Automatic Identification of Multi-document Relations

Erick Galani Maziero, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguistica Computacional – NILC
Instituto de Ciências Matemáticas e de Computação – ICMC
Av. Trabalhador São-carlense, 400. P.O.Box. 668. 13560-970 – São Carlos/SP, Brazil
{erickgm,taspardo}@icmc.usp.br

Abstract. The multi-document analysis is indispensable in the current scenario of the electronic media, in which several documents are produced about the same topic, especially considering the explosion of information allowed by the web. Both readers and applications are benefited for the multi-document analysis through which relations between portions of texts are showed, for example, equivalence, contradictions or background relations. In this work, to achieve this automatic analysis is adopted the CST (Cross-document Structure Theory [13]). This type of multi-document knowledge allows i) the appropriated treatment of phenomena such as redundancy, complementarity and contradiction, and consequently ii) the production of better systems of text processing, such as more intelligent web search engines and automatic multi-document summarizers. This paper presents a methodology used to identify this relationship using traditional and hierarchical machine learning algorithms and hand-crafted rules. Finally, a parser is generated using the classifiers and rules.

Keywords: Multi-document analysis, Cross-document Structure Theory

1 Introduction

In the electronic media, there are many sources reporting the same topic from the same or different perspectives. Online newspapers are an example: the same event is reported on different news portals. In general, these documents are produced soon after the event and, subsequently, other documents are generated to update the news. Therefore, readers interested on a current event will find an endless number of texts, and it will be crucial to pick just a few to read. This requires a great effort on the part of readers. Since these texts are produced by different sources at different moments, many phenomena can occur, as contradictory or redundant portions of text. For instance, the two sentences below, S1 and S2, from different documents, are contradictory regarding the number of bombs in an attack, but both also present overlapping information (that there was a bomb):

S1: The downtown Public Finance Department building was hit by three homemade bombs.
S2: The Public Finance Department was also hit by a bomb.

An automatic analysis of groups of texts about the same topic, showing the relationship between portions of these texts is needed. This information can be used for applications that aim to manipulate many documents to guide the user's reader, allowing them the panoramic view of the topic, making it easier to find what they are looking for.

It is believed that, when readers know how the parts of multiple documents are related, they can ignore redundancy, find contradictions, and understand the temporal evolution of a fact or event, which would allow them to approach the information in which they are interested in a more organized way. In another vein, this type of knowledge might also be useful for several computer applications, such as web browsers and automatic summarizers, which would have more information available to produce their results and meet the users' needs more efficiently. Some theories or models on multi-document relationships have been proposed for this purpose. One of the most used is the Cross-document Structure Theory (CST) [13].

In this work, we propose to investigate the automatic identification of the relations among portions of several texts suggested by the CST, developing an automated multi-document parser. We explore, in particular, the use of traditional (flat) and hierarchical machine learning techniques in this task, using a corpus of news texts written in Brazilian Portuguese, already annotated according to CST, which allows applying machine learning techniques and testing them. Also, a set of hand-crafted rules are applied to the sentences. The results obtained improved the state of the art.

One of the hypothesis that guided this work is that is possible to establish a generic typology of relations for multi-document analysis, which can be applied to any group of documents. Another hypothesis is that the CST is applicable to Portuguese language and their relations can be automatically detected with good results, enabling the development of identification of the CST relations. It is believed that hybrid strategies are needed to identify the relations, due to relations frequency. Some treated by statistical techniques and others with symbolic techniques.

This type of multi-document analysis is unheard, and researches that aim such information show the value of this work.

In Section 2, related work on multi-document parsing is briefly presented. In Section 3, we describe the proposed architecture for the multi-document parser and discuss the methodology for identifying multi-document relations. In Section 4, the results are presented for the classifiers and rules. Finally, Section 5 provides conclusions and future work.

2 Related work

Not many researchers have defined and applied multi-document representation models, since instantiating such models with real texts is a difficult task. Pioneer work was carried out by [15] and [16]. They employed a set of relations to (manually) structure scientific text portions and their relations in semantic networks. [14] used relations among parts of several texts to perform multi-document summarization. These previous works and the work of [8] were the basis for the CST discourse model

proposed by [13]. In a different line, [1] proposed a methodology to define and identify multi-document relations, using ontology and a set of related semantic templates.

Using CST, [3] developed the first step of multi-document parsing for the Portuguese language, when they detected pairs of sentences to be associated. Later, these and other authors [4], [6] built an annotated corpus of news texts, which is called CSTNews. During the manual parsing of this corpus, it was perceived that the relations could be organized in a typology that takes into account some features that the defined relation groups have in common [9].

The works of [18] and [19] consist of an attempt to automate the CST parsing for the English language. These authors handled only some CST relations and obtained average values of 45% for precision, 31% for coverage, and 35% for f-measure. Several other works tried to identify some relations for varied purposes. [11] tried to identify *Equivalence* and *Transition* (very similar to the CST *Contradiction*) relations between sentence pairs, using a Support Vector Machine (SVM) classifier, and obtained an f-measure of 75.5% for *Equivalence* and of 45.6% for *Transition*. [17] also used SVM to identify the relations *Identity*, *Paraphrase*, *Subsumption*, *Overlap*, and *Elaboration*, but did not report any evaluation. [12] dealt with the identification of the relations *Entailment*, *Contradiction*, *Confinement* (which represents the union of *Entailment* and *Contradiction* relations) and *Unknown* for Japanese. Interpreting semantic templates extracted from the sentences, these authors reported that the *Confinement* relation is recognized with an f-measure of 61%. [9] reported state of the art results with the application of a decision tree algorithm (J48) to identify a large group of relations, achieving average precision, recall and f-measure of 44%.

3 The multi-document analysis

In this work, a multi-document parser was developed following CST. Figure 1 illustrates its architecture.

A group of texts on the same topic (coming from web portals, as GoogleNews, for instance) is the input of the process. These texts are automatically segmented in sentences. As this work considers CST relations among pairs of sentences, any combination between sentences in the several documents is checked in accordance with the measure word overlap (= number of words in common in S1 and S2 / (number of words in S1 + number of words in S2)). This measure generates a value for each pair of sentences, and the pairs with a value above a pre-established threshold are selected for the following steps. This is done because it was observed that CST relations occur between sentences with some lexical similarity [19]. We use a threshold of 0.12, since this was the value used for English [19] and that showed to be good for Portuguese too [3].

The selected pairs are then analyzed by several tools (part of speech tagger, syntactic parser, named entity recognizer) assisted by several resources (thesaurus, list of verbs of attribution – e.g., “say” and “announce”) in order to extract relevant features from each sentence pair. The result of this step is an attribute-value table used by classifiers that, after training, identify the existing relations between sentence pairs.

These pairs are then passed through the rules to identify other relations. The result of the multi-document parsing process is a graph, whose nodes are sentences from the several documents under analysis and the edges are the identified relations.

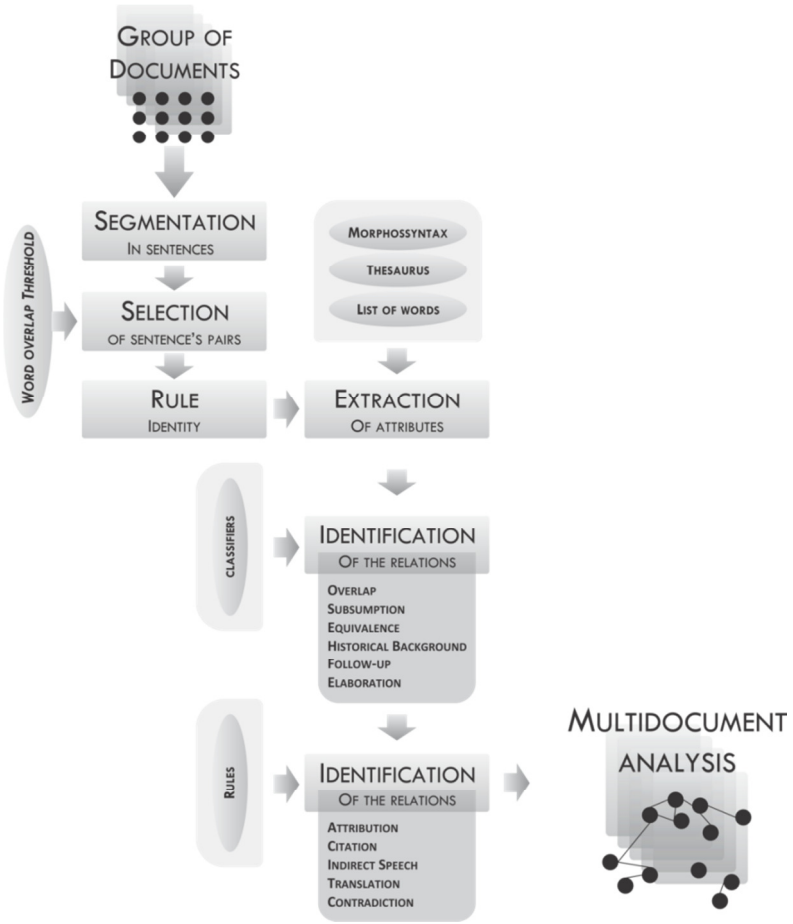

Fig. 1. Architecture of the parser

Several machine learning techniques with different configurations for producing classifiers have been explored, so that it was possible to choose the best scenario for the task. 14 features/attributes were extracted from each sentence pair used to generate the classifiers. After this extraction, all features are normalized to avoid possible classification discrepancies.

The first six features were obtained using only text surface information, working with the words from the sentences. These features help to identify information overlap between sentences, since sentences with redundant information present word overlap. To extract features 7 to 12, it were employed a part of speech tagger for Brazilian Portuguese [2], with a precision of more than 96%. These features do not check the

word itself, but the amount of words in the same class that has been found in the sentences as a sign that there is a content relation between sentences in a pair.

The features 13 and 14 were obtained using the syntactic parser Palavras [5], which lemmatize verbs as well. To the feature 14, a synonym database was used [10]. In this feature, a list of synonym identifiers was compiled for each word, ignoring the stop words. The synonym database is fundamental to identify overlap when the words are not identical, but belong to the same set of synonyms.

The CSTNews corpus used in the experiments contains 50 clusters of texts, totalizing 140 texts, 2,088 sentences and 47,240 words. The corpus was manually annotated according the CST by four experts and the kappa agreement values [7] for this task were computed and the result was 0.50. Given the subjectivity of this task, such values are considered good. The corpus has 1,651 pairs of sentences with some CST relation, and these pairs were used to train the classifiers and create the rules.

The techniques to develop the classifiers were: NaiveBayes, Support Vector Machine (SVM), and decision tree (J48). In this paper we show the results only for the techniques that produced the best results. Some scenarios have been explored during the development of classifiers. Two multi-class classifiers (they seek to identify one among several classes for each instance) were created. The typology of CST relations allowed for the development of two hierarchical classifiers (top-down and big-bang approaches, which take into consideration the hierarchy of the relations). Finally, we explored the development of binary classifiers (they consider only two classes in the decision process) for the most frequent relations in the corpus. In this paper we show the results for only three experiments, namely, the multi-class classifier for content relations, the hierarchical classifiers, and the binary classifiers. These classifiers produced the best results. We used ten-fold cross-validation over the CSTNews corpus.

We replicated all the experiments for balanced data, in order to see its impact in the decision process. For balancing the data, we used the traditional approach of systematically duplicating the instances of each class until that each class has the same number of instances of the majority class. Although running these tests, we think this is not a good strategy, since the data is too unbalanced and some instances have to be duplicated several times, causing the classifiers to be potentially biased, suffering from overfitting. Such approach also results in losing the fact that such classes are really unbalanced in actual occurrences in the language.

Some relations do not have sufficient frequency (in the used corpus) for the creation of the classifiers, but are likely to be identified by hand-crafted rules. The rules are applied to every pair of sentence that is analyzed by the classifiers.

The *Identity* relation indicates total equality between two sentences. The relations *Attribution* and *Citation* are similar in their definition; they deal with the authorship of redundant information between sentences. The difference is that *Attribution* gives the authorship to some author present in another sentence of the pair and the Citation gives the authorship to another document, the owner of the other sentence of the pair. The identification of the relation Indirect Speech is obtained with the use of patterns that indicate the form of the discourse: one sentence with direct speech and another with indirect speech. The relation *Contradiction* is identified with the verification of numerical differences between sentences with redundant information. Finally, the

Translation relation is identified by checking the same information but in different languages, shared by two sentences. The rules where obtained from the manual analysis of the examples of the relations from the corpus. This process was evaluated measuring the precision and recall of the corpus examples.

4 Results

For comparison and validation procedures, we simulated a baseline method for parsing that simply assigns the most frequent relation (*Overlap*) to every sentence pair (Table 2). It produces a general accuracy of 32.74% and 16.67%, respectively to unbalanced and balanced data. These results are outperformed by the classifiers explored in this work. Finally, we also performed attribute selection before generating the classifiers, but only some attributes were ignored and the results were the same.

Table 1. Results for the classifiers

Strategy	**Technique**		
	NB	**SVM**	**J48**
Unbalanced data			
Multi-class "content" relations	0.3906	0.4158	0.4109
Binary	-	-	0.7051
Hierarchical top-down	-	-	0.4270
Hierarchical big-bang	-	-	0.6150
Baseline	0.3274		
Balanced data			
Multi-class "content" relations	0.4525	0.4804	0.7287
Binary	-	-	0.6028
Hierarchical top-down	-	-	0.7416
Hierarchical big-bang	-	-	0.7070
Baseline	0.1667		

The training and testing of the classifiers were made over the examples of the corpus using a ten-fold cross validation procedure. In the evaluation of the rules, the same pairs of sentences used to develop the rules were used to evaluate it. The best result for the unbalanced data is using the binary classifiers: 70.51% of general accuracy (percentage of corrected classified instances). The hierarchical big-bang also had a good result: 61.50%, but this classifier does not achieve the leaves of the hierarch at all times. The results are much better to balanced data, but we believe that it is a consequence of overffiting, since the examples of some classes are replicated many times.

In the Table 3 are presented the results for the hand-crafted rules for some relations. The rule for relation Translation obtained 0.5 of F-measure, because there are only two examples of this relation and one of them has an expression that is not found on translator used. The rule for *Contradiction* obtained the worst result due to diffi-

culty of identification of contradictory information between two sentences and only contradictions among numerical values where treated.

Table 2. Results for the rules

Rule	Recall	Precision	F-Measure
Indirect Speech / Attribution / Citation	0.6322	0.5288	0.5759
Translation	0.5000	0.5000	0.5000
Contradiction	0.1765	0.2728	0.2143
Average	**0.4363**	**0.4339**	**0.4301**

The general accuracy of the parser is 68.57%. This result was obtained weighting the results from classifiers and rules, accordingly to the number of instances in the corpus. The result is better than the current state of the art for other works in the area.

5 Conclusions and future work

This work allows the automatic handling of multiple documents in Portuguese. Both users and computer applications will benefit from it.

The CST was chosen by its computational tractability, and because it is widely used in others researches, which require multi-document treatment. During this research, a corpus was generated and the relations where organized in a typology, used by the hierarchical classifiers.

The methodology described can be adapted to others languages, being necessary the training of the classifiers and creation of another rules.

The main limitation is the unbalancing of the relations in the corpus (it is natural in this task). This limitation was overcome to some relations through the creation of the rules, even with few examples in the corpus.

Some approaches can be explored, as semi supervised learning, increasing the number of examples per relation.

References

1. Afantenos, S.D.; Doura, I.; Kapellou, E.; Karkaletsis, V. (2004). Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In the *Proceedings of SETN*, pp. 410-419.
2. Aires, R.V.X.; Aluísio, S.M.; Kuhn, D.C.S.; Andreeta, M.L.B.; Oliveira Jr., O.N. (2000). Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In the *Proceedings of the Brazilian AI Symposium*, pp. 20-22.
3. Aleixo, P. and Pardo, T.A.S. (2008a). Finding Related Sentences in Multiple Documents for Multi-document Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana – TIL*, pp. 298-303. Vila Velha, Espírito Santo.
4. Aleixo, P. and Pardo, T.A.S. (2008b). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Semântico-Discursiva Multi-documento CST (Cross-document*

Structure Theory). Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP, May, 12p.
5. Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis. Aarhus University. Denmark University Press.
6. Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 1-18. October 26, Cuiabá/MT, Brazil.
7. Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, Vol. 22, N. 2, pp. 249-254.
8. Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.
9. Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010). Identifying Multi-document Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science - NLPCS*, pp. 60-69.
10. Maziero, E.G.; Pardo, T.A.S.; Di Felippo, A.; Dias-da-Silva, B.C. (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL, pp. 390-392. Vila Velha, Espírito Santo. October, 26-28.
11. Miyabe, Y.; Takamura, H.; Okumura, M. (2008). Identifying cross-document relations between sentences. In the *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp 141-148.
12. Ohki, M.; Nichols, E.; Matsuyoshi, S.; Murakami, K.; Mizuno, J.; Masuda, S.; Inui, K.; Matsumoto, Y. (2011). Recognizing Confinement in Web Texts. In the *Proceedings of the Ninth International Conference on Computational Semantics*, pp. 215-224.
13. 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*.
14. Radev, D.R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, N. 3, pp. 469-500.
15. Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community*. PhD dissertation. Department of Computer Science, University of Maryland.
16. Trigg, R. and Weiser, M. (1987). TEXTNET: A network-based approach to text handling. *ACM Transactions on Office Information Systems*, Vol. 4, N. 1, pp. 1-23.
17. Zahri, N. and Fukumoto, F. (2011). Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences. *Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science, Vol. 6609, pp. 328-338.
18. Zhang, Z.; Otterbacher, J.; Radev, D.R. (2003). Learning Cross-document Structural Relationships using Boosting. In the *Proceedings of the Twelfth International Conference on Information and knowledge Management*, pp. 124-130.
19. Zhang, Z. and Radev, D.R. (2004). Combining Labeled and Unlabeled Data for Learning Cross-Document Structural Relationships. In *Proceedings of IJCNLP*, pp. 32-41.

Acknoledgments.

The authors are grateful to FAPESP for supporting this work.