Dual Monolingual Cross-Entropy-Delta Filtering of Noisy Parallel Data

Amittai Axelrod
DiDi AI Labs
Los Angeles, CA
amittai@didiglobal.com

Anish Kumar

Steve Sloto

Abstract

We introduce a purely monolingual approach to filtering for parallel data from a noisy corpus in a low-resource scenario. Our work is inspired by Junczys-Dowmunt (2018), but we relax the requirements to allow for cases where no parallel data is available. Our primary contribution is a dual monolingual cross-entropy delta criterion modified from Cynical data selection (Axelrod, 2017), and is competitive (within 1.8 BLEU) with the best bilingual filtering method when used to train SMT systems. Our approach is featherweight, and runs end-to-end on a standard laptop in three hours.

1 Introduction

The 2018 WMT shared task on parallel corpus filtering (Koehn et al., 2018) required participants to select subcorpora of 10M and 100M words from an extremely noisy 1B word German-English parallel corpus from Paracrawl (Buck and Koehn, 2016). These subcorpora were then used to train machine translation systems, and evaluated on held-out test sets. The best submission (Junczys-Dowmunt, 2018) comprised:

1. a filter based on language ID
2. a dual conditional cross-entropy filter to determine whether the halves of a sentence pair were of roughly equal translation probability
3. a cross-entropy difference filter to prioritize in-domain sentence pairs

The 2019 WMT shared task on parallel corpus filtering (Koehn et al., 2019) was set for low-resource conditions, with the goal of translating Wikipedia texts both Sinhala-to-English and Nepali-to-English (Guzmán et al., 2019).

We participated only in the Sinhala-English track, basing our system on that of Junczys-Dowmunt (2018) but extensively modified for the 2019 low-resource scenario. As compared to their work, ours comprised: a minor upgrade of their first element, a relaxation of the second, a modern replacement for the third, and an additional length-based filter. The resulting entirely monolingual pipeline to filter noisy parallel data proved to be competitive with the other multilingual entries when used to train downstream SMT systems.

2 Related Work

We now describe the Junczys-Dowmunt (2018) system that was the inspiration for ours.

2.1 2018 Language ID Filter

The first feature used the `langid` Python module to classify the language of each half of each sentence pair to a language. Any sentence pair where either half was classified as being in an incorrect language was removed, and sentence pairs with correctly-classified halves were kept.

2.2 2018 Dual Conditional Cross-Entropy

The dual conditional cross-entropy filtering method rewards sentence pairs with minimal *symmetric translation disagreement*. That is the difference in average (per-word) conditional cross-entropy of the sentence pair halves:

$$|H_{F\to E}(s_E|s_F) - H_{E\to F}(s_F|s_E)|$$

For a sentence pair (s_E, s_F), the per-word conditional cross-entropy $H_{F\to E}(s_E|s_F)$ of one half of the sentence pair is computed by a translation model $F \to E$, and the corresponding $H_{E\to F}(s_F|s_E)$ of the other half of the sentence pair is computed by a translation model in the opposite direction. The two translation models are trained in inverse directions on the same parallel corpus, so they should be equally expressive.

However, the difference in translation scores does not take into account whether the scores

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 247–253
Florence, Italy, August 1-2, 2019.

are good or not. A perfectly translated sentence pair where the translation models agree perfectly would have the same score as a poorly translated sentence pair where the translation models also agree. This same weakness is found in the cross-entropy difference criterion (Moore and Lewis, 2010) on which the conditional cross-entropy difference is based. To force the better sentence pair to have a lower feature score than the other pair, Junczys-Dowmunt (2018) add a term consisting of the average per-word conditional cross-entropy of the two halves. Worse sentences have higher entropy, so a score of 0 remains ideal. The equation for the dual conditional cross-entropy is thus:

$$h(s_E, s_F) = \left| \underset{F \to E}{H}(s_E|s_F) - \underset{E \to F}{H}(s_F|s_E) \right| + \frac{1}{2}\left(\underset{F \to E}{H}(s_E|s_F) + \underset{E \to F}{H}(s_F|s_E) \right) \quad (1)$$

The first term is the translation disagreement, and the second term is the average entropy. The score is exponentiated so that good sentence pairs have a feature score of 1, and bad sentence pairs have a score of 0:

$$f(s_E, s_F) = e^{-h(s_E, s_F)}$$

In describing their approach, Junczys-Dowmunt (2018) criticize the Moore and Lewis (2010) cross-entropy difference method for "missing" an adequacy component. This is misguided, as the Moore-Lewis method was originally designed for language modeling and was only later repurposed for machine translation. In MT, the two halves of a sentence pair might be fluent but not express the same thing, and so the notion of *adequacy* is used to describe how well the halves correspond in meaning. In language modeling, there is no such thing as a sentence pair, and there should not be much doubt that a sentence rather adequately (and tautologically) manages to express exactly that which it *does* express. It would be more proper to state that the omission of adequacy is a weakness of the bilingual extension of Moore-Lewis to machine translation by Axelrod et al. (2011).

2.3 2018 Moore-Lewis Filtering

The third and final feature in the best 2018 system was a monolingual (English) cross-entropy difference (Moore and Lewis, 2010) score:

$$H_{in}(s_E) - H_{out}(s_E) \quad (2)$$

The cross-entropies H were computed using language models trained on 1M sentences of WMT news data as in-domain, and 1M random Paracrawl sentences as out-of-domain data. This is an ideal setup for cross-entropy difference, as Equation 2 fundamentally assumes that the two corpora are as different as possible.

3 Cynical Data Selection

Both the relaxation of the dual conditional cross-entropy filter and our replacement of the cross-entropy difference filter are based on Cynical data selection (Axelrod, 2017), described below. The Moore-Lewis cross-entropy difference approach fundamentally views the training data as being either *in*-domain or *out/general*-domain. This stark distinction is not realistic. Cynical data selection relaxes that assumption, and starts with one corpus of *representative* data (REPR), and one of *available* data (AVAIL). The representative data is exactly that: representative of what we would like to be translating. The *available* data is similarly the data pool from which one can select a subcorpus. No relationship is assumed between the *representative* and *available* corpora, nor between the domains they cover.

The algorithm incrementally grows a corpus of sentences, selecting from AVAIL, in order to better model REPR. First, it estimates the perplexity of a language model trained on the already-selected data and evaluated on the REPR corpus. Next, for each sentence still available, it estimates the change in that perplexity (or entropy, ΔH) that would result from adding it as a new sentence to the LM training data and re-training the LM (Sethy et al., 2006). The sentence with the lowest cross-entropy delta is removed from AVAIL, added to the selected pile, and the process repeats. Identifying the next single sentence to add is $O(n^2)$ and not computationally practical, but it is efficient to find the best word v in the vocabulary V_{REPR} to add once to the selected data. From there, it is now practical to pick the best sentence still in AVAIL that contains that word. The $n + 1^{th}$ iteration, after selecting n sentences, is:

1. Find the single word $v \in V_{repr}$ that would most lower the entropy (evaluated on REPR) of a language model, trained on the n already-selected sentences plus the one-word sentence "v".

2. Find the single sentence $s \in$ AVAIL containing v that would (also) most lower the entropy (evaluated on REPR) of a language model trained on the n sentences plus s.

3. Remove s from AVAIL, update the language model with the count c of all words in s, and add s to the selected sentences.

The cross-entropy delta ΔH is the change in the entropy of a language model, evaluated on a constant test set, after adding a new entry to the language model's training corpus. This is straightforward to compute, as there is a entropic length penalty for increasing the size of the training corpus, and an entropy gain for adding new information to the training set. This was first formulated by Sethy et al. (2006) as "relative entropy", and clarified by Axelrod (2017) as:

$$\begin{aligned}\underset{n\to n+1}{\Delta H} &= H_{n+1} - H_n \\ \underset{n\to n+1}{\Delta H} &= \underbrace{\log \frac{W_n + w_{n+1}}{W_n}}_{Penalty} \\ &\quad + \underbrace{\sum_{v \in V_{\text{REPR}}} \frac{C_{\text{REPR}}(v)}{W_{\text{REPR}}} \log \frac{C_n(v)}{C_n(v) + c_{n+1}(v)}}_{Gain}\end{aligned} \tag{3}$$

The penalty term depends on the length w_{n+1} of the $n + 1^{th}$ line, and the size in words W_n of the already-selected data. The gain term depends on the empirical probability of each word v in the REPR corpus, and then the count C_n of the word so far in the n selected lines, and the count c_{n+1} of the word in the $n + 1^{th}$ line.

4 Sinhala-English Data

The 2019 iteration of the shared task focused exclusively on filtering a noisy parallel corpus for low-resource language pairs, and had considerably less data than the 2018 German-English task. Table 1 shows that only 645k lines of parallel Sinhala-English were provided in total– less than the small 1M German-English sentence pair subsets used to train the dual NMT engines for the scoring function of Junczys-Dowmunt (2018).

4.1 Data

The 2019 Si-En parallel data was drawn from conversations and technical manuals, unlike the wiki-based evaluation data. Larger and more relevant, yet monolingual, corpora were provided from both Wikipedia and Common Crawl, detailed in Table 2.

Corpus	Lines	Tok (Si)	Tok (En)
Open Subtitles	601,164	3.4M	3.6M
Ubuntu	45,617	175k	151k
Total	646,781	3.5M	3.7M

Table 1: Parallel Data for Sinhala-English

Corpus	Lines	Tokens
Sinhala Wikipedia	156k	4.7M
English Wikipedia	67.8M	1.9B
Sinhala Common Crawl	5.2M	110M
English Common Crawl	380M	8.9B
English Subset Wikipedia	150k	5.5M
English Subset Common Crawl	6M	123M

Table 2: Corpus statistics for provided monolingual data in Sinhala and English, and an English subset of comparable size to the Sinhala data.

The provided monolingual English data was several orders of magnitude larger than the Sinhala data, which would have made it difficult to create equally strong (or weak) monolingual models used in this work. We therefore assembled a monolingual English corpus comparable in size and content to the Sinhala one by randomly selecting 150k lines from Wikipedia and 6M lines from Common Crawl. We used SentencePiece (Kudo and Richardson, 2018), with `model_type=word`, to preprocess the Sinhala and English sides separately, producing a fairly word-like vocabulary of 100k subwords for each language. Each SentencePiece model was trained on 1M lines of monolingual data: 150k Wiki + 850k Common Crawl.

5 Our Submission

We used the feature framework from Junczys-Dowmunt (2018) as the basis for ours. For each sentence pair (s_{Si}, s_{En}) in the noisy corpus, we computed a final score $f(s_{Si}, s_{En}) \in [0, 1]$ by multiplying each of the individual feature scores for the sentence pair:

$$f(s_{Si}, s_{En}) = \prod_i f_i(s_{Si}, s_{En}) \tag{4}$$

The feature scores, and therefore the final score, all had the same range of $[0, 1]$. For evaluation,

the lines were sorted from highest to lowest overall score, and then selected in that order until the number of selected English words reached the evaluation threshold. Any feature score being 0 effectively removed that sentence pair from consideration. The selected subsets were then submitted for evaluation by the task organizers. The following are the monolingual features we used to score the noisy parallel data.

5.0 Length Ratio Feature

We added one feature as compared to Junczys-Dowmunt (2018), based on the length ratio of the two halves of the sentence pair, penalizing sentence pairs with sides of disparate lengths. The provided clean, parallel, training data in Table 1 is inconclusive regarding the expected Si-to-En length ratio, as one of the parallel corpora had more English tokens than Sinhala, and the other had the reverse. The ratios were approximately inverses, so we set the desired ratio to be 1 and penalized sentence pair scores according to how divergent the parallel segment's length ratio was from 1. A sentence pair with a length ratio within two orders of magnitude, *i.e.* $e^{-2} < \frac{si}{en} < e^{2}$, or $|ln(\frac{si}{en})| < 2$, received a feature score of 1, or no penalty. The feature score was set to 0.5 if $2 < |ln(\frac{si}{en})| < 3$. For $3 < |ln(\frac{si}{en})|$ the feature was 0.35. For pairs where both segments contained fewer than six tokens, we applied less strict penalties as ratios are more likely to vary with shorter segment lengths. For such pairs, we assigned a score of 0.5 if the ratio is greater than 4 orders of magnitude, 0.75 if between 3 and 4, and 0.9 if within 2-3 factors. We also observed that large numbers of non-parallel Paracrawl sentence pairs contained mostly numbers on one side. Any sentence pair where at least 15% of either half was only numerals received a score of 0.

5.1 Language ID Feature

As with the 2018 task, a considerable quantity of the provided Paracrawl data was not in the correct language. Following Junczys-Dowmunt (2018), we classified the halves of the sentence pair using the `langid` Python module and assigned 0 to any sentence pair with an incorrectly-labeled half. If the correct languages were selected, then the feature value was the product of the `langid` confidence scores. Inspecting the filter output showed that it was not strong enough. The `langid` classification had many false positives, as well as source-side (Sinhala) sentences that were mixed with a significant amount of English. The shared task's hard limit on the number of selectable words made it important to minimize the amount of English on the Sinhala side. The languages have non-overlapping writing scripts, so it was easy to detect erroneous characters. We therefore multiplied the lang_id score by the proportion of characters (excluding numerals and punctuation) in each sentence that belong to the correct Unicode block, resulting in an overall language ID feature that slightly extends the original.

5.2 Dual Monolingual Cross-Entropy Deltas

Junczys-Dowmunt (2018) trained MT systems on clean parallel data for the 2018 task, but used only the translation probability of each to score the Paracrawl data and not the translation output itself. The motivation for training the dual NMT systems on the same parallel corpus was to ensure that the models would have similar BLEU scores and translation probabilities for the halves of a truly parallel sentence pair.

We did not have enough good parallel data for Sinhala and English, which ruled out training models on identical information. However, perhaps the translation models themselves were not inherently necessary as long as similar scores could be obtained. Language models require less training data than an MT engine to be reliable, and can also output an average per-word probability for a sentence– and we were provided with good monolingual data. We set out to construct language models with similar *amounts* of information hoping they might have similar perplexities for the halves of a parallel sentence pair, and different perplexities for a non-parallel pair. The result was a relaxation of the dual conditional translation cross-entropy feature that only required monolingual data, and used equal relative informativeness instead of equal translation probability.

5.2.1 Setting Up Language Models

N-gram language models in different languages are not comparable. Differences in morphology can lead to significant differences in word counts, data sparsity, and thus how well a fixed model architecture can represent the language. Instead of multilingual word embeddings using sparse data (Artetxe and Schwenk, 2019), we simply used SentencePiece to force the Sinhala and English corpora to have the same size vocabulary (100k

subwords). First, we hoped a standard lexicon size would mitigate the effect of morphology differences affecting sentence length and word sparsity. Secondly, we hoped it would encourage language models trained on similar– but not parallel– English and Sinhala texts to have similar perplexities over each half of a parallel test set.

This would mean the two LMs had similar estimates of how much information is in each half of the parallel data. The two halves of the parallel corpus presumably contain the same amount of actual information, but two LMs would only come up with the same estimate if they themselves contained comparable amounts of information, even if they did not the same information.

To test this, we trained 4-gram language models using KenLM (Heafield, 2011) on the Sinhala monolingual data and the restricted-size English data in Table 2, both unprocessed and after tokenizing with SentencePiece. The models were evaluated on the provided parallel `valid` and `test` sets. Table 3 shows that, indeed, forcing English and Sinhala LMs to have identical vocabulary sizes was enough to obtain nearly identical perplexities on the halves of a parallel corpus, even though the language models were trained on similar but not parallel data.

Corpus	`valid`	`test`
Sinhala, untok	1,759.8	1,565.9
English, untok	1,069.2	985.3
Sinhala, tok=SP	**320.5**	**299.2**
English, tok=SP	**302.5**	**292.7**

Table 3: Using SentencePiece to equalize LM perplexities in different languages on the dev sets.

5.2.2 Parallel Scoring with Monolingual LMs

Wse used the "greedy cross-entropy delta" term from Cynical data selection in a novel way: to score each side of the Paracrawl data as a memoryless stream of text. In this setup, we had a language model trained on the monolingual Wikipedia data, which is the REPR corpus, and representative of the kind of data the organizers will evaluate on. We compute the ΔH of adding each sentence s in Paracrawl to the REPR corpus, retraining a LM on REPR+s, and recomputing the perplexity on REPR corpus. After computing all of the ΔH scores for the Paracrawl data, cynical data selection would normally extract the best one, incorporate it into the training set, and iterate. Instead, we modified the public implementation[1] of Cynical data selection to not update anything, and the scoring is done in a single pass as done by Sethy et al. (2006).

The difference between a LM perplexity and the ΔH score is that the LM quantifies the likelihood, and the ΔH score quantifies informativeness. The ΔH score estimates, for a fixed REPR corpus: does this next line contain any information at all about REPR that we do not already know? A negative score would indicate a estimated decrease in entropy, so adding this line should improve a model trained on the selected data.

We constructed monolingual Sinhala and English LMs with similar perplexities on a parallel test set that resembled the task evaluation, so we hoped that sentences with equal ΔH scores according to these two models could be parallel. Or, at least, that sentences with disparate ΔH scores would be deemed not-parallel, and filtered out.

One could simply replace the translation system conditional cross-entropies in Equation 1 with the cross-entropies from the two comparable language models just described. However, that would only characterize the fluency, without any sense of the content. It is not clear whether identical perplexities or identical ΔH scores is a better indicator of "these sentences are parallel": being equally likely and being equally informative are each positive indicators. The goal of the shared task was to assemble the parallel corpus that produced the best downstream MT system for Wikipedia test; prioritizing informative sentences seemed more important here than prioritizing likely ones. Our version of Equation 1 thus used Equation 3's ΔH scores, dual monolingual cross-entropy deltas, for each sentence pair (s_{Si}, s_{En}), instead of dual bilingual conditional cross-entropies:

$$|\Delta H_{En}(s_{En}|\text{REPR}_{En}) - \Delta H_{Si}(s_{Si}|\text{REPR}_{Si})| + \frac{1}{2}\big(\Delta H_{En}(s_{En}|\text{REPR}_{En}) + \Delta H_{Si}(s_{Si}|\text{REPR}_{Si})\big) \quad (5)$$

This was exponentiated to be in the range of $[0, 1]$.

5.3 Dual Monolingual Cynical Data Selection

The final feature from (Junczys-Dowmunt, 2018) was a monolingual Moore-Lewis score, intended to bias the filtering towards in-domain news data.

[1] `github.com/amittai/cynical`

However, the Moore-Lewis method for data selection has some notable flaws, as described in Axelrod (2017). The biggest is that it has no sense of *sufficiency*: while it is not helpful to see an identical sentence pair 10,000 times, the Moore-Lewis criterion will assign the same score to all copies. Cynical data selection selects sentences only if they contribute new information to the set of sentences already selected, and has previously been shown to help domain adaptation in SMT (Santamaría and Axelrod, 2017) and NMT (Zhang et al., 2019), and a variation of it was used for the 2018 corpus filtering task (Erdmann and Gwinnup, 2018). As a side effect, Cynical selection eliminates the need for explicit vocabulary coverage features that were used in the previous shared task (Lo et al., 2018; Azpeitia et al., 2018).

For each language, we used Cynical data selection to rank the sentences in the noisy corpus. We set the Paracrawl data to be the *Available* set, and the clean monolingual Wikipedia data to be the *Representative* set. This selects the subset of Paracrawl that best models monolingual Wikipedia. The re-ranked Paracrawl corpus was then scored by converting the Cynical ranking to a percentage and subtracted from 1. Thus the sentence selected as number 15,000 out of 3M would have a score of $1 - \frac{15k}{3M}$, and 1 would be the best score and 0 the worst. The ranking score for a sentence pair was the product of the monolingual rank scores for each half.

6 Results and Discussion

Our submission was entirely monolingual, and used parallel data only to sanity-check the language models trained in Section 5.2.1. Furthermore, all of the preprocessing, language modeling, data selection, and feature computation in this work was run on a laptop. As such, we had no expectations for whether our method would be effective compared against bilingual or multilingual methods trained for days on GPU machines.

We tried to predict results using NMT systems after the submission deadline, thanks to the scripts, code, and standard settings provided by the organizers, but all of our system BLEU scores (Papineni et al., 2002) were under 0.20 and worse than a random baseline. While the evaluation campaign cutoff was set to be 1M or 5M English words, the Sinhala sides of our filtered corpus contained only 740k and 3.6M words respectively. Our length ratio feature was overly complicated and not aggressive enough; the Si→En NMT systems tended to stutter to produce English sentences of appropriate length. Discarding anything with a length difference $> 20\%$ would probably have been better.

The official evaluation results were a pleasant surprise. Table 4 shows the top and bottom scores for each evaluation category, providing context for our submission. We were in the bottom third of the SMT systems, yet within 1.8 BLEU of the best system at 1M, and 1.3 BLEU of the best system at 5M. This is rather competitive for a gratuitously-monolingual approach to a bilingual task!

Our submitted system, like roughly 30% of the submissions, was not suitable for filtering data for a low-resource NMT pipeline. However, the NMT systems trained on 1M words were several BLEU points better than systems trained on 5M words, so training an NMT system on small amounts of data is unpredictable. Better feature engineering would certainly help.

System	1M SMT	1M NMT	5M SMT	5M NMT
Rank 1	4.27	6.39	4.94	4.44
DiDi	2.53	0.19	3.70	0.20
Rank 10	0.92	0.03	2.73	0.10

Table 4: Bleu scores on `test` for systems trained on subsets with 1M and 5M English words of the noisy Paracrawl data.

7 Conclusion

We presented a purely monolingual method, based on cynical data selection (Axelrod, 2017), for filtering noisy parallel data. Our approach is a relaxation of the dual conditional cross-entropy method of Junczys-Dowmunt (2018), that does require any parallel data. As secondary contributions, we have used Cynical data selection in a streaming scenario for the first time, and used relative *informativeness* to judge the relationship between the halves of a sentence pair. While our method does not outperform most parallel approaches, it is competitive, and more suitable for scenarios with little or no parallel data. Furthermore, our work is also undemanding of computational resources, as it ran end-to-end on a single laptop in a couple hours, and should integrate well into a feature ensemble for real-world deployment.

References

Mikel Artetxe and Holger Schwenk. 2019. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. *ACL (Association for Computational Linguistics)*.

Amittai Axelrod. 2017. Cynical Selection of Language Model Training Data. *arXiv [cs.CL]*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation Via Pseudo In-Domain Data Selection. *EMNLP (Empirical Methods in Natural Language Processing)*.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez garcia. 2018. STACC, OOV Density and N-gram Saturation: Vicomtech's Participation in the WMT 2018 Shared Task on Parallel Corpus Filtering. *WMT Conference on Statistical Machine Translation*.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 Bilingual Document Alignment Shared Task. *WMT Conference on Statistical Machine Translation*.

Grant Erdmann and Jeremy Gwinnup. 2018. Coverage and Cynicism: The AFRL Submission to the WMT 2018 Parallel Corpus Filtering Task. *WMT Conference on Statistical Machine Translation*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two New Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English. *arXiv [cs.CL]*.

Kenneth Heafield. 2011. KenLM : Faster and Smaller Language Model Queries. *WMT (Workshop on Statistical Machine Translation)*.

Marcin Junczys-Dowmunt. 2018. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. *WMT Conference on Statistical Machine Translation*.

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. *WMT Conference on Statistical Machine Translation*.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel Forcada. 2018. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. *WMT Conference on Statistical Machine Translation*.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. *EMNLP (Empirical Methods in Natural Language Processing) System Demonstrations*.

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. *WMT Conference on Statistical Machine Translation*.

Robert C Moore and William D Lewis. 2010. Intelligent Selection of Language Model Training Data. *ACL (Association for Computational Linguistics)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *ACL (Association for Computational Linguistics)*.

Lucía Santamaría and Amittai Axelrod. 2017. Data Selection with Cluster-Based Language Difference Models and Cynical Selection. *IWSLT (International Workshop on Spoken Language Translation)*.

Abhinav Sethy, Panayiotis G. Georgiou, and Shrikanth Narayanan. 2006. Text Data Acquisition for Domain-Specific Language Models. *EMNLP (Empirical Methods in Natural Language Processing)*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum Learning for Domain Adaptation in Neural Machine Translation. *NAACL (North American Association for Computational Linguistics)*.