Bilingual Sentence Alignment of a Parallel Corpus by Using English as a Pivot Language

Josafá de Jesus Aguiar Pontes
National Polytechnic School of Ecuador, Quito, Ecuador
Ladrón de Guevara E11-253, Quito 170517
josafa@furui.cs.titech.ac.jp

Abstract

Statistically training a machine translation model requires a parallel corpus containing a huge amount of aligned sentence pairs in both languages. However, it is not easy to obtain such a corpus when English is not the source or the target language. The European Parliament parallel corpus contains only English sentence alignments with 20 European languages, missing alignments for other 190 language pairs. A previous method using sentence length information is not enough reliable to produce alignments for training statistical machine translation models. Hybrid methods combining sentence length and bilingual dictionary information may produce better results, but dictionaries may not be affordable. Thus, we introduce a technique which aligns non-English corpora from the European Parliament by using English as a pivot language without a bilingual dictionary. Our technique has been illustrated with French and Spanish, resulting on an equivalent performance with the existing one in the original English-French and English-Spanish corpora.

1 Introduction

Obtaining a parallel corpus of aligned sentence pairs is an important task to further work for human translators and several natural language processing applications such as statistical machine translation (Brown et al, 1990; Melamed, 1998), cross-lingual information retrieval (Davis and Dunning, 2995; Landauer and Littman, 1990; Oard, 1997) and lexical acquisition (Gale and Church, 1991; Melamed, 1997), to mention some. Bilingual corpora are useful for human translators to search for a chunk of text in a source language and to find its corresponding translation into a target language. From the machine's standpoint, one of the most common applications is on training statistical models for machine translation. In the translation domain, no matter human or machine, they both need a very huge amount of aligned sentence pairs in order to find appropriate word combination that enable them to produce good translations.

Each language is a world of symbols made of its own set of words and their possible combinations that lead to a meaning from the native speakers' point of view. A parallel corpus comes as a map in between two languages, indicating which set of word combinations in a source language produces another set of words in a target language. Being so, we assume that the more sentence pairs there are in a corpus, the better is the mapping between the two languages and consequently, the better are the derived translations from it. Therefore, a huge amount of translated sentence pairs is essential.

Due to this growing demand, a number of parallel corpora have become available within the last decade, for instance the Europarl corpus (Koehn, 2005[1]), the News from OPUS[2], the JRC-

[1] Europarl, 2005. www.statmt.org/europarl/

[2] Tiedemann, 2009 http://opus.lingfil.uu.se/ECB.php

Acquis corpus3, the MultiUN corpus4 and the EU Official Journal EU Official Journal Multilingual Legal Text in 22 European Languages (Gale and Church, 1993), which are freely downloadable for research purposes. The Europarl corpus in particular is a parallel corpus extracted from the proceedings of the European Parliament. It consists of texts in 21 European languages, where English is the only language with which the other languages are aligned. Some of the remaining resources above mentioned do contain alignments between all combinations of language pairs; however, the quality of these alignments is questionable given that the alignment method utilized for most of them is solely based on sentence length information (Varga et al., 2005). Our experiments show that such alignments may present around 90% of precision. Obviously, the performance depends on the internal arrangement of the sentences being provided as input. Although the information of a good bilingual dictionary may be used to enhance the performance of an aligner (Schmid, 1994[5]), it is not normally available for free, even less when none of the two languages involved is English. In other words, most of the freely available non-English parallel corpora have not been aligned with the use of the respective bilingual dictionaries and therefore the quality relies basically on sentence length information.

Although the Europarl corpus has also been aligned with sentence length feature, there are underlying alignment information and noise removal which make the final quality to be very high. First, its alignment is simplified by the fact that the texts are originally available in a paragraph aligned format. Second, each paragraph is typically small, containing from 2 to 5 sentences only. Third, much noise is removed by discarding an utterance of a speaker when the number of paragraphs in it differs in the two languages being aligned. The prior data preparation done by the underlying paragraph information combined with the noise removal technique leads to an alignment of excellent quality. According to our experiments its precision reaches more than 99%.

Each corpus contains approximately 2 million English sentences and it is pairwise aligned with 20 other European languages. Since each parallel corpus is independently aligned, the number of sentences in each bitext is not the same across the language pairs. Most of the difference is due to the utterance removal process described above which occurred prior to the alignment. Consequently, not all the English sentences of a corpus (e.g. the English part from the English-French bitext) are present in the other corpus (e.g. the English part from the English-Spanish bitext). In other words, considering the English-French and the English-Spanish corpora for example, not all of the English sentences from the former can be found in the latter and viceversa. It means there are sentence insertions, deletions and substitutions when we consider two English corpora coming from diferent aligned language pairs of the same Europarl corpus.

It is unreasonable to expect the same alignment precision of two non-English texts from the Europarl corpus just by using the sentence length information. The prior sentence insertions, deletions and substitutions introduce an observable noise when comparing a pair of non-English texts, making harder the work of the aligner. In fact, our experiments point out to a precision of only 90% given an amount of such a data. As previously stated, a bilingual dictionary may be helpful to improve this figure, but unfortunately, good ones are very expensive[6] to be affordable by developing countries for research purposes.

Taking these constraints into consideration, we have developed a sentence alignment method which exempts the use a bilingual dictionary when a multilingual corpus has previously and efficiently been aligned with English. This is the case of the Europarl corpus which contains only English sentence alignments with other languages. This paper is organized in the following

[3] Ralf et al., 2006. http://ipsc.jrc.ec.europa.eu/index.php?id=198

[4] Eisele and Chen, 2010 www.dfki.de/lt/publication_show.php?id=4790

[5] www.cis.unimuenchen.de/~schmid/tools/TreeTagger/

[6] ELRA: SCI-FRES-EURADIC http://catalog.elra.info/product_info.php?cPath=42_45&products_id=668

way: In the Section 2 we describe our method. Section 3 contains the experiments for validating the method. Section 4 brings the results and the related discussions. In Section 5 we point out to conclusions and future work.

Bilingual Sentence Alignment Algorithm

This section is divided into two parts. First, we define the core algorithm and explain which type of corpus is needed in order to utilize the method. Then we provide additional details of the algorithm for implementation.

Assumptions and the Core of the Algorithm

We assume that we use a multilingual corpus which has previously been aligned with at least one language. Let's say that English is the pivot language. We want to obtain sentence alignments between any two foreign (non-English) languages of this data. Let's illustrate our method with French and Spanish. By assumption, there are available an English-French and an English-Spanish corpora, where each corpus is individually sentence aligned with English as the pivot language. Although the majority of the English sentences of both corpora are the same, not all of them need to be so. In other words, we allow for insertions, deletions and substitutions of English sentences on both sides and therefore the number of sentences in both bitexts are different. This is the case of the Europarl corpus.

Our method is very simple. It basically consists of creating a new alignment between two English corpora while keeping the reference to the original alignment information in order to map from one foreign language to the other. For instance, suppose that we need to obtain a French-Spanish sentence alignment. Since English is the common language for both English-French and English-Spanish corpora, the English texts are first aligned with each other. The original English-French and the English-Spanish alignment information is the basis for the new English-English sentence alignment to work properly.

Four cases are possible during this alignment process. First, the simplest cases consist of those sentences which are exactly the same in both corpora (one-to-one cases). Second, the first side of the corpus contains a short sentence which needs to be concatenated with one or more adjacent sentences in order to produce the same sequence of characters as the second side (many-to-one cases). Third, the first side of the corpus contains a long sentence while the second contains a short sentence which needs to be concatenated with one or more adjacent sentences in order to result in the same sequence of characters as the first side (one-to-many cases). And finally, there are cases where a sentence of a side is not a substring of the sentence from the other side or vice-versa and therefore these sentence pairs are not easily aligned (one-to-zero or zero-to-one cases).

In spite of this, we still try to find an alignment for them, given that we allow for insertions, deletions and substitutions of English sentences in the input data at both sides. In such a case, our algorithm temporarily stores the sentence positions of both unaligned sentences in order to perform the following procedures. A pointer to the sentence of the first side refers to a string that is compared with each one of the next 500 sentences of the second side. If found somewhere, an alignment is obtained and the algorithm proceeds from the next sentence position on, at both sides. Otherwise, a pointer to the sentence of the second side is used for comparison with each of the next 500 sentences of the first side. If found somewhere, an alignment is obtained and the algorithm proceeds from the next sentence position on, at both sides. However, when no alignment can be obtained after trying these thousand times, we assume there is no way of aligning those pointed sentences with any other adjacent sentence of the opposite side. Then it continues the execution of the aligner from the next sentence positions on, right after the pointers.

Note that we assume the number 500 as a generous search limit between the two texts, given that during the preparation of the Europarl corpus, each paragraph typically contained only a few sentences and the discarded utterances occurred only when the number of paragraphs in them differed in the original two languages being aligned.

During the execution of this algorithm, the history of all sentence positions having successful English-English alignments is stored. We call it ladder alignment history, making reference to the

Hunalign tool developed by Varga et al. It contains a list of pair of numbers, representing the sentence position of both English corpora having successful alignment with each other. This is the main information needed for aligning the pairs of French and Spanish sentences of our example. Note that the sentence positions on the left stand for the English corpus originally aligned with French, while the sentence positions on the right stand for the English corpus originally aligned with Spanish. Therefore, each pair of numbers represents the alignment between French and Spanish sentences. While the number of lines in the ladder alignment history represent the number of newly aligned sentences.

Also note that the sentence positions of the new alignment are relative to the original alignments in the English-French and English-Spanish parallel corpora. It means that the original alignment errors are also preserved. A new alignment error is produced whenever an x English sentence is correctly aligned with French but incorrectly aligned with Spanish or vice-versa. This is a one-to-one error type, and it is due to a single bad pre-existing alignment which is found either in the English-French or in the English-Spanish corpus. Now, let's consider the case where a y English sentence is originally misaligned with both French and Spanish at the same time. The newly produced alignment accounts for both as a single error, given that the French sentence is misaligned with a single Spanish sentence. This is a two-to-one error type.

$$\text{\#New alignment errors} \leq \sum(\text{\#alignment errors of Pivot-Foreign1}) + \sum(\text{\#alignment errors of Pivot-Foreign2}) \qquad \text{(Equation 1)}$$

It implies that the number of alignment errors produced by our algorithm is usually less than the sum of all misalignments for each original bitext. In the worst case, there is no two-to-one error type, i.e. the sentences of both parallel corpora do not contain any overlapping misalignments. In such a case, the number of new alignment errors is the sum of all misalignments present in both original corpora. This idea is expressed by Equation (1), where Pivot indicates the common language of the original alignments (i.e. English), while Foreign1 and Foreign2 represent the pair of foreign languages that our algorithm aligns, being illustrated here by the French and Spanish languages.

Additional Details of the Algorithm

1. Now that we have presented the core of our algorithm, we introduce some further details which allow our algorithm to work efficiently. When an English-English alignment is one-to-many or many-to-one, a special symbol is added in between two adjacent sentences. The amount of special symbols indicates how many short adjacent sentences are concatenated together in order to correspond to the same string of characters as the long sentence. We also store the information whether the concatenated short sentences are on the left (English-French corpus) or on the right (English-Spanish corpus), so that our algorithm can later reproduce the same number of sentence concatenations to the adjacent sentences of a corpus. This information is stored in the ladder alignment history as a pair of numbers, where the first one stands for the number of concatenated English sentences originating from the English-French corpus while the second is the number of concatenated English sentences originating from the English-Spanish corpus.

However, the ladder alignment history at this point is not yet ready. Some wrong alignment might have been introduced during the English-English sentence alignment process, which is normal for any aligner. We do here a post-processing which confirms whether every pair of aligned English sentences contains exactly the same string of characters. The wrongly aligned sentences are removed. This is the way we use for automatically validating the produced alignments. We do so by fetching the respective pair of English sentences whose indexes are present in the ladder alignment history. They are extracted from both English texts, respectively from the English-French and the English-Spanish corpora. Then, we remove the special symbols used for sentence concatenation of one-to-many or many-to-one cases in order to perform the string comparisons. Finally, we preserve only those lines containing exactly the same English sentences, and consequently producing a clean ladder alignment history.

We use the sentence alignment information present in it to obtain the aligned foreign sentence pairs. It tells the sentence index of the first foreign language which matches with the sentence index of the second one. It also tells on a sentence basis how many adjacent sentences of a corpus need to be concatenated in order to fully correspond to its translation. The work of the algorithm from this point on is basically to read the pieces of data from the following three files: ladder alignment history, first and second foreign language corpora. It combines the sentences together in order to produce the aligned parallel corpus. It finalizes the process by removing null sentence pairs and those having null translations.

Experiments

2. We want to quantify the efficiency of our algorithm to produce an aligned parallel corpus of non-English language pairs given that the sentences in both languages have been previously aligned with English. We illustrate the performance of our method by using the French and the Spanish texts from the Europarl corpus, which had been previously aligned with English on an individual basis. The English-French and the English-Spanish parallel corpora are freely available for download.

We have created a reference French-Spanish parallel corpus from the Europarl data. We extracted the first 14,941 sentences from the French corpus and the first 14,356 sentences from the Spanish corpus, totalizing 29,297 monolingual sentences. This alignment has been done in three steps. First, each corpus has been individually lemmatized by using the TreeTagger software. Second, we utilized a sentence aligner software called Hunalign to produce the sentence alignments, providing both lemmatized corpora as input and a French-Spanish bilingual dictionary of 69,231 entries. Finally, we manually revised all the automatically produced alignments by the tool. Although a considerable part of the alignments were correct, we still had to apply manual corrections on about 2,000 alignments. As result, we obtained 13,847 pairs of correctly aligned French-Spanish sentences. This is the gold data for the evaluation.

Once we have the reference alignments ready, we align the sentences based on two previous methods. For that, we utilize the Hunalign software. This tool can perform the work based only on sentence length information (6). In this case, the input data is the pair of texts to be aligned. This first method produces our baseline alignments. In addition, the software can also align sentences based on the combination of sentence length and bilingual dictionary information. In this case, the input data is the same pair of texts and a good bilingual dictionary. This second method is supposed to produce better results than the baseline.

Finally, we are ready to evaluate our algorithm. It receives as input the 14,941 non-lemmatized English sentences coming from the English-French corpus and the 14,356 non-lemmatized English sentences coming from the English-Spanish corpus. Initially, it produces 14,855 non-validated English-English sentence alignments. We call it non-validated because at this point our algorithm still needs to confirm whether every pair of aligned English sentences matches exactly the same string of characters for both corpora. After the validation process has taken place, it produces a total amount of 13,711 English-English sentence alignments in the clean ladder alignment history.

4 Results and Discussions

First, we want to check the performance of the alignment based only on sentence length information, which is our baseline. For this, we provide the Hunalign tool with 14,941 lemmatized sentences from the French corpus and the 14,356 lemmatized sentences from the Spanish corpus. Consequently, it produces 13,459 true positives out of 13,847 and 1,354 false positives. This outcome indicates a precision rate of 0.908. The number of false negatives is 388 (13,847-13,459), resulting on a recall rate of 0.972. Table 1 shows these results under the column Baseline.

Second, we want to check the performance of the alignment based on sentence length combined with bilingual dictionary information. Now, the tool receives as input the same lemmatized parallel corpus and our French-Spanish bilingual dictionary having 69,231 entries. It produces 13,704 true positives out of 13,847, while the number of false positives is 1,146. As for the precision rate, it raises to 0.923. The number of false negatives decreases to 143 (13,847-13,704) cases, producing a recall rate of 0.989. The results of this experiment are summarized on the SL+Dic column of Table 1.

Third, after obtaining the 13,711 sentence pairs described in the last paragraph of Section 3, our algorithm removes the null sentence pairs and those having null translations. Finally we obtain a French-Spanish parallel corpus having 13,640 entries. Then we compare our alignments with the reference. On the one hand, we obtain a result of 13,542 correct alignments and 98 incorrect ones. In other words, the number of true positives is 13,542 instances while the number of false positives is just 98 cases. This result indicates a very good precision rate of 0.993. On the other hand, the algorithm misses 305 (13,847-13,542) alignments that are still possible. This figure represents the instances of false negatives, which leads to a recall rate of 0.978. Table 1 contains these results under its last column.

For this particular data, the misses of correct alignments is more than 3 times the number of false positives, representing a loss rate of 2.2% of all possible correct alignments. This implies that if the size of a parallel corpus for training a statistical machine translator model is very large, the loss would be irrelevant since the amount of training data would still be very large. For such a purpose and under such conditions, an excellent precision rate is much more relevant than a perfect recall. Note that the highest possible precision rate is essential because otherwise wrong sentence alignments necessarily produce wrong word misalignments and consequently wrong translations. However when the number of wrong sentence alignments present in the parallel corpora is minimal (i.e. less than 1%), lesser will be the errors introduced to the posterior training of word alignments. In fact, good translation models depend not only on the size of a parallel corpus, but also on the high quality of the sentence alignments. In Table 1, we present the results of the evaluation by using three methods: 1) sentence length (SL) information (baseline), 2) sentence length + bilingual dictionary (SL+Dic) information and 3) our method, which is based on the high quality of existing alignments with the pivot language. Note also that the method proposed by Gale and Church, 1991 is indicated as a baseline when there is no other source of information available than the sentences themselves. However, when a good bilingual dictionary is available, an improvement is observed and the precision rate rises in 15% = (100-(1,146*100/1,354))/100 for the tested data. But an even better result is obtained when a high quality alignment has been previously performed with a pivot language. The improvement we could observe from applying our method was 92% = (100-(98*100/1,354))/100 for the tested data. This excellent result suggests that our method is efficient to transfer the original alignment information from a pair of parallel corpora sharing a common language to aligning the new pair of languages in question.

	Baseline	SL+Dic	Our method
True positives	13,459	13,704	13,542
False positives	1,354	1,146	98
False negatives	388	143	305
Precision	0.908	0.923	*0.993*
Recall	0.972	0.989	0.978

Table 1: French-Spanish sentence alignment using three methods

Conclusions and Future Work

3. A number of natural language processing applications heavily depend upon the availability of a parallel corpus. Statistical machine translation for instance requires a parallel corpus containing a huge amount of aligned sentence pairs in both languages. However, the lack of availability of almost perfectly aligned non-English parallel corpus makes unfeasible the development of such applications and researches.

Nevertheless, the relatively recent availability of the Europarl corpus which aligns English sentences with other 20 European languages has shed light on the development of our new method for obtaining such a training data. We have introduced a technique, which allows for sentence alignments of non-English texts based on the original English alignments, given a multilingual parallel corpus such as the Europarl.

Our method has been evaluated and tested against two previous methods: the first one utilizing sentence length information (baseline), while the second one, combining sentence length with bilingual dictionary information. Our method has proved to be much more efficient to align French and Spanish sentences than the other two previous methods. By applying our method, we could observe an error rate reduction of false positives of 92% in comparison with the baseline. Of course, this is due to the good quality of the original alignments, which are present in the Europarl corpus. Unfortunately, the proposed approach of aligning corpora at the sentence level cannot be applied to all sorts of bilingual data as it needs the source and target already aligned with a pivot language. This is a limitation of course, but even more limiting is when there is no reliable parallel corpus available at all for the desired language pairs.

Further work on this area stands for applying our method over all the 20 European languages of the Europarl texts. The use of our method will allow for building up to 190 new language pairs out of these corpora. We intend to develop mechanisms to process all this data and make the non-English parallel corpora available for future research and development of natural language processing applications. We hope this contribution will foster research and innovation in order to help on the development of machine translation systems for language pairs which data is not affordable or cannot be easily obtained.

Acknowledgments

This scientific work has been financed by the Prometeo Project of the Secretaría de Educación Superior de Ciencia, Tecnología e Innovación, SENESCYT of the Republic of Ecuador under the grant 20130943. We also would like to thank researcher Philipp Koehn who contributed by providing valuable information for validating the novelty of this research.

References

Brown, P.F. et al.: A Statistical Approach to Machine Translation. Computational Linguistics, 16(2), 79-85 (1990).

Davis, M., Dunning T.: A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval. Fourth Text Retrieval Conference (TREC-4). NIST (1995).

Eisele, A., Chen, Y.: MultiUN: A Multilingual Corpus from United Nation Documents. Proceedings of the Seventh conference on International Language Resources and Evaluation, Pages 2868-2872, La Valletta, Malta, European Language Resources Association (ELRA), 5/2010, www.dfki.de/lt/publication_show.php?id=4790.

ELRA: SCI-FRES-EURADIC French-Spanish Bilingual Dictionary. Catalog Reference : ELRA-M0035, http://catalog.elra.info/product_info.php?cPath=42_45&products_id=668.

EU Official Journal Multilingual Legal Text in 22 European Languages, http://apertium.eu/data

Gale, W.A., Church, K. W.: Identifying Word Correspondences in Parallel Texts. Fourth DARPA Workshop on Speech and Natural language, Asilomar, California (1991).

Gale, W.A., Church, K. W.: A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(1) (1993).

Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit (2005), www.statmt.org/europarl/

Landauer, T.K., Littman, M. L.: Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing. Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, pp. 31-38, UW Centre for the New OED and Text Research, Waterloo, Ontario (1990)

Melamed, I.D.: Word-to-word Models of Translation Equivalence. IRCS technical report #98-08, University of Pennsylvania (1998)

Melamed, I.D.: Automatic Discovery of Noncompositional Compounds in Parallel Data. Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97), Brown University (1997).

Oard, D.W.: Cross-language Text Retrieval Research in the USA. Third DELOS Workshop. European Research Consortium for Informatics and Mathematics (1997).

Ralf, S. et al. : The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, (2006), http://ipsc.jrc.ec.europa.eu/index.php?id=198.

Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994. www.cis.unimuenchen.de/~schmid/tools/TreeTagger/.

Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia (2009), http://opus.lingfil.uu.se/ECB.php

Varga, D. et al. : Parallel Corpora for Medium Density Languages. In Proceedings of the RANLP 2005, pages 590-596 (2005)