Sentence Alignment of Brazilian Portuguese and English Parallel Texts

Helena de Medeiros Caseli and Maria das Graças Volpe Nunes

NILC- ICMC- USP CP 668P, 13560-970 São Carlos, SP, Brazil
{helename,gracan}@icmc.usp.br

Abstract. Parallel texts – texts in one language and their translations to other languages – are becoming more and more available nowadays on the Web. Aligning these texts means to find some correspondence between them, in sentence level, for instance. In this paper we describe some experiments done with Brazilian Portuguese and English parallel texts using five well known sentence alignment methods. The results show that most of them performed very well on the four corpora used for testing, with 85.89%-100% of precision.

1 Introduction

Parallel texts – texts with the same content written in different languages – are becoming more and more available nowadays, mainly on the Web. These texts are extremely important for applications such as machine translation, bilingual lexicography and multilingual information retrieval. Furthermore, their importance increases considerably when correspondencies between the two halves of a bitext – source and target (source's translation) parts – are identified.

One way of identifying these correspondencies is by means of alignment. Aligning two (or more) texts means to find correspondencies (translations) between segments of the source text and segments of its translation (the target text). These segments can be the whole text or its parts such as: chapters, sections, paragraphs, sentences, words or even characters. In this paper, the focus is on sentence alignment methods.

The most frequent sentence alignment category is 1-1, in which one sentence in the source text is translated exactly to one sentence in the target text. However, there are other alignment categories, such as omissions (1-0 or 0-1), expansions (n-m, with $n < m$; $n, m \geq 1$), contractions (n-m, with $n > m$; $n, m \geq 1$) or unions (n-n, with $n \geq 1$).

In the last years, the importance of sentence aligned corpora has increased a lot due to their use in Example Based Machine Translation (EBMT) systems. In this case, parallel texts can be used by machine learning algorithms to extract translation rules ([1], [8]).

Although automatic sentence alignment is a quite approached problem, the purpose of this paper is to report the results of PESA[1] (Portuguese-English Sentence Alignment) project, which aimed to investigate, implement and evaluate some

[1] The URL for PESA project is: http://www.nilc.icmc.usp.br/nilc/projects/pesa.htm.

sentence alignment methods on Brazilian Portuguese (BP) and English parallel texts. As far as we know, PESA is the first work, in this area, involving BP and it is also a first effort to propose a new sentence alignment method.

This paper is organized as following: Section 2 gives an overview of sentence alignment methods, with special attention to those evaluated in PESA project, Section 3 describes the linguistic resources developed to support this project and Section 4 reports the results of the five sentence alignment methods evaluated on BP-English parallel corpora. In Section 5 some ideas for a new sentence alignment method are given and, in Section 6, some conluding remarks are made.

2 Sentence Alignment Methods

Parallel text alignment can be done on different levels of resolution: from the whole text to its parts (paragraphs, sentences, words, etc). In sentence alignment, given two parallel texts, a sentence alignment method try to find the best correspondencies between source and target sentences. In this process, the methods can use information about sentences' length, cognate and anchor words, POS tags, and other clues. These information stands for the methods' alignment criteria.

The sentence alignment methods evaluated in PESA project as well as their alignment criteria are shown in Table 1.

Table 1. Sentence alignment methods evaluated in PESA project and their alignment criteria

Methods	Alignment Criteria
GC ([2], [3])	Sentence length correlation
GMA ([6], [7])	Word correspondence based only on cognates
GSA+ ([6], [7])	Word correspondence based on cognates and an anchor word list[2]
Piperidis et al ([9])	Semantic load based on POS tagging
TCA ([5])	Sentence length correlation, word correspondence based on cognates, an anchor word list, etc

GC is a sentence alignment method based on a simple statistical model of sentence lengths, in characters. It relies only on the length of the two sets of sentences under consideration to determine the correspondence between them. The main idea is that longer sentences in the source language tend to have longer translations in the target language, and that shorter sentences tend to be translated into shorter ones. GC is the most referenced sentence alignment method and one with the best performance considering its simplicity.

GMA and GSA+ use a pattern recognition technique to find the alignments between sentences. Their main idea is that the two halves of a bitext - source

[2] An anchor word list is a list of words in source language and their translations in the target language. If a pair (source_word, target_word) that occurs in this list appears in the source and target sentence, respectively, it is taken as a point of correspondence between these sentences.

sentences and target sentences - are the axes of a rectangular bitext space and, in this bitext space, each token is associated with the position of its middle character. When a token at position x on the source text and a token at position y on the target text correspond to each other, it is said to be a point of correspondence (x, y).

They use two algorithms for aligning sentences: SIMR (Smooth Injective Map Recognizer) and GSA (Geometric Segment Alignment). The SIMR algorithm produces points of correspondence that are the best approximation of the true bitext maps - the correct translations - and GSA aligns the segments based on these resultant bitext maps and information about segment boundaries. The difference between GMA and GSA+ methods is that in the former, SIMR considers only cognate words to find points of correspondence, while, in the latter, a bilingual anchor word list is also considered.

The Piperidis et al's method is based on the critical issue in translation: meaning preservation. Traditionally, the four major classes of content words (or open class words) - verb, noun, adjective and adverb - carry the most significant amount of meaning. So, the alignment criterion used by this method is based on the semantic load of a sentence[3], i.e., two sentences are aligned if, and only if, the semantic loads of source and target sentences are similar.

Finally, TCA method relies on several alignment criteria to find the correspondence between source and target sentences, such as a bilingual anchor word list, words with an initial capital (candidates for proper nouns), special characters (such as question and exclamation marks), cognates and sentence length.

These methods were chosen to take part in PESA project due to some facts: a) they have different alignment criteria (as shown in Table 1); b) they are well known sentence alignment methods; c) they had shown good performance on other languages pairs. Furthermore, neither of them had already been evaluated on the specific case of BP-English parallel texts and, for this purpose, some linguistic resources, described in the next section (Section 3), had to be developed.

3 Linguistic Resources

The linguistic resources developed to support PESA project can be divided in two groups: corpora and anchor word lists[4]. For testing and evaluation purposes, three BP-English parallel corpora were built: CorpusPE, CorpusALCA and CorpusNYT.

CorpusPE is composed of 130 authentic (non-revised) academic parallel texts (65 abstracts in BP and 65 in English) in Computer Science. This corpus generated another corpus with the same 130 texts after being revised by a human translator (pre-edited corpus). They were named Authentic CorpusPE and Pre-edited CorpusPE, respectively.

Authentic CorpusPE has 855 sentences and 21432 words, while Pre-edited CorpusPE has 849 sentences and 21492 words. These two corpora were also used to

[3] Semantic load of a sentence is defined, in this case, as the union of all open classes that can be assigned to the words of this sentence.

[4] For more details of linguistic resources developed in PESA project, see http://www.nilc.icmc.usp.br/nilc/download/NILC-TR-02-07.zip (in Portuguese).

investigate the methods' behavior in texts with (Authentic CorpusPE) and without (Pre-edited CorpusPE) noise (grammatical and translation errors).

CorpusALCA, by its turn, is composed of 4 official documents of Free Trade Area of the Americas (FTAA)[5] written in BP and in English and has 725 sentences and 22069 words. Finally, CorpusNYT is composed of 7 articles in English and their translation to BP from "The New York Times"[6] journal and has 422 sentences and 10595 words.

Table 2 details the number of words in each corpus for each language (BP and English).

Table 2. Number of words per language (BP and English) in each corpus

Number of Words	Authentic CorpusPE	Pre-edited CorpusPE	CorpusALCA	CorpusNYT
BP	11349	11306	11217	5410
English	10083	10186	10852	5185
Total	**21432**	**21492**	**22069**	**10595**

These parallel corpora were chosen for two reasons: they come from different domains (scientific, law and journalistic) and have different lengths: on average, there are 7 sentences per text in CorpusPE; 91 sentences per text in CorpusALCA; and 30 sentences per text in CorpusNYT. Parallel texts' lengths influence alignment task since the greater the number of sentences, the greater will be the number of combinations among sentences to be tryed during alignment.

Test and reference corpora were built based on these four corpora (Authentic CorpusPE, Pre-edited CorpusPE, CorpusALCA and CorpusNYT) and used, respectively, to test and evaluate the methods. Text (<text> and </text>), paragraphs (<p> and </p>) and sentences (<s> and </s>) boundaries of the texts in test corpora were tagged before being aligned by the sentence alignment methods. The texts in reference corpora, besides these boundary tags, have attributes for sentence (**id**) and correspondence (**corresp**) identification in their initial sentence tag (<s>). These attributes were inserted by a semi-automatic process of sentence alignment (done by a human specialist) and are supposed to be correct, so they were used as reference in the evaluation task. These two pre-process tasks (automatic tagging of text, paragraphs and sentences boundaries and semi-automatic sentence alignment) were done using the pre-processor tool TagAlign[7].

A pair of parallel texts from the reference corpora (more specifically Pre-edited CorpusPE) is shown in Table 3 in which BP text is on the left and English text on the right. In Table 4, all alignment categories found in the four reference corpora are shown.

[5] Available in http://www.ftaa-alca.org/alca_e.asp.

[6] Available in http://www.nytimes.com and http://ultimosegundo.ig.com.br/useg/nytimes, in English and BP versions, repectively.

[7] For more details of TagAlign see http://www.nilc.icmc.usp.br/nilc/download/NILC-TR-02-09.zip (in Portuguese).

Table 3. Pair of parallel texts from the reference corpora

BP	English
<text lang=pt id=quali3R> <p><s id=quali3R.1.s1 corresp=quali3A.1.s1>Este trabalho propõe uma modelagem lingüística dos itens lexicais do português do Brasil, uma modelagem relacional e sua implementação na forma de uma Base de Dados Lexicais.</s><s id=quali3R.1.s2 corresp=quali3A.1.s2>O recurso de PLN resultante favorece padronização, centralização e reutilização dos dados, facilitando o que é considerado uma das etapas mais difíceis no processo de desenvolvimento: a aquisição de conhecimento lingüístico necessário.</s> </p> </text>	<text lang=en id=quali3A> <p><s id=quali3A.1.s1 corresp=quali3R.1.s1>This dissertation proposes a linguistic modeling of lexical items of Brazilian Portuguese, a relational modeling and its implementation in the form of a Lexical Database.</s><s id=quali3A.1.s2 corresp=quali3R.1.s2>The resulting NLP resource favors the standardization, centralization, and reuse of data, aiming at facilitating one of the most difficult stages in the development process: the linguistic knowledge acquisition.</s> </p> </text>

Table 4. All alignment categories found in reference corpora

Alignment Category	Authentic CorpusPE	Pre-edited CorpusPE	CorpusALCA	CorpusNYT
0-1 or 1-0	6	2	1	1
1-1	353	395	362	195
1-2 or 2-1	41	17	-	7
2-2	4	2	-	-
2-3	1	-	-	-
Total	**405**	**416**	**363**	**203**

Besides the corpora, other linguistic resources developed to support PESA project were an anchor word list for each corpora domain: scientific (CorpusPE), law (CorpusALCA) and journalistic (CorpusNYT), named as LPA_PE, LPA_ALCA and LPA_NYT, respectively. Table 5 presents an extract of LPA_PE in which BP words are on the left, English words on the right and the character * indicates that a suffix can be added at the end of the word.

Table 5. LPA_PE extract

BP	English
abordagem	approach
além	beyond
algoritmo	algorithm
algumas	some, several
alguns	some, several
ambient*	environment*
ambos	both
análise	analysis
ao	to the, for the, at the

4 Evaluation and Results

In this experiment, it was applied the same metrics used by Véronis and Langlais for the evaluation of sentence and word alignment methods: precision, recall and F-measure ([10]). These metrics are used to evaluate the quality of a given alignment (generated automaticaly) regarding a reference (reference corpora) by counting the number of correct alignments, as shown in (1), (2) and (3).

$$precision = \frac{NumberOfCorrectAlignments}{NumberOf\,\mathrm{Pr}\,oposedAlignments}. \tag{1}$$

$$recall = \frac{NumberOfCorrectAlignments}{NumberOf\,\mathrm{Re}\,ferenceAlignments}. \tag{2}$$

$$F = 2\frac{recall \times precision}{recall + precision}. \tag{3}$$

Methods' precision, recall and F-measure for test corpora (see Section 3) are shown in Table 6, Table 7 and Table 8, respectively. It is important to say that only GMA, GSA+ and TCA methods were evaluated on CorpusNYT, since the other two methods did not present a good performance in the previous experiments (done with the other 3 corpora).

Table 6. Methods' precision

Methods	Authentic CorpusPE	Pre-edited CorpusPE	CorpusALCA	CorpusNYT
GC	0.9125	0.9759	0.9917	-
GMA	0.9485	0.9904	0.9876	0.8788
GSA+	0.9507	0.9904	0.9876	0.8832
Piperidis et al	0.8589	0.9784	0.9833	-
TCA	0.9017	0.9420	1.0000	0.9190

Based on Table 6, it can be noticed that methods' precision are between 85.89% and 100%, and the best methods, considering this metric, were: GMA/GSA+ (Authentic and Pre-edited CorpusPE) and TCA (CorpusALCA and CorpusNYT).

Table 7. Methods' recall

Methods	Authentic CorpusPE	Pre-edited CorpusPE	CorpusALCA	CorpusNYT
GC	0.9012	0.9736	0.9890	-
GMA	0.9556	0.9928	0.8788	0.8571
GSA+	0.9531	0.9928	0.8788	0.8571
Piperidis et al	0.8716	0.9784	0.9725	-
TCA	0.9062	0.9375	1.0000	0.9507

As shown in Table 7, it can be observed that methods' recall is between 85.71% and 100%, and the best methods, considering this metric, were the same: GMA/GSA+ (Authentic and Pre-edited CorpusPE) and TCA (CorpusALCA and CorpusNYT).

Table 8. Methods' F-measure

Methods	Authentic CorpusPE	Pre-edited CorpusPE	CorpusALCA	CorpusNYT
GC	0.9068	0.9747	0.9903	-
GMA	0.9520	0.9916	0.9300	0.8678
GSA+	0.9519	0.9916	0.9300	0.8700
Piperidis et al	0.8652	0.9784	0.9778	-
TCA	0.9039	0.9398	1.0000	0.9346

In Table 8, it is possible to notice that methods' F-measure are between 86.52% and 100%, and, as expected, considering this and the other metrics, the best methods were: GMA/GSA+ (Authentic and Pre-edited CorpusPE) and TCA (CorpusALCA and CorpusNYT).

Methods' precision, recall and F-measure are graphically presented in Figure 1, Figure 2 and Figure 3, respectively.

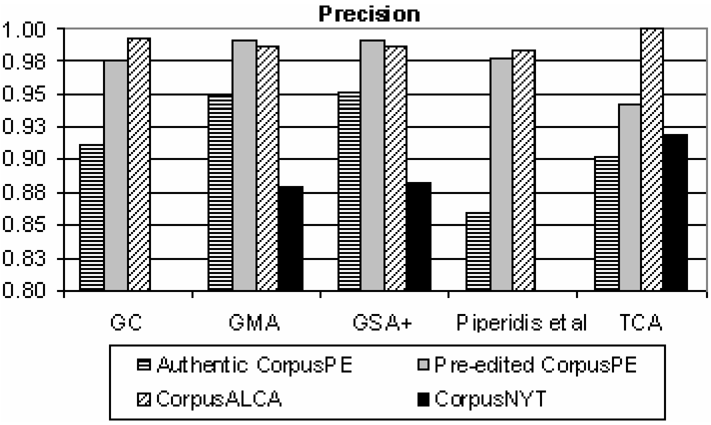

Fig. 1. Methods' precision (see Table 6)

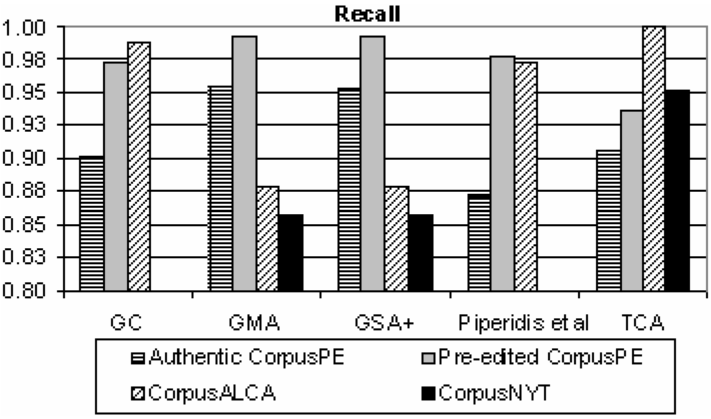

Fig. 2. Methods' recall (see Table 7)

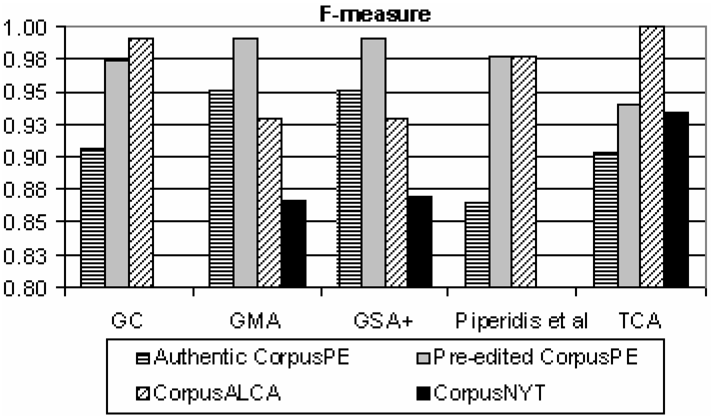

Fig. 3. Methods' F-measure (see Table 8)

Taking into account these results, it is possible to notice that all methods performed better on Pre-edited CorpusPE than on Authentic one, as already indicated in other experiments ([4]). These two corpora have some features which distinguish them apart from the other two (CorpusALCA and CorpusNYT). Firstly, the average text length (in words) in the former two is much smaller than in the latter two (BP=175, E=155 on Authentic CorpusPE and BP=173, E=156 on Pre-edited CorpusPE versus BP=2804, E=2713 on CorpusALCA and BP=772, E=740 on CorpusNYT).

Secondly, the data in CorpusPE was translated with more complex alignments than those in law and journalistic corpora. For example, CorpusPE contains six 2-2 alignments while 99.7% and 96% of all alignments in CorpusALCA and CorpusNYT, respectively, are 1-1 (see Section 3, Table 4).

In a manner of speaking, differences between Authentic/Pre-edited CorpusPE and CorpusALCA/CorpusNYT probably causes different methods' performance evaluated on these corpora.

Besides these three metrics, methods were also analyzed considering the error rate per alignment category. The major error rate was in: 2-3, 2-2 and omissions (0-1 and 1-0). The error rate in 2-3 alignments was of 100% in all methods (i.e., none of them correctly aligned the unique 2-3 alignment in Authentic CorpusPE). In 2-2 alignments, only GC and GMA didn't have 100% of error (their error rate was 83.33%).

TCA had the lower error rate in omissions (40%), followed by GMA and GSA+ (80% each), while the other methods had 100% of error on these cases. It can be noticed here that only the methods that consider cognate words as an alignment criterion had success in omissions. In [3], Gale and Church had already mentioned the necessity of considering language-specific methods to deal adequately with this category and this point was confirmed by results reported in this paper.

As expected, all methods works best on 1-1 alignments and their error rate in this category was between 2.88% and 5.52%.

5 Looking for a New Sentence Alignment Method

The work related above was the first step towards a new sentence alignment method. Although the five methods evaluated on BP-English parallel texts in PESA project had presented good scores, it is possible to change some methods parameters aiming to improve the sentence alignment precision/recall on BP-English parallel texts.

Firstly, some distinguished resources will be considered as alignment criteria: bilingual anchor word lists, special characters, cognates and sentences' length, among others. Then, we will investigate the best way to combine them using linear regression, statistics and/or machine learning algorithms.

The resulting methods will be evaluated on the four parallel corpora presented in Section 3. An environment where the user could choose an arbitrary set of parameters (resources used by the method) is also a goal for future work. Finally, other language pairs, such as BP-Spanish, will also be considered.

6 Conclusions

This paper has described some experiments done with five sentence alignment methods for BP-English parallel texts, as part of PESA project. Based on the evaluation results, we can conclude that, considering the task of sentence alignment, GMA/GSA+ performed better than the others in CorpusPE (Authentic and Pre-edited), while TCA was the best in CorpusALCA and CorpusNYT.

The obtained precision scores were all above 95%, which is the average value related in the literature. However, due to the very similar performances of the methods, at this moment it is not possible to choose one of them as the best sentence alignment method for BP-English parallel texts. More tests are necessary (and will be done) to determine the influence of alignment categories, texts' length and domain on methods' performance.

Some computational (five sentence aligners and the TagAlign) and linguistic (parallel corpora and anchor word lists) resources were developed. These resources, mainly the linguistic ones, may be used to support other projects on word alignment and machine translation.

Acknowledgments

We would like to thank Monica S. Martins for her help on developing CorpusPE; Marcela F. Fossey for her attention; CAPES and CNPq for financial support.

References

1. Carl, M.: Inducing probabilistic invertible translation grammars from aligned texts. In: Proceedings of CoNLL-2001. Toulouse, France (2001) 145–151
2. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL). Berkley (1991) 177–184
3. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. Computational Linguistics, Vol. 19(3). (1993) 75–102
4. Gaussier, E., Hull, D., Aït-Mokthar, S.: Term alignment in use: Machine-aided human translation. In: Véronis, J. (ed.): Parallel text processing: Alignment and use of translation corpora. Kluwer Academic Publishers (2000) 253–274
5. Hofland, K.: A program for aligning English and Norwegian sentences. In: Hockey, S., Ide, N., Perissinotto, G. (eds.): Research in Humanities Computing. Oxford University Press, Oxford (1996) 165–178
6. Melamed, I.D.: A Geometric Approach to Mapping Bitext Correspondence. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia, Pennsylvania (1996) 1–12
7. Melamed, I.D.: Pattern recognition for mapping bitext correspondence. In: Véronis, J. (ed.): Parallel text processing: Alignment and use of translation corpora. Kluwer Academic Publishers (2000) 25–47

8. Menezes, A., Richardson, S.D.: A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In: Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics (ACL'01). Toulouse, France (2001) 39–46
9. Piperidis, S., Papageorgiou, H., Boutsis, S.: From sentences to words and clauses. In: Véronis, J. (ed.): Parallel text processing: Alignment and use of translation corpora. Kluwer Academic Publishers (2000) 117–138
10. Véronis, J., Langlais, P.: Evaluation of parallel text alignment systems: The ARCADE Project. In: Véronis, J. (ed.): Parallel text processing: Alignment and use of translation corpora. Kluwer Academic Publishers (2000) 369–88