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. 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 mining task and the UN reconstruction task by more than 10 F1 and 30 precision points, respectively. 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

While Neural Machine Translation (NTM) has obtained breakthrough improvements in standard benchmarks, it is known to be particularly sensitive to the size and quality of the training data (Koehn and Knowles, 2017; Khayrallah and Koehn, 2018). In this context, effective approaches to mine and filter parallel corpora are crucial to apply NMT in practical settings.

Traditional parallel corpus mining has relied on heavily engineered systems. Early approaches were mostly based on metadata information from web crawls (Resnik, 1999; Shi et al., 2006). More recent methods focus on the textual content instead. For instance, Zipporah learns a classifier over bag-of-word features to distinguish between ground truth translations and synthetic noisy ones (Xu and Koehn, 2017). STACC 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). Many of these approaches rely on cross-lingual document retrieval (Utiyama and Isahara, 2003; Munteanu and Marcu, 2005, 2006; Abdul-Rauf and Schwenk, 2009) or machine translation (Abdul-Rauf and Schwenk, 2009; Bouamor and Sajjad, 2018).

More recently, a new research line has shown promising results using multilingual sentence embeddings alone[1] (Schwenk, 2018; Guo et al., 2018). These methods use an NMT inspired encoder-decoder to train sentence embeddings on existing parallel data, which are then directly applied to retrieve and filter new parallel sentences using nearest neighbor retrieval over cosine similarity with a hard threshold (España-Bonet et al., 2017; Hassan et al., 2018; Schwenk, 2018).

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. This issue was also pointed out by Guo et al. (2018), who learn an encoder to score known translation pairs above synthetic negative examples and train a separate model to dynamically scale and shift the dot product on held out supervised data. In contrast, our

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

[1] Multilingual entence embeddings have also been used as part of a larger system, either to obtain an initial alignment that is then further filtered (Bouamor and Sajjad, 2018) or as an intermediate representation of an end-to-end classifier (Grégoire and Langlais, 2017).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3197–3203
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

(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 method. We show the nearest neighbors 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 method tackles this issue by considering the margin between a given candidate and the rest of the k nearest neighbors.

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.

2 Multilingual sentence embeddings

Figure 1 shows our encoder-decoder architecture to learn multilingual sentence embeddings, which is based on Schwenk (2018). The encoder consists of a bidirectional LSTM, and our sentence embeddings are obtained by applying a max-pooling operation over its output. These embeddings are fed into an LSTM decoder in two ways: 1) they are used to initialize its hidden and cell state after a linear transformation, and 2) they are concatenated to the input embeddings at every time step. We use a shared encoder and decoder for all languages with a joint 40k BPE vocabulary learned on the concatenation of all training corpora.[2] The encoder is fully language agnostic, without any explicit signal of the input or output language, whereas the decoder receives an output language ID embedding at every time step. Training minimizes the cross-entropy loss on parallel corpora, alternating over all combinations of the languages involved. We train on 4 GPUs with a total batch size of 48,000 tokens, using Adam with a learning rate of 0.001 and dropout set to 0.1. We use a single layer for both the encoder and the decoder with a hidden size of 512 and 2048, respectively, yielding 1024 dimensional sentence embeddings. The input embeddings size is set to 512, while the language ID embeddings have 32 dimensions. After training, the decoder is discarded, and the encoder is used to map a sentence to a fixed-length vector.

3 Scoring and filtering parallel sentences

The multilingual encoder can be used to mine parallel sentences 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 (Schwenk, 2018), we argue that it suffers from the scale of cosine similarity not being 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 example B, 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.[4] We next describe our scoring method inspired by this idea in Section 3.1, and discuss our candidate generation and filtering strategy in Section 3.2.

[2]Prior to BPE segmentation, we tokenize and lowercase the input text using standard Moses tools. As the only exception, we use Jieba (https://github.com/fxsjy/jieba) for Chinese word segmentation.

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

[4]As a downside, this approach will penalize sentences with many 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 major harm.

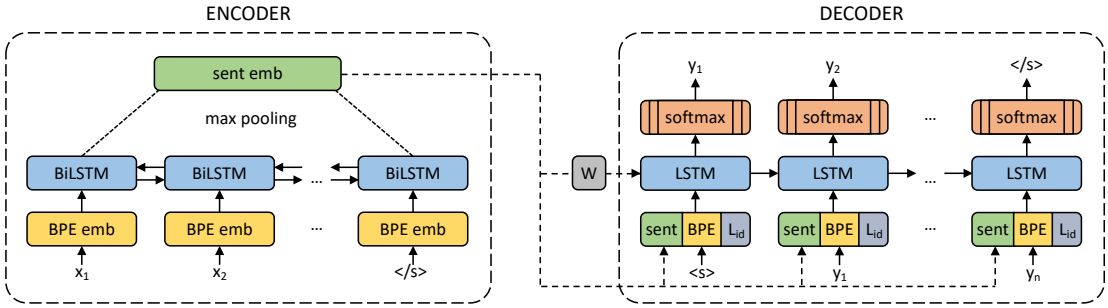

Figure 1: Architecture of our system to learn multilingual sentence embeddings.

3.1 Margin-based scoring

We consider 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 excluding duplicates,[5] and analogously for $\text{NN}_k(y)$. We explore the following variants of this general definition:

- **Absolute** ($\text{margin}(a, b) = a$): Ignoring the average. This is equivalent to cosine similarity and thus our baseline.
- **Distance** ($\text{margin}(a, b) = a - b$): Subtracting the average cosine similarity from that of the given candidate. This is proportional to the CSLS score (Conneau et al., 2018), which was originally motivated to mitigate the hubness problem on Bilingual Lexicon Induction (BLI) over cross-lingual word embeddings.[6]
- **Ratio** ($\text{margin}(a, b) = \frac{a}{b}$): The ratio between the candidate and the average cosine of its nearest neighbors in both directions.

3.2 Candidate generation and filtering

When mining parallel sentences, we explore the following strategies to generate candidates:

- **Forward**: Each source sentence is aligned with exactly one best scoring target sentence.[7] Some target sentences may be aligned with multiple source sentences or with none.
- **Backward**: Equivalent to the forward strategy, but going in the opposite direction.
- **Intersection** of forward and backward candidates, which discards sentences with inconsistent alignments.
- **Max. score**: Combination of forward and backward candidates that, instead of discarding all inconsistent alignments, it selects those with the highest score.

These candidates are then sorted according to their margin scores, and a threshold is applied. This can be either optimized on the development data, or adjusted to obtain the desired corpus size.

4 Experiments and results

We next present our results on the BUCC mining task, UN corpus reconstruction, and machine translation over filtered ParaCrawl. All experiments use an English/French/Spanish/German multilingual encoder trained on Europarl v7 (Koehn, 2005) for 10 epochs. To cover all languages in BUCC, we use a separate English/French/Russian/Chinese model trained on the UN corpus (Ziemski et al., 2016) for 4 epochs.

4.1 BUCC mining task

The shared task of the workshop on Building and Using Comparable Corpora (BUCC) is a well-established evaluation framework for bitext mining (Zweigenbaum et al., 2017, 2018). The task is

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

[6] While our work is motivated by thresholding, which is not used in BLI, 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.

[7] For efficiency, 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.

Func.	Retrieval	EN-DE			EN-FR		
		P	R	F1	P	R	F1
Abs. (cos)	Forward	78.9	75.1	77.0	82.1	74.2	77.9
	Backward	79.0	73.1	75.9	77.2	72.2	74.7
	Intersection	84.9	80.8	82.8	83.6	78.3	80.9
	Max. score	83.1	77.2	80.1	80.9	77.5	79.2
Dist.	Forward	94.8	94.1	94.4	91.1	**91.8**	91.4
	Backward	94.8	94.1	94.4	91.5	91.4	91.4
	Intersection	94.9	94.1	94.5	91.2	**91.8**	91.5
	Max. score	94.9	94.1	94.5	91.2	**91.8**	91.5
Ratio	Forward	95.2	**94.4**	**94.8**	**92.4**	91.3	91.8
	Backward	95.2	**94.4**	**94.8**	92.3	91.3	91.8
	Intersection	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**
	Max. score	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**

Table 2: BUCC results (precision, recall and F1) on the training set, used to optimize the filtering threshold.

to mine for parallel sentences between English and four foreign languages: German, French, Russian and Chinese. 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 on the training set.[8] Our results show that multilingual sentence embeddings already achieve competitive performance using standard forward retrieval over cosine similarity, which is in line with Schwenk (2018). Both of our bidirectional retrieval strategies achieve substantial improvements over this baseline while still relying on cosine similarity, with *intersection* giving the best results. Moreover, our proposed margin-based scoring brings large improvements when using either the *distance* or the *ratio* functions, outperforming cosine similarity by more than 10 points in all cases. The best results are achieved by *ratio*, which outperforms *distance* by 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.

Table 3 reports the results on the test set for both the Europarl and the UN model in comparison to previous work.[9] Our proposed system outperforms all previous methods by a large margin,

[8]Note that the gold standard information was exclusively used to optimize the filtering threshold for each configuration, making results comparable across different variants.

[9]We use the *ratio* margin function with *maximum score* retrieval for our method. 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.

	en-de	en-fr	en-ru	en-zh
Azpeitia et al. (2017)	83.7	79.5	-	-
Azpeitia et al. (2018)	85.5	81.5	81.3	77.5
Bouamor and Sajjad (2018)	-	76.0	-	-
Schwenk (2018)	76.9	75.8	73.8	71.6
Proposed method (Europarl)	**95.6**	**92.9**	-	-
Proposed method (UN)	-	-	**92.0**	**92.6**

Table 3: BUCC results (F1) on the test set. We use the *ratio* function with *maximum score* retrieval and the filtering threshold optimized on the training set.

	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)

obtaining improvements of 10-15 F1 points and showing very consistent performance across different languages, including distant ones.

4.2 UN corpus reconstruction

So as to compare our method to the similarly motivated system of Guo et al. (2018), we mimic their experiment on aligning the 11.3M sentences of the UN corpus. This task does not require any filtering, so we use *forward* retrieval with the *ratio* margin function. As shown in Table 4, our system outperforms that of Guo et al. (2018) by a large margin despite using only a fraction of the training data (2M sentences from Europarl in contrast with over 400M sentences from Google's internal data).

4.3 Filtering ParaCrawl for NMT

Finally, we filter the English-German ParaCrawl corpus and evaluate NMT models trained on them. Our NMT models use `fairseq`'s implementation of the big transformer model (Vaswani et al., 2017), using the same configuration as Ott et al. (2018) and training for 100 epochs. Following common practice, we use newstest2013 and newstest2014 as our development and test sets, respectively, and report both tokenized and detokenized BLEU scores as computed by `multi-bleu.perl` and sacreBLEU. We decode with a beam size of 5 using an ensemble of the last 10 epochs. One single model is only slightly worse.

Given the large size of ParaCrawl, we first preprocess it to remove all duplicated sentence pairs,

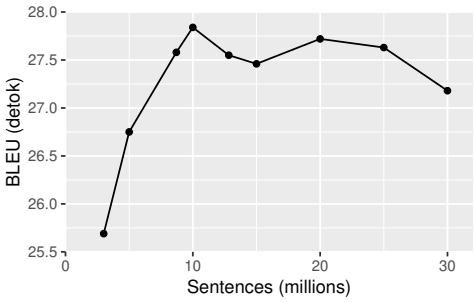

Figure 2: English-German Dev results (newstest2013) using different thresholds to filter ParaCrawl.

	#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.

sentences for which the fastText language identification model[10] 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 from 4.59 billion to 64.4 million sentence pairs, mostly due to deduplication. We then score each sentence pair with the *ratio* function, processing the entire corpus in batches of 5 million sentences, and take the top scoring entries up to the desired size. Figure 2 shows the development BLEU scores of the resulting system for different thresholds, which peaks at 10 million sentences. As shown in Table 5, this model clearly outperforms the two official filtered versions of ParaCrawl in the test set.

Finally, Table 6 compares our results to previous works in the literature using different training data. In addition to our ParaCrawl system, we include an additional one combining it with all parallel data from WMT18 except CommonCrawl. As it can be seen, our system outperforms all previous systems but 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) by nearly 2 points despite using the same configuration and training data, so

[10] 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.

our improvement can be attributed to a better filtering of ParaCrawl.[11]

5 Conclusions and future work

In this paper, we propose a new method for parallel corpus mining based on multilingual sentence embeddings. We use 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. 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.

The code of this work is freely available as part of the LASER toolkit, together with an additional single encoder which covers 93 languages.[12]

[11] 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.

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

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*.

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. In *EMNLP*, pages 489–500.

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. In *ICML*, pages 1243–1252.

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. In *WMT*, pages 165–176.

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. In *WMT*, pages 1–9.

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

Holger Schwenk. 2018. Filtering and Mining Parallel Data in a Joint Multilingual Space. In *ACL*, pages 228–234.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In *NAACL*, pages 464–468.

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.

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.

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*.