Empirical Evaluation of Language Agnostic Filtering of Parallel Data for Low Resource Languages

Praveen Dakwale[1] **Talaat Khalil**[2] **Brandon Denis**[3]
Huawei Technologies R&D
Amsterdam, Netherlands
praveen.dakwale@huawei.com[1],
khalil.talaat@gmail.com[2],
brandon.james.denis@huawei.com[3]

Abstract

Most of the available resources for low resource languages are crawled from the web. In order to obtain reasonable machine translation performance with such datasets, it is important to filter low quality samples from the training data. In this paper we explore the use of language agnostic sentence representations for filtering parallel data for low resource language pairs: Pashto-English, Khmer-English, Nepali-English and Sinhalese-English. We determine the quality of the samples based on embedding similarity between source and target sentences. Our experiments show that when preceded by language filtering using language agnostic embeddings significantly improves the performance of neural machine translation (NMT) and achieve performance competitive to language specific approaches.

1 Introduction

Neural machine Translation models are known to be data hungry (Koehn and Knowles, 2017). Training a high-quality NMT model requires a very large amount of data, usually in the order of millions of sentence pairs. For the majority of language pairs, parallel training data is compiled by aligning web-crawls in source and target languages using various heuristics based methods (Munteanu and Marcu, 2005). These web crawled datasets are often noisy due to alignment errors between source and target sentences. These misalignments lead to models with poor translation performance (Khayrallah and Koehn, 2018). Therefore, it is important to ensure the quality of the training data by filtering out noisy samples.

Various filtering techniques have been proposed in the machine translation literature which focus on different types of noise. Significant gains can be achieved by applying simple rules that do not take into consideration the semantic similarity between source and target sentences. Such rules include: removal of sentence pairs that are identified with language codes that are not aligned with source and target languages, length based pruning, removal of repetitive strings, and other heuristics (Barbu and Barbu Mititelu, 2018). However, a major type of noise is the non-equivalence of source and target sentences. Here, non-equivalence implies that the source and target sentences are not correct translations of each other (Khayrallah and Koehn, 2018). These non-equivalence cases are much harder to identify by simple heuristics. To detect noisy samples of this type, an oracle model is required to calculate the semantic similarity between source and target sentence pairs.

In the last few years, there has been a growing interest in developing advanced parallel data filtering techniques, resulting in a shared task for "parallel corpus filtering" focusing on high and low resource language pairs (Koehn et al., 2019; Koehn et al., 2020). The majority of these approaches focus on rule-based pre-filtering followed by scoring with an oracle model trained on high-quality parallel data for the same language pairs (Esplà-Gomis et al., 2020). One of the early reported approaches with high accuracy (Junczys-Dowmunt, 2018) is based on predictor models that are applied in both forward and reverse directions. The models are trained on a given "good quality" data and cross-

entropy scores from both models are combined to calculate the quality of a given "noisy" sentence pair. Such techniques rely on assumed "good quality" data to train the oracle models which is not always available, especially in case of low-resource languages. On the other hand, there has been considerable research on learning "language agnostic" sentence representations (embeddings) which aim to produce equivalent representations of semantically equal sentences across languages. These models are trained on large monolingual and multilingual data and are shown to generate reasonable language independent representations, including for languages not included in their training data.

In this paper, we investigated the use of two well-known "language agnostic" sentence embedding models to asses training samples quality, namely: "Language-Agnostic Sentence Representations" (LASER) (Schwenk and Douze, 2017) and "Language-agnostic BERT Sentence Embedding" (LaBSE) (Feng et al., 2020a). We conducted experiments on the low resource language pairs used in the "Parallel Corpus filtering" shared tasks in WMT-19 and WMT-20. The performance of different filtering methods was evaluated by training MT models on size varying datasets drawn from the top ranked sentence pairs.

2 Related work

Some recent works have explored the use of cross-lingual representations for parallel corpus filtering. (Herold et al., 2021) experimented with various agreement scores to compute source-target sentence similarity based on word embeddings. Another work in the same line explored the use of multilingual BERT for filtering parallel corpora (Zhang et al., 2020). The authors leveraged the ability of the aforementioned model to project multilingual sentences into a shared space. Both these works can be considered the closest to ours. However, our work offers different contributions which can be listed as follows:

- They have mainly experimented with high resource languages, on the other hand, we focused on the low resource language pairs from the WMT "Parallel Corpus filtering" task namely: Khmer-English (KM-EN), Pashto-English (PS-EN), Sinhalese-English (SI-EN) and Nepali-English (NE-EN).
- Our work is centered around comparing the most well established models for generating language agnostic sentence representations while these other works experimented with either word embedding based approaches or models that are not explicitly optimized to align latent representations of parallel sentence pairs. Such models are proven to be inferior to the models that we experiment with in cross-lingual similarity tasks.
- To be able to mimic real world settings, we conduct large scale experiments by evaluating the best performing methods on a dataset that is complied by combining all the publicly available datasets for a given language pair which we collect from OPUS (Tiedemann, 2012). Furthermore, we complement our work by conducting human evaluations on chosen languages pairs.

3 Data filtering

In this paper, our aim is to explore the use of language agnostic sentence embedding models as oracles to determine the translation quality of parallel sentences in training datasets. For this purpose, we obtain language independent representations for each source and target sentence and determine the translation quality by calculating the cosine similarity between these embeddings. Sentence pairs are scored, ranked and then filtered based on a similarity threshold or by selecting a pre-defined percentage of the original data size. We experiment with two well-known sentence representation models: Language Agnostic Sentence Representations (*LASER*) and Language agnostic Bert sentence embeddings (*LaBSE*). We briefly explain these representations in the following subsections.

3.1 LASER

Language-Agnostic Sentence Representations (LASER) (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019) is based on exploiting neural machine translation architectures to learn join representations for different language pairs. They use separate encoders and decoders for each language and encourage bridging the representation gap using a multi-task learning framework with various training configurations. Their configurations vary based on the number of used encoders and decoders per batch update as follows: one-to-one,

many-to-one, one-to-many, and many-to-many. Such training strategies are claimed to push the encoder representations of equivalent sentences in different languages closer to each other in the embedding space. At inference time, decoders are discarded and sentence embeddings are obtained by applying a max-pooling operation over the output of the encoder. LASER supports 93 languages belonging to 23 different scripts.

3.2 LaBSE

Language Agnostic Bert Sentence Embeddings (LaBSE) (Feng et al., 2020b), is based on combining some of the latest representation learning approaches namely: Masked Language Modeling (MLM) (Devlin et al., 2019), Translation Language Model (TLM) (Lample and Conneau, 2019), dual encoder translation ranking (Guo et al., 2018), and the use of additive margin softmax loss (Yang et al., 2019). Embeddings for source and target sentences were generated from a transformer model (Vaswani et al., 2017) pretrained using MLM and TML, they were then separately fed into a shared 12-layer transformer network. Finally, the output representations of parallel sentences were optimized to be similar to each other and distant from negative samples using an additive margin softmax loss. The latest LaBSE model supports 112 languages and is claimed to outperform LASER in parallel text retrieval.

4 Experimental setting

In WMT-19, the parallel corpus filtering task was organized for Sinhalese-English (SI-EN) and Nepali-English (NE-EN) and in WMT-20 the same task was organized for Khmer-English (KM-EN) and Pashto-English (PS-EN). In WMT-19, no scores were provided with the noisy training data to be used as a baseline system. In WMT-20, noisy training data was released along with the similarity scores computed using LASER embeddings to be used as the baselines.

We conducted two sets of experiments on the aforementioned language pairs. In the first set of experiments (see 4.1), we applied multiple filtering techniques on the noisy training datasets released by WMT parallel corpus filtering shared tasks, we then evaluated the performance of MT models trained on a varying number of the top ranked sentences according to the filtering systems as defined in the shared task descriptions. In the second set of experiments (see 4.2), we included additional training datasets that are publicly available and applied the approaches that resulted in the best results in our first set of experiments. We studied the performance of the MT systems trained on different data sub-samples to be able to determine an optimal similarity score threshold. The choice of the languages posed an interesting zero shot challenge because not all languages were used in training of both the models. As a result, not all four languages are supported by both the models. LASER only supports Khmer and Sinhalese, while LaBSE only supports Khmer, Nepali and Sinhalese. Pashto is not supported by either of the models. However it has been observed that both the models can be generalized well even to minority languages that are not supported by the models. This observation could possibly be attributed to transfer learning from languages in the training data which are closely related to these four languages.

	Year	Train	Valid	Test
KM-EN	WMT-20	4.1 m	2378	2320
PS-EN	WMT-20	1 m	3162	2719
SI-EN	WMT-19	3.3 m	2898	2766
NE-EN	WMT-19	2.2 m	2559	2835

Table 1: Parallel corpus filtering task WMT-19 and WMT-20 statistics. Data sizes are in number of sentence pairs.

4.1 WMT noisy data filtering

In this set of experiments, we aimed to replicate the setup of the "parallel corpus filtering" shared tasks in WMT-19 and WMT-20. Given a noisy corpus for a low resource language pair, the participating teams were required to submit quality scores for all the training samples. Filtering systems were then evaluated according to the performance of MT systems trained on varying size datasets (in terms of target tokens) sampled from the top ranked pairs according to the submitted quality scores. The MT systems were trained by the task organizers and used fixed model and hardware configurations to guarantee comparable assessment. Dataset statistics for each language pair are provided in Table 1

To establish a baseline, we train two MT systems on the entire data without any re-ranking or sub-sampling as follows:

- No filtering: The entire released noisy data is used for training without any filtering.

- Language ID filtering: Language filtering is applied to the training data using fastText toolkit (Joulin et al., 2016). Only the examples where the detected source and target language codes are not consistent with source and target languages are filtered out.

Sub-sampling experiments are setup as follows: sample sizes of [1, 2, 5] million target tokens are used for Nepali-English and Sinhalese-English and sample sizes of [2, 3, 5, 7] millions are used for Khmer-English and Pashto-English. Sub-sampling was performed using the script provided by the WMT organizers. The following filtering methods configurations are reported:

- Language ID filtering + LASER scoring: after applying language ID filtering, the remaining sentence pairs were scored and ranked according to the cosine similarity of the LASER embeddings.

- Language ID filtering + LaBSE scoring: same as the previous point but used LaBSE embeddings instead of LASER embeddings.

- Best performing systems in WMT tasks: To compare LASER and LaBSE filtering with the state-of-the-art language specific techniques, we selected AFRL (Erdmann and Gwinnup, 2019) for WMT-19 languages and Huawei (Açarçiçek et al., 2020) for WMT-20 languages. These models were chosen since they are the best performing models in the respective tasks where publicly available scores for the training samples are provided.[1]

4.2 Variable/mixed quality data filtering

Based on the first set of experiments as described in 4.1, we determined the best technique (out of LASER and LaBSE) for each language pair and applied it on a larger corpus that consists of samples of "unknown" quality. For each language pair, we first applied the standard language-id filtering, followed by scoring and re-ranking with the source-target embedding similarity score. We filtered the ranked corpus using different threshold values [0.5, 0.6, 0.7, 0.8, 0.9] of the similarity scores and trained NMT models on each filtered sub-sample. In this way we determined the best similarity threshold per language pair. We mainly used the datasets from OPUS repository and combined these datasets with the noisy corpus provided in the relevant WMT task. The details of the used datasets are provided in Table 2.

4.3 Training details

All the experiments in this paper were conducted using the same model architecture and training configuration. A TransformerBase model with the default configuration was trained using Opennmt-tf translation toolkit[2]. A Shared vocabulary of 10000 sub-words is trained using sentencepiece tokenizer(Kudo and Richardson, 2018). Models were trained for 1 million steps for the WMT data experiments and until convergence for the larger scale experiments. Convergence was defined as no significant change in the validation set performance according to BLEU scoring at 100,000 step increments.

Note that the NMT toolkit and the training configuration we used in this paper are different from those used in the WMT parallel corpus filtering tasks. This is because, in this paper, our purpose is not to do a direct comparison of a proposed methods with the results or methods reported in the WMT tasks but to empirically analyze utility of language agnostic embeddings for corpus filtering.

4.4 Evaluation metrics

BLEU score (Papineni et al., 2002) is the most commonly used automatic evaluation metric for machine translation performance. However, it has been recently criticized due its failure to correlate with human judgement. A recent study (Kocmi et al., 2021) conducted an extensive comparison of various MT evaluation metrics and found out that BLEU is inferior to other automatic metrics with respect to correlation to human judgements. They found that other metrics such as COMET (Rei et al., 2020) and ChrF (Popović, 2015) correlate much better. Therefore, following their recommendations, we reported model performances on COMET and ChrF in addition to BLEU. BLEU is calculated using sacreBLEU python implementation (Post, 2018).

ChrF is a character level n-gram F-score between generated translation and reference. Similar to BLEU, it calculates n-gram matches between

[1]Data can be downloaded from the websites of the respective WMT tasks

[2]https://github.com/OpenNMT/OpenNMT-tf

	Datasets	Sentences
KM-EN	CCAligned, GNOME, KDE4, Paracrawl, QED, wikimedia, XLEnt. WMT-20	4.8 M
PS-EN	CCAligned, GNOME, KDE4, Paracrawl, QED, wikimedia, XLEnt, WMT-20	1.4 M
SI-EN	CCAligned, CCMatrix, GNOME, KDE4, OpenSubtitles, Paracrawl, QED, Ubuntu, Wikimatrix, wikimedia, XLENT, WMT019	11 M
NE-EN	CCAligned, CCMatrix, GNOME, KDE4, OpenSubtitles, Paracrawl, QED, Ubuntu, Wikimatrix, wikimedia, XLENT, bible-uedin, WMT-19	2.1 M

Table 2: Combined data statistics for all language pairs for mixed data experiments.

translation and reference, however, at character level. ChrF is calculated using the same sacreBLEU implementation. For both BLEU and ChrF, statistical significance was measured using bootstrap re-sampling (Koehn, 2004) with 1000 samples. For the noisy-data-only experiments, we calculate statistical significance between LaBSE and LaSER based filtering as well as between best of language agnostic filtering with that of the competitor baseline as described in subsection 4.1. For the mixed-quality-data experiments, we compare the statistical significance between the no-filtering baseline with all other experiments (subsampling at various threshold values).

COMET is a neural network framework in which large pre-trained cross-lingual language models such as XLM-RoBERTa (Lample and Conneau, 2019) were fine-tuned on [source, hypothesis, reference] pairs in order to predict annotated human evaluation scores. We used a reference-based regression model which is built on top of XLM-R *wmt-comet-da*. This model covers all the languages in our study.

5 Results

5.1 Khmer-English

Table 3 shows the results for noisy data filtering experiment for KM-EN language pair. The performance of all models using the three evaluation metrics is monotonically consistent, i.e., higher performance with respect to one metric also means higher performance with respect to other metrics. The model achieves the lowest performance when trained on the entire data without any filtering. Filtering using language identification provides significant improvements. For sample sizes of 5 million and 7 million, sub-sampling based on LaBSE scoring performs the best, while for sample sizes of 2m and 3m, Huawei filtering (Açarçiçek et al., 2020) performs the best. This is consistent for both dev and test sets. Moreover, for all filtering methods, using a sample size of 7 million target tokens seems to perform the best, while using 2 million tokens seem to perform the worst.

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 58 million					
NF	4.2	17.9	-1.17	4.8	20.1	-1.02
	target tokens = 15 million					
lg	6.3	24.8	-0.94	6.9	27.6	-0.85
	target tokens =7 million					
HW	**8.7**	32.1	-0.60	**10.2**	36.9	-0.44
LS	6.1	30.1	-0.75	7.0	33.1	-0.65
LB	8.5†	**32.2**†	**-0.59**	**10.2**†	**37.3**†	**-0.42**
	target tokens =5 million					
HW	8.0	32.4	-0.61	10.0	37.0	-0.45
LS	6.6	29.8	-0.75	7.2	32.8	-0.68
LB	**8.4***†	**32.5***	**-0.58**	**10.9***†	**38.1***	**-0.40**
	target tokens =3 million					
HW	**8.7***	**32.9***	**-0.58**	**10.5***	**38.0***	**-0.43**
LS	5.7	28.6	-0.82	6.5	32.2	-0.74
LB	8.2†	32.4†	-0.60	9.9†	37.2†	-0.45
	target tokens =2 million					
HW	**8.0***	**32.1***	**-0.63**	**9.6***	**36.9***	**-0.49**
LS	4.7	27.1	-0.90	5.4	30.4	-0.83
LB	7.3†	30.9†	-0.67	8.7†	35.3†	-0.55

Table 3: KM-EN WMT noisy data filtering. **NF=** No filtering, lg = language id filtering only, **HW=**Huawei system, **LS** = language id + LASER scoring, **LB=** language id + LaBSE scoring. Values in bold indicate the highest ranking system for each subsample category. * represents a statistically significant comparison between HW and best of the language agnostic method and † represents the same between LASER and LaBSE at $p < 0.01$.

Given that LaBSE performs significantly better than LASER when using the noisy data, we applied LaBSE scoring along with language filtering on a combined set of noisy and clean training data as described in 4.2. Table 4 describes the results for the mixed data experiments for KM-EN. Filtering only on language ID drops the performance on the development set when compared to using the full training data but the performance remained al-

	devtest			test		
	BLEU	ChrF++	COME	BLEU	ChrF	COME
NF	7.9	23.5	-0.97	8.7	27.5	-0.78
lg	**6.2**	25.5	-0.93	8.8	**30.7**	-0.70
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	**10.2**	**33.9**	-0.52	**12.1**	**39.1**	-0.33
0.6	10.2	33.8	-0.51	12.1	39.4	-0.33
0.7	10.0	33.8	-0.52	11.8	39.6	-0.32
0.8	9.9	34.3	-0.52	11.1	39.7	-0.32
0.9	.6	32.7	-0.60	9.8	37.5	-0.45

Table 4: KM-EN mixed data experiments. τ = LaBSE similarity score threshold. All results are statistically significance wrt no filtering baseline. Values in bold indicate training corresponding to highest score.

most the same on the test set. However, when combined with LaBSE filtering, it provided significant improvements compared to no filtering at all. An embedding similarity score threshold of 0.6 seems to work the best on such dataset.

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 12.9 million					
NF	7.7	31.6	-0.66	5.9	28.1	-0.80
	target tokens = 7.3 million					
lg	7.4	30.5	-0.67	5.7	27.2	-0.81
	target tokens = 7 million					
HW	**10.7**	**37.1**	**-0.43**	**8.8**	**34.2**	**-0.55**
LS	7.9	31.1	-0.65	5.8	27.7	-0.81
LB	8.2	31.7	-0.62	6.4	28.8	-0.75
	target tokens = 5 million					
HW	**10.2**	**37.3**	**-0.42**	**8.7**	**35.0**	**-0.52**
LS	7.3	31.6	-0.66	5.6	28.9	-0.78
LB	9.7	35.7	-0.47	8.0	33.2	-0.58
	target tokens = 3 million					
HW	**10.1***	**37.0***	**-0.43**	**9.3***	**35.4***	**-0.53**
LS	7.2	31.7	-0.68	6.0	30.2	-0.76
LB	**10.1**	**37.0**	-0.44	9.2	35.2	-0.54
	target tokens = 2 million					
HW	**9.3**	**35.9**	-0.49	**8.4**	**34.2**	-0.58
LS	6.4	31.3	-0.71	5.7	30.2	-0.77
LB	9.3†	35.5†	-0.51	8.3†	33.9†	-0.60

Table 5: PS-EN WMT noisy data filtering. Legends have same meaning as Table 3.

5.2 Pashto-English

Table 5 summarizes the results for noisy data experiments for PS-EN. Applying only language ID filtering causes some slight performance drop as compared to using the entire training data. Scoring and sub-sampling using LASER embeddings performs the worst for all sub-sampled sizes. For sample sizes of 7m and 5m, Huawei filtering technique performed the best while for sample sizes of 3m and 2m, both Huawei and LaBSE based filtering perform almost equally. The differences in performance between models are consistent across metrics and test sets. Regardless of the fact that Pashto is not supported by LaBSE, it's performance is comparable to the language specific filtering technique (Huawei).

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
NF	10.3	34.5	-0.52	8.3	31.0	-0.67
lg	8.8	32.3	-0.60	6.7	28.9	-0.75
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	11.0	37.8	-0.39	9.8	36.0	-0.48
0.6	10.7	37.8	-0.40	9.8	36.1	-0.48
0.7	**11.2**	**38.3**	-0.37	**9.9**	**36.4**	-0.48
0.8	10.7	37.7	-0.42	9.5	35.9	-0.50
0.9	6.1	30.8	-0.73	5.8	29.8	-0.80

Table 6: PS-EN mixed data experiments. Legends have same meaning as Table 4.

LaBSE performed significantly better than laser on the noisy data experiments therefore, for the experiments with mixed quality data, we applied LaBSE based filtering. As observed in Table 6, for mixed data experiments, filtering only with the language ID seemed to drop the performance significantly. However, applying language ID filtering in combination with LaBSE based scoring with similarity thresholds in the range of [0.5, 0.8] provided substantial improvements as compared to using all the training data. However, with a score threshold of 0.9, the performance dropped even below that of only language ID filtering which implies that with a very high similarity threshold, a substantial amount of useful training samples get filtered out.

5.3 Sinhalese-English

Table 7 presents the results for Sinhalese-English noisy data filtering. An important observation for this language pair is the very low performance when no filtering is applied which might indicate high noise level in the crawled dataset. The performance drops further with language ID filtering. However, for this language pair, scoring with LaBSE outperforms both LASER as well as the best reported language specific approach (AFRL). The difference in scores is consistent across test sets as well as metrics. The best performance

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 45 million					
NF	3.8	21.7	-0.86	3.1	19.7	-0.91
	target tokens = 11 million					
lg	2.8	21.3	-0.91	2.0	19.9	-0.93
	target tokens =5 million					
AF	5.8	30.4	-0.53	5.2	29.5	-0.52
LS	**6.1***	**31.8***†	**-0.48**	**5.6***	**31.3**$^{*}\dagger$	**-0.45**
LB	6.0†	31.3†	-0.50	5.4†	30.4†	-0.47
	target tokens =2 million					
AF	5.5	30.5	-0.57	5.0	29.8	-0.55
LS	5.7	32.0	-0.51	5.4	31.6	-0.50
LB	**7.3**$^{*}\dagger$	**34.2**$^{*}\dagger$	**-0.39**	**6.8**$^{*}\dagger$	**33.7**$^{*}\dagger$	**-0.37**
	target tokens =1 million					
AF	4.3	28.6	-0.64	4.0	28.3	-0.62
LS	3.3	27.4	-0.70	3.1	27.4	-0.67
LB	**6.4**$^{*}\dagger$	**32.5**$^{*}\dagger$	**-0.47**	**5.6**$^{*}\dagger$	**31.6**$^{*}\dagger$	**-0.46**

Table 7: SI-EN WMT noisy data filtering. **AF** = AFRL filtering. Other legends have same meaning as Table 3.

for the LaBSE model was observed with 2m samples which supports our hypothesis that the initial dataset is of lower quality than the other language pairs that we experimented with. Since LaBSE

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
NF	18.2	46.7	0.11	16.3	43.4	0.03
lg	19.2	47.3	0.14	16.5	43.7	0.05
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	19.9	49.0	0.21	18.8	48.2	0.22
0.6	**20.2**	**49.4**	0.22	**19.3**	**48.7**	0.23
0.7	20.2	49.5	0.23	19.0	48.6	0.23
0.8	19.5	49.0	0.21	18.5	48.2	0.22
0.9	15.8	45.6	0.06	14.9	44.7	0.06

Table 8: SI-EN mixed data experiments. Legends have same meaning as Table 4.

based filtering performed the best for the noisy data experiments, we applied it for mixed dataset filtering. Table 8 shows the mixed data filtering results for Sinhalese-English. The absolute scores using all metrics are substantially higher than those for the previous two language pairs. Simply applying language ID filtering provided significant improvements compared to using all the data without filtering. Further filtering using LaBSE provided additional improvements for all threshold values except for the threshold value of 0.9. The highest performance was observed when using a threshold score of 0.7.

5.4 Nepali-English

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
	target tokens = 35 million					
NF	0.7	13.2	-1.25	1.0	13.9	-1.17
	target tokens = 9 million					
lg	1.5	18.2	-1.07	1.8	18.9	-1.00
	5 million					
AF	**2.7***	**23.0***	**-0.85**	**2.8***	**24.5***	**-0.78**
LS	2.1	22.0	-0.92	2.7	23.8	-0.82
LB	2.4†	22.3†	-0.86	2.5†	23.5†	**-0.78**
	target tokens = 2 million					
AF	3.6	26.5	-0.78	3.8	28.5	-0.69
LS	2.4	24.2	-0.88	2.7	25.8	-0.78
LB	**5.2***†	**29.6***†	**-0.62**	**5.9***†	**31.6***†	**-0.54**
	target tokens = 1 million					
AF	2.7	25.4	-0.82	2.9	27.3	-0.75
LS	0.8	19.6	-1.03	1.2	20.7	-0.98
LB	**5.2***†	**29.4***†	**-0.64**	**6.1***†	**31.6***†	**-0.55**

Table 9: NE-EN WMT noisy data filtering. **AF** = AFRL filtering. **AF** = AFRL filtering. Other legends have same meaning as Table 3.

Table 9 shows the Nepali-English noisy data results. The absolute scores without filtering are the lowest when compared to other languages. Applying language ID filtering slightly improved the performance as compared to no filtering. LaBSE filtering performs significantly better that the other methods according to the majority of the evaluation metrics when using sample sizes of 1m and 2m samples. The language specific approach performs better than LaBSE on the 5m sub-sample however, the results on this sub-sample were the worst for all the approaches.

LaBSE based filtering results for the mixed quality dataset are presented in Table 10. The best performance was observed using a similarity threshold of 0.8. Consistent with the observations

	devtest			test		
	BLEU	ChrF	COME	BLEU	ChrF	COME
NF	9.1	30.1	-0.61	10.7	33.3	-0.50
lg	8.6	29.4	-0.63	10.3	32.6	-0.52
τ	BLEU	ChrF	COME	BLEU	ChrF	COME
0.5	11.3	36.5	-0.33	12.8	39.9	-0.19
0.6	11.6	38.0	-0.27	14.1	42.2	-0.12
0.7	12.3	39.4	-0.22	14.7	43.4	-0.09
0.8	**12.8**	**40.4**	-0.19	**15.3**	**44.3**	-0.07
0.9	10.9	38.1	-0.32	12.9	41.8	-0.19

Table 10: NE-EN mixed data experiments. Legends have same meaning as Table 4.

for other language pairs, a very high threshold of 0.9 dropped the performance significantly as compared to other lower values.

5.5 Human evaluation

In order to further verify the certainty of model performances calculated using automatic scores, we additionally performed human evaluations for some experiments. Due to the low availability of human evaluators, we performed human evaluations only for Pashto-English and Sinhalese-English for WMT noisy data experiments corresponding to the results reported in Table 5 and 7 for the 5 million sub-sample task. For Pashto-English and Sinhalese-English, we randomly sampled 100 sentences from the development set which were rated by native speakers of the corresponding languages. The raters were directed to assign an integer adequacy score between [1,5] to each hypothesis translation (Koehn and Monz, 2006). The final average scores are shown in Table 11. For both language pairs, as expected, the reference translations scored the highest. For Pashto-English, the Huawei filtering method scored significantly higher than both LASER and LaBSE based filtering. For Sinhalese-English, while LASER model performed significantly lower, Huawei and LaBSE filtering performed approximately equally. These observations for both of the language pairs were consistent with the automatic evaluations in Table 5 and 7.

6 Discussion

In this paper, we presented an empirical evaluation of the use of language agnostic sentence representations to filter parallel data for low resource neural machine translation. Our experiments show that using similarity scores based on language agnostic embeddings to compute the quality of the sentence pairs performs competitively when compared to state-of-the-art language specific techniques for low resource languages.

Filtering out sentences based on automatic language detection seems to give inconsistent results, we think that this happens because of the accuracy differences of the used language detection tool across different languages. Further analysis needs to be done for better understanding.

Data filtering thresholds based on the similarity score or a pre-defined number of tokens seems to vary across languages and datasets. This can be attributed to two main factors namely: The inherent quality of the dataset and the performance of the cross-lingual embeddings when it comes to the language pair under evaluation. Further analysis needs to be conducted to understand the per language pair effects.

Based on our experiments, language agnostic approaches perform competitively and provide a simple and a hassle-free way of filtering parallel datasets. However, this isn't the case when the language pair is not supported by the cross-lingual embeddings models as shown in the PS-EN experiments. Further research is needed to develop and test approaches for incremental language addition to the cross-lingual embedding based models.

References

Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Açarçiçek, Haluk, Talha Çolakoğlu, pınar ece aktan hatipoğlu, Chong Hsuan Huang, and Wei Peng. 2020. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation*, pages 940–946, Online, November. Association for Computational Linguistics.

Barbu, Eduard and Verginica Barbu Mititelu. 2018. A hybrid pipeline of rules and machine learning to filter web-crawled parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 867–871, Belgium, Brussels, October. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Erdmann, Grant and Jeremy Gwinnup. 2019. Quality and coverage: The afrl submission to the wmt19 parallel corpus filtering for low-resource conditions task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 269–272, Florence, Italy, August. Association for Computational Linguistics.

Esplà-Gomis, Miquel, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of*

	Reference	Huawei/AFRL	LASER	LaBSE
Pashto-English	4.4	2.4	1.7	1.8
Sinhalese-English	4.6	2.26	1.88	2.23

Table 11: Human evaluation average scores for Pashto-English and Sinhalese-English

the Fifth Conference on Machine Translation, pages 952–958, Online, November. Association for Computational Linguistics.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020a. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020b. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Guo, Mandy, 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 *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium, October. Association for Computational Linguistics.

Herold, Christian, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2021. Data filtering using cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–172, Online, June. Association for Computational Linguistics.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Junczys-Dowmunt, Marcin. 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, October. Association for Computational Linguistics.

Khayrallah, Huda 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, July. Association for Computational Linguistics.

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, Online, November. Association for Computational Linguistics.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, Philipp and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Koehn, Philipp, Francisco Guzmán, Vishrav Chaudhary, and Juan 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 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy, August. Association for Computational Linguistics.

Koehn, Philipp, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco GuzmÃ¡n. 2020. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online, November. Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Kudo, Taku and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, Eduardo and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. ACL '02, page 311–318, USA. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Schwenk, Holger and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada, August. Association for Computational Linguistics.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, Yinfei, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In Kraus, Sarit, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5370–5378. ijcai.org.

Zhang, Boliang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online, July. Association for Computational Linguistics.