Evaluation of Sentence Alignment Methods on Portuguese-English Parallel Texts

Helena de Medeiros Caseli[1], Maria das Graças Volpe Nunes[1]

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

ABSTRACT

Parallel texts, i.e., texts in one language and their translations to other languages, are very useful nowadays for many applications such as machine translation and multilingual information retrieval. If these texts are aligned in sentence level, for instance, their relevance increases considerably. In this paper we describe some experiments that have being done with Portuguese and English parallel texts using five well known sentence alignment methods. Four corpora were used for testing, achieving 85.89% to 100% of precision.

KEYWORDS: Parallel texts, sentence alignment, Portuguese and English.

RESUMO

Textos paralelos – textos acompanhados de suas traduções – são muito úteis em diversas aplicações como tradução automática e recuperação de informação envolvendo várias línguas. Além disso, a relevância desses textos aumenta consideravelmente se estiverem alinhados, por exemplo, no nível sentencial. Neste artigo, são apresentados alguns experimentos realizados com textos paralelos

ISSN 0104 – 1770

escritos em Português e em Inglês e cinco métodos de alinhamento sentencial bem referenciados na literatura. Quatro corpora foram utilizados para teste alcançado uma precisão de 85,89% a 100%.

PALAVRAS CHAVE: Textos paralelos, alinhamento sentencial, português e inglês.

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 useful for applications such as machine translation, bilingual lexicography and multilingual information retrieval. Furthermore, their relevance increases considerably when correspondencies between the source and the target (source's translation) texts 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 >= 1), contractions (n-m, with n > m; n, m >= 1) or unions (n-n, with n >= 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 ((Carl 2001), (Menezes and Richardson 2001)).

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

This paper is organized as following: Section 2 presents 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 Portuguese-English parallel corpora. Finally, in Section 5 some concluding remarks are made.

2 Sentence Alignment Methods

Parallel text alignment can be done on different levels: from the whole text to its parts (paragraphs, sentences, words, etc). In sentence level, given two parallel texts, a sentence alignment method tries 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 were named: GC ((Gale and Church 1991), (Gale and Church 1993)), GMA and GSA+ ((Melamed 1996), (Melamed 2000)),

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

Piperidis et al (Piperidis, Papageorgiou, and Boutsis 2000) and TCA (Hofland 1996).

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

These methods 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[2] is also considered.

The Piperidis et al.'s method is based on a 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.

[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.

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.

The above sentence alignment methods as well as their alignment criteria are shown in Table 1.

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

Methods	Alignment Criteria
GC	Sentence length correlation
GMA	Word correspondence based only on cognates
GSA+	Word correspondence based on cognates and an anchor word list
Piperidis et al.	Semantic load based on POS tagging
TCA	Sentence length correlation, word correspondence based on cognates, an anchor word list, etc

These five methods were chosen because: a) they have different alignment criteria (as shown in Table 1); b) they are well known sentence alignment methods; and c) they had shown good performance on other languages pairs. Among these methods only TCA had already been evaluated on texts written in Portuguese (European version) with 97.1% of precision (Santos and Oksefjell 2000). Other researches on alignment involving Portuguese (also European version) are (Ribeiro, Lopes, and Mexia 2000a) and (Ribeiro, Lopes, and Mexia 2000b). Neither of them had already been evaluated on the specific case of Brazilian Portuguese

[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 (Piperidis, Papageorgiou, and Boutsis 2000)

texts and, for this purpose, some linguistic resources, described in the next section, 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 Portuguese-English parallel corpora were built: CorpusPE, CorpusALCA and CorpusNYT.

CorpusPE is composed of 130 authentic (non-revised) academic parallel texts (65 abstracts in Portuguese and 65 in English) on Computer Science. A revised (by a human translator) version of this corpora was also generated. They were named Authentic CorpusPE and Pre-edited CorpusPE, respectively.

Authentic CorpusPE has 855 sentences, while Pre-edited CorpusPE has 849 sentences. These two corpora were also used to investigate the methods behavior in texts with (Authentic CorpusPE) and without (Pre-edited CorpusPE) noise (grammatical and translation errors).

CorpusALCA is composed of 4 official documents of Free Trade Area of the Americas (FTAA)[5] written in Portuguese and in English and has 725 sentences. Finally, CorpusNYT is composed of 8 articles in English and their translation to Portuguese from the journal "The New York Times"[6] and has 492 sentences.

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

The above parallel corpora were chosen for two reasons: a) they come from different genres (scientific, law and journalistic) and b) they have different lengths - on average, there are 7 sentences per text in CorpusPE; 91 sentences per text in CorpusALCA; and

[4]For more details of linguistic resources developed in PESA project, see (Caseli and Nunes 2002) (in Portuguese).

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

[6]Available in http://www.nytimes.com (English version) and http://ultimosegundo.ig.com.br/useg/nytimes (Portuguese version).

31 sentences per text in CorpusNYT. Parallel texts' lengths have influence in alignment task since the greater the number of sentences are, the greater will be the number of combinations among sentences to be tryed during alignment.

Tabela 2. Number of words per language (Portuguese and English) in each corpus

Number of Words	Authentic CorpusPE	Pre-edited CorpusPE	CorpusALCA	CorpusNYT
Portuguese	11349	11306	11217	6023
English	10083	10186	10852	5758
Total	21432	21492	22069	11516

To test and evaluate the methods, the Test and the Reference corpora were built based on the four corpora (Authentic CorpusPE, Pre-edited CorpusPE, CorpusALCA and CorpusNYT). Texts in the Test corpora were given as input for the five sentence alignment methods. After the alignment, the aligned parallel texts were compared with the texts in the Reference corpora (correctly aligned parallel texts) to evaluate the methods.

Text (<text> and </text>), paragraphs (and) and sentences (<s> and </s>) boundaries of the texts in Test corpora were automatically 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 expert) 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 a pre-processor tool specially built for this: TagAlign (Caseli, Feltrim, and Nunes 2002).

In Table 3 all alignment categories found in the four reference corpora are shown. One can note that most of the alignments

(94%) are of type 1-1 while omissions, expansions, contractions and unions (n-n, with n > 1) are quite rare.

Tabela 3. 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	226
1-2 or 2-1	41	17	-	7
2-2	4	2	-	-
2-3	1	-	-	-
Total	405	416	363	234

Other linguistic resources developed to support PESA project were an anchor word list for each corpora genre: scientific (CorpusPE), law (CorpusALCA) and journalistic (CorpusNYT).

The anchor word lists were created from (parallel or not) texts of the same genre of the three test corpora. Firstly, the most frequent words in these texts were identified through the WordSmith tool[7] and stoplists[8]. Secondly, the lists generated by WordSmith were manually analyzed to compose pairs of words in Portuguese and English that are mutual translations. Finally, generalizations were made (indicated by *) to optimize the list. The character * indicates that a suffix can be added at the end of the word.

Table 4 presents an extract of the anchor word list built for scientific genre in which Portuguese words are on the left and English words on the right.

After building the linguistic resources presented in this section, the five sentence alignment methods were evaluated with the four parallel corpora, as described in the next section.

[7]Available in: http://www.lexically.net/wordsmith/.

[8]Lists with stopwords, i.e., closed class words (prepositions, pronouns, articles, etc). Once Portuguese/English words like "a", "an", "um"and "uma"often produce spurious points of correspondence.

Tabela 4. Extract of an anchor word list

Portuguese	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

The experiments described in this paper used the metrics precision, recall and F-measure to evaluate the sentence alignment methods. These metrics were used for the evaluation of sentence and word alignments in (Véronis and Langlais 2000) and they evaluate the quality of a given automatically generated alignment regarding a reference (from a reference corpora) by counting the number of correct alignments, as shown in (1), (2) and (3).

$$precision = \frac{NumberOfCorrectAlignments}{NumberOfProposedAlignments} \quad (1)$$

$$recall = \frac{NumberOfCorrectAlignments}{NumberOfReferenceAlignments} \quad (2)$$

$$F = 2\frac{recall \times precision}{recall + precision} \quad (3)$$

Precision, recall and F-measure for Test corpora (see Section 3) are shown in Table 5. These values are also graphically presented in Figure 1.

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).

Tabela 5. Precision, Recall and F-measure of Sentence Alignment Methods

Corpus	Metric	GC	GMA	GSA+	Piperidis et al.	TCA
Authentic	Precision	0.9125	0.9485	0.9507	0.8589	0.9017
CorpusPE	Recall	0.9012	0.9556	0.9531	0.8716	0.9062
	F-measure	0.9068	0.9520	0.9519	0.8652	0.9039
Pre-edited	Precision	0.9759	0.9904	0.9904	0.9784	0.9420
CorpusPE	Recall	0.9736	0.9928	0.9928	0.9784	0.9375
	F-measure	0.9747	0.9916	0.9916	0.9784	0.9398
	Precision	0.9917	0.9876	0.9876	0.9833	1.0000
CorpusALCA	Recall	0.9890	0.8788	0.8788	0.9725	1.0000
	F-measure	0.9903	0.9300	0.9300	0.9778	1.0000
	Precision	-	0.8788	0.8832	-	0.9190
CorpusNYT	Recall	-	0.8571	0.8571	-	0.9507
	F-measure	-	0.8678	0.8700	-	0.9346

Based on Table 5, it can be noticed that precision ranges between 85.89% and 100% and recall is between 85.71%. The best methods, considering these metrics were: GMA/GSA+ (Authentic and Pre-edited CorpusPE) and TCA (CorpusALCA and CorpusNYT).

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 (Gaussier, Hull, and Aït-Mokthar 2000). 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 (P=175, E=155 on Authentic CorpusPE and P=173, E=156 on Pre-edited CorpusPE versus P=2804, E=2713 on CorpusALCA and P=753, E=720 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 3).

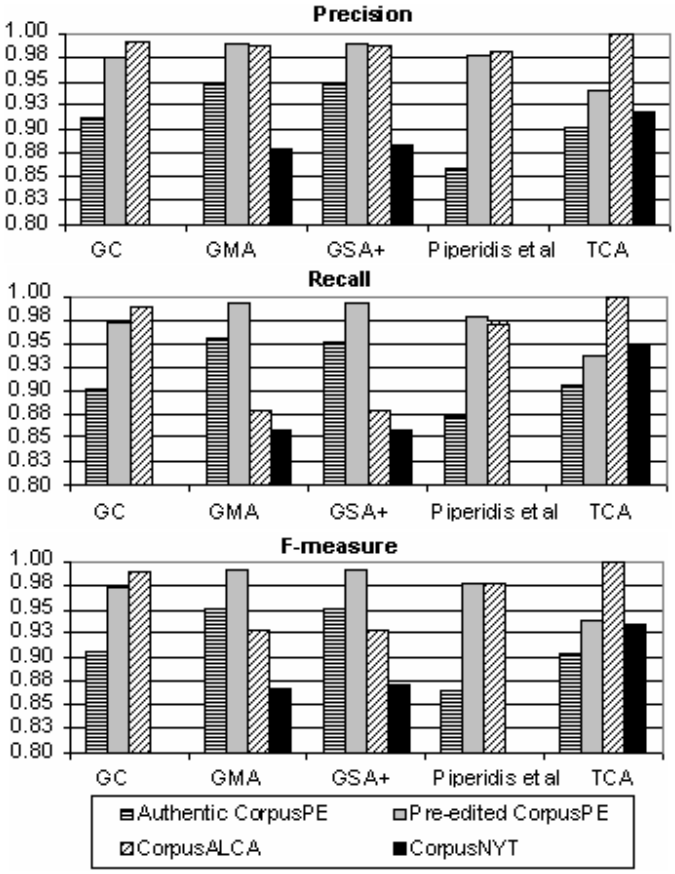

Figura 1. Precision, recall and F-measure per corpus

Differences between Authentic/Pre-edited CorpusPE and CorpusALCA/CorpusNYT probably causes different methods performance evaluated on these corpora.

Besides these three metrics, the 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 did not 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 that only the methods that consider cognate words as an alignment criterion had success in omissions. In (Gale and Church 1993), 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 the 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%.

The influence of the anchor word list on methods' performance was also investigated. As can be noticed in Table 5, the use of an anchor word list did not improve the performance of GSA+ method. This is due to the fact that GSA+ looks for candidate tokens in an anchor word list only if the matching predicate could not generate enough candidate correspondence points based only on cognates.

On the other hand, for TCA method, the quality of the anchor word list influenced considerably its performance. For example, trying to align CorpusALCA with journalistic or scientific lists the method achieved 100% of precision, but its recall decreased to 96.42%. On CorpusNYT the results were worse: using juridic list, precision and recall decreased to 71.43% and 51.28%, respectively; and using the scientific list these values decreased to 69.28% and 45.30%, respectively.

5 Conclusions

This paper has described some experiments carried out with five sentence alignment methods on Portuguese-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 and recall scores for all methods in almost all corpora are above 95%, which is the average value related in the literature (Véronis and Langlais 2000). 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 Portuguese-English parallel texts. More tests are necessary (and will be done) to determine the influence of alignment categories, texts' length and genre on methods' performance.

This work has specially contributed to researches on computational linguistic involving Brazilian Portuguese by implementing, evaluating and distributing a great number of potential resources which can be useful for important applications such as machine translation and information retrieval.

Acknowledgments

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

Referências

Carl, M. (2001). Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of CoNLL-2001*, pp. 145–151. Toulouse, France.

Caseli, H.M., V.D. Feltrim, and M.G.V. Nunes (2002). Tagalign: Uma ferramenta de pré-processamento de textos. Série de Relatórios do NILC NILC-TR-02-09,

NILC, http://www.nilc.icmc.usp.br/nilc/download/NILC-TR-02-09.zip.

Caseli, H.M. and M.G.V. Nunes (2002). A construção dos recursos lingüísticos do projeto pesa. Série de Relatórios do NILC NILC-TR-02-07, NILC, http://www.nilc.icmc.usp.br/nilc/download/NILC-TR-02-07.zip.

Gale, W.A. and K.W. Church (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 177–184. Berkley.

Gale, W.A. and K.W. Church (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics* **19**, 75–102.

Gaussier, E., D. Hull, and S. Aït-Mokthar (2000). Term alignment in use: Machine-aided human translation. In J. Véronis (Ed.), *Parallel text processing: Alignment and use of translation corpora*, pp. 253–274. Kluwer Academic Publishers.

Hofland, K. (1996). A program for aligning english and norwegian sentences. In S. Hockey, N. Ide, and G. Perissinotto (Eds.), *Research in Humanities Computing*, pp. 165–178. Oxford: Oxford University Press.

Melamed, I.D. (1996). A geometric approach to mapping bitext correspondence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1–12. Philadelphia, Pennsylvania.

Melamed, I.D. (2000). Pattern recognition for mapping bitext correspondence. In J. Véronis (Ed.), *Parallel text processing: Alignment and use of translation corpora*, pp. 25–47. Kluwer Academic Publishers.

Menezes, A. and S.D. Richardson (2001). 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 (ACL01), pp. 39–46. Toulouse, France.

Piperidis, S., H. Papageorgiou, and S. Boutsis (2000). From sentences to words and clauses. In J. Véronis (Ed.), *Parallel text processing: Alignment and use of translation corpora*, pp. 117–138. Kluwer Academic Publishers.

Ribeiro, A., G. Lopes, and J. Mexia (2000a). Linear regression based alignment of parallel texts using homograph words. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI2000)*, pp. 446–450. Berlin, Germany.

Ribeiro, A., G. Lopes, and J. Mexia (2000b). Using confidence bands for parallel texts alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, pp. 432–439. Hong Kong, China.

Santos, D. and S. Oksefjell (2000). An evaluation of the translation corpus aligner, with special reference to the language pair english-portuguese. In *Proceedings from the 12th "Nordisk datalingvistikkdager"*, pp. 191–205. Trondheim, Department of Linguistics, NTNU.

Véronis, J. and P. Langlais (2000). Evaluation of parallel text alignment systems: The arcade project. In J. Véronis (Ed.), *Parallel text processing: Alignment and use of translation corpora*, pp. 369–388. Kluwer Academic Publishers.