Alibaba Submission to the WMT18 Parallel Corpus Filtering Task

Jun Lu, Xiaoyu Lv, Yangbin Shi, Boxing Chen
Machine Intelligence Technology Lab, Alibaba Group
Hangzhou, China
{joelu.luj, anzhi.lxy, taiwu.syb, boxing.cbx}@alibaba-inc.com

Abstract

This paper describes the Alibaba Machine Translation Group submissions to the WMT 2018 Shared Task on Parallel Corpus Filtering. While evaluating the quality of the parallel corpus, the three characteristics of the corpus are investigated, i.e. 1) the bilingual/translation quality, 2) the monolingual quality and 3) the corpus diversity. Both rule-based and model-based methods are adapted to score the parallel sentence pairs. The final parallel corpus filtering system is reliable, easy to build and adapt to other language pairs.

1 Introduction

The parallel corpus is an essential resource for machine translation and multilingual natural language processing. Apart from the quantity and domain, the quality of parallel corpus is also very important in MT system training (Koehn and Knowles, 2017; Khayrallah and Koehn, 2018). The Internet contains a large number of multilingual resources, including parallel and comparable sentences (Resnik and Smith, 2003). Many successful machine translation systems are built using the corpus crawled from the web. But in practice, this kind of parallel corpus may be very noisy. The task of Parallel Corpus Filtering tackles the problem of cleaning noisy parallel corpus.

In this task, we can divide the corpus cleaning task into three parts. Firstly, a *high-quality* parallel sentence pair 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 quality (also called bilingual score) and accuracy of the sentence pair. Secondly, the quality of the target and/or source sentences of the parallel corpus should also be evaluated. In this work, the target side sentences are concerned a lot for their importance in NMT. Thirdly, as described by the *Parallel Corpus Filtering* task, the participants should not pay attention to the domain-relatedness. We need to focus on all the domains so that the resulting MT system can be widely used. So the diversity should be evaluated while subsampling the parallel corpus. Finally, the three characteristics of the parallel corpus are combined to build the final clean corpus.

The paper is structured as follows: Section 2 describes our methods which are used in parallel corpus filtering. Section 3 specifies the experiments and results. The dataset for building model-based methods is also detailed in this section. Conclusions are drawn in Section 4.

2 Parallel Sentence Pairs Scoring Methods

In this section, three kinds of scoring/filtering methods are detailed.

2.1 Bilingual Quality Evaluation

Here, we describe the noisy corpus filtering rules and two kinds of translation quality evaluation methods: (1) Word Alignment Based bilingual scoring and (2) Bitoken CNN Classifier based bilingual scoring(Chen et al., 2016).

Rule-based Filtering

A series of heuristic rules are applied to filter *bad* sentence pairs. They are simple but efficient, which are described below.

- The length ratio of source sentence to target sentence. Sentence length is calculated as the number of tokens/words. In our system, the ratio is set between 0.4 and 2.5.
- The edit distance between the source token sequence and the target token sequence. A

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 917–922

https://doi.org/10.18653/v1/W18-64109

small edit distance indicates that the source and target sentences are very similar. This kind of corpus harms the performance of the NMT system a lot (Khayrallah and Koehn, 2018). Besides, the edit distance can be normalized by the average length of source and target sentence length, which represents the *edit distance ratio*. Both edit distance and edit distance ratio are used to filter sentence pairs in which the source and target sentence are similar. In our system, a sentences pair will be dropped if its edit distance is less than 2 or edit distance ratio is less than 0.1.

- 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 an email address, URL, and a big Arabic number.

Word Alignment-based Bilingual Scoring

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:

$$score(s,t) = \frac{1}{m} \sum_{s_i,t_j \in a_{s2t}} \log p(t_j|s_i) + \frac{1}{n} \sum_{s_i,t_j \in a_{t2s}} \log p(s_i|t_j) \quad (1)$$

In Equation (1), s and t represent the source and target sentences respectively, $p(w_1|w_2)$ indicates the word translation probability, a_{s2t} indicates the source words to target words alignment, m and n are the lengths of source and target sentences.

In this task, the word alignment model is trained on a clean parallel corpus provided by *WMT18 New Translation Task*. We use the fast_align toolkit (Dyer et al., 2013) to train the model, and get the forward and reverse word translation probability tables.

This model is also called alignment scoring model.

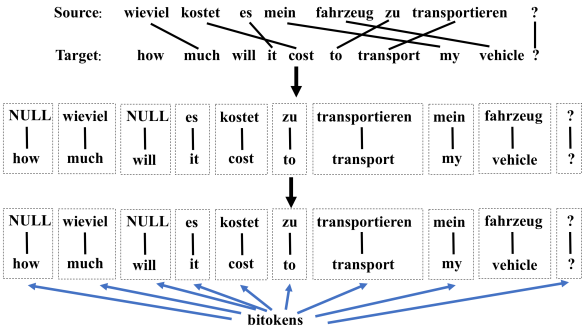

Figure 1: Bitoken sequence

Bitoken CNN Classifier-based Bilingual Scoring

Following the work of (Chen et al., 2016), a bitoken CNN based scoring model is built for translation quality evaluation.

In this model, the *bitokens* are extracted from aligned sentence pairs. Figure 1 shows how a bitoken sequence can be obtained from a word-aligned sentence pair. Each bitoken in the sequence is treated as a word, and each bitoken sequence is treated as a normal sentence. Then these bitoken sentences are fed to the CNN Classifier to build the bilingual scoring model. For every candidate sentence pair, this model will give two probabilities: p_{pos} and p_{neg}, and the quality score is treated as $score_{bitoken} = p_{pos} - p_{neg}$. For the train data set, the bitoken sequences obtained from the high-quality corpus are labeled as positive. As for the negative train data, we manually construct some noisy data based on the clean data.(Lample et al., 2017) For example, shuffle the target side sentences of the clean parallel corpus, or randomly delete the source or target sentence's words. So the negative bitoken sequences could be obtained from this unparallel corpus.

This scoring model can also be called bitoken_CNN scoring model.

2.2 Monolingual Quality Evaluation

Rule based Filtering

A few rules are applied to filtering the sentence pairs whose source or target side are *not good*. These rules are:

- The length of the sentence which is too short ($\leq$ 2 words) or too long ($>$ 80 words) will be dropped.

- The ratio of valid tokens counts to the length of the sentence. Here, valid tokens are the tokens 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 if its valid-tokens ratio is less than 0.2.

- Language filtering. For German-English parallel corpus, the source and target sentences' languages should be English and German. We can detect the sentence's language by using a language detection tool we developed[1]. The sentences pair is filtered if the languages of its source and target sides are not German and English.

Language Model Scoring

We use the language model to evaluate the quality of sentences. The language model is successfully used to select domain-related corpus (Yasuda et al., 2008; Moore and Lewis, 2010). Besides, the language model can also be used to filter out ungrammatical data (Denkowski et al., 2012; Allauzen et al., 2011), which is suitable for this task.

In our corpus filtering system, we focus on the quality of target sentences, i.e. English sentences, as they are more important in NMT. Firstly, a large language model is built on all available English monolingual corpus provided by WMT18. The training corpus is cleaned using some rules mentioned above. Then the normalized-length language model score can be regarded as the monolingual quality score. But in practice, this method has a shortcoming: it gives lower scores for the good sentences that contain rare words. The training corpus needs to be generalized to overcome this shortness, for example, we can replace the words that occur less than 10 times in LM train corpus with their part of speech tag(Axelrod et al., 2015). Finally, the language model is re-built on the generalized corpus.

2.3 Corpus Diversity

Rule-based Filtering

We could use a simple rule to reduce the number of similar sentence pairs. Firstly, source and target sentences should be generalized. In our experiment, for the English sentence, the generalization is done by removing all the characters except for English letters. Also, a similar operation is done for generalizing German sentences. After that, if some sentence pairs have the same generalized source or target sentences, the sentence pair that has the highest quality score will be selected.

N-gram based Diversity Scoring

In this method, we aim to sub-select a corpus which contains a variety of N-grams. Such a corpus is regarded as high diversity. We follow the work of (Ambati, 2011; Biçici and Yuret, 2011), with the motivation for introducing a feature decay function for the n-gram weight. In our system, after selecting a subset S_1^{j-1}, the next sentence s_j's diversity score is given by:

$$f(s_j|S_1^{j-1}) = \frac{\sum_{n=1}^{N}\sum_{ng\in NG(s_j,n)} weight(ng, j-1)}{norm(s_j)} \quad (2)$$

$$weight(ng, j-1) = Freq(ng, S) * e^{-\lambda * Freq(ng, S_1^{j-1})},$$

where S_1^{j-1} represents the set of selected sentences which contains 1^{st} to $(j-1)^{th}$ sentences, and S is the whole sentences pool to be selected.

$f(s_j|S_1^{j-1})$ is the diversity score of sentence s_j under the condition that corpus S_1^{j-1} is selected.

$NG(s_j, n)$ is all n-grams of size n in sentence s_j. $|NG(s_j, n)|$ is the size of the $NG(s_j, n)$.

$norm(s_j)$ is the normalization factor for sentence s_j, and equals $\sum_{n=1}^{N} |NG(s_j, n)|$.

$Freq(ng, S)$ is the frequency of n-gram in selection data S.

λ is the exponential decay hyper parameter, $\lambda = 1$ in our experiment.

The equation (2) indicates that the n-gram is weighted by its frequency in the pool set S and selected set S_1^{j-1}. The higher the frequency of n-grams in the selected set, the lower the weight; the higher the frequency of n-gram in the pool set, the higher the weight. In practice, firstly, the sentences pairs in the pool S are sorted by their quality scores(combined by bilingual and monolingual score) in descending order. Then the selection method described above is carried out on the target side of the bilingual corpus.

Parallel Phrases Diversity Scoring

The N-gram based Diversity Scoring is commonly used for selecting monolingual sentences with high diversity. Here we aim to sub-select a bilingual corpus which contains a variety of parallel

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

phrases. With this kind of corpus, the MT model will learn more translation knowledge.

Firstly, we use the fast_align toolkit to train a word alignment model. And then the phrase table of the corpus can be extracted by using the Moses toolkit. Next, we can obtain the parallel phrases pairs for each sentence pair from the phrase table using the methods of maximum matching. Finally, following the method described in section *N-gram based Diversity Scoring*, the same selection procedure (in which, N-gram is replaced by phrase pairs) is used for sentence pairs' scoring. In our system, it works best when the phrase length is less than 7.

2.4 Methods Combination and corpus sampling

In our corpus filtering system, all the methods are combined into a pipeline.

First of all, we apply all the bilingual and monolingual rules to filter very noisy sentence pairs. Then, two bilingual scores and target side language model score could be produced by the above corresponding models. These three scores are individually normalized and then linearly combined to produce a single quality score. Here, the weights of these scores are selected with gird search method(Hsu et al., 2003). After that, we sort the sentence pairs by their corresponding quality scores in the descending order. The diversity method is then used to re-score/re-order the corpus. Finally, we select two sets of the top-N sentence pairs that contain totally 10 million words and 100 million words.

3 Experiments and Results

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

3.1 Corpora and Settings

The selection data pool[2] is provided by *WMT18 Corpus Filtering Task*, which contains about 100 million sentences pairs. It is very noisy. The task's participants are asked to sub-select sentence pairs that amount to (a) 100 million words and (b) 10 million words.[3] The quality of the resulting subsets is determined by the BLEU scores of a statistical machine translation (Moses, phrase-based) and neural machine translation system (Marian) trained on this data. In our SMT and NMT experiments, we used the SMT and NMT configuration that are provided by the task organizer[4], as well as the development and test set.

While building the alignment scoring model, after using the bilingual and monolingual filtering rules, 4,337,154 sentence pairs are selected from the corpora provided by the *WMT18 English-German news translation task*. Next, the fast_align tool is used to build the word alignment model on the clean corpus, and then we can obtain the forward and reverse word translation probability tables.

When building the bitoken_CNN scoring model, 20,000 positive labeled bitoken sequences and 20,000 negative labeled bitoken sequences are constructed. The fast_align toolkit is also used here. Then, we use the *CONTEXT*[5] toolkit to train the CNN models. The bitokens' embedding vectors are trained by *word2vec*[6], and the size of each vector was set to 200.

For target sentences' quality evaluation, we use the KenLM(Heafield et al., 2013) toolkit to train the normal and generalized LM. The clean training corpus contains 60 million English sentences, which are sub-selected from the corpora provided by *WMT18 News Translation Task*.

3.2 Experimental Results

Firstly, the whole corpus which contains about 100 million sentence pairs was evaluated by training the SMT and NMT system. The final BLEU scores are 21.21 and 7.8 respectively. This experiment shows that the whole corpus is really noisy.

Other experimental results are detailed in Table 1. The randomly sub-selected corpus' performance is also very poor. The *sys_1* system uses the bilingual/monolingual rules and alignment scoring, which performed much better. We replace the alignment scoring method by bitoken_CNN method and then build the *sys_2* system. We find that the alignment scoring method and bitoken_CNN method are very similar in sentences pairs scoring. As a result, a lot of sentence pairs (about 70% in the subset) are selected by both methods. The two methods are combined in *sys_3*, which has a little improvement. While combining,

[2]http://www.statmt.org/wmt18/parallel-corpus-filtering-data/data.gz

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

[4]http://www.statmt.org/wmt18/parallel-corpus-filtering-data/dev-tools.tgz

[5]http://riejohnson.com/cnn_download.html

[6]https://code.google.com/archive/p/word2vec/

System ID	Method	10M words subset			100M words subset		
		sentence pairs count ($\times 10^6$)	SMT	NMT	sentence pairs count ($\times 10^6$)	SMT	NMT
-	Random subset	1.31	15.25	7.73	8.23	18.21	7.57
sys_1	bilingual & monolingual rules + Alignment scoring	1.29	20.57	23.23	7.56	25.15	30.02
sys_2	bilingual & monolingual rules + bitoken_CNN scoring	1.09	21.02	23.69	6.45	25.19	30.33
sys_3	bilingual & monolingual rules + Alignment + bitoken_CNN	0.46	21.93	24.14	5.05	25.13	30.43
sys_4	sys_3 + Language Model	0.76	23.53	25.01	5.41	25.77	31.44
sys_5	sys_4 + Diversity Evaluation	0.64	23.79	25.34	5.41	25.77	31.44

Table 1: Methods used in Corpus selection and their performance

the original scores are normalized to the interval $[0, 1]$, and then the linear model is used to produce a new score. In *sys_3* system, the weights of alignment score and bitoken_CNN score are 0.4 and 0.6 respectively.

The *sys_4* introduced language mode score based on *sys_3*. The weights of the alignment score, bitoken_CNN score, and the language model score are 0.4, 0.6 and 0.8 respectively. It shows that the language model is useful in selecting clean sentences pairs.

Finally, based on *sys_4*, the corpus diversity filtering rules and scoring are introduced in *sys_5*. We find that the diversity method (only *Parallel Phrases Diversity Scoring* is used in *sys_5* system) works well in selecting the smaller subset corpus, e.g. the 10 million words corpus. For large subset corpus selection, it almost has no improvement. We attribute this to the sufficiently high diversity of larger subset corpus.

4 Conclusions

In this paper, we present our corpus filtering system for the *WMT 2018 Corpus Filtering Task*. In our system, sentence pairs are evaluated in three aspects: (1) the bilingual translation quality, (2) the monolingual quality of the source and target sentences and (3) the diversity of the sub-selected corpus. Our experiments show that all the methods are contributed to building a cleaner parallel corpus.

References

Alexandre Allauzen, Hélene Bonneau-Maynard, Hai-Son Le, Aurélien Max, Guillaume Wisniewski, François Yvon, Gilles Adda, Josep M Crego, Adrien Lardilleux, Thomas Lavergne, et al. 2011. Limsi@ wmt11. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 309–315. Association for Computational Linguistics.

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

Amittai Axelrod, Philip Resnik, Xiaodong He, and Mari Ostendorf. 2015. Data selection with fewer words. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 58–65, Lisbon, Portugal. Association for Computational Linguistics.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283. Association for Computational Linguistics.

Boxing Chen, Roland Kuhn, George Foster, Colin Cherry, and Fei Huang. 2016. Bilingual methods for adaptive training data selection for machine translation. In *Proc. of AMTA*, pages 93–103.

Michael Denkowski, Greg Hanneman, and Alon Lavie. 2012. The cmu-avenue french-english translation system. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 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.

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

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. A practical guide to support vector classification.

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, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224. Association for Computational Linguistics.

Philip Resnik and Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

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.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.