BUCC2017: A Hybrid Approach for Identifying Parallel Sentences in Comparable Corpora

Sainik Kumar Mahata$^1$
sainik.mahata@gmail.com
Dipankar Das$^2$
ddas@cse.jdvu.ac.in
Sivaji Bandyopadhyay$^3$
sbandyopadhyay@cse.jdvu.ac.in
$^{1,2,3}$ Department of Computer Science and Engineering, Jadavpur University, India

Abstract

A Statistical Machine Translation (SMT) system is always trained using large parallel corpus to produce effective translation. Not only is the corpus scarce, it also involves a lot of manual labor and cost. Parallel corpus can be prepared by employing comparable corpora where a pair of corpora is in two different languages pointing to the same domain. In the present work, we try to build a parallel corpus for French-English language pair from a given comparable corpus. The data and the problem set are provided as part of the shared task organized by BUCC 2017. We have proposed a system that first translates the sentences by heavily relying on Moses and then group the sentences based on sentence length similarity. Finally, the one to one sentence selection was done based on Cosine Similarity algorithm.

1 Introduction

Statistical Machine Translation (SMT) analyzes the output of human translators using statistical methods and extracts information about the translation process from corpora of translated texts. SMT has shown good results for many language pairs and is responsible for the recent surge in terms of popularity of Machine Translation among the research communities. But, for a SMT system to work efficiently, it has to be fed with large parallel corpus, for producing high quality phrase table and translation models (Brown et al., 1991; Church et al., 1993; Dagan et al., 1999). Since availability of large parallel corpus is an issue for low resourced languages, building one from scratch involves high manual labor and cost (Pal et al., 2014; Tan and Pal, 2014; Mahata et al., 2016). This is the reason why lot of research has gone into the concept of building parallel corpus, from comparable corpus (Jagarlamudi et al., 2011; Kay and Roscheisen, 1993; Kupiec, 1993; Lardilleux et al., 2012). A comparable corpus is a pair of monolingual corpus in the same domain, where the sentences in the both the corpus are not aligned. The proposed work deals with identifying parallel sentences from such a comparable corpus provided by BUCC 2017$^1$ shared task. Sample, training and test data contain monolingual corpora split into sentences, in the format, “utf-8 text, with UNIX end-of-lines; identifiers are made of a two-letter language code + 9 digits, separated by a dash ‘-’:

- Monolingual EN corpus (where EN stands for English), one tab-separated sentence_id + sentence per line.
- Monolingual FR corpus (where FR stands for Foreign, e.g. French), one tab-separated sentence_id + sentence per line.
- Gold standard list of tab-separated EN-FR sentence_id pairs (held out for the test data)

The algorithm of the proposed work has been constructed primarily using Moses (Koehn, 2015) toolkit that has been fed with parallel corpus from Europarl$^2$, with French as the source language and English as the target language. Also, the similarity based on sentence length has been used for the preliminary alignment because equivalent sentences in comparable corpus may roughly correspond with respect to length. Cosine Similarity algorithm was used for the final alignment. Section 2 will discuss the proposed algorithm in detail and will be followed by results and discussions in Section 3 and Section 4, respectively.

1https://comparable.limsi.fr/bucc2017/bucc2017-task.html
2http://www.statmt.org/europarl/
One sentence from the translated English corpus is taken and is matched with the selected sentences in English corpus. Example of this process is shown in Figure 1. The formula used in our approach is as follows:

\[
\text{Similarity} = \cos(\theta) = \frac{\langle A, B \rangle}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

(1)

Where "A" and "B" are the translated English sentence and one of the English sentences from the test data found out using the preliminary alignment system, respectively. One sentence from the translated English corpus is taken and is matched with the selected sentences in English corpus from the other language. This idea forms the basis of our preliminary alignment system, which tries to align sentence pairs based on their length. We have found out the length of the translated English sentence and have found matches in the sentences of the English text from the test data. This results in one-to-many relationship between the translated English and the English sentences from the test data. The variance in this step is kept as 4, which means if the length of the English sentences of the test data exceeds or falls behind by a factor of 4, when compared to the translated English sentence, they are also included in this step. This is done for reducing the time complexity of the Cosine Similarity search algorithm. Example of this step is shown in Figure 4.

2.3 Final alignment using Cosine Similarity Algorithm

Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. The formula used in our approach is as follows:

\[
\text{Similarity} = \cos(\theta) = \frac{\langle A, B \rangle}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

(1)

Figure 1: Translation of French sentences into English sentences using Moses.

Figure 2: Appending sentence_id's to translated English sentence
the test data, using the Cosine Similarity algorithm.

The sentence pair with the highest Cosine Similarity value is considered as the final alignment. Sentence ID's of the selected sentence pair are extracted and given as output. An example of the output format is shown in Figure 3.

Figure 3: Final alignment using Cosine Similarity

Sentence 1: 'Like all "Ghulam Maro" campaigns, "Prophets" contains a co-operative role-playing portion and a competitive Player versus Player (PvP) portion.'
Sentence 2: 'I. Afghanistan is a country with plains in the north and southwest. Afghanistan is located within South Asia and Central Asia.'

Evaluation

BUCC 2017 provided us with an evaluation script and a gold standard data to calculate the Precision, Recall and F-Score. This is shown in Figure 5. The calculation was done using value TP, FP and FN, where TP (true positive) is a pair of sentences that is present in the gold standard, FP (false positive) is a pair of sentences that is not present in the gold standard and FN (false negative) is a pair of sentences present in the gold standard but absent from system. We submitted 38,736 sentence pair alignment. Table 1 shows the results.

Table 1: Evaluation Results

| Proposed System | TP  | FP  | FN  | Precision | Recall | F-Score |
|-----------------|-----|-----|-----|-----------|--------|---------|
|                 | 1011 pairs | 37725 pairs | 8032 pairs | 0.0261 | 0.1118 | 0.0423 |

Table 2: Second evaluation Results.

As a future prospect, we would like to align the sentences based on Named-Entity and Edit distance approach.

Discussion

We tested the proposed approach by training Moses for translating English to French as well. The English data from the test data corpus was translated to Spanish. After preliminary alignment, Cosine Similarity was sought for translated Spanish corpus of the test data. After testing the system with the gold standard, we found out only one match.
5 Conclusion

The paper proposes a Hybrid approach for sentence alignment in comparable corpora. Moses toolkit was used for building the baseline translation system along with similarity based on sentence length and Cosine Similarity algorithms. The evaluation of the proposed method yielded results as Precision: 0.0261 Recall: 0.1118 and F-Score: 0.0423.

References

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL ’91, pages 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenneth Ward Church. 1993. Char align: A program for aligning parallel texts at the character level. In Proceedings of the 31st Annual Conference of the Association for Computational Linguistics, pages 1–8.

I. Dagan, K. Church, and W. Gale, 1999. Natural Language Processing Using Very Large Corpora, chapter Robust Bilingual Word Alignment for Machine Aided Translation, pages 209–224. Springer Netherlands, Dordrecht.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL ’91, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jagadeesh Jagarlamudi, Hal Daume, III, and Raghavendra Udupa. 2011. From bilingual dictionaries to interlingua document representations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume2, HLT ’11, pages 147–152, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Kay and Martin Roscheisen. 1993. Text-translation alignment. Comput. Linguist.,19(1):121–142, March.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics, ACL ’93, pages 17–22, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adrien Lardilleux, Franois Yvon, and Yves Lepage.2012. Hierarchical sub-sentential alignment with Anyalign. pages 279–286, Trento, Italy.

Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro.1993. Structural matching of parallel texts. In Proceedings of the 31st Annual Meeting on Association for Computational Linguistics, ACL ’93, pages 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, and Sudip Kumar Naskar.2014. Automatic building and using parallel resources for smt from comparable corpora. Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL, pages 48–57.

Alexandre Patry and Philippe Langlais. 2005. Automatic Identification of Parallel Documents With Light or Without Linguistic Resources, pages 354–365. Springer Berlin Heidelberg, Berlin, Heidelberg.

Michel Simard, George F. Foster, and Pierre Isabelle.1992. Using cognates to align sentences in bilingual corpora. In In proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, pages 67–81.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. Proceedings of the Ninth Workshop on Statistical Machine Translation, pages201–206.

Thuy Vu, Ai Ti Aw, and Min Zhang. 2009. Feature based method for document alignment in comparable news corpora. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL ’09, pages 843–851, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Huang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Sainik Kumar Mahata, Dipankar Das, and Santanu Pal. 2016. WMT2016: A Hybrid Approach to Bilingual Document Alignment. In proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 724–727, Berlin, Germany, August 11-12, 2016.