Filtering Noisy Parallel Corpus using Transformers with Proxy Task Learning

Haluk Acarcicek[1], Talha Colakoglu[1], Pinar Ece Aktan[1], Chongxuan Huang[2], Wei Peng[2] *
[1]Turkey R&D AI Enablement Department, Huawei Technologies
{haluk.acarcicek,talha.colakoglu,ece.aktan.hatipoglu}@huawei.com
[2]Artificial Intelligence Application Research Center, Huawei Technologies
{huang.chongxuan,peng.wei1}@huawei.com

Abstract

This paper illustrates Huawei's submission to the WMT20 low-resource parallel corpus filtering shared task. Our approach focuses on developing a proxy task learner on top of a transformer-based multilingual pre-trained language model to boost the filtering capability for noisy parallel corpora. Such a supervised task also helps us to iterate much more quickly than using an existing neural machine translation system to perform the same task. After performing empirical analyses of the finetuning task, we benchmark our approach by comparing the results with past years' state-of-the-art records. This paper wraps up with a discussion of limitations and future work. The scripts for this study will be made publicly available. [1]

1 Introduction

Crawling web has been regarded as a *de facto* approach to produce bitexts, yet the crawled texts are under-qualified often in some aspects to train a proper machine translation system. Under-qualified bitexts present misalignments, no alignments, wrong language pairs, sentences mostly composed of numbers and mathematical formulas, etc. Parallel corpus filtering in this manner holds a critical research area to improve the performance of machine translation systems. WMT organizes a shared task for parallel corpus filtering since 2018 intending to filter our noisy bitexts to this end. The challenge targets low-resource language pairs since 2019.

Many existing filtering methods require multiple layers of elimination by implementing manually engineered features such as length filtering, language identification, normalizing, etc. These hand-picked features work well for a language pair but don't generalize well to another language pair or domain and often bring algorithmic complexity to the overall system.

The LASER (Artetxe and Schwenk, 2019) model achieved state-of-the-art (SOTA) records at the WMT19 shared task on low-resource parallel corpus filtering (Chaudhary et al., 2019). The sentence representation model implemented in LASER provides a means for measuring the similarity between a source and a target sentence. As stated in the future work at Artetxe and Schwenk (2019), there is still space to improve. Utilizing a self-attention mechanism remains future work as the LASER was not built upon the latest transformer architecture (Vaswani et al., 2017). We are also interested in designing a filtering tool that can be efficiently applied to a wide range of language pairs. Pre-trained multilingual language models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), are exploited to this end.

We make two contributions to the field in this manner. The first contribution is a proposal of approaching the filtering problem as a discrimination task that can be trained with a proxy task and synthetic training data generation (see in Section 3.1). The other contribution is the empirical knowledge learned from an analysis of the finetuning pre-trained multilingual language models on cross-lingual discrimination tasks.

2 Related Work

In the WMT18 shared task, participants mostly used similar techniques in components as prefiltering, scoring the sentence pairs, and using a classifier for feature functions. Teams applied prefiltering rules to eliminate noisy data, including:

- short or lengthy sentences;
- sentence pairs with few words and unbalanced token lengths;

*Corresponding author
[1]https://github.com/WPti/proxy-filter

Proceedings of the 5th Conference on Machine Translation (WMT), pages 940–946
Online, November 19–20, 2020.

- sentence pairs with unmatched names, numbers, web addresses, etc.;

- sentences where a language identifier fails to identify a source or target language type.

Scoring functions were mostly used to correlate qualified texts. Participants also used sentence embeddings (Bouamor and Sajjad, 2018; Axelrod et al., 2011; Artetxe and Schwenk, 2019) altogether with a similarity function to detect the similarity of pairs. The WMT19 shared task focused on low-resource languages, namely Nepali-English and Sinhala-English. Participants mostly applied basic filtering techniques similar to those used in 2018. Chaudhary et al. (2019) used sentence embeddings that were trained on parallel sentence pairs. Another approach was to train a machine translation system on the clean data and then used it to translate the non-English side to make a comparison. Several metrics were used to match sentence pairs such as METEOR, Levenshtein distance, and BLEU.

We found that our work relates to the submission from Bernier-Colborne and Lo (2019). However, their submission was unable to show the effectiveness of the proposed method due to potential issues in the pretraining process. Besides the parallel corpus filtering task, we come across several works utilizing a similar approach. In Yang et al. (2019), BERT rescoring method is more effective at bitext mining than heuristic scoring methods, i.e., marginal cosine distance. In Grégoire and Langlais (2018), a similar negative random sampling technique has been used for generating synthetic bad pairs. Also, attempts to create harder negative pairs were proven effective in bitext mining (Guo et al., 2018).

3 Methodology

Transformer models are currently state-of-the-art systems on most NLP classification and regression tasks. With the emergence of multilingual pretrained models, their cross-lingual capabilities can be exploited with little effort.

3.1 Proxy Task

To treat this problem as a supervised one, we design a proxy learner to model this task. The correctly aligned pairs can be regarded as positive samples in a simple sense for binary classification.

Most of the noise in the corpus originate from ill-aligned sentence pairs. The intuitive idea is to treat the misalignments as synthetic negative samples for our proxy task learner.

Taking random samples of the target sentences for all source sentences was the easiest way to create negative samples. But this results in an easily-classifiable training data which offers little assistance to the low-resource bitext filtering task. We need to create more valuable training data, which is referred to as harder examples.

3.1.1 Generating Harder Examples

Instead of training transformers with easily-discernible random negative samples, we need to create harder examples to confuse the model to boost its performance on the filtering task. We try the following ways to generate harder examples:

Neighborhood Awareness The neighbor sentences in the corpus have a higher chance of sharing common semantics and topics than those randomly extracted from corpus-wide. Alignment slips are most likely to occur in this context. This concept of neighborhood awareness inspires us to generate harder training data. For every positive pair, we create two negative pairs by pairing adjacent sentences of that target sentence with the source sentence. Incorporating this simple strategy may help to boost filtering performance.

Fuzzy String Matching Sampling Instead of randomly sampling negative examples from bitexts, we develop a new sampling strategy inspired by KNN (the k-nearest neighbors algorithm). To create harder examples for finetuning, we sampled lexically similar but semantically different sentences using a fuzzy string search method. [2] For each one of the source sentences (S), we perform a fuzzy search and identify the N similar sentence respecting to the fuzzy string score (F). We set a limit (L) on the F and ignore sentences with similarities over this limit (L) to avoid duplicated or highly related candidates. Then we pair the corresponded target sentences of those N similar sentences with the source sentences to create N negative pairs. We apply a setting with an L value of 60 (in a 100 scale) and N values of 2 and 3 to generate the validation and training data.

Model Architecture	Siamese	Finetuning
Bert-base-Multi-cased	0.62	0.69
Xlm-Roberta-Base	0.84	0.86
Xlm-Roberta-Large	0.88	0.92

Table 1: Model performances on proxy task as accuracy in F1 scores.

3.1.2 Architecture

We explore two candidate architecture in this study, one of which is a Siamese network (Reimers and Gurevych, 2019). The other model is a pre-trained transformer with a binary classification learner to differentiate ok-aligned sentence pairs with their negative counterparts. A comparison between the performance of architecture can be seen in the Table 1.

Sentence Transformers Reimers and Gurevych (2019) adopt a Siamese architecture, which allows us to feed sentence pairs separately to a transformer network like BERT. Each sentence pairs are encoded into fixed-size embeddings connected to a classifier network. Embeddings can be compared using a cosine similarity function at the inference stage. We reach on par performance to the LASER in the WMT19 parallel corpus filtering task (Table 3).

Transformer Finetuning with Pair Classification BERT is a language model introduced by Devlin et al. (2018). A pre-trained BERT model can be finetuned by adding an extra output layer to address many NLP tasks. One of BERT's derivatives is RoBERTa (Liu et al., 2019), and it is essentially very similar to its successor in structure. The authors of RoBERTa discarded the next sentence prediction (NSP) task and altered the mask language modeling task.

We compare multilingual variations of BERT and RoBERTa, which contains both Khmer (km) and Pashto (ps) monolingual data in the pretraining. The multilingual version of the RoBERTa, aka XLM-R (Conneau et al., 2019), performs far superior as it leverages more data in training (Table 1).

3.1.3 Amount of Parallel Data

To observe the effect of the amount of the available parallel corpus on this proxy learner's performance, we try two different data regimes. The orange line in Figure 1 represents a very low resource setting, and we subsampled 2k parallel pairs to mimic that. The blue line represents a 10k subsampled version of the training data. As can be seen from Figure 1, the more we increase the number of parallel sentences used in training the proxy task, the more performance we observe for the proxy task. Other than that, a system using almost as little as 2k parallel sentence pair is enough to beat the benchmark results. The proposed approach is promising for other low-resource domains and applications.

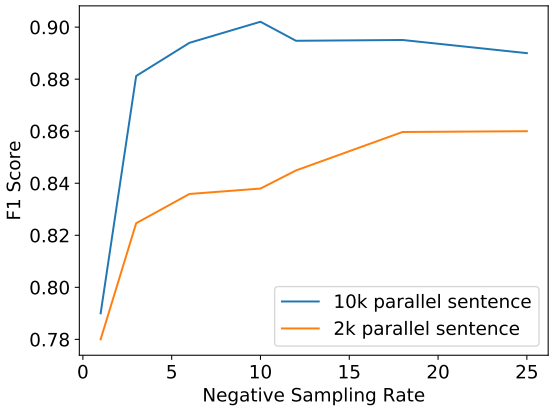

Figure 1: Proxy task validation performance in the face of changing volumes of training data (Pashto - English).

3.1.4 Negative Sampling Ratio

The amount of negative data that can be used in training is analyzed in the prior works (Section 2). Into our observations from Figure 1 and Figure 2, using larger negative ratios leads to better performances. However, it is better to keep the positive/negative ratio to $1:10$ for our datasets with a presence of more parallel data.

We oversample the positive pairs in the finetuning step to balance the positive-negative ratio. But it didn't make a noticeable change in proxy task performance or filtering performance. The immunity of the pre-trained transformer models to the class-imbalance up to 20x is very surprising.

3.1.5 Learning Rate

To prevent the catastrophic forgetting problem in the transformers, we apply a very small ($2e^{-6}$) learning rate with the inverse root scheduler and a warmup step of $1,000$. We also try other learning rate schedulers like cyclic learning rate scheduler (CLR) from (Lee et al., 2020) but couldn't observe any benefit for this task. We suspect CLR may not

[2]https://github.com/seatgeek/fuzzywuzzy

apply to a finetuning process with a small epoch number (i.e., 2 epochs in this study).

3.1.6 Finetuning and Scoring

We add a classification layer on top of XLM-R having $2,048$ hidden units with RELU activations and dropout. On single Nvidia V100 GPU, we finetune our models for 2 epochs without any early stopping. It takes about 6 hours to finetune on the generated datasets. The scoring step is just getting the probability of that pair being positive. Scoring a sentence pair takes $5ms$ on average.

3.2 Rescoring

Bidirectional Scoring Similar to the bidirectional scoring in Chaudhary et al. (2019), we reverse source and target sentences and train two different networks, which produce two different scores (SRC-TRG and TRG-SRC) for a pair. We then combine these two scores under (min, mean, max) strategies. In the "min" strategy, we aim to filter false-positive pairs by keeping the lowest score from the (SRC-TRG and TRG-SRC) for each pair. In "max" strategy, we use the highest score for each pair. And in the "mean" strategy, an average of the scores are applied. We observe that filtering on the "max" score can turn some of the false-negative sentences into true-positives, which increases NMT performance (Table 2).

Strategy	BLEU
SRC-TRG	12.97
TRG-SRC	12.65
Mean	12.42
Min	12.93
Max	13.17

Table 2: NMT results of systems trained after filtering based on different bidirectional scoring strategies (Pashto - English)

Ensembling We ensemble our top 3 trained transformer models under (min, max, mean) strategies and observe a minor improvement on the Pashto-English (ps-en) dataset. On the Khmer-English (km-en) dataset, there is no improvement (Table 3).

3.3 Heuristic Filters

Heuristic filters like overlap filters, length ratio, min-max length, and language identification are applied. For the Pashto-English setup, this step is not beneficial to the overall performance. For the Khmer-English setup, we observe a minor gain (Table 3). It appears that our scoring method can learn heuristic filtering on the fly without reliant on hard-coded heuristic filters.

4 Results

There is a relationship between F1 scores of the proxy task and the final NMT system performance (Figures 1-2). Improvements of the final NMT in the proxy task peaks along with the negative sampling rate and decreases potentially due to overfitting. By looking at the same ratio presented in Figures 1-2, we can conclude a correlation between the performance of the proxy task and that for the filtering task, showcasing the proposed approach's effectiveness.

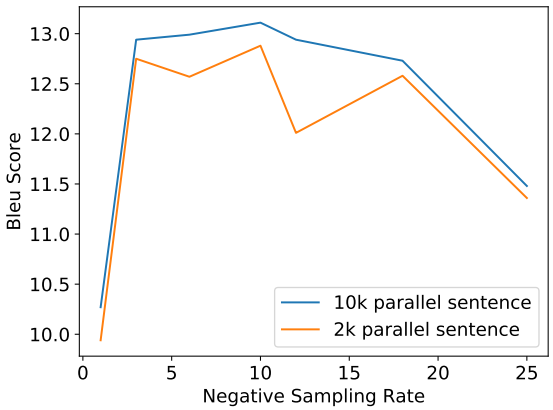

Figure 2: MBART performance of the filtering model (Pashto - English).

WMT20 Here we have presented our NMT performances of the submitted filtering systems in Table 3. Note that we measure all of the development cycles and improvements with the MBART finetuning (Liu et al., 2020). We do not replicate every experiment with training from scratch regime due to resource constraints. As shown in Table 3, our method outperforms the LASER baseline without needing any prefiltering rules and costly marginal KNN scoring method in solving the hubness problem for both language settings.

4.1 Older Tasks

To find how our method generalizes across different filtering scenarios, we test it for the past generations of this shared task.

WMT18 We use the same neural machine translation system defined by the organizers. Our NMT

Method	Pashto-English		Khmer-English	
	Scratch	MBART*	Scratch	MBART*
Baseline(LASER)	9.6	12.2	7.1	10.4
Sentence Transformers	9.7	12.5	7.5	10.6
XML-R finetuning	10.1	12.6	7.7	10.8
+Neighbourhood Awareness	-	12.9	-	-
+Fuzzy String Matching	-	13.0	-	-
+Bidirectional Scoring	-	13.2	-	11.5
+Ensemble Scoring	**10.9**	**13.3**	-	11.5
+Heuristic Filters (3.3)	10.7	13.2	**8.7**	**11.7**

Table 3: NMT scores (BLEU) of the models that trained on a corpus filtered by the specified methods on WMT20 test sets. The bold fonts indicate the SOTA results. * indicates finetuning of the pretrained MBART model which is provided by the organizers.

model using the submission by Junczys-Dowmunt (2018) couldn't reach the reported scores (can be observed in Table 4 for the $10M$ subsampled set). Although our method couldn't match the SOTA results under these settings, it achieves a reasonable score. Note that we only used 10% of the available clean parallel data to accomplish this result. Also, instead of finetuning a multilingual pre-trained model, bilingual models can be tried to avoid the curse of multilinguality (Conneau et al., 2019).

WMT19 Our NMT model using the submission by Chaudhary et al. (2019) couldn't reach the reported scores, as shown in Table 4 for the Nepalese-English (ne-en) set. The mismatches mentioned above with WMT18 and WMT19 are possibly due to a result of using multiple GPUs with distributed optimizers like stated in Koehn et al. (2019). In the low-resource setting, our method can surpass the SOTA results (Table 4).

Task	SOTA	OURS
WMT18 (de-en)	***27.9 (28.62)**	27.53
WMT19 (ne-en)	*6.9 (7.1)	**7.5**

Table 4: WMT18 and WMT19 filtering tasks test results. Note that numbers with "*" represent the submitted score performance under our NMT setup. Those in parenthesis are the reported scores.

5 Conclusions and Future Work

We illustrate our submission to the WMT20 low-resource parallel corpus filtering task. By developing a proxy task learner on top of a transform-based pre-trained language model XLM-R, We are able to improve the filtering capability for noisy data, achieving SOTA results.

The parallel corpus filtering task is recall-oriented. Therefore our model may not be suitable for high-precision jobs. The model has limitations in dealing with short sentences. It can be improved by finetuning on dictionaries or phase-based bitexts. The model performs better in low-resource and high-recall settings.

In our experiments depicted in the subsection 3.1.6, we observe low performances several times. It may appear the model is suffering from the random seeds caused fragility mentioned in Risch and Krestel (2020). A close look ascribes these abnormal results to the randomness in the sampling strategy. We leave this issue to future work.

Different kinds of synthetic noise generation techniques can be adapted to increase the robustness and accuracy of the model. For example in the filtered data we observed several false-positive cases which contains mis-translated numbers:

en reference:

"3) Sonar coverage: 45K at 200KHz"

ps to en translation:

"4) Sonar coverage: 90 at 125KHz"

Training an NMT model on this type of data hurts the translation performance. But this kind of noise can be fixed by altering the numerical values in the clean training data to sample negative pairs for our proxy task. Moreover, all the other synthetically generatable errors like a typo error, one to many alignment errors, etc. can be incorporated into the training data. But its not viable to model those kinds of errors independent from

the language or domain with the naive assumptions and inventing heuristic rules. We believe further researches should focus on domain invariant noise generation techniques.

Acknowledgments

We would like to show our gratitude to colleagues from HTRDC AIE and AARC, Huawei for their support during this work.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Gabriel Bernier-Colborne and Chi-kiu Lo. 2019. NRC parallel corpus filtering system for WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 252–260, Florence, Italy. Association for Computational Linguistics.

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*.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. *CoRR*, abs/1806.05559.

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

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

Philipp Koehn, 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. Association for Computational Linguistics.

Cedric K. M. Lee, Jianfeng Liu, and Wei Peng. 2020. Applying cyclical learning rate to neural machine translation. *ArXiv*, abs/2004.02401.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julian Risch and Ralf Krestel. 2020. Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, Marseille, France. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yinfei Yang, 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. *CoRR*, abs/1902.08564.