Low-Resource Corpus Filtering using Multilingual Sentence Embeddings

Vishrav Chaudhary♦ **Yuqing Tang**♦ **Francisco Guzmán**♦ **Holger Schwenk**♦ **Philipp Koehn**■
♦Facebook AI ■Johns Hopkins University
{vishrav,yuqtang,fguzman,schwenk}@fb.com phi@jhu.edu

Abstract

In this paper, we describe our submission to the WMT19 low-resource parallel corpus filtering shared task. Our main approach is based on the LASER toolkit (Language-Agnostic SEntence Representations), which uses an encoder-decoder architecture trained on a parallel corpus to obtain multilingual sentence representations. We then use the representations directly to score and filter the noisy parallel sentences without additionally training a scoring function. We contrast our approach to other promising methods and show that LASER yields strong results. Finally, we produce an ensemble of different scoring methods and obtain additional gains. Our submission achieved the best overall performance for both the Nepali–English and Sinhala–English 1M tasks by a margin of 1.3 and 1.4 BLEU respectively, as compared to the second best systems. Moreover, our experiments show that this technique is promising for low and even no-resource scenarios.

1 Introduction

The availability of high-quality parallel training data is critical for obtaining good translation performance, as neural machine translation (NMT) systems are less robust against noisy parallel data than statistical machine translation (SMT) systems (Khayrallah and Koehn, 2018). Recently, there is an increased interest in the filtering of noisy parallel corpora (such as Paracrawl[1]) to increase the amount of data that can be used to train translation systems (Koehn et al., 2018).

While the state-of-the-art methods that use NMT models have proven effective in mining parallel sentences (Junczys-Dowmunt, 2018) for high-resource languages, their effectiveness has not been tested in low-resource languages. The implications of low availability of training data for parallel-scoring methods is not known yet.

For the task of low-resource filtering (Koehn et al., 2019), we are provided with a very noisy 40.6 million-word (English token count) Nepali–English corpus and a 59.6 million-word Sinhala–English corpus crawled from the web as part of the Paracrawl project. The challenge consists of providing scores for each sentence pair in both noisy parallel sets. The scores will be used to subsample sentence pairs that amount to 1 million and 5 million English words. The quality of the resulting subsets is determined by the quality of a statistical machine translation (Moses, phrase-based (Koehn et al., 2007)) and the neural machine translation system `fairseq` (Ott et al., 2019) trained on this data. The quality of the machine translation system will be measured by BLEU score using SacreBLEU (Post, 2018) on a held-out test set of Wikipedia translations for Sinhala–English and Nepali–English from the `flores` dataset (Guzmán et al., 2019).

In our submission for this shared task, we use of multilingual sentence embeddings obtained from LASER[2] which uses an encoder-decoder architecture to train a multilingual sentence representation model using a relatively small parallel corpus. Our experiments demonstrate that the proposed approach outperforms other existing approaches. Moreover we make use of an ensemble of multiple scoring functions to further boost the filtering performance.

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

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

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 263–268
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

2 Methodology

The WMT 2018 shared task for parallel corpus filtering (Koehn et al., 2018)[3] introduced several methods to tackle a high-resource German-English data condition.While many of these methods were successful to filter out noisy translations, few have been tried under low-resource conditions. In this paper, we address the problem of low-resource sentence filtering using sentence-level representations and compare them to other popular methods used in high-resource conditions.

The LASER model (Artetxe and Schwenk, 2018a) makes use of multilingual sentence representations to gauge the similarity between the source and the target sentence. It has provided state-of-the-art performance on the BUCC corpus mining task and has also been effective in filtering WMT Paracrawl data (Artetxe and Schwenk, 2018a). However, these tasks only considered high-resource languages, namely French, German, Russian and Chinese. Fortunately, this technique has also been effective on zero-shot cross-lingual natural language inference in the XNLI dataset (Artetxe and Schwenk, 2018b) which makes it promising for the low resource scenario being focused in this shared task. In this paper, we propose to use an adaptation of LASER to low-resource conditions to compute the similarity scores to filter out noisy sentences.
For comparison to LASER, we also establish initial benchmarks using Bicleaner and Zipporah, two popular baselines which have been used in the Paracrawl project; and dual conditional cross-entropy, which has proven to be state-of-the-art for the high-resource corpus filtering task (Koehn et al., 2018). We explore the performance of the techniques under similar pre-processing conditions regarding language identification filtering and lexical overlap. We observe that LASER scores provide a clear advantage for this task. Finally, we perform ensembling of the scores coming from different methods. We observe that when LASER scores are included in the mix, the boost in performance is relatively minor. In the rest of this section we discuss the settings for each of the methods applied.

2.1 LASER Multilingual Representations

The underlying idea is to use the distances between two multilingual representations as a notion of parallelism between the two embedded sentences (Schwenk, 2018). To do this, we first train an encoder that learns to produce a multilingual, fixed-size sentence representation; and then compute a distance between two sentences in the learned embedding space. In addition, we use a *margin* criterion, which uses a k nearest neighbors approach to normalize the similarity scores given that cosine similarity is not globally consistent (Artetxe and Schwenk, 2018a).

Encoder The multilingual encoder consists of a bidirectional LSTM, and our sentence embeddings are obtained by applying max-pooling over its output. We use a single encoder and decoder in our system, which are shared by all languages involved. For this purpose, we trained multilingual sentence embeddings on the provided parallel data only (see Section 3.2 for details).

Margin We follow the definition of *ratio*[4] from (Artetxe and Schwenk, 2018a). Using this, the similarity score between two sentences (x, y) can be computed as

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

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

Neighborhood Additionally, we explored two ways of sampling k nearest neighbors. First a *global* method, in which we used the neighborhood comprised of the noisy data along with the clean data. Second a *local* method, in which we only scored the noisy data using the noisy neighborhood, or the clean data using the clean neighborhood.[5]

[3] http://statmt.org/wmt18/parallel-corpus-filtering.html

[4] We explored the *absolute*, *distance* and *ratio* margin criteria, but the latter worked best

[5] this last part was only done for training an ensemble

2.2 Other Similarity Methods

Zipporah (Xu and Koehn, 2017; Khayrallah et al., 2018), which is often used as a baseline comparison, uses language model and word translation scores, with weights optimized to separate clean and synthetic noise data. In our setup, we trained Zipporah models for both language pairs Sinhala–English and Nepali–English. We used the open source release[6] of the Zipporah tool without modifications. All components of the Zipporah model (probabilistic translation dictionaries and language models) were trained on the provided clean data (excluding the dictionaries). Language models were trained using KenLM (Heafield et al., 2013) over the clean parallel data. We are not using the provided monolingual data, as per default setting. We used the development set from the `flores` dataset for weight training.

Bicleaner (Sánchez-Cartagena et al., 2018) uses lexical translation and language model scores, and several shallow features such as: respective length, matching numbers and punctuation. As with Zipporah, we used the open source Bicleaner[7] toolkit unmodified out-of-the-box. Only the provided clean parallel data was used to train this model. Bicleaner uses a rule-based component to identify noisier examples in the parallel data and trains a classifier to learn how to separate them from the rest of the training data. The use of language model features is optional. We only used models without a language model scoring component.[8]

Dual Conditional Cross-Entropy One of the best performing methods on this task was dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018), which uses a combination of forward and backward models to compute a cross-lingual similarity score. In our experiments, for each language pair, we used the provided clean training data to train neural machine translation models in both translation directions: source-to-target and target-to-source. Given such a translation model M, we force-decode sentence pairs (x, y) from the noisy parallel corpus and obtain the cross-entropy score

$$H_M(y|x) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p_M(y_t|y_{[1,t-1]}, x) \quad (1)$$

Forward and backward cross entropy scores, $H_F(y|x)$ and $H_B(x|y)$ respectively, are then averaged with an additional penalty on a large difference between the two scores $|H_F(y|x) - H_B(x|y)|$.

$$\begin{aligned} \text{score}(x, y) = & \frac{H_F(y|x) + H_B(x|y)}{2} \\ & - |H_F(y|x) - H_B(x|y)| \end{aligned} \quad (2)$$

The forward and backward models are five-layer encoder/decoder transformers trained using `fairseq` with parameters identical to the ones used in the baseline `flores` model [9]. The models were trained on the clean parallel data for 100 epochs. For the Nepali-English task, we also explored using Hindi-English data without major differences in results. We used the `flores` development set to pick the model that maximizes BLEU scores.

2.3 Ensemble

To leverage over the strengths and weaknesses of different scoring systems, we explored the use of a binary classifier to build an ensemble. While it's trivial to obtain positives (e.g. the clean training data), mining negatives can be a daunting task. Hence, we use positive-unlabeled (PU) learning (Mordelet and Vert, 2014), which allows us to obtain classifiers without having to curate a dataset of explicit positive and negatives. In this setting our positive labels come from the clean parallel data while the unlabeled data comes from the noisy set.

To achieve this, we apply bagging of 100 weak, biased classifiers (i.e. with a 2:1 bias for unlabeled data vs. positive label data). We use support vector machines (SVM) with a radial basis kernel, and we randomly sub-sample the set of features for training each base classifier, helping keep them diverse and low-capacity.

We ran two iterations of training of this ensemble. In the first iteration we used the original positive and unlabeled data described above. For the second iteration, we used the learned classifier to re-label the training data. We explored several re-labeling approaches (e.g. setting a threshold that maximizes F_1 score). However, we found that setting a class boundary to preserve the original positives-to-unlabeled ratio worked best. We also observed that the performance deteriorated after two iterations.

[6] https://github.com/hainan-xv/zipporah

[7] https://github.com/bitextor/bicleaner

[8] We found that including a LM as a feature resulted in almost all sentence pairs receiving a score of 0.

[9] https://github.com/facebookresearch/flores#train-a-baseline-transformer-model

3 Experimental Setup

We experimented with various methods using a setup that closely mirrors the official scoring of the shared task. All methods are trained on the provided clean parallel data (see Table 1). We did not use the given monolingual data. For development purposes, we used the provided `flores` dev set. For evaluation, we trained machine translation systems on the selected subsets (1M, 5M) of the noisy parallel training data using `fairseq` with the default `flores` training parameter configuration. We report SacreBLEU scores on the `flores` devtest set. We selected our main system based on the best scores on the devtest set for the 1M condition.

	si-en	**ne-en**	**hi-en**
Sentences	646k	573k	1.5M
English words	3.7M	3.7M	20.7M

Table 1: Available bitexts to train the filtering approaches.

3.1 Preprocessing

We applied a set of filtering techniques similar to the ones used in LASER (Artetxe and Schwenk, 2018a) and assigned a score of -1 to the noisy sentences based on incorrect language on either the source or the target side or having an overlap of at least 60% between the source and the target tokens. We used fastText[10] for language id filtering. Since LASER computes similarity scores for a sentence pair using these filtering techniques, we experimented by adding these to the other models we used for this shared task.

3.2 LASER Encoder Training

For our experiments and the official submission, we trained a multilingual sentence encoder using the permitted resources in Table 1. We trained a single encoder using all the parallel data for Sinhala–English, Nepali–English and Hindi-English. Since Hindi and Nepali share the same script, we concatenated their corpora into a single parallel corpus. To account for the difference in size of the parallel training data, we over-sampled the Sinhala–English and Nepali/Hindi-English bitexts in a ratio of 5:3. This resulted in roughly 3.2M training sentences for each language direction, i.e. Sinhala and combined Nepali-Hindi.

The models were trained using the same setting as the public LASER encoder which involves normalizing texts and tokenization with Moses tools (falling back to the English mode). We first learn a joint 50k BPE vocabulary on the concatenated training data using fastBPE[11]. The encoder sees Sinhala, Nepali, Hindi and English sentences at the input, without having any information about the current language. This input is always translated into English.[12] We experimented with various techniques to add noise to the English input sentences, similar to what is used in unsupervised neural machine translation, e.g. (Artetxe et al., 2018; Lample et al., 2018), but this did not improve the results.

The encoder is a five-layer BLSTM with 512 dimensional layers. The LSTM decoder has one hidden layer of size 2048, trained with the Adam optimizer. For development, we calculate similarity error on the concatenation of the `flores` dev sets for Sinhala–English and Nepali–English. Our models were trained for seven epochs for about 2.5 hours on 8 Nvidia GPUs.

4 Results

From the results in Table 2, we observe several trends: (*i*) the scores for the 5M condition are generally lower than for the 1M condition. This condition appears to be exacerbated by the application of language id and overlap filtering. (*ii*) LASER shows consistently good performance. The *local* neighborhood works better than the *global* one. In that setting, LASER is on average 0.71 BLEU above the best non-LASER system. These gaps are higher for the 1M condition (0.94 BLEU). (*iii*) The best ensemble configuration provides small improvements over the best LASER configuration. For Sinhala–English the best configuration includes every other scoring method (ALL). For Nepali–English the best configuration is an ensemble of LASER scores. (*iv*) Dual cross entropy shows mixed results. For Sinhala–English, it only works once the language id filtering is enabled which is consistent with previous observations (Junczys-Dowmunt, 2018). For Nepali–English, it provides scores well below the rest of the scoring methods. Note that we did not perform an architecture exploration.

[10] 9https://fasttext.cc/docs/en/language-identification.html

[11] https://github.com/glample/fastBPE

[12] This means that we have to train an English autoencoder. This didn't seem to hurt, since the same encoder also handles the three other languages

Method	**ne-en**		**si-en**	
	1M	**5M**	**1M**	**5M**
Zipporah				
base	5.03	2.09	4.86	4.53
+ LID	5.30	1.53	5.53	3.16
+ Overlap	5.35	1.34	5.18	3.14
Dual X-Ent.				
base	2.83	1.88	0.33	4.63^{+}
+ LID	2.19	0.82	6.42	3.68
+ Overlap	2.23	0.91	6.65	4.31
Bicleaner				
base	5.91	2.54^{+}	6.20	4.25
+ LID	5.88	2.09	6.36	3.95
+ Overlap	6.12^{+}	2.14	6.66^{+}	3.26
LASER				
local	*7.37**	**3.15**	*7.49**	5.01
global	6.98	*2.98**	7.27	4.76
Ensemble				
ALL	6.17	2.53	**7.64**	**5.12**
LASER *glob. + loc.*	**7.49**	2.76	7.27	*5.08**

Table 2: SacreBLEU scores on the `flores` devtest set. In **bold**, we highlight the best scores for each condition. In *italics**, we highlight the runner up. We also signal the best non-LASER method with $^{+}$.

Submission For the official submission, we used the *ALL* ensemble for the Sinhala–English task and the LASER *global* + *local* ensemble for the Nepali–English task. We also submitted the LASER *local* as a contrastive system. As we can see in Table 3, the results from the main and contrastive submissions are very close. In one case, the contrastive solution (a single LASER) model yields better results than the ensemble. These results placed our 1M submissions 1.3 and 1.4 BLEU points above the runner ups for the Nepali–English and Sinhala–English tasks, respectively. As noted before, our systems perform worse on the 5M condition. We also noted that the numbers in Table 2 differ slightly from the ones reported in (Koehn et al., 2019). We attribute this difference to the effect of training in 4 (ours) gpus vs. 1 (theirs).

Method	**ne-en**		**si-en**	
	1M	**5M**	**1M**	**5M**
Main - Ensemble	6.8	2.8	**6.4**	4.0
Constr. - LASER *local*	**6.9**	2.5	6.2	3.8
Best (other)	5.5	**3.4**	5.0	**4.4**

Table 3: Official results of the main and secondary submissions on the `flores` test set evaluated with the NMT configuration. For comparison, we include the best scores by another system.

4.1 Discussion

One natural question to explore is how would the LASER method benefit if it had access to additional data. To explore this, we used the LASER open-source toolkit, which provides a trained encoder covering 93 languages, but does not include Nepali. In Table 4, we observe that the pre-trained LASER model outperforms the LASER *local* model by 0.4 BLEU. For Nepali–English the situation reverses: LASER *local* provides much better results. However, the results of the pre-trained LASER are only slightly worse that those of Bicleaner (6.12) which is the best non-LASER method. This suggests that LASER can function well in zero-shot scenarios (i.e. Nepali–English), but it works even better when it has additional supervision for the languages it is being tested on.

Method	**ne-en**		**si-en**	
	1M	**5M**	**1M**	**5M**
Pre-trained LASER	6.06	1.49	**7.82**	**5.56**
LASER *local*	**7.37**	**3.15**	7.49	5.01

Table 4: Comparison of results on the `flores` devtest set using the constrained and the pre-trained vesions of LASER.

5 Conclusions and Future Work

In this paper, we describe our submission to the WMT low-resource parallel corpus filtering task. We use of multilingual sentence embeddings from LASER to filter noisy sentences. We observe that LASER can obtain better results than the baselines by a wide margin. Incorporating scores from other techniques and creating an ensemble provides additional gains. Our main submission to the shared task is based on the best of the ensemble configuration and our contrastive submission is based on the best LASER configuration. Our systems perform the best on the 1M condition for the Nepali–English and Sinhala–English tasks. We analyze the performance of a pre-trained version of LASER and observe that it can perform the filtering task well even in zero-resource scenarios, which is very promising.

In the future, we want to evaluate this technique for high-resource scenarios and observe whether the same results transfer to that condition. Moreover we plan to investigate how the size of training data (parallel, monolingual) impact low-resource sentence filtering task.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Mikel Artetxe and Holger Schwenk. 2018a. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. *arXiv preprint arXiv:1811.01136*.

Mikel Artetxe and Holger Schwenk. 2018b. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv preprint arXiv:1812.10464*.

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*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 901–908, Belgium, Brussels. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Huda Khayrallah, Hainan Xu, and Philipp Koehn. 2018. The JHU parallel corpus filtering systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 909–912, Belgium, Brussels. Association for Computational Linguistics.

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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar Chris Dyer, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demo session*.

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, Volume 2: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 5039–5049, Belgium, Brussels. Association for Computational Linguistics.

Fantine Mordelet and J-P Vert. 2014. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209.

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 NAACL-HLT 2019: Demonstrations*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation (WMT), Volume 1: Research Papers*, volume 1804.08771, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Víctor M Sánchez-Cartagena, Marta Bañón, Sergio Ortiz Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

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 (Short Papers)*, pages 228–234, Australia, Melbourne. 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, Denmark, Copenhagen. Association for Computational Linguistics.