Quality and Coverage: The AFRL Submission to the WMT19 Parallel Corpus Filtering For Low-Resource Conditions Task

Grant Erdmann, Jeremy Gwinnup
Air Force Research Laboratory
grant.erdmann@us.af.mil,jeremy.gwinnup.1@us.af.mil

Abstract

The WMT19 Parallel Corpus Filtering For Low-Resource Conditions Task aims to test various methods of filtering noisy parallel corpora, to make them useful for training machine translation systems. This year the noisy corpora are from the relatively low-resource language pairs of English-Nepali and English-Sinhala. This papers describes the Air Force Research Laboratory (AFRL) submissions, including preprocessing methods and scoring metrics. Numerical results indicate a benefit over baseline and the relative effects of different options.

1 Introduction

For this task the participants were provided with a corpus of parallel data in English-Nepali (en-ne) and English-Sinhala (en-si). Both parallel and monolingual training datasets were provided in these languages. The task organizers built statistical machine translation (SMT) and neural machine translation (NMT) systems from the scores produced, based on parallel training sets of 1M (one million) and 5M English words.

Subset selection techniques often strive to reduce a set to the most useful. For the shared task one should avoid selecting:

- A line with undue repetition of content of other selected lines. This repetition can extend training times and/or skew the translation system to favor this type of line.
- Long lines, which will be ignored in training the MT systems.

In addition to adapting the corpus to the building of a general-purpose MT system, we must also deal with significant noise. The main types of noise present in the given data are:

- Not natural language
- One or both languages are incorrect
- Lines are not translations of each other

In contrast to our WMT18 submission (Erdmann and Gwinnup, 2018), we include a text quality metric in the subcorpus-building process, rather than combining it afterward.

2 Preprocessing

As a first step, a rough preprocessing filter is applied to the data.

We remove lines where either language text contains more than 80 words, since the test systems use a maximum of 80 words per line. We also remove lines where the language ID probabilities from fastText (Joulin et al., 2016b,a) do not match the expected languages (using the pre-built language ID models of the authors).

This preprocessed text is used to generate the scores that determine a line's usefulness. We note that there are many fewer preprocessing steps than our previous system (Erdmann and Gwinnup, 2018). We can simplify preprocessing because inclusion of a text quality metric during subcorpus-building will avoid other forms of noise in the process.

3 Coverage Metric

Our metric for subcorpus-building uses both a coverage metric and a text quality metric.

We first give our coverage metric (Gwinnup et al., 2016). Let us select a subcorpus S from a larger corpus C to maximize its similarity to a representative corpus T. Let our preferred subselected subcorpus size be τ times the size of T. Let $\mathcal{V}$ be a set of vocabulary elements of interest. Defining $c_v(X)$ to be the count of the occurrence

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 267–270
Florence, Italy, August 1-2, 2019.

of feature $v \in \mathcal{V}$ in a given corpus X, the coverage g is given by

$$g(S,T,\tau) = \frac{\sum_{v\in\mathcal{V}} f(\min(c_v(S), c_v^\tau(T)))}{\sum_{v\in\mathcal{V}} f(c_v^\tau(T)) + p_v(S,T,\tau)} \quad (1)$$

where the oversaturation penalty $p_v(S,T,\tau)$ is

$$\max(0, c_v(S) - c_v^\tau(T))\,[f(c_v^\tau(T)+1) - f(c_v^\tau(T))]\,.$$

Here f can be any submodular function, but we choose exclusively $f(x) = \log(1+x)$. The scaled count $c_v^\tau(T) = \tau c_v(T)$ accounts for the preferred size of the selected subcorpus differing from the size of T.

4 Text Quality Metric

To create a text quality metric, we use the given clean parallel data to create a MT system. We use the MT system to translate both pre-filtered noisy parallel corpora into English.

This allows us to compute the Meteor (Denkowski and Lavie, 2014) score of the given English lines, using the translated English as a reference. The Meteor metric was chosen due to its using deeper linguistic information than BLEU. The text quality metric of a subcorpus is given by its average:

$$h(S) = \frac{\sum_{s\in S} m(s)}{\sum_{s\in S} 1} \quad (2)$$

where $m(s)$ is the text quality metric (e.g., Meteor) score of line s. This corpus metric is defined to be zero for the empty corpus: $h(\emptyset) = 0$.

The overall score of a subcorpus is given by the product of the coverage metric (1) and the quality metric (2):

$$F(S,T,\tau) = g(S,T,\tau)h(S) \quad (3)$$

5 Subcorpus-Building Algorithm

To build a subcorpus, we iterate the following two steps until the selected subcorpus is large enough:

1. Add the line that has the best effect on the overall score F from (3).
2. If removal of any line would improve F, find the line with the largest improvement. Remove it, unless infinite cycling would result.

This is a greedy algorithm, with review after each selection.

6 Application

This section outlines the particulars of the method applied to the given data for this task. Pre-filtering removed a significant percentage of the noisy parallel corpora prior to scoring. The thresholds for language identification were set empirically. For en-ne we used 40% for English and 1% for Nepali. For en-si we used 10% for both English and Sinhala. After filtering for language identification and a maximum of 80 words, 0.9M of the 2.2M lines remained for en-ne and 1.2M of the 3.4M lines remained for en-si.

We trained phrase-based Moses (Koehn et al., 2007) systems with the small amount of "clean" training data provided by the organizers. These training corpora were normalized as necesssary to remove systematic representation oddities, mostly in punctuation. The Moses systems employ a hierarchical reordering model (Galley and Manning, 2008) and 5-gram operation sequence model (Durrani et al., 2011). The 5-gram English language model used by both systems was trained with the constrained monolingual corpus from our WMT15 (Gwinnup et al., 2015) efforts.

These Moses MT systems were used to translate the pre-filtered datasets. The Meteor score of the given English lines was computed, using the translated English as a reference.

The pre-filtered parallel corpora were lowercased and tokenized with tools from Moses. We built a 2000-word-vocabulary SentencePiece (Kudo and Richardson, 2018) model on the given monolingual corpora for each language. The pre-filtered parallel corpora were processed with these models prior to subcorpus-building.

Our subcorpus-building procedure was followed, producing a subcorpus that we ranked by the order a line was added to the subcorpus. This can produce too few scored lines for the 1M-word or 5M-word subcorpora, so we order the scores of the remaining lines by their text quality metric (i.e., Meteor) scores alone. We submitted scores generated by two values of τ for each language pair. The smaller value of τ produced a 50k-line subcorpus, and the larger value of τ produced 150k lines. Our expectation was that the smaller subcorpus would be best in the 1M-word case, and the larger subcorpus in the 5M-word case. For these cases the selected corpora were roughly the same size as the training sets.

7 Numerical Results

The official results of the WMT19 Parallel Filtering Task are given by Bojar et al. (2019).

Here we give some general findings by using the given Moses-EMS configuration for the task. Tables 1–2 give numerical results of this test. BLEU scores are uncased and produced during the Moses-EMS run. We see that the parallel filtering methods we expected to be best do in fact improve on the Zipporah (Xu and Koehn, 2017) baseline.

The smaller, 50k-line subcorpus shows increases of by 0.24 BLEU for 1M en-ne and 0.15 BLEU for 1M en-si. The larger, 150k-line subcorpus shows increases of by 0.11 BLEU for 5M en-ne and 0.32 BLEU for 5M en-si. Picking the best results over all our experiments shows greater improvements over baseline: 0.48 BLEU for 1M en-ne, 0.46 BLEU for 1M en-si, 0.11 BLEU for 5M en-ne, and 0.44 BLEU for 5M en-si.

The tables show that the subcorpus-building process normally improves over scoring by the text quality metric score alone (the row labelled "quality", which is equivalent to either building an empty subcorpus or choosing $F = h$ in (3)). These improvements are largest and most consistent in the 1M-word tests. We expect that the larger sets might be struggling to find helpful data in the noisy corpora, essentially converging to the text-quality-metric-only score.

We tested excluding the text quality metric from the selection process (i.e., choosing $F = g$ in (3)), and these tests are given in the table rows labelled "coverage". As in (Erdmann and Gwinnup, 2018), we saw great benefit from including the text quality using an MT system, even in this low-resource setting.

Varying the number of grams considered in the subcorpus-building algorithm's vocabulary yielded small and inconsistent changes over unigram selection. We have no insight into which linguistic or corporeal features make it beneficial to consider 2-grams in English-Nepali but slightly detrimental in English-Sinhala.

Table 1: Results for English-Nepali. Line counts are in thousands and (English) word counts in millions. The two bolded rows are the official AFRL submissions.

Type	Lines selected	Words selected	1M SMT BLEU	5M SMT BLEU
quality	N/A	N/A	2.91	4.26
coverage	50	1.4	1.79	4.17
1-gram	**50**	**1.0**	**3.64**	**4.14**
2-gram	50	1.1	3.88	4.21
3-gram	50	1.2	3.84	4.17
4-gram	50	1.2	3.78	4.23
1-gram	75	1.4	3.50	4.25
1-gram	100	1.9	3.47	4.12
coverage	150	3.8	1.24	3.84
1-gram	**150**	**3.1**	**3.55**	**4.33**
1-gram	225	4.8	3.53	4.12
Zipporah	N/A	N/A	3.40	4.22

Table 2: Results for English-Sinhala. Line counts are in thousands and (English) word counts in millions. The two bolded rows are the official AFRL submissions.

Type	Lines selected	Words selected	1M SMT BLEU	5M SMT BLEU
quality	N/A	N/A	3.26	5.07
coverage	50	1.4	1.98	5.17
1-gram	**50**	**0.8**	**4.31**	**5.16**
2-gram	50	1.0	4.26	5.15
3-gram	50	1.0	4.22	4.98
4-gram	50	1.1	4.30	5.04
1-gram	75	1.2	4.54	5.21
1-gram	100	1.6	4.49	5.19
coverage	150	4.0	1.40	3.43
1-gram	**150**	**2.6**	**4.62**	**5.09**
1-gram	225	4.3	4.57	4.91
Zipporah	N/A	N/A	4.16	4.77

We expect the optimal choices in our method to vary significantly with language pairs and noisy corpora. This might be in parameters (language ID thresholds, τ, n-gram levels, etc.) or the combination of coverage and metric metrics (product, sum, etc.), the design of the MT system(s) used for the text quality metric (e.g., phrase-based or neural, with their myriad design choices) or the text quality metric itself (Meteor, BEER (Stanojević and Sima'an, 2015), chrF (Popović, 2015), etc.).

Building a machine translation system in each direction would provide us with two text quality metric scores to incorporate into the overall score. We expect this would decrease dependence on the language ID thresholds and produce a somewhat better subcorpus.

8 Conclusions

We have presented the techniques we used in our submissions to the WMT19 Parallel Corpus Filtering For Low-Resource Conditions Task. Numerical results show our method to be a fraction of a BLEU point better than the Zipporah baseline for training the SMT system.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 12 Jun 2019. Originator reference number RH-19-119920. Case number 88ABW-2019-2964.

References

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pages 1045–1054, Portland, Oregon.

Grant Erdmann and Jeremy Gwinnup. 2018. Coverage and cynicism: The AFRL submission to the WMT 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 872–876, Belgium, Brussels. Association for Computational Linguistics.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Katherine Young, Michaeel Kazi, Elizabeth Salesky, and Brian Thompson. 2016. The AFRL-MITLL WMT16 news-translation task systems. In *Proceedings of the First Conference on Machine Translation*, pages 296–302, Berlin, Germany. Association for Computational Linguistics.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, Katherine Young, Christina May, Michaeel Kazi, Elizabeth Salesky, and Brian Thompson. 2015. The AFRL-MITLL WMT15 system: There's more than one way to decode it! In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 112–119, Lisbon, Portugal. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401, Lisbon, Portugal. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.