Automatic Alignment of Common Information in Comparable Sentences of Portuguese

Eloize Rossi Marques Seno, Maria das Graças Volpe Nunes
NILC-ICMC – University of São Paulo
São Carlos – SP, Brazil
Caixa Postal 668 - 13560-970
+55 16 33739700
{eloize,gracan}@icmc.usp.br

ABSTRACT

The ability to recognize distinct word sequences which refer to the same meaning is of extreme relevance for many applications in NLP, such as automatic summarization, question answering, generation, etc. In this paper we describe our first attempt at aligning common information between portuguese similar sentences. We propose a method based on lexical and syntatic information and some paraphrase rules to find different strings with the same meaning. A preliminary experiment suggests that the method has potential for identifying strings which are semantically related but lexically different, as is the case of lexical paraphrases.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing – *language parsing and understanding*.

General Terms

Languages.

Keywords

Common Information Alignment, Paraphrase Rules, Comparable Sentences.

1. INTRODUCTION

Recognizing distinct word sequences which refer to the same meaning (e.g. synonyms and paraphrases) is of the greatest importance in many Natural Language Processing (NLP) tasks, such as machine translation [9], automatic summarization [1] [2], question answering [7], generation [11], among others. In this work, aligning of common information is a vital step towards generating novel sentences by fusing similar information shared by a set of comparable sentences.

In the last years, several approaches for the alignment of common information have been proposed in the literature for foreign languages (e.g. [2], [7], [8], [9], [11]). Such approaches may be distinguished into two categories, depending on the type of sentences used: a) comparable sentences, which come from different source on the same event and b) parallel sentences, which are multiple translations of the same source. In [9], for example, a model has been proposed for the alignment of parallel sentence pairs at the level of syntatic trees. In such work, words of the same part-of-speech are considered paraphrases. Althougth that approach has provided quite satisfactory results in the context of which it has been proposed[1], it does not address the problem of recognizing paraphrases between phrases (e.g. *capital paulista* and *capital de São Paulo*)[2].

Unlike [9], some researches on alignment of comparable sentences were based on the similarity between the structure of dependency trees. In [2], for instance, alignment of common information is performed between pairs of dependency trees making use of two different kind of information: similarity of syntatic relations (e.g. subject-verb) and similarity between words and phrases. Word similarity is given by a thesaurus, while phrase similarity is given by a paraphrase lexicon automatically induced from corpora. Nevertheless, this approach has a drawback when two strings can not be recognized either by the thesaurus or by the paraphrase lexicon. In addition, a large volume of paraphrastic sentences is necessary in order to obtain a representative paraphrase lexicon, which is hard to get. While [2] ignore part-of-speech information, in [11] only pairs of lexically matched words with the same part-of-speech and the same syntatic dependency traces are aligned. However, such approach may not also treat alignments involving phrases properly, since none paraphrase information or thesaurus are used[3]. Paraphrastic phrases are the most commom type of paraphrases mainly in comparable sentences (see Section 2) and unfortunately also very difficult to handle.

[1] In [9], paraphrase recogniton has been used to generate paraphrastic sentences for evaluation of machine translation systems.

[2] This kind of paraphrase is usually called syntactic paraphrases (see [5]).

[3] No evaluation result for the alignment process in specific is presented by those works, since alignment of common information is an intermediate process.

Following the same idea of some previous works (e.g. [2], [7] and [11]), the method described in this paper combines lexical, part-of-speech and syntatic dependency information to find links between common words and phrases from comparable sentences. The main difference from those works is the phrase alignment which is solved by applying some paraphrase rules extracted from corpora.

The remainder of this paper is organized as follows: section 2 presents an overview of paraphrase rules extraction and section 3 describes how our approach works. Section 4 presents some preliminar experiments and their results. Finally, section 5 presents some final remarks.

2. PARAPHRASE RULES EXTRACTION

In this first version of the method, paraphrase rules have been extracted manually from corpora. We analized 30 pairs of comparable sentences randomly selected from a set of ≈ 670 pairs. For the building of the comparable sentence set, we have selected 50 news document collections with an average of 4 documents on the same topic per collection[4]. Then, each document collection has been submitted to a clustering system described in [10], which has identified and grouped similar sentences into cluster. For each sentence cluster we have generated all possible combinations of comparable sentence pairs, resulting in ≈ 670 pairs. For each pair of sentence randomly selected, the paraphrases have been identified, totalizing 81 for all pairs. Following the Hoey's paraphrase definition [6], we have considered two distinct word sequences as paraphrases whether one may substitute for the other, in a given context, with no discernible change in meaning.

Some examples of paraphrases are presented in Table 1. 26.3% of the identified paraphrases occur between words (e.g. (a), (g)), while 73.7% occur between phrases (e.g. (b), (c), (d), (f)) or between one word and one phrase (e.g. (e), (h)).

Table 1. Examples of paraphrases[5]

a. chocou; bateu	e. acordo; acordo financeiro
b. tucano Geraldo Alckmin; candidato tucano Geraldo Alckmin	f. mercado moscovita; mercado Cherskisov de Moscou
c. capital russa; capital da Rússia	g. deputados; parlamentares
d. estudos finais que estão sendo realizados pela Infraero; estudos finais da Infraero	h. grupo; grupo criminoso

From corpus analysis, 27 paraphrase rules were obtained. Some of them are presented in Table 2 (where ADJ: adjective; DET: determiner; N: noun; PRP: preposition; PROP: proper; REL: relative clause; V: verb; ?: zero or one occurrence; and numbers mean similar lexical units). R1 covers the examples (c) and (f); R2 covers examples (e) and (h); R3 and R4 cover the examples (b) and (d), respectively. For the examples (a) and (g) there are no rules, since they are lexical paraphrases.

Table 2. Examples of paraphrase rules

R1. N1 ADJ ; N1 PROP? PRP DET? PROP	R3. N PROP1; N ADJ PROP1
R2. N1 ; N1 ADJ	R4. N1 ADJ? REL V V V? PRP DET? PROP1 ; N1 ADJ? PRP DET? PROP1

3. ALIGNMENT APPROACH

The alignment algorithm works with part-of-speech (POS) tagging and dependency analysis provided by Palavras parser [3]. The dependency traces hold between tokens and include dependencies such as head/subject, head/modifier, subject-verb, etc. Figure 1 shows an example of the parser output for the portuguese sentence "*O prazo foi definido pela Mesa Diretora da Câmara*". In that example, *prazo* (token #2) is the subject (@SUBJ) of the verb (V) *foi* (token #3) where #2->3 means token #2 depends on token #3. Lemmatizing is also performed by the parser, where lemmas are represented by square brakets.

```
O [o] <artd> DET M S @>N  #1->2
prazo [prazo] <per> <temp> N M S @SUBJ>  #2->3
foi [ser] <fmc> <aux> V PS 3S IND VFIN @FS-STA  #3->0
definido [definir] <mv> V PCP M S @ICL-AUX<  #4->3
por [por] <sam-> PRP @<PASS  #5->4
a [o] <artd> <-sam> DET F S @>N  #6->7
Mesa=Diretora [Mesa=Diretora] <org> PROP F S @P<  #7-
>5
de [de] <sam-> <np-close> PRP @N<   #8->7
a [o] <artd> <-sam> DET F S @>N  #9->10
Câmara [câmara] <prop> <Lh> <HH> N F S @P<  #10->8
$. #11->0
```

Figure 1. Example of Palavras's output

Given a pair of comparable sentences, the algorithm tries to find the best alignment between words and phrases which share common information. We consider only one-to-one alignments, that is, every string in one sentence is linked to at most one string in the other sentence. The alignment differs considerably from alignment perfomed in other NLP tasks (e.g. machine translation), since just a subset of strings in one sentence aligns with a subset of strings in the other. Moreover, only nouns, verbs, adverbs and adjectives are aligned. Words of closed class (e.g. determiners and prepositions) only take part in phrase alignments (e.g. *capital russa* and *capital da Rússia*).

[4] This corpus has been manually collected from several web news agencies.

[5] The examples used in this paper have been kept in Brazilian Portuguese in order to avoid noise in the translation.

For each word in a source sentence, the algorithm looks for possible candidates for the alignment in the target sentence. For this, it uses as âncor words with the same lemma, synonyms and cognates. Synonym relations are given by Wordnet-BR [4] and cognates are obtained by using the longest common subsequence ratio (LCSR)[6]. By using LCSR, the algorithm can treats small changes in the strings (e.g. *Hezbolla* and *Hisbola*) and different forms of the same proper name (e.g. *Rui Pimenta* and *Rui Costa Pimenta*). We do not employ LCSR for verbs in order to avoid cases like *correr* and *morrer* which, in spite of the high LCSR value (i.e. 0.84), are totally different in meaning. Furthermore, candidate words have to be of the same POS of the source word.

After candidates have been found, the algorithm tries to retrieve the corresponding phrases for source word and each candidate (e.g. *Câmara* (Figure 1) pertains to noun phrase *Mesa Diretora da Câmara*). Retrieving phrases is facilitated by making use of dependency traces. Then, the algorithm calculates the alignment probability between source word and each candidate. For cases in which both source and candidate are one-word phrases, the probability is 1 if they were identical or 0.3 if they were synonyms or cognates. So, it priorizes alignments between literal string match. For verbs, it also considers subject-verb relations, since frequently there are more than one synonym with the same chance. If corresponding subjects are aligned, the probability is increased in 0.1, or it is penalized in -0.1, otherwise. Thus, at first iteration, the algorithm priorizes alignments involving nouns and proper names to find links between subjects. In cases in which source and/or candidate are not one-word phrases, the alignment probability is obtained by applying paraphrase rules. Paraphrastic phrases have probability equal 0.5 and identical phrases have probability equal 1. Those values were determined empirically.

Finally, the algorithm tries to align the remaining unaligned words and phrases for which no one paraphrase rule could be applied. This is performed based only on dependency relations, for instance, between subject-verb and verb-subject. So, if two subjects were aligned to each other and corresponding verbs were still not aligned, they are aligned and vice-versa.

Next, we show some examples of alignments produced by the algorithm and a brief description about how they have been found.

- (reforma de Guarulhos : reforma de Guarulhos) : a alignment between two identical phrases;
- (Secretaria de Estado da Fazenda : Secretaria da Fazenda): a alignment between two proper names with different forms, but that refer to the same entity. In this case, they were found by using LCSR. Here, the LCSR value is 19/27 = 0.7, which is greater than the threshold of 0.65 used by the algorithm;
- (enchentes : inundações): a alignment between two synonym words;
- (aviação israelense : aviação de Israel): a alignment between two paraphrastic phrases found by applying the paraphrase rule N1 ADJ ; N1 PROP? PRP DET? PROP;
- (mafia das ambulâncias : mafia dos Sanguessugas): a alignment between two paraphrastic phrases which were found by applying the paraphrase rule N1 PRP DET? N : N1 PRP DET? (N|PROP);
- (voltou : recomeçou): a alignment between two different verbs whose corresponding subjects were aligned to each other. In this case, they were not found in Wordnet-BR and the LCSR value (i.e. 3/9 = 0.33) were smaller than threshold;
- (Lula : presidente Luiz Inácio Lula da Silva): a alignment between two different strings (both subjects) whose corresponding verbs were aligned to each other. In this case, the LCSR value (i.e. 4/31 = 0.12) were smaller than threshold.

4. EXPERIMENTAL EVALUATION

A preliminary experiment has been performed in order to evaluate the quality of the alignments produced by the algorithm. For the evaluation, a reference corpus has been built with 20 pairs of comparable sentences randomly selected from corpus described in Section 2 (such sentences are different from those used for identification of paraphrase rules). The 20 pairs of sentences have been manually aligned by two annotators. Inter-annotator agreement has been calculated with respect to the number of alignments in common divided by the total number of alignments produced by both annotators. An agreement rate of 87% indicates that the annotations are reasonably reliable. So, discrepant alignments have been reviewed by the annotators and mistaken or forgotten alignments have been corrected.

In order to verify the contribution of paraphrase rules for the alignment, we have compared our method with two different baselines: a) baseline1, which is based on lexical information only (i.e. just synonym relations and cognates) and b) baseline2, which is based on lexical information plus dependency relations (without paraphrase rules). Alignments produced by the algorithm have been evaluated using the well-known Precision, Recall and F-measure metrics.

Let *R* be the set of reference alignments, *A* the set of predicted automatic alignments and $|A \cap R|$ the set of correctly predicted automatic alignments. Precision, Recall and F-measure (the harmonic mean between Recall and Precision) are given by Formulas (1), (2) and (3), respectively.

$$\text{Precision} = \frac{|A \cap R|}{|A|} \qquad (1)$$

$$\text{Recall} = \frac{|A \cap R|}{|R|} \qquad (2)$$

$$\text{F-measure} = \frac{2 \text{ x } Precision \text{ x } Recall}{Precision + Recall} \qquad (3)$$

[6] The LCSR of two words is computed by dividing the length of their longest common subsequence by the length of the longer word. For instance, the LCSR of the words *bujão* and *botijão* is 0.57 (i.e. 4/7) as their longest common subsequence is *b-j-ã-o*.

Table 3 summarizes the results obtained for each alignment algorithm.

Table 3. Results obtained for Precision, Recall and F-measure on automatic alignment

System	Precision	Recall	F-measure
Baseline1	0.77	0.72	0.74
Baseline2	0.77	0.72	0.74
Proposed algorithm	0.86	0.81	0.83

The results presented in Table 3 show that the algorithm improves the baselines in 9% on the overall performance. The baselines already achieve a high score (74% F-measure), which may be explained by the high frequency of literal string overlap in comparable sentences (i.e. 72% of the total of alignments). It is worth noting that no difference on performance has been achieved by the baseline2 (with respect to baseline1) when we plus dependency traces.

Based on these results, Precision, Recall and F-measure have been also calculated considering alignments involving paraphrases only (both lexical and syntactic paraphrases). Table 4 summarizes the results achieved.

Table 4. Results obtained for Precision, Recall and F-measure considering paraphrase alignments only

System	Precision	Recall	F-measure
Baseline1	0.63	0.12	0.20
Baseline2	0.50	0.17	0.25
Proposed algorithm	0.63	0.45	0.53

As we can see in the results presented in Table 4, the algorithm improves substantially the baselines on the overall performance (i.e. a 33% and 28% increase with respect to baseline1 and baseline2, respectively), when we regard paraphrase alignment only. The use of dependency relations in baseline2 (without paraphrase rules) improves few points on the overall performance, but loses a lot on precision, when compared to baseline1 (without dependency relations). Such results suggest that only lexical and dependency traces similarity can not deal more complex paraphrases properly such as syntactic paraphrases.

5. FINAL REMARKS

This paper has presented an alignment method which aligns words and phrases in common between Portuguese comparable sentences based on lexical and syntactic similarity and some language-dependent paraphrase rules.

A preliminary experiment has been carried out in order to analyze the method performance. The results achieved (i.e. 83% F-measure) improve the baselines in 9% and are close to results reported by other literature works on alignment of common information among parallel sentences (see [7]), which has achieved a F-measure of 85%. With respect to paraphrase alignment only (that is, excluding literal string overlap), the algorithm improves the baselines up to 33% on the overall performance. Regarding just paraphrase alignment, no one result has been reported by the literature works.

So far, the algorithm do not deals with dependencies between verb-object and object-verb. In order to increase the method performance, novel dependency traces must be considered and novel paraphrase rules will be investigated. Furthermore, future works include automatic extraction of paraphrases from corpora. A larger corpus of comparable sentences is already been built for this purpose. As a long-term goal, the alignment method will be part of a system which aims at generating a single sentence by fusing common information from comparable sentences.

6. ACKNOWLEDGMENTS

We thank CNPq for financial support. Our thanks also to Helena de Medeiros Caseli for her help in the reference corpus annotation.

7. REFERENCES

[1] Barzilay, R. 2003. Information Fusion for Multidocument Summarization: Paraphrasing and Generation. Phd. Thesis, Columbia University, New York, 221 p.

[2] Barzilay, R. and McKeown, K. 2005. Sentence Fusion for Multi-document News Summarization. Computational Linguistics, Vol. 31, nº 3, pp. 297–327.

[3] Bick, Eckhard. 2000. The Parsing System “Palavras” - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press.

[4] Dias-da-Silva, B.C., Di Felippo, A., and Hasegawa, R. 2006. Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses and Conceptual-Semantic Relations. In Proceedings of the 7th Workshop on Computational Processing of the Portuguese Language - Written and Spoken - PROPOR (Lecture Notes in Artificial Intelligence, 3960), pp. 120–130.

[5] Dras, M. 1999. Tree Adjoining Grammar and the Reluctant Paraphrasing of Text. Phd. Thesis, Macquarie University, Australia, 282 p.

[6] Hoey, M. 1991. Patterns of Lexis in Text. Oxford : Oxford University Press.

[7] Marsi, E. and Krahmer, E. 2005. Explorations in Sentence Fusion. In Proceedings of the 10th European Workshop on Natural Language Generation – ENLG’ 2005, pp. 109–117.

[8] Marsi, E. and Krahmer, E. 2007. Annotating a parallel monolingual treebank with semantic similarity relations. In Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories – TLT'07.

[9] Pang, B., Knight, K., and Marcu, D. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In Proceedings of the

Human Language Technology Conference – HLT/NAACL, pp. 102–109.

[10] Seno, E.R.M. and Nunes, E.R.M. 2008. Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese. In: Proceedings of the International Conference on Computational Processing of Portuguese Language - PROPOR (Lecture Notes in Artificial Intelligence, 5190), pp. 133–144.

[11] Shen, S., Radev, D. R., Patel, A., and Erkan, G. 2006. Adding Syntax to Dynamic Programming for Aligning Comparable Texts for the Generation of Paraphrases. In Proceedings of the COLING/ACL 2006, pp. 747–754.