Improving Parallel Data Identification using Iteratively Refined Sentence Alignments and Bilingual Mappings of Pre-trained Language Models

Chi-kiu Lo and **Eric Joanis**
Multilingual Text Processing
Digital Technologies Research Centre
National Research Council Canada (NRC-CNRC)
1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
{chikiu.lo,eric.joanis}@nrc-cnrc.gc.ca

Abstract

The National Research Council of Canada's team submissions to the parallel corpus filtering task at the Fifth Conference on Machine Translation are based on two key components: (1) iteratively refined statistical sentence alignments for extracting sentence pairs from document pairs and (2) a crosslingual semantic textual similarity metric based on a pretrained multilingual language model, XLM-RoBERTa, with bilingual mappings learnt from a minimal amount of clean parallel data for scoring the parallelism of the extracted sentence pairs. The translation quality of the neural machine translation systems trained and fine-tuned on the parallel data extracted by our submissions improved significantly when compared to the organizers' LASER-based baseline, a sentence-embedding method that worked well last year. For re-aligning the sentences in the document pairs (component 1), our statistical approach has outperformed the current state-of-the-art neural approach in this low-resource context.

1 Introduction

The aim of the Fifth Conference on Machine Translation (WMT20) shared task on parallel corpus filtering (Koehn et al., 2020) is essentially the same as the two previous editions (Koehn et al., 2018b, 2019): identifying high-quality sentence pairs in a noisy corpus crawled from the web using ParaCrawl (Koehn et al., 2018a), in order to train machine translation (MT) systems on the clean data.

This year, the low-resource language pairs being tested are Khmer–English (km–en) and Pashto–English (ps–en). Specifically, participating systems must produce a score for each sentence pair in the test corpora indicating the quality of that pair. Then samples containing the top-scoring 5M words are used to train MT systems. While using the filtered parallel data to train a FAIRseq (Ott et al., 2019) neural machine translation (NMT) system remains the same as last year, the organisers are no longer building statistical machine translation (SMT) systems as part of the task evaluation. Instead, as an alternative evaluation, the filtered parallel corpus is used to fine-tune an MBART (Liu et al., 2020) pretrained NMT system. Participants were ranked based on the performance of these MT systems on a test set of Wikipedia translations (Guzmán et al., 2019), as measured by BLEU (Papineni et al., 2002). A few small sources of parallel data, covering different domains, were provided for each of the two low-resource languages. Much larger monolingual corpora were also provided for each language (en, km and ps). In addition to the task of computing quality scores for the purpose of filtering, there is also a sub-task of re-aligning the sentence pairs from the original crawled document pairs.

Cleanliness or quality of parallel corpora for MT systems is affected by a wide range of factors, e.g., the parallelism of the sentence pairs, the fluency of the sentences in the output language, etc. Previous work (Goutte et al., 2012; Simard, 2014) showed that different types of errors in the parallel training data degrade MT quality in different ways. Cross-lingual semantic textual similarity is one of the most important properties of high-quality sentence pairs. Lo et al. (2016) scored cross-lingual semantic textual similarity in two ways, either using a semantic MT quality estimation metric, or by first translating one of the sentences using MT, and then comparing the result to the other sentence, using a semantic MT evaluation metric. At the WMT18 parallel corpus filtering task, Lo et al. (2018)'s supervised submissions were developed for the same MT evaluation pipeline using a new semantic MT metric, YiSi-1 (Lo, 2019) (see also section 2.3). At the WMT19 parallel corpus filtering task, Bernier-

Proceedings of the 5th Conference on Machine Translation (WMT), pages 972–978

Colborne and Lo (2019) exploited the quality estimation metric YiSi-2 using bilingual word embeddings learnt in a supervised manner (Luong et al., 2015) from clean parallel training data or a weakly supervised manner (Artetxe et al., 2016) from bilingual dictionary. Lo and Simard (2019) further showed that using YiSi-2 with multilingual BERT (Devlin et al., 2019) on fully unsupervised parallel corpus filtering (i.e. without access of any parallel training data) achieved similar results to those in Bernier-Colborne and Lo (2019).

This year, the National Research Council of Canada (NRC) team submitted one system to the parallel corpus filtering task and one to the alignment task. The two systems share the same components in scoring the parallelism of the noisy sentence pairs, i.e., the pre-filtering rules and the quality estimation metric YiSi-2. For the parallel corpus aligning task, we use an iterative statistical alignment method to align sentences from the given document pairs before passing the aligned sentences to the scoring pipeline.

Our internal results show that MT systems trained on pre-aligned sentences filtered by our scoring pipeline outperform those trained on the organizers' LASER-based baseline (Chaudhary et al., 2019) by 0.2–1.4 BLEU. Training MT systems on re-aligned sentences using our iterative statistical alignment method achieve further gains of 0.3–1.8 BLEU.

2 System architecture

There are a wide range of factors that determine whether a sentence pair is good for training MT systems. Some of the more important properties of a good training corpus include:

- High parallelism in the sentence pairs, which affects translation adequacy.
- High fluency and grammaticality, especially for sentences in the output language, which affect translation fluency.
- High vocabulary coverage, especially in the input language, which helps make the translation system more robust.
- High variety of sentence lengths, which should also improve robustness.

In previous years, we explicitly tried to maximize all four of these properties, but this year we focused only on the first two in the scoring presented in section 2.3 below.

2.1 Iterative statistical sentence alignment

Our iterative statistical sentence alignment method as detailed in Joanis et al. (2020) uses `ssal`, a reimplementation and extension of Moore (2002) which is part of the Portage statistical machine translation toolkit (Larkin et al., 2010).

First, we train an IBM-HMM model (Och and Ney, 2003) on the clean parallel training data and the subsampled noisy corpora (see Table 1 for statistics) and use it to align paragraphs in the given document pairs, as Moore (2002) does. The subsampled noisy corpora are those obtained by applying our filtering baseline as described in sections 2.2 and 2.3 (and denoted as "nrc.baseline" in table 2). Then, we segment the paragraphs in both languages into sentences using the Portage sentence splitter. Finally, we align sentences within aligned paragraphs using the IBM model again. In this process, both the data used in training the IBM-HMM model and the noisy document pairs for alignment are punctuation tokenized using the Portage tokenizer.

In past work on sentence alignment (Joanis et al. (2020) and other unpublished experiments), we have found that first aligning paragraphs and then aligning sentences within aligned paragraphs outperforms approaches that align sentences without paying attention to paragraph boundaries.

2.2 Initial filtering

The pre-filtering steps of our submissions are mostly the same as those in Bernier-Colborne and Lo (2019). We remove:

1. duplicates after masking email, web addresses and numbers,
2. sentence pairs with a majority of number mismatches,
3. sentence pairs with either side in the wrong language according to the `pyCLD2` language detector[1],
4. sentence pairs where over half of the source sentence is non-alphabetical or target language characters, and
5. sentence pairs where over half of the target sentence is non-alphabetical characters.

An additional pre-filtering rule included in this year's submissions is the removal of pairs where over 50% of the target English sentence is directly

[1] https://github.com/aboSamoor/pycld2

Lang(s)	Training data sources	#sentence pairs	#source tokens	#target tokens
clean parallel				
km–en	JW300, Bible, GNOME/KDE/Ubuntu, Tatoeba, Global Voices	290k	6M	4M
ps–en	Bible, GNOME/KDE/Ubuntu, Wikimedia, TED Talks, Tatoeba	123k	792k	662k
filtered noisy				
km–en	ParaCrawl	288k	2M	5M
ps–en	ParaCrawl	393k	6M	5M

Table 1: Data used to train the IBM-HMM model used in the iterative statistical sentence alignment.

copying from the source Khmer or Pashto sentence.

2.3 Sentence pair scoring

The core of our sentence pair scoring component is the semantic MT quality estimation metric, YiSi-2. YiSi (Lo, 2019) is a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. YiSi-1 measures the similarity between a machine translation and human references by aggregating weighted distributional (lexical) semantic similarities, and optionally incorporating shallow semantic structures. YiSi-2 is the bilingual, reference-less version, which uses bilingual word embeddings to evaluate cross-lingual lexical semantic similarity between the input and MT output or, in this task, between the source and target sentences.

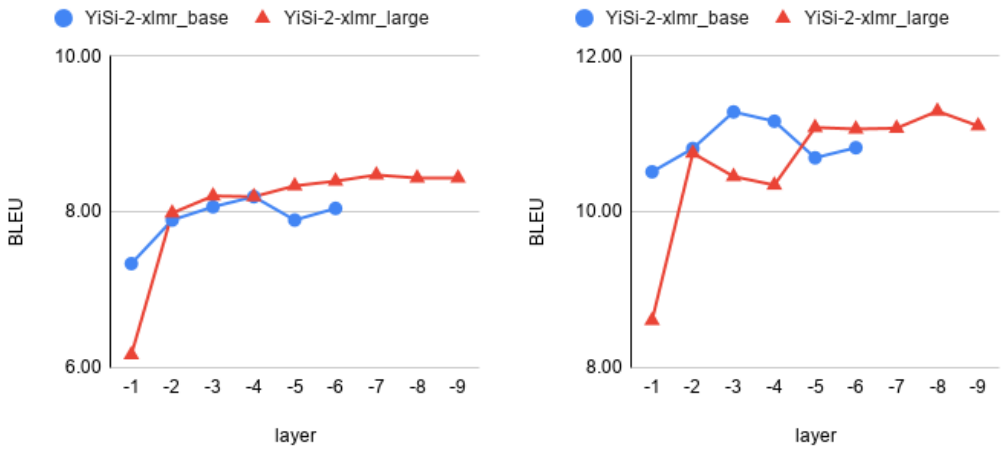
YiSi-2 relies on a crosslingal language representation to evaluate the crosslingual lexical semantic similarity. Previously, it used pre-trained multilingual BERT (Devlin et al., 2019) for this purpose. In this work, we instead experiment with XLM-RoBERTa (Conneau et al., 2020) because (1) at the time this work was done, it was the only pre-trained multilingual language encoder that covers both Khmer, Pashto and English; and (2) it shows better performance with lower-resource languages than BERT.

As suggested by Devlin et al. (2019); Peters et al. (2018); Zhang et al. (2020), we experiment with using contextual embeddings extracted from different layers of the multilingual language encoder to find out the layer that best represents the semantic space of the language.

YiSi is semantic oriented. In the past, we noticed that YiSi-based scoring functions failed to filter out sentence pairs with disfluent target text. Following Zhao et al. (2020), we experiment with improving the sentence pair scoring function by linearly combining YiSi score with the language model (LM) scores of the target text obtained from the multilingual language model used in YiSi. However, instead of using an additional pretrained language model—GPT-2 (Radford et al., 2019)—as in Zhao et al. (2020), we use the left-to-right LM scores obtained from XLM-RoBERTa while computing the crosslingual lexical semantic similarity. The advantages of using the same pretrained model for computing the crosslingual lexical semantic similarity and the language model scores are 1) it costs less in both memory and computation; 2) it is more portable to languages other than English. We combined the LM scores in the probability domain linearly with the semantic similarity scores with a weight of 0.1 assigned to the LM scores.

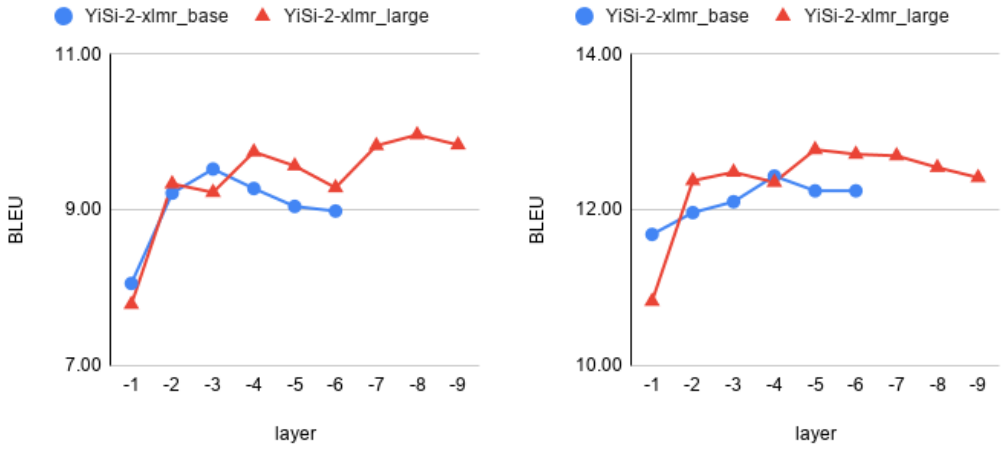
In the WMT19 metrics shared task (Ma et al., 2019), we saw a very significant performance degradation between YiSi-1 and YiSi-2. This suggests that current multilingual language models construct a shared multilingual space in an unsupervised manner without any direct bilingual signal, in which representations of context in the same language are likely to cluster together in part of the subspace and there is a language segregation in the shared multilingual space. Inspired by Artetxe et al. (2016) and Zhao et al. (2020), we sample 5k clean sentence pairs and use the token pairs aligned by maximum alignment of their semantic similarity to train a cross-lingual linear projection that would transform the source embeddings into the target embeddings subspace.

Lo and Larkin (2020) provide a detailed correlation analysis of YiSi-2 with all the improvements mentioned above and human judgment on



(a) km–en FAIRseq

(b) km–en MBART



(c) ps–en FAIRseq

(d) ps–en MBART

Figure 1: BLEU scores on the Khmer–English dev set for (a) FAIRseq and (b) MBART and the Pashto–English dev set for (c) FAIRseq and (d) MBART trained on 5M-word parallel subsample extracted according to the scoring functions as shown: on the x-axis, layer = $-n$ means YiSi-2 based on the embeddings of the n^{th} layer, counting from the last, of XLM-RoBERTa$_{base}$ (blue circles) or XLM-RoBERTa$_{large}$ (red triangles).

MT reference-less evaluation.

3 Experiments and results

We used the software provided by the task organizers to extract the 5M-word samples from the original test corpora according to the scores generated by each alignment and/or filtering system. We then trained a FAIRseq MT system or fine-tuned an MBART pretrained NMT using the extracted subsamples. The MT systems were then evaluated on the official dev set ("dev-test").

We exhaustively experimented with the last few layers of both XLM-RoBERTa$_{base}$ and XLM-RoBERTa$_{large}$ in order to find out the model and layer best representing crosslingual semantic similarity. Figure 1 shows the plots of the change in BLEU scores of each MT system using the embeddings extracted from the n^{th} layer, counting from the last, of the multilingual LM for evaluating crosslingual lexical semantic similarity. In general, we see a trend of rising performance as we roll back from the last layer. The performance peaks at some point and starts to fall when we roll back too far from the end. For XLM-RoBERTa$_{base}$, the peak performance of the MT systems is achieved by the 3$^{\text{rd}}$ or 4$^{\text{th}}$ last layer (out of 12 layers). For XLM-RoBERTa$_{large}$, the peak performance of the MT systems is achieved by the 8$^{\text{th}}$ last layer (out of 24 layers). The peak performance of MT systems trained on sentences filtered by XLM-RoBERTa$_{large}$ based YiSi-2 is better than that by XLM-RoBERTa$_{base}$ based YiSi-2.

system	alias	km–en		ps–en	
		FAIRseq	MBART	FAIRseq	MBART
filtering only					
LASER	baseline	7.10	10.13	9.77	11.03
+ filter rules		7.55	10.44	9.87	11.91
YiSi-2-xlmr_large (layer -8) + filter rules	nrc.baseline	8.43	11.29	**9.96**	12.54
+ LM score		8.53	11.31	9.61	**12.82**
+ LM score + CLP_{5k}	nrc.filtering	**8.54**	**11.58**	9.93	12.80
re-aligning and filtering					
iterative alignment + nrc.filtering	nrc.alignment	***8.82***	11.17	***11.73***	***13.21***

Table 2: BLEU scores of selected systems. The two final submitted systems are labelled nrc.filtering and nrc.alignment.

Table 2 shows the results of the experiments described in section 2.3. First, we show an improved version of the organizers' baseline by simply adding our initial filtering rules. This shows that our initial filtering rules are able to catch bad parallel sentences which are hard to filter by an embedding-based filtering system.

Next, we see that using YiSi-2 with XLM-RoBERTa$_{large}$'s 8th last layer as parallelism scoring function outperforms the LASER baseline by 0.1–0.9 BLEU in different translation directions and MT architectures. This is our "nrc.baseline" system, and the baseline used for filtering the noisy corpus in training the IBM-HMM alignment model for the "nrc.alignment" system. Adding the LM score to the scoring function shows small improvements. Learning the cross-lingual linear projection matrix to transform the source embeddings in the target language subspace shows more improvements overall. This is our "nrc.filtering" submission to the parallel corpus filtering task.

At last, we show that using our iterative statistical alignment method to redo the alignment of sentences from the given document pairs improves the translation quality of the resulting MT systems significantly. This is our "nrc.alignment" submission to the parallel corpus filtering task.

4 Conclusion and Future Work

In this paper, we presented the NRC's two submissions to the WMT20 Parallel Corpus Filtering and Alignment for Low-Resource Conditions task. Our experiments show that YiSi-2 is a scoring function of parallelism that is very competitive, and that a statistical sentence alignment method is still able to provide better alignment results than neural ones in low resource situations. Further analysis is required to understand the characteristics of the sentence pairs aligned by the baseline vecalign and our iterative statistical sentence alignment and how the latter achieves better translation quality for the trained MT systems.

It is worth highlighting that in this task, as well as in our Inuktitut–English corpus alignment work (Joanis et al., 2020), a well-tuned statistical sentence-alignment system outperformed a state-of-the-art neural one. We hypothesise that this is a low-resource effect, but further work is still needed to explore the best low-resource corpus alignment methods. In particular, we intend to integrate YiSi-2 into our sentence aligner to test whether it's our iterative alignment methodology that makes the difference or the fact that the underlying scoring function is statistical (we use IBM-HMM models for sentence pair scoring in our aligner). It's possible that the statistical approach might continue to win here, because in the low-resource context there might not be enough training data to tune the orders of magnitude more parameters of the neural models; a counter-argument is that YiSi-2 did better on the scoring task than statistical scoring functions. Our future work will explore the trade-offs between these two approaches, and consider hybrid methods.

Acknowledgements

We thank Samuel Larkin and Marc Tessier for their help in setting up the FAIRseq and MBART baselines using the LASER scores; and Patrick Littell for discussion and feedback on the Pashto test set. We also thank the reviewers for their comments and suggestions, and Roland Kuhn for his comments and feedback on the paper.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. 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.

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. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

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, Kenneth Heafield, Mikel L. Forcada, Miquel Esplà-Gomis, Sergio Ortiz-Rojas, Gema Ramírez Sánchez, Víctor M. Sánchez Cartagena, Barry Haddow, Marta Bañón, Marek Střelec, Anna Samiotou, and Amir Kamran. 2018a. ParaCrawl corpus version 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel Forcada. 2018b. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Samuel Larkin, Boxing Chen, George Foster, Ulrich Germann, Eric Joanis, Howard Johnson, and Roland Kuhn. 2010. Lessons from NRC's Portage system at WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 127–132, Uppsala, Sweden. 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.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo, Cyril Goutte, and Michel Simard. 2016. CNRC at SemEval-2016 task 1: Experiments in crosslingual semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 668–673, San

Diego, California. Association for Computational Linguistics.

Chi-kiu Lo and Samuel Larkin. 2020. MT reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Chi-kiu Lo and Michel Simard. 2019. Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 206–215, Hong Kong, China. Association for Computational Linguistics.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 908–916, Belgium, Brussels. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, pages 135–144.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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.

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*, Philadelphia, Pennsylvania, USA.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Michel Simard. 2014. Clean data for training statistical MT: the case of MT contamination. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas*, pages 69–82, Vancouver, BC, Canada.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.