UNIVERSIDADE DE SÃO PAULO

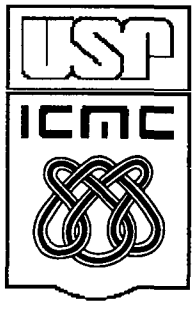
Instituto de Ciências Matemáticas e de Computação

Using Multiword Lists for Lexically Aligning Brazilian Portuguese and English Texts

Aline M. da Paz Silva
Maria das Graças Volpe Nunes

N° 77

NOTAS



São Carlos - SP

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2577

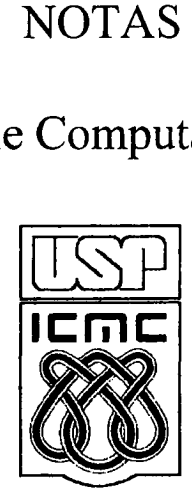
Using Multiword Lists for Lexically Aligning Brazilian Portuguese and English Texts

Aline M. da Paz Silva
Maria das Graças Volpe Nunes

Nº 77

NOTAS

Série Computação



São Carlos – SP
Mar./2004

Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Using Multiword Lists for Lexically Aligning Brazilian Portuguese and English Texts

Aline M. da Paz Silva

Maria das Graças Volpe Nunes

NILC-TR-04-02

Março 2004

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Using Multiword Lists for Lexically Aligning Brazilian Portuguese and English texts

Aline Maria da Paz Silva
Maria das Graças Volpe Nunes

NILC-ICMC-USP, CP 668P, 13560-970 São Carlos, SP, Brazil
{alinepaz, gracan}@icmc.usp.br
http://www.nilc.icmc.usp.br

Abstract. Multiword unit treatment has been receiving considerable amount of attention. There is, nowadays, an increasing interest in the evaluation of multiword units in real-world applications. In this paper we describe some experiments that have been carried out with Brazilian Portuguese and English parallel texts by the use of well known lexical alignment methods. We are interested in determining the influence of the use of multiwords lists on the performance of the lexical alignment methods.

Keywords: Lexical alignment, parallel corpus, multiword extraction, evaluation.

1. Introduction

In order to translate and analyze texts in a language it is necessary to take into account the fact that these texts are not composed only of single words, but also have some complex units: multiword units, or simply multiwords - word groups that express ideas and concepts that can not be explained or defined by a single word. Multiwords may be, for example, phrasal verbs (e.g., turn on) or nominal compounds (e.g., telephone box). Thus, some tools have been developed to identify and extract multiwords from electronic corpora.

One of the challenges involved in lexical alignment is the treatment of multiword units, which, once detected, should be aligned as if they were single units. One way to do this is through the use of predefined lists of multiwords – one for each language involved – which are searched for by the alignment algorithm. These lists are often domain-dependent and can be either manually or automatically extracted from corpora (one for each language). Manual extraction is often subjective and time consuming. Automatic extraction of multiword units has been the focus of several works and some empiric and hybrid methods have been proposed and evaluated for various languages. As far as we know, this is the first work on evaluation of multiword extraction involving Brazilian Portuguese (BP). An alternative way to deal with multiwords is to provide the alignment algorithm with mechanisms to detect them during the alignment process. This method is more independent on corpora but requires more linguistic information than the former, which can consider only frequency information.

In this work we are interested in determining the influence of the use of predefined multiword lists on the performance of the lexical alignment of BP and English parallel texts.

2. Linguistic Resources

Linguistic resources for lexical alignment methods can be divided into (i) test corpus, (ii) reference corpus, (iii) closed-class unit lists and (iv) multiword lists.

The test corpus (corpus *PE*) is composed of texts used as input for the alignment methods. On the other hand, the reference corpus is composed of aligned texts (manually or with some automatic tool) by a human translator. This corpus is used for comparison with the alignments proposed by the alignment methods.

The corpora, both test and reference corpus, are composed of 65 BP-English pairs of parallel texts (Martins, Caseli & Nunes, 2001) with some correction done by a human translator. These modifications have been done to extract from texts possible ambiguities, mistakes, grammatical and translations errors.

Closed-class unit lists contain grammatical words, namely articles, prepositions, pronouns and conjunctions. For BP, the list was generated from the lexical database Diadorim[1]. For English, the lists were obtained via the SIMR algorithm.

Multiword lists contain those multiwords to be considered during the alignment process. For the task of domain-dependent multiword extraction, two corpora of texts in Computer Science in English and BP were compiled. The English portion (704.915 words) contains texts from the ACM Journals. The BP portion (809.708 words) contains academic texts from Brazilian Universities. Two additional corpora with journalistic texts were also compiled in order to extract domain-independent, colloquial multiword units.

3. Alignment Methods

On the lexical level, the alignment can be divided into two steps: (a) the identification of lexical units (words and multiwords) in both texts of each parallel pair and (b) the establishment of correspondences between the identified units. However, in practice, the modularization of these tasks is not usually simple considering that a single unit can correspond to a multiword unit.

Two alignment methods were selected as a basis for the implementation of our lexical aligners: SIMR (Melamed, 1997, 2000; Melamed, Al-Adhaleh & Kong, 2001) and LWA (Ahrenberg, Andersson & Merkel, 1998, 2000, 2002). SIMR is based on pattern recognition and uses cognate measures to establish correspondences between the units of bitexts. This method only considers single words. LWA uses co-occurrence (frequency of a pair of units in bitexts) statistical measures and knowledge-lite linguistic modules to determine correspondences between bitext units. Unlike SIMR, this method has a module for treating multiword units whose lists are generated in a preprocessing phase.

4. Multiword units extraction

For the purpose of lexical alignment, it is necessary to find ways to identify the units that cannot be aligned word-by-word, i.e., multiword units. One of the ways to treat these units is using multiword lists to be consulted during the alignment process and help the identification of multiwords inserted in the corpus. These lists can be manually or automatically extracted from corpora.

In the automatic extraction task we experimented with two techniques: the Mutual Expectation (Dias & Kaalep, 2002) and NSP packages[2]. Four multiword lists (SE – Specific English, SBP – Specific Brazilian Portuguese, GE – General English and GBP –

[1] Available at http://nilc.icmc.usp.br/nilc/tools/intermed.htm

[2] Available at http://www.d.umn.edu/ tdeperse/code.html

General Brazilian Portuguese) were generated by each technique. Table 1 and Figure 1 show the contrast between the numbers of units in the lists generated.

Table 1 - Number of units in the lists generated by the extraction techniques.

Corpus	Mutual Expectation	NSP
SE	8694	30516
SP	12197	90503
GE	17226	73938
GBP	23196	185650

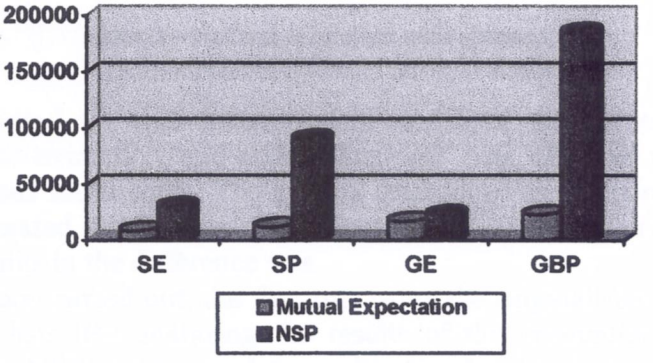


Figure 1 – Number of units generated by multiword extraction techniques.

Analyzing the lists, it was possible to observe that several units are not relevant units; these units should not be considered as multiwords. So, it was necessary to develop an elimination task to ensure that only units actually satisfying multiword properties would remain in the lists.

Multiwords are defined as word groups whose meaning or connotation cannot be computed from its components. Thus, some criteria were established to consider a unit as valid multiword. These criteria are: non-compositionality, non-substitutability and non-modifiability (Schone & Jurafsky, 2001). The entire elimination task was manually carried out in a post-processing phase.

Following these criteria, after elimination eight new lists were generated. Table 2 shows the number of units for each new list, which is usually twice greater for NSP. Figure 2 allows visualizing this difference.

Table 2 - Number of units in the lists obtained after the elimination process.

Corpus	Mutual Expectation	NSP
SE	140	262
SBP	153	260
GE	348	645
GBP	1115	1556

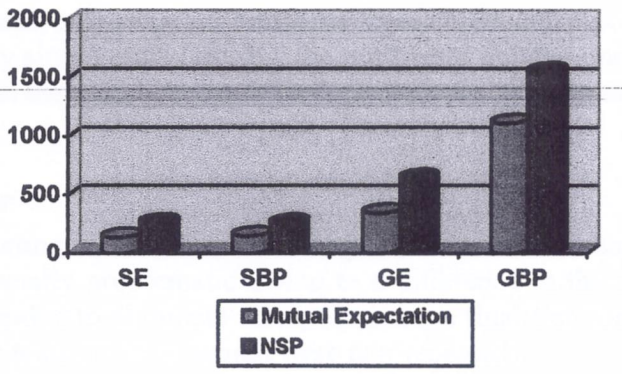


Figure 2 – Number of relevant units obtained.

We thus finished up with sixteen multiword lists, namely eight original lists, obtained by automatic extraction, and eight "skimmed" lists, obtained thereof via manual elimination of spurious multiwords. The criteria established to compare the lists were (i) number of units generated, (ii) number of eliminated units, (iii) number of coincident units and (iv) number of units in the difference sets.

Two tasks were carried out, the first one using the original lists and the other one using the skimmed lists. By analyzing the results of this evaluation, we came to the conclusion that the NSP package, despite its simpler heuristic, obtained a better performance than Mutual Expectation, which produced half as many relevant multiwords[3].

Finally, in the manual extraction task, two multiword lists were built manually by an expert, one for each language. The BP and English lists contained 304 and 297 multiwords, respectively.

5. Evaluation of lexical aligners

The lexical aligners were evaluated as to precision, recall and F-measure. Precision means the proportion of correct alignments in all those generated, recall means the number of correct alignments divided by the number of expected correct alignments, even if not generated, and F-measure is the combination of these two metrics (Verónis & Langlais, 2000).

5.1. SIMR aligner

In order to evaluate the SIMR aligner the test corpus used was corpus *PE*, with 11341 words in source texts and 10217 words in target texts. The number of proposed alignments was 2046. The results of the evaluation are shown in Table 3.

Table 3 - SIMR performance

	Recall	Precision	F-measure
SIMR	0.1921	0.8670	0.3145

[3] For more information consult (Silva, 2004) (in Brazilian Portuguese).

As we can observe, this aligner has a very low recall value. This can be explained by two facts: 1) it uses only cognate measures to establish correspondences; and 2) this aligner considers only single units (words), the multiword units are not taken into account, generating mistakes in the establishment of correspondence between the units in the parallel texts.

5.2. LWA aligner

During the first experiments with the LWA aligner, we observed that 19 parallel pairs in corpus *PE* were especially problematic owing to a difference in the number of sentences. Consequently, we decided to eliminate them, obtaining a final test corpus of 46 pairs, with 7327 words in source texts (1718 of which were non-repeated).

The LWA aligner was used in four different alignment tasks, one for each multiword list automatically extracted by the Mutual Expectation and NSP packages. LWA's performance is shown in Table 4.

Table 4 - LWA performance using automatically extracted lists.

	Mutual Expectation		**NSP**	
	SBP/SE	**GBP/GE**	**SBP/SE**	**GBP/GE**
Recall	0.3865	0.4150	0.4424	0.4476
Precision	0.3394	0.3576	0.3886	0.3854
F-measure	0.3614	0.3841	0.4137	0.4142

It is possible to observe that all figures are too low, the best of which were nevertheless obtained using the lists generated with the NSP package. We also carried out an evaluation task using the manually-extracted multiword lists. The results are shown in Table 5.

Table 5 - LWA performance using manually extracted lists.

	Recall	**Precision**	**F-measure**
LWA	0.5193	0.4818	0.4998

Unlike an evaluation reported on in (Ahrenberg, Andersson & Merkel, 2000), in which LWA presented good precision (between 83.9% and 96.7%) and lower recall (between 50.9% and 67.1%), our results were really not good in this experiment. So we decided to develop a task to identify the factors hampering LWA's performance.

It is important to say that the LWA aligner was also evaluated in ACADE project, along with other four systems (CEA, LILA, RALI, XEROX), the best precision obtained being 75% (Verónis & Langlais, 2000). In this task, LWA's performance was 60% precision, 57% recall and 58% F-measure. These numbers hint at the prominence of language in text alignment: when the language pair used was English and French (the latter being a Latin language, like Brazilian Portuguese), LWA's performance was somewhat worse than when pairing English and Swedish, both Germanic languages.

Other experiments were next carried out in which only the manually-extracted multiword lists were used. Some points were emphasized, namely (i) the influence of non-literal translations, (ii) corpus size, and (iii) influence of automatic POS tagging.

Through these experiments, we have concluded that using a subset of corpus *PE* composed of 10 pairs of parallel texts with more "literal" translations, recall increased from 51.23% to 57.82%, but precision decreased from 48.18% to 47.56%. As can be noticed, removing texts with freer translations did not improve precision. On the other hand, changes in corpus size influenced performance. With a 47% increase in corpus size, LWA's precision and recall respectively rose to 51.48% and 55.20% from their original figures of 48.18% and 51.23%. This improvement of almost 3% in precision and 4% in recall is due to one of LWA's alignment criteria: co-occurrence of word pairs. These results are close to those of an experiment carried out for French, and we believe that a significant increase in corpus size in future experiments will narrow the difference further still due to the common nature of BP and French.

Finally, we addressed the influence of automatic tagging in alignment. As it was necessary to have a correctly-tagged test corpus, we built a smaller subcorpus, with 10 pairs of parallel texts (the same used in the evaluation of the influence of non-literal translations) manually checked on POS tagging. If we were to observe an influence of tagging precision in this preliminary experiment, we intended to verify the tagging of the rest of our corpus. The results showed that recall increased from 51.23% to 57.92%, with a slight decrease in precision from 48.18% to 47.59%. These values are almost the same achieved in the first experiment of this investigation. We concluded that our current POS tagging precision was not a significant hindrance to our aligners.

Based on the results obtained in this investigation, it was possible to identify two aspects that have a significant influence on LWA's performance: multiword automatic extraction and corpus size. The precision of automatic multiword extraction directly reflects on the final precision, and increasing corpus size leads to better precision.

6. Conclusions

Despite the fact that our evaluation experiments were carried out with small corpora, the values of precision for LWA were much lower than that achieved by SIMR, which does not use multiword lists. However, SIMR's recall score is far below LWA's (also low) score. This is certainly due to the use of multiword lists. Moreover, it is worth noticing the improvement of these figures caused by the use of better-quality (manually-extracted) lists. Indeed, the low performance of the multiword extraction algorithms for these corpora had strongly and negatively affected the alignment algorithms.

SIMR showed best precision, but its recall was very low and its algorithm does not deal with multiwords. LWA, on the other hand, achieved best recall and is able to deal with multiwords, but its precision was not so good as SIMR's. Considering multiword units, the literature has not yet established standards for recall and precision, but it is clear by now, which this work makes a point of stressing, that corpus size and the considered language pair have great influence on alignment performance.

References

G. Dias, H. Kaalep, Automatic Extraction of Multiword Units for Estonian: Phrasal Verbs. *In*: H. Metslang, M. Rannut (eds.) Languages in Development, Linguistic Edition 41, Lincom-Europa, München (2002).

I. D. Melamed, A portable algorithm for mapping bitext correspondence. *In*: Proceeding of the 35th Annual Meeting of the Association for Computational Linguistics (1997), 305-312.

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

I. D. Melamed, M. H. Al-Adhaileh, T. E. Kong, 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).

J. Verónis, P. Langlais, Evaluation of parallel texts alignment systems: *The ARCADE project*. *In*: Véronis, J., ed. *Parallel text processing: Alignment and use of translation corpora*. s.l.: Kluwer Academic Publishers (2000), 369-388.

L. Ahrenberg, M. Andersson, M. Merkel, A simple hybrid aligner for generating lexical correspondences in parallel texts. *In*: Proceedings of Association for Computational Linguistics (1998), 29-35.

L. Ahrenberg, M. Andersson, M. Merkel, A knowledge-lite approach to word alignment. *In*: Véronis, J., ed. *Parallel text processing: Alignment and use of translation corpora*. s.l.: Kluwer Academic Publishers (2000), 97-116.

L. Ahrenberg, M. Andersson, M. Merkel, A System for Incremental and Interactive Word Linking. *In*: Third International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas, 29-31 May (2002), 485-490.

M. S. Martins, H. M. Caseli, M. G. V. Nunes, A construção de um corpus de textos paralelos inglês-português. Série de relatórios do NILC. NILC-TR-01-5 (2001), 62p.

P. Schone, D. Jurafsky, "Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?". *In*: Proceedings of the 2001 Empirical Methods in Natural Language Processing (EMNLP2001), Pittsburgh, PA (2001), 100-108.

NOTAS DO ICMC

SÉRIE COMPUTAÇÃO

076/2004 BÍSCARO, H.H; CASTELO FILHO, A.; NONATO, L.G. – A topological approach to curve reconstruction from scattered points.

075/2004 NONATO, L.G.; CASTELO FILHO, A.; CAMPOS, J.E.P.P.; BÍSCARO, H.; MINGHIM, R. – Topological tetrahedron characterization with applications in volumetric reconstruction.

074/2003 TOMÉ, M.F.; CUMINATO, J.A.; CASTELO, A.; FERREIRA, V.G.; McKEE, S. - Recent Developments in the MAC Technique.

073/2003 SOUZA, F.S.; MANGIAVACCHI, N.; NONATO, L.G.; CASTELO FILHO, A.; TOMÉ, M.F.; FERREIRA, V.G.; CUMINATO, J.A.; McKEE, S. – A front-tracking finite difference method to solve the 3D navier-stokes equations for multi-fluid flows with free surfaces.

072/2003 MANGIAVACCHI, N.; CASTELO FILHO, A.; TOMÉ, M.F.; CUMINATO, J.A.; OLIVEIRA, M.L.B.; McKEE, S. – A numerical technique for incluiding surface tension effects for axisymmetric and planar flows using the gensmac method.

071/2003 TOMÉ, M.F.; GROSSI, L.; CASTELO, A.; CUMINATO, J.A.; MANGIAVACCHI, N.; FERREIRA, V.G.; SOUSA, F.S.; McKEE, S. – A numerical method for solving three-dimensional generalized Newtonian free surface flows.

070/2003 SANTOS, F.L.P.; MANGIAVACCHI, N.; CASTELO, A.; TOMÉ, M.F.; CUMINATO, J.A. – A novel technique for free surface 2D multiphase flows.

069/2003 VIANNA, A.C.G.; ARENALES, M.N.; GRAMANI, M.C.N. – Two-stage and constrained two-dimensional guillotine cutting problems.

068/2003 ARAUJO, S.A.; ARENALES, M.N.; CLARK, A.R. – A lot-sizing and scheduling problem in a foundry.

067/2003 ARAUJO, S.A.; ARENALES, M.N. – Dimensionamento de lotes e programação do forno numa fundição automatizada de porte médio.