Dual Conditional Cross Entropy Scores and LASER Similarity Scores for the WMT20 Parallel Corpus Filtering Shared Task

Felicia Koerner **Philipp Koehn**
Center for Language and Speech Processing
Johns Hopkins University
{fkr, phi}@jhu.edu

Abstract

This paper describes our submission to the WMT20 Parallel Corpus Filtering and Alignment for Low-Resource Conditions Shared Task. This year's corpora are noisy Khmer-English and Pashto-English, with 58.3 million and 11.6 million words respectively (English token count). Our submission focuses on filtering Pashto-English, building on previously successful methods to produce two sets of scores: LASER_LM, a combination of the LASER similarity scores provided in the shared task and perplexity scores from language models, and DCCEF_DUP, dual conditional cross entropy scores combined with a duplication penalty. We improve slightly on the LASER similarity score and find that the provided clean data can successfully be supplemented with a subsampled set of the noisy data, effectively increasing the training data for the models used for dual conditional cross entropy scoring.

1 Introduction

Machine translation systems require large amounts of high quality parallel corpora for training. Neural machine translation models in particular have been found to both require more data (Koehn and Knowles, 2017), and be more sensitive to noise in training data (Khayrallah and Koehn, 2018) than statistical machine translation models. While these data can be acquired from online sources, the resulting crawled texts are often noisy and require filtering to produce large amounts of sufficiently clean training data.

2 Related Work

We refer readers to (Koehn et al., 2019) for a more detailed overview of methods for parallel corpus filtering, here we describe the most relevant methods to this work.

2.1 Rule-based Filtering

Most filtering methods employ some rule-based filtering, usually to prepare the data for other scoring methods, based on language models, classifiers, or other translation models. (Sánchez-Cartagena et al., 2018) apply hard rules to filter out data before using a classifier to score sentence pairs. (Rossenbach et al., 2018) use many rules, including limits on sentence length, Levenshtein distance, length ratio, and token ratio. We use basic language ID and overlap rules only for the Dual Conditional Cross Entropy Scores, this is described in more detail in subsection 5.1. The LASER similarity scores provided by the shared task organizers also apply a language ID filter (assigning the pair a score of 0 if either of the sentences are not recognized as the expected language).

2.2 Dual Conditional Cross Entropy Scores

The most successful scoring method in the WMT18 Shared Task on Parallel Corpus Filtering was Dual Conditional Cross Entropy Filtering (dccef) (Junczys-Dowmunt, 2018). This method trains an NMT model in both translation directions, uses these to calculate the cross-entropy for each sentence, and finally produces a score based on their agreement. As this year's task deals with low-resource languages (contrary to WMT18, which was En-De), we explore a method to bootstrap the available clean data, thus producing more training data for the intermediate NMT models required for the method (described in more detail subsection 5.2).

2.3 LASER Similarity Scores

LASER similarity scoring was the most successful scoring method of the WMT19 Shared Task on Parallel Corpus Filtering for Low-Resource Languages (Chaudhary et al., 2019). This method embeds parallel sentences with Language Agnostic SEntence

Proceedings of the 5th Conference on Machine Translation (WMT), pages 966–971
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

Representations (LASER) (Artetxe and Schwenk, 2018), and uses these to compute cosine similarity scores. This work attempts to augment LASER similarity scores with language model scores (described in more detail in subsection 4).

3 Shared Task

For this year's shared task on Parallel Corpus Filtering and Alignment for Low-Resource conditions, participants are asked to produce scores for each of the sentence pairs in the provided noisy 58.3 million-word (English token count) Khmer-English corpus and 11.6 million-word Pashto-English corpus. These scores are used to subsample sentence pairs amounting to 5 million English words. The resulting subset is evaluated by the quality of an NMT system (`fairseq` (Ott et al., 2019)) trained on this data.

Participants were given the scripts to either train the evaluation system from scratch, or use the data to fine-tune a provided pretrained MBART model. The MBART model was trained on monolingual data, the details of which are described in (Liu et al., 2020). The performance of the NMT system is measured by BLEU score on a held-out test set of Wikipedia translations. Participants may also provide re-alignments of the source and target sentences. The organizers provide clean parallel and monolingual data for both of the language pairs, as well as LASER similarity scores, a previously successful method in low-resource conditions (Chaudhary et al., 2019), (Koehn et al., 2019).

We participated in the Pashto-English track only, after finding that the model-based methods we explored did not produce meaningful scores for Khmer-English. We did not submit sentence realignments, focusing instead on sentence filtering. Our submission builds on previously successful methods from past WMT shared tasks on parallel corpus filtering to produce two scores: LASER_LM, a combination of the LASER similarity scores and perplexity scores from language models, and DCCEF_DUP, dual conditional cross entropy scores combined with a a duplication penalty. All BLEU scores listed in this paper come from systems trained from scratch and run on the provided development data.

4 LASER_LM

A shortcoming of LASER similarity scores is that they may produce a false positive in the event that the source and target embeddings are similar to each other, but not good translations of each other. Consider, for example, a source and target pair in which the target is simply a copy of the source. This is clearly not a good translation; nothing has been translated. However, the embeddings would be exactly the same, and thus appear to be a very good match. This exact scenario is easily remedied by the use of a language identification filter, but other instances may be more difficult to root out. For example, a source and target sentence in which the target sentence is a string of literal word-for-word translations of the source sentence. To complement the LASER similarity scores and introduce some measure of fluency we train a language model for both English and Pashto.

4.1 LASER Similarity Scores

The LASER similarity scores provided are produced using the methodology outlined in the WMT19 submission (Chaudhary et al., 2019). A language identification filter is applied, and sentences pairs with an overlap between source and target of greater than 60% are discarded. The similarity scores are based on the cosine similarity between the multilingual sentence embeddings in the learned embedding space, and normalized with a margin using the k nearest neighbors approach.

4.2 Language Model Scores

Language models were trained on the provided clean monolingual data. For the English language model was trained on the Wikipedia corpus with 67,796,935 sentences. The Pashto language model was trained on a concatenation of the CommonCrawl and Wikipedia corpora, with the CommonCrawl oversampled by a factor of 64 to produce a dataset of 9,273,763 sentences. The shuffled datasets were split 90/10 (train/test) with test split into 90/10 (dev/test). The language models were trained using `fairseq` (Ott et al., 2019) with the same settings as the WikiText103 example [1].

The language model, M, was used to produce per-sentence perplexity scores for each of the sentences in the corpus. Where $s = w_1, w_2, ..., w_n$ is a sentence of length n:

$$PPL_M(s) = 2^{-\frac{1}{n}\log P(w_1,w_2,...,w_n)} \qquad (1)$$

[1] https://github.com/pytorch/fairseq/blob/master/examples/language_model/README.md

scoring	BLEU (%)
LASER	9.67
LASER + 0.4 PPL_SCORE	9.82
LASER + 0.5 PPL_SCORE	9.81
LASER + 0.6 PPL_SCORE	9.62
LASER + 0.7 PPL_SCORE	9.75
LASER + 0.8 PPL_SCORE	9.88
LASER + 0.9 PPL_SCORE	9.94
LASER + 1.0 PPL_SCORE	9.57

Table 1: Results on development data (training from scratch) for different scaling factors of the PPL_SCORE.

Perplexity scores for both sides (Pashto and English, $H_{ps}(x)$ and $H_{en}(x)$ respectively) are then added together.

$$\begin{aligned}\text{PPL\_SCORE}(x) = {} & PPL_{M_{en}}(s_{en}) \\ & + PPL_{M_{ps}}(s_{ps})\end{aligned} \tag{2}$$

4.3 Combining LASER and LM Scores

The language model scores and LASER similarity scores were combined to produce LASER_LM. Both scores were normalised to fall in the range $[0, 1]$ and the PPL_SCORE subtracted from 1.0, such that lower perplexity corresponded to a higher score. Finally, the two scores were added together to produce the final score in the range $[0, 2]$. We experimented with different scaling factors f for the PPL_SCORE.

$$\begin{aligned}\text{LASER\_LM} = {} & \text{LASER} \\ & + \text{f} \cdot (1.0 - \text{PPL\_SCORE})\end{aligned} \tag{3}$$

Table 1 shows the range of factors f explored to select the scaling factor used in the final score. Since the BLEU scores produced differed only slightly, we also evaluated the models on some of the provided clean data, randomly selecting 2500 lines (roughly the size of the provided devset) from each of the clean corpora, as well as 2500 lines of a shuffled concatenation (concat) of the clean corpora. Results are shown in table 2. For the most part, they did not vary greatly, and where they did there was no consistent winner across corpora. We choose a factor of 0.5, as the model resulting from these scores generally performed well, and, importantly, performed well on the provided devset.

5 DCCEF_DUP

The dual conditional cross entropy scores produced state-of-the-art performance on the WMT18 shared task on filtering corpora for high-resource languages. However, this method requires two translation models trained in both the forward and backward direction. This presents a challenge in low-resource conditions due to the limited training data available. We find that the model quality can be improved by supplementing the provided clean data with a subsampled set consisting of 1M English tokens of the noisy data, subsampled based on the LASER similarity scores.

5.1 Preprocessing

Sentence pairs in which one or both of the sentences did not match the expected language (English or Pashto) as determined by `fastText` [2] were given a score of 0, effectively removing this pair from consideration. This is a harsh filter, removing around 45% of sentence pairs.

The resulting scores were scaled by the overlap between source and target sentence tokens, producing a sort of non-word token matching score. Note that this does not reward pairs that copy large portions of the source sentence to the target, as these are already removed by the language identification filtering.

5.2 Dual Conditional Cross Entropy Scores

Dual Conditional Cross Entropy Filtering (Junczys-Dowmunt, 2018) was found to be state of the art in the WMT18 high-resource data filtering task (Koehn et al., 2018). The method uses two translation models in the forward and backward direction, which are used to compute crosslingual similarity scores. Given the translation model M, sentence pairs (x, y) from the noisy corpus were force-decoded and a cross-entropy score produced:

$$H_M(y|x) = \frac{1}{|y|}\sum_{t=1}^{|y|} \log p_M(y_t|y_{[1,t-1],x}) \tag{4}$$

Cross-entropy scores for both directions (source-to-target and target-to-source, $H_F(y|x)$ and $H_B(x|y)$ respectively) are then averaged with a penalty on a large difference between the scores to produce the overall score:

$$\begin{aligned}\text{DCCEF}(x, y) = {} & \frac{H_F(y|x) + H_B(x|y)}{2} \\ & - |H_F(y|x) - H_B(x|y)|\end{aligned} \tag{5}$$

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

factor	Concat	Bible	GNOME	KDE4	Tatoeba	Ubuntu	Wikimedia	TED Talks
0.4	5.84	1.79	13.08	6.98	5.21	10.73	4.65	5.76
0.5	6.28	1.07	14.09	7.72	10.31	11.02	4.83	5.17
0.6	6.76	1.62	13.89	7.41	11.39	10.36	4.34	5.43
0.7	6.43	1.18	14.02	7.98	11.02	11.42	4.28	5.21
0.8	6.20	1.00	13.71	7.87	6.66	10.92	5.11	5.71
0.9	6.25	1.80	12.99	7.32	5.80	10.32	6.02	6.15
1.0	6.67	1.63	13.77	7.44	9.53	10.75	4.54	5.69

Table 2: Results (BLEU(%)) on subsamples of clean data (training from scratch) for different scaling factors of the PPL_SCORE.

Translation models were trained using `fairseq` (Ott et al., 2019) with the same parameters used in the baseline flores model [3].

We used the provided clean training data to train translation models in both directions, and used these models to produce a dccef score as described above. Initially **only** the dccef scores were used to filter the noisy data and train a system, we did not perform the preprocessing as described in 5.1. The BLEU score produced by this system is shown in 3 under clean.

We then supplemented the clean training data with a subsample of the noisy data and trained translation models in both directions on the augmented data. The subsample of 1 million English tokens and their translations was selected based on the provided LASER similarity score. Again, for this experiment only the dccef scores were used to filter the noisy data, no preprocessing was performed. As shown in Table 3, supplementing the training data with the subsampled set resulted in an overall increase in 3.37 BLEU points.

Finally, we preprocessed the noisy data as described in 5.1 and used both sets of systems (one set trained on clean data, and one set trained on augmented data) to score the preprocessed data. As shown in Table 3, there were further, significant gains, from preprocessing, and the dccef scores from the systems trained on augmented data outperformed the dccef scores from the systems trained on just the clean data. Prepocessing also reduced the gap between the performance of dccef scores produced by systems trained on just the clean data and the performance of dccef scores produced by systems trained on augmented data.

5.3 Duplication Penalty

The scores were scaled by a duplication penalty for duplicate (greater than one) occurrences of either one or both of the target or source sentence of a pair in the corpus as follows:

$$\text{dup\_penalty} = \begin{cases} 1.0 & \text{neither side duplicate} \\ 0.9 & \text{one side duplicate} \\ 0.8 & \text{both sides duplicate} \end{cases} \tag{6}$$

This resulted in a minor improvement in BLEU score on the development data, as seen in Table 3.

6 Results

Various other combinations of the aforementioned scores were explored, and the results are listed in Table 4. Interestingly, the results suggest that the duplication penalty did not improve the LASER_LM score, and combining the LASER_LM and DCCEF_DUP scores did not result in a better BLEU score. However, it should be noted that the differences in BLEU scores resulting from different combinations are generally minor and may not be statistically significant.

None of the filtering methods significantly outperformed the LASER-based method, but the improved dccef filtering method can at least match the LASER-based method when the training data is augmented, and the preprocessing steps and duplication penalty are applied.

7 Conclusion

This paper describes the our submission to the WMT20 Parallel Corpus Filtering Shared Task for low-resource conditions. We find that filtering based on dccef scores can compete with filtering based on LASER similarity scores when the models trained for the dccef scores are augmented with a subsample of the noisy data. This suggests that

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

training data for H_{en},H_{ps}	scoring method	BLEU (%)
clean	dccef	3.97
clean + top 1M noisy	dccef	7.34
clean	dccef + preprocessing	8.93
clean + top 1M noisy	dccef + preprocessing	9.68
clean + top 1M noisy	(dccef + preprocessing) · dup_penalty	**9.94**

Table 3: Results on development data (training from scratch) for dccef scores.

training data for H_{en},H_{ps} (dccef)	scoring method	BLEU (%)
N/A	laser	9.67
N/A	laser + 0.5LM	**9.81**
N/A	(laser + 0.5LM) · dup_penalty	9.74
clean	dccef	8.93
clean + top 1M noisy	dccef	9.68
clean + top 1M noisy	(dccef · dup_penalty)	**9.94**
clean + top 1M noisy	(dccef · dup_penalty) + laser	9.30
clean + top 1M noisy	(dccef · dup_penalty) + laser + 0.5LM	9.58

Table 4: Results on development data (training from scratch). Bolded scores are the two scores submitted. All dccef scores reported in this table were combined with preprocessing as described in 5.1

challenges posed by limited data for model-based filtering methods can be somewhat mitigated by bootstrapping additional data from the noisy corpus.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.

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.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, 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.

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.

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, Belgium, Brussels. Association for Computational Linguistics.

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

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.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney. 2018. The RWTH aachen university filtering system for the WMT 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 946–954, 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.