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

Lo, Chi-Kiu; Simard, Michel; Stewart, Darlene; Larkin, Samuel; Goutte, Cyril; Littell, Patrick

Publisher's version / Version de l'éditeur:
https://doi.org/10.18653/v1/W18-6482
Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 2, pp. 921-929, 2018-11-01

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=6a17ac14-d76b-4b32-9343-93b03c77ca0d
https://publications-cnrc.canada.ca/fra/voir/objet/?id=6a17ac14-d76b-4b32-9343-93b03c77ca0d

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

Chi-kiu Lo **Michel Simard** **Darlene Stewart**
Samuel Larkin **Cyril Goutte** **Patrick Littell**
NRC-CNRC
Multilingual Text Processing
National Research Council Canada
1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
Firstname.Lastname@nrc-cnrc.gc.ca

Abstract

We present our semantic textual similarity approach in filtering a noisy web crawled parallel corpus using YiSi—a novel semantic machine translation evaluation metric. The systems mainly based on this supervised approach perform well in the WMT18 Parallel Corpus Filtering shared task (4th place in 100-million-word evaluation, 8th place in 10-million-word evaluation, and 6th place overall, out of 48 submissions). In fact, our best performing system—`NRC-yisi-bicov` is one of the only four submissions ranked top 10 in both evaluations. Our submitted systems also include some initial filtering steps for scaling down the size of the test corpus and a final redundancy removal step for better semantic and token coverage of the filtered corpus. In this paper, we also describe our unsuccessful attempt in automatically synthesizing a noisy parallel development corpus for tuning the weights to combine different parallelism and fluency features.

1 Introduction

The WMT18 shared task on parallel corpus filtering (Koehn et al., 2018b) challenged teams to find clean sentence pairs from ParaCrawl, a humongous high-recall, low-precision web crawled parallel corpus (Koehn et al., 2018a), for training machine translation (MT) systems. Data cleanliness 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 at different levels. Intuitively, the crosslingual semantic textual similarity of the sentence pairs in the corpora is one of the most important factors affecting the parallelism of the target sentence pairs. Lo et al. (2016) scored crosslingual semantic textual similarity crosslingually, using a semantic MT quality estimation metric with fewer resource requirements, or monolingually, using a pipeline of MT system and semantic MT evaluation metric with better performance. The core of the National Research Council of Canada (NRC) supervised submissions (`NRC-yisi-bicov` and `NRC-yisi`) of the parallel corpus filtering shared task were developed in the same philosophy using a new semantic MT evaluation metric, YiSi (Lo, 2018).

The participants of the parallel corpus filtering shared task were given a large set of "clean" German-English monolingual and bilingual training corpora for the WMT18 news translation shared task (except a filtered version of ParaCrawl) and tasked to score the cleanliness of each sentence pair in the "dirty" ParaCrawl corpus. Our supervised submissions used the given parallel data to train an MT system to translate the German side of the dirty corpus into English. The provided version of the dirty ParaCrawl corpus contains raw data crawled from the web with minimal de-duplication processing only, and includes non-parallel, or even non-linguistic data. It contains 104 million German-English sentence pairs, with 1 billion English tokens and 964 million German tokens before punctuation tokenization. A 10-million-word (10M-word) and a 100-million-word (100M-word) corpus sub-selected by the participating cleanliness scoring system were used to train statistical machine translation (SMT) and neural machine translation (NMT) systems. The success of the participating scoring systems was determined by the quality of the MT output from the four MT systems as measured by BLEU (Papineni et al., 2002) on some in-domain and out-of-domain evaluation sets.

In this paper, we describe the efforts in developing our supervised submissions: the initial fil-

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 921–929

https://doi.org/10.18653/v1/W18-6482

tering steps for scaling down the size of the given ParaCrawl dirty corpus, the wide range of features experimented for measuring parallelism, fluency and grammaticality, the failed attempt to combine useful features and the final redundancy removal for improving token coverage of the filtered corpus. Despite the simple single-feature architecture used in the NRC best-performing supervised submission (`NRC-yisi-bicov`), it performed well in the MT quality evaluation compared to other participants. It ranked 4th in the 100-million-word evaluation, 8th in the 10-million-word evaluation and 6th overall among 48 submissions. It is one of the only four submissions ranked top 10 in both evaluations.

2 System architecture

There are a wide range of factors constituting a good parallel sentence pair for training MT systems. Some of the more important factors for a good general MT system parallel training corpus include:

- High parallelism in the sentence pairs
- High fluency and grammaticality, especially for sentences in the output language
- High token coverage, especially in the input language
- High variety of sentence lengths

The NRC supervised and unsupervised submissions shared the same general skeleton for the system architecture. The systems consisted of: initial filtering to remove obvious noise and to prevent selections constituted of a large collection of short sentences; feature scoring for measuring parallelism, fluency and grammaticality; feature combination (only in the `NRC-mono` and `NRC-mono-bicov` submissions); and final redundancy removal (only in the `NRC-*-bicov` submissions) to improve token coverage.

2.1 Initial filtering

Although the given "dirty" corpus had already been deduplicated, we did an additional deduplication step in which email and web addresses were replaced with a placeholder token, before deciding which sentences were duplicates. Sentence pairs were filtered out if the pair was seen before or if the input side was exactly the same as the output side.

We also observed that many sentences in the corpus, although parallel, were rather similar and uninformative, especially numerical data such as long lists of page numbers or dates. We observed that using measurements that preferred such sentences resulted in comparatively poor MT performance, likely because the MT systems did not get enough varied data. To mitigate this, we ran two additional filtering steps regarding numbers. First, over 50% of the numbers on each side of the sentence pair had to have a match, otherwise it was filtered out as a bad translation. Next, we removed all the numbers and punctuation and, similar to the previous deduplication step, filtered out sentence pairs if their non-number parts had been seen before, or if the non-number input side was exactly the same as the non-number output side.

A common error found in web crawled corpora is sentences that are in the wrong language. We therefore ran the `pyCLD2` language detector[1] on each side of the sentence pair and filtered out pairs whose input side was non-German with a confidence score over 0.5, or whose output side was non-English with a confidence score over 0.5.

Our final filtering step was to remove unreasonably long sentences. Another common error in web crawled corpora is that they contain non-linguistic data, such as tables or computer code. We therefore punctuation-tokenized both sides of the sentence pairs and removed the pair if either side was more than 150 tokens.

The above mentioned steps removed obvious and uninteresting noise and significantly scaled down the size of the original ParaCrawl corpus for more resource demanding feature scoring. The corpus was scaled down from 104 million sentence pairs originally to 28 million sentence pairs.

2.2 Feature scoring

We experimented with a large collection of feature models to address the factors for good general MT training data mentioned at the beginning of this section. Below is a selected list of features that performed reasonably well in our internal sanity check.

2.2.1 Parallelism

YiSi-1: monolingual semantic MT evaluation metric We first used the "clean" WMT18 news translation task monolingual and parallel training data (tokenized and lowercased) to train an

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

SMT system using Portage (Larkin et al., 2010), a conventional log-linear phrase-based SMT system. The translation model of the SMT system uses IBM4 word alignments (Brown et al., 1993) with grow-diag-final-and phrase extraction heuristics (Koehn et al., 2003). The system has two n-gram language models: a 5-gram mixture language model (LM) trained on the four corpora components using SRILM (Stolcke, 2002), and a pruned 6-gram LM trained on the WMT monolingual English training corpus built using KenLM (Heafield, 2011). The SMT system also includes a hierachical distortion model, a sparse feature model consisting of the standard sparse features proposed in Hopkins and May (2011) and sparse hierarchical distortion model features proposed in Cherry (2013), and a neural network joint model, or NNJM, with 3 words of target context and 11 words of source context, effectively a 15-gram LM (Vaswani et al., 2013; Devlin et al., 2014). The parameters of the log-linear model were tuned by optimizing BLEU on the development set (newstest2017) using the batch variant of margin infused relaxed algorithm (MIRA) by Cherry and Foster (2012). Decoding uses the cube-pruning algorithm of Huang and Chiang (2007) with a 7-word distortion limit. We then translated the German side of the filtered ParaCrawl into English.

We also used the monolingual English data to train word embeddings using `word2vec` (Mikolov et al., 2013) for evaluating monolingual lexical semantic similarity.

YiSi is new a semantic MT evaluation metric inspired by MEANT 2.0 (Lo, 2017). YiSi-1 is equivalent to MEANT 2.0-nosrl. It measures the segmental semantic similarity. The segmental semantic precision and recall divide the inverse-document-frequency weighted sum of the n-gram lexical semantic similarity of the MT output and the English sentence of the target pair by the weighted count of n-grams in the MT output and the English sentences, respectively. In this work, we set the n-gram size to two. Precisely, YiSi-1 is computed as follows:

$$
\begin{aligned}
w(e) &= \text{inverse document freq. of token } e \\
w(\overrightarrow{e}) &= \sum_k w(e_k) \\
v(e) &= \text{word embedding of token } e \\
s(e,f) &= cos(v(e), v(f))
\end{aligned}
$$

$$
s_p(\overrightarrow{e},\overrightarrow{f}) = \frac{\sum_a w\left(\overrightarrow{e_{a..a+n-1}}\right) \cdot \max_b \frac{\sum_{k=0}^{n-1} w(e_{a+k}) \cdot s(e_{a+k}, f_{b+k})}{n-1}}{\sum_a w\left(\overrightarrow{e_{a..a+n-1}}\right)}
$$

$$
s_r(\overrightarrow{e},\overrightarrow{f}) = \frac{\sum_b w\left(\overrightarrow{f_{b..b+n-1}}\right) \cdot \max_a \frac{\sum_{k=0}^{n-1} w(f_{b+k}) \cdot s(e_{a+k}, f_{b+k})}{n-1}}{\sum_b w\left(\overrightarrow{f_{b..b+n-1}}\right)}
$$

$$
\begin{aligned}
\text{precision} &= s_p(\overrightarrow{e_{\text{sent}}}, \overrightarrow{f_{\text{sent}}}) \\
\text{recall} &= s_r(\overrightarrow{e_{\text{sent}}}, \overrightarrow{f_{\text{sent}}}) \\
\text{YiSi-1} &= \frac{\text{precision} \cdot \text{recall}}{\alpha \cdot \text{precision} + (1-\alpha) \cdot \text{recall}}
\end{aligned}
$$

YiSi-1_srl measures the semantic similarity with additional frame semantic or semantic role labeling (srl) information. It uses a more principle way to compute the precision and recall of semantic similarity between the translation output and the reference when comparing to MEANT 2.0. Instead of aggregating the precision and recall at the segmental semantic similarity level, YiSi-1_srl precision is the weighted sum of the segmental semantic precision and the frame semantic precision and similarly, for YiSi-1_srl recall. The frame semantic precision is the weighted sum of the segmental semantic precision of the semantic role fillers according to the shallow semantic structure parsed by the `mateplus` (Roth and Woodsend, 2014) English semantic parser over the weighted counts of roles and frames according to the shallow semantic structure of the MT output and similarly, for the frame semantic recall. Precisely, YiSi-1_srl is computed as follows:

$$
\begin{aligned}
q^0_{i,j} &= \text{ARG } j \text{ of aligned frame } i \text{ in MT} \\
q^1_{i,j} &= \text{ARG } j \text{ of aligned frame } i \text{ in REF} \\
w^0_i &= \frac{\text{\#tokens filled in aligned frame } i \text{ of MT}}{\text{total \#tokens in MT}} \\
w^1_i &= \frac{\text{\#tokens filled in aligned frame } i \text{ of REF}}{\text{total \#tokens in REF}} \\
w_j &= \text{count (ARG } j \text{ in REF)} \\
w_t &= 0.25 * \text{count (predicate in REF)} \\
\text{srl}_p &= \frac{\sum_i w^0_i \frac{w_t s_p(\overrightarrow{e_{i,t}}, \overrightarrow{f_{i,t}}) + \sum_j w_j s_p(\overrightarrow{e_{i,j}}, \overrightarrow{f_{i,j}})}{w_t + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i} \\
\text{srl}_r &= \frac{\sum_i w^1_i \frac{w_t s_r(\overrightarrow{e_{i,t}}, \overrightarrow{f_{i,t}}) + \sum_j w_j s_r(\overrightarrow{e_{i,j}}, \overrightarrow{f_{i,j}})}{w_t + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i}
\end{aligned}
$$

$$
\begin{aligned}
\text{precision} &= \beta \cdot \text{srl}_p + (1-\beta) \cdot s_p(\overrightarrow{e_{\text{sent}}}, \overrightarrow{f_{\text{sent}}}) \\
\text{recall} &= \beta \cdot \text{srl}_r + (1-\beta) \cdot s_r(\overrightarrow{e_{\text{sent}}}, \overrightarrow{f_{\text{sent}}})
\end{aligned}
$$

$$\text{YiSi-1\_srl} = \frac{\text{precision} \cdot \text{recall}}{\alpha \cdot \text{precision} + (1-\alpha) \cdot \text{recall}}$$

When we evaluate MT output in practice, YiSi score is a weighted harmonic mean of the precision and recall. However, in this work, we segregated the precision and recall of YiSi into separate features as we planned to let the regression decide suitable weights to combine them. Further details of YiSi are provided in Lo (2018).

YiSi-2: crosslingual semantic MT evaluation metric For the crosslingual version of YiSi, YiSi-2, instead of training a German-English MT system, we used the "clean" WMT18 news translation task parallel training data to train bilingual word embeddings using `bivec` (Luong et al., 2015) for evaluating crosslingual lexical semantic similarity.

Similar to YiSi-1, YiSi-2 precision and recall are the weighted sum of the crosslingual lexical semantic similarity of the sentence pairs over the weighted count of tokens in the German and English sentences respectively. In this work, we set the n-gram size to one.

YiSi-2_srl precision and recall are the weighted sum of the crosslingual lexical semantic similarity according to the shallow semantic structure parsed by `mateplus` German and English semantic parser over the weighted counts of roles and frames according to the shallow semantic structure of the German and the English sentence, respectively. We also segregated the precision and recall of YiSi-2 and YiSi-2_srl into separate features for the same reason mentioned above.

Alignment scores The SMT model trained on the "clean" WMT18 news translation task parallel training data for YiSi score computation include several alignment models as components, from which probabilities $p(d|e)$ and $p(e|d)$ were computed. We find the hidden markov model (HMM) alignment models (Vogel et al., 1996) are reliably useful for scoring parallelism of the sentence pairs in the target corpus.

Perplexity ratio of input sentences and output sentences The perplexity ratio reflects the different amounts of information contained in each side of the sentence pairs. This is computed by dividing the smaller perplexity score of the two sentences in the target pair by the larger one. Thus, the ratio ranged from 0 to 1, where a larger value represents better parallelism.

Perplexity ratio of the part-of-speech (POS) tags of the input sentences and output sentences Similar to the previous feature, the perplexity ratio of the input and output sentences POS tags is computed by dividing the smaller POS perplexity score of the two sentences in the target pair by the larger one.

Distance of sentence vectors Sentence vectors were trained using `sent2vec` (Pagliardini et al., 2018) on each side of the "clean" parallel WMT18 news translation task parallel training data. Further details on how to compute these features are described in Littell et al. (2018).

2.2.2 Fluency and grammaticality

Perplexity 6-gram LMs of the input and output languages were built using KenLM (Heafield, 2011) on the WMT18 news translation task German (263 million sentences) and English (303 million sentences) monolingual corpora.

Perplexity of POS tags We parsed the German and the English monolingual training data using `mateplus` and built 6-gram LMs based on the POS tags using KenLM.

2.3 Feature combination

2.3.1 Synthetic noisy data generation

We used the WMT09-13 test sets (Callison-Burch et al., 2009, 2010, 2011, 2012; Bojar et al., 2013) as the basis of our development set, as we believe that all the test sets in the previous years are clean and highly parallel, as opposed to the "clean" training data where glitches may occur (especially in the Europarl and CommonCrawl corpora). We introduced several types of synthetic errors into the development set as negative examples and assigned scores according to the severity of each error.

We added the output from the best and the worst participating systems in each year as the mostly parallel but less fluent sentence pairs. We also constructed error sentence pairs by offsetting or deleting tokens on either side, or introducing tokens in the wrong language. The target scores of these pairs are proportional to the percentage of tokens offset, deleted or introduced. Lastly, misaligned sentence pairs were added as fluent but non-parallel negative examples. The resulting development set had 11k sentence pairs of positive and synthetic negative examples.

2.3.2 Regression

In order to benefit from multiple features, we first experimented with linear feature combination. Using the scores generated in §2.2 as features, and the data described in the previous section as modeling data, we trained a linear model with L_1 regularization. The amount of regularization was set by optimizing a 10-fold cross-validation estimator of the generalization error on the modeling data. On the synthetic data, it turns out that the optimal level of regularization is minimal, suggesting the overfitting is minimal with this amount of data. We also tried building a linear combination of a subset of the most relevant features, selected from the results of the regularized model built on the full set of features (essentially removing features for which combination weights were not significantly different from zero). The linear features combination models yield marginal improvements according to the cross-validation estimator built from the synthetic data. However, there was no gain in precision when evaluated on our small annotated set or in MT quality when training MT system using data sub-selected by the combined model, so we ended up not submitting the combined results.

2.4 Redundancy filtering

Our scoring mechanisms naturally tend to assign higher scores to semantically similar sentences without paying attention to their usefulness for MT. As a result, we observe much redundancy and a somewhat limited vocabulary coverage in the top-ranking sentences, such as numerous perfectly translated dateline. To compensate for this effect, we applied a form of redundancy filtering after scoring sentence pairs: going down the re-ranked corpus, we filtered out any sentence pair that did not contain at least one "new" source-language word bigram, i.e., a pair of consecutive source-language tokens not observed in previous pairs. This had the effect of excluding sentences that were too similar to one another. Because it was applied post-scoring on the re-ranked corpus, it tended to retain higher-scoring sentence pairs.

3 Experiments and results

3.1 Sanity check

We annotated about 300 random sentence pairs from the filtered target corpus, labeling 93 as correct translations and the rest as non-parallel. We did not tune any parameters to this set, since it was small and also doing so would violate the competition guidelines, but used it to sanity check our feature engineering. We computed the precision of each experimented feature by dividing the number of true positives in the top 93 pairs (scored by the feature) by 93.

features	precision
baselines	
random	0.312
hunalign	0.624
parallelism	
YiSi-1 precision	**0.796**
YiSi-1 recall	0.763
YiSi-1_srl (β=1) precision	0.559
YiSi-1_srl (β=1) recall	0.559
YiSi-2 precision	0.753
YiSi-2 recall	0.731
YiSi-2_srl (β=1) precision	0.441
YiSi-2_srl (β=1) recall	0.452
HMM $p(d\|e)$	0.753
HMM $p(e\|f)$	0.753
s2v d100 cosine	0.435
s2v d300 Mahalanobis	0.634
perplexity ratio	0.538
POS perplexity ratio	0.441
fluency and grammaticality	
German perplexity	0.419
English perplexity	0.355
German POS perplexity	0.376
English POS perplexity	0.462
feature combination	
regression	0.763

Table 1: Precision on the 300-annotated sentence pairs.

Table 1 shows the precision of the experimented features. We also include the results from a random scoring baseline and the given hunalign scores (Initial filtering was integrated into both baselines). YiSi-1 precision was the best performing feature with close to 80% true positive rate in its top ranking sentence pairs. In general, we can see that supervised parallelism features achieved over 73% precision. It is expected that the structural semantic options of YiSi were less accurate as standalone features due to the fact the score for a sentence pair would be zero when the shallow semantic parser failed to find a semantic frame on either side. Our original plan was to combine these features with other semantic features and bias the combined scores to prefer longer sentences with

	SMT				NMT			
	10M-word		100M-word		10M-word		100M-word	
system	dev.	test	dev.	test	dev.	test	dev.	test
random	17.52	20.28	22.06	26.88	19.58	24.06	27.27	34.63
HMM $p(e\|f)$	19.09	23.55	24.42	29.73	21.16	26.59	31.53	39.52
HMM $p(e\|f)$ bicov	20.42	25.31	24.68	29.98	23.17	29.08	31.98	39.66
YiSi-1 precision (`NRC-yisi`)	21.56	24.68	24.47	30.10	24.24	30.75	32.49	40.27
YiSi-1 precision bicov (`NRC-yisi-bicov`)	**22.19**	**27.41**	**24.84**	**30.46**	**26.69**	**33.56**	**33.20**	**40.98**
regression bicov	21.86	26.97	**24.84**	30.27	25.28	31.94	31.30	39.34

Table 2: BLEU scores of SMT and NMT systems trained on the 10M- and 100M-word corpora subselected by the scoring systems. "bicov" indicates that the final bigram coverage step (§2.4) was performed. The development set is newstest2017 and the test set is newstest2018.

semantic structure recognized by the parser. However, as we can see, the regression hurt the precision on the 300-annotated subset of data. This was the first hint that our feature combination was not a promising avenue.

3.2 MT quality check

We used the official software to extract the 10M-word and 100M-word corpora from the original ParaCrawl according to the feature scores. We then trained SMT and NMT systems using the extracted data. The SMT systems were trained using Portage with components and parameters similar to the German-English SMT system in Williams et al. (2016). The NMT systems were transformer models with self-attention (Vaswani et al., 2017) trained using Sockeye-1.18.20 (Hieber et al., 2017) with default parameter settings[2], except for the maximum sequence length, which was reduced to 60:60, and we also clip gradients to 1. We used newstest2017 and newstest2018 as the MT development and test set.

Table 2 shows the BLEU scores for MT systems trained on the ParaCrawl data subselected by our scoring features. We have also included the random scoring feature (with initial filtering) as a baseline. The MT quality trained on data subselected by the feature scores showed the same trend as the results of the sanity check. That is to say, a feature that performed better in the sanity check indeed was able to pick "cleaner" data to train better MT systems.

We noticed that the differences in BLEU of MT systems trained on the 100M-word corpus subselected by our features were very small. This shows that our supervised features were successful in identifying parallel data.

[2] https://github.com/awslabs/sockeye/blob/arxiv_1217/arxiv/code/transformer/sockeye/train-transformer.sh

In addition, the results on MT quality confirmed again that our feature combination was not performing as planned. Compared to the systems trained on data subselected by the best feature (YiSi-1 precision bicov), those trained on data subselected by the regression score list had their performance decreased by 0.2-0.5 BLEU on SMT and 1.6 BLEU on NMT.

Systems in which we applied redundancy removal are labeled "bicov". On the larger (100M words) selections, the redundancy removal had virtually no effect when applied after YiSi scoring. However, on the smaller (10M words) selection, it allowed for substantial BLEU score increases: +1.61 BLEU for SMT systems on average and +2.44 BLEU for NMT systems.

4 Official Results

Table 3 presents the results of the official BLEU scores on seven development and test sets (devtests) in four training conditions, the average scores across the seven devtests for each of the four training conditions, the average scores across all the devtests for the 10M-word and 100M-word training conditions and the average scores across all the test documents and all training conditions. Our best performing supervised submission—`NRC-yisi-bicov` ranked 4th in the 100M-word evaluation, 8th in the 10M-word evaluation and 6th overall, out of 48 submissions. In fact, it is one of the only four submissions ranked top 10 in all four training conditions.

Our supervised systems perform strongly on the 100M-word conditions with most of the results in the top 10 (among 48 submissions) and very small differences from the highest score of each test set. Similar to the results from our internal MT quality check, the performance differences of our supervised systems on the 100M-word conditions were very small. In other words, the redundancy re-

SMT

10M-word

	dev.	test						
domain	news	news	speech	laws	medical	news	IT	
system \ test set	newstest17	newstest18	iwslt17	Acquis	EMEA	Global Voices	KDE	average
highest scores	**23.23 (1)**	**29.59 (1)**	**22.16 (1)**	**21.45 (1)**	**28.70 (1)**	**22.67 (1)**	**25.51 (1)**	**24.58 (1)**
NRC-yisi-bicov	**22.03 (8)**	**28.72 (6)**	**21.34 (7)**	19.66 (12)	26.35 (21)	**22.06 (4)**	**25.21 (3)**	**23.89 (6)**
NRC-yisi	21.34 (20)	27.97 (12)	**21.12 (9)**	19.26 (19)	26.00 (22)	**21.79 (8)**	**24.99 (5)**	**23.52 (10)**
100M-word								
highest scores	**25.80 (1)**	**31.35 (1)**	**23.17 (1)**	**22.51 (1)**	**31.45 (1)**	**24.00 (1)**	**26.93 (1)**	**26.49 (1)**
NRC-yisi-bicov	**25.76 (3)**	**31.35 (1)**	22.80 (15)	**22.36 (9)**	**31.11 (7)**	**23.84 (5)**	**26.93 (1)**	**26.40 (5)**
NRC-yisi	**25.63 (7)**	**31.04 (9)**	**23.16 (2)**	**22.46 (5)**	30.83 (18)	**23.93 (3)**	**26.82 (5)**	**26.37 (6)**

NMT

10M-word

	dev.	test						
domain	news	news	speech	laws	medical	news	IT	
system \ test set	newstest17	newstest18	iwslt17	Acquis	EMEA	Global Voices	KDE	average
highest scores	**29.44 (1)**	**36.04 (1)**	**25.64 (1)**	**25.57 (1)**	**32.72 (1)**	**26.72 (1)**	**28.25 (1)**	**28.62 (1)**
NRC-yisi-bicov	**27.61 (8)**	**33.93 (9)**	**24.37 (9)**	23.20 (12)	29.75 (13)	**25.44 (7)**	**27.75 (4)**	**27.41 (8)**
NRC-yisi	26.62 (11)	32.72 (12)	23.89 (11)	22.22 (19)	28.55 (19)	24.83 (12)	**26.81 (8)**	26.50 (12)
100M-word								
highest scores	**32.41 (1)**	**39.85 (1)**	**27.43 (1)**	**28.43 (1)**	**36.72 (1)**	**29.26 (1)**	**30.92 (1)**	**32.06 (1)**
NRC-yisi-bicov	**31.97 (3)**	**39.59 (4)**	**26.95 (9)**	**28.35 (4)**	**36.59 (3)**	**29.09 (3)**	**30.70 (5)**	**31.88 (4)**
NRC-yisi	31.53 (11)	**39.30 (9)**	**27.13 (4)**	27.91 (13)	36.28 (12)	**29.01 (6)**	**30.92 (1)**	**31.76 (6)**

system	10M-word average	100M-word average	all average
highest scores	**26.54 (1)**	**29.27 (1)**	**27.90 (1)**
NRC-yisi-bicov	**25.65 (8)**	**29.14 (4)**	**27.39 (6)**
NRC-yisi	25.01 (11)	**29.07 (5)**	**27.04 (9)**

Table 3: BLEU scores (and ranking, out of 48 submissions) of NRC's supervised submissions: "bicov" indicates that the final bigram coverage step (§2.4) was performed. The highest scores of each testing conditions are included for reference. Results in the top 10 performers are bolded.

moval had virtually no effect on the larger selections.

Compared to other top-ranking submissions, both of our supervised submissions have weaker MT performance in the 10M-word training conditions although still rank above the median system on all test sets. This suggests that our systems are generally good at identifying parallel sentences for the 100M-word training set but relatively weaker at ranking the sentence pairs according to the usefulness-for-MT beyond parallelism. Although the redundancy removal heuristic appeared to play a more significant role in the 10M-word training conditions, the improvements on the official test sets are less substantial than what we observed in our internal experiments. This is potentially due to the differences in architecture between our MT systems and the MT systems built in the official evaluation.

5 Conclusion

In this paper, we presented the NRC supervised submissions (`NRC-yisi-bicov` and `NRC-yisi`) to the WMT18 parallel corpus filtering task. The core of the submissions used YiSi – a novel semantic machine translation (MT) evaluation metric to score the semantic textual similarity between the translated German side and the English of the target sentence pair. Despite failing to combine with other fluency or grammaticality oriented features, the YiSi-based system with redundancy removal performed well in the shared task, particularly in the 100M-word evaluation (4th place out of 48 submitted systems). This shows that using an adequacy oriented scoring measure is a reliable method to identify good sentence pairs for training MT systems. At the same time, the slightly worse performance in the 10M-word evaluation (8th place out of 48 submitted systems) also suggests that fluency or grammaticality oriented features are useful for fine-grained ranking of MT training data quality. Thus, future work includes investigating other feature combination methodologies, such as more realistic tuning example generation.

References

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics. Revised August 2010.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of NAACL HLT 2013*.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proc. 2012 Conf. of the N. American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1370–1380, Baltimore, Maryland.

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*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362. Association for Computational Linguistics.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proc. 45th Annual Meeting of the Assoc. for Comp. Linguistics*, pages 144–151, Prague, Czech Republic.

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.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Samuel Larkin, Boxing Chen, George Foster, Uli Germann, Eric Joanis, J. Howard Johnson, and Roland Kuhn. 2010. Lessons from NRC's Portage System at WMT 2010. In *5th Workshop on Statistical Machine Translation (WMT 2010)*, pages 127–132.

Patrick Littell, Samuel Larkin, Darlene Stewart, Michel Simard, Cyril Goutte, and Chi-kiu Lo. 2018. Measuring sentence parallelism using Mahalanobis distances: The NRC unsupervised submissions to the WMT18 Parallel Corpus Filtering shared task. In

Proceedings of the Third Conference on Machine Translation (WMT 2018).

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo. 2018. The NRC metric submission to the WMT18 metric and parallel corpus filtering shared task. In *Arxiv*.

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.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Modeling for NLP*, Denver, United States.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. 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 528–540. 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 *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, Pennsylvania.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413, Doha, Qatar. Association for Computational Linguistics.

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.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Processdings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1387–1392.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. 2016. Edinburgh's statistical machine translation systems for wmt16. In *Proceedings of the First Conference on Machine Translation*, pages 399–410, Berlin, Germany. Association for Computational Linguistics.