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

Helena de Medeiros Caseli,
Aline Maria da Paz Silva, and Maria das Graças Volpe Nunes

NILC-ICMC-USP, CP 668P, 13560-970 São Carlos, SP, Brazil
{helename,alinepaz,gracan}@icmc.usp.br
http://www.nilc.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 a sentence or lexical level their relevance increases considerably. In this paper we describe some experiments that have being carried out with Brazilian Portuguese and English parallel texts by the use of well known alignment methods: five methods for sentence alignment and two methods for lexical alignment. Some linguistic resources were built for these tasks and they are also described here. The results have shown that sentence alignment methods achieved 85.89% to 100% precision and word alignment methods, 51.84% to 95.61% on corpora from different genres.

Keywords: Sentence alignment, Lexical alignment, Brazilian Portuguese

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 correspondences between the source and the target (source's translation) parts are tagged.

One way of identifying these correspondences is by means of alignment. Aligning two (or more) texts means to find correspondences (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: chapters, sections, paragraphs, sentences, words or even characters. In this paper, the focus is on sentence and lexical (or word) alignment methods.

The importance of sentence and word aligned corpora has increased mainly 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 or templates ([1], [2]).

A.L.C. Bazzan and S. Labidi (Eds.): SBIA 2004, LNAI 3171, pp. 184–193, 2004.

The purpose of this paper is to report the results of experiments carried out on sentence and lexical alignment methods for Brazilian Portuguese (BP) and English parallel texts. As far as we know this is the first work on aligners involving BP. Previous work on sentence alignment involving European Portuguese has shown similar values to the experiment for BP described in this paper. In [3], for example, the Translation Corpus Aligner (TCA) has shown 97.1% precision on texts written in English and European Portuguese.

In a project carried out to evaluate sentence and lexical alignment systems, the ARCADE project, twelve sentence methods have been evaluated and it was achieved over 95% precision while the five lexical alignment methods have achieved 75% precision ([4]).

The lower precision for lexical alignment is due to its hard nature and it still remains problematic as shown in previous evaluation tasks, such as ARCADE. Most alignment systems deal with the stability of the order of translated segments, but this property does not stand to lexical alignment due to the syntactic difference between languages[1].

This paper is organized as following: Section 2 presents an overview of alignment methods, with special attention to the five sentence alignment methods and the two lexical alignment methods considered in this paper. Section 3 describes the linguistic resources developed to support these experiments and Section 4 reports the results of the seven alignment methods evaluated on BP-English parallel corpora. Finally, in Section 5 some concluding remarks are presented.

2 Alignment Methods

Parallel text alignment can be done on different levels: from the whole text to its parts (paragraphs, sentences, words, etc). In the sentence level, given two parallel texts, a sentence alignment method tries to find the best correspondences 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 alignment criteria of these methods.

In the lexical level, the alignment can be divided into two steps: a) the identification of word units in the source and in the target texts; b) the establishment of correspondences between the identified units. However, in practice the modularization of these tasks is not quite simple considering that a single unit can correspond to a multiword unit. A multiword unit is a word group that expresses ideas and concepts that can not be explained or defined by a single word, such as phrasal verbs (e.g., "turn on") and nominal compounds (e.g., "telephone box").

In both sentence and lexical alignments the most frequent alignment category is 1-1, in which one unit (sentence or word) in the source text is translated exactly to one unit (sentence or word) 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

[1] Gaussier, E., Langé, J.-M.: Modèles statistiques pour l'extraction de lexiques bilingues. T.A.L. 36 (1–2) (1995) 133–155 apud [5].

(n-n, with n > 1). In the lexical level, categories different from 1-1 are more frequent than in the sentence level as can be exemplified by multiword units.

2.1 Sentence Alignment Methods

The sentence alignment methods evaluated here were named: GC ([6], [7]), GMA and GSA+ ([8], [9]), Piperidis et al. ([10]) and TCA ([11]).

GC (its authors' initials) is a sentence alignment method based on a simple statistical model of sentence lengths, in characters. 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 method in the literature and it presents the best performance considering its simplicity.

GMA and GSA+ methods use a pattern recognition technique to find the alignments between sentences. The 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 the position x in the source text and a token at the 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 (lexical alignments) that are the best approximation of the correct translations (bitext maps) 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 out the 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. 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 (Translation Corpus Aligner) relies on several alignment criteria to find out 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), cognate words and sentence lengths.

[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.

[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 ([10]).

2.2 Lexical Alignment Methods

The lexical alignment methods evaluated here were: SIMR ([12], [9], [13]) and LWA ([14], [15], [16]).

The SIMR method is the same used in sentence alignment task (see Section 2.1). This method considers only single words (not multiword units) in its alignment process.

The LWA (Linköping Word Aligner) is based on co-occurrence information and some linguistic modules to find correspondences between source and target lexical units (words and multiwords). Three linguistic modules were used by this method: the first one is responsible for the categorization of the units, the second one deals with multiword units using multiword unit lists and the last one establishes an area (a correspondence window) within the correspondences will be looked for.

3 Linguistic Resources

3.1 Linguistic Resources for Sentence Alignment

The required linguistic resources for sentence alignment methods can be divided into two groups: corpora and anchor word lists ([17]). For testing and evaluation purposes, three BP-English parallel corpora of different genres – scientific, law and journalistic – 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) 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, 21432 words and 7 sentences per text on average. Pre-edited CorpusPE has 849 sentences, 21492 words and also 7 sentences per text on average. These two corpora were used to investigate the methods' performance on texts with (authentic) and without (pre-edited) noise (grammatical and translation errors).

CorpusALCA is composed of 4 official documents of Free Trade Area of the Americas (FTAA)[4] written in BP and in English with 725 sentences, 22069 words and 91 sentences per text on average.

Finally, CorpusNYT is composed of 8 articles in English and their translation to BP from the journal "The New York Times"[5]. It has 492 sentences, 11516 words and 30 sentences per text on average.

To test and evaluate the methods, two corpora were built (test and reference) based on the four previous corpora. Texts in the test corpora were given as input for the five sentence alignment methods. Reference corpora – composed of correctly aligned parallel texts – were built in order to calculate precision and recall metrics for the texts of test.

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

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

The texts of test and reference corpora have been tagged to distinguish paragraphs and sentences. Tags for aligned sentences were also manually introduced in the reference corpora. A tool for aiding this pre-processing was especially implemented [18].

Most of the alignments in the reference corpora (94%), as expected, are of type 1-1 while omissions, expansions, contractions and unions are quite rare.

Other linguistic resources developed include an anchor word list for each corpus genre: scientific, law and journalistic. Examples of BP/English anchor words found in these lists are: "abordagem/approach", "algoritmo/algorithm" (in scientific list); "adoção/adoption", "afetado/affected" (in law list) and "armas/weapons", "ataque/attack" (in journalistic list).

3.2 Linguistic Resources for Lexical Alignment

The linguistic resources for lexical alignment methods can be divided into two groups: corpora and multiword unit lists.

For testing and evaluation purposes, three corpora were used: pre-edited CorpusPE[6], CorpusALCA and CorpusNYT, the same corpora built for the sentence alignment task (see Section 3.1). Texts in the test corpora were automatically tagged with word boundaries and reference corpora were also built with alignments of words and multiwords.

Multiword unit lists contain the multiwords that have to be considered during the lexical alignment process. For the extraction of these lists, were used the following corpora: texts on Computer Science from the ACM Journals (704915 English words); academic texts from Brazilian Universities (809708 BP words); journalistic texts from the journal "The New York Times" (48430 English words and 17133 BP words) and official texts from ALCA documentation (251609 English words and 254018 BP words).

The multiword unit lists were built using automatic extraction algorithms followed by a manual analysis done by a human expert. The algorithms used for automatic extraction of multiword units were NSP (N-gram Statistic Package)[7] and another which was implemented based on the Mutual Expectation technique [19]. Through this process, three lists (for each language) were generated by each algorithm and the final English and BP multiword lists have 240 and 222 units respectively.

Some examples of multiwords in these lists are: "além disso", "nações unidas" and "ou seja" for BP; "as well as", "there are" and "carry out" for English[8].

4 Evaluation and Results

The experiments described in this paper used the precision, recall and F-measure metrics to evaluate the alignment methods. Precision stands for the number of

[6] It is important to say that CorpusPE was evaluated with 64 pairs rather than 65 because we note that one of them was not parallel at lexical level.

[7] Available in http://www.d.umn.edu/ tdeperse/code.html.

[8] For more details of automatic extraction of multiword units lists see [20].

correct alignments per the number of proposed alignments; recall stands for the number of correct alignments per the number of alignments in the reference corpus; and F-measure is the combination of these two previous metrics [4].

The values for these metrics range between 0 and 1 where a value close to 0 indicates a bad performance of the method while a value close to 1 indicates that the method performed very well.

4.1 Evaluation and Results of Sentence Alignment Methods

Precision, recall and F-measure for each corpus of test corpora (see Section 3.1) are shown in Table 1.

Table 1. Precision, Recall and F-measure of Sentence Alignment Methods

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

It is important to say that only GMA, GSA+ and TCA methods were evaluated on CorpusNYT because this corpus was evaluated later and only the methods which had had better performance where considered in this last experiment.

It can be noticed that precision ranges between 85.89% and 100% and recall is between 85.71% and 100%. The best methods considering these metrics were GMA/GSA+ for CorpusPE (authentic and pre-edited) and TCA for CorpusALCA and CorpusNYT.

Taking into account these results, it is possible to notice that all methods performed better on pre-edited CorpusPE than on the authentic one, as already evidenced by other experiments [21]. These two corpora have some features which distinguish them from the other two. 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, texts in CorpusPE have 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.

These differences between authentic/pre-edited CorpusPE and CorpusALCA /CorpusNYT probably causes the differences in methods' performance on these corpora. It is important to say that text lengths affected the alignment task since the greater the number of sentences are, the greater will be the number of combinations among sentences to be tried during alignment.

Besides the three metrics, the methods were also evaluated by considering the error rate per alignment category. The major error rate was in 2-3, 2-2 and omissions (0-1 and 1-0) categories. 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, the error rate for GC and GMA was 83.33% while for the remaining methods it was 100%.

TCA had the lowest error rate in omissions (40%), followed by GMA and GSA+ (80% each), while the other methods had 100% of error in this category. It can be noticed that only the methods that consider cognate words as an alignment criterion had success in omissions. In [7], Gale and Church had already mentioned the necessity of considering language-specific methods to deal adequately with this alignment category and this point was confirmed by the results reported in this paper.

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

4.2 Evaluation and Results of Lexical Alignment Methods

Precision, recall and F-measure for each corpus of test corpora (see Section 3.2) are shown in Table 2.

SIMR method had a better precision (91.01% to 95.61%) than LWA (51.84% to 62.15%), but its recall was very low (16.79% to 20%) what can be a problem for many applications such as bilingual lexicography. The high precision, on the other hand, can be explained by its very accurate alignment criterion based only on cognate words.

LWA had a better distribution between precision and recall: 51.84% to 62.15% and 59.38% and 65.14% respectively. These values are quite different from that obtained in an experiment carried out on English-Swedish pair in which LWA has achieved 83.9% to 96.7% precision and 50.9% to 67.1% recall ([15]) but are close to that obtained in another experiment carried out on English-French pair in which LWA has achieved 60% precision and 57% recall ([4]). So, for languages with common nature like French and BP the values were very close.

The LWA's partially correct link proposals were also evaluated using the metrics proposed in [22]. With these metrics precision improved 12% to 16% (from 51.84%–62.15% considering only totally correct alignments to 66.87%–74.86% considering also partially correct alignments) while recall improved almost 1% (from 59.38%–65.14% to 59.81%–65.82% considering totally and partially correct alignments respectively).

Table 2. Precision, Recall and F-measure of Lexical Alignment Methods

Corpus	Metric	SIMR	LWA
Pre-edited CorpusPE	Precision	0.9383	0.5888
	Recall	0.1832	0.6514
	F-measure	0.3065	0.6185
CorpusALCA	Precision	0.9561	0.6215
	Recall	0.2000	0.5983
	F-measure	0.3308	0.6097
CorpusNYT	Precision	0.9101	0.5184
	Recall	0.1679	0.5938
	F-measure	0.2835	0.5535

5 Some Conclusions

This paper has described some experiments carried out on five sentence alignment methods and two lexical alignment methods for BP-English parallel texts.

The obtained precision and recall values for all sentence alignment methods in almost all corpora are above 95%, which is the average value related in the literature [4]. 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 the alignment categories, the text lengths and genre on methods' performance.

For lexical alignment, SIMR was the method that presented the best precision, but its recall was very low and it does not deal with multiwords. LWA, on the other hand, achieved a better recall and it is able to deal with multiwords, but its precision was not so good as SIMR's one. Considering multiword units, the literature has not yet established an average value for precision and recall, but it has been clear and this work has stressed that corpus size and the pair of language have great influence on the aligners' performance ([15], [4]).

The results for sentence alignment methods have stressed the values related in the literature while the results for lexical alignment methods have demonstrated that there are still some improvement to be achieved.

In spite of this, 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 FAPESP, 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. 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
3. Santos, D., Oksefjell, S.: An evaluation of the translation corpus aligner, with special reference to the language pair English-Portuguese. In: Proceedings of the 12th "Nordisk datalingvistikkdager", Trondheim, Departmento de Lingüística, NTNU (2000) 191–205
4. 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) 369–388
5. Kraif, O.: From translation data to constrative knowledge: Using bi-text for bilingual lexicons extraction. International Journal of Corpus Linguistic **8:1** (2003) 1–29
6. 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
7. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. Computational Linguistics **19** (1993) 75–102
8. 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
9. 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
10. 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
11. Hofland, K.: A program for aligning English and Norwegian sentences. In Hockey, S., Ide, N., Perissinotto, G., eds.: Research in Humanities Computing, Oxford, Oxford University Press (1996) 165–178
12. Melamed, I.D.: A portable algorithm for mapping bitext correspondence. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. (1997) 305–312
13. Melamed, I.D., Al-Adhaileh, M.H., Kong, T.E.: Malay-English bitext mapping and alignment using SIMR/GSA algorithms. In: Malaysian National Conference on Research and Development in Computer Science (REDECS'01), Selangor Darul Ehsan, Malaysia (2001)
14. Ahrenberg, L., Andersson, M., Merkel, M.: A simple hybrid aligner for generating lexical correspondences in parallel texts. In: Proceedings of Association for Computational Linguistics. (1998) 29–35
15. Ahrenberg, L., Andersson, M., Merkel, M.: A knowledge-lite approach to word alignment. In Véronis, J., ed.: Parallel text processing: Alignment and use of translation corpora. (2000) 97–116

16. Ahrenberg, L., Andersson, M., Merkel, M.: A system for incremental and interactive word linking. In: Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas (2002) 485–490
17. Caseli, H.M., Nunes, M.G.V.: 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 (2002)
18. Caseli, H.M., Feltrim, V.D., Nunes, M.G.V.: 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 (2002)
19. Dias, G., Kaalep, H.: Automtic extraction of multiword units for Estonian: Phrasal verbs. In Metslang, H., Rannut, M., eds.: Languages in Development. Number 41 in Linguistic Edition, Lincom-Europa, München (2002)
20. Silva, A.M.P., Nunes, M.G.V.: Extração automática de multipalavras. Série de Relatórios do NILC NILC-TR-03-11, NILC, http://www.nilc.icmc.usp.br/nilc/download/NILC-TR-03-11.zip (2003)
21. 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
22. Ahrenberg, L., Merkel, M., Hein, A.S., Tiedemann, J.: Evaluation of word alignment systems. In: Proceedings of 2nd International Conference on Language Resources & Evaluation (LREC 2000). (2000) 1255–1261