WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia

Holger Schwenk Facebook AI schwenk@fb.com
Vishrav Chaudhary Facebook AI vishrav@fb.com
Shuo Sun Johns Hopkins University ssun32@jhu.edu
Hongyu Gong Univ. of Illinois Urbana-Champaign hgong6@illinois.edu
Francisco Guzmán Facebook AI fguzman@fb.com

Abstract

We present an approach based on multilingual sentence embeddings to automatically extract parallel sentences from the content of Wikipedia articles in 96 languages, including several dialects or low-resource languages. We systematically consider all possible language pairs. In total, we are able to extract 135M parallel sentences for 1620 different language pairs, out of which only 34M are aligned with English. This corpus is freely available.[1]

To get an indication on the quality of the extracted bitexts, we train neural MT baseline systems on the mined data only for 1886 languages pairs, and evaluate them on the TED corpus, achieving strong BLEU scores for many language pairs. The WikiMatrix bitexts seem to be particularly interesting to train MT systems between distant languages without the need to pivot through English.

1 Introduction

Most of the current approaches in Natural Language Processing are data-driven. The size of the resources used for training is often the primary concern, but the quality and a large variety of topics may be equally important. Monolingual texts are usually available in huge amounts for many topics and languages. However, multilingual resources, i.e. sentences which are mutual translations, are more limited, in particular when the two languages do not involve English. An important source of parallel texts is from international organizations like the European Parliament (Koehn, 2005) or the United Nations (Ziemski et al., 2016). Several projects rely on volunteers to provide translations for public texts, e.g. news commentary (Tiedemann, 2012), OpensubTitles (Lison and Tiedemann, 2016) or the TED corpus (Qi et al., 2018)

Wikipedia is probably the largest free multilingual resource on the Internet. The content of Wikipedia is very diverse and covers many topics. Articles exist in more than 300 languages. Some content on Wikipedia was human translated from an existing article into another language, not necessarily from or into English. Eventually, the translated articles have been later independently edited and are not parallel anymore. Wikipedia strongly discourages the use of unedited machine translation,[2] but the existence of such articles cannot be totally excluded. Many articles have been written independently, but may nevertheless contain sentences which are mutual translations. This makes Wikipedia a very appropriate resource to mine for parallel texts for a large number of language pairs. To the best of our knowledge, this is the first work to process the entire Wikipedia and systematically mine for parallel sentences in all language pairs.

In this work, we build on a recent approach to mine parallel texts based on a distance measure in a joint multilingual sentence embedding space (Schwenk, 2018; Artetxe and Schwenk, 2018a), and a freely available encoder for 93 languages. We approach the computational challenge to mine in almost six hundred million sentences by using fast indexing and similarity search algorithms.

The paper is organized as follows. In the next section, we first discuss related work. We then summarize the underlying mining approach. Section 4 describes in detail how we applied this approach to extract parallel sentences from Wikipedia in 1620 language pairs. In section 5, we assess the quality of the extracted bitexts by training NMT systems for a subset of language pairs and evaluate them on the TED corpus (Qi et al., 2018) for 45 languages. The paper concludes with a discussion of future research directions.

[1] https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix

[2] https://en.wikipedia.org/wiki/Wikipedia:Translation

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1351–1361
April 19 - 23, 2021.

2 Related work

There is a large body of research on mining parallel sentences in monolingual texts collections, usually named *"comparable coprora"*. Initial approaches to bitext mining have relied on heavily engineered systems often based on metadata information, e.g. (Resnik, 1999; Resnik and Smith, 2003). More recent methods explore the textual content of the comparable documents. For instance, it was proposed to rely on cross-lingual document retrieval, e.g. (Utiyama and Isahara, 2003; Munteanu and Marcu, 2005) or machine translation, e.g. (Abdul-Rauf and Schwenk, 2009; Bouamor and Sajjad, 2018), typically to obtain an initial alignment that is then further filtered. In the shared task for bilingual document alignment (Buck and Koehn, 2016), many participants used techniques based on n-gram or neural language models, neural translation models and bag-of-words lexical translation probabilities for scoring candidate document pairs. The STACC method uses seed lexical translations induced from IBM alignments, which are combined with set expansion operations to score translation candidates through the Jaccard similarity coefficient (Etchegoyhen and Azpeitia, 2016; Azpeitia et al., 2017, 2018). Using multilingual noisy web-crawls such as ParaCrawl[3] for filtering good quality sentence pairs has been explored in the shared tasks for high resource (Koehn et al., 2018) and low resource (Koehn et al., 2019) languages.

In this work, we rely on massively multilingual sentence embeddings and margin-based mining in the joint embedding space, as described in (Schwenk, 2018; Artetxe and Schwenk, 2018a,b). This approach has also proven to perform best in a low resource scenario (Chaudhary et al., 2019; Koehn et al., 2019). Closest to this approach is the research described in España-Bonet et al. (2017); Hassan et al. (2018); Guo et al. (2018); Yang et al. (2019). However, in all these works, only bilingual sentence representations have been trained. Such an approach does not scale to many languages, in particular when considering all possible language pairs in Wikipedia. Finally, related ideas have been also proposed in Bouamor and Sajjad (2018) or Grégoire and Langlais (2017). However, in those works, mining is not solely based on multilingual sentence embeddings, but they are part of a larger system. To the best of our knowledge, this work is the first one that applies the same mining approach to all combinations of many different languages, written in more than twenty different scripts. In follow up work, the same underlying mining approach was applied to a huge collection of Common Crawl texts (Schwenk et al., 2019). Hierarchical mining in Common Crawl texts was performed by El-Kishky et al. (2020).[4]

Wikipedia is arguably the largest comparable corpus. One of the first attempts to exploit this resource was performed by Adafre and de Rijke (2006). An MT system was used to translate Dutch sentences into English and to compare them with the English texts, yielding several hundreds of Dutch/English bitexts. Later, a similar technique was applied to Persian/English (Mohammadi and GhasemAghaee, 2010). Structural information in Wikipedia such as the topic categories of documents was used in the alignment of multilingual corpora (Otero and López, 2010). In another work, the mining approach of Munteanu and Marcu (2005) was applied to extract large corpora from Wikipedia in sixteen languages (Smith et al., 2010). Otero et al. (2011) measured the comparability of Wikipedia corpora by the translation equivalents on three languages Portuguese, Spanish, and English. Patry and Langlais (2011) came up with a set of features such as Wikipedia entities to recognize parallel documents, and their approach was limited to a bilingual setting. Tufis et al. (2013) proposed an approach to mine bitexts from Wikipedia textual content, but they only considered high-resource languages, namely German, Spanish and Romanian paired with English. Tsai and Roth (2016) grounded multilingual mentions to English Wikipedia by training cross-lingual embeddings on twelve languages. Gottschalk and Demidova (2017) searched for parallel text passages in Wikipedia by comparing their named entities and time expressions. Finally, Aghaebrahimian (2018) propose an approach based on bilingual BiLSTM sentence encoders to mine German, French and Persian parallel texts with English. Parallel data consisting of aligned Wikipedia titles have been extracted for twenty-three languages.[5] We are not aware of other attempts to systematically mine for parallel sentences in the textual content of Wikipedia for a large number of languages.

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

[4] http://www.statmt.org/cc-aligned/

[5] https://linguatools.org/tools/corpora/wikipedia-parallel-titles-corpora/

3 Distance-based mining approach

The underlying idea of the mining approach used in this work is to first learn a multilingual sentence embedding. The distance in that space can be used as an indicator of whether two sentences are mutual translations or not. Using a simple absolute threshold on the cosine distance was shown to achieve competitive results (Schwenk, 2018). However, it has been observed that an absolute threshold on the cosine distance is globally not consistent, e.g. (Guo et al., 2018). This is particularly true when mining bitexts for many different language pairs.

3.1 Margin criterion

The alignment quality can be substantially improved by using a margin criterion (Artetxe and Schwenk, 2018a). The margin between two candidate sentences x and y is defined as the ratio between the cosine distance between the two sentence embeddings, and the average cosine similarity of its nearest neighbors in both directions:

$$M(x,y) = \frac{\cos(x,y)}{\sum_{z \in \mathrm{NN}_k(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in \mathrm{NN}_k(y)} \frac{\cos(y,z)}{2k}} \quad (1)$$

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 used $k = 4$ in all experiments.

We follow the *"max"* strategy of Artetxe and Schwenk (2018a): the margin is first calculated in both directions for all sentences in language L_1 and L_2. We then create the union of these forward and backward candidates. Candidates are sorted and pairs with source or target sentences that were already used are omitted. We then apply a threshold on the margin score to decide whether two sentences are mutual translations or not.

The complexity of a distance-based mining approach is $O(N \times M)$, where N and M are the number of sentences in each monolingual corpus. This makes a brute-force approach with exhaustive distance calculations intractable for large corpora. The languages with the largest Wikipedia are English and German with 134M and 51M sentences, respectively. This would require 6.8×10^{15} distance calculations.[6] We show in Section 3.3 how to tackle this computational challenge.

[6] Strictly speaking, Cebuano and Swedish are larger than German, yet mostly consist of template/machine translated text

3.2 Multilingual sentence embeddings

Distance-based bitext mining requires a joint sentence embedding for all the considered languages. One may be tempted to train a bi-lingual embedding for each language pair, e.g. (España-Bonet et al., 2017; Hassan et al., 2018; Guo et al., 2018; Yang et al., 2019), but this is difficult to scale to thousands of language pairs present in Wikipedia. Instead, we chose to use one single massively multilingual sentence embedding for all languages, namely the one proposed by the open-source LASER toolkit (Artetxe and Schwenk, 2018b). Training one joint multilingual embedding on many languages at once also has the advantage that low-resource languages can benefit from the similarity to other languages in the same language family. For example, we were able to mine parallel data for several Romance (minority) languages like Aragonese, Lombard, Mirandese or Sicilian although data in those languages was not used to train the multilingual LASER embeddings. The reader is referred to Artetxe and Schwenk (2018b) for a detailed description how LASER was trained.

3.3 Fast similarity search

In this work, we use the open-source FAISS library[7] which implements highly efficient algorithms to perform similarity search on billions of vectors (Johnson et al., 2017). Our sentence representations being 1024-dimensional, all English sentences require $134{\cdot}10^6 \times 1024 \times 4 = 536$ GB of memory. Therefore, dimensionality reduction and data compression are needed for efficient search. We chose a rather aggressive compression based on a 64-bit product-quantizer (Jégou et al., 2011), and partitioning the search space in 32k cells.[8] We build one FAISS index for each language.

The compressed FAISS index for English requires only 9.2GB, i.e. more than fifty times smaller than the original sentences embeddings. This makes it possible to load the whole index on a standard GPU and to run the search in a very efficient way on multiple GPUs in parallel, without the need to shard the index. The overall mining process for German/English requires less than 3.5 hours on 8 GPUs, including the nearest neighbor search in both direction and scoring all candidates.

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

[8] FAISS index type `OPQ64,IVF32768,PQ64`, see https://github.com/facebookresearch/faiss/wiki/Faiss-indexes

4 Bitext mining in Wikipedia

For each Wikipedia article, it is possible to get the link to the corresponding article in other languages. This could be used to mine sentences limited to the respective articles. On one hand, this **local mining** has several advantages: 1) mining is very fast since each article usually has a few hundreds of sentences only; 2) it seems reasonable to assume that a translation of a sentence is more likely to be found in the same article than anywhere in the whole Wikipedia. On the other hand, we hypothesize that the margin criterion will be less efficient since one article usually has few sentences which are similar. This may lead to many sentences in the overall mined corpus of the type *"NAME was born on DATE in CITY"*, *"BUILDING is a monument in CITY built on DATE"*, etc. Although those alignments may be correct, we hypothesize that they are of limited use to train an NMT system.

The other option is to consider the whole Wikipedia for each language: for each sentence in the source language, we mine in all target sentences. This **global mining** has the advantages that we can try to align two languages even though there are only a few articles in common. A drawback is a potentially increased risk of misalignment. In this work, we chose the global mining option.

4.1 Corpus preparation

Extracting the textual content of Wikipedia articles in all languages is a rather challenging task, i.e. removing all tables, citations, footnotes or formatting markup. There are several ways to download Wikipedia content. In this study, we use the so-called `CirrusSearch dumps` since they directly provide the textual content without any meta information.[9] We downloaded this dump in March 2019. A total of about 300 languages are available, but the size obviously varies a lot between languages. We applied the following processing: 1) extract the textual content; 2) split the paragraphs into sentences; 3) remove duplicate sentences; and 4) perform language identification and remove sentences which are not in the expected language.

It should be pointed out that sentence segmentation is not a trivial task. Some languages do not use specific symbols to mark the end of a sentence, namely Thai. We are not aware of a freely available sentence segmenter for Thai and we had to exclude it. We used a freely available Python tool[10] to detect sentence boundaries. Regular expressions were used for most of the Asian languages, falling back to English for the remaining languages. This gives us 879 million sentences in 300 languages. The margin criterion to mine for parallel data requires that the texts do not contain duplicates. This removes about 25% of the sentences.[11]

L_1 (French)	*Ceci est une très grande maison*
L_2 (German)	*Das ist ein sehr großes Haus*
	This is a very big house
	Ez egy nagyon nagy ház
	Ini rumah yang sangat besar

Table 1: Illustration how sentences in the wrong language can hurt the alignment process with a margin criterion. See text for a detailed discussion.

LASER's sentence embeddings are totally language agnostic. This has the side effect that the sentences in other languages (e.g. citations or quotes) may be considered closer in the embedding space than a potential translation in the target language. Table 1 illustrates this problem. The algorithm would not select the German sentence although it is a perfect translation. The sentences in the other languages are also valid translations which would yield a very small margin. To avoid this problem, we perform language identification (LID) on all sentences and remove those which are not in the expected language. LID is performed with fasttext[12] (Joulin et al., 2016). Fasttext does not support all the 300 languages present in Wikipedia and we disregarded the missing ones (which typically have only a few sentences anyway). After deduplication and LID, we dispose of 595M sentences in 182 languages. English accounts for 134M sentences, and German with 51M sentences is the second largest language. The sizes for all languages are given in Tables 3 and 5 (in the appendix).

4.2 Threshold optimization

Artetxe and Schwenk (2018a) optimized their mining approach for each language pair on a provided corpus of gold alignments. This is not possible when mining Wikipedia, in particular when con-

[9] https://dumps.wikimedia.org/other/cirrussearch/

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

[11] The Cebuano and Waray Wikipedia were largely created by a bot and contain more than 65% of duplicates.

[12] https://fasttext.cc/docs/en/language-identification.html

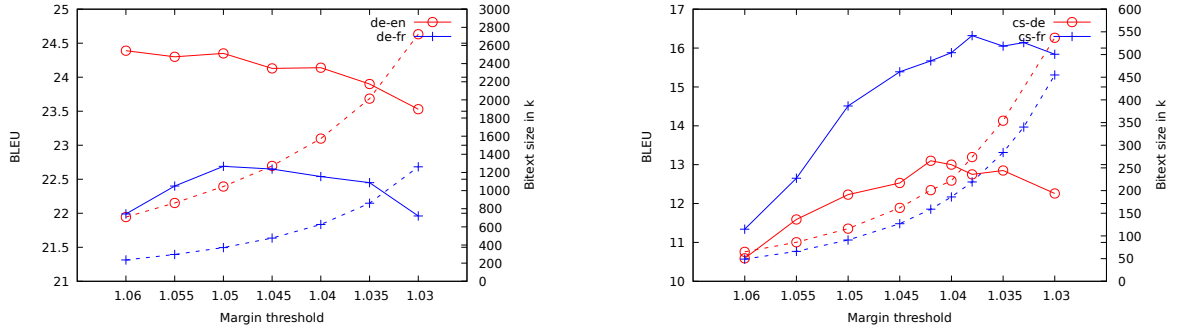

Figure 1: BLEU scores (continuous lines) for several NMT systems trained on bitexts extracted from Wikipedia for different margin thresholds. The size of the mined bitexts are depicted as dashed lines.

sidering many language pairs. In this work, we use an evaluation protocol inspired by the WMT shared task on parallel corpus filtering for low-resource conditions (Koehn et al., 2019): an NMT system is trained on the extracted bitexts – for different thresholds – and the resulting BLEU scores are compared. We choose `newstest2014` of the WMT evaluations since it provides an N-way parallel test sets for English, French, German and Czech. We favoured the translation between two morphologically rich languages from different families and considered the following language pairs: German/English, German/French, Czech/German and Czech/French. The size of the mined bitexts is in the range of 100k to more than 2M (see Table 2 and Figure 1). We did not try to optimize the architecture of the NMT system to the size of the bitexts and used the same architecture for all systems: the encoder and decoder are 5-layer transformer models as implemented in `fairseq` (Ott et al., 2019). The goal of this study is not to develop the best performing NMT system for the considered languages pairs, but to compare different mining parameters.

The evolution of the BLEU score in function of the margin threshold is given in Figure 1. Decreasing the threshold naturally leads to more mined data – we observe an exponential increase of the data size. The performance of the NMT systems trained on the mined data seems to change as expected: the BLEU score first improves with increasing amounts of available training data, reaches a maximum and than decreases since the additional data gets more and more noisy, i.e. contains wrong translations. It is also not surprising that a careful choice of the margin threshold is more important in a low-resource setting. Every additional parallel sentence is important. According to Figure 1, the optimal value of the margin threshold seems to be

Bitexts	**de-en**	**de-fr**	**cs-de**	**cs-fr**
Europarl	1.9M	1.9M	568k	627k
	21.5	23.6	14.9	21.5
	1.0M	370k	200k	220k
	21.2	21.1	12.6	19.2
Mined Wikipedia	1.0M	372k	201k	219k
	24.4	22.7	13.1	16.3
Europarl + Wikipedia	3.0M	2.3M	768k	846k
	25.5	*25.6*	17.7	24.0

Table 2: Comparison of NMT systems trained on the Europarl corpus and on bitexts automatically mined in Wikipedia by our approach at a threshold of 1.04. We give the number of sentences (first line) and the BLEU score (second line of each bloc) on `newstest2014`.

1.05 when many sentences can be extracted, in our case German/English and German/French. When less parallel data is available, i.e. Czech/German and Czech/French, a value in the range of 1.03–1.04 seems to be a better choice. Aiming at one threshold for all language pairs, we chose a value of 1.04. It seems to be a good compromise for most language pairs. However, for the open release of this corpus, we provide all mined sentence with a margin of 1.02 or better. This enables end users to choose an optimal threshold for their particular applications. However, it should be emphasized that we do not expect that many sentence pairs with a margin as low as 1.02 are good translations.

For comparison, we also trained NMT systems on the Europarl corpus V7 (Koehn, 2005), i.e. professional human translations, first on all available data, and then on the same number of sentences as the mined ones (see Table 2). With the exception of Czech/French, we were able to achieve better

BLEU scores with the mined bitexts in Wikipedia than with Europarl of the same size. Adding the mined bitexts to the full Europarl corpus, leads to further improvements of 1.1 to 3.1 BLEU.

5 Result analysis

We run the alignment process for all possible combinations of languages in Wikipedia. This yielded 1620 language pairs for which we were able to mine at least ten thousand sentences. Remember that mining $L_1 \rightarrow L_2$ is identical to $L_2 \rightarrow L_1$, and is counted only once. We propose to analyze and evaluate the extracted bitexts in two ways. First, we discuss the amount of extracted sentences (Section 5.1). We then turn to a qualitative assessment by training NMT systems for all language pairs with more than twenty-five thousand mined sentences (Section 5.2).

5.1 Quantitative analysis

Due to space limits, Table 3 summarizes the number of extracted parallel sentences only for languages which have a total of at least five hundred thousand parallel sentences (with all other languages at a margin threshold of 1.04). Additional results are given in Table 5 in the Appendix.

There are many reasons which can influence the number of mined sentences. Obviously, the larger the monolingual texts, the more likely it is to mine many parallel sentences. Not surprisingly, we observe that more sentences could be mined when English is one of the two languages. Let us point out some languages for which it is usually not obvious to find parallel data with English, namely Indonesian (1M), Hebrew (545k), Farsi (303k) or Marathi (124k sentences). The largest mined texts not involving English are Russian/Ukrainian (2.5M), Catalan/Spanish (1.6M), or between the Romance languages French, Spanish, Italian and Portuguese (480k–923k), and German/French (626k).

It is striking to see that we were able to mine more sentences when Galician and Catalan are paired with Spanish than with English. On one hand, this could be explained by the fact that LASER's multilingual sentence embeddings may be better since the involved languages are linguistically very similar. On the other, it could be that the Wikipedia articles in both languages share a lot of content, or are obtained by mutual translation.

Services from the European Commission provide human translations of (legal) texts in all the 24 official languages of the European Union. This N-way parallel corpus enables training of MT system to directly translate between these languages, without the need to pivot through English. This is usually not the case when translating between other major languages, for example in Asia. Some interesting language pairs for which we were able to mine more than 100k sentences include: Korean/Japanese (222k), Russian/Japanese (196k), Indonesian/Vietnamese (146k), or Hebrew/Romance languages (120–150k sentences).

Overall, we were able to extract at least ten thousand parallel sentences for 96 different languages. For several low-resource languages, we were able to extract more parallel sentences with other languages than English. These include, among others, Aragonse with Spanish, Lombard with Italian, Breton with several Romance languages, Western Frisian with Dutch, Luxembourgish with German or Egyptian Arabic and Wu Chinese with the respective major language.

Finally, Cebuano (ceb) falls clearly apart: it has a rather huge Wikipedia (17.9M filtered sentences), but most of it was generated by a bot, as for the Waray language.[13] This certainly explains that only a very small number of parallel sentences could be extracted. Although the same bot was also used to generate articles in the Swedish Wikipedia, our alignments seem to be better for that language.

5.2 Qualitative evaluation

Aiming to perform a large-scale assessment of the quality of the extracted parallel sentences, we trained NMT systems on the bitexts. We identified a publicly available dataset which provides test sets for many language pairs: translations of TED talks as proposed in the context of a study on pretrained word embeddings for NMT[14] (Qi et al., 2018). We would like to emphasize that we did not use the training data provided by TED – we only trained on the mined sentences from Wikipedia. The goal of this study is not to build state-of-the-art NMT system for for the TED task, but to get an estimate of the quality of our extracted data, for many language pairs. In particular, there may be a mismatch in the topic and language style between Wikipedia texts and the transcribed and translated TED talks.

NMT systems are trained with a transformer model from `fairseq` (Ott et al., 2019) with the

[13] https://en.wikipedia.org/wiki/Lsjbot

[14] https://github.com/neulab/word-embeddings-for-nmt

ISO	Name	Lang. Family	size	az	ba	be	bg	bn	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	gl	he	hi	hr	hu	id	is	it	ja	kk	ko	lt	mk	ml	mr	ne	nl	no	oc	pl	pt	ro	ru	sh	si	sk	sl	sq	sr	sv	sw	ta	te	tl	tr	tt	uk	vi	zh	total
ar	Arabic	Arabic	**6516**	17	15	11	54	40	34	94	67	53	99	66	999	37	174	40	24	58	53	163	50	68	38	38	60	90	18	123	83	11	48	33	52	32	32	12	73	58	9	74	157	71	125	35	32	32	39	30	49	58	13	27	27	18	69	14	70	93	86	**4435**
az	Azerbaijani	Turkic	**1873**	7	3	14	8	9	15	17	11	34	10	71	9	31	10	7	16	14	29	10	12	9	10	15	17	4	25	23	6	11	9	8	5	4	2	22	13	2	22	25	14	47	9	3	9	8	9	10	19	5	11	5	2	42	8	19	10	19	**885**
ba	Bashkir	Turkic	**536**		2	14	4	9	17	16	13	27	10	28	9	28	8	6	9	12	29	12	9	4	10	12	12	3	26	13	4	5	7	9	1	1	1	21	15	3	19	24	15	42	10	1	10	10	7	10	20	4	3	1	1	11	8	15	7	10	**697**
be	Belarusian	Slavic	**1690**			16	4	7	16	14	9	20	8	33	9	28	7	5	5	10	24	9	10	3	9	11	9	3	24	12	2	5	7	8	2	2	1	14	10	2	19	23	13	161	8	2	8	7	5	10	16	2	3	2	1	8	2	80	8	8	**803**
bg	Bulgarian	Slavic	**3327**				38	34	76	79	53	132	62	357	40	122	43	25	37	61	117	43	58	30	47	68	60	17	102	71	11	38	42	86	29	35	13	84	58	9	96	114	69	270	41	31	43	46	26	65	63	12	21	23	21	56	12	126	60	60	**3770**
bn	Bengali	Indo-Aryan	**1412**					21	41	47	33	70	36	280	27	81	26	14	20	37	68	27	34	21	23	41	36	8	64	38	4	20	21	23	7	8	3	50	35	4	52	76	46	62	20	8	25	26	17	25	54	6	12	8	4	33		37	31	31	**1991**
bs	Bosnian	Slavic	**1060**						44	43	32	71	33	210	24	70	25	16	20	36	60	32	30	16	164	39	38	11	52	36	6	22	23	39	19	20	8	45	36	6	48	62	37	59	178	16	25	34	19	130	38	7	15	12	13	33	8	39	38	31	**2348**
ca	Catalan	Romance	**8332**							100	86	180	90	1205	81	1580	54	77	44	83	490	268	84	37	57	92	107	23	316	103	12	52	45	61	45	56	17	144	102	57	121	358	110	169	52	52	50	57	34	67	102	14	30	35	31	77	16	98	106	90	**8047**
cs	Czech	Slavic	**7634**								75	233	70	519	75	181	62	36	45	95	185	54	72	38	63	105	78	23	161	105	15	53	55	51	36	41	15	139	86	11	176	153	82	186	50	37	474	64	30	63	97	15	31	34	25	75	14	104	74	80	**5325**
da	Danish	Germanic	**2881**									180	54	436	39	140	45	26	29	75	142	44	55	25	43	69	63	20	115	76	9	37	35	39	30	35	12	110	303	8	89	123	70	109	37	32	39	40	23	43	168	11	20	21	23	55	11	62	68	57	**4006**
de	German	Germanic	**50944**										95	1573	186	418	106	53	66	163	626	80	109	57	87	192	107	34	388	217	23	82	78	64	51	58	21	472	207	17	285	294	129	368	68	50	94	106	51	81	216	20	58	57	32	127	23	165	107	134	**9692**
el	Greek	Hellenic	**3211**											620	39	145	41	23	35	55	137	48	56	26	43	64	73	15	119	69	6	35	34	52	27	32	9	76	60	8	77	144	78	114	38	31	35	46	28	52	62	11	16	18	20	56	9	68	75	62	**3728**
en	English	Germanic	**134431**												298	3377	243	119	303	375	2757	446	545	231	259	488	1019	85	2126	851	20	306	157	395	71	124	15	796	636	37	668	2461	631	1661	224	115	178	318	180	395	546	51	95	91	75	477	32	681	1073	786	**34321**
eo	Esperanto	constructed	**2371**													149	31	25	23	46	134	46	39	22	29	57	46	15	101	48	7	26	28	30	28	28	9	81	47	9	77	91	43	81	26	28	41	32	19	36	53	8	16	19	17	37	8	50	42	39	**2817**
es	Spanish	Romance	**25202**														89	154	83	155	905	610	153	71	94	167	198	42	671	219	26	108	76	92	65	98	25	272	181	35	235	923	183	393	81	84	81	93	53	107	181	21	57	71	48	147	26	187	206	174	**14980**
et	Estonian	Uralic	**2303**															22	24	70	85	32	39	20	33	56	41	14	75	57	6	29	35	32	20	21	8	72	49	7	73	76	48	96	27	19	34	35	18	34	58	9	16	16	15	43	7	56	40	44	**2511**
eu	Basque	Isolate	**2259**																14	33	65	43	25	13	21	35	27	10	54	33	5	18	19	19	11	10	5	44	29	6	43	58	30	47	19	9	20	20	12	21	38	6	13	11	8	29	5	30	25	23	**1699**
fa	Farsi	Iranian	**3954**																	34	71	25	36	20	24	39	46	9	64	46	8	26	20	27	11	10	6	49	32	4	50	77	40	72	21	9	21	24	17	30	42	8	21	12	7	42	9	41	38	42	**2122**
fi	Finnish	Uralic	**7428**																		156	47	64	28	48	90	63	22	131	87	9	43	47	40	29	30	10	126	86	10	119	131	69	139	39	27	50	46	25	46	126	12	23	24	21	72	10	76	60	64	**3998**
fr	French	Romance	**34494**																			154	136	60	85	164	161	38	744	214	24	89	71	83	62	83	25	331	166	124	255	558	206	410	72	74	83	86	48	92	186	19	56	65	42	130	26	170	165	157	**12441**
gl	Galician	Romance	**2221**																				41	21	33	50	56	14	120	50	9	28	27	35	29	39	11	66	52	17	65	227	56	84	30	36	29	33	20	39	54	8	15	17	22	43	12	51	58	46	**3969**
he	Hebrew	Semitic	**6962**																					28	41	65	63	17	121	82	7	43	35	42	26	25	9	86	64	8	84	133	67	131	35	21	36	38	23	45	67	10	21	25	13	54	8	73	66	62	**3666**
hi	Hindi	Indo-Aryan	**1353**																						21	33	31	7	56	35	3	18	16	24	6	11	12	40	27	4	44	63	35	56	17	6	18	21	14	22	40	7	13	18	3	29	4	33	26	30	**1781**
hr	Croatian	Slavic	**2229**																							58	47	14	80	48	8	27	31	52	24	24	10	65	48	7	71	85	51	85	666	19	35	53	21	205	56	9	16	16	17	42	9	55	46	42	**3629**
hu	Hungarian	Uralic	**7702**																								70	20	146	99	11	49	48	47	27	28	10	121	75	10	126	148	87	149	46	26	56	55	27	53	88	13	29	30	20	75	9	83	74	75	**4357**
id	Indonesian	Malayo-Pol.	**3899**																									16	146	77	8	45	33	55	25	23	10	101	83	8	93	204	94	127	43	23	37	46	32	56	79	13	24	19	21	79	11	73	146	83	**4743**
is	Icelandic	Germanic	**488**																										31	18	2	9	12	12	6	5	3	27	22	3	27	35	20	30	13	4	13	13	9	13	28	4	7	5	5	16	3	18	16	14	**995**
it	Italian	Romance	**21025**																											179	24	83	62	73	58	78	24	240	150	20	219	480	161	303	67	68	72	81	48	83	153	19	44	58	41	112	23	144	143	137	**10084**
ja	Japanese	Japonic	**31614**																												14	222	47	48	21	23	8	123	81	9	128	175	79	196	40	19	48	50	28	51	96	12	37	31	12	84	12	92	75	267	**5382**
kk	Kazakh	Turkic	**1684**																													6	5	6	2	2	1	18	11	1	17	22	12	32	8	1	8	7	6	8	16	3	3	1	1	10	4	14	6	9	**581**
ko	Korean	Koreanic	**5400**																														23	26	10	10	4	56	41	5	63	93	47	89	23	9	25	26	17	29	51	7	17	13	5	47	5	48	49	57	**2539**
lt	Lithuanian	Baltic	**2031**																															28	16	16	6	57	39	6	70	64	39	107	25	15	30	30	16	29	46	8	13	11	12	36	6	57	33	35	**2091**
mk	Macedonian	Slavic	**1487**																																21	22	9	53	46	6	56	93	56	88	52	19	29	39	25	106	48	9	14	15	16	43	8	57	57	45	**2761**
ml	Malayalam	Dravidian	**272**																																	3	1	41	32	7	41	58	35	46	21	2	22	21	13	20	44	6	4	3	1	16	2	26	15	17	**1388**
mr	Marathi	Indo-Aryan	**503**																																		1	46	35	5	47	86	47	50	22	3	25	25	13	19	56	4	5	3	1	17	1	26	14	17	**1648**
ne	Nepali	Indo-Aryan	**235**																																			17	13	4	17	21	14	19	10	1	10	10	6	9	17	3	1	1		7	1	11	6	6	**584**
nl	Dutch	Germanic	**14091**																																				133	13	177	218	96	199	53	42	66	64	34	61	151	16	37	35	29	90	18	104	84	88	**6364**
no	Norwegian	Germanic	**1364**																																					9	103	161	74	121	42	28	43	51	26	47	270	12	24	23	24	61	13	69	79	63	**4771**
oc	Occitan	Romance	**472**																																						13	24	11	14	6	5	6	6	4	7	12	2	3	3	3	7	3	9	8	7	**706**
pl	Polish	Slavic	**16270**																																							200	97	285	56	40	81	68	35	69	121	16	39	42	28	92	16	172	84	92	**6035**
pt	Portuguese	Romance	**13354**																																								177	312	74	76	71	85	47	101	155	20	42	54	45	140	23	156	213	165	**10932**
ro	Romanian	Romance	**3504**																																									136	44	43	42	49	30	58	75	15	23	27	29	72	13	82	96	74	**4515**
ru	Russian	Slavic	**32537**																																										70	42	85	78	44	114	140	17	54	55	29	119	25	2486	122	148	**11311**
sh	Serbo-Croat	Slavic	**2069**																																											17	27	46	19	373	46	8	14	13	17	35	9	45	42	36	**3337**
si	Sinhala	Indo-Aryan	**320**																																												22	22	10	15	48	4	4	4	2	15	1	21	14	16	**1464**
sk	Slovak	Slavic	**1904**																																														36	17	34	51	8	14	15	16	37	9	51	38	38	**2726**
sl	Slovenian	Slavic	**1836**																																															19	47	50	9	15	16	17	39	8	50	48	42	**2623**
sq	Albanian	Albanian	**740**																																																25	31	6	13	10	10	27		29	30	24	**1515**
sr	Serbian	Slavic	**4226**																																																	51	9	19	18	14	43	8	71	56	45	**3510**
sv	Swedish	Germanic	**17906**																																																		16	33	39	35	72	17	82	74	73	**5107**
sw	Swahili	Niger-Congo	**235**																																																			6	3	4	12		12	10	11	**652**
ta	Tamil	Dravidian	**1629**																																																				6	2	29	5	30	19	27	**1315**
te	Telugu	Dravidian	**1509**																																																					1	21	1	30	16	20	**1315**
tl	Tagalog	Malayo-Pol.	**312**																																																						12	1	16	17	10	**1015**
tr	Turkish	Turkic	**4067**																																																							10	67	77	69	**3698**
tt	Tatar	Turkic	**449**																																																								11	7	10	**603**
uk	Ukrainian	Slavic	**12871**																																																									73	72	**7043**
vi	Vietnamese	Vietic	**6727**																																																										89	**4660**
zh	Chinese	Chinese	**12306**																																																											**4415**

Table 3: WikiMatrix: number of extracted sentences for each language pair (in thousands), e.g. Es-Ja=219 corresponds to 219,260 sentences (rounded). The column *"size"* gives the number of lines in the monolingual texts after deduplication and LID.

Src/Trgar	bg	bs	cs	da	de	el	en	eo	es	frfr-ca	gl	he	hr	hu	id	it	ja	ko	mk	nb	nl	pl	ptpt-br	ro	ru	sk	sl	sr	sv	tr	uk	vizh-cnzh-tw		
ar	4.9	1.8	3.0	4.5	3.8	6.7	20.3	4.1	13.2	12.2	9.0	5.6	3.5	2.2	2.7	9.2	9.9	4.2	5.3	5.5	4.9	4.4	3.0	12.0	12.2	5.6	5.6	1.5	2.7	1.2	4.0	2.4	4.5	12.3	8.2	4.9
bg	3.0	4.2	6.7	8.7	8.5	10.0	25.3	7.7	16.3	14.7	11.9	8.4	3.2	6.4	4.7	10.3	12.2	4.0	5.4	16.2	8.1	7.7	6.3	14.7	15.4	9.8	12.4	4.0	6.4	2.7	7.1	2.4	10.4	11.7	6.4	4.7
bs	1.2	6.1	4.1	3.7	5.8	4.5	21.7	9.9	6.7	4.6	5.7	1.0	30.9	2.4	5.3	6.5	1.4	12.9	3.8	2.9	9.9	10.9	5.7	4.8	3.1	10.4	10.6	4.4	1.5	5.4	5.8	2.8	1.8			
cs	1.9	7.8	3.7	7.1	8.3	6.4	20.0	10.4	12.6	11.4	8.6	5.0	2.5	6.5	4.9	7.8	9.6	4.1	5.5	5.0	6.3	7.6	8.1	10.8	12.1	6.3	9.4	28.1	6.7	1.6	7.0	2.6	7.8	9.0	6.6	4.7
da	2.0	8.9	4.0	5.2	14.0	9.0	32.9	6.7	16.7	16.7	12.8	7.3	3.5	4.4	4.7	10.8	13.4	4.7	6.0	6.2	33.1	12.4	4.8	14.0	16.2	8.5	7.8	3.1	5.2	1.4	25.8	2.7	6.3	11.2	7.3	4.9
de	2.4	9.7	4.9	8.1	16.9	7.8	24.5	15.9	17.4	18.3	14.7	6.8	4.3	5.5	7.2	8.6	13.5	6.4	6.6	5.6	11.5	17.6	6.8	14.2	15.2	8.7	9.2	5.4	8.6	1.5	12.7	3.6	7.8	11.3	9.2	4.3
el	4.1	11.2	4.8	5.5	9.7	6.7	27.9	8.0	18.8	16.3	13.5	10.1	4.0	5.6	5.3	13.1	15.3	5.5	6.1	10.2	9.3	8.7	5.1	18.0	18.3	10.4	8.4	3.1	6.4	2.0	7.2	3.4	7.2	14.4	8.7	5.3
en	11.9	23.9	14.7	15.5	30.9	20.4	27.1	22.6	35.8	32.6	25.1	24.3	17.3	18.8	13.5	28.8	29.5	10.2	18.6	21.8	31.8	25.1	12.0	31.4	37.0	20.4	17.4	13.8	16.5	5.5	29.1	10.3	17.6	26.9	18.0	10.7
eo	1.8	6.4	7.3	7.4	13.5	8.1	23.1	16.1	17.6	12.7	10.5	1.9	3.0	4.8	9.2	13.9	2.5	4.8	6.2	7.2	12.1	6.7	12.4	16.0	6.9	8.6	7.4	5.6	0.6	8.8	1.8	5.4	7.7	5.2	3.6	
es	6.2	14.3	6.8	8.0	15.7	12.9	16.4	33.2	13.5	25.6	19.9	30.1	8.1	8.7	7.7	16.1	23.8	7.9	9.9	11.9	13.3	14.1	8.1	27.6	27.8	14.7	11.6	6.0	8.1	2.6	13.7	5.2	10.0	17.8	12.3	6.6
fr	6.0	12.7	5.0	8.3	15.6	14.5	15.4	31.6	18.6	26.4	17.2	6.8	6.9	7.7	15.0	24.6	7.3	8.7	10.7	13.1	15.9	8.1	24.2	26.0	16.5	12.7	5.8	7.1	2.1	13.8	4.2	9.8	17.0	11.1	6.6	
fr-ca	4.9	12.5	2.8	7.8	14.3	12.9	15.4	27.8	14.0	23.7	18.1	6.8	6.8	7.9	13.7	23.4	7.5	8.8	10.3	15.1	7.2	18.6	23.2	15.3	11.3	6.1	6.6	3.3	12.5	5.0	9.7	15.4	6.2			
gl	2.6	7.3	4.7	2.9	7.4	5.3	8.5	23.4	9.2	34.4	16.0	15.2	2.5	3.5	2.8	9.8	19.3	3.6	4.2	5.6	6.9	6.5	4.3	22.4	23.7	7.7	5.9	1.7	3.1	0.2	4.3	2.3	3.9	9.4	5.8	4.2
he	3.4	5.7	1.8	3.7	6.1	5.4	6.3	25.7	4.2	15.3	13.6	10.3	5.1	2.5	3.2	8.9	11.2	4.1	4.5	4.6	5.4	6.5	3.7	13.7	15.0	6.3	8.0	1.9	2.6	1.0	5.2	1.8	5.1	10.7	6.9	4.6
hr	1.6	8.7	29.9	7.0	6.5	5.8	6.6	24.4	4.4	12.6	9.9	7.8	5.7	1.5	4.3	8.2	10.4	1.8	4.0	14.1	5.3	6.1	5.4	12.1	12.6	7.3	8.3	4.6	12.6	11.7	6.0	1.9	6.9	9.4	4.8	2.9
hu	1.6	5.6	2.6	5.9	7.0	5.5	16.7	6.5	10.8	10.9	9.3	3.9	2.1	3.6	6.6	8.2	4.4	6.2	3.9	4.5	6.0	4.3	9.7	9.9	7.1	5.8	3.4	4.2	1.2	5.3	3.0	4.4	8.5	6.5	4.2	
id	4.1	9.1	4.2	5.2	10.1	6.7	11.1	24.9	8.2	16.4	15.1	11.1	9.9	5.1	5.6	5.2	12.7	5.8	9.1	7.0	10.0	9.4	5.5	14.6	16.7	9.8	8.1	3.6	5.6	1.8	9.3	4.6	7.3	18.5	11.0	6.2
it	5.3	11.7	5.0	6.9	13.1	11.5	14.5	30.0	13.9	26.4	24.9	20.0	19.3	6.2	7.0	6.7	14.0	7.3	9.0	9.9	13.3	12.8	7.3	22.8	24.9	13.3	10.2	5.4	7.1	2.3	11.9	4.6	8.8	15.3	10.7	5.8
ja	1.4	1.9	0.7	1.8	3.1	2.7	2.5	7.9	2.2	6.0	6.0	4.7	2.3	1.4	1.3	2.0	3.5	4.6	16.9	1.6	2.6	3.1	1.8	5.1	4.9	2.8	2.7	1.2	1.8	0.5	2.6	1.9	2.2	5.9		
ko	0.9	1.7	1.3	2.0	1.7	1.7	8.7	1.3	4.7	4.4	3.4	1.5	0.9	0.7	1.5	3.1	3.2	9.2	1.2	1.3	1.9	1.4	4.3	4.6	2.0	2.1	1.0	1.5	0.4	1.5	1.4	1.5	4.3			
mk	2.4	18.2	12.0	5.4	7.2	4.3	10.3	23.4	8.9	15.2	11.5	10.0	5.4	4.0	12.6	3.7	8.9	11.1	3.5	4.7	7.5	6.7	4.5	13.9	15.6	8.3	7.5	3.5	7.1	3.7	5.6	2.0	6.4	10.9	6.3	4.3
nb	2.7	8.5	5.4	32.7	9.8	8.9	35.1	9.5	17.0	14.6	8.0	3.1	3.9	4.5	9.9	15.0	5.5	5.7	6.7	4.7	14.2	9.5	7.5	3.2	3.9	0.6	26.3	1.9	6.2	7.8						
nl	2.4	8.2	2.9	5.9	14.2	16.1	8.4	26.5	13.4	16.8	16.7	13.5	7.3	3.9	4.8	5.3	11.4	13.3	5.2	5.9	5.9	5.3	13.8	15.4	7.8	7.6	4.1	5.1	1.6	11.1	3.2	6.1	10.6	8.0	5.1	
pl	1.8	7.4	2.9	8.2	6.6	6.5	5.4	15.1	7.5	11.4	11.3	8.6	5.2	2.3	4.8	4.0	7.5	8.6	4.2	5.7	5.2	4.1	6.2	9.6	9.9	6.2	9.5	6.2	5.3	1.5	5.6	2.4	8.9	7.7	6.3	3.6
pt	7.4	15.2	7.0	8.0	15.5	13.2	18.7	35.0	12.0	32.4	26.7	19.0	23.0	8.8	10.3	8.4	17.5	24.4	7.8	10.6	13.1	14.3	14.9	8.9	15.3	11.7	6.6	8.1	2.0	14.3	6.2	9.3	18.0	12.9	5.8	
pt-br	6.5	14.7	7.4	8.6	16.8	12.9	17.6	37.3	16.0	31.0	26.6	20.3	23.0	8.7	9.8	8.1	18.6	24.8	7.8	10.7	12.5	14.8	8.5	15.1	11.8	6.4	8.9	2.8	14.6	5.3	10.8	18.8	13.2	6.7		
ro	3.2	9.7	3.7	5.1	9.4	7.5	10.4	25.0	6.7	18.8	19.3	14.6	10.0	4.0	5.7	5.9	11.0	15.5	4.3	6.4	7.3	8.0	8.1	5.2	15.4	17.7	8.0	3.6	5.0	1.9	7.0	3.3	6.6	12.7	7.8	4.9
ru	3.3	12.6	4.2	7.7	8.5	8.8	8.3	18.7	9.9	14.3	14.5	11.0	6.0	4.9	6.8	5.6	9.5	11.7	6.1	7.7	7.4	8.0	8.1	8.9	12.4	13.9	8.2	5.8	5.4	2.7	8.2	2.9	22.5	11.5	9.1	5.2
sk	0.7	5.1	2.7	27.0	4.3	5.7	3.2	16.9	9.3	9.4	8.5	6.7	2.7	1.0	5.1	3.7	4.9	6.9	2.2	3.9	3.5	4.9	5.0	7.1	7.8	8.5	3.5	6.6	5.0	1.5	4.3	1.6	5.4	5.4	2.3	2.5
sl	1.2	6.2	7.6	5.5	4.7	7.6	5.8	17.3	5.9	11.4	8.5	6.4	3.2	1.2	11.2	3.9	6.5	7.8	2.7	4.2	6.3	4.3	5.5	4.8	9.9	4.8	5.9	3.6	1.9	4.1	2.2	4.3	7.4	3.8	2.6	
sr	1.8	7.6	33.2	5.1	3.7	3.8	5.8	22.8	3.2	11.9	9.1	7.5	3.5	1.1	30.4	2.7	6.1	8.2	1.2	2.9	13.7	3.3	3.3	3.9	10.8	11.3	5.6	7.2	2.8	9.5	3.0	1.1	5.7	6.8	4.3	3.1
sv	2.2	7.4	4.8	5.9	26.5	12.6	8.1	31.8	11.0	16.9	15.7	10.7	7.1	3.3	5.5	5.4	11.6	13.3	4.8	6.2	5.2	25.4	11.5	5.8	15.0	17.4	7.9	8.1	3.8	4.7	1.0	3.2	6.9	12.9	7.8	4.8
tr	2.2	3.5	2.0	2.6	3.9	4.1	4.7	15.9	2.9	9.4	7.7	6.7	3.6	1.6	2.1	3.4	6.7	6.4	4.3	7.0	3.5	3.1	4.2	2.5	9.0	8.4	4.6	4.0	1.8	2.3	0.8	3.5		3.3	8.2	6.7	4.4
uk	2.9	12.3	5.3	7.4	7.5	7.5	8.4	20.7	6.5	14.2	14.1	11.2	5.5	3.5	6.6	4.7	9.5	11.2	4.9	5.8	7.2	6.3	6.9	9.6	12.9	7.2	23.5	4.9	5.7	2.6	6.9	2.6	11.4	7.9	4.9	
vi	4.2	7.5	4.0	4.7	8.5	6.0	8.8	20.2	7.3	13.7	13.2	9.9	6.5	4.6	4.9	4.7	14.7	10.7	5.6	9.3	6.9	5.7	7.3	4.5	13.0	14.1	8.5	7.2	3.4	4.6	1.7	8.2	4.0	6.7		9.9	6.7
zh-cn	2.1	3.2	1.0	2.2	3.8	3.2	4.5	11.8	3.8	8.2	7.6	3.2	1.7	1.9	3.0	6.6	6.0			3.4	3.8	2.2	7.1	7.9	4.1	4.1	1.6	2.4	0.9	3.1	2.3	3.0	10.8			
zh-tw	2.2	3.1	1.1	2.1	3.7	2.8	3.9	10.7	3.4	7.5	7.2	6.1	2.8	1.8	1.6	3.0	6.2	5.4		2.8	3.5	2.3	6.3	6.9	3.5	3.9	1.4	2.1	0.9	3.0	2.4	2.9	10.0			

Table 4: BLEU scores on the TED test set as proposed in (Qi et al., 2018). NMT systems were trained on bitexts mined in Wikipedia only (with at least twenty-five thousand parallel sentences). No other resources were used.

parameter settings shown in Figure 2 in the appendix. Since the TED development and test sets were already tokenized, we first detokenize them using Moses. We trained NMT systems for all possible language pairs with more than 25k mined sentences. This gives us in total 1886 language pairs in 45 languages. We train $L_1 \rightarrow L_2$ and $L_2 \rightarrow L_1$ with the same mined bitexts L_1/L_2. Scores on the test sets were computed with SacreBLEU (Post, 2018), see Table 4. Some additional results are reported in Table 6 in the annex. 23 NMT systems achieve BLEU scores over 30, the best one being 37.3 for Brazilian Portuguese to English. Several results are worth mentioning, like Farsi/English: 16.7, Hebrew/English: 25.7, Indonesian/English: 24.9 or English/Hindi: 25.7 We also achieve interesting results for translation between various non English language pairs for which it is usually not easy to find parallel data, e.g. Norwegian $\leftrightarrow$ Danish $\approx$33, Norwegian $\leftrightarrow$ Swedish $\approx$25, Indonesian $\leftrightarrow$ Vietnamese $\approx$16 or Japanese / Korean $\approx$17.

Our results on the TED set give an indication on the quality of the mined parallel sentences. These BLEU scores should be of course appreciated in context of the sizes of the mined corpora as given in Table 3. Finally, we would like to point out that we run our approach on all available languages in Wikipedia, independently of the quality of LASER's sentence embeddings for each one.

6 Conclusion

We have presented an approach to systematically mine for parallel sentences in the textual content of Wikipedia, for all possible language pairs. We use a mining approach based on massively multilingual sentence embeddings (Artetxe and Schwenk, 2018b) and a margin criterion (Artetxe and Schwenk, 2018a). The same approach is used for all language pairs without the need for a language-specific optimization. In total, we make available 135M parallel sentences in 96 languages, out of which only 34M sentences are aligned with English. We were able to mine more than ten thousands sentences for 1620 different language pairs. This corpus of parallel sentences is freely available.[15] We also performed a large scale evaluation of the quality of the mined sentences by training 1886 NMT systems and evaluating them on the 45 languages of the TED corpus (Qi et al., 2018). This approach was recently extended to mine in Common Crawl texts (Schwenk et al., 2019).

[15] https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix

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.

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 (WMT)*.

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.

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

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

Francis Grégoire and Philippe Langlais. 2017. BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora. In *BUCC*, pages 46–50.

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*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. https://arxiv.org/abs/1607.01759.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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

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.

P Otero, I López, S Cilenis, and Santiago de Compostela. 2011. Measuring comparability of multilingual corpora extracted from Wikipedia. *Iberian Cross-Language Natural Language Processings Tasks (ICL)*, page 8.

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, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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, Belgium, Brussels. 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.

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, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2019. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *https://arxiv.org/abs/1911.04944*.

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

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*.

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*.

A Appendix

Table 5 provides the amounts of mined parallel sentences for languages which have a rather small Wikipedia. Aligning those languages obviously yields to a very small amount of parallel sentences. Therefore, we only provide these results for alignment with high resource languages. It is also likely that several of these alignments are of low quality since the LASER embeddings were not directly trained on most these languages, but we still hope to achieve reasonable results since other languages of the same family may be covered.

ISO	Name	Language Family	size	ca	da	de	en	es	fr	it	nl	pl	pt	sv	ru	zh	total
an	Aragonese	Romance	**222**	24	7	12	23	33	16	13	9	10	14	9	11	6	**324**
arz	Egyptian Arabic	Arabic	**120**	7	6	11	18	12	12	10	8	9	10	8	12	7	**278**
as	Assamese	Indo-Aryan	**124**	8	6	11	7	11	12	10	9	9	8	8	9	3	**216**
azb	South Azerbaijani	Turkic	**398**	6	4	9	8	9	10	9	7	6	8	6	7	3	**172**
bar	Bavarian	Germanic	**214**	7	6	41	16	12	12	10	8	9	10	8	10	5	**261**
bpy	Bishnupriya	Indo-Aryan	**128**	2	1	4	4	3	4	2	2	3	2	2	3	1	**71**
br	Breton	Celtic	**413**			20	16	22	23	22			19		16	6	**200**
ce	Chechen	Northeast Caucasian	**315**	2	1	2	2	2	2	2	2	2	2	2	2	1	**56**
ceb	Cebuano	Malayo-Polynesian	**17919**	14	9	22	29	27	24	24	15	17	20	55	21	9	**594**
ckb	Central Kurdish	Kur-Iranian	**127**	2	2	6	8	5	5	4	4	4	4	3	6	4	**113**
cv	Chuvash	Turkic	**198**	4	3	5	4	6	6	7	5	4	6	5	8	2	**129**
dv	Maldivian	Indo-Aryan	**52**	2	2	5	6	4	4	3	3	3	3	3	5	3	**96**
fo	Faroese	Germanic	**114**	13	12	14	32	21	18	15	11	11	17	12	13	6	**335**
fy	Western Frisian	Germanic	**493**	13	8	16	32	21	18	17	38	12	18	13	14	5	**453**
gd	Gaelic	Celtic	**66**	1	1	1	1	1	1	1	1	1	1	1	1	1	**41**
ga	Irish	Irish	**216**		2	3	4	3	3	3	2	2	3	2	3	1	**70**
gom	Goan Konkani	Indo-Aryan	**69**	9	7	10	8	13	13	13	9	9	11	9	10	4	**240**
ht	Haitian Creole	Creole	**60**	2	1	3	4	3	4	3	2	3	2	2	3	1	**72**
ilo	Iloko	Philippine	**63**	3	2	4	5	4	4	4	3	3	4	3	4	2	**96**
io	Ido	constructed	**153**	5	3	6	11	7	7	5	5	5	6	5	5	3	**143**
jv	Javanese	Malayo-Polynesian	**220**	8	5	8	13	12	10	11	8	7	11	8	8	3	**219**
ka	Georgian	Kartvelian	**480**	11	7	15	12	16	17	16	12	11	14	12	13	5	**288**
ku	Kurdish	Iranian	**165**	5	4	8	5	8	7	8	7	6	7	6	6	3	**222**
la	Latin	Romance	**558**	12	9	17	32	20	18	17	12	13	18	13	14	6	**478**
lb	Luxembourgish	Germanic	**372**	12	7	26	22	19	18	15	11	11	16	12	11	4	**305**
lmo	Lombard	Romance	**147**	6	3	7	10	7	7	11	6	5	7	5	5	3	**144**
mg	Malagasy	Malayo-Polynesian	**263**	6	5	9	13	9	12	8	7	7	7	8	7	4	**199**
mhr	Eastern Mari	Uralic	**61**	3	2	4	3	4	4	5	3	3	4	3	4	2	**96**
min	Minangkabau	Malayo-Polynesian	**255**	4	2	6	7	5	5	5	4	4	4	5	5	2	**121**
mn	Mongolian	Mongolic	**255**	4	3	7	5	6	6	7	6	5	5	5	5	3	**197**
mwl	Mirandese	Romance	**64**	6	3	4	10	8	6	5	3	4	34	3	4	2	**154**
nds nl	Low German/Saxon	Germanic	**65**	5	4	6	10	7	7	6	15	5	6	5	5	3	**151**
ps	Pashto	Iranian	**89**		2	3	2	3	3	3	3	3	3	3	3	1	**73**
rm	Romansh	Italic	**57**	2	2	10	5	4	4	3	2	3	3	3	3	1	**86**
sah	Yakut	Turkic/Sib	**134**	4	3	7	5	6	6	6	5	5	5	5	6	3	**134**
scn	Sicilian	Romance	**81**	5	3	6	9	7	7	11	5	5	6	5	5	2	**143**
sd	Sindhi	Iranian	**115**		3	9		8	8	7	7	6	7	5	8	5	**152**
su	Sundanese	Malayo-Polynesian	**120**	4	3	5	7	6	5	6	4	4	5	4	4	2	**117**
tk	Turkmen	Turkic	**56**	2	2	3	3	4	3	4	2	2	4	2	3	1	**76**
tg	Tajik	Iranian	**248**	5	4	11	15	9	9	8	8	7	8	6	10	6	**192**
ug	Uighur	Turkic	**83**	4	3	9	10	7	8	6	6	5	6	5	9	6	**168**
ur	Urdu	Indo-Aryan	**150**	2	2	3	5	3	3	3	3	3	3	3	3	2	**123**
wa	Walloon	Romance	**56**	3	2	4	5	5	4	4	3	3	4	3	3	2	**93**
wuu	Wu Chinese	Chinese	**75**	8	6	11	17	12	11	10	8	9	11	9	10	43	**283**
yi	Yiddish	Germanic	**131**	3	2	4	3	4	4	5	3	3	4	3	4	1	**92**

Table 5: WikiMatrix (part 2): number of extracted sentences (in thousands) for languages with a rather small Wikipedia. Alignments with other languages yield less than 5k sentences and are omitted for clarity.

Table 2 gives the detailed configuration which was used to train NMT models on the mined data in Section 5. An 5000 subword vocabulary was learnt using SentencePiece (Kudo and Richardson, 2018). Decoding was done with beam size 5 and length normalization 1.2.

```
--arch transformer
--share-all-embeddings
--encoder-layers 5
--decoder-layers 5
--encoder-embed-dim 512
--decoder-embed-dim 512
--encoder-ffn-embed-dim 2048
--decoder-ffn-embed-dim 2048
--encoder-attention-heads 2
--decoder-attention-heads 2
--encoder-normalize-before
--decoder-normalize-before
--dropout 0.4
--attention-dropout 0.2
--relu-dropout 0.2
--weight-decay 0.0001
--label-smoothing 0.2
--criterion label_smoothed_cross_entropy
--optimizer adam
--adam-betas '(0.9, 0.98)'
--clip-norm 0
--lr-scheduler inverse_sqrt
--warmup-update 4000
--warmup-init-lr 1e-7
--lr 1e-3 --min-lr 1e-9
--max-tokens 4000
--update-freq 4
--max-epoch 100
--save-interval 10
```

Figure 2: Model settings for NMT training with `fairseq`

Finally, Table 6 gives the BLEU scores on the TED corpus when translating into and from English for some additional languages.

Lang	xx → en	en → xx
et	15.9	14.3
eu	10.1	7.6
fa	16.7	8.8
fi	10.9	10.9
lt	13.7	10.0
hi	17.8	21.9
mr	2.6	3.5

Table 6: BLEU scores on the TED test set as proposed in (Qi et al., 2018). NMT systems were trained on bitexts mined in Wikipedia only. No other resources were used.