Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning

Erick Galani Maziero and Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

{erickgm,taspardo}@icmc.usp.br

Abstract. *Multi-document handling is essential today, when many documents on the same topic are produced, especially considering the Web. Both readers and computer applications can benefit from a discourse analysis of this multi-document content, since it demonstrates clearly the relations among portions of these documents. This work aims to identify such relations automatically using machine learning techniques. Particularly, this work focuses on the identification of relations predicted by the Cross-document Structure Theory (CST). The obtained results improve the state of the art.*

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, they may contain contradictory or redundant portions. 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.

It is believed that, when readers know how the parts of multiple documents are related, they can, for example, 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) (Radev, 2000).

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. The results obtained are that the performance of some

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 1–10,
Cuiabá, MT, Brazil, October 24–26, 2011.

classifiers improved the state of the art. It is also demonstrated that the task in question can be characterized as a hierarchical problem.

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, introducing the experiments carried out with machine learning. Finally, Section 4 provides conclusions and future work.

2. Related work

Although useful, 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 Trigg (1983) and Trigg & Weiser (1987). They employed a set of relations to (manually) structure scientific text portions and their relations in semantic networks. Radev & Mckeown (1998) used relations among parts of several texts to perform multi-document summarization. These previous works and the work of Mann & Thompson (1987) were the basis for the CST discourse model proposed by Radev (2000). In a different line, Afantenos *et al.* (2004) proposed a methodology to define and identify multi-document relations, using an ontology and a set of related semantic templates.

Using CST, Aleixo & Pardo (2008a) 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 (Aleixo & Pardo, 2008b; Cardoso *et al.*, 2011) built an annotated corpus of news texts, which is called CSTNews and is used in this paper. 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 (Maziero *et al.*, 2010). This typology is illustrated in Figure 1, where the relations are at the lowest level of the hierarchy.

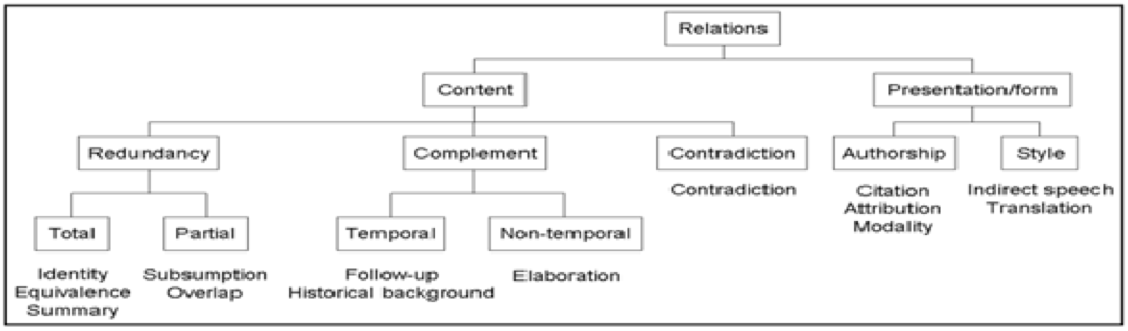

Figure 1. Typology of CST relations

Some relations deal mainly with the "content" of sentences. In this group, the "contradiction" among contents, the "redundancy", which can be "total" or "partial", and the "complementarity" among contents are analyzed, taking into consideration the "temporal" or "non-temporal" aspect of this feature. Other relations deal mainly with the "form" of sentences, i.e., how they convey information, considering their "source/authorship" and the writing "style". The relations in this group can occur jointly with some other relation from the "content" group, since the sentence pair under analysis will always have a similar content.

The works of Zhang *et al.* (2003) and Zhang & Radev (2004) consist of an attempt to automate the CST parsing for the English language. These authors carried out a two-step CST parsing: first, they developed a classifier to determine if any two segments (sentences) from different texts are possibly related. Next, they used other classifier to determine which relation there is between the segments. 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. Miyabe *et al.* (2008) 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*. Zahri & Fukumoto (2011) also used SVM to identify the relations *Identity*, *Paraphrase*, *Subsumption*, *Overlap*, and *Elaboration*, but did not report any evaluation. Ohki *et al.* (2011) 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%. Maziero *et al.* (2010) reported state of the art results with the application of a decision tree algorithm (J48) to identify a large group of relations (the "content" group in the CST typology reproduced in Figure 1), achieving average precision, recall and f-measure of 44%. This last work is the basis for this paper, which builds on it by proposing new ways of exploring machine learning techniques.

3. The multi-document parser

In this work, a multi-document parser was developed following CST. Figure 2 illustrates its initial architecture.

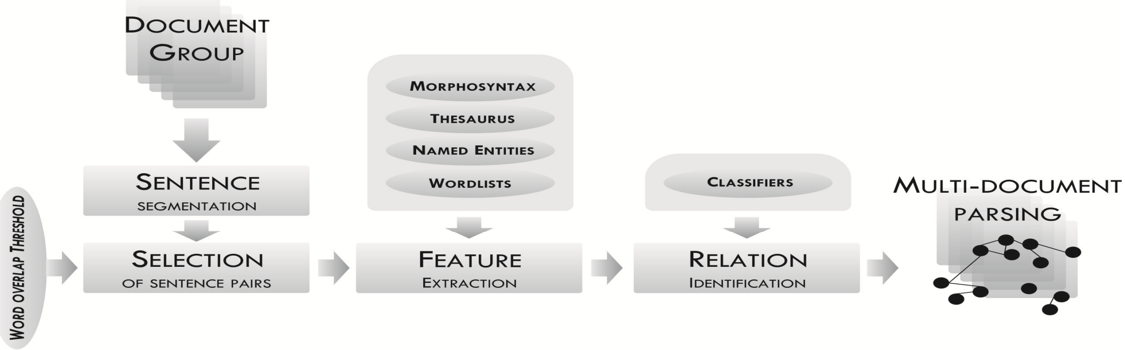

Figure 2. Multi-document Parser Architecture

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 (Zhang & Radev, 2004). This is also a strategy to make the task tractable (avoiding a combinatorial explosion), since it is virtually possible to find a relation between every pair of sentences. We use a threshold of 0.12, since this was the value used for English (Zhang & Radev, 2004) and that showed to be good for Portuguese too (Aleixo & Pardo, 2008a).

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. 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. It is interesting to notice that this graph is probably not fully connected, since not all sentence pairs present CST relations.

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. Table 1 shows the features/attributes extracted from each sentence pair used to generate the classifiers. After this extraction, all features are normalized to avoid possible classification discrepancies.

Table 1. Classifier attributes

1. Difference in word size (S1-S2)
2. Percentage of words in common in S1
3. Percentage of words in common in S2
4. Position of S1 in text (0- beginning, 2- end, 1- middle)
5. Position of S2 in text (0- beginning, 2- end, 1- middle)
6. Number of words in the longest substring between S1 and S2
7. Difference in the number of nouns between S1 and S2
8. Difference in the number of adverbs between S1 and S2
9. Difference in the number of adjectives between S1 and S2
10. Difference in the number of verbs between S1 and S2
11. Difference in the number of proper nouns between S1 and S2
12. Difference in the number of numerals between S1 and S2
13. Difference in the number of verbs of attribution between S1 and S2
14. Number of possible synonyms in common in S1 and S2
15. Number of matching named entities in S1 and S2

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 was employed a part of speech tagger for Brazilian Portuguese (Aires *et al.*, 2000), 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* (Bick, 2000), which lemmatize verbs as well. The lemmatization was necessary to perform a search in a list of verbs of attribution to calculate the value of the attribute 13. This feature aims particularly at identifying the relations *Attribution* and *Citation*, where there is an attribution indicator, especially a verb of attribution or other indicator of attribution, such as “according to…”, “as stated by…”, etc. As to the feature 14, a synonym database was used (Maziero *et al.*, 2008). In this feature, a list of synonym identifiers was compiled for each word, ignoring the *stopwords*. After matching the lists for each word from each sentence, the feature value was generated. The synonym database is fundamental to identify overlap when the words are not identical, but belong to the same set of synonyms. It is important to notice that, at this moment, the syntactic parser could be replaced by a part of speech tagger and a lemmatizer. However, in future work, we envision the creation of syntax-based symbolic rules for identifying some relations. So, the syntactic analysis, used to generate some features, is stored to be used in the future. The last feature was calculated after the processing performed by the named entity recognizer Rembrandt (Cardoso, 2008), which identified in each sentence pair under analysis the common named entities to generate the feature value. This sharing of named entities may indicate that the sentences are about the same topic.

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 (Carletta, 1996) for this task were computed for the following 3 items: relations, directionality of relations and type of relations (according to the typology shown in Figure 1). The results were 0.50, 0.44 and 0.61, respectively (zero represents the worst case and 1 the optimal agreement). Given the subjectivity of this task, such values are considered good.

In each experiment that was carried out in this paper, classifiers were developed and compared using the tool WEKA (Witten & Frank, 2005). In this comparison, a paired T-test was applied (with a 95% confidence interval) to point out the best classifier. The techniques to develop the classifiers were: NaiveBayes, Support Vector Machine (SVM), and decision tree (J48). NaiveBayes is a probabilistic technique; SVM is mathematical; and J48 is symbolic. 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: one considering all CST relations and other with just some content relations – the most frequent ones. Given that more than one relation may happen for a sentence pair, a multi-label classifier (it identifies more than one class for the same instance) was developed. The typology of CST relations also allowed for the development of two hierarchical classifiers (they take into consideration the hierarchy). 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.

The decision for developing some classifiers that considered only the content relations – which were the most frequent ones in the corpus – was due to the unbalanced data found in the corpus (the frequency of relations is shown in Figure 3). Table 2 displays the results from the classifier for each content relation (using the algorithm J48), except for *Summary*, and *Contradiction*, given their low frequency in the corpus.

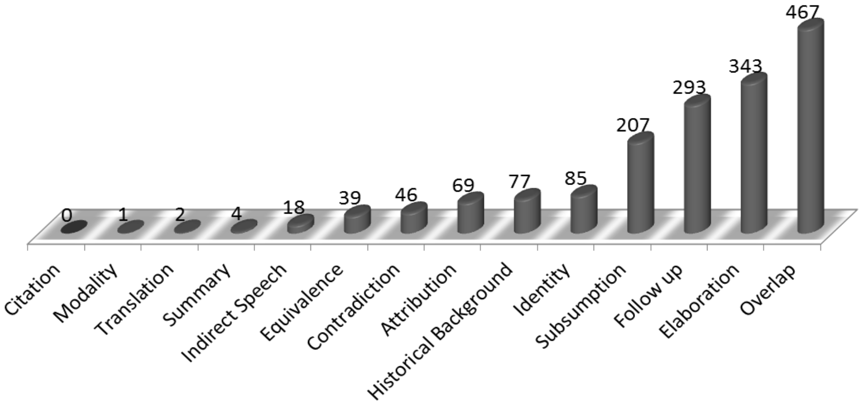

Figure 3. Frequency of relations in the CSTNews corpus

The relations *Overlap* and *Subsumption* returned f-measures higher than 49%, for they are very frequent in the corpus. As to *Historical background*, its value is 23%, since its frequency is low. Moreover, due to its characteristics, this relation was not satisfactorily identified by the employed features. A 40.2% general accuracy was obtained (for all the relations, differently from the numbers from Table 2, which were computed for each class independently).

Table 2. Results from the multi-class classifier for content relations

Relation	Precision	Coverage	F-Measure
Subsumption	0.485	0.560	0.520
Elaboration	0.386	0.382	0.384
Overlap	0.486	0.503	0.494
Follow-up	0.317	0.287	0.301
Historical background	0.243	0.221	0.231
Identity	0.941	0.941	0.941
Equivalence	0.207	0.154	0.176
Average	0.438	0.435	0.435

Exploring the hierarchy of the typology of relations, two hierarchical classifiers have been developed according to the Top-Down and Big-Bang approaches. For a comparison between these approaches, see Freitas & Carvalho (2007). In the Top-Down approach, a classifier is used at each level of the typology. For example, at the first level, a classifier is used to choose between "content" and "form" groups. Supposing that the "content" is the selected group, another classifier is then used to choose between "redundancy", "complement" and "contradiction" subgroups, and so on for each typology branch. When the lowest level of the hierarchy is reached, the process ends with the choice of a CST relation. Table 3 shows the results for each classifier produced according to the Top-Down approach, using the J48 technique. One may see, for instance, that the first classifier (classifier A) decides if a sentence pair contains a "content" or a "form" relation with f-measures of 95.3% and 48.8% for each of these classes, respectively. All average f-measures were higher than 45%. This proves the potential of this approach to identify relations and corroborates that the CST typology makes sense. However, some relations still produced very low results, such as *Modality*, *Translation* and *Summary*.

Regarding the unbalanced data for CST relations, when these are grouped according to the typology discussed, the scenario becomes better. When "form" group is chosen, for instance, the examples in this branch will not be so unbalanced, since the relations of this type will be judged separately from the content relations (by another classifier). It is interesting to notice that, when the "form" branch is followed and one relation of this group is chosen, it is worth following the "content" group too, since "form" relations usually happen with "content" relations. This Top-Down method allows such strategy to be applied. Following the presented procedure, the classifiers were combined and evaluated, obtaining 35.3% of general accuracy (differently from the numbers in Table 3, which were computed for each class independently).

The alternative hierarchical method – Big-Bang – performs the relation identification in just one step, taking into account the relation hierarchy as a whole. This approach consists of an alteration in the C4.5 algorithm (Quinlan, 1993), described by Clare (2003). Using this method, a general accuracy of 58.7% was obtained for the classification. However, differently from the previous approach, the Big-Bang technique may not reach the leaves of the hierarchy, stopping the classification when it is more advantageous. At the moment, this strategy does not fit well with our intended task, but is nonetheless interesting, since it may indicate a relation type instead of the relations themselves.

Finally, for some relations, binary classifiers were also developed. These classifiers test whether a sentence pair presents a specific relation or "other" relation. This way, each sentence pair must be checked by each classifier, i.e., the existence of each relation is checked independently. When used by the parser, each classifier will then assign a label to each sentence pair. For choosing a class, it will be necessary only to decide among labels different from "other". As a selection criterion, a confidence value provided by the classifier is used in the choice of a class, and the option will be for the highest. The results for each binary classifier (using the J48 machine learning method) are shown in Table 4.

Table 3. Hierarchical Classifiers

Classifier A – content vs. form

Relation	Precision	Coverage	F-Measure
Content	0.933	0.974	0.953
Form	0.640	0.394	0.488
Average	0.786	0.684	0.720

Classifier B – redund. vs. compl. vs. contradiction

Relation	Precision	Coverage	F-Measure
Redundancy	0.696	0.701	0.699
Complement	0.661	0.679	0.670
Contradiction	0.045	0.022	0.029
Average	0.470	0.467	0.466

Classifier C – total vs. partial

Relation	Precision	Coverage	F-Measure
Total	0.953	0.985	0.969
Partial	0.905	0.742	0.815
Average	0.929	0.864	0.892

Classifier D – identity vs. equivalence vs. summary

Relation	Precision	Coverage	F-Measure
Identity	0.882	0.965	0.921
Equivalence	0.800	0.718	0.757
Summary	0	0	0
Average	0.561	0.561	0.559

Classifier E – subsumption vs. overlap

Relation	Precision	Coverage	F-Measure
Subsumption	0.609	0.541	0.573
Overlap	0.806	0.846	0.825
Average	0.707	0.693	0.699

Classifier F – temporal vs. non-temporal

Relation	Precision	Coverage	F-Measure
Temporal	0.440	0.286	0.346
Non-temporal	0.851	0.918	0.884
Average	0.654	0.602	0.615

Classifier G – follow-up vs. historical back.

Relation	Precision	Coverage	F-Measure
Follow-up	0.871	0.922	0.896
Hist. back.	0.617	0.481	0.540
Average	0.744	0.701	0.718

Classifier H – source vs. style

Relation	Precision	Coverage	F-Measure
Source	0.853	0.829	0.841
Style	0.400	0.444	0.421
Average	0.626	0.636	0.631

Classifier I – attribution vs. modality

Relation	Precision	Coverage	F-Measure
Attribution	0.986	1	0.993
Modality	0	0	0
Average	0.493	0.500	0.496

Classifier J – indirect speech vs. translation

Relation	Precision	Coverage	F-Measure
Ind. speech	0.895	0.944	0.919
Translation	0	0	0
Average	0.447	0.472	0.459

The relations *Citation*, *Modality*, *Translation*, *Summary*, *Indirect speech*, and *Identity* were not considered due to their low frequency in the corpus. It is possible to see in the table that the average f-measures were above 65%, showing that this approach is promising to identify the most frequent relations in a corpus, i.e., it produces better results than the other scenarios tested. Obviously, when binary classifiers are created, the class "other" (which contains all other relations in the corpus, except for that the binary classifier aims to identify) becomes prevalent (that is, it contains a higher number of examples) and obtains higher results. Except for the *Contradiction* relation, the f-measure values for the relations in each classifier were above 48%. For example, the classifier of *Attribution* identifies this relation with a precision of 60%. Although the relation "other" is identified with more than 90%, the precision in the identification of *Attribution* is higher than in the multi-class classifier for all CST relations (not shown in this paper). The set of classifiers achieved a general accuracy of 27.5%. Such approach would also allow the identification of both "form" and "content" relations. Supposing that a "form" relation would be better scored and selected for a sentence pair, it would be possible to select the "content" relation that was best scored.

Table 4. Binary Classifiers

Attribution classifier

Relation	Precision	Coverage	F-Measure
Other	0.950	0.976	0.963
Attribution	0.600	0.413	0.489
Average	0.775	0.694	0.726

Follow-up classifier

Relation	Precision	Coverage	F-Measure
Other	0.814	0.887	0.848
Follow-up	0.668	0.529	0.590
Average	0.741	0.708	0.719

Contradiction classifier

Relation	Precision	Coverage	F-Measure
Other	0.957	0.994	0.976
Contradiction	0.700	0.228	0.344
Average	0.828	0.611	0.660

Historical Background classifier

Relation	Precision	Coverage	F-Measure
Other	0.953	0.968	0.960
Hist. back.	0.608	0.513	0.556
Average	0.780	0.740	0.758

Elaboration classifier

Relation	Precision	Coverage	F-Measure
Other	0.810	0.844	0.827
Elaboration	0.677	0.622	0.648
Average	0.743	0.733	0.737

Overlap classifier

Relation	Precision	Coverage	F-Measure
Other	0.759	0.763	0.761
Overlap	0.697	0.693	0.695
Average	0.728	0.728	0.728

Equivalence classifier

Relation	Precision	Coverage	F-Measure
Other	0.972	0.986	0.979
Equivalence	0.593	0.410	0.485
Average	0.782	0.698	0.732

Subsumption classifier

Relation	Precision	Coverage	F-Measure
Subsumption	0.749	0.698	0.723
Other	0.915	0.933	0.924
Average	0.832	0.815	0.823

Nowadays, the classifiers are attributing only one relation to each sentence pair. This is the cause for the low general accuracy values (despite of the good values for precision and recall for each class), since sentence pairs with more than one relation (which is usual) will have only one relation detected (the other possible relations will be considered as misclassified instances).

Compared with the results obtained by Zhang *et al.* (2003) and Zhang & Radev (2004), the values of the classifiers investigated in this research were better, demonstrating the potential of the developed methodology. It is worth emphasizing that these are results obtained for different languages and training corpora, even though this comparison may show possible paths to improve results in the field, also giving an idea of the state of the art in the area. The results of this work also show that the hierarchical classification of multi-document relations is promising. Its results, as well as the results for the binary classifiers, were better than the ones reported by Maziero *et al.* (2010). It is also worth noticing that decision trees (J48) produced the best results in all the cases. SVM also did well in some cases, with statistical tests showing that the differences were not significant.

Finally, we replicated all the previous 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. Table 5 shows the general accuracy for the classifiers with and without data balancing. One may see the improved results. However, as it happens for overfitting, we expect that these classifiers have a poor performance on unseen data.

Table 5. General accuracies for unbalanced and balanced data

Classification strategy	Unbalanced data	Balanced data
Multi-class	40.2%	72.5%
Hierarchical	35.3%	75.6%
Binary	27.5%	72.4%

For comparison and validation procedures, we also simulated a baseline method for parsing that simply assigns the most frequent relation (*Overlap*) to every sentence pair. It produces a general accuracy of 28.3%, which is outperformed by some 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.

5. Conclusions and future work

This work will allow for the automatic handling of multiple documents in Portuguese. Both users and computer applications will benefit from it. Regarding the techniques employed to develop the parser, as some relations have a low frequency in the corpus, symbolic rules to identify these relations are being manually prepared. Some rules look simple to develop, since some relations have clear signals in text. For instance, the *Indirect speech* relation might be identified by finding its direct counterpart, usually marked by a dash symbol in the text, and the *Translation* relation might be easily detected by using a multi-lingual dictionary. We believe that approaching the problem with classifiers and rules in a hybrid approach may produce better results. In fact, we think that the best approach is to use classification for the “content” relations and the rules to “form” relations and some low frequency “content” relations (e.g., the *Contradiction* relation). This way, the multi-label problem might be naturally solved, since these two strategies might be simultaneously applied.

Acknowledgments

The authors are grateful to FAPESP for supporting this work.

References

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.

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.

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.

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.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* PhD thesis. Aarhus University. Denmark University Press.

Cardoso, N. (2008). REMBRANDT – Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In C. Mota e D. Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

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.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, Vol. 22, N. 2, pp. 249-254.

Clare, A. (2003). *Machine learning and data mining for yeast functional genomics*. PhD thesis. University of Wales Aberystwyth.

Freitas, A. and Carvalho, A.C.P.F. (2007). A Tutorial on Hierarchical Classification with Applications in Bioinformatics. In D. Taniar (ed.), *Research and Trends in Data Mining Technologies and Applications: Advances in Data Warehousing and Mining*, pp. 176-209.

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

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.

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.

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.

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.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

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*.

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.

Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community*. PhD dissertation. Department of Computer Science, University of Maryland.

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.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

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.

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.

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.