H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings

Houda Bouamor[1] **and Hassan Sajjad**[2]
[1]Carnegie Mellon University in Qatar, [2]Qatar Computing Research Institute, HBKU, Qatar
hbouamor@cmu.edu, hsajjad@qf.org.qa

Abstract

This paper presents our solution for the BUCC 2018 Shared Task on parallel sentence extraction from comparable corpora. Our system identifies parallel sentence pairs in French-English corpora by following a hybrid approach pairing multilingual sentence-level embeddings, neural machine translation, and supervised classification. Our system consists of a two-step process. In the first step, to reduce the size and the noise of the candidate sentence pairs, we filter the target translation candidates using the continuous vector representation of each source-target sentence pair learned using a bilingual distributed representation model. Then we select the best translation using a neural machine translation system or a binary classification model. We achieve an F_1-score of up to 75.2 and 76.0 on the BUCC18 train and test sets respectively.

Keywords: Comparable Corpora, Parallel Sentences, Multilingual Embeddings, Neural Machine Translation, Supervised Classification

1. Introduction

Building standard machine translation (MT) systems require a large amount of sentence-aligned parallel corpora. While these resources are available for mainstream languages (i.e., English, French, German, and Arabic) and domains, unfortunately, many low resourced languages and specialized domains suffer from the scarcity of such corpora. The manual generation of parallel data for several language pairs needs human expertise, a costly and time-consuming task. Although this problem can be alleviated by exploiting a pivot language to bridge the source and target languages (Cohn and Lapata, 2007; El Kholy et al., 2013; Sajjad et al., 2013; Cheng et al., 2017), the performance of such systems is never comparable to the ones built using parallel corpora. The scarcity of these resources pushed researchers to investigate the use of comparable corpora (Bouamor et al., 2013a; Rapp et al., 2016).

Comparable corpora include non-aligned sentences, phrases or documents that are not an exact translation of each other but share common features such as domain, genre, sampling period, etc. (Wu and Fung, 2005). Wikipedia articles describing the same topic, but written in two different languages (Barrón-Cedeño et al., 2015) and news topics covered in different newspapers appearing the same day, reporting about the same event or describing the same subject, are both good examples of comparable corpora. These resources could be leveraged to automatically extract parallel sentence pairs and build a parallel corpus between two languages. In recent years, there has been a body of work related to MT based on non-parallel comparable corpora. Rapp et al. (2016) gives a detailed survey of the use of comparable corpora in MT and several other NLP tasks.

In this work, we present our solution for the BUCC 2018 Shared Task on parallel sentence extraction from comparable corpora. Our system identifies parallel sentence pairs in French-English corpora by defining a hybrid approach pairing multilingual sentence-level embeddings, neural machine translation, and supervised classification.

The two monolingual corpora provided in the shared task are of approximately 370K and 270K sentences. Here, every target sentence is a candidate translation of every source sentence. The search space for the number of comparisons is is very large. To tackle this, we propose a two-step process. In the first step, in order to reduce the size of the candidate sentences, we filter the English translation candidates using the continuous vector representation of each French-English sentence pair learned using a bilingual distributed representation model. Then we select the best translation by leveraging the output of a neural machine translation system or a supervised classification model.

The remainder of this paper is organized as follows: We give a detailed description of our approach in Section 2.. Then, we present our experimental setup in Section 3.. We finally report and discuss our system results in Section 4..

2. Approach

When dealing with comparable corpora, every sentence in the target corpus can be a potential translation of every source sentence. Given a source corpus of S sentences and a target corpus of T sentences, the number of comparisons required to find translation pairs are $S \times T$. Given the large size of S and T, the search space becomes very large to find translation pairs from the corpus efficiently. In this work, we split the process of parallel sentence extraction into two steps: The first step reduces the search space from millions of comparisons to a few hundreds of top candidate pairs. In the second step, we select the best translation from the list of candidate pairs.

In the first step, we use multilingual sentence embeddings to identify top N closest target sentences to a source sentence. In the second step, we use machine translation, a machine translation evaluation metric, and binary classifier to select the best translation from the list of N candidate pairs.

2.1. Bilingual Distributed Representations

Monolingual distributed word representations have shown great potential in boosting the performance in several NLP tasks (Iacobacci et al., 2015; Guzmán et al., 2016; Santos et al., 2017). The use of word embeddings was further extended to include multilingual tasks (Zou et al., 2013;

Adams et al., 2017; Ammar et al., 2016), where distributed representations are induced over different language-pairs and thus serve as an effective way of capturing linguistic regularities in words that share same semantic and syntactic space, across languages (Gouws et al., 2015). However, there is a major problem with using monolingual word embeddings in a multilingual scenario. The models are usually trained independently for each of the languages using vector spaces. Thus, measuring the similarity between words is a challenging task, even for similar words.

Much research work has been conducted to address this problem, following several approaches (Luong et al., 2015): (i) *Bilingual mapping*, where word representations are trained for each language independently, and a linear mapping is then learned to transform representations from one language to another (Mikolov et al., 2013; Grégoire and Langlais, 2017); (ii) *Monolingual adaptation* that relies on pre-trained embeddings of the source language when learning target representations (Zou et al., 2013); and (iii) *Bilingual training* aiming at jointly learning representations for both languages using a parallel corpus, benefiting from word alignments (Luong et al., 2015) or without word alignments (Gouws et al., 2015).

In our model, we exploit the power of bilingual distributed representations to identify highly similar sentences in a $fr-en$ comparable corpus. For this, we use multivec (Bérard et al., 2016), an implementation of (Luong et al., 2015)'s bivec model for bilingual distributed representations. This toolkit is used for computing continuous representations for text at different granularity levels (word-level or sequences of words).[1]

Similarly to word2vec (Mikolov et al., 2013), for each pair of sentences in a parallel corpus, bivec tries to predict words in the same sentence, but also uses words in the source sentence to predict words in the target sentence (and conversely).

Following this approach, we first train multivec on a large $fr-en$ parallel corpus, to build a bilingual sentence level embedding model in the same vector space. Then, we use the model to learn a continuous representation for each source and target sentences from the train and test datasets provided in the shared task.

Our system detects a parallel sentence pair by measuring the cosine similarity between a sentence vector $\vec{fi}$ of each French sentence fr_i (in the source corpus) and each vector $\vec{e_j}$ corresponding to a possible en_j candidate (in the target corpus). We define each sentence embedding defined as an average of the source word embeddings of its constituent words. We create our set of candidate pairs by keeping the top N most similar target sentences for each source sentence fr_i (as per the cosine similarity measure) .[2]

2.2. Candidate Filtering

We follow two approaches to filter further the parallel sentence candidates obtained using the multilingual vector similarity: machine translation and supervised classification.

2.2.1. Machine Translation

To this point, we have a list of translation candidate sentences for every source sentence. We have reduced our search space of comparison from thousands of options to 10 and 100 options by using the bilingual distributed representations. In order to choose the best translation for each source sentence, the ideal scenario would require a reference sentence against which we can compare the candidate translations and keep the closest one. We use machine translation to produce a "reference" translation for each source sentence.

We hypothesize that given a machine translation system of decent quality, translation of a source sentence should be closest to its parallel sentence in the target language. To achieve this, we translate all French sentences in the comparable corpora to English using the French to English machine translation system. Then, for every French sentence, we compare its translation against all the English candidate sentences. The candidate sentence that gives the highest BLEU (Papineni et al., 2002) above certain threshold is selected as a translation of the source sentence. We use a high threshold above 50 BLEU point to discard source sentences that do not have any matching translations among the candidate translations.

2.2.2. Supervised Binary Classification

Machine translation systems are not perfect and can induce translation errors and noise, which impacts the quality of the sentence pairs identified. In order, to experiment with a more straightforward approach that leverages only the source-target sentence pairs without any intermediate step, we explore the use of supervised classification.

After obtaining the top N candidate source-target parallel sentence pairs from the first step (described in Section 2.1.), we build a bi-class classifier to identify parallel sentences among them, without translating the source sentences into the target language. Our system takes as input a French sentence FR and each of its English candidate EN_n (considered here as a possible translation) and outputs a score for each pair FR-EN_n estimating a kind of translation quality. The parallel sentence pair FR-EN_{best} selected is the one that has the highest quality score.

For this, we use a Support Vector Machine (SVM) classifier and exploit a rich set of features to represent a French source language sentence and each of its English translation candidates.

Learning features: We use the following group of features which have been used in work related to translation quality estimation for several languages (Bouamor et al., 2013b; Specia et al., 2015).

- **General features**: For each sentence, we use different features modeling its length in terms of words, the ratio of source-target length, source-target punctuation marks, numerical characters, and source-target content words.[3]

[1] https://github.com/eske/multivec

[2] Since we are working with vector representations, doing the Cartesian product is possible.

[3] As the English candidates are not the output of a machine translation system, there was no need to use language modeling or MT-based features (such as perplexity scores or number of OOVs)

- **Morphosyntactic features:** We use features to model the difference of sequences of POS tags for a pair of source-target sentences. These features measure the POS preservation between a source sentence and its target candidate. We compute the absolute difference between the number of different POS tags. We also indicate the percentage of nouns, verbs, and adjectives in the source and target sentences. The source and target sentences were tagged respectively using the French and English distributions of the Stanford coreNLP pipeline (Manning et al., 2014).

- **Named Entity features** A pair of parallel sentences usually contains the same number and type of Named Entities (NEs) (a translation/transliteration of each other). We use this hypothesis to measure the difference in number of various types of named entities in the source-target candidate parallel sentences. We use the CoreNLP named entity recognizer to extract persons, locations, organizations, and dates.

3. Experimental Setup

We experiment with different configurations and following several approaches. We present in this section our experimental settings and describe the datasets and tools used.

3.1. Dataset

In addition to the $fr-en$ training and testing datasets (BUCC18$_{train}$ and BUCC18$_{test}$) provided in the Shared Task, we use the $fr-en$ Europarl parallel corpus (Koehn, 2005) containing 2 million sentence pairs, as well as a News corpus made available from WMT 2016 with 183,000 aligned $fr-en$ sentence pairs (Bojar et al., 2016)(Europarl+News). All the corpora (French and English) are preprocessed through the following steps: Tokenization, POS tagging, and Name-Entity recognition. These preprocessing steps are completed using The Stanford CoreNLP Toolkit (Manning et al., 2014).

3.2. Model Training Settings

Continuous Vector Modeling: to train our bilingual model, we use the parallel $fr-en$ Europarl+News corpus described above, with the default configuration of the multivec tool. The model was trained using a learning rate α set to 0.05, a sample (a threshold on words' frequency) set to 0.001 and a window size of 5.

Machine Translation: we use the OpenNMT toolkit (Klein et al., 2017) to train a 2-layered LSTM encoder-decoder with attention (Bahdanau et al., 2015). In order to keep the training and test time low, we restrict ourself to uni-directional LSTM model. We use the default settings: embedding layer size: 512, hidden layer size: 1,000.
We limit the vocabulary to 40,000 words using BPE (Sennrich et al., 2016) with 40,000 operations. The sub-word units help us to map various morphological variations of a word to known sub-units. It also fixes the mismatch of vocabulary between our training corpus of machine translation and comparable corpus by splitting the unknowns in the comparable corpus into known sub-word units of the training corpus.

Binary Classification: we use the models described in Section 2.2.2. to build a Support Vector Machine (SVM) binary classifier using the LinearSVC implementation of scikit-learn[4].
To train our classifier we needed a gold standard corpus where a pair of $fr-en$ sentence is labeled as having high or low translation quality.
In order to build this dataset, we use the Europarl-News parallel corpus. Each sentence pair in this corpus is considered as a positive example (high translation quality). We then built a set of negative training examples (low translation quality), by selecting sentences from the French part of the corpus and randomly assigning a sentence from the English part to them. 80% of this corpus is used for training and 20% for testing. None of the sentences provided in the Shared task are used in building this classification model.

3.3. Evaluation Protocol

We evaluate our models, after obtaining the final predicted $fr-en$ parallel sentence pairs, using precision (P), recall (R), and F_1 score, defined in the shared task as follows:

$$P = \frac{TP}{TP+FP};\ R = \frac{TP}{TP+FN};\ F_1 = \frac{2PR}{P+R}$$

Where TP stands for the number of $fr-en$ sentence pairs that are present in the gold standard provided. A false positive (FP) is a pair of sentences that are not present in the gold standard. And a false negative (FN) is a pair of sentences present in the gold standard, but absent from system results.

4. Evaluation and Results

We tested several configurations:

Baseline: Our baseline consists of selecting $fr-en$ sentence pairs predicted only by the cosine similarity between sentence embedding pairs (described in Section 2.1.) with N=10. Since our method looks for a translation for every French sentence, we have a large number of false positives. Later, we use machine translation and classification to filter out these false positive pairs.

Machine translation: We take N=10 best candidates from our baseline system. For every French sentence, we compare its English translation generated automatically using a machine translation system against the ten candidate sentences. We sort the candidates based on BLEU and choose a translation with the best BLEU score above a certain threshold. Table 1 shows the results on the BUCC18$_{train}$ set when tested for different values of BLEU. The multivec-10best shows the highest initial recall of the list before applying BLEU-based filtering. The system achieved best f-score at BLEU value 0.57. It is interesting to see that a small difference in BLEU threshold dropped the recall by more than two points. This could be due to the nature of the BLEU metric that prefers exact ngram matches and penalizes words that are only different

[4]available at:`http://scikitlearn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

by a small morphological change. We suspect that BLEU at the sub-word level or Meteor would be less sensitive to small threshold changes and may result in a better balance between precision and recall.

Method	**P**	**R**	**F_1**
multivec-10best	-	83.4	-
BLEU-0.00	2.8	82.3	5.3
BLEU-0.50	60.6	77.3	67.9
BLEU-0.52	62.2	74.5	72.2
BLEU-0.55	79.0	71.7	75.2
BLEU-0.57	83.9	69.1	75.8
BLEU-0.59	87.5	65.8	75.1

Table 1: Precision, recall and F_1 on BUCC18$_{train}$, when filtered for various BLEU thresholds. **multivec-10best** shows the oracle recall that our system can achieve.

Classification: We measure the accuracy of our classifier on the external dataset (Europarl+News$_{test}$) as well as the train and test sets provided for the French-English task: BUCC18$_{train}$ and BUCC18$_{test}$. The source-target pairs that exist in the training and testing gold standards have been considered as positive examples, and an equivalent number from the rest of the pairs, generated by applying the multilingual word embedding based approach are considered as negative examples. Table 2 reports the accuracy of the classifier on different test sets of different nature and various sizes. The results obtained are encouraging, as we only exploit a group of basic features and do not include any semantic features such as sentence vector similarity or machine translation features.

H2@BUCC-2018 Results: We submitted three runs of our system:

- **Run1**: 10 best candidates with a BLEU filtering threshold of 0.52;
- **Run2**: 10 best candidates with a BLEU filtering threshold of 0.55;
- **Run3**: 10 best candidates with the SVM binary classification model output.

Table 3 summarizes the results for the three runs. Because of the time constraint, we reduced the number of candidate sentences to 10 only. This caused a loss of more than 16% in recall. In future, we would like to increase the candidate list to 100 candidates. This would slow down the filtering process but would result in better F_1 score.
The best machine translation results mentioned in Table 1 dropped the recall by 14 points. In future, we would like to consider other metrics like classification and sentence embedding in combination with MT results to improve the loss in recall.

	#of examples	**Accuracy**
Europarl+News$_{test}$	437,103	**81.05**
BUCC18$_{train}$	18,178	72.60
BUCC18$_{test}$	18,086	72.73

Table 2: Accuracy of the classifier on different test sets. The size of each test set is indicated.

	P	**R**	**F_1**
Run1	71	**75**	73
Run2	**82**	72	**76**
Run3	70	64	67

Table 3: Official results of our system on the BUCC2018 Testset

5. Conclusion

In this work, we presented our system to extract parallel sentences from comparable corpora. Initially, we learned sentence embedding vectors of the source and target languages using a parallel corpus. For every source sentence, we found the closest target sentence embeddings to create a list of candidate sentences. We then chose the best translation from the candidate translations by considering it either as a machine translation evaluation task or a binary classification task of choosing the best translation given a source sentence. Our method achieved an F_1-score of up to 75.2 and 76.0 on the BUCC18 train and test sets respectively.

6. References

Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-Lingual Word Embeddings for Low-Resource Language Modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain.

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively Multilingual Word Embeddings. *CoRR*, abs/1602.01925.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate.

Barrón-Cedeño, A., España Bonet, C., Boldoba, J., and Màrquez, L. (2015). A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13, Beijing, China.

Bérard, A., Servan, C., Pietquin, O., and Besacier, L. (2016). MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portoroẑ, Slovenia.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Bouamor, D., Popescu, A., Semmar, N., and Zweigenbaum, P. (2013a). Building Specialized Bilingual Lexicons Using Large Scale Background Knowledge. In *Proceedings of the 2013 Conference on Empirical Methods in*

Natural Language Processing, pages 479–489, Seattle, Washington, USA.

Bouamor, H., Mohit, B., and Oflazer, K. (2013b). SuMT: A Framework of Summarization and MT. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 270–278, Nagoya, Japan.

Cheng, Y., Yang, Q., Liu, Y., Sun, M., and Xu, W. (2017). Joint training for pivot-based neural machine translation. In *Proceedings of IJCAI*.

Cohn, T. and Lapata, M. (2007). Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic.

El Kholy, A., Habash, N., Leusch, G., Matusov, E., and Sawaf, H. (2013). Language Independent Connectivity Strength Features for Phrase Pivot Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Sofia, Bulgaria.

Gouws, S., Bengio, Y., and Corrado, G. (2015). BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In *Proceedings of the 32nd International Conference on Machine Learning*.

Grégoire, F. and Langlais, P. (2017). BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 46–50, Vancouver, Canada.

Guzmán, F., Bouamor, H., Baly, R., and Habash, N. (2016). Machine Translation Evaluation for Arabic using Morphologically-enriched Embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1398–1408, Osaka, Japan.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Luong, T., Pham, H., and Manning, C. D. (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, June. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Morristown, NJ, USA.

Rapp, R., Sharoff, S., and Zweigenbaum, P. (2016). Recent Advances in Machine Translation using Comparable Corpora. *Natural Language Engineering*, 22(4):501–516.

Sajjad, H., Darwish, K., and Belinkov, Y. (2013). Translating Dialectal Arabic to English. In *Proceedings of the 51st Conference of the Association for Computational Linguistics (ACL)*.

Santos, L., Corrêa Júnior, E. A., Oliveira Jr, O., Amancio, D., Mansur, L., and Aluísio, S. (2017). Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1284–1296, Vancouver, Canada.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Specia, L., Paetzold, G., and Scarton, C. (2015). Multilevel Translation Quality Prediction with QuEst++. In *ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China.

Wu, D. and Fung, P. (2005). Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-comparable Corpora. In *International Conference on Natural Language Processing*, pages 257–268. Springer.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA.