Adaptation of Discourse Parsing Models for the Portuguese Language

Erick G. Maziero[1,2], Graeme Hirst[1]
[1]Department of Computer Science
University of Toronto
Toronto, ON, M5S 3G4, Canada
Email: 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
Email: taspardo@icmc.usp.br

***Abstract*—Discourse parsing in Portuguese has two critical limitations. The first is that the task has been explored using only symbolic approaches, i.e., using manually extracted lexical patterns. The second is related to the domain of the lexical patterns, which were extracted through the analysis of a corpus of academic texts, generating many domain-specific patterns. For English, many approaches have been explored using machine learning with features based on a prominent lexicon-syntax notion of dominance sets. In this paper, two works were adapted to Portuguese, improving the results, outperforming the baselines and previous works for Portuguese, considering the task of rhetorical relation identification.**

I. 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, mainly to those that use deep linguistic information. A process to recognize, automatically, the coherent or discursive (or also rhetorical) structure of a text needs to be robust, given that this task has so much subjectivity. This process is named discourse parsing (DP).

The Rhetorical Structure Theory (RST) proposed by Mann and Thompson [1] is one of the most used discursive theories in Natural Language Processing (NLP). In RST, the text is segmented into elementary discourse units (EDUs) and these units are related to each other by relations explaining the coherence of the text. For example, consider one sentence in Figure 1. It is segmented into three EDUs, numbered from 1 to 3. EDUs 2 and 3 are related by the *Enablement* relation, forming a new span of text, that is related to 1 by the *Attribution* relation. In each relation, EDUs can be *Nucleus* (more essential) or *Satellites* to the writer's purpose.

DP has been addressed using different approaches, as the use of lexical patterns [3]–[7] and machine learning algorithms [2], [8]–[16]. The majority of the cited works are for the English language and have good results. However for Portuguese, DP has two critical limitations. The first is that the task was only approached with the use of lexical patterns, and the second is that the available discourse parser was obtained through the analysis of a corpus of academic texts [6], being domain-specific. To treat a new text genre, new lexical patterns must be extracted through corpus analysis, which is an expensive process.

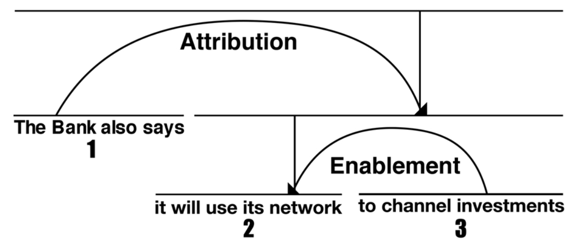

Fig. 1. Sentence-level structure according to RST. The leaves of the tree are the EDUs, which are related by rhetorical relations and they are defined as nucleus or satellite. The arrows depart from the satellite to the nucleus. Example extracted from [2].

To overcome these limitations, two important works [2], [14] were adapted to Portuguese in order to improve the F-score of DP. The chosen works introduced an important notion, called dominance sets, used in many works after [2], which achieved near human F-score in intra-sentential DP with manual segmentation and syntax analysis. The adaptation of [2] obtained a low F-score, but the use of lexicon-syntax features in combination with other superficial features obtained good results (as in [14]), in intra-sentential relation identification for Portuguese.

In the next Section, the adapted works are detailed. In Section 3, the corpora used in the adaptation are presented and the adaptation of [2] and [14] to Portuguese is detailed. Then, in Section 4, the performed experiments are discussed. Lastly, conclusions and future directions are presented.

II. Related work

Many approaches have been used in DP, the majority of them using machine learning algorithms, such as probabilistic models [2], SVMs [8], [10], [14], [15], and dynamic conditional random fields [16]. Soricut and Marcu [2] developed a discourse parser (called SPADE) which uses two probabilistic models, one to segment the text into EDUs, and another to identify the discursive relations between the segments. The features used in the models are based on the dominance sets, which represent the lexical and syntax information obtained

978-1-5090-0016-6/15 $31.00 © 2015 IEEE
DOI 10.1109/BRACIS.2015.24

in the attachment point between two EDUs. This work is limited to intra-sentential analysis and achieves human levels of performance in the task when the segmentation and syntax analysis are performed manually.

To obtain the dominance sets, the syntax tree is lexicalized, i.e., the internal nodes receive words of the sentence following the canonical lexical head projection rules, as in [17]. For each occurrence of a relation R, two probabilities are calculated, one with lexical information Pr and other only with the syntax labels Ps. Pr is the probability of a relation R given information θ_1. Ps is the probability of a structure S between two segments given information θ_2.

θ_1 and θ_2 are defined in Equation 1 and 2, respectively, where $S1$ and $S2$ are the segments related by the relation R. θ_1 encodes the lexical-syntax information in the attachment point between the segments (LH refers to Lexical Head and ST to Syntax Tag), and θ_2 encodes only the syntax information (ST) in the attachment point. LH_1 is extracted from the head word of $S1$ and LH_2, from the head word of $S2$. The dominance relation ($\prec$) between information of $S1$ and $S2$ indicates the order of the segments in the relationship.

$$\theta_1 = (S2, LH_2, ST_2) \prec (S1, LH_1, ST_1) \quad (1)$$

$$\theta_2 = (S2, ST_2) \prec (S1, ST_1) \quad (2)$$

To identify the R relation, the model uses Equation 3, choosing the maximum $Pr \times Ps$ among all the candidates in the training corpora.

$$R = argmax(\prod Pr \times Ps) \quad (3)$$

The authors report a F-score of 0.49 in a set of 18 RST relations (some of them are groups of two or more rhetorical relations). The human performance, in this same task was 0.77. The authors, then, use the probabilistic model with manual segmentation and syntactical trees to see the impact of this information in the DP, and the result increased to 0.75, indicating that segmentation and syntactical parsing are important in the task. It is important to note that the cited results are related to DP for each sentence, including segmentation, relation identification and rhetorical tree building.

Hernault et al. [14] developed the HILDA (High-Level Discourse Analyser) parser, which uses an expanded set of features based on the notion of dominance sets [2] and superficial features to train SVM classifiers to analyze the entire text, not only each sentence separately. In relation labeling, HILDA achieves an average F-score of 0.47.

For Portuguese, there is only a symbolic approach (called DiZer) based on lexical patterns to identify the discursive relations [7], [18]. The lexical patterns were extracted from a corpus of scientific texts, called CorpusTCC [19], forming a set of more than 700 patterns. Table I contains an example of a lexical pattern to identify the *Cause-Result* relation (grouping *Volitional-result*, *Volitional-cause*, *Non-volitional-result* and *Non-volitional-cause* relations) which is composed of discursive markers, their position in the EDUs and nuclearity. For example, if in the first segment the marker *Dado que* (Given that) occurs in the beginning, and in the beginning of the second segment the marker *possvel* (is possible) occurs, the relation *Cause-Result* will be chosen with the first segment as satellite and the second as nucleus of the relation.

	Segment 1	**Segment 2**
Discursive marker	*Dado que*	*possvel*
Marker position	Beginning	Beginning
Nuclearity	Satellite	Nucleus

TABLE I
EXAMPLE OF LEXICAL PATTERN FROM DIZER. THE PATTERN SPECIFIES THREE FIELDS AND THEIR CONTENTS IN THE DETECTION OF A *Cause-Result* RELATION BETWEEN TWO EDUS.

In [18], this approach achieves a F-score of 0.625 in relation detection when evaluated with academic texts. When evaluated in news texts, it achieves a F-score of 0.405, given that many patterns, created from academic texts, do not generalize well to other domains. When evaluated in the test set used in the experiments reported in this paper, the F-score decreased to 0.22, as will be explained latter.

III. MACHINE LEARNING DP FOR PORTUGUESE

A. Corpora used

Before describing the adaptation of the models, the data used in the experiments is detailed briefly. The RST set of corpora in Portuguese is composed of the CSTNews corpus [20], Summ-it [21], and two-thirds of Rhetalho [22], which are composed of news texts, and the corpus CorpusTCC [19] and one-third of Rhetalho, which are composed of scientific texts. In Table II the number of documents and words per corpus are presented.

This work is focused on the identification of rhetorical relations at the sentence-level, and as is common since the work of [2], the relations were grouped according to Table III-A. At sentence-level relationship, 29 rhetorical relations were found and grouped into 16 groups, following the work of [2] and [1]. The imbalance of the relations in discourse parsing is a natural characteristic, 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 the identification of the *Summary* relation very difficult. The examples of sentence-level rhetorical relations were separated into training and test set, following a stratified proportion of 7/10 for training and 3/10 for test.

B. Adaptation process

The adaptation of SPADE is hereafter called SPADE-PT, and the first step in the adaptation process was the choice of a syntax parser. The often-used parser Palavras [23] uses a formalism (constraint grammar) different to the traditional grammars in syntax parsers, like grammars used by [24] and [25], and produces flatter dependency trees, making the

Corpus	Documents	Words
CSTNews	140	47,240
Rhetalho	50	2,903
Summ-it	50	16,704
CorpusTCC	100	53,000
Total	340	119,847

TABLE II
NUMBER OF DOCUMENTS AND WORDS IN THE SET OF CORPORA FOR PORTUGUESE (COMPOSED OF CSTNEWS, RHETALHO, SUMM-IT AND CORPUSTCC).

Relation	Frequency
Attribution	799
Antithesis Concession Contrast	256
Background Circumstance	362
Volitional-Result Non-Volitional-Result Volitional-Cause Non-Volitional-Cause	449
Comparison	37
Condition Otherwise	104
Elaboration	1491
Enablement Motivation Purpose	695
Evidence Justify Explanation	194
Interpretation Evaluation Conclusion	40
List	703
Means	73
Restatement	28
Same-unit	731
Sequence	199
Summary	2

TABLE III
GROUPING OF 29 RHETORICAL RELATIONS INTO 16 GROUPS AND THEIR FREQUENCIES. THESE ARE SENTENCE-LEVEL RELATIONS FOUND IN RST-DT-PT

extraction of the dominance sets difficult. Therefore, the LX-parser [26], which is based on the Stanford parser [25], was used in this work. The lexical head projection had to be adapted according to the set of tags used by LX-parser. For example, the lexicalized syntax tree of the text in Figure 1 is presented in Figure 2.

The attachment point (indicated by circles) between EDUs 1 and 2 contains the following dominance set: $(2, SBAR) \prec$

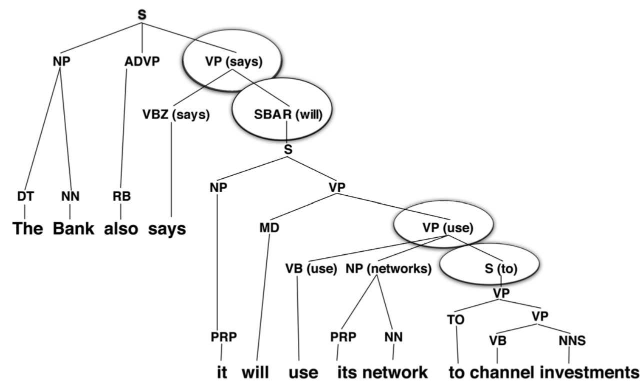

Fig. 2. Lexicalized syntactic tree used by SPADE. The circles indicate the node used as the most indicative information to identify the rhetorical relation and structure.

$(1, VP)$, indicating that EDU 2 is dominated by 1. After lexicalization of the tree, we can add lexical information to the dominance set: $(2, SBAR, will) \prec (1, VP, says)$. To avoid sparseness during the learning, given the size of the corpora, we opted to work only with one pair of EDUs per relation. Therefore, the tree in Figure 2 generates two dominance sets, between EDUs 1 and 2 and between 2 and 3. As a syntax tree is given for each sentence, only intra-sentential relations were considered.

The results obtained in this adaptation were very low, possibly due to insufficient amount of annotated data. So, we adapted HILDA, which expands the information (dominance sets) used by SPADE.

HILDA uses the feature set shown in Table IV. The first group of features (textual organization) uses tokens and EDU information, like distances of EDUs (in number of tokens and EDUs) to the beginning of the sentence and to the beginning of the text. The second group (related to the dominance sets) uses the lexicalized syntax tree to extract POS tags and lexical heads of the attachment points between EDUs. The scope of each feature may be for each EDU (E), or for the pair (P) of EDUs. This method also was adapted only to intra-sentential relations and is hereafter called HILDA-PT.

[14] used SVM to create classifiers to identify the rhetorical relations. But, during the adaptation, with some experiments, it was found that the decision tree algorithm J48 [27] performed better than SVM, and it was chosen to create the classifiers to identify the rhetorical relations. To use this type of machine learning algorithm, the string features need to be converted to numerical values, and, during this procedure, some generalizations were made. For example, words with numbers, symbols or punctuation were replace by labels (NUM for numbers, SYM for symbols, and PUNC for punctuation) in order to decrease the size of the word vector.

IV. EXPERIMENTS

For comparison, two baselines were considered. One was obtained by labeling each pair of segments with the *Elaboration*

Feature name	Scope
Textual organization	
Same sentence	P
Same paragraph	P
Number of sentence boundaries	E
Number of paragraph boundaries	E
Length in tokens	E
Length in EDUs	E
Distance to beginning of sentence in tokens	E
Size of span over sentence in tokens	E
Size of span over sentence in EDUs	E
Size of both spans over sentence in EDUs	P
Distance to beginning of sentence in EDUs	E
Distance to beginning of text in tokens	E
Distance to end of sentence in tokens	E
Syntax - dominance sets	
Distance to the root of syntax tree	E
Distance to common ancestor in syntax tree	E
Delta of distances to common ancestor	P
Dominating node's lexical head in span	E
Common ancestor's POS tag	P
Common ancestor's lexical head	P
Dominating node's POS tag	P
Dominating node's lexical head	P
Dominated node's POS tag	P
Dominated node's lexical head	P
Dominated node's sibling's POS tag	P
Dominated node's sibling's lexical head	P
Relative position of lexical head in sentence	E

TABLE IV
FEATURE SET USED IN HILDA ADAPTATION (HILDA-PT). THE FEATURES ARE GROUPED IN TWO SETS: TEXTUAL ORGANIZATION AND SYNTAX (RELATED TO DOMINANCE SETS).

relation, given that this is the most frequent relation in the corpora. The other baseline was the parser DiZer (the unique DP for Portuguese) for relation identification. The results obtained for intra-sentential rhetorical relation are presented in Table V.

SPADE-PT obtained a low F-score of 0.35, given that only 18% of the test set was classified, due to the sparseness of the generated model. The precision of this model was 0.53, but the recall was only 0.26. The *Elaboration* baseline had a F-score of 0.26 (which is the percentage of that relation in the test set), and performed better than DiZer baseline, which obtained a F-score of 0.22 in the test set of this experiment. DiZer had a good precision of 0.61, but the recall was very low (0.14), since the lexical patterns were extracted from an academic corpus and this experiment uses more news than academic texts.

One of the reasons for the low F-score of SPADE-PT is the overlapping in the generated model. Consider, for example, the following dominance set: $(2, CONJP, and) \prec (2, CONJP, or)$. This is used to identify both the *Restatement* relation and the group of relations formed by *Interpretation*,

Adaptation	F-score
SPADE-PT	0.35
HILDA-PT	**0.52**
Elaboration	0.26
DiZer	0.22

TABLE V
F-SCORE OF EACH ADAPTED METHOD (*SPADE-PT* AND *HILDA-PT*) AND BASELINES (*Elaboration* AND *DiZer*)

Relation	F-score
Attribution	0.550
Antithesis Concession Contrast	0.485
Background Circumstance	0.380
Volitional-Result Non-Volitional-Result Volitional-Cause Non-Volitional-Cause	0.229
Comparison	0.083
Condition Otherwise	0.361
Elaboration	0.654
Enablement Motivation Purpose	0.787
Evidence Justify Explanation	0.216
Interpretation Evaluation Conclusion	0.000
List	0.409
Means	0.000
Restatement	0.000
Same-unit	0.692
Sequence	0.094
Summary	0.000
All relations	**0.521**

TABLE VI
F-SCORE FOR EACH CLASS TREATED BY HILDA-PT, CONSIDERING THE PREVIOUS GROUPING OF RELATIONS, IN TABLE III-A. THE F-SCORE FOR ALL RELATIONS IS WEIGHTED ACCORDING TO THE FREQUENCY OF THE RELATIONS IN THE TEST SET.

Evaluation and *Conclusion*. The use of an expanded set of features (Table IV) decreased this problem and improved the F-score of the relation identification. HILDA-PT performed better than all the other methods, achieving a F-score of 0.52 and showing the potential of this approach. Better results should be obtained if more annotated data were available.

Considering the F-score of some relations in Table IV, the *Comparison* relation obtained a low result, given its low frequency in the test set (only 11 examples). The group formed by relations *Interpretation*, *Evaluation* and *Conclusion*, and the relations *Means*, *Restatement* and *Summary* have, respectively, 12, 21, 8 and 2 examples in the test set and these relations obtained zero F-score in the evaluation. The group formed by relations *Enablement*, *Motivation* and *Purpose* obtained better results (0.787) than the most frequent relation *Elaboration*, even though that group is less than half as frequent (208 examples) as *Elaboration* (447 examples).

The experiments related in this paper treat only relation identification. Aiming a complete DP, a classifier of nuclearity was trained (using the same feature set in Table IV) and obtained a F-score of 0.86 (close to 0.87, obtained by [16]).

V. Conclusion

The state-of-art in discourse parsing for Portuguese was advanced in this work, using supervised learning. It is important given that many applications have used discourse knowledge and may be fully automated.

Using the adapted models, a workflow using never-ending semi-supervised learning (SSNEL) was proposed, achieving near human F-score [28] for Portuguese language.

As some relations or group of relations reached low F-scores, a semi-supervised approach may be used to obtain new instances of the low frequency relations and improve their identification. Also, new features may be explored, such as various types of discourse signals (beyond discourse markers) proposed by [29] and the use of semantic knowledge, such as polarity and synonymy.

To treat deterministic occurrences of structural relations like *Parenthetical*, *Same-unit* and *Attribution*, rules may be created to be used along with the classifiers to improve their performance. For example, rules may be created to identify texts between parentheses and appositions (indicating *Parenthetical* relation) and lexicon-syntax patterns of attribution (indicating *Attribution* relation), among others.

As done by Feng and Hirst [15], a better set of features will be selected to identify relations at inter-sentential level. Also, a similar procedure to tree building used by Feng and Hirst [15] will be employed in the future DP, which will use the sentence-level relation identification related in this paper.

Acknowledgment

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

[1] W. Mann and S. Thompson, "Rhetorical Structure Theory: Toward a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1987.

[2] R. Soricut and D. Marcu, "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*, vol. 1, 2003, pp. 149–156.

[3] D. Marcu, "The rhetorical parsing, summarization, and generation of natural language texts," Ph.D. dissertation, University of Toronto, 1997.

[4] S. Corston-Oliver, "Computing representations of the structure of written discourse," Ph.D. dissertation, University of California, Santa Barbara, CA, USA, 1998.

[5] F. Schilder, *Robust discourse parsing via discourse markers, topicality and position*. Cambridge University Press, June 2002, vol. 8, pp. 235–255.

[6] T. Pardo, "Métodos para análise discursiva automática," Ph.D. dissertation, University of Sao Paulo, June 2005.

[7] E. Maziero, T. Pardo, I. da Cunha, J. Torres-Moreno, and E. SanJuan, "DiZer 2.0 - an adaptable on-line discourse parser," in *Proceedings of 3rd RST Brazilian Meeting*, 2011, pp. 1–17.

[8] D. Reitter, "Rhetorical analysis with rich-feature support vector models," Master's thesis, University of Postdam, 2003.

[9] R. Mahmud and A. Ramsay, *Finding Discourse Relations in Student Essays*. Berlin, Heidelberg: Springer-Verlag, February 2005, ch. Finding Discourse Relations in Student Essays, pp. 116–119.

[10] D. duVerle and H. Prendinger, "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*, vol. 2, 2009, pp. 665–673.

[11] K. Sagae, "Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing," in *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*. Paris, France: Association for Computational Linguistics, October 2009, pp. 81–84.

[12] R. Subba and B. Di Eugenio, "An effective discourse parser that uses rich linguistic information," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, 2009, pp. 566–574.

[13] B. Wellner, J. Pustejovsky, C. Havasi, A. Rumshisky, and R. Saur, "Classification of discourse coherence relations: An exploratory study using multiple knowledge sources," in *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 2009, pp. 117–125.

[14] H. Hernault, H. Prendinger, D. duVerle, and M. Ishizuka, "HILDA: A discourse parser using support vector machine classification," *Dialogue and Discourse*, vol. 1, no. 3, pp. 1–33, 2010.

[15] V. Feng and G. Hirst, "Text-level discourse parsing with rich linguistic features," in *Proceedings of 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2012, pp. 60–68.

[16] S. Joty, G. Carenini, and R. T. Ng, "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*, ser. EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 904–915.

[17] D. Magerman, "Statistical decision-tree models for parsing," in *Proceedings of Association for Computational Llinguistics 1995*, Cambridge, Massachusetts, June 1995, pp. 276–283.

[18] T. Pardo and M. Nunes, "On the development and evaluation of a Brazilian Portuguese discourse parser," *Journal of Theoretical and Applied Computing*, vol. 15, pp. 43–64, 2008.

[19] T. Pardo, "Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em Português do Brasil," University of São Paulo, Tech. Rep. 231, 2004.

[20] P. Cardoso, E. Maziero, M. Jorge, E. Seno, A. Di Felippo, L. Rino, M. Nunes, and T. Pardo, "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*, Cuiaba/Brazil, October 2011, pp. 85–105.

[21] S. Collovini, T. Carbonel, J. Coelho, J. Fuchs, and R. Vieira, "Summ-it: um corpus anotado com informações discursivas visando à sumarização automática," *Congresso Nacional da SBC*, pp. 1605–1614, 2007.

[22] T. Pardo and E. Seno, "Rhetalho: um corpus de referência anotado retoricamente," in *Proceedings of V Encontro de Corpora*, 2005.

[23] E. Bick, "The parsing system "Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework," Ph.D. dissertation, Aarhus University., Denmark University Press, October 2000.

[24] E. Charniak, "A maximum-entropy-inspired parser," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, ser. NAACL 2000. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 132–139.

[25] D. Klein and C. D. Manning, "Accurate unlexicalized parsing." in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.

[26] J. Silva, A. Branco, S. Castro, and R. Reis, "Out-of-the-box robust parsing of portuguese," in *Proceedings of 9th International Conference*

on the Computational Processing of Portuguese, ser. PROPOR'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 75–85.

[27] R. Quinlan, *C4.5: Programs for Machine Learning*. Sao Francisco, CA, USA: Morgan Kaufmann Publishers, Inc., 1993.

[28] E. Maziero, G. Hirst, and T. Pardo, "Semi-supervised never-ending learning in rhetorical relation identification," in *Proceeding of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, September 2015.

[29] M. Taboada and D. Das, "Annotation upon annotation: Adding signalling information to a corpus of discourse relations," *Dialogue and Discourse*, vol. 4, no. 2, pp. 249–281, 2013.