Recent works on Parallel Sentence Extraction from Comparable Corpora

Sakshi Maskara
Indian Institute of Technology, Bombay
sakshi@cse.iitb.ac.in

Pushpak Bhattacharyya
Indian Institute of Technology, Bombay
pb@cse.iitb.ac.in

Abstract

In this global era, the demand for translation is rapidly growing in various scenes, and it is impossible to translate everything manually. Machine translation (MT), as a powerful tool to improve the efficiency and reduce the cost of translation, is quite important to promote globalization. But the most popular MT systems like Statistical MT and Neural MT are data driven. The translation knowledge is automatically learned from parallel corpora. Therefore, the quality and quantity of parallel corpora is of utmost important.

Parallel Corpora is not a widely available resource for many language pairs and hence, extraction of parallel corpora from comparable corpora is a major area of research. This paper compiles all the recent works done in the field of extraction of parallel sentences from comparable corpora, using word embedding based, machine translation based, and deep learning based approaches.

1 Introduction

In today's world, with the growth of numerous fields and technologies, large amount of information is present over the web. Due to this, machine translation systems have become a major requirement as the people accessing the web, come from different language backgrounds. The major bottleneck in machine translation systems like Statistical MT and Neural MT is the increasing amount of parallel corpora required to train the systems. Since manual construction of parallel corpora requires a lot of manpower and resources, it is not a practical approach which is why automatic creation of parallel corpora has become an attractive field of research.

Comparable corpora is a pair of bilingual documents which are not sentence aligned but topic aligned, that is, they come from the same domain. They are widely present for various language pairs and in variety of domains. One such source of comparable corpora is Wikipedia (Smith et al., 2010). Wikipedia has a huge collection of articles on a large variety of domains and in various languages.

With the advent of neural network and continuous vector representation of words, deep learning systems are replacing traditional machine learning approaches in all natural language processing tasks. Word embedding models (Mikolov et al., 2013; Bengio et al., 2003; Bojanowski et al., 2016) capture many linguistic features between words in vector space. Hence, parallel corpora extraction systems also uses these new techniques like word embedding models, deep learning approaches etc. to extract parallel data from comparable corpora.

In our survey, we focus on the recent works being done in the field of parallel sentence extraction from comparable corpora based on word embeddings, machine translation systems and deep learning based approaches.

2 Traditional Methods

Many research work has been conducted in the domain of constructing parallel corpora from comparable corpora. Some researchers have used bilingual news articles (Munteanu et al., 2004; Abdul Rauf and Schwenk, 2011; Bing Zhao and Vogel, 2002; Utiyama and Isahara, 2003), while others have used Wikipedia (Smith et al., 2010; Stefanescu and Ion, 2013;

Chu et al., 2014, 2016) as a source of comparable corpora to generate parallel corpora.

Resnik and Smith (2003), working on web pages, use STRAND, which is their structural filtering system, to recognize parallel pairs. In order to do so, they specify a set of pair-specific values and experiment on English-Chinese corpus, reporting precision and recall of 98 percent and 61 percent, respectively.

Munteanu et al. (2004) used publication dates and vector-based similarity to identify similar news articles, then did a Cartesian product of sentences from these article pairs to generate candidate sentence pairs. They passed these sentence pairs through a word overlap based filter and then trained a maximum entropy based classifier to classify them as parallel or non-parallel. They evaluated the quality of the extracted parallel sentences by improving the performance of a Statistical Machine Translation system.

Fung and Cheung (2004) used a bootstrapping approach where parallel sentences are extracted in iterations. In each iteration, new word translations were learnt from the intermediate output of the previous parallel sentence extraction. This refines their bilingual lexicon with more in-domain data, and hence the overall extraction process.

Koehn (2005) extract parallel texts for 11 languages from the proceedings of the European Parliament to be used as the training data for building SMT systems.

Smith et al. (2010) used a first order linear chain Conditional Random Field (CRF) for aligning parallel sentences within aligned document pairs from Wikipedia. This model is taken from discriminative CRF-based word alignment model described by Blunsom and Cohn (2006). They use HMM based word alignment features and position of aligned sentences in document pairs to train their model.

Stefanescu and Ion (2013) used Wikipedia to build English-German, English-Romanian and English-Spanish parallel corpus. They first aligned similar documents using cross-lingual Wikipedia links, then used a tool called LEXACC developed by ACCURAT project which measures similarity between sentence pairs, to extract parallel sentences.

Chu et al. (2014) have generated a Chinese-Japanese parallel corpus by using common Chinese characters as additional features for filtering and content and non-content word features for classification.

3 Word Embedding based approaches

Bouamor and Sajjad (2018) uses a hybrid approach comprising multilingual sentence level embeddings, neural machine translation and supervised classification. They use a two step process to extract parallel sentences from candidate ones.

1. **Multilingual sentence level embedding** In this step, they have used some filtering mechanism to reduce the search space from millions of comparisons to hundreds. In this process, they have used multilingual word embeddings instead of monolingual word embeddings. The motivation behind using multi-lingual word embeddings is that the models for monolingual word embeddings are trained separately for each language using different vector spaces. Hence, even similar words have different vector representation and so capturing similarity between them becomes very challenging. So, following this, they use a multivec toolkit developed by Berard et al. (2016) and build a bilingual word embeddings.

   Then, they use this model to learn a continuous representation for each source and target sentences from the train and test datasets in the shared task. They learn sentence embeddings by averaging the word embeddings of each word in the sentence. Parallel sentence pairs are recognized by measuring cosine similarity between candidate English and French sentence pairs, and for every source sentence, top N most similar target sentences are kept. They used two different approaches to further filter the candidate parallel sentences.

2. **Neural MT** In this approach, they used a French-English translation system to translate all French sentences to English and calculated the BLEU score between the translated French sentence and the English sentence. All sentence pairs below 50 BLEU score were discarded.

3. **Supervised Classification** In this approach, they classified sentence pairs using a SVM classifier. They built a rich feature set using context similarity, morpho-syntactic features and named entity features and use to classify sentence pairs as parallel or non-parallel.

Leong et al. (2018) also uses a two step approach comprising of alignment candidate identification and classification models. (Check figure 1)

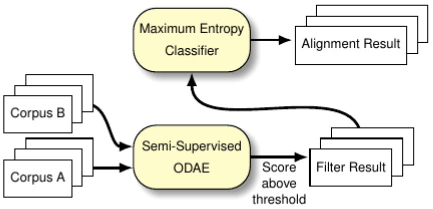


Figure 1: Model Arcitecture used by Leong et al. (2018)

Alignment candidate identification They use a semi-supervised orthogonal denoising autoencoder to transform the embedding features of candidate parallel sentences into shared and private latent spaces, with an objective to better capture the translation correspondences of parallel sentences. The autoencoder gives scores to all the candidate parallel sentence pairs and those which are scored above a certain threshold are selected, and rest are discarded.

Classification Models In this step, a maximum entropy classifier is employed to determine and select the parallel sentences from the candidate list, which is also done using a scoring mechanism. The features used by the classifier include the length-based features (Gale and Church, 1993), alignment-based features (Munteanu et al., 2004; Dyer et al., 2013) and the anchor text Patry and Langlais (2011). During the alignment process, one source sentence is only allowed to align to a target sentence once. The candidate with the highest score is considered.

4 Translation based approaches

Chu et al. (2016) uses neural network features into a parallel sentence extraction system, which consists of a parallel sentence candidate filter and a binary classifier for parallel sentence identification.

The steps followed in this method are:

- In the first step, all articles in Wikipedia based on the same topic are aligned via the inter language links

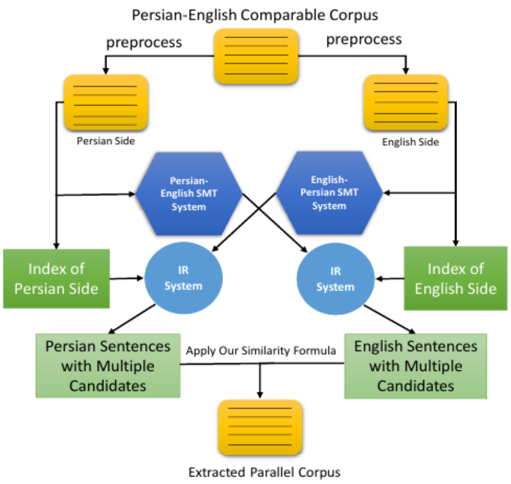
- Next, they do a simple Cartesian product of the topic aligned articles obtained from the above step, and generate all possible sentence pairs. The pairs which do not fulfill the condition of the filter are discarded to make the candidate pairs more reliable.

- Next, a classifier is trained on a small number of parallel sentences from a seed parallel corpus to identify the parallel sentences from the candidate pairs.

- In the last step, a NMT model is trained on the identified parallel sentences and obtain the neural network features which are fed into the above classifier to improve its performance.

They train four neural translation models. For each translation direction, they train character and word based models using the parallel corpus. After training a neural translation model, it is used to produce a score for a sentence pair, where the neural translation model is viewed as a bilingual language model. These four scores are used as the neural network features for the classifier.

Karimi et al. (2017) uses Machine translation and an information retrieval system to extract parallel sentences from English and Persian document aligned Wikipedia. (Check figure 2) They use two machine translation systems to translate from English to Persian and vice versa, then uses a Lucene IR system to measure the similarity of the sentence pairs. The similarity scores are calculated for both English and its translation from Persian and, Persian and its translation. For every English sentence, the persian sentence which has the highest similarity score is retained, and others are dropped. These extracted sentences when used along with seed parallel corpus to train a machine translation system, improves its performance.



Persian-English Comparable Corpus
preprocess
preprocess
Persian Side
English Side
Persian-English SMT System
English-Persian SMT System
IR System
IR System
Index of Persian Side
Index of English Side
Persian Sentences with Multiple Candidates
Apply Our Similarity Formula
English Sentences with Multiple Candidates
Extracted Parallel Corpus

Figure 2: The architecture of the bidirectional method for parallel sentence extraction. (Karimi et al., 2017)

Mahata et al. (2017) implemented a method that translates a French corpus with a Machine Translation system, selects candidate sentence pairs with a suitable length ratio, and chooses the final sentence pairs based on Cosine similarity.

Zhang and Zweigenbaum (2017) used a bilingual dictionary to perform word-level translation of the ZH corpus, complemented by calls to an on-line Machine Translation system. They used the Solr search engine to index sentences and search for similar sentences, collecting a number of candidate translations for each source sentence. They selected the best translation (or none) by training a classifier with Solr score and rank, word overlap, and sentence length features.

Azpeitia et al. (2017) used probabilistic dictionaries acquired by word alignment of parallel corpora to translate each corpus. They used the Lucune search engine to index sentences and search for similar sentences, collecting a number of candidate translations for each source sentence, in both directions. Final sentence similarity is computed by their STACC method (SetTheoretic Alignment for Comparable Corpora), which extends basic word overlap by taking into account non-matched words that share a long enough common prefix, as well as numbers and capitalized true-cased tokens. STACC measures word overlap with the Jaccard coefficient. They refined the STACC method by taking into account lexical weights that penalize frequent words.

5 Deep learning based approach

As the deep learning based models gain popularity for NLP tasks and other classification tasks, there is a deep learning based approach reported for parallel sentence extraction task as well.

Grégoire and Langlais (2017) experimented with a deep learning framework. They trained bilingual word embeddings with BilBOWA (Bilingual Bag-of-Words without Alignments (Gouws and Søgaard, 2015) on the Europarl parallel corpus, represented source and target sentences in this common space and used Cosine similarity to select candidate parallel sentence pairs. They also trained a bidirectional recurrent neural network with gated recurrent units on both the source and target languages to build sentence level continuous representations. They learned a linear transformation of these representations from one language to the other and decided on the parallelism of two sentences based on the comparison of their continuous representations through this transformation.

6 Conclusion

In this paper, we have covered various recent works done in the field of parallel sentence extraction from comparable corpora. We have described these past works on the basis of traditional, word-embedding based, machine translation based and deep-learning based approaches. In summary, a lot of work has been done in the field of parallel sentence extraction, in particular past approaches have used different types of features based on context similarity, word alignment based and distortion based to train their classifiers. The classifier used in most works are maximum entropy based classifier and support vector machine classifier.

Some research works have used information retrieval based mechanisms like Solr search engine and Lucene search engine to find similar sentences. Finally, we have also touched upon recent trends in parallel sentence extraction, where researcher have also tried with deep learning architecture and have shown improvement in performance as compared to statistical baseline approaches.

References

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. 25:341–375.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Alexandre Berard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. Multivec: a multilingual and multilevel representation learning toolkit for nlp. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Bing Bing Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 745–748.

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 65–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Houda Bouamor and Hassan Sajjad. 2018. H2@bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2016. Parallel sentence extraction from comparable corpora with neural network features. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a chinese-japanese parallel corpus from wikipedia. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *In Proc. NAACL*.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390. Association for Computational Linguistics.

Francis Grégoire and Philippe Langlais. 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. Association for Computational Linguistics.

Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2017. Extracting an english-persian parallel corpus from comparable corpora. *CoRR*, abs/1711.00681.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Chongman Leong, Derek F. Wong, and Lidia S. Chao. 2018. Um-paligner: Neural network-based parallel sentence identification model. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2017. Bucc2017: A hybrid approach for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 56–59. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Alexandre Patry and Philippe Langlais. 2011. Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 87–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411. Association for Computational Linguistics.

Dan Stefanescu and Radu Ion. 2013. Parallel-wiki: A collection of parallel sentences extracted from wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013)*.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 72–79, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zheng Zhang and Pierre Zweigenbaum. 2017. znlp: Identifying parallel sentences in chinese-english comparable corpora. In *BUCC@ACL*.