Alibaba Submission to the WMT20 Parallel Corpus Filtering Task

Jun Lu, Xin Ge, Yangbin Shi, Yuqi Zhang
Machine Intelligence Technology Lab, Alibaba Group
Hangzhou, China
{joelu.luj, shiyi.gx, taiwu.syb, chenwei.zyq}@alibaba-inc.com

Abstract

This paper describes the Alibaba Machine Translation Group submissions to the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In the filtering task, three main methods are applied to evaluate the quality of the parallel corpus, i.e. a) Dual Bilingual GPT-2 model, b) Dual Conditional Cross-Entropy Model and c) IBM word alignment model. The scores of these models are combined by using a positive-unlabeled (PU) learning model and a brute-force search to obtain additional gains. Besides, a few simple but efficient rules are adopted to evaluate the quality and the diversity of the corpus. In the alignment-filtering task, the extraction pipeline of bilingual sentence pairs includes the following steps: bilingual lexicon mining, language identification, sentence segmentation and sentence alignment. The final result shows that, in both filtering and alignment tasks, our system significantly outperforms the LASER-based system.

1 Introduction

The parallel corpus is an essential resource for building a high quality machine translation(MT) system. It has been shown that, the higher the corpus quality, the better the performance of a MT system(Koehn and Knowles, 2017; Khayrallah and Koehn, 2018). Many successful machine translation systems are built on the corpus crawled from the web. In practice, this kind of parallel corpus may be very noisy. The task of Parallel Corpus Filtering is aimed at tackling the problem of cleaning noisy parallel corpora.

We form the bilingual sentences quality in the following aspects. Firstly, a *high-quality* parallel sentence pair(also called bitext) should have the property that its target sentence precisely translates the source sentence, and vice versa. In this task, we attempt to quantify the translation accuracy (also called bilingual score) of bilingual sentence pairs. Secondly, the monolingual quality of the target and source sentences of a parallel corpus should also be considered. In our system, we evaluate the monolingual quality (also called monolingual score) of a target sentence due to its importance for the MT procedure. Finally, the bilingual and monolingual scores are combined to evaluate bilingual sentence pairs and filter out the ones with low quality.

The paper is structured as follows. Section 2 describes our methods which are used in the parallel corpus filtering. In Section 3, we briefly outline the pipeline of parallel sentence extraction. Section 4 specifies the experiments and results as well as the dataset for building model-based methods. Conclusions are drawn in Section 5.

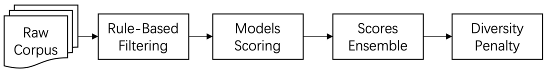

Figure 1: Framework of parallel corpus filtering

2 Parallel Corpus Filtering Methods

Figure 1 shows the framework of parallel corpus filtering. The raw parallel corpus is firstly filtered by heuristic rules so that the very noisy sentence pairs will be removed. Then, the bilingual & monolingual models are built to score all the remaining sentence pairs. By using an ensemble model, the partial scores of each sentence pair are combined to a single quality score.

2.1 Rule-based Filtering

A series of heuristic rules(Lu et al., 2018) are applied to filter low quality sentence pairs. They are simple, (almost) language independent but efficient, which are described below.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 979–984
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

Monolingual Rules

- The length of the sentence which is too short (≤ 2 tokens) or too long (> 200 tokens) will be dropped. In our system, sentences(English, Khmer and Pashto) are tokenized by SentencePiece[1].

- The ratio of the valid tokens count to the length of the sentence. Here, valid tokens are the ones which contain the letters in the corresponding language. For example, a valid token in English should contain English letters. In our system, the sentence is filtered out if its valid-tokens ratio is less than 0.2.

- Language filtering. For the Pashto-English parallel corpus, the languages of source and target sentences should be Pashto and English. We detect the language of a sentence by using a language detection tool we developed[2]. A sentence pair is dropped when its source language and target language are not Pashto and English, respectively.

Bilingual Rules

- The length ratio of a source sentence to a target sentence. The sentence length is calculated by the number of sentencepiece tokens. In our system, the ratio is set between 0.2 and 5.0 for both language pairs.

- The edit distance between the source token sequence and the target token sequence. A small edit distance indicates that the source and target sentences are very similar, which harms the performance of the NMT system a lot (Khayrallah and Koehn, 2018).

- The consistency of special tokens (Taghipour et al., 2010). For example, the high-quality sentence pairs should contain the same email address in both source and target sentences (if exists). In this task, special tokens are email addresses, URLs, and big Arabic numbers.

2.2 Dual Bilingual GPT-2 Model

Inspired by the *Cross-lingual Language Model Pre-training* work of (Lample and Conneau, 2019), we propose a Translation Language Model(called Bilingual GPT-2 model) based on the GPT-2 model(Radford et al., 2019). As illustrated in Figure 2, the Bilingual GPT-2 model is trained with both monolingual and parallel sentences. For parallel sentence pairs, we concatenate the source and target sides to obtain a long sentence and then feed it to the model. For monolingual sentences, we convert them to *fake sentences pairs* by assigning the corresponding side sentence with a unique token. For example, when an English sentence ”Hello word.” is used in the English-Khmer bilingual GPT-2 model training, a fake sentence pair, (“Hello word”, “<KM>”), will be used. Here, the English sentence is the source and “<KM>” is the target. While training, a large number of fake bilingual corpora are firstly used to *pre-train* the model. Then, the real clean parallel sentence pairs are used to *fine-tune* the model. In this task, we trained two Bilingual GPT-2 models for each language pair, i.e., source-to-target and target-to-source models. The two translation quality scores from the Dual Bilingual GPT-2 model are given precisely by:

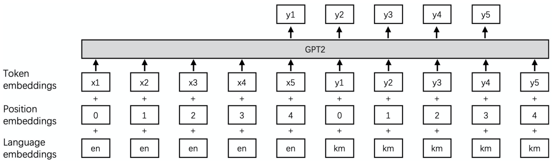

Figure 2: Bilingual GPT-2 model structure

$$
\begin{aligned}
score_1(x, y) = & \frac{1}{2}(\sum_{t \in |y|} \log p_{s2t}(y_t) \\
& + \sum_{t \in |x|} \log p_{t2s}(x_t))
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
score_2(x, y) = & \frac{1}{2}(\sum_{t \in |y|} \log p_{s2t}(y_t) - \log p_{t2s}(y_t) \\
& + \sum_{t \in |x|} \log p_{t2s}(x_t) - \log p_{s2t}(x_t))
\end{aligned} \tag{2}
$$

In Equation (1) and (2), x and y are the source and target sentences. $logp_{s2t}(y_t)$ represents the cross-entropy loss of the target side token y_t, which is obtained by the source-to-target model. $\log p_{t2s}(x_t)$ represents the cross-entropy loss of the source side token x_t, which is obtained from the target-to-source model.

We don't use the BERT model here, as it is hard for computing the cross-entropy loss efficiently.

[1] https://github.com/google/sentencepiece

[2] This tool is similar to Google's CLD2: https://github.com/CLD2Owners/cld2

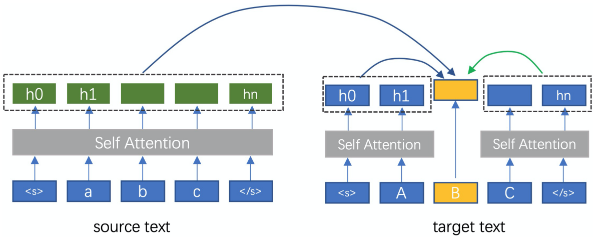

Figure 3: Optimized cross entropy model (source-to-target)

2.3 Dual Conditional Cross-Entropy Model

The dual conditional cross-entropy model(Junczys-Dowmunt, 2018) has been proven effective in parallel corpus filtering, which uses a combination of forward and backward models to compute a force-decoding score. In our system, the model is optimized to better evaluate the quality of the parallel sentences in low-resource languages.

Specifically, Figure 3 shows the structure of our model. Each token in a target side sentence is predicted by its left and right context and the source text. Hence, the cross-entropy score of a sentence pair is given below:

$$H_M(y|x) = \frac{1}{|y|} \sum_{t \in |y|} \log p_M(y_t|x, y_{<t}, y_{>t})$$

The final bilingual quality score combines the source to target and target to source cross-entropy scores as below:

$$\begin{aligned} score(x, y) = & \frac{1}{2}(H_{Fwd}(y|x) + H_{Bck}(x|y)) \\ & + |H_{Fwd}(y|x) + H_{Bck}(x|y)| \end{aligned}$$

As shown in Figure 3, the source sentence and target context are encoded by two 12-layer transformer models with hidden size 768. In fact, the target side model can be regarded as a bidirectional GPT-2 model. In our system, the source and target side transformer models are pre-trained by using large amount of monolingual data. Then, the models are fine-tuned by clean bilingual sentences pairs.

2.4 IBM Word Alignment Model

The word alignment model can be used for evaluating the translation quality of bilingual sentence pairs (Khadivi and Ney, 2005; Taghipour et al., 2010; Ambati, 2011). Inspired by the work of (Khadivi and Ney, 2005), we simplify the original algorithm, and the translation score of sentence pairs is given below:

$$\begin{aligned} score(s, t) = & \frac{1}{|s|} \sum_{s_i, t_j \in a_{s2t}} \log p(t_j|s_i) \\ & + \frac{1}{|t|} \sum_{s_i, t_j \in a_{t2s}} \log p(s_i|t_j) \end{aligned} \quad (3)$$

In Equation (3), s and t represent the source and target sentences respectively, $p(w_1|w_2)$ indicates the word translation probability, and a_{s2t} indicates the source words to target words alignment.

In this task, by using the *fast_align* toolkit (Dyer et al., 2013), the word alignment model is trained on a clean parallel corpus as described in Section 4.1 to get the forward and reverse word translation probability tables. This model is also called alignment scoring model.

2.5 GPT-2 Language Model

In this task, GPT-2 language model is applied to compute the monolingual scores of source and target sentences. We train GPT-2 models for each language by using the *HuggingFace Transformers* toolkit (Wolf et al., 2019) with the monolingual data provided by the task organizers. The training data is cleaned by the rules described in the Section 2.1. The configuration of the GPT-2 model is also the same with the *GPT2-large model* described in the work of (Radford et al., 2019).

2.6 Ensemble

Each sentence pair in the noisy parallel corpus is scored by each of the models described above. As a result, each sentence pair would obtain a few partial scores. We need a single score based on the partial scores to rank the sentence pairs.

At first, we turn the scores from each model to the values between 0 and 1. Specifically, the scores are normalized with the method described in (Junczys-Dowmunt, 2018), which is based on the entropy information.

Then, a single score $f(x, y)$ is produced as the product of partial scores $f_i(x, y)$. Since the different importance of the partial scores, the lower boundary value of the scores is represented as θ, where $0 \leq \theta \leq 1$, which results in a new normalization range $[\theta, 1]$. The more important the model is, the closer to 0 the θ is. It means that the scores from this model could distribute from 0 to 1, which would affect a lot on the final score. On the contrary, when we set θ close to 1, the model has minor impact on the final score whatever its distribution is. Hence, the single score is given by:

$$f(x, y) = \prod_i f_i(x, y), f_i(x, y) \in [\theta_i, 1] \quad (4)$$

We applied the brute-force search to find the best θs for the models. Compared to the pure production of the partial scores, our method has improved 0.5% - 1.0% BLEU score(Papineni et al., 2002).

In addition, the ensemble could also be treated as a Positive-Unlabeled classification task (Chaudhary et al., 2019). We use the officially released high quality data and the sentence pairs which are ranked top by our models mentioned above as the positive samples. Meanwhile, the sentence pairs from the noisy parallel corpus are treated as the unlabeled samples. As a result, the PU classification based on the random forest models has contributed 0.1% - 0.2% improvement on the development data.

In our final submissions, the brute-force search method and PU-classification are used in Khmer-English and Pashto-English filtering tasks respectively.

3 Pipeline of Parallel Corpus Extraction

Bilingual Lexicon Extraction. In the first step, by using the word alignment model, parallel token pairs are extracted from the clean parallel corpus. Specifically, after tokenization, the parallel corpus is fed to the fast align toolkit to obtain the mutual translation probabilities dictionary. We then extract the token pairs with forward and backward translation probabilities higher than 0. The bilingual lexicon (i.e., the collection of parallel token pairs) will iteratively be updated after more bitexts are mined, since the lexicon is the cornerstone of bitexts mined from aligned documents which is described below.

Language Identification. The second step is to identify the language of each document by using a language detection tool we developed. In this way, a document pair will be discarded if its detection results do not match the expected languages.

Sentence Segmentation. This step is to split sentences in documents with rules or models. A few rules based on end-of-sentence punctuations are used to split sentences of language Pashto and Khmer. For English sentence segmentation, a segmentation model is built via nltk toolkit[3].

Language	Sentences	English Words
Khmer-English	270K	4.2M
Pashto-English	106k	1.9M

Table 1: Clean bitexts used in bilingual models training

Sentence Alignment. In this step, a dynamic programming framework based on bilingual lexicon (Ma, 2006) is built to mine parallel sentence pairs.

Corpus Filtering. Finally, the extracted bitexts are cleaned by using the methods described in section 2. And as mentioned above, we mix the new mined bitexts with the provided bitexts from WMT2020 to iteratively run the fast_align model to update bilingual lexicon.

4 Experiments and Results

In this section, we specify the experimental settings and results in the corpus filtering and alignment task.

4.1 Corpora and Settings

The selection data pool[4] (which we called noisy dataset) is provided by *WMT20 Corpus Filtering and Alignment Task*. It contains 1.02 million sentences pairs of Pashto-English corpus and 4.17 million sentences pairs of Khmer-English corpus. These parallel corpora are very noisy. The task's participants are asked to sub-select sentence pairs that amount to 5 million English words for each of the noisy parallel sets. The quality of the resulting subsets is determined by the BLEU scores of a neural machine translation system[5] trained on selected data. In our NMT experiments, we use the NMT configuration that is provided by the task organizers[6] as well as the development and test sets.

In addition, organisers provide the permissible third-party sources of parallel corpora, which we called "official parallel data". Additional monolingual corpora are also provided for English, Khmer and Pashto languages. For sentence pair alignment task, the organisers also provide the document pairs

[3] Natural Language Toolkit: https://github.com/nltk/nltk

[4] http://www.statmt.org/wmt20/parallel-corpus-filtering.html

[5] https://github.com/pytorch/fairseq.git

[6] http://data.statmt.org/wmt20/filtering-task/dev-tools.tgz

Method	km-en			ps-en		
	pairs counts ($\times 10^4$)	normal train	finetune	pairs counts ($\times 10^4$)	normal train	finetune
LASER(Baseline)	24.1	7.35	10.4	22.5	9.66	10.76
LASER +Rules	24.7	7.56	10.89	22.9	9.88	11.13
IBM word align +Rules	25.7	8.25	11.04	35.6	10.37	12.49
Dual X-Ent +Rules	34.6	8.12	10.71	43.4	9.9	12.14
Dual bi-GPT-2_1 +Rules	33.3	8.3	10.86	30.2	9.95	11.62
Dual bi-GPT-2_2 +Rules	38.1	8.5	10.95	34.9	10.04	12.17
Ensemble + Rules	25.8	8.61	11.34	37.5	10.84	12.75
Alignment + Ensemble + Rules	-	-	-	21.2	11.36	13.29

Table 2: Main results for corpus filtering and alignment task

in which the participants can extract bilingual sentence pairs.

In Section 2, we introduced 3 sub-models for translation quality scoring, i.e. Dual Bilingual GPT-2 Model, Dual Conditional Cross-Entropy and IBM Word Alignment Model. These models can be trained with the monolingual and clean parallel corpus. In particular, the clean parallel data is more important in training. Unfortunately, both Khmer-English and Pashto-English are low-resource language pairs and lack parallel corpus. Therefore, in order to expand the clean parallel dataset, the high quality sentence pairs are selected/extracted from the noisy dataset or the parallel document pairs by using an iterative process in our filtering system. Specifically, the corpus filtering models are initially trained by using the official parallel data. Then, these models are used to estimate the quality of the sentence pairs in the noisy dataset and parallel documents. Finally, by applying some rules and strict threshold value, the high quality sentence pairs are selected and combined with official parallel data to train the new version of filtering models. The process described above was repeated 3 times and achieved larger clean parallel corpora as detailed in Table 1.

For text preprocessing, we built two joint SentencePiece models for Khmer-English and Pashto-English respectively with the 60k vocabulary size. Then, monolingual and bilingual texts are tokenized by the corresponding SentencePiece models.

4.2 Experimental Results

Our main results are shown in Table 2. All NMT experiments were done in the same environment with 2 GPUs for normal training (i.e., NMT training from scratch) and 1 GPU for MBART-based fine-tuning[7]. The LASER scores provided by the organisers were used as baseline scores, which achieved reasonable results in both normal training and MBART-based fine-tuning. Our rules proposed above were firstly used to filter very noisy sentence pairs and achieved a slightly better performance. Then, the 3 main bitexts scoring models were combined with rules respectively to test their effectiveness in experiments. We found that, the IBM word alignment model was reliable in most cases and Dual Bilingual GPT-2 model slightly outperformed the Dual Conditional Cross-Entropy model. Finally, the ensemble model obtained the highest BLEU scores in the filtering task.

In the task of sentence pairs alignment, we only submitted the results of Pashto-English. While extracting sentence pairs, 13,976 bilingual word pairs were firstly obtained from the clean parallel corpus. As a result, we mined 723,414 sentence pairs from 45,307 document pairs and achieved an improvement of 0.5 BLEU score.

5 Conclusions

In this paper, we present our corpus filtering system for the *WMT 2020 Corpus Filtering Task*. In our system, Dual Bilingual GPT-2 model, Dual Conditional Cross-Entropy model and IBM word alignment model are combined to filter the noisy parallel corpus. Besides, a parallel sentence pairs extraction system is built to re-align the bilingual sentences. The experiments show that, compared

[7] The MBART pre-trained models were provided by the organizers and described here, http://www.statmt.org/wmt20/parallel-corpus-filtering.html.

to the baseline system, our filtering and extraction system achieve much better results.

Acknowledgments

This work is supported by National Key RD Program of China (2018YFB1403202).

References

Vamshi Ambati. 2011. *Active learning and crowdsourcing for machine translation in low resource scenarios*. Ph.D. thesis, University of Southern California.

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, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 261–266. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. 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, Volume 2: Shared Task Papers*, pages 901–908, Belgium, Brussels. Association for Computational Linguistics.

Shahram Khadivi and Hermann Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *International Conference on Application of Natural Language to Information Systems*, pages 263–274. Springer.

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 1–10.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jun Lu, Xiaoyu Lv, Yangbin Shi, and Boxing Chen. 2018. Alibaba submission to the wmt18 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 930–935, Belgium, Brussels. Association for Computational Linguistics.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Kaveh Taghipour, Nasim Afhami, Shahram Khadivi, and Saeed Shiry. 2010. A discriminative approach to filter out noisy sentence pairs from bilingual corpora. In *Telecommunications (IST), 2010 5th International Symposium on*, pages 537–541. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.