanese, tokenization was performed with Kytea and then scored with `multi-bleu.perl`.

We outperform the best performing system at WAT'20,[8] in particular when translating into Japanese. On one hand, the participants in WAT were constrained to only use the provided resources. But on the other hand, Russian/English and Japanese/English were included and participants were encouraged to train multilingual models, and use techniques like monolingual pre-training or back-translation. Therefore, our results are not directly comparable, but remain a positive indicator of the quality of our mined bitexts.

6 Conclusion

We show that margin-based mining in a joint multilingual sentence embedding space can be scaled to monolingual texts of more than 71 billion unique sentences in 90 languages, including several low resource languages. This procedure yields 10.8 billion parallel sentences, out of which only 2.9 billions are aligned with English. We performed an extensive evaluation of the quality of the mined bitexts by training NMT systems for many language pairs. Training only on mined data, we outperform the best single NMT systems at WMT'19 for translations between German, Russian, and Chinese with English, as well as between German and French. We also achieve state-of-the-art BLEU scores for translation between Russian and Japanese at WAT'20.

All mined data is freely available.[9] We hope this will enable widespread research on multilingual NMT, particularly on languages where training data is not currently available.

[8] See results at `http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html`

[9] `https://github.com/facebookresearch/LASER/tree/master/tasks/CCMatrix`

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the Use of Comparable Corpora to Improve SMT performance. In *EACL*, pages 16–23.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Ahmad Aghaebrahimian. 2018. Deep neural networks at the service of multilingual parallel sentence extraction. In *Coling*.

Mikel Artetxe and Holger Schwenk. 2018a. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. *https://arxiv.org/abs/1811.01136*.

Mikel Artetxe and Holger Schwenk. 2018b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In *https://arxiv.org/abs/1812.10464*.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted Set-Theoretic Alignment of Comparable Sentences. In *BUCC*, pages 41–45.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2018. Extracting Parallel Sentences from Comparable Corpora with STACC Variants. In *BUCC*.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings. In *BUCC*.

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*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2019. A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2020. 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.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, pages 1340–1348.

Miquel Esplà-Gomis and Mikel L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 9:77–86.

Thierry Etchegoyhen and Andoni Azpeitia. 2016. Set-Theoretic Alignment for Comparable Corpora. In *ACL*, pages 2009–2018.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *JMLR*.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Simon Gottschalk and Elena Demidova. 2017. Multiwiki: Interlingual text passage alignment in Wikipedia. *ACM Transactions on the Web (TWEB)*, 11(1):6.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. https://arxiv.org/abs/1802.06893.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective Parallel Corpus Mining using Bilingual Sentence Embeddings. *arXiv:1807.11906*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv:1803.05567*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*.

H. Jégou, M. Douze, and C. Schmid. 2011. Product quantization for nearest neighbor search. *IEEE Trans. PAMI*, 33(1):117–128.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan M. 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 2: Shared Task Papers*, 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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*, pages 66–71.

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

P. Lison and J. Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC*.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *IJCNLP*, pages 553–561.

Mehdi Zadeh Mohammadi and Nasser GhasemAghaee. 2010. Building bilingual parallel corpora based on Wikipedia. In *2010 Second International Conference on Computer Engineering and Applications*, pages 264–268.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Pablo Gamallo Otero and Isaac González López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. *arXiv:1806.00187*.

Alexandre Patry and Philippe Langlais. 2011. Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *https://openai.com/blog/better-language-models/*.

Philip Resnik. 1999. Mining the Web for Bilingual Text. In *ACL*.

Philip Resnik and Noah A. Smith. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *ACL*, pages 228–234.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *http://arxiv.org/abs/1907.05791*.

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*, pages 1715–1725.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *NAACL*, pages 403–411.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 374–381, Florence, Italy. Association for Computational Linguistics.

J. Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598.

Dan Tufis, Radu Ion, Ștefan Daniel, Dumitrescu, and Dan Ștefănescu. 2013. Wikipedia as an smt training corpus. In *RANLP*, pages 702–709.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *ACL*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CCNet: extracting high quality monolingual datasets from web crawl data. *https://arxiv.org/abs/1911.00359*.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *https://arxiv.org/abs/1902.08564*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *LREC*.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the Third BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*.