Noisy Parallel Corpus Filtering through Projected Word Embeddings

Murathan Kurfalı*
Department of Linguistics
Stockholm University
murathan.kurfali@ling.su.se

Robert Östling*
Department of Linguistics
Stockholm University
robert@ling.su.se

Abstract

We present a very simple method for parallel text cleaning of low-resource languages, based on projection of word embeddings trained on large monolingual corpora in high-resource languages. In spite of its simplicity, we approach the strong baseline system in the downstream machine translation evaluation.

1 Introduction

With the advent of web-scale parallel text mining, quality estimation and filtering is becoming an increasingly important step in multilingual NLP. Existing methods focus on languages with relatively large amounts of parallel text available (Schwenk, 2018; Artetxe and Schwenk, 2018), but scaling down to languages with limited amounts of parallel text poses new challenges. We present a method based on projecting word embeddings learned from a monolingual corpus in a high-resource language, to the target low-resource language through whatever parallel text is available.

The goal of participants in the WMT 2019 parallel corpus filtering shared task is to select the 5 million words of parallel sentences producing the highest-quality machine translation system, given a set of automatically crawled sentence candidates of varying quality. It is the continuation of the last year's task (Koehn et al., 2018), except that this year two low-resource languages are used: Nepali and Sinhalese.

2 Related Work

We refer readers to Koehn et al. (2018) for a more thorough review of the methods used in the WMT 2018 parallel corpus filtering shared task, and here review only a few studies of particular relevance to our model.

The Zipporah model of Xu and Koehn (2017) is used as a (strong) baseline in this year's shared task. It aims to find sentences pairs with high adequacy, according to dictionaries generated from an aligned corpora, and fluency modeled by n-gram language models.

Zariņa et al. (2015) use existing parallel corpora to learn word alignments and identify parallel sentences on the assumption that non-parallel sentences have few or none word alignments. In preliminary experiments we also evaluated a variant of this method, but found the resulting machine translation system to produce worse results than the simple approach described below.

Similar to the our model, Bouamor and Sajjad (2018) perform parallel sentence mining through sentence representations obtained by averaging bilingual word embeddings. Based on the cosine similarity, they create a candidate translation list for each sentence on the source side. Then, finding the correct translation is modelled as either a machine translation or binary classification task.

3 Data

In this section, we summarize the target noisy data and the allowed third-party resources where we train our model.

3.1 Target Noisy Corpora

The target noisy parallel corpora provided by the WMT 2019 organizers come from the Paracrawl project[1], and is provided before the standard filtering step to ensure high-recall, low-precision retrieval of parallel sentences.

The noisy corpora have 40.6 million words on the English side (English-Nepali) and 59.6 million words (English-Sinhala). The task is thus to se-

* Authors contributed equally.

[1] https://paracrawl.eu/

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 279–283
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

Language	Word Count	Sentence Count
Sinhala	3,745,282	646,781
Nepali	3,738,058	581,297

Table 1: Word and sentences counts of the "clean" parallel text

lect the approximately 10% highest-quality parallel text.

3.2 Training Data

Participants are allowed to use only the resources provided by the organizers to train systems. The permissible resources include supposedly clean parallel data, consisting of bible translations, Ubuntu localization files as well as movie subtitles. Larger monolingual corpora based on Wikipedia and common crawl data were also provided.[2]

To train our model, we use all the parallel data available for the English-Sinhala and English-Nepali pairs (summarized in Table 1) and the English Wikipedia dump which contains about 2 billion words. We modified the Nepali-English dictionary so that multiple translations were split into separate lines. As manual inspection revealed some problems in this data as well, we ran the same pre-filtering pipeline on it as we used for the noisy evaluation data (see Section 4.1)

4 Method

In this section, we present the components our model used to score the nosiy parallel data.

4.1 Pre-filtering Methods

As many types of poor sentence pairs are easy to detect with simple heuristics, we begin by applying a series of pre-filters. Before pre-filtering, the corpus is normalized through punctuation removal and lowercasing. We pre-filter all parallel data, both the (supposedly) clean and the noisy evaluation sets, using a set of heuristics based heavily on the work of Pinnis (2018):

- **Empty sentence filter:** Remove pairs where either sentence is empty after normalization.
- **Numeral filter:** Remove pairs where either sentence contains 25% or more numerals.
- **Sentence length filter:** Remove pairs where sentence lengths differ by 15 or more words.
- **Foreign writing filter:** Remove pairs where either sentence contains 10% or more words written in the wrong writing system.
- **Long string filter:** Remove pairs containing any token longer than 30 characters.
- **Word length filter:** Remove pairs where either sentence has an average word length of less than 2.

The statistics of each individual filter on the training data and the noisy data are provided in Table 2 and Table 3. In total, the pre-filtering step removed 2,790,557 pairs for the English–Sinhala data and 1,778,339 pairs for English–Nepali. Of all filters, foreign writing and numeral filter seem to be the most useful ones in terms of removing poor data.

Although almost 150 thousand sentence pairs are filtered out in the training data, the rate is considerably less than that of the raw noisy data suggesting that our pre-filters have a low rate of false positives. We further tested our pre-filters on the development data for the MT system evaluation (discarding the result), and found that less than 3% is removed.

4.2 Multilingual word vectors

We first train 300-dimensional FASTTEXT vectors (Bojanowski et al., 2017) with its default parameters using the provided English Wikipedia data.

Our first goal is now to create word vectors for the low-resource languages Sinhala and Nepali, in the same space as the English vectors.

After pre-filtering, we perform word alignment of the provided parallel text using the EFLOMAL tool (Östling and Tiedemann, 2016) with default parameters. Alignment is performed in both directions, and the intersection of both alignments is used. The vector v_i^f for word i in the non-English language f is computed as

$$v_i^f = \sum_j c(i,j) v_j^e$$

that is, the weighted sum of the vectors v_j^e of all aligned English word types j, which have been aligned to the non-English type i with frequency $c(i,j)$. Word types which are aligned less than 20% of the most commonly aligned type are not

[2] http://www.statmt.org/wmt19/parallel-corpus-filtering.html

	SINHALA		NEPALI	
	Count	**Percentage**	**Count**	**Percentage**
Before filtering	646,781	100	581,297	100
Word length filter	3,149	-0.49	4,133	-0.71
Long string filter	90	-0.01	77	-0.01
Numeral filter	4,803	-0.74	11,981	-2.06
Empty sentence filter	1,859	-0.29	410	-0.07
Sentence length filter	1,140	-0.18	4,501	-0.77
Foreign writing filter	38,965	-6.02	96,161	-16.54
Remaining	596,775	92.27	464,034	79.83

Table 2: Result of pre-filtering the "clean" parallel data.

	SINHALA		NEPALI	
	Count	**Percentage**	**Count**	**Percentage**
Before filtering	3,357,018	100.0	2,235,512	100.0
Word length filter	-7,981	-0.2	-3,015	-0.1
Long string filter	-2,782	-0.1	-4,848	-0.2
Numeral filter	-1,202,438	-35.8	-556,491	-24.9
Empty sentence filter	-7,672	-0.2	-4,378	-0.2
Sentence length filter	-216,486	-6.4	-272,567	-12.2
Foreign writing filter	-1,353,198	-40.3	-937,040	-41.9
Remaining	566,461	16.87	457,173	20.45

Table 3: Result of pre-filtering the noisy data.

counted, to compensate for potentially noisy word alignments. In other words, we let $c(i, j) = 0$ if the actual count is less than $0.2 \max_{j'} c(i, j')$. On average, the vector of each Sinhala word type is projected from 1.66 English word types, and each Nepali word from 1.83 English words types.

	1 Million	5 Million
Sinhala	3.59 (4.65)	0.53 (3.74)
Nepali	4.55 (5.23)	1.21 (1.85)

Table 4: BLEU scores of the NMT system trained on the released development sets. Numbers within parenthesis refer to the baseline scores

4.3 Sentence similarity

Given a sentence pair x and y, our task is to assign a score of translation equivalence. The multilingual word vectors learned in Section 4.2 provide a measure of *word-level* translational equivalence, by using the cosine similarity between the vectors of two words. Since sentence-level equivalence correlates strongly with word-level equivalence, we can approximate the former by looking at pairwise cosine similarity between the words in the sentence pair: $\cos(v_i^e, v_j^f)$. A good translation should tend to have a high value of $\max_j \cos(v_i^e, v_j^f)$ since most English words w_i^e (with vector v_i^e) should have a translationally equivalent word w_j^f (with vector v_j^f) in the other language, and these vectors should be similar.

However, this naive approach suffers from the so-called hubness problem in high-dimensional spaces (Radovanović et al., 2010), where some words tend to have high similarity to a large number of other words. This can be compensated for by taking the distribution of vector similarities for each word into account (as done in similar contexts by e.g. Conneau et al., 2017; Artetxe and Schwenk, 2018). We use this information in two ways. First, all words which have an average cosine similarity higher than 0.6 to the words in the English sentence are removed since they are unlikely to be informative. We then use as our score the ratio between the highest and the second highest similarity within the sentence, averaged over all remaining words in the sentence.[3]

[3]Sentences with vectors for less than half of their words are removed, since we are unable to make a reliable estimate.

	Nepali		Sinhala	
	Sentence Count	Word Count	Sentence Count	Word Count
1 Million	46,529	793,233	55,293	897,198
5 Million	272,605 (248,765)	3,737,250 (3,456,614)	250,767 (279,503)	4,119,591 (3,327,811)

Table 5: Word and sentence counts in the 1 million and 5 million sub-samples according to our model. Numbers in parenthesis refer to the counts of the baseline system (Xu and Koehn, 2017) which is only available only for 5 million sub-sample

5 Results

The quality of the sub-sampled data is assessed according to the BLEU scores of the statistical and neural machine translation systems trained on them.

Here, we present the BLEU scores of the NMT system (Guzmán et al., 2019) which will be used in the official evaluation on the released development set. We evaluate our model via two different sub-samples, one with 1 million and one with 5 million words on the English side. See Table 5 for statistics on the filtered data.

Table 4 presents our results using the NMT system. For Nepali, the performance of our model approaches the strong baseline on both the 1 million and 5 million sub-samples, whereas the NMT system fails completely using the 5 million word Sinhala sub-sample. All BLEU scores are below 6, for our system as well as for the baseline, indicating that there is insufficient data for the NMT system to learn a useful translation model.

6 Conclusion

We have described our submission to the WMT 2019 parallel corpus filtering shared task. Our submission explored the use of multilingual word embeddings for the task of parallel corpus filtering. The embeddings were projected from a high-resource language, to a low-resource language without sufficiently large monolingual corpora, making the approach suitable for a wide range of languages.

Acknowledgments

We would like to thank NVIDIA for their GPU grant.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Houda Bouamor and Hassan Sajjad. 2018. H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proc. Workshop on Building and Using Comparable Corpora*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

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.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Marcis Pinnis. 2018. Tilde's parallel corpus filtering methods for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 939–945, Belgium, Brussels. Association for Computational Linguistics.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950.

Ieva Zariņa, Pēteris Ņikiforovs, and Raivis Skadiņš. 2015. Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.