Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

Mikel Artetxe
University of the Basque Country (UPV/EHU)*
mikel.artetxe@ehu.eus

Holger Schwenk
Facebook AI Research
schwenk@fb.com

Abstract

Machine translation is highly sensitive to the size and quality of the training data, which has led to an increasing interest in collecting and filtering large parallel corpora. In this paper, we propose a new method for this task based on multilingual sentence embeddings. Our approach uses an encoder-decoder trained over an initial parallel corpus to build multilingual sentence representations, which are then incorporated into a new margin-based method to score, mine and filter parallel sentences. In contrast to previous approaches, which rely on nearest neighbor retrieval with a hard threshold over cosine similarity, our proposed method accounts for the scale inconsistencies of this measure, considering the margin between a given sentence pair and its closest candidates instead. Our experiments show large improvements over existing methods. We outperform the best published results on the BUCC shared task on parallel corpus mining by more than 10 F1 points. We also improve the precision from 48.9 to 83.3 on the reconstruction of 11.3M English-French sentence pairs of the UN corpus. Finally, filtering the English-German ParaCrawl corpus with our approach, we obtain 31.2 BLEU points on newstest2014, an improvement of more than one point over the best official filtered version.

1 Introduction

Given the prevalence of corpus-based approaches to machine translation in the last decades, parallel corpus mining has long been a central task in natural language processing. The topic has got a renewed interest after the recent irruption of Neural Machine Translation (NMT), which is known to be particularly sensitive to the size and quality of the training data. This way, even if NMT has been shown to be vastly superior to Statistical Machine Translation (SMT) in standard benchmarks (Wu et al., 2016), SMT still beats NMT when the training corpus is relatively small (Koehn and Knowles, 2017), and it is also substantially more robust to noisy training data (Khayrallah and Koehn, 2018), thus increasing the need for large and high-quality parallel corpora.

While traditional parallel corpus mining has relied on heavily engineered systems (Utiyama and Isahara, 2003; Etchegoyhen and Azpeitia, 2016), often based on metadata information (Resnik, 1999; Shi et al., 2006), a recent line of work has shown promising results using multilingual sentence embeddings alone (Schwenk, 2018; Guo et al., 2018). The basic approach is to use a Recurrent Neural Network (RNN) encoder to map any input sentence to a fixed-length vector representation, which is coupled with an auxiliary RNN decoder and trained on initial parallel corpus. Having done that, sentence pairs are scored using the cosine similarity of their respective embeddings, so parallel corpora can be mined through nearest neighbor search and filtered by setting a hard threshold over this cosine score.

In this paper, we argue that this retrieval method suffers from the scale of cosine similarity not being globally consistent. As illustrated by the example in Table 1, some sentences without any correct translation have overall high cosine scores, making them rank higher than other sentences with a correct translation. Our proposed method tackles this issue by considering the margin between the cosine of a given sentence pair and that of its respective k nearest neighbors.

Our experiments support the effectiveness of our approach, showing large improvements over existing methods in different scenarios. For instance, our system obtains 92.6 F1 on the English-Chinese BUCC mining task and 83.3 P@1 on the

*This work was performed during an internship at Facebook AI Research.

(A)	*Les produits agricoles sont constitués de thé, de riz, de sucre, de tabac, de camphre, de fruits et de soie.*
0.818	Main crops include wheat, sugar beets, potatoes, cotton, tobacco, vegetables, and fruit.
0.817	The fertile soil supports wheat, corn, barley, tobacco, sugar beet, and soybeans.
0.814	Main agricultural products include grains, cotton, oil, pigs, poultry, fruits, vegetables, and edible fungus.
0.808	The important crops grown are cotton, jowar, groundnut, rice, sunflower and cereals.

(B)	*Mais dans le contexte actuel, nous pourrons les ignorer sans risque.*
0.737	But, in view of the current situation, we can safely ignore these.
0.499	But without the living language, it risks becoming an empty shell.
0.498	While the risk to those working in ceramics is now much reduced, it can still not be ignored.
0.488	But now they have discovered they are not free to speak their minds.

Table 1: Motivating example of the proposed scoring method. We show the nearest neighbor list of two French sentences on the BUCC training set along with their cosine similarities. Only the nearest neighbor of B is a correct translation, yet that of A has a higher cosine similarity. We argue that this is caused by the cosine similarity of different sentences being in different scales, making it a poor indicator of the confidence of the prediction. Our proposed scoring method tackles this issue by considering the margin between a given candidate and the rest of the k nearest neighbors, so it can correctly rank the nearest neighbor of B above that of A.

English-French UN corpus reconstruction task, an absolute improvement of nearly 15 and 35 points over the previous best results, respectively. In addition to that, we show that our method can also bring large improvements to downstream machine translation. An NMT model trained on our filtered version of ParaCrawl obtains 31.2 BLEU points on the standard English-German newstest2014 benchmark, an improvement of more than one point over an equivalent system trained on the official filtered version.

2 Related work

There is a large body of work on **parallel corpus mining** using a variety of different approaches.

A common strategy is to exploit **metadata** from web crawled comparable corpora. For instance, the STRAND algorithm relies on HTML markup to extract the structure of the document, which is linearized and aligned across different languages. It uses a threshold over the strength of this alignment to filter spurious candidates (Resnik, 1999; Resnik and Smith, 2003). Shi et al. (2006) go further and consider the full DOM hierarchy to align documents at the tree level. In addition to structural information, other metadata like document titles (Yang and Li, 2002) and URL patterns (Resnik and Smith, 2003) have also been used.

Nevertheless, it has been argued that metadata information is not always available nor necessarily reliable (Uszkoreit et al., 2010), which has motivated alternative methods that rely on the **textual content** of the comparable documents instead. For instance, 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). This basic approach was later extended to incorporate a word weighting scheme (Azpeitia et al., 2017) as well as a better handling of named entities (Azpeitia et al., 2018). The Zipporah model learns a logistic regression classifier over bag-of-word features to distinguish between ground truth translations and synthetic noisy ones (Xu and Koehn, 2017). Many of these content-based approaches rely on cross-lingual document retrieval to some extent (Utiyama and Isahara, 2003; Munteanu and Marcu, 2005, 2006; Abdul-Rauf and Schwenk, 2009), typically to obtain an initial alignment that is then further filtered. In addition to that, there have been several proposals that rely on machine translation to score parallel sentence candidates using automatic metrics like BLEU and TER (Abdul-Rauf and Schwenk, 2009; Bouamor and Sajjad, 2018).

Closer to our work, a recent research line has been successful at applying **multilingual sentence embeddings** to mine parallel corpora. These sentence embeddings have often been used as part of a larger system, either to obtain an initial alignment that is further filtered by a separate system (Bouamor and Sajjad, 2018), or as an intermediate representation of an end-to-end classifier (Grégoire and Langlais, 2017). More in line with our proposal, some authors have used an NMT inspired encoder-decoder to independently train

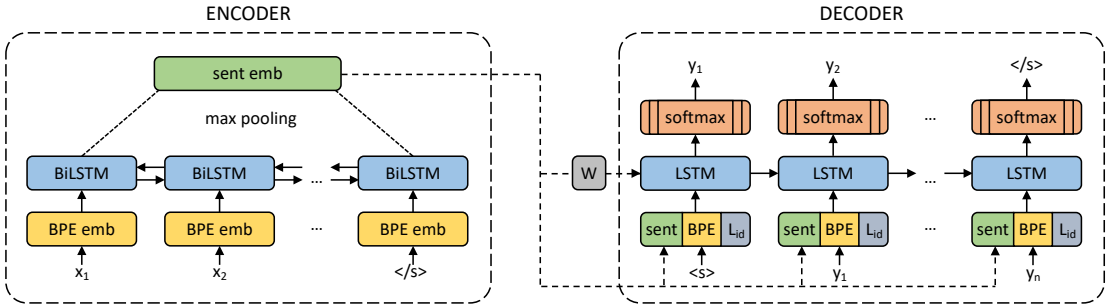

Figure 1: Architecture of our system to learn multilingual sentence embeddings.

multilingual sentence embeddings over existing parallel corpora, which are then directly applied to retrieve and filter parallel sentences with cosine similarity (España-Bonet et al., 2017; Hassan et al., 2018; Schwenk, 2018). Our proposed architecture is closest to Schwenk (2018) in the use of max-pooling to combine the hidden states of the encoder, which is shared across multiple languages with a common BPE vocabulary, but differs in the use of a shared decoder as opposed to having a separate decoder for each language. However, all these approaches use nearest neighbor retrieval and a hard threshold over cosine similarity, which we show that suffers form scale inconsistency issues, and our proposed margin-based score outperforms them by a large margin.

In relation to that, the recent work by Guo et al. (2018) also pointed out that the similarity score of different sentence pairs tends to be in different ranges when using the dot product, making it a **globally inconsistent** measure to apply a hard threshold over. Unlike previous methods, their architecture gets rid of the decoder, and instead trains the encoder to score known translation pairs above either random or automatically selected hard negative samples. With the goal to obtain a calibrated confidence score that is globally consistent, they train a separate model to dynamically scale and shift the dot product on held out supervised data. In comparison, our approach is conceptually simpler, and obtains substantially better results (e.g. 48.9 vs 83.3 P@1 on English-French UN reconstruction).

3 Multilingual sentence embeddings

Following recent trend, we use a sequence-to-sequence encoder-decoder to learn our sentence embeddings (Schwenk and Douze, 2017; España-Bonet et al., 2017; Hassan et al., 2018; Schwenk, 2018). During training, the entire system is trained end-to-end on a small parallel corpus. The decoder is then discarded, and the encoder is used to map any arbitrary sentence to a fixed-length continuous vector representation.

Figure 1 illustrates the architecture of the proposed system, which is based on Schwenk (2018). As it can be seen, the encoder consists of a bidirectional LSTM, and our sentence embeddings are obtained by applying a max-pooling operation over its output. The resulting sentence embeddings are fed into the decoder in two ways: first, they are used to initialize the hidden and cell state of the decoder LSTM after applying a linear transformation to each, and second, they are concatenated to the input embeddings at every time step. Note that there is no other connection between the encoder and the decoder, as we want all relevant information of the input sequence to be captured by our sentence embeddings.

We use a single encoder and decoder in our system, which are shared by all languages involved. For that purpose, we use a joint byte-pair encoding (BPE) vocabulary for all languages with 40K operations, which is learned on the concatenation of all training corpora.[1] Thanks to this, the encoder is fully language agnostic, meaning that it has no signal on what the source or target language is other than the input sequence itself, thus encouraging it to learn language independent representations. Nevertheless, the decoder does need to know what language it has to generate. Therefore, we use a separate embedding that identifies the target language, which is concatenated to the input and sentence embeddings at every time step as described above.

[1] Prior to BPE segmentation, we tokenize and lowercase the input text using standard Moses tools. As the only exception, we use Jieba for Chinese word segmentation.

Training minimizes the cross-entropy loss on parallel corpora, alternating over all combinations of the languages involved. For that purpose, we use Adam with a constant learning rate of 0.001 and dropout set to 0.1, and train for a fixed number of epochs. Our implementation is based on `fairseq`,[2] and we make use of its multi-GPU support to train on 4 GPUs with a total batch size of 48,000 tokens. We use a single layer for both the encoder and the decoder with a hidden size of 512 and 2048, respectively, thus yielding 1024 dimensional sentence embeddings. The input embedding size is set to 512, while the language ID embeddings have 32 dimensions.

4 Scoring and filtering parallel sentences

Having learned a multilingual sentence encoder as described in Section 3, parallel sentences can be mined by taking the nearest neighbor of each source sentence in the target side according to cosine similarity, and filtering those below a fixed threshold.

While this approach has been reported to be competitive in previous work (Schwenk, 2018), we argue that it suffers from the fact that the scale of cosine similarity is not globally consistent across different sentences.[3] For instance, Table 1 shows an example where an incorrectly aligned sentence pair has a larger cosine similarity than a correctly aligned one, thus making it impossible to filter it through a fixed threshold. In that case, all four nearest neighbors have equally high values. In contrast for the example 2, there is a big gap between the nearest neighbor and its other candidates. As such, we argue that the margin between the similarity of a given candidate and that of its k nearest neighbors is a better indicator of the strength of the alignment, as a correctly aligned candidate would be expected to have a higher cosine similarity than an incorrectly aligned one for a particular sentence, even if the absolute values across different sentences might vary. As a downside, this approach will tend to penalize sentences with many valid paraphrases in the corpus. While possible, we argue that such cases rarely happen in practice and, even when they do, filtering them is unlikely to cause any significant harm. We next describe our scoring method inspired by this idea in Section 4.1, and discuss our candidate generation and filtering strategy in Section 4.2.

4.1 Margin-based scoring

In order to account for the relative scale of cosine similarity and provide a globally consistent measure, our proposed scoring function considers the margin between the cosine of a given candidate and the average cosine of its k nearest neighbors in both directions as follows:

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

where $\text{NN}_k(x)$ denotes the k nearest neighbors of x in the other language,[4] and analogously for $\text{NN}_k(y)$. Note that this list of nearest neighbors does not include duplicates, so even if a given sentence has multiple occurrences in the corpus, it would have (at most) one entry in the list.

This general definition allows for different implementations of the margin function. In this work, we explore the followings ones, which in some cases generalize existing measures:

- **Absolute** ($\text{margin}(a, b) = a$): This margin function returns the absolute similarity of the given candidate ignoring the average. This is equivalent to the standard cosine similarity used in previous work. As such, it can be seen as the baseline that we aim to improve upon.

- **Distance** ($\text{margin}(a, b) = a - b$): This margin function subtracts the average cosine similarity from that of the given candidate. Interestingly, the resulting score is proportional to the cross-domain similarity local scaling (CSLS) measure proposed by Conneau et al. (2018). CSLS was originally motivated as a way to mitigate the hubness problem when inducing bilingual dictionaries from cross-lingually mapped word embeddings. Our main motivation is different, as bilingual dictionary induction does not require any filtering and, as such, having globally inconsistent scores is not a problem in the sense that there is no need to apply any thresh-

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

[3] Note that, even if cosine similarity is normalized in the (-1, 1) range, it is still susceptible to concentrate around different values.

[4] Unless otherwise indicated, we use $k = 4$.

old.[5] In any case, this connection points out a related problem that our approach also addresses: even when the source sentence is fixed, the potentially different scales of its target candidates might also affect their relative ranking, which ultimately causes the hubness problem. Thanks to its bidirectional nature, our proposed scoring method penalizes target sentences with overall high cosine similarities, so it can learn better alignments that account for this factor.

- **Ratio** ($\text{margin}(a, b) = \frac{a}{b}$): This margin function takes the ratio between the given candidate and the average cosine similarity of its nearest neighbors in both directions. While the motivation is the same as for the distance function, its behavior is slightly different for similarities in different ranges, so the lower the similarities are, the more significant their differences are considered. As shown in our experiments, this gives slightly better results than the distance function.

4.2 Candidate generation and filtering

So as to mine parallel text from comparable corpora, we score sentence pairs as described in Section 4.1, and then filter them by setting a threshold over this score. We first generate a candidate list using one of the following strategies, and then apply the filtering over it:

- **Forward**: For each source sentence, this strategy picks the target sentence with the highest margin-based score.[6] Each source sentence will be aligned with one and only one target sentence, but it is still possible that some target sentences are aligned with multiple source sentences, or even not aligned at all.

- **Backward**: Equivalent to the forward strategy, but going in the opposite direction.

- **Intersection**: This strategy takes the intersection between the forward and backward candidates, thus discarding sentences with inconsistent alignments. As such, it combines the retrieval in both directions to filter less reliable candidates but, by doing so, some correct alignments may be discarded.

- **Max. score**: This strategy takes the union between the forward and backward candidates, and sorts them in decreasing order according to their score. Having done that, sentence pairs are visited one by one: if both the source and the target side are new, a candidate is generated for them and, if not, the sentence pair is discarded. This strategy aims to find a compromise between the previous ones: it combines both retrieval directions to obtain non-overlapping alignments but, instead of discarding all inconsistent ones, it selects those with the highest score.

Having done that, candidates are filtered by applying a threshold over their respective scores. Depending on the scenario, the threshold can be adjusted to obtain the desired corpus size, or tuned to optimize some evaluation metric on a separate training set. In order to do this tuning efficiently, we sort all candidates in decreasing order according to their margin-based score, and select the optimal threshold accordingly. This can be done in $O(n \log n)$ time.

5 Experiments and results

We next present our results on the BUCC mining task, UN corpus reconstruction and downstream machine translation over the filtered ParaCrawl corpus. Unless otherwise indicated, all experiments use an English/French/Spanish/German multilingual encoder trained on Europarl v7 (Koehn, 2005) for 10 epochs as described in Section 3. As the only exception, we use a separate English/French/Russian/Chinese model to cover all languages in BUCC. This model is trained on the United Nations (UN) Parallel Corpus v1.0 (Ziemski et al., 2016) for 4 epochs given the larger size of this corpus.

5.1 BUCC mining task

In order to perform an intrinsic evaluation of our method for parallel corpus mining, we use the standard dataset from the 2017 shared task of the workshop on Building and Using Comparable Corpora (BUCC) (Zweigenbaum et al., 2017). It

[5] In fact, the source side average term does not have any effect in the dictionary induction method of Conneau et al. (2018) which is not the case for our parallel corpus mining algorithm given the thresholding over the resulting scores.

[6] To speed up computations, only the k nearest neighbors over cosine similarity are considered, where the neighborhood size k is the same as that used for the margin-based scoring.

Margin funct.	Retrieval	EN-DE			EN-FR		
		P	R	F1	P	R	F1
Absolute (Cosine)	Forward	78.94	75.09	76.97	82.09	74.19	77.94
	Backward	78.96	73.07	75.90	77.24	72.24	74.66
	Intersection	84.89	80.76	82.78	83.60	78.33	80.88
	Max. score	83.14	77.18	80.05	80.86	77.53	79.16
Distance	Forward	94.79	94.09	94.44	91.05	**91.83**	91.44
	Backward	94.78	94.11	94.44	91.46	91.36	91.41
	Intersection	94.90	94.09	94.50	91.15	91.81	91.48
	Max. score	94.90	94.09	94.50	91.15	91.82	91.49
Ratio	Forward	95.18	94.39	94.79	92.37	91.29	91.83
	Backward	95.18	**94.42**	94.80	92.32	91.31	91.81
	Intersection	95.27	94.39	94.83	**92.43**	91.27	**91.85**
	Max. score	**95.28**	94.41	**94.84**	**92.43**	91.28	**91.85**

Table 2: Results on the BUCC mining task for different margin functions and retrieval strategies. We report the precision, recall and F1 score on the training set, used to optimize the filtering threshold for each variant.

was also used in the 2018 edition (Zweigenbaum et al., 2018). The task is to mine for parallel sentences in English and four foreign languages: German, French, Russian and Chinese, respectively. There are 150K to 1.2M sentences for each language split into a sample, training and test set. About 2–3% of the sentences are parallel.

Table 2 reports precision, recall and F1 scores obtained by different variants of our proposed method on the BUCC training set. Note that the gold standard information was exclusively used to optimize the filtering threshold for each configuration, making results comparable across different variants. Our results show that multilingual sentence embeddings already achieve competitive performance using standard forward retrieval over cosine similarity. This is in line with the findings of Schwenk (2018). Our numbers are slightly better under these settings, probably thanks to the new implementation and the use of a shared decoder. Moreover, both of our bidirectional retrieval strategies achieve substantial improvements over this baseline while still relying on cosine similarity, with *intersection* giving the best results.

More importantly, our proposed margin-based scoring brings large improvements when using both the *distance* and the *ratio* functions, outperforming cosine similarity by more than 10 points in all cases. The best results are achieved by the *ratio* function, which outperforms the *distance* function by a small but consistent margin of 0.3-0.5 points. Interestingly, the retrieval strategy has a very small effect in both cases, suggesting that the proposed scoring is more robust than cosine, yet both bidirectional variants still give marginally better results than *forward* and *backward*.

In order to put these numbers into perspective, Table 3 reports the results for both the Europarl and the UN model in comparison to previous work. We use the *ratio* margin function with *maximum score* retrieval for our method, which achieves the best results on the training set. The filtering threshold was optimized to maximize the F1 score on the training set for each language pair and model. The gold-alignments of the test set are not publicly available - these scores on the test set are calculated by the organizers of the BUCC workshop. We have done one single submission.

As it can be seen, our proposed system outperforms all previous methods by a large margin, obtaining improvements of more than 10 F1 points in all cases, and up to 15 in some. We achieve F1 scores above 92 points on the test set for all language pairs, reaching 95.58 in the best case (English-German). Note, moreover, that the dataset was automatically created and is known to contain false negatives – Zweigenbaum et al. (2017) find that the F1 score for French-English could be underestimated by up to 1.5 points. This means that our real performance should be slightly better . At the same time, it is remarkable that

	TRAIN				TEST			
	de-en	**fr-en**	**ru-en**	**zh-en**	**de-en**	**fr-en**	**ru-en**	**zh-en**
Azpeitia et al. (2017)	83.33	78.83	-	-	83.74	79.46	-	-
Grégoire and Langlais (2017)	-	20.67	-	-	-	20	-	-
Zhang and Zweigenbaum (2017)	-	-	-	43.48	-	-	-	45.13
Azpeitia et al. (2018)	84.27	80.63	80.89	76.45	85.52	81.47	81.30	77.45
Bouamor and Sajjad (2018)	-	75.2	-	-	-	76.0	-	-
Chongman Leong and Chao (2018)	-	-	-	58.54	-	-	-	56
Schwenk (2018)	76.1	74.9	73.3	71.6	76.9	75.8	73.8	71.6
Proposed method (Europarl)	**94.84**	**91.85**	-	-	**95.58**	**92.89**	-	-
Proposed method (UN)	-	90.75	**90.92**	**91.04**	-	-	**92.03**	**92.57**

Table 3: F1 scores on the BUCC mining task. Our proposed method uses the *ratio* margin function with *maximum score* retrieval, and the filtering threshold was optimized on the training set.

our method generalizes well to distant languages. For instance, our French-English results are only 0.3 points better than our Chinese-English results, while the difference is above 4 points for both Schwenk (2018) and Azpeitia et al. (2018). Finally, it is interesting that our Europarl model achieves better results than our UN model for French-English despite being trained on less data (2M vs 11M sentences). This suggests that the domain of the Europarl corpus is more appropriate.

5.2 UN corpus reconstruction

As discussed in Section 2, the recent work by Guo et al. (2018) is also based on multilingual sentence embeddings and, similar to our proposal, addresses the issue of the global inconsistency of cosine similarity. In order to compare our system to theirs, and to test our proposed method in a different scenario, we mimic their UN corpus reconstruction experiments. This task consists in aligning the 11.3M sentences that comprise the UN corpus. Given that the task does not require any filtering, we use *forward* retrieval with the *ratio* margin function over the Europarl model.

Table 4 reports the results obtained by both systems, using the precision at 1 as the evaluation measure. As it can be seen, our proposed method outperforms that of Guo et al. (2018) by a large margin, obtaining 83.27% and 85.78% precision for English-French and English-Spanish, respectively. Moreover, our model uses only a fraction of the training data: 2M sentences from the Europarl corpus in comparison to more than 400M sentences from Google's internal data Guo et al. (2018).

	EN-FR	**EN-ES**
Guo et al. (2018)	48.90	54.94
Proposed method	**83.27**	**85.78**

Table 4: Results on UN corpus reconstruction (P@1)

5.3 Filtering ParaCrawl for NMT

In order to understand the effect of our improved parallel corpus mining on downstream machine translation, we build different filtered versions of the English-German ParaCrawl corpus,[7] and evaluate NMT models trained on them. For that purpose, we use `fairseq`'s implementation of the big transformer model (Vaswani et al., 2017), using the exact same configuration as Ott et al. (2018) and training on 8 DGX-1 nodes with 8 Nvidia V100 GPUs each for 100 epochs. Following common practice, we use newstest2013 and newstest2014 as our development and test sets, respectively, and report both tokenized BLEU scores as computed by `multi-bleu.perl` as well as the official detokenized BLEU scores as computed by sacreBLEU.[8]. We decode with a beam size of 5 using an ensemble of the last 10 epochs. One single model is only slightly worse.

With 4.59 billion sentence pairs, the full raw re-

[7] https://paracrawl.eu/

[8] sacreBLEU signature: BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+test.SET+tok.13a+version.1.2.10 with SET ∈ {wmt13, wmt14/full} This is equivalent to the official `mteval-v13a.pl` script from WMT.

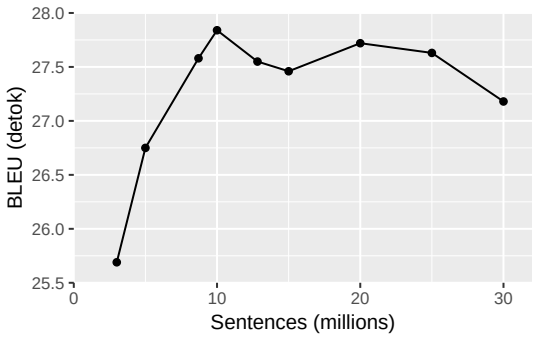

Figure 2: English-German dev results (newstest2013) using different thresholds to filter ParaCrawl.

	#SENT	BLEU	
		tok	detok
BiCleaner v1.2	17.4M	30.05	29.37
Zipporah v1.2	40.5M	24.78	24.38
Proposed method	10.0M	**31.19**	**30.53**

Table 5: Results on English-German newstest2014 for different filtered versions of the ParaCrawl corpus.

lease of ParaCrawl is very large and noisy. For that reason, we first preprocess it to remove all duplicated sentence pairs, sentences for which the fastText language identification model[9] predicts a different language, those with less than 3 or more than 80 tokens, or those with either an overlap of at least 50% or a ratio above 2 between the source and target tokens. This reduces the corpus size to 64.4M million sentence pairs, mostly due to deduplication. Having done that, we score each sentence pair with our margin-based method using the *ratio* function, processing the entire corpus in batches of 5 million sentences, and producing different filtered versions by taking the top scoring entries up to the desired size.

Figure 2 shows the development BLEU scores of the resulting system for different thresholds. Not surprisingly, results initially improve as more sentences are added, with a peak at 10 million sentences, slightly fluctuate afterwards, and start to degrade with too many noisy entries. Table 5 shows the final results for our best development model in comparison to the two official filtered versions of the last ParaCrawl release, which are based on Zipporah (Xu and Koehn, 2017) and Bi-

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

	DATA	BLEU	
		tok	detok
Wu et al. (2016)	wmt	26.3	-
Gehring et al. (2017)	wmt	26.4	-
Vaswani et al. (2017)	wmt	28.4	-
Ahmed et al. (2017)	wmt	28.9	-
Shaw et al. (2018)	wmt	29.2	-
Ott et al. (2018)	wmt	29.3	28.6
Ott et al. (2018)	wmt+pc	29.8	29.3
Edunov et al. (2018)	wmt+nc	35.0	33.8
Proposed method	pc	31.2	30.5
	wmt+pc	31.8	31.1

Table 6: Results on English-German newstest2014 in comparison to previous work. *wmt* for WMT parallel data (excluding ParaCrawl), *pc* for ParaCrawl, and *nc* for monolingual News Crawl with back-translation.

Cleaner, a heavily engineered system combining hard rules and a dictionary-based classifier. As it can be seen, our filtered version achieves substantially better results, outperforming BiCleaner and Zipporah by more than 1 and 6 BLEU points, respectively.

Finally, Table 6 compares our results to previous works in the literature using different training data. In addition to our main system trained on ParaCrawl, we include an additional one combining it with all parallel data from WMT18 with the exception of CommonCrawl. As it can be seen, our system achieves very good results, outperforming all but one previous systems by a considerable margin. The only exception is Edunov et al. (2018), who use a large in-domain monolingual corpus through back-translation, making both works complementary. Quite remarkably, our full system outperforms Ott et al. (2018), which was the previous state-of-the-art when training on parallel corpora only, by nearly 2 points. Note that both systems use the same configuration and training data, so our improvement can be attributed to a better filtering of ParaCrawl.[10]

[10] To confirm this, we trained a separate model on WMT data, obtaining 29.4 tokenized BLEU. This is on par with the results reported by Ott et al. (2018) for the same data (29.3 tokenized BLEU). This shows that the difference cannot be attributed to implementation details.

6 Conclusions and future work

In this paper, we propose a new method for parallel corpus mining based on multilingual sentence embeddings. Our approach uses a sequence-to-sequence architecture to train a multilingual sentence encoder on an initial parallel corpus, and a novel margin-based scoring method that overcomes the scale inconsistencies of cosine similarity.

Our experiments show large improvements over previous methods in all the tested scenarios. Our system obtains the best published results on the BUCC mining task, outperforming previous systems by more than 10 F1 points for all the four language pairs. In addition, our method obtains up to 85% precision at reconstructing the 11.3M sentence pairs from the UN corpus, improving over the similarly motivated method of Guo et al. (2018) by more than 30 points. Finally, we show that our improvements also carry over to downstream machine translation, as we obtain 31.2 BLEU points for English-German newstest2014 training on our filtered version of ParaCrawl, an improvement of more than one point over the best performing official release.

References

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

Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. 2017. Weighted Transformer Network for Machine Translation. *arXiv:1711.02132*.

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*.

Derek F. Wong Chongman Leong and Lidia S. Chao. 2018. UM-pAligner: Neural Network-Based Parallel Sentence Identification Model. In *BUCC*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *ICLR*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. *arXiv:1808.09381*.

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.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. *arXiv:1705.03122*.

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*.

Huda Khayrallah and Philipp Koehn. 2018. On the Impact of Various Types of Noise on Neural Machine Translation. In *WNMT*, pages 74–83.

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

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *WNMT*, pages 28–39.

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

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In *ACL*, pages 81–88.

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

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 and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *ACL workshop on Representation Learning for NLP*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. *arXiv:1803.02155*.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A DOM Tree Alignment Model for Mining Parallel Data from the Web. In *ACL*, pages 489–496.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. In *COLING*, pages 1101–1109.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 6000–6010.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144*.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora. In *EMNLP*, pages 2945–2950.

Christopher C Yang and Kar Wing Li. 2002. Mining English/Chinese Parallel documents from the world wide web. In *Proceedings of the 11th International World Wide Web Conference*.

Zheng Zhang and Pierre Zweigenbaum. 2017. zNLP: Identifying Parallel Sentences in Chinese-English Comparable Corpora. In *BUCC*, pages 51–55.

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. 2017. Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. In *BUCC*, pages 60–67.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the Third BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. In *BUCC*.