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/

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 56–59,
Vancouver, Canada, August 3, 2017.

```
Le pays est un carrefour de l'Asie qui voit passer de nombreux peuples par son territoire.  The country is a crossroads of Asia which sees
many peoples pass through its territory.
Cette région est cependant le noyau de vastes empires comme l'Empire bactrien, l'Empire kouchan ou encore l'Empire ghaznévide.  This region,
however, is the nucleus of vast empires such as the Bactrian Empire, the Kushan Empire or the Empire Ghazn Empire.
Le pays devient ainsi un État tampon de 1879 à 1919, demeurant indépendant sur le plan de la politique intérieure.  The country thus becomes
a buffer state from 1879 to 1919, remaining independent on the domestic policy level.
En 1996 un gouvernement islamiste, celui des talibans, prend le pouvoir et est chassé par une coalition internationale en 2001. In 1996, an
Islamist government, the Taliban government, took power and was expelled by an international coalition in 2001.
L'Afghanistan est un pays montagneux avec des plaines au Nord et au Sud-Ouest.  Afghanistan is a mountainous country with plains to the north
and southwest.
De grandes parties du pays sont arides, et l'eau fraîche est limitée.  Large parts of the country are arid, and fresh water is limited.
```

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

```
fr-000000001    The country is a crossroads of Asia which sees many peoples pass through its territory.
fr-000000002    This region, however, is the nucleus of vast empires such as the Bactrian Empire, the Kushan Empire or the Empire Ghazn
Empire.
fr-000000003    The country thus becomes a buffer state from 1879 to 1919, remaining independent on the domestic policy level.
fr-000000004    In 1996, an Islamist government, the Taliban government, took power and was expelled by an international coalition in
2001.
fr-000000005    Afghanistan is a mountainous country with plains to the north and southwest.
fr-000000006    Large parts of the country are arid, and fresh water is limited.
```

Figure 2: Appending sentence_id's to translated English sentence

2 Proposed System

2.1 Building baseline Statistical Machine Translation Model

Moses is a statistical machine translation system that allows you to automatically train translation models for any language pair, when trained with a large collection of translated texts (parallel corpus). Once the model has been trained, an efficient search algorithm quickly finds the highest probability translation among the exponential number of choices. For the given system, Moses was trained with French (Fr) as the source language and English (En) as the target language. The En-Fr parallel corpus that was used to train Moses has been downloaded from Europarl Corpus. The language model training of Moses was done by concatenating the English corpus of Europarl and the English text of the test data provided by BUCC 2017. The French corpora from the given test data was taken and sentences were extracted barring the sentence_id's. The extracted French sentences were then fed to Moses to get translated English sentences as output. Example of this process is shown in Figure 1. The segregated sentence_id's from the previous step were again appended to the translated English sentences. Example of this process is shown in Figure 2.

2.2 Sentence similarity based on sentence length

Gale and Church (1991) in their paper, proposed a system for aligning corresponding sentences in a parallel corpora, based on the principle that equivalent sentences should roughly correspond in length—that is, longer sentences in one language should correspond to longer sentences in 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.

$$Similarity = \cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

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 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.

```
fr-000000001    en-000121474
fr-000000002    en-000313524
fr-000000003    en-000292858
fr-000000004    en-000043944
fr-000000005    en-000193935
fr-000000006    en-000210237
fr-000000007    en-000269236
fr-000000008    en-000193986
fr-000000009    en-000218701
fr-000000010    en-000315531
```

Figure 3: Final alignment using Cosine Similarity

```
en-000000001    Like all "Guild Wars" campaigns, "Prophecies" contains a co-operative role-playing portion and a competitive Player versus
Player (PvP) portion.

fr-000000004    In 1996, an Islamist government, the Taliban government, took power and was expelled by an international coalition in 2001.
fr-000000014    The Afghans consider the medieval name of their country is Khorassan which currently designates a region of northeastern
Iran.
fr-000000029    The founder of Afghanistan also bears the title of "Bábó Málà", which, in pachto, means father of the Nation.
fr-000000034    Having left no instructions or protocol on his succession, Ahmad Shâh had complicated the succession to the Afghan throne.
fr-000000035    For the leaders of the time, it was no secret that Timour Shâh had the preference of his father.
fr-000000039    The young Timur was able to enter the city of Kandahar and be crowned Padishah of the Afghan Empire.
fr-000000048    The sudden death of Timur Shah Durrani opens an era of war and tears for succession to the throne.
fr-000000064    The Soviet Union unilaterally decided to leave the country in February 1989, leaving Nadjibullah in control of the country.
fr-000000070    Members of Hezb-Ã © -islami (party of Hekmatyar) enter the government of President Rabbani while Hekmatyar becomes prime
minister.
fr-000000087    Afghanistan will receive about 400 million euros in royalties per year for 30 years, the duration of the concession.
fr-000000089    Its vineyards are so abundant that the grains are given, for three months of the year, to the cattle.
fr-000000105    The exploitation of iron ores is not currently on the agenda, but represents a huge potential for the country.
```

Figure 4: Finding corresponding sentences with respect to Gale and Church algorithm.

```
fr-000000005    en-000193935    L'Afghanistan est un pays montagneux avec des plaines au Nord et au Sud-Ouest.  A landlocked mountainous
country with plains in the north and southwest, Afghanistan is located within South Asia and Central Asia. 1   3
fr-000000007    en-000269236    L'Afghanistan a un climat continental, avec des étés chauds et des hivers froids.   Ruse has a continental
climate () with very hot summers and relatively cold winters.    1   3
fr-000000079    en-000222220    La guerre d'Afghanistan est particulièrement liée au conflit armé du Nord-Ouest du Pakistan.    The
conflict in Afghanistan also forced millions of Afghan refugees into Pakistan, particularly in the northwestern regions.    1   2
fr-000000134    en-000251190    Il existe 40 langues répertoriées en Afghanistan dont 2 langues officielles nationales, le dari et le
pachto.   Pashto and Dari are both designated as the official languages of Afghanistan.   1   3
fr-000000138    en-000315065    L'Afrique du Sud est aussi la première puissance politique et militaire en Afrique. South Africa is the
largest economic and military power in the SADC.    1   4
fr-000000213    en-000087663    Le taux de violence sexuelle en Afrique du Sud était, en 2000, le plus élevé au monde.  The rate of sexual
violence in South Africa is among the highest in the world.  1   4
```

Figure 5: Result of evaluation.

3 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.

Proposed System	
TP	10111 pairs
FP	37725 pairs
FN	8032 pairs
Precision	0.0261
Recall	0.1118
F-Score	0.0423

Table 1: Evaluation Results.

4 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 and Spanish corpus of the test data. After testing the system with the gold standard, we found out only one match.

Second Evaluation	
TP	3 pairs
FP	20779 pairs
FN	9040 pairs
Precision	0.0001
Recall	0.0003
F-Score	0.0002

Table 2: Second evaluation Results.

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

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 Anymalign. 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 12thConerence 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 Hoang, 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.