Tilde's Parallel Corpus Filtering Methods for WMT 2018

Mārcis Pinnis
Tilde / Vienibas gatve 75A, Riga, Latvia
marcis.pinnis@tilde.lv

Abstract

The paper describes parallel corpus filtering methods that allow reducing noise of noisy "parallel" corpora from a level where the corpora are not usable for neural machine translation training (i.e., the resulting systems fail to achieve reasonable translation quality; well below 10 BLEU points) up to a level where the trained systems show decent (over 20 BLEU points on a 10 million word dataset and up to 30 BLEU points on a 100 million word dataset). The paper also documents Tilde's submissions to the WMT 2018 shared task on parallel corpus filtering.

1 Introduction

Parallel data filtering for statistical machine translation (SMT) has shown to be a challenging task. Stricter filtering does not always yield positive results (Zariņa et al., 2015). This phenomenon can be explained with the higher robustness to noise of SMT systems, i.e., it does not harm the model if there are some incorrect translation candidates for a word or a phrase if the majority are still correct. However, there are also positive examples where data filtering allows improving SMT translation quality (Xu and Koehn, 2017). Neural machine translation (NMT), on the other hand, is much more sensitive to noise that is present in parallel data (Khayrallah and Koehn, 2018). From our own experience (as also shown by the experiments below), stricter filtering allows NMT models to show faster training tendencies and reach higher overall translation quality.

In this paper, we describe Tilde's methods for parallel data filtering for NMT system development and Tilde's submissions to the WMT 2018 shared task on parallel data filtering.

The paper is further structured as follows: Section 2 describes the data used in the filtering experiments, Section 3 provides details on the filtering methods that were applied to filter the parallel corpus of the shared task, Section 4 describes NMT experiments performed to evaluate the different filtering methods, Section 5 discusses the evaluation results, and Section 6 concludes the paper.

2 Data

The parallel data filtering experiments were performed on a German-English corpus that was provided by the WMT 2018 organisers. The corpus was a raw deduplicated subset[1] of the German-English ParaCrawl corpus[2]. It consists of one billion words and 104,002,521 sentence pairs.

For filtering, we require source-to-target and target-to-source probabilistic dictionaries. The dictionaries for the WMT 2018 experiments were acquired by 1) performing word alignment of the parallel corpora from the WMT 2018 shared task on news translation[3] (excluding the filtered ParaCrawl corpus) using *fast_align* (Dyer et al., 2013), and 2) performing raw probabilistic dictionary filtering using the transliteration-based probabilistic dictionary filtering method by Aker et al. (2014).

3 Filtering Methods

Although the filtering task required to score sentence pairs and not filter invalid sentence pairs out of the dataset, we start by filtering sentence pairs out of the raw corpus, after which we score each sentence pair and produce the scored output for submission. In order to filter the rather noisy "parallel" corpus, we use a combination of pre-existing parallel data filtering methods from the Tilde MT

[1] The corpus can be found online at http://www.statmt.org/wmt18/parallel-corpus-filtering.html.
[2] https://paracrawl.eu/download.html
[3] http://www.statmt.org/wmt18/translation-task.html

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 952–958

https://doi.org/10.18653/v1/W18-6487

	Filtering step	Sentence pairs	Proportion of the raw corpus
	Raw corpus	*104,002,521*	*100.00%*
Tilde MT filters for SMT systems			
1.1.	Identical source and target sentence filter	7,102,840	6.83%
1.2.	Sentence length ratio filter	5,276,660	5.07%
1.3.	Maximum sentence length filter	415,995	0.40%
1.4.	Maximum word length filter	286,485	0.28%
1.5.	Maximum word count filter	0	0.00%
1.6.	Unique sentence pair filter	20,821,646	20.02%
1.7.	Foreign word filter	14,983,927	14.41%
Additional Tilde MT filters for NMT systems			
2.1.	Empty sentence filter	222	0.00%
2.2.	Token count ratio filter	1,430,818	1.38%
2.3.	Corrupt symbol filter	33,519	0.03%
2.4.	Digit mismatch filter	20,534,497	19.74%
2.5.	Invalid character filter	630,818	0.61%
2.6.	Invalid language filter	1,229,434	1.18%
2.7.	Stricter sentence length ratio filter	1,710,401	1.64%
2.8.	Low content overlap filter	352,474	0.34%
Additional filters for the filtering task			
3.1.	Non-translated sentence filter	2,781,252	2.67%
3.2.	Maximum alignment filter	12,663,101	12.18%
	Sentence pairs after filtering	**13,748,432**	**13.22%**

Table 1: Statistics of sentence pairs removed by individual filtering steps

platform (Pinnis et al., 2018) and methods specifically developed to address the noisy nature of the ParaCrawl corpus. Some of the filtering methods feature hyperparameters, which were set empirically in parallel corpora filtering experiments. The first part of the filters were originally developed to increase SMT system quality. The filters are applied in the following order (for statistics of each individual filtering step, refer to Table 1):

1. **Identical source and target sentence filter** - validates whether the source sentence and the target sentence in a sentence pair are not identical. Although it may very well be that a sentence translates into the same sentence, it is also a strong indicator of non-translated sentence pairs.

2. **Sentence length ratio filter**. The filter validates whether the longest sentence (in terms of characters) is less than three times longer than the shortest sentence. This filter is meant to identify partially translated sentences. However, it has to be noted that this filter has been tested only for language pairs with Latin-based, Cyrillic-based, and Greek alphabets.

3. **Maximum sentence length filter** - validates whether neither the source nor the target sentence is longer than 1000 characters long.

4. **Maximum word length filter** - validates whether neither the source nor the target sentence contains tokens that are longer than 50 characters and do not contain directory separator characters. When extracting data from, e.g., PDF or image files, it may happen that word boundaries are not captured correctly. This may result in long words being formed in sentences. This filter is intended to remove such sentence pairs.

5. **Maximum word count filter** - validates whether neither the source nor the target sentence contains more than 400 tokens.

6. **Unique sentence pair filter** - validates whether a sentence pair is unique. The shared task organisers claimed that deduplication was performed[4], however, this filter removes

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

all white-spaces and punctuation marks, replaces all digit sequences with a numeral placeholder, and lowercases the sentence before validating the uniqueness of a sentence pair. Therefore, it is able to identify more redundant data.

7. **Foreign word filter** - validates whether the source sentence contains only words written in the alphabet of the source language and whether the target sentence contains only words written in the alphabet of the target language.

The filtering steps, which had been originally developed for SMT systems, removed a total of 48,887,553 sentence pairs. After these steps, 55,114,968 sentence pairs were left in the corpus.

As NMT systems have shown to be more sensitive to noise (Khayrallah and Koehn, 2018), the Tilde MT platform implements additional filtering steps that are stricter compared to the previous filters. Together with the parallel data noise, these filters may also remove valid sentence pairs. However, as shown by the results in Section 5, the amount of the parallel data is less important than the quality of the data. The following are the additional filtering steps that are used when preparing data for NMT systems:

1. **Empty sentence filter** - validates whether neither the source nor the target sentence is empty (or contains only white-space characters) after decoding HTML entities.

2. **Token count ratio filter** - The filter validates whether the token count ratio of the shortest sentence and the longest sentence is greater than or equal to 0.3 (in other words, if one sentence has three times as many tokens as the other sentence, then the sentence pair is considered invalid).

3. **Corrupt symbol filter** - validates whether neither the source nor the target sentence contains words that contain question marks between letters (e.g., '*flie?en*' instead of '*fließen*', '*gr??ere*' instead of '*größere*', etc.). Such words indicate encoding corruption in data, therefore, sentences containing such words are deleted.

4. **Digit mismatch filter** - validates whether all digits that can be found in the source sentence can also be found in the target sentence (and vice versa). Although this filter removes all sentence pairs where numbers that are written in digits have been translated into numbers written in words, it is effective for 1) identification of sentence breaking issues that are caused by incorrect handling of punctuation marks (e.g., cardinal numbers in some languages are written with the full stop character), and 2) identification of non-parallel content. By ensuring numeral writing consistency in parallel data, we can also ensure that digits will always be translated by the NMT systems as digits and numbers written in words as words.

5. **Invalid character filter** - validates whether neither the source nor the target sentence contains characters that have shown to indicate of encoding corruption issues. As most of potentially invalid (due to encoding corruption) sentence pairs are captured by the *foreign word filter* and the *corrupt symbol filter*, this filter provides just a minor addition - the list of invalid characters that are not included in valid alphabets consists of just four characters. However, this minor addition invalidates over 600 thousand sentence pairs.

6. **Invalid language filter** - validates whether the source sentence is written in the source language and whether the target sentence is written in the target language using a language detection tool (Shuyo, 2010). As language detection tools tend not to work well for shorter segments, this filter is applied only if the content overlap score (see below) between the source and target sentences is less than a trustworthy content alignment threshold (in the experiments set to 0.3) and the longest (source or target) sentence is at most two times longer than the shortest sentence.

7. **Stricter sentence length ratio filter** - validates whether the longest sentence (in terms of characters is less than two times longer than the shortest sentence.

8. **Low content overlap filter** - validates whether the content overlap according to the cross-lingual alignment tool *MPAligner* (Pinnis, 2013) is over a threshold. Because the content overlap metric produced by

MPAligner represents the level of parallelity, it is used to score sentence pairs. Therefore, the threshold was also set to a low value (0.01).

This far, a total of 74,809,736 were removed from the corpus, leaving a total of 29,192,785 sentence pairs remaining in the corpus.

When training NMT systems with the subsampled datasets, we identified that there were frequent (wrong) many-to-many alignments left in the corpus even after filtering. We also found that the corpus contained many entries with text in both languages on one side (i.e., imagine a translation where some of the source words are translated, but the majority is just copied over from the source segment and left untranslated), which contribute to parallel data noise. Therefore, we introduced two additional filters that address these issues:

1. **Non-translated sentence filter** - validates whether more than half of the source words have been translated (i.e., are not present in the target sentence).

2. **Maximum alignment filter** - keeps only those sentence pairs where the target sentence is the highest scored target sentence for the source sentence (according to the content overlap scores) and vice versa.

After all filtering steps, there were 13,748,432 sentence pairs left in the *Max Filtered+* corpus. In order to compare whether the full filtering workflow produces better results than a part of the workflow, we also prepared the following intermediate datasets:

1. *Filtered* - the corpus filtered up to and including the *low content overlap filter*. The dataset consists of 29,192,785 sentence pairs.

2. *Max Filtered* - the corpus filtered using all filters except the *Non-translated sentence filter*. The dataset consists of 15,613,062 sentence pairs.

3. *Filtered+* - the corpus filtered up to and including the *non-translated sentence filter*. The dataset consists of 26,411,533 sentence pairs.

4. *Max Filtered+ Rescored* - the corpus filtered using all filters and rescored by ranking sentences with a Round-robin-based method according to source sentence lengths. I.e., all sentence pairs were separated into different lists according to sentence lengths and sorted according to the content overlap scores in a descending order. Then, sentences were ranked by assigning the highest score to the best-scored unigram sentence, the second highest score to the best-scored bigram sentence, etc. We performed such rescoring, because the filtering assigned higher scores to shorter segments, thereby skewing the sentence length statistics towards shorter sentences. The dataset consists of 16,529,684 sentence pairs.

In each of the datasets (except for the *Max Filtered+ Rescored* dataset), sentence pairs were scored using the content overlap metric produced by *MPAligner*. In order to create scores for the raw dataset (i.e., to create submissions for the shared task), we scored each sentence pair in the raw dataset as follows: if a sentence pair was found in a particular filtered dataset, the sentence pair was scored using the score produced by *MPAligner* (or the rescoring method), otherwise the sentence pair received the score '*0*'. This means that all sentence pairs that were filtered out by any of the filtering steps, received the score '*0*'.

4 Trained Systems

To evaluate, which of the datasets allows achieving higher translation quality, we performed subsampling of the filtered datasets into 10 million and 100 million word datasets. For this, we used the *subselect.perl* script, which was provided by the organisers in the *dev-tools* package[5]. Then, we trained attention-based NMT systems with gated recurrent units in the recurrent layers using the Marian toolkit (Junczys-Dowmunt et al., 2018). All systems were trained using the configuration that is provided in the same package until convergence.

In addition to the filtered dataset systems, we trained four baseline systems. The first two baseline systems were trained on datasets, which were subsampled using the Hunalign (Varga et al., 2007) scores that were provided by the organisers. For the other two systems, data subsampling was performed on randomly assigned scores.

The NMT system training progress (in terms of BLEU scores on the raw tokenised development

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

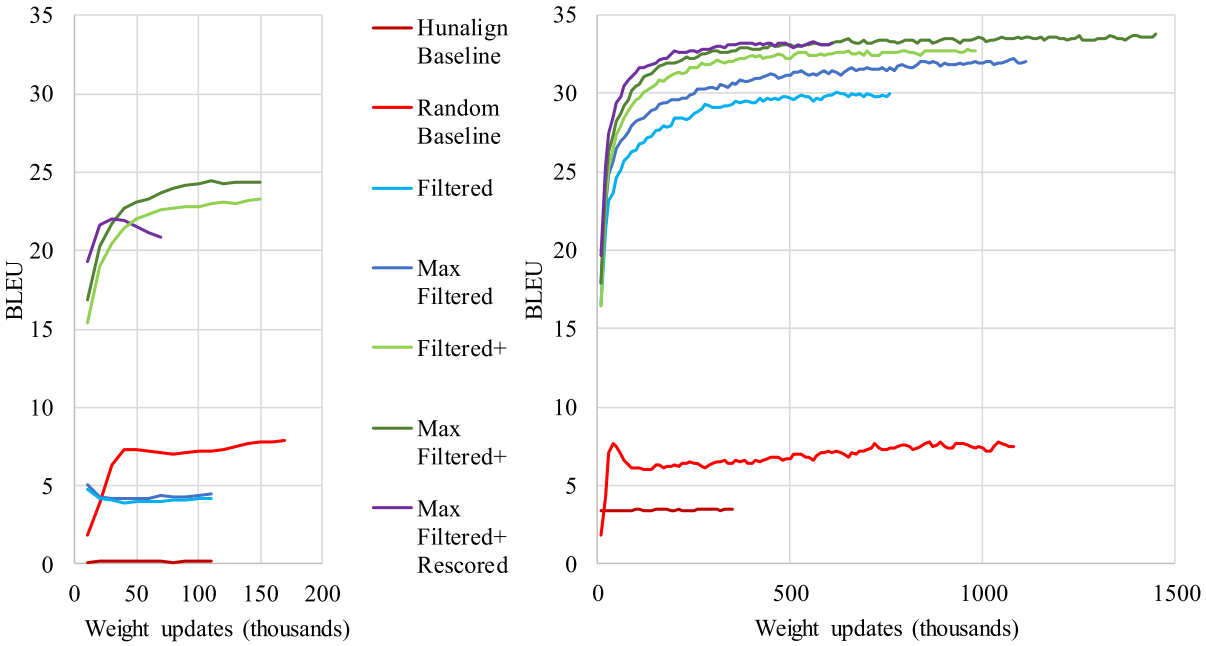

Figure 1: Training progress of NMT systems (10 million word systems - left; 100 million word systems - right)

set) is depicted in Figure 1. The figure shows that for the small dataset systems, only the systems with the *non-translated sentence filter* were able to achieve results of over 20 BLEU points. All other systems show rather poor performance, indicating the necessity of careful data cleaning. It is also evident that the *Filtered* and *Max Filtered* datasets contain too much noise among the highest scored sentence pairs. The reason for this is because the content overlap filter (by design) does not look at whether a sentence pair is a reciprocal translation. It tries to identify, just like a word alignment tool, which words in the source sentence correspond to which words in the target sentence, and non-translated words can be paired easily.

Although for the large dataset systems the *Filtered* and *Max Filtered* datasets contain higher levels of noise (compared to the more filtered datasets), they show comparative (however, lower) results to the more filtered datasets. The fact that the datasets are approximately 10 times larger than the smaller datasets allowed for higher quality sentence pairs to be included in the data sub-selected for NMT system training.

The figure also shows an interesting tendency for the *Max Filtered+ Rescored* dataset. In both experiments (10 million and 100 million word systems) the quality increases at the beginning, but then it starts to drop – very noticeably for the small system and slightly for the large system.

5 Results

Automatic evaluation results in terms of BLEU (Papineni et al., 2002) scores are provided in Table 2. For all systems, we used the '*test.sh*' script that was provided by the organisers in order to translate the test set and evaluate each model's translation quality.

The evaluation results illustrate the same dataset rankings as the training progress chart. The best results are achieved by using the *Max Filtered+* dataset.

We were also interested in seeing whether the filtering methods (by improving the parallel data quality) also allow improving out-of-vocabulary (OOV) word rates on the development set. It is evident in Table 2 that the OOV rate decreases by adding more filtering steps. However, there is one exception – the translation quality of the NMT systems, which were trained using the *Max Filtered+ Rescored* dataset, decreases although the OOV rate drops (especially when calculated for unique tokens). There may be multiple explanations for the quality decrease. For instance, for the smaller (10 million word) dataset, the rescoring introduced a higher percentage of lower quality sentence pairs due to the fact that the frequency of longer sentences is naturally lower than that of shorter sentences. E.g., there are 746,480 English

System	BLEU	BLEU-C	Development data OOV rate (running)	(unique)
10 million token experiments				
Hunalign Baseline	0.15	0.14	8.27%	32.08%
Random Baseline	8.41	7.74	3.31%	13.25%
Filtered	4.86	4.32	6.25%	25.28%
Max Filtered	5.00	4.43	5.99%	24.63%
Filtered+	21.35	19.75	4.54%	18.44%
Max Filtered+	**21.95**	**20.42**	4.27%	17.25%
Max Filtered+ Rescored	20.10	18.75	**3.29%**	**12.87%**
100 million token experiments				
Hunalign Baseline	3.64	3.28	1.78%	7.16%
Random Baseline	7.26	6.75	1.32%	5.43%
Filtered	27.72	26.14	1.39%	5.65%
Max Filtered	29.06	27.46	1.28%	5.17%
Filtered+	30.24	28.59	1.32%	5.24%
Max Filtered+	**30.83**	**29.14**	**1.31%**	5.10%
Max Filtered+ Rescored	30.40	28.78	1.32%	**4.95%**

Table 2: Evaluation results of NMT systems trained using different sub-sampled filtered datasets (the table shows case-insensitive BLEU and case-sensitive BLEU (BLEU-C))

sentences that consist of five tokens, compared to just 2673 sentences of 80 tokens in the *Max Filtered+* dataset (which was used to acquire the rescored dataset). This means that the rescoring method was forced to select lower quality longer sentence pairs simply because of insufficient sentence pairs to select from. For the larger dataset, the results also show that the running OOV rate is slightly larger than the unique token OOV rate. However, the issue with the limited number of longer sentences did affect also the larger system as the sub-sampled dataset included all sentence pairs that were longer than or equal to 42 tokens regardless of their quality. For future work, it could be beneficial to investigate whether a fixed content overlap threshold could allow the rescoring method to perform better.

For the WMT 2018 shared task, we submitted the following three datasets:

1. *tilde-isolated (Filtered+)* – this dataset represents isolated sentence filtering where only individual sentence pairs are passed to the filtering method.

2. *tilde-max (Max Filtered+)* – this dataset represents full corpus filtering where (in addition to the filtering results of a particular sentence pair) also information about other sentence pairs is used to decide whether to keep a sentence pair or not.

3. *tilde-max-rescored (Max Filtered+ Rescored)* – this dataset represents both full corpus filtering and (a rather simple) data selection method.

6 Conclusion

The paper presented parallel corpus filtering methods that allow reducing the noise in noisy "parallel" corpora to a level where the corpus is usable in neural machine translation system development. Most of the filtering methods are simple (except for the *low content overlap filter*) and do not require any machine learning methods to be implemented (except for the *invalid language filter*). We showed that, by applying stricter filtering methods, NMT system quality increases.

For the WMT 2018 shared task on corpus filtering, we submitted three scored datasets that represent isolated sentence filtering (*Filtered+*), full corpus filtering (*Max Filtered+*), and (a rather simple method for) full corpus filtering with data selection (*Max Filtered+ Rescored*).

The filtering methods are integrated into the Tilde MT platform and serve its users when they require SMT and NMT system training.

For future work, it may be beneficial to perform ablation experiments, to identify, which of the individual filtering methods contributes the most in order to acquire a higher quality parallel corpus.

Acknowledgments

The research has been supported by the European Regional Development Fund within the research project "Neural Network Modelling for Inflected Natural Languages" No. 1.1.1.1/16/A/215.

References

Ahmet Aker, Monica Lestari Paramita, Mārcis Pinnis, and Robert Gaizauskas. 2014. Bilingual Dictionaries for All EU Languages. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)*, pages 2839–2845, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, June, pages 644–648, Atlanta, USA.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Mārcis Pinnis. 2013. Context Independent Term Mapper for European Languages. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*, pages 562–570, Hissar, Bulgaria.

Mārcis Pinnis, Andrejs Vasiļjevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš, and Valters Šics. 2018. Tilde MT Platform for Developing Client Specific MT Solutions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nakatani Shuyo. 2010. Language detection library for java.

Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor Tron. 2007. Parallel corpora for medium density languages. *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, 292:247.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950.

Ieva Zariņa, Pēteris Ņikiforovs, and Raivis Skadiņš. 2015. Word Alignment Based Parallel Corpora Evaluation and Cleaning Using Machine Learning Techniques. In *Proceedings of the Eighteenth Annual Conference of the European Association for Machine Translation (EAMT 2015)*, pages 185–192, Antalya. European Association for Machine Translation.