Semi-Supervised Never-Ending Learning in Rhetorical Relation Identification

Erick G. Maziero[1,2], **Graeme Hirst**[1]
[1]Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
erick,gh@cs.toronto.edu

Thiago A. S. Pardo[2]
[2]Department of Computer Science
University of São Paulo
São Carlos, SP, 13566-590, Brazil
taspardo@icmc.usp.br

Abstract

Some languages do not have enough labeled data to obtain good discourse parsing, specially in the relation identification step, and the additional use of unlabeled data is a plausible solution. A workflow is presented that uses a semi-supervised learning approach. Instead of only a predefined additional set of unlabeled data, texts obtained from the web are continuously added. This obtains near human perfomance (0.79) in intra sentential rhetorical relation identification. An experiment for English also shows improvement using a similar workflow.

1 Introduction

A text is composed of coherent propositions (phrases and sentences, for example) ordered and connected according to the intentions of the author of the text. This composition may be recognized and structured according to many theories and this type of information is valuable to many natural language processing applications. A process to recognize, automatically, the coherent or discursive (or also rhetorical) structure of a text is named discourse parsing (DP).

The most prominent theory in Computational Linguistics to structure the discourse of a text is the Rhetorical Structure Theory (RST) proposed by Mann and Thompson (1987). In this theory, the text is segmented into elementary discourse units (EDUs), which each contain a proposition (basic idea) of the text. The theory proposes a set of rhetorical relations that may hold between the EDUs, explicating the intentions of the author. For example, consider the sentence in Figure 1. It is segmented into three EDUs, numbered from 1 to 3. EDUs 2 and 3 are related by the relation *Enablement*, forming a new span of text, which is related to 1 by the relation *Attribution*. In each relation, EDUs can be *Nucleus* (more essential) or *Satellite* to the writer's purpose.

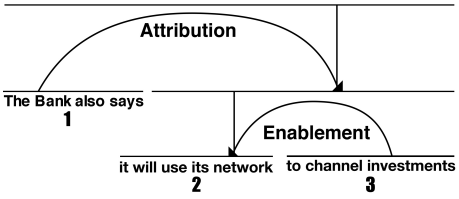

Figure 1: An example of sentence-level structure according to RST. From Soricut and Marcu (2003).

Many approaches have been used in DP, the majority of them using machine learning algorithms, such as probabilistic models (Soricut and Marcu, 2003), SVMs (Reitter, 2003; duVerle and Prendinger, 2009; Hernault et al., 2010; Feng and Hirst, 2012) and dynamic conditional random field (Joty et al., 2012). To obtain acceptable results, these approaches need plenty of labeled data. But even more than other levels of linguistic information, such as morphology or syntax, the annotation of discourse is an expensive task. Given this fact, what can we do when there is not enough data to perform effective learning of DP, as in languages with little annotated data?

This paper describes a methodology to overcome the problem of insufficient labeled data in the task of identifying rhetorical relations between

Proceedings of Recent Advances in Natural Language Processing, pages 436–442,
Hissar, Bulgaria, Sep 7–9 2015.

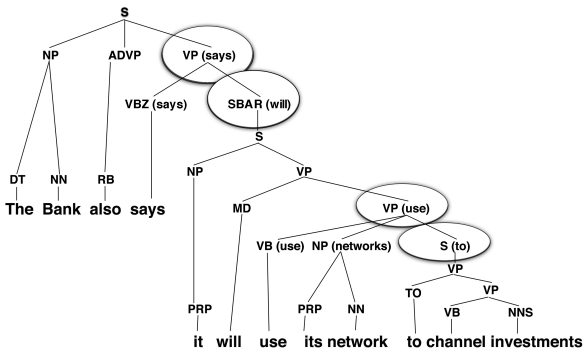

Figure 2: Lexicalized syntactic tree used by SPADE. The circles indicate the node used as the most indicative information to identify the rhetorical relation and structure.

EDUs, which is the most important step during DP. The language used in our work is Portuguese and two well-known systems of DP for English were adapted to this language. Portuguese is a language with insufficient annotated data to obtain a good discourse parser, but has all the tools to adapt some English discourse parsers. A framework of semi-supervised never-ending learning (SSNEL) (see Section 2.2 below) was created and evaluated with the adapted models. The results show that this approach improved the results to achieve near-human perfomance, even with the use of automatic tools (syntax parser and discourse segmenter).

2 Related Work

2.1 Supervised Discourse Parsing

Soricut and Marcu (2003) use two probabilistic models to perform a sentence-level analysis, one for segmentation and other to identify the relations and build the rhetorical structure. The parser is named SPADE (Sentence-level Parsing of DiscoursE) and the authors base their model on lexical and syntactic information, extracting features from a lexicalized syntactic tree. They assume that the features extracted at the jointing point of two discursive segments are the most indicative information to identify the rhetorical structure of the sentence. For example, in Figure 2, the circled nodes correspond to the most indicative cues to identify the structure and relation between each two adjacent segments.

The authors report a F-measure of 0.49 in a set of 18 RST relations. The human performance in this same task is 0.77 (measured by inter-annotation agreement). The authors, then, use the probabilistic model with manual segmentation and syntactic trees to see the impact of this information in the parsing and the model achieves 0.75.

Hernault et al. (2010) use support vector machine (SVM) classifiers to perform DP. This discourse parser is named HILDA (HIgh-Level Discourse Analyser). This work used a set of 41 rhetorical relations and achieves a F-measure of 0.48 in the step of relation identification, both intra-sentential and inter-sentential.

Feng and Hirst (2012) improve HILDA by incorporating new proposed features and some adapted from Lin et al. (2009). Another important decision was the specification of features for intra-sentential and inter-sentential relationships and the use of contextual features in the building of the rhetorical tree. Considering the approach to intra-sentential relation identification, with 18 RST relations this work achieves a macro average F-measure of 0.49 and weighted average F-measure of 0.77 in relation identification.

Joty et al. (2012) use a joint modelling approach to identify the structure and the relations at the sentence-level using DCRFs (dynamic conditional random fields) and a non-greedy bottom-up method in the construction of the rhetorical structure. The features used in this work were similar to those used by HILDA. They achieve a F-measure of 0.77, using manual segmentation, and 0.65 using automatic segmentation.

Some languages, such as Portuguese, do not have enough data to train a good DP and there is no work treating this limitation in this language. The first attempt to perform DP in Portuguese was made by Pardo and Nunes (2006), who used an approach based on lexical patterns extracted from an RST-annotated corpus of academic texts to create DiZer (Discourse analyZer). More than 740 lexical patterns were manually extracted from the corpus. A lexical pattern is composed of the discursive markers, its position in the EDU, and corresponding nuclearity. The use of lexical patterns is a unique approach for Portuguese, and achieves a F-measure of 0.625 in relation detection when evaluated in academic texts; in news texts, DiZer achieves an F-measure of 0.405.

2.2 Semi-supervised Discourse Parsing

All the above cited approaches to DP use annotated data to extract discursive knowledge and are limited to the availability of this resource, which is expensive to obtain. Specially, it is important to note that, to obtain good performance in the task more data is necessary. Semi-supervised learning (SSL) is employed in scenarios in which there is some labeled data and large availability of unlabeled data, and manual annotation is an expensive task (Zhu, 2008).

Related to the use of SSL in DP, Marcu and Echihabi (2002) used naive Bayes to train binary classifiers to distinguish between some types of relations, as *Elaboration* vs. *Cause-Explanation-Evidence*. For example, for this binary classifier, applying SSL, the accuracy increased from approximately 0.6 to 0.95 after the use of millions of new instances. Chiarcos (2012) used SSL to develop a probabilistic model mapping the occurrence of discourse markers and verbs to rhetorical relations. For Italian, Soria and Ferrari (1998) conducted work in the same direction. Sporleder and Lascarides (2005) performed similar work to Marcu and Echihabi, with similar results for a different set of relations and a more sophisticated classifier. Building on this, there is an interesting idea, known as never-ending learning (NEL) by Carlson et al. (2010), in which they apply SSL with infinite unlabeled data. The needed data is widely and freely available on the web. Their architecture runs 24 hours per day, forever, obtaining new information and performing a learning task.

With the aim of surpassing the limitation of labeled RST in Portuguese to develop a good DP, we employ SSNEL in the task by adapting the work of Soricut and Marcu (2003) and Hernault et al. (2010). This choice for SSLNEL was made considering the large and free availability of news texts on the web.

3 RST Corpora

RST-DT (RST Discourse TreeBank) (Carlson et al., 2001) is the most widely used corpus annotated with RST in English. Table 1 compares it with available Portuguese corpora labeled according to RST (these corpora will be referred to as RST-DT-PT hereafter). The corpora CSTNews (Cardoso et al., 2011), Summ-it (Collovini et al., 2007) and two-thirds of Rhetalho (Pardo and Seno, 2005) are composed of news texts, and the corpus CorpusTCC (Pardo and Nunes, 2004) and the reminder of Rhetalho are composed of scientific texts. The RST-DT contains more documents (45) and many more words (55,536) than RST-DT-PT.

Corpus	Language	Documents	Words
RST-DT-PT	PT	340	120,847
CSTNews		140	47,240
Rhetalho		50	2,903
Summ-it		50	16,704
CorpusTCC		100	53,000
RST-DT	EN	385	176,383

Table 1: Size of the RST-DT-PT and its components, and of the RST-DT.

This work focuses on the identification of rhetorical relations at the sentence level, and as is common since the work of Soricut and Marcu (2003), fine-grained relations were grouped: 29 sentence-level rhetorical relations were found and grouped into 16 groups. The imbalance of the relations is a natural characteristic in discourse and, to avoid overfitting of a learning model on the less-frequent relations, no balancing was made. The relation *Summary*, for example, occurs only 2 times, and *Elaboration* occurs 1491 times, making very difficult the identification of the *Summary* relation.

4 Adapted Models

Syntactic information is crucial in SPADE (Soricut and Marcu, 2003) and for Portuguese the parser most similar to that used by Soricut and Marcu is the LX-parser (Stanford parser trained to Portuguese (Silva et al., 2010)). After the parsing of the text by the syntactic parser, the same lexicalization procedure (Magerman, 1995) was applied and adapted according to the tagset used by LX-parser. In this adaptation, only pairs of adjacent segments at sentence-level were considered, and nuclearity was not considered, in order to avoid sparseness in the data. Training the adapted model (here called SPADE-PT) using the RST-DT-PT achieved F-measure of 0.30. The precision was 0.69, but the recall was only 0.19.

The same features used by HILDA (Hernault et al., 2010) were extracted from the pairs of adjacent segments at sentence-level and many machine learning algorithms were tested, besides the SVM, which was used in the original work. The algorithm which obtained the best F-measure was

J48, an implementation of decision trees (Quinlan, 1993). The RST-DT-PT corpora was used and the adaptation (here called of HILDA-PT) achieved an F-measure of 0.548, which is much better than that of SPADE-PT. A possible explanation is that the feature set in SPADE is composed only of syntactic tags and words. The resulting probabilistic model is sparse and many equal instances may indicate different relations (classes). However HILDA adds more features over which the classifier can work better, even when some values are absent.

Given the results of the adapted models, HILDA-PT was chosen to be incorporated into the SSNEL, explicated below.

5 Semi-supervised Never-ending Learning Workflow

Here, an adaptation of Carlson et al. (2010) self-training algorithm was used. Two different approaches to relation identification are used, that is to say, a lexical pattern set *LPS* (the relation identification module of DiZer), and a multi-class classifier C generated according to some machine learning algorithm. All the new instances obtained from the lexical module are used together with the more confident classifications of C to retrain this last. For each classification, J48 returns a confidence value used to choose the most confident classifications.

Also, there is interest in observing the behaviour of the classifier in each iteration of the semi-supervision, searching for the best F-measure it may achieve. In this way, a workflow of never-ending learning (NEL) was proposed and is presented in Figure 3. Workflow 1 is presented as an alternative visualization to the illustration in Figure 3. Continuously, a crawler gets pages from online news on the web and performs cleaning to obtain the main text (*Text*). In a first iteration, a *Segmenter* (Maziero et al., 2007) is applied to obtain the EDUs in each sentence and, for each pair of adjacent EDUs (*PairEDUs*), the C_1 classifier (C_1 initially trained with the $LabeledData_1$ from the RST-DT-PT) and the lexical pattern set *LPS* are used to identify the relations between the segments. To retrain C_1, all the new instances from the lexical pattern set *LabeledDataLPS* (as *LPS* does not provide a confidence value, all the labelled instances are

Data: $LabeledData_1$ and *Text*

train a classifier C_1 using $LabeledData_1$

while *exist some Text* **do**
  get one *Text* from *NewsTexts*
  apply *Segmenter* on *Text* to obtain *PairEDUs*
  $Index \leftarrow 1$
  forall the *PairEDUs* **do**
    apply *LPS* to obtain *LabeledDataLPS*
    apply C_{Index} to obtain *LabeledDataC*
    forall the *LabeledDataC as newInstanceC* **do**
      if *confidence of newInstanceC* ≥ 0.7 **then**
        $LabeledDataCConfident \leftarrow$ *newInstance*
      end
    end
    $LabeledData_{Index+1} \leftarrow$ *LabeledDataLPS*+ *LabeledDataCConfident*
    train a new classifier $C_{Index+1}$ using $LabeledData_{Index+1}$
    apply *Monitor* and obtain $FmC_{Index+1}$
    plot $FmC_{Index+1}$ in the graph G **if**
    $FmC_{index+1} < FmC_{Index}$ **then**
      discard $C_{Index+1}$
      $C_{Index+1} \leftarrow C_{Index}$
    end
  end
end

Workflow 1: Workflow of the SSNEL using two models to identify rhetorical relations between each *PairEDUs*.

used in the semi-supervision) and the classifications *LabeledDataC* with confidence greater than 0.7 by C_1 are joined with $LabeledData_1$ to obtain $LabeledData_2$ ($LabeledData_2 = LabeledDataLPS + LabeledDataC$). After the retraining, a *Monitor* verifies the new F-measure of C_2 (FmC_2, obtained using 10-fold cross validation) and, if it decreased compared with the F-measure of C_1 (FmC_1), C_2 is discarded and, for the next iteration, C_1 will continue to be used. If FmC_1 did not decrease, C_2 will be used in the next iteration. *Monitor* also plots a graph G to present the behaviour of FmC during SSNEL. This process continues iteratively.

It is important to note that, given the small size of the training data, we opted to use 10-fold cross-validation during the training and testing of the

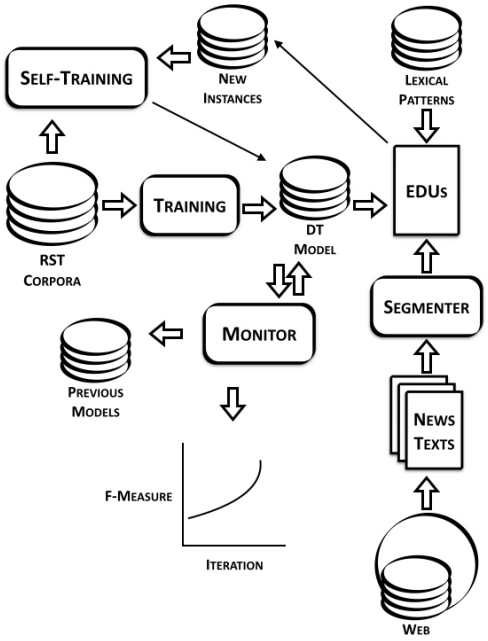

Figure 3: SSNEL workflow.

classifiers, instead of separating the data into three sets (training, development, and test). The total number of instances was 6163 and some relations, such as *Restatement* with 28 instances, would have few relations when split into three sets.

During the semi-supervision of SPADE-PT, the model of relation identification was incrementally obtained at each iteration, since the addition of a new instance only modifies the probabilities of the instances already present in the model. If the instance is new, it is added to the model and the probabilities are adjusted. However, in the semi-supervision of the HILDA-PT, the algorithm J48 does not allow incremental learning. There are some implementations of incremental decision trees, but the resulting models are not as accurate as J48 because they work with an incomplete set of training instances. As we want the best F-measure for relation identification, the algorithm J48 was employed, even though it is not an incremental learning.

Another important decision is to monitor the concept-drift (CD) (Klinkenberg, 2004) during the SSNEL, given that a concept may change over time. In this work, CD refers to different sources and topics to which the classifier is applied. To treat CD, the algorithm may detect the evolution of the concept and be able to modify the model to accommodate the concept, avoiding the decrease

Method	F-measure		Instances
	Initial	Final	
DiZer	0.22	-	-
Elaboration Relation	0.26	-	-
SPADE-PT	0.30	0.34	1,592
HILDA-PT	0.55	**0.79**	21,740

Table 2: Comparison of results considering the two adapted models (SPADE-PT and HILDA-PT) with two baselines (Elaboration Relation and DiZer).

in the performance of the model being generated. One technique to monitor the CD is *statistical process control* (SPC) (Gama et al., 2004). This technique constantly analyses the error during the learning: if the F-measure drops, it may indicate some changes in the concept and the model needs to be modified. In the SSNEL workflow, this is treated by the *Monitor*, which discards new instances used to retrain the model if its F-measure decreases, ensuring that the learned model always acquires correct new learning.

6 Experiments

Considering Workflow 1, the two adapted models were instantiated as *C*, and many iterations were executed. After 1,640 iterations and the addition of 1,592 new training instances, the F-measure of SPADE-PT increased only 0.05. HILDA-PT, after 180 iterations and with the addition of 21,740 new instances, increased 0.24, achieving 0.79 using automatic segmentation. Table 2 presents a summary of the results. As explained in Section 4, the features used by SPADE-PT lead to a sparse model (when there is not enough initial data), and this is the reason that, during 1,640 iterations, only 1,592 new instances were acquired, compared to the number of iterations and new instances during the experiment with HILDA-PT.

To evaluate the parsers, two baselines were considered. One of them (Elaboration Relation) is the labeling of all the instances with the most frequent relation in the corpus (*Elaboration*); the second is the use of *LPS* (DiZer) applied to all *PairEDUs* in RST-DT-PT. SPADE-PT, even after many iterations in SSNEL, performed lower than the two baselines. HILDA-PT, since even before the use of SSNEL, performed better than the baselines.

The class composed of relations *Interpretation*, *Evaluation* and *Conclusion* had 40 labeled exam-

ples, initially. After the iterations, its F-measure increased from 0.054 to 0.916. Except for *Comparison* and *Summary*, all the other relations increased their F-measures. This reinforces the results obtained by Marcu and Echihabi (2002), which increased (the result of a binary classifier to distinguish between two relations) from 0.6 to 0.95 after the use of millions of new instances. The relation *Summary*, however, with only 2 labeled instances, continued its zero F-measure.

SSNEL of HILDA-PT was executed for 23 days. Documents used had on average 28 sentences and 749 words. The choice of only 10 documents per batch is to have a fine-grained control over the new instances, given that if a new classifier decreases the F-measure, it is discarded. Out of 70 generated classifiers were discarded.

As the use of 10-fold cross-validation in the SSNEL may lead to some overfitting on the data which was already classified in the workflow, two other SSNEL experiments were performed, for English and Portuguese, with separated training and test sets. These experiments had less time to run, and, in order to determine whether the improvements during the SSNEL were statistically significant, paired T-tests were employed to compare initial classifier and the best classifier obtained during iterations in the workflow. The test shown improvements (at the level $p < .1$), even though they are low for both experiments. Probably, with many more iterations the results would be better. Table 3 shows the improvements in the accuracy during the SSNEL, the number of iterations, and the number of new instances incorporated in the training data. Although a direct comparison between the experiments is not fair, due to different corpora, the improvements show that this workflow is promising to increase the accuracy of classifiers with unlabeled data.

The experiment with SSNEL for English was realized in order to see the results that could be obtained when large annotated corpora are available. In the SSNEL for English, only decision-tree classifiers were used to classify new instances. For Portuguese, a symbolic model (lexical patterns) was also used together with the classifiers.

The improved results presented in Table 2 and 3 are very different due to differing evaluation strategies. Using separated test data, we tried to avoid possible overfitting on training data, but the size of test data may not lead to a fair evaluation of some relations with very few examples.

Experiment	Accuracy		Instances	Iterations
	Initial	Final		
Portuguese	0.531	**0.556**	1,247	200
English	0.635	**0.645**	565	25

Table 3: Results of SSNEL applied to Portuguese and English languages using training and test sets.

We do not compare our results to those of Soricut and Marcu (2003) or Joty et al. (2012), since HILDA-PT used different corpora (RST-DT-PT instead of RST-DT), and some reported results are for the complete DP. However, our results show the potential of the SSNEL workflow when not enough labeled data is available for supervised learning, since the same approach for relation identification of Hernault et al. (2010) was used in HILDA-PT and 0.531 was initially obtained. These results constitute the state of art for rhetorical relation identification for Portuguese and it is believed that with more time (iterations in SSNEL), the results may increase.

7 Conclusion

Even though the results obtained in the SSNEL were satisfactory, new features will be added to the HILDA-PT, for example, types of discourse signals, beyond the discourse markers (Taboada and Das, 2013), and the use of semantic information, as synonymity. Also, given that the number of features will increase, feature selection may be applied to select the most informative features in each iteration of the SSNEL.

Since this work treats only rhetorical relations, without nuclearity, a classifier of nuclearity was trained (with the same features of HILDA-PT) and obtained a F-score of 0.86. As done by Feng and Hirst (2012), a better set of features will be selected to identify relations between inter-sentential spans. A procedure similar to tree building used by Feng and Hirst (2012) will be employed in the future DP.

Acknowledgments

This work was financially supported by grant♯2014/11632-0, São Paulo Research Foundation (FAPESP), the Natural Sciences and Engineering Research Council of Canada and by the University of Toronto.

References

Paula C.F. Cardoso, Erick G. Maziero, Maria L.C. Jorge, Eloize M.R. Seno, Ariani Di Felippo, Lucia H.M. Rino, Maria G.V. Nunes, and Thiago A.S. Pardo. 2011. CSTNews: A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 85–105. Cuiaba/Brazil.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of Association for the Advancement of Artificial Intelligence*, volume 5, pages 1306–1313.

Lynn Carlson, Daniel Marcu, and Mary E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of Second SIGdial Workshop on Discourse and Dialogue*, volume 16, pages 1–10.

Christian Chiarcos. 2012. Towards the unsupervised acquisition of discourse relations. In *Proceedings of 50th Annual Meeting of the Association for Computational Linguistics*, pages 213–217.

Sandra Collovini, Thiago I. Carbonel, Jorge C.B. Coelho, Juliana T. Fuchs, and Renata Vieira. 2007. Summ-it: um corpus anotado com informações discursivas visando à sumarização automática. *Congresso Nacional da SBC*, pages 1605–1614.

David A. duVerle and Helmut Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proceedings of Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2, pages 665–673.

Vanessa W. Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 60–68.

João Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. 2004. Learning with drift detection. In *Proceedings of 17th Brazilian symp. on Artif. Intell. SBIA*, pages 286–295.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Shafiq Joty, Giuseppe Carenini, and Raymon T. Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 904–915. Association for Computational Linguistics, Stroudsburg, PA, USA.

Ralf Klinkenberg. 2004. Learning drifting concepts: Example selection vs. example weighting. *Intelligent Data Analysis*, 8(3):281–300.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1, pages 343–351.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of Association for Computational Llinguistics 1995*, pages 276–283. Cambridge, Massachusetts.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375.

Erick G. Maziero, Thiago A.S. Pardo, and Maria G.V. Nunes. 2007. Identificação automática de segmentos discursivos: o uso do parser Palavras. Technical Report 305, University of Sao Paulo.

Thiago A.S. Pardo and Maria G.V. Nunes. 2004. Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em Português do Brasil. Technical Report 231, University of Sao Paulo.

Thiago A.S. Pardo and Maria G.V. Nunes. 2006. Review and evaluation of DiZer: An automatic discourse analyzer for Brazilian Portuguese. In *Proceedings of 7th Workshop on Computational Processing of Written and Spoken Portuguese - PROPOR (Lecture Notes in Computer Science 3960)*, pages 180–189.

Thiago A.S. Pardo and Eloize R.M. Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente. In *Proceedings of V Encontro de Corpora*.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

David Reitter. 2003. Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models.

Joao Silva, António Branco, Sérgio Castro, and Reis Reis. 2010. Out-of-the-box robust parsing of portuguese. In *Proceedings of 9th International Conference on the Computational Processing of Portuguese*, PROPOR'10, pages 75–85. Springer-Verlag, Berlin, Heidelberg.

Claudia Soria and Giacomo Ferrari. 1998. Lexical marking of discourse relations - some experimental findings. In *Proceedings of ACL-98 Workshop on Discourse Relations and Discourse Markers*, pages 36–42.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 149–156.

Caroline Sporleder and Alex Lascarides. 2005. Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-05)*, pages 157–166. Bulgaria.

Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse*, 4(2):249–281.

Xiaojin Zhu. 2008. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin-Madison.