Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts

Priscila Aleixo, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400 – Centro. 13560-970 - São Carlos – SP

per_pri@yahoo.com.br, taspardo@icmc.usp.br

ABSTRACT

Based on Cross-document Structure Theory (CST), we investigate the problem of finding related sentences from multiple documents on the same topic. We test some lexical similarity measures from related literature and improve them with language specific resources. The conclusions are that for Portuguese a different measure from English is the best one and that the knowledge resources we use affect the results in different ways.

Categories and Subject Descriptors

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

General Terms

Languages and Theory

Keywords

Natural Language Processing; discourse parsing

1. INTRODUCTION

Many researches on how to automatically establish relationships between different parts of a text have been carried out in the last decade. In the beginning of the 80's emerged one of the most used discourse theories, RST (Rhetorical Structure Theory) [10], which is largely used until today. This theory identifies relations between parts of a text (e.g., cause-effect, contrast, and elaboration relations) and, with the existence of many RST discourse parsers, the theory helped many applications of Natural Language Processing (NLP) such as summarization ([11][13][15], text generation [19], essay scoring [3] and others.

Recently, with the increasing amount of information mainly in electronic format, dealing with multiple documents turned to be necessary, for instance, to grasp the main facts about an event described in time, as a terrorism attack or some international soccer championship. As for single-document relationship, a discourse theory was suggested for multidocument analysis: the CST (Cross-document Structured Theory) [17], the only formal theory about cross-document relationship found in literature. Such theory gives NLP tasks the valuable ability to deal with redundant, complementary, contradictory, and temporal information. It establishes relations among segments of different documents that are about the same topic. Some examples of CST relations are paraphrase, contradiction, and subsumption. For instance, sentences S1 and S2 below come from different documents and present a subsumption relation, where S2 subsumes S1 (S2 presents all the information in S1 and some additional material):

S1: *John Doe was found guilty of the murder.*

S2: *The court found John Doe guilty of the murder of Jane Doe last August and sentenced him to life.*

Although CST and multidocument analysis usefulness is a consensus in the area, few NLP tools involving CST analysis have been developed, since discourse parsers for this theory are still rare. In fact, as far as we know, only one poor CST parser for English exists [22].

Discovering cross-document relations can be an exhaustive and hard work for both humans and machines. For example, consider that we are looking for relationships between sentences of 2 documents with 20 sentences each one. It will be necessary to analyze 400 sentence pairs for determining the ones that present some relation and, in this case, to decide among a set of relations which ones are most appropriate. The scenario is much worse if we consider that we are analyzing relationships between clauses or phrases segments instead of sentences or that we have more than 3 documents under analysis at the same time.

One first step to enable efficient discourse parsing for CST is to restrict the possibilities for combining segments from different documents. This is the strategy followed for English [18]. This is possible because CST authors claim that the cross-document relations usually happen between segments that present some lexical similarity. Therefore, automatic lexical similarity measures may be used for determining which segments to combine.

Although some segment pairs may be eventually lost, the benefits of pruning the possibilities are worth, as is argued by the authors.

In this paper, we investigate the problem of finding related sentence pairs from multiple documents written in Brazilian Portuguese. We evaluate some lexical similarity measures used in [18] and some variations of them using language specific resources, like thesaurus relations, lemmatization and stoplist. Our evaluation is carried out over a reference corpus of CST-annotated news texts built by experts in the theory. Our purpose is to discover the measure that best fits news texts in Portuguese.

The rest of this paper is organized as follows: related work on lexical similarity measures is described in Section 2, while our proposal for Brazilian Portuguese is shown in Section 3; Section 4 brings our evaluation setup and results; some final remarks are presented in Section 5.

2. RELATED WORK

Countless researches have used lexical similarity measures in many different NLP tasks. In general, they make use of the vector space model [20] to represent sentences, sometimes with variations for specific applications (see, e.g., [2]).

Cosine measure is probably the most used similarity measure (see, e.g., [7][9][21][23]). Other usual measures are word overlap (see, e.g., [21]), the traditional edit distance (see, e.g., [2]), and Longest Common Subsequence (LCS hereafter) (see, e.g., [18]). Variations of the above measures are also easily found (see, e.g., [5]). They include, for instance, the use of Princeton WordNet [8], part-of-speech tags, ontology concepts, compound nouns and word frequency information. It is not rare to find machine learning techniques working over similarity measures (see, e.g., [2][4]).

Specifically for CST discourse parsing, the cosine, word overlap, LCS and BLEU [14] measures were evaluated [18]. As in this work, the measures were used to restrict the possible number of sentence pairs to analyze. Such measures were tested in a small reference CST-annotated news corpus with 3 documents and 45 sentences. The authors tested the measures with diverse thresholds. A threshold is simply the score above which some CST relation is supposed to occur between the sentences under measurement. The performance of each measure for finding related sentences was computed by the authors in terms of the traditional measures precision, recall and f-measure: precision=K/S, recall=K/T, and f-measure=(2*precision*recall)/(precision+recall), where K stands for the number of sentence pairs correctly found by the measure to have some CST relation; S stands for the number of sentence pairs indicated by the measure to have some CST relation, correct or not; T stands for the number of sentence pairs that have some CST relation in the reference CST-annotated corpus.

The authors determined that the best choice would be the measure that would select as many correct sentence pairs as possible and that would filter out a large number of incorrect sentence pairs. Therefore, the authors preferred recall over precision, but did not ignore the latter. Their conclusions were that (a) word overlap measure was the best one with a threshold of 0.12 (with recall of 87.5% and precision not reported by the authors), (b) cosine measure is good but not the best, and (c) LCS and BLEU were the worst measures, since they use high order n-gram matching and this is not frequent in the task under focus. In a later experiment, with a bigger corpus, the authors observed similar results.

In this paper, for news texts written in Brazilian Portuguese, we chose to evaluate the two best measures according the above work: word overlap and cosine measure. Additionally, we extend such measures by using language specific resources, namely: thesaurus relations, lemmatization and stoplist, as described in the next section. In the future, we plan to test other measures and techniques, as latent semantic analysis, for instance.

3. LEXICAL SIMILARITY MEASURES: A GENERAL FRAMEWORK

The measures we evaluate work according to the generic framework in Figure 1.

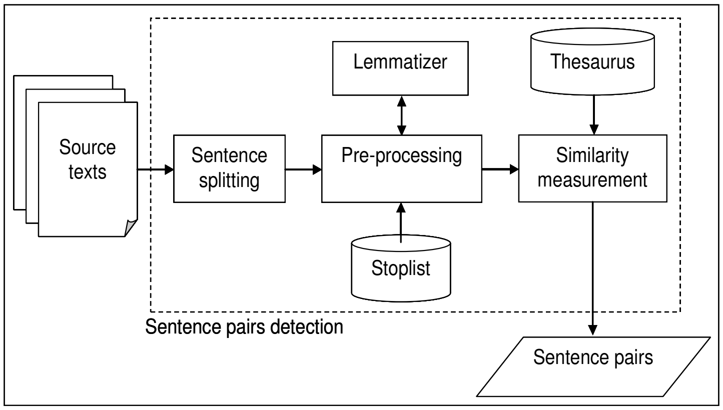

Figure 1. Generic framework for detection of CST-related sentence pairs

Initially, the source texts have their sentences delimited in the sentence splitting process. Then, a pre-processing step is optionally carried out, lemmatizing and/or removing stopwords from the sentences. Finally, a lexical similarity measure is applied to all possible sentence pairs from the documents. This last process may optionally use synonym relations from a thesaurus. In the end, the output data consists of sentence pairs supposed to present CST relations. Sentence splitting is performed by the system SENTER [16]. Stoplist removal may be useful for reducing non-important words counting in the similarity measurement, while lemmatization is usually applied for changing inflected word forms into their unique base form, allowing the system to recognize them as similar words. In this work, we use a generic stoplist for Brazilian Portuguese and NILC lemmatizer [12]. Finally, we use a Brazilian Portuguese thesaurus [6] in order to identify synonyms and, therefore, to consider them similar words in the computation of the lexical similarity measures.

The measures we evaluate in this work are word overlap and cosine measure, as defined in what follows:

$$\mathrm{WordOverlap(S1,S2)} = \frac{\#\mathrm{CommonWords(S1,S2)}}{\#\mathrm{Words(S1)} + \#\mathrm{Words(S2)}}$$

$$\mathrm{Cosine(S1,S2)} = \frac{\sum_{\mathrm{word}} (\mathrm{freq(word,S1)} * \mathrm{freq(word,S2)})}{\sqrt{\sum_{\mathrm{word}} \mathrm{freq(word,S1)}^2 * \sum_{\mathrm{word}} \mathrm{freq(word,S2)}^2}}$$

In the formulas, S1 and S2 represent the sentences for which we want to compute the measures, *#CommonWords* is the number of words in common (or their synonyms if thesaurus is used) between the sentences, *#Words* is the number of words in a sentence, and *freq(w,S)* is a function that outputs the frequency of a word *w* (or its synonyms) in a sentence *S*.

4. EXPERIMENTS

For evaluating the measures for the CST parsing of Brazilian Portuguese texts, we used a reference corpus of CST-annotated news texts in Brazilian Portuguese. The corpus, named CSTNews [1], is composed of 50 document clusters, with a total of 195 documents from diverse news sources, 3.534 sentences and 72.148 words. Each cluster has in average 4 documents, all of them about the same topic. The corpus was manually annotated by 2 experts in CST.

We randomly selected 2 clusters from the corpus for our experiments, which amount to 6 documents (3 per cluster), 134 sentences, 2.440 words, 2.658 possible sentence pairs to relate, and 91 of these pairs with some CST relation. One cluster is about a terrorism attack in USA; the other one is about taxes in Brazil.

We tested word overlap and cosine measures as originally proposed and their variations using thesaurus, lemmatization and stoplist. We tried all possible combinations of these resources with the measures. We computed the same metrics used in [18], namely, precision, recall and f-measure in terms of the correct related sentence pairs found by the similarity measures. We also tried different thresholds for each measure: from 0.1 to 0.5 for word overlap (since 0.5 is its maximum value) and from 0.1 to 1 for cosine measure (since 1 is its maximum value).

Tables 1 to 8 show the average results for word overlap measure and its variations. Assuming the assumption that recall is more important than precision (as it is done in [18]), one can see that:

- using the stoplist increased precision but penalized recall, which indicates that stoplist is not a useful resource for this task;
- using thesaurus, lemmatization or their combination (without using the stoplist) did not change the results significantly;
- the best option is the original word overlap measure, with threshold between 0.1 and 0.2 and a recall value between 93 and 53%, as verified for English.

If one considers that recall and precision are equally important, then word overlap with stoplist is the best option for a threshold of 0.2.

Table 1. Word overlap

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.14	0.17	0.45	0.80	0.00
Recall	0.93	0.53	0.35	0.21	0.00
F-measure	0.24	0.25	0.39	0.31	0.00

Table 2. Word overlap + stoplist

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.16	0.50	0.66	1.00	0.00
Recall	0.44	0.44	0.32	0.12	0.00
F-measure	0.23	0.47	0.43	0.21	0.00

Table 3. Word overlap + lemmatization

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.13	0.15	0.40	0.35	0.00
Recall	0.93	0.53	0.44	0.35	0.00
F-measure	0.23	0.23	0.41	0.31	0.00

Table 4. Word overlap + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.13	0.14	0.38	0.80	0.00
Recall	0.93	0.53	0.44	0.21	0.00
F-measure	0.23	0.22	0.40	0.31	0.00

Table 5. Word overlap + stoplist + lemmatization

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.16	0.39	0.46	1.00	0.00
Recall	0.48	0.44	0.32	0.12	0.00
F-measure	0.24	0.40	0.37	0.21	0.00

Table 6. Word overlap + stoplist + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.14	0.34	0.41	0.60	0.00
Recall	0.48	0.44	0.32	0.12	0.00
F-measure	0.22	0.37	0.34	0.16	0.00

Table 7. Word overlap + lemmatization + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.13	0.13	0.38	0.80	0.00
Recall	0.93	0.53	0.44	0.21	0.00
F-measure	0.22	0.21	0.40	0.31	0.00

Table 8. Word overlap + stoplist + lemmatization + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5
Precision	0.19	0.27	0.40	0.57	0.00
Recall	0.79	0.44	0.32	0.16	0.00
F-measure	0.31	0.28	0.33	0.22	0.00

Tables 9 to 16 show the average results for cosine measure and its variations. Assuming the same assumption that recall is more important than precision, it is possible to realize that:

- stoplist had the same effect of its use with word overlap, i.e., it increased precision and penalized recall;
- lemmatization alone increased precision but did not penalized recall;
- the best option looks to be the cosine measure with lemmatization, with threshold between 0.1 and 0.2 and a recall value between 100 and 93%.

If one considers that recall and precision are equally important, then cosine measure with lemmatization is still the best option, but with a high threshold of 0.6.

Table 9. Cosine

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.12	0.14	0.21	0.24	0.31	0.55	0.66	0.80	0.50	0.00
Recall	1.00	0.93	0.81	0.53	0.53	0.44	0.32	0.21	0.05	0.00
F-measure	0.21	0.24	0.28	0.27	0.39	0.49	0.43	0.31	0.08	0.00

Table 10. Cosine + stoplist

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.14	0.18	0.28	0.50	0.66	0.66	0.88	1.00	0.00	0.00
Recall	0.74	0.48	0.44	0.44	0.39	0.32	0.28	0.12	0.00	0.00
F-measure	0.23	0.26	0.34	0.47	0.50	0.43	0.42	0.21	0.00	0.00

Table 11. Cosine + lemmatization

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.16	0.18	0.22	0.21	0.39	0.63	0.83	0.80	0.50	0.00
Recall	1.00	0.93	0.81	0.53	0.53	0.44	0.32	0.21	0.05	0.00
F-measure	0.27	0.29	0.33	0.28	0.44	0.51	0.46	0.31	0.08	0.00

Table 12. Cosine + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.11	0.13	0.17	0.14	0.24	0.36	0.44	0.80	0.50	0.00
Recall	1.00	0.93	0.86	0.53	0.53	0.44	0.37	0.21	0.09	0.00
F-measure	0.20	0.23	0.27	0.22	0.33	0.39	0.39	0.31	0.15	0.00

Table 13. Cosine + stoplist + lemmatization

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.20	0.18	0.28	0.40	0.39	0.50	0.58	1.00	0.00	0.00
Recall	0.88	0.50	0.44	0.44	0.32	0.33	0.28	0.26	0.00	0.00
F-measure	0.31	0.27	0.34	0.41	0.35	0.40	0.37	0.21	0.00	0.00

Table 14. Cosine + stoplist + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.11	0.14	0.17	0.28	0.36	0.33	0.41	1.00	0.00	0.00
Recall	0.74	0.57	0.44	0.44	0.39	0.32	0.32	0.16	0.00	0.00
F-measure	0.19	0.22	0.24	0.34	0.35	0.31	0.34	0.28	0.00	0.00

Table 15. Cosine + lemmatization + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.11	0.13	0.17	0.14	0.24	0.36	0.44	0.80	0.50	0.00
Recall	1.00	0.93	0.86	0.53	0.53	0.44	0.37	0.21	0.09	0.00
F-measure	0.21	0.23	0.27	0.22	0.34	0.39	0.39	0.31	0.15	0.00

Table 16. Cosine + stoplist + lemmatization + thesaurus

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precision	0.13	0.18	0.23	0.27	0.44	0.40	0.50	1.00	0.00	0.00
Recall	0.88	0.73	0.51	0.37	0.32	0.25	0.25	0.17	0.00	0.00
F-measure	0.23	0.29	0.31	0.31	0.35	0.29	0.34	0.29	0.00	0.00

Interestingly, for Brazilian Portuguese, the overall best measure is the cosine measure with lemmatization, considering that recall is more important. It showed to be significantly better than the word overlap measure (93-100% recall over 53-93%) for the same threshold interval (0.1-0.2). It is also interesting to notice that recall values were significantly better for Portuguese than for English. The same conclusions can be drawn if we consider that recall and precision are equally important. It is also important to say that text genre did not interfere in the fact that results were different for English and Portuguese languages. In fact, text genre was the same for the experiments for both languages.

As argued in [18], we also believe that recall is more important than precision for CST parsing. This makes us to choose the cosine measure with lemmatization for such task, assuming 0.1 as threshold. The fact that precision is not high causes several sentence pairs that do not present CST relations to be selected.

However, the next CST parsing step, i.e., relation determination for these pairs, may eventually discard some false pairs if a CST relation may not be determined between the sentences.

Besides reproducing for Portuguese the evaluation carried out for English, we also computed the general accuracy of each measure. While the previous evaluation considers the correct pairs that had CST relations, we consider now all pairs that have relations and also those that have not. This is the traditional accuracy measure (shown in what follows), which corresponds to the number of pairs with relations that were correctly predicted to have relations plus the pairs without relations that were correctly predicted not to have relations over the number of possible pairs.

$$\text{Acc} = \frac{\#\text{correct pairs with relation} + \#\text{correct pairs without relation}}{\#\text{sentence pairs}}$$

Now we may judge the general predictive power of the measures and see if they are good in dealing with both pairs that present relations and pairs that do not. Tables 17 and 18 show average accuracy results for word overlap and cosine measure. The first line in the tables show all possible threshold values and the first column show the measures with their variations. It is possible to see that the cosine measure with lemmatization (with a threshold of 0.1) is still the best choice, with accuracy of 89%.

Table 17. Accuracy for word overlap

Threshold	0.1	0.2	0.3	0.4	0.5
No resource is used	0.75	0.30	0.07	0.03	0.00
Stoplist	0.25	0.07	0.04	0.01	0.00
Lemmatization	0.74	0.32	0.07	0.03	0.00
Thesaurus	0.77	0.33	0.08	0.03	0.00
Stoplist + lemmatization	0.24	0.07	0.04	0.01	0.00
Stoplist + thesaurus	0.32	0.08	0.04	0.01	0.00
Lemmatization + thesaurus	0.79	0.36	0.08	0.03	0.00
Stoplist + lemmatization + thesaurus	0.33	0.09	0.04	0.01	0.00

Table 18. Accuracy for cosine measure

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
No resource is used	0.89	0.74	0.53	0.30	0.14	0.07	0.04	0.03	0.01	0.00
Stoplist	0.55	0.24	0.11	0.07	0.05	0.04	0.03	0.01	0.00	0.00
Lemmatization	0.89	0.75	0.54	0.33	0.14	0.07	0.04	0.03	0.01	0.00
Thesaurus	0.87	0.76	0.56	0.33	0.16	0.08	0.06	0.03	0.01	0.00
Stoplist + lemmatization	0.55	0.23	0.12	0.07	0.06	0.04	0.03	0.01	0.00	0.00
Stoplist + thesaurus	0.59	0.33	0.16	0.09	0.06	0.06	0.04	0.01	0.00	0.00
Lemmatization + thesaurus	0.86	0.75	0.57	0.35	0.16	0.08	0.06	0.03	0.01	0.00
Stoplist + lemmatization + thesaurus	0.60	0.34	0.17	0.10	0.06	0.06	0.04	0.01	0.00	0.00

5. FINAL REMARKS

This paper introduced the first effort towards building a CST discourse parser for Brazilian Portuguese texts. The task under analysis was the detection of related sentence pairs from multiple documents about the same topic. We tested several measures and showed that the best measure for Brazilian Portuguese news texts is the cosine measure with the use of lemmatization, differently of what was observed in previous work for English language.

A multidocument discourse parser is a valuable tool for dealing with tough questions in language processing. Its advent may be very useful for advancing the state of the art in the area. To the best of our knowledge, this work is not only the first one on CST for Brazilian Portuguese, but also the first one on multidocument discourse analysis for this language.

6. ACKNOWLEDGMENTS

Our thanks to FAPESP, CAPES, and CNPq for supporting this work.

7. REFERENCES

[1] Aleixo, P. and Pardo, T.A.S. (2008). *CSTNews: um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory).* Technical Report NILC-TR-08-05.

[2] Bilenko, M. and Mooney, R.J. (2003). Adaptive Duplicate Detection Using Learnable String Similarity Measures. In the *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 39-48.

[3] Burstein, J.; Marcu, D.; Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, pp. 32-39.

[4] Cohen, W. (1996). Learning trees and rules with set-valued features. In the *Proceedings of the Fourteenth National Conference on Artificial Intelligence*.

[5] De Boni, M. and Manandhar, S. (2003). An Analysis of Clarification Dialogue for Question Answering. In the *Proceedings of HLT-NAACL*, pp. 48-55

[6] Dias da Silva, B.C.; Oliveira, M.F.; Moraes, H.R.; Hasegawa, R.; Amorim, D.; Paschoalino, C.; Nascimento, A.C. (2000). A Construção de um Thesaurus Eletrônico para o Português do Brasil. In *Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*.

[7] Erkan, G. and Radev, D. (2004). LexPageRank: Prestige in Multi-Document Text Summarization. In the *Proceedings of EMNLP*.

[8] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

[9] Hatzivassiloglou, V., Klavans, J.L., Holcombe, M.L., Barzilay, R., Kan, M., McKeown, K.R. (2001). SimFinder: A Flexible Clustering Tool for Summarization. In the *Proceedings of the Workshop on Automatic Summarization at NAACL*, pp. 41-49.

[10] Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

[11] Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts

[12] Nunes, M.G.V. et al. (1996). The Design of a Lexicon for Brazilian Portuguese: Lessons Learned and Perspectives. In the *Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese*, pp. 61-70.

[13] O'Donnell, M. (1997). Variable-Length On-Line Document Generation. In the *Proceedings of the 6th European Workshop on Natural Language Generation*.

[14] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318. Philadelphia, PA.

[15] Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

[16] Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.

[17] Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong.

[18] Radev, D.R.; Otterbacher, J.; Zhang, Z. (2004). CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In the *Proceedings of Fourth International Conference on Language Resources and Evaluation*.

[19] Rino, L.H.M. (1996). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-USP. São Carlos - SP.

[20] Salton, G. and Lesk, M.E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, Vol. 15, N. 1, pp. 8-36.

[21] Seno, E.R. and Nunes, M.G.V. (2008). Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese. In the *Proceedings of the International Conference on Computational Processing of Portuguese*. Aveiro and Curia, Portugal.

[22] Zhang, Z.; Otterbacher, J.; Radev, D.R. (2003). Learning Cross-document Structural Relationships using Boosting. In the *Proceedings of ACM CIKM*. New Orleans, LA.

[23] Yuan, S-T and Sun, J. (2005). Ontology-Based Structured Cosine Similarity in Document Summarization: With Applications to Mobile Audio-Based Knowledge Management. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, V. 35, N. 5.