Evaluation of Sentence Alignment Methods for Brazilian Portuguese and English Parallel Texts

Helena de Medeiros Caseli and Maria das Graças Volpe Nunes

Núcleo Interinstitucional de Lingüística Computacional – NILC
CP 668 – ICMC-USP, 13560-970
São Carlos, SP, Brazil
{helename,gracan}@icmc.usp.br

Abstract *Parallel texts – texts in one language and their translation in other – are becoming plentiful and available nowadays on the WWW. Aligning these texts means to find the correspondences between them in sentence or word level. In this paper we describe some experiments done with two sentence alignment methods – Gale and Church's method [Gale and Church 1991], [Gale and Church 1993] and Geometric Mapping and Alignment (GMA) [Melamed 1996a], [Melamed 2000] – for Brazilian Portuguese and English parallel texts. The results show that both methods performed very well, but, as already evidenced in other experiments, GMA had a better performance with precision of 96-99%.*

1. Introduction

Parallel texts[1] – texts with the same content written in different languages – are becoming plentiful and available mainly on the WWW. These texts are extremely important for applications such as machine translation, bilingual lexicography and multilingual information retrieval. Furthermore, their importance increases considerably when the correspondences between the two halves of a bitext – source and target (source's translation) parts – are identified.

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

The most common category of sentence alignment 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 categories of alignment such as omission (1-0 or 0-1), expansion (n-m, with n, $m \geq 1$, $n < m$), contraction (n-m, with n, $m \geq 1$, $n > m$) or union (n-m, with n, $m > 1$, $n = m$).

Although automatic sentence alignment is a quite approached problem, our interest is to evaluate two well known algorithms for Brazilian Portuguese (BP) and English parallel texts and make evident (or not) the results already documented for other pairs. As far as we know, this is the first work with BP, mainly due to the lack of available parallel texts involving this language.

[1] Parallel texts are also called bitexts when only two languages are involved.

This paper is organized as following: Section 2 gives a brief overview about the related work on sentence alignment of parallel texts, Section 3 presents the methods evaluated in this experiment, Section 4 describes the evaluation and its results, and Section 5 provides a brief conclusion and some proposals for future research.

2. Related Work

The research on parallel text alignment started in the end of the 50's but just in the end of the 80's the first alignment method was proposed due to the improvement on computational store and processing.

As mentioned before, parallel text alignment can be done on different levels of resolution: from the whole text to its parts. In this paper, we focus on sentence alignment.

Since the 80's, a great number of sentence alignment methods have been proposed, most of them derived from two groups of initial studies: Gale and Church (1991, 1993) and Brown, Lai and Mercer (1991) on the one hand, and Kay and Röscheisen (1988, 1993) on the other.

The first group relied on the fact that the length of a source sentence is highly correlated with the length of its target text translation. In their turn, the second group assumed that, in order to have correspondence between sentences in a translation, their words must also correspond.

Other methods have been proposed based on the first two groups. The hybrid method presented in [Moore 2002] is based on word correspondences and also on the empirical sentence length method proposed by Brown et al (1991).

Furthermore, some new alignment methods present a novel characteristic: they try to cope with the problem of re-optimization of parameters for every new pair of languages. The re-optimization process is carried on at the same time as alignment in a many-steps process. Chuang et al, for example, proposed aligning paragraphs and from this alignment an estimative for sentence alignment is produced [Chuang et al. 2002].

The importance of sentence aligned corpora has increased a lot in the last years due to their use in example based machine translation systems (EBMT). These texts can be used by machine learning algorithms to extract translation rules [Carl 2001], [Menezes and Richardson 2001].

3. Sentence Alignment Methods

This paper describes experiments with one method of each group of the initial studies mentioned in the previous section: Gale and Church's method [Gale and Church 1991], [Gale and Church 1993], from now on GC, and Geometric Mapping and Alignment (GMA) [Melamed 1996a], [Melamed 2000]. The first relies on sentence length correlation while the second relies on lexical anchoring.

Our interest in studying and evaluating these methods is due to some facts: a) they have different alignment criteria (sentence length correlation and lexical anchoring); b) they are well known sentence alignment methods; c) they had shown good performance for other languages pairs. Furthermore, neither of them had already been evaluated for the specific case of BP-English.

GC and GMA are two of the five sentence alignment methods that will be evaluated for BP and English parallel texts in the scope of project PESA[2]. The next sections bring an overview of them.

3.1. GC Method

GC is a sentence alignment method based on a simple statistical model of sentence lengths, in characters. It is the most referenced sentence alignment method and one with the best performance compared to its simplicity.

GC 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.

In the sentence alignment task, a probabilistic score is assigned to each proposed alignment, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences.

Dynamic programming is a technique for optimizing a family of problems where global answers are constructed from successive local choices, but where an optimal purely local choice may not be part of a globally optimal answer [Campbell et al. 1998].

Figure 1 shows an extract of parallel texts aligned by GC. The number pointed by an arrow near each sentence indicates their length since this is the only alignment criterion used by this method.

BP	English
<s id=art2R.1.s1 corresp=art2A.1.s1>O crescimento do mercado de software acarreta o aumento do uso de ➡ 118 técnicas de desenvolvimento, muitas vezes informais.</s><s id=art2R.1.s2 corresp=art2A.1.s2>A manutenção de softwares torna-se problemática, uma vez que sua documentação raramente ➡ 118 reflete o código implementado.</s>	<s id=art2A.1.s1 corresp=art2R.1.s1>The growth of the software market brings about an increasing use of development ➡ 117 techniques, which are often informal.</s> <s id=art2A.1.s2 corresp=art2R.1.s2>The maintenance of software is problematic, since its documentation rarely reflects the code implemented.</s> ➡ 105

Figure 1. Example of two alignments 1-1 produced by GC.

In Figure 1, source sentences (in BP) are on the left and target sentences (in English) on the right. XML tags were included to indicate beginning (<s>) and end (</s>) of sentences. Alignment between sentences is indicated by two attributes in <s> tag: **id** and **corresp**. **id** has the sentence identification, while **corresp** has target sentences ids, which can be zero, one or more. The id is unique for each sentence in the text and is divided in three parts separated by a full-stop, they are: the name of the file where the text is stored, the number of the paragraph which contains the sentence and, finally, the character "s" followed by sentence number (its position in the paragraph).

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

The method was evaluated and the results are described in Section 4. For more details of GC method see [Gale and Church 1991] and [Gale and Church 1993].

3.2. GMA Method

GMA is a sentence alignment method which uses a pattern recognition technique to find the alignments between sentences. GMA's main idea is that the two halves of a bitext – source sentences and target sentences – are the axes of a rectangular bitext space, as shown in Figure 2. 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).

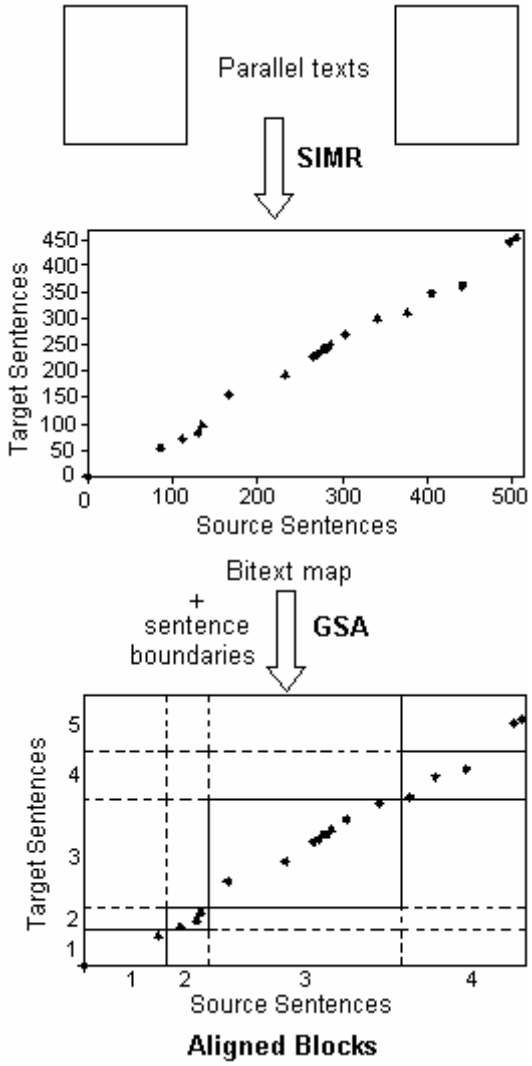

Figure 2. GMA alignment process.

GMA uses 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. In this mapping task SIMR can use many heuristics. In this experiment it uses a pattern recognition technique which is based only on cognates.

The resultant sets of points of correspondence (a bitext map) and information about segment boundary are then processed by the second algorithm, GSA, to align the segments. Figure 2 brings an overview of this process.

In Figure 2, for example, a point of correspondence inside the cell (1, 1) indicates that some token in sentence 1 in the source text corresponds to some token in sentence 1 in the target text. In other words, these sentences are correspondent. Points that are not on the main diagonal can also be found by SIMR, but they were removed in Figure 2 to avoid misunderstanding. The grid formed by the sentence boundaries over the bitext space indicates aligned blocks that are the output of GSA and, consequently, GMA.

To generate a bitext map, SIMR needs some parameters that had to be re-optimized for BP and English parallel texts, as described in [Melamed 1996b]. The re-optimization was done using simulated annealing algorithm [Vidal 1993 apud Melamed 1996b] and two bitexts manually aligned at sentence level. These bitexts used in the re-optimization process were not the same ones which were used for testing (test corpora). The construct set of parameters is shown in Table 1.

Table 1. The re-optimized set of SIMR's parameters for BP and English parallel texts.

Parameter	Value
Chain size	6
Min. cognate length ratio	4
Max. point ambiguity	0.11
Max. linear regression error	15
Max. angle deviation	0.65

The chain size parameter determines the maximum number of points of correspondence that form a chain. The minimum cognate length ratio is the threshold for the cognate metric used to find points of correspondence – the Longest Common Subsequence Ratio (LCSR). The last three parameters – maximum point ambiguity, maximum linear regression error and maximum angle deviation – are used to choose chains. The best chains are those closer to the bitext spaces's main diagonal or, in other words, that whose points are closer to the true bitext maps (see Figure 2).

BP	English
<s id=art2R.1.s1 corresp=art2A.1.s1>O crescimento do mercado de software acarreta o aumento do uso de técnicas de desenvolvimento, muitas vezes informais.</s><s id=art2R.1.s2 corresp=art2A.1.s2>A manutenção de softwares torna-se problemática, uma vez que sua documentação raramente reflete o código implementado.</s>	<s id=art2A.1.s1 corresp=art2R.1.s1>The growth of the software market brings about an increasing use of development techniques, which are often informal.</s><s id=art2A.1.s2 corresp=art2R.1.s2>The maintenance of software is problematic, since its documentation rarely reflects the code implemented.</s>

Figure 3. Example of two alignments 1-1 produced by GMA.

The same extract given in Figure 1 is shown in Figure 3, but now aligned by GMA method. Underlined words indicate the points of correspondence set by SIMR during bitext mapping.

The method was evaluated and the results are described in the next section (Section 4). For more details of GMA method see [Melamed 1996] and [Melamed 2000].

4. Evaluation and Results

The parallel corpus used for testing the two sentence alignment methods described in the previous section is composed of 65 pairs of academic parallel texts (abstracts) in Computer Science. The corpus was divided into two groups: one with 65 pairs composed of original transcriptions (authentic corpus); other with the same 65 pairs, but after a revision done by a human translator to remove grammatical and translation errors (pre-edited corpus). They were named CAT and CPT, respectively, and were used to evaluate the sentence alignment methods.

CAT has 416 BP sentences and 439 English sentences with an average of 6.4 and 6.75 sentences per file, respectively. CPT has 418 BP sentences and 431 English sentences with an average of 6.43 and 6.63 sentences per file, respectively.

This division aims to investigate the behavior of the methods in texts with (CAT) and without (CPT) noise (grammatical and translation errors) and confirm (or not) what was already revealed: "automatic sentence alignment is effective if the parallel texts are relatively clean and come from technical domains where literal translations are expected" [Gaussier et al. 2000].

Besides these two corpora, other two were built as reference to evaluate the methods. The reference corpora CAR and CPR are composed of the same parallel texts of CAT and CPT, respectively, but after a semi-automatic process of sentence alignment. They are meant to be correct aligned so they were used as reference in the evaluation task.

In this experiment, we apply the same metrics used by Véronis and Langlais in [Véronis and Langlais 2000] for the evaluation of some sentence and word alignment methods: precision, recall and F-measure given bellow as (1), (2) and (3), respectively. These metrics were used to evaluate the quality of a given alignment in regard to a reference (CAR and CPR) by counting the number of correct alignments.

$$precision = \frac{NumberOfCorrectAlignments}{NumberOf\ \mathrm{Pr}oposedAlignments} \quad (1)$$

$$recall = \frac{NumberOfCorrectAlignments}{NumberOf\ \mathrm{Re}ferenceAlignments} \quad (2)$$

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

Table 2 shows the values of these metrics for both methods on both test corpora (CAT and CPT).

Table 2. Metrics for GC and GMA methods.

	GC		GMA	
Metrics	**CAT**	**CPT**	**CAT**	**CPT**
precision	0.9125	0.9759	0.9485	0.9904
recall	0.9012	0.9736	0.9556	0.9928
F	0.9068	0.9747	0.9520	0.9916

In Table 2 we can notice that GMA performed better than GC in CAT, with 95% of F-measure against 91% in GC. In CPT, GMA also performed better but with a little advantage of 99% against 97% for GMA and GC, respectively.

Furthermore, both methods performed better on the pre-edited corpus (CPT) than on authentic one (CAT), as already evidenced in other experiments [Gaussier et al. 2000].

Although these metrics are a good measurement of methods performance, they do not take into account the fact that some alignments can be partially correct. An example of alignment partially correct is given in Table 3 and its reference alignment (totally correct) is given in Table 4.

Table 3. Example of an alignment partially correct.

BP	English
<s id=art1R.1.s4 corresp=art1A.1.s4>Também são apresentadas heurísticas para a evolução do modelo de requisitos para modelos de análise, exemplificadas através do estudo de caso apresentado.</s>	<s id=art1A.1.s4 corresp=art1R.1.s4>Heuristics to evolve from the requirements model to the analysis are also presented.</s>
	<s id=art1A.1.s5 corresp="">An example to illustrates the approach is also presented.</s>

Table 4. Example of a reference alignment.

BP	English
<s id=art1R.1.s4 corresp='art1A.1.s4 art1A.1.s5'>Também são apresentadas heurísticas para a evolução do modelo de requisitos para modelos de análise, exemplificadas através do estudo de caso apresentado.</s>	<s id=art1A.1.s4 corresp=art1R.1.s4>Heuristics to evolve from the requirements model to the analysis are also presented.</s><s id=art1A.1.s5 corresp=art1R.1.s4>An example to illustrates the approach is also presented.</s>

Arrows in Table 3 indicate the alignment between source sentence art1R.1.s4 and target sentence art1A.1.s4 and alignment between no source sentence and target sentence art1A.1.s5. While arrows in Table 4 indicate the alignment between source sentence art1R.1.s4 and target sentences art1A.1.s4 and art1A.1.s5. The correct alignment is of type (1-2) in which sentence art1R.1.s4 corresponds to sentences art1A.1.s4 and art1A.1.s5. The partial alignment is of type (1-1) and (0-1) in which sentence art1R.1.s4 corresponds only to art1A.1.s4 and the other target sentence, art1A.1.s5, doesn't correspond with any source sentence.

Partial correctness of the alignments was measured for each method and the results for GC and GMA are shown in Table 5.

Table 5. Evaluation of GC with respect to partial correctness.

	GC		GMA	
Correctness	**CAT**	**CPT**	**CAT**	**CPT**
Partially correct	6.25%	1.45%	2.94%	0.72%
Totally correct	91.25%	97.59%	94.85%	99.04%
Erroneous	2.50%	0.96%	2.21%	0.24%

From Table 5 we can notice that GC had 6.25% of CAT sentences and 1.45% of CPT sentences partially correct aligned; while GMA had 2.94% and 0.72% of CAT and CPT sentences, respectively, partially correct aligned.

Alignments were also analyzed regarding their category. GC found 1-1, 1-2, 2-1 and 2-2 alignments, while 1-0 or 0-1 alignments where totally missed (100% of error). GC's major error rate was in 2-2 alignments (83.33%) and minor in 1-1 (3.88%). GMA also found omissions (0-1 and 1-0) but the major part of them were misaligned (75%). As in GC, the GMA's major error rate was in 2-2 alignments (83.33%) and minor in 1-1 (0.80%).

5. Conclusion and Future Work

This paper has described an experiment with two sentence alignment methods for Brazilian Portuguese (BP) and English parallel texts. From the results we can conclude that GMA performed better than GC in the task of sentence alignment in both test corpora: with and without noise. This fact has just confirmed what was expected and had been described elsewhere [Melamed 2000].

Both methods had good performance, but they have problems that can be solved in order to improve accuracy. GC has a constraint of only aligning sets with the maximum of two sentences. In other words, it can produce alignments 1-2, 2-1 or 2-2; but any alignment of m-n, with m, n>2 is incorrectly aligned. In our test corpora, we have only one 2-3 alignment which was misaligned by GC. This error did not cause a huge decrease in overall performance because it was just one case. However, in a different set of texts it could represent a major problem.

GC also presented problems with omission category (0-1 or 1-0). This category had already been pointed by the authors as that in which could be necessary to consider language-specific methods in order to deal adequately with it [Gale and Church 1991] [Gale and Church 1993]. This fact was confirmed by the GMA's performance in this category. Using only cognates, GMA correctly aligned 25% of omission cases.

GMA can also be improved by using an anchor word list[3] as one of its criteria to find points of correspondence between sentences in source and target texts. Such list can be used as an alignment criterion in that way: 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. This improvement will be done as future work in project PESA.

As future work we will also test these methods using a corpus of texts in law domain. These BP and English parallel texts are documents of the Free Trade Area of

[3] An anchor word list is a list composed of words in source language and their translations to the target language.

the Americas (FTAA) available on the Internet[4] and will be used to test the sentence alignment methods in a different domain. By doing this we will be able to evaluate the methods on a more reliable basis.

Other well known sentence alignment methods such as those proposed by Piperidis et al. (2000) and Hofland (1996) are being implemented and evaluated for BP and English parallel texts and their results will be compared to those presented above. Aiming to produce a good domain-independent sentence aligner for BP-English parallel texts, some improvements for those methods will be proposed in a following step.

Acknowledgments

We would like to thank Monica S. Martins for her help on building the test corpora, Marcela F. Fossey for the text revision, and CAPES and CNPq for the financial support.

References

Brown, P. F., Lai, J. C. and Mercer, R. (1991) "Aligning sentences in parallel corpora", In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkley, p. 169-176.

Campbell, J. A., Chatterjee, N. and Dawkins, N. (1998) "Experiments in Automated Alignment of Text over Several Languages", In: Proceedings of the International Conference on Computational Linguistics, Speech and Document Processing, Indian Statistical Institute, Calcutta, C-47-C-54.

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

Chuang, T. C, You, GN; Chang, J. S. (2002) "Adaptive Bilingual Sentence Alignment", In: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA 2002), Tiburon, CA, USA, p. 21-30.

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

Gale, W. A. and Church, K. W. (1993) "A program for aligning sentences in bilingual corpora", Computational Linguistics 19(3), p. 75-102.

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

Hofland, K. (1996) "A program for aligning English and Norwegian sentences", Research in Humanities Computing, S. Hockey, N. Ide, and G. Perissinotto, Oxford, Oxford University Press, p. 165-178.

Kay, M. and Röscheisen, M. (1988) "Text-translation alignment", Technical Report, Xerox Palo Alto Research Center.

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

Kay, M. and Röscheisen, M. (1993) “Text-translation alignment”, Computational Linguistics 19(1), p. 121-142.

Melamed, I. D. (1996a) “A Geometric Approach to Mapping Bitext Correspondence”, In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, Pennsylvania, p. 1-12.

Melamed. I. D. (1996b) “Porting SIMR to New Language Pairs”, IRCS Technical Report, p. 96-26.

Melamed, I. D. (2000) “Pattern recognition for mapping bitext correspondence”, Parallel text processing: Alignment and use of translation corpora, J. Véronis, s.l., Kluwer Academic Publishers, p. 25-47.

Menezes, A. and Richardson, S. D. (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 (ACL’01), Toulouse, France, p. 39-46.

Moore, R. C. (2002) “Fast and Accurate Sentence Alignment of Bilingual Corpora”, In: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA 2002), Tiburon, CA, USA, p. 135-144.

Piperidis, S., Papageorgiou, H. and Boutsis, S. (2000) “From sentences to words and clauses”, Parallel text processing: Alignment and use of translation corpora, J. Véronis, s.l., Kluwer Academic Publishers, p. 117-138.

Véronis, J. and Langlais, P. (2000) “Evaluation of parallel text alignment systems: The ARCADE Project”, Parallel text processing: Alignment and use of translation corpora, J. Véronis, s.l., Kluwer Academic Publishers, p. 369-388.

Vidal, R. V. V, Applied Simulated Annealing, Heidelberg: Springer-Verlag, 1993.