On the contribution of discourse structure to topic segmentation

Paula C. F. Cardoso[1], Maite Taboada[2], Thiago A. S. Pardo[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400 - Centro
Caixa Postal: 668 – CEP: 13566-970 – São Carlos/SP

[2]Department of Linguistics – Simon Fraser University
8888 University Dr., Burnaby, B.C., V5A 1S6 - Canada

pcardoso@icmc.usp.br, mtaboada@sfu.ca, taspardo@icmc.usp.br

Abstract

In this paper, we describe novel methods for topic segmentation based on patterns of discourse organization. Using a corpus of news texts, our results show that it is possible to use discourse features (based on Rhetorical Structure Theory) for topic segmentation and that we outperform some well-known methods.

1 Introduction

Topic segmentation aims at finding the boundaries among topic blocks in a text (Chang and Lee, 2003). This task is useful for a number of important applications such as information retrieval (Prince and Labadié, 2007), automatic summarization (Wan, 2008) and question-answering systems (Oh et al., 2007).

In this paper, following Hearst (1997), we assume that a text or a set of texts develop a main topic, exposing several subtopics as well. We also assume that a topic is a particular subject that we write about or discuss (Hovy, 2009), and subtopics are represented in pieces of text that cover different aspects of the main topic (Hearst, 1997; Hennig, 2009). Therefore, the task of topic segmentation aims at dividing a text into topically coherent segments, or subtopics. The granularity of a subtopic is not defined, as a subtopic may contain one or more sentences or paragraphs.

Several methods have been tested for topic segmentation. There are, however, no studies on how discourse structure directly mirrors topic boundaries in texts and how they may contribute to such task, although such possible correlation has been suggested (e.g., Hovy and Lin, 1998).

In this paper, we follow this research line, aiming at exploring the relationship of discourse and subtopics. In particular, our interest is mainly on the potential of Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) for this task. We propose and evaluate automatic topic segmentation strategies based on the rhetorical structure of a text. We also compare our results to some well-known algorithms in the area, showing that we outperform these algorithms. Our experiments were performed using a corpus of news texts manually annotated with RST and subtopics.

The remainder of this paper is organized as follows. Section 2 gives a brief background on text segmentation. Section 3 describes our automatic strategies to find the subtopics. The corpus that we use is described in Section 4. Section 5 presents some results and Section 6 contains the conclusions and future work.

2 Related work

Several approaches have tried to measure the similarity across sentences and to estimate where topic boundaries occur. One well-known approach, that is heavily used for topic segmentation, is TextTiling (Hearst, 1997), which is based on lexical cohesion. For this strategy, it is assumed that a set of lexical items is used during the development of a subtopic in a text and, when that subtopic changes, a significant proportion of vocabulary also changes.

Passoneau and Litman (1997), in turn, have combined multiple linguistic features for topic segmentation of spoken text, such as pause, cue words, and referential noun phrases. Hovy and Lin (1998) have used various complementary

Proceedings of the SIGDIAL 2013 Conference, pages 92–96,
Metz, France, 22-24 August 2013.

techniques for topic segmentation, including those based on text structure, cue words and high-frequency indicative phrases for topic identification in a summarization system. Although the authors do not mention an evaluation of these features, they suggested that discourse structure might help topic identification. For this, they suggested using RST.

RST represents relations among propositions in a text and discriminates nuclear and satellite information. In order to present the differences among relations, they are organized in two groups: subject matter and presentational relations. In the former, the text producer intends that the reader recognizes the relation itself and the information conveyed, while in the latter the intended effect is to increase some inclination on the part of the reader (Taboada and Mann, 2006). The relationships are traditionally structured in a tree-like form (where larger units – composed of more than one proposition – are also related in the higher levels of the tree).

To the best of our knowledge, we have not found any proposal that has directly employed RST for topic segmentation purposes. Following the suggestion of the above authors, we investigated how discourse structure mirrors topic shifts in texts. Next section describes our approach to the problem.

3 Strategies for topic segmentation

For identifying and partitioning the subtopics of a text, we developed four baseline algorithms and six other algorithms that are based on discourse features.

The four baseline algorithms segment at paragraphs, sentences, random boundaries (randomly selecting any number of boundaries and where they are in a text) or are based on word reiteration. The word reiteration strategy is an adaptation of TextTiling[1] (Hearst, 1997) for the characteristics of the corpus that we used (introduced latter in this paper).

The algorithms based on discourse consider the discourse structure itself and the RST relations in the discourse tree. The first algorithm (which we refer to as Simple Cosine) is based on Marcu's idea (2000) for measuring the "goodness" of a discourse tree. He assumes that a discourse tree is "better" if it exhibits a high-level structure that matches as much as possible the topic boundaries of the text for which that structure was built. Marcu associates a clustering score to each node of a tree. For the leaves, this score is 0; for the internal nodes, the score is given by the lexical similarity between the immediate children. The hypothesis underlying such measurements is that better trees show higher similarity among their nodes. We have adopted the same idea using the cosine measure. We have proposed that text segments with similar vocabulary are likely to be part of the same topic segment. In our case, nodes with scores below the average score are supposed to indicate possible topic boundaries.

The second algorithm (referred to as Cosine Nuclei) is also a proposal by Marcu (2000). It is assumed that whenever a discourse relation holds between two textual spans, that relation also holds between the most salient units (nuclei) associated with those spans. We have used this formalization and measured the similarity between the salient units associated with two spans (instead of measuring among all the text spans of the relation, as in the previous algorithm).

The third (Cosine Depth) and fourth (Nuclei Depth) algorithms are variations of Simple Cosine and Cosine Nuclei. For these new strategies, the similarity for each node is divided by the depth where it occurs, traversing the tree in a bottom-up way. These should guarantee that higher nodes are weaker and might better represent topic boundaries. Therefore, we have the assumption that topic boundaries are more likely to be mirrored at the higher levels of the discourse structure. We also have used the average score to find out less similar nodes. Figure 1 shows a sample RST tree. The symbols N and S indicate the nucleus and satellite of each rhetorical relation. For this tree, the score between nodes 3 and 4 is divided by 1 (since we are at the leaf level); the score between Elaboration and node 5 is divided by 2 (since we are in a higher level, 1 above the leaves on the left); and the score between Sequence and Volitional-result is divided by 3 (1 above the leaves on the right).

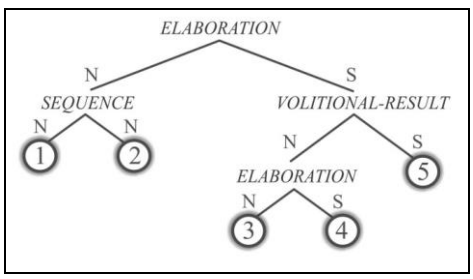

Figure 1. Example of an RST structure

[1] We have specifically used the block comparison method with block size=2.

The next algorithms are based on the idea that some relations are more likely to represent topic shifts. For estimating this, we have used the CSTNews (described in next section), which is manually annotated with subtopics and RST.

In this corpus, there are 29 different types of RST relations that may connect textual spans. In an attempt to characterize topic segmentation based on rhetorical relations, we recorded the frequency of those relations in topic boundaries. We realized that some relations were more frequent on topic boundaries, whereas others never occurred at the boundaries of topics. Out of the 29 relations, 16 appeared in the reference annotation. In topic boundaries, Elaboration was the most frequent relation (appearing in 60% of the boundaries), followed by List (20%) and Non-Volitional Result (5%). Sequence and Evidence appeared in 2% of the topic boundaries, and Background, Circumstance, Comparison, Concession, Contrast, Explanation, Interpretation, Justify, and Non-Volitional Cause in 1% of the boundaries.

We used this knowledge about the relations' frequency and attributed a weight associated with the possibility that a relation indicates a boundary, in accordance with its frequency on topic boundaries in the reference corpus. Figure 2 shows how the 29 relations were distributed. One relation is weak if it usually indicates a boundary; in this case, its weight is 0.4. One relation is medium because it may indicate a boundary or not; therefore, its weight is 0.6. On the other hand, a strong relation almost never indicates a topic boundary; therefore, its weight is 0.8. Such values were empirically determined. Another factor that may be observed is that all presentational relations are classified as strong, with the exception of Antithesis. This is related to the definition of presentational relations, and Antithesis was found in the reference segmentation with a low frequency.

Class	Relations
Weak (0.4)	Elaboration, Contrast, Joint, List
Medium (0.6)	Antithesis, Comparison, Evaluation Means, Non-Volitional Cause, Non-Volitional Result, Solutionhood, Volitional Cause, Volitional Result, Sequence
Strong (0.8)	Background, Circumstance, Concession, Conclusion, Condition, Enablement, Evidence, Explanation, Interpretation, Justify, Motivation, Otherwise, Purpose, Restatement, Summary

Figure 2. Classification of RST relations

From this classification we created two more strategies: Relation_Depth and Nuclei_Depth_Relation. Relation_Depth associates a score to the nodes by dividing the relations weight by the depth where it occurs, in a bottom-up way of traversing the tree. We also have used the average score to find out nodes that are less similar. As we have observed that some improvement might be achieved every time nuclei information was used, we have tried to combine this configuration with the relations' weight. Hence, we computed the scores of the Nuclei Depth strategy times the proposed relations weight. This was the algorithm that we called Nuclei_Depth_Relation. Therefore, these two last algorithms enrich the original Cosine Depth and Nuclei Depth strategies with the relation strength information.

The next section presents the data set we have used for our evaluation.

4 Overview of the corpus

We used the CSTNews corpus[2] that is composed of 50 clusters of news articles written in Brazilian Portuguese, collected from several sections of mainstream news agencies: Politics, Sports, World, Daily News, Money, and Science. The corpus contains 140 texts altogether, amounting to 2,088 sentences and 47,240 words. On average, the corpus conveys in each cluster 2.8 texts, 41.76 sentences and 944.8 words. All the texts in the corpus were manually annotated with RST structures and topic boundaries in a systematic way, with satisfactory annotation agreement values (more details may be found in Cardoso et al., 2011; Cardoso et al., 2012). Specifically for topic boundaries, groups of trained annotators indicated possible boundaries and the ones indicated by the majority of the annotators were assumed to be actual boundaries.

5 Evaluation

This section presents comparisons of the results of the algorithms over the reference corpus.

The performance of topic segmentation is usually measured using Recall (R), Precision (P), and F-measure (F) scores. These scores quantify how closely the system subtopics correspond to the ones produced by humans. Those measures compare the boundary correspondences without considering whether these are close to each other: if they are not the same (regardless of wheth-

[2] www2.icmc.usp.br/~taspardo/sucinto/cstnews.html

er they are closer or farther from one another), they score zero. However, it is also important to know how close the identified boundaries are to the expected ones, since this may help to determine how serious the errors made by the algorithms are. We propose a simple measure to this, which we call Deviation (D) from the reference annotations. Considering two algorithms that propose the same amount of boundaries for a text and make one single mistake each (having, therefore, the same P, R, and F scores), the best one will be the one that deviates the least from the reference. The best algorithm should be the one with the best balance among P, R, F, and D scores.

The results achieved for the investigated methods are reported in Table 1. The first 4 rows show the results for the baselines. The algorithms based on RST are in the last 6 rows. The last row represents the human performance, which we refer by topline. It is interesting to have a topline because it possibly indicates the limits that automatic methods may achieve in the task. To find the topline, a human annotator of the corpus was randomly selected for each text and his annotation was compared with the reference one.

As expected, the paragraph baseline was very good, having the best F values of the baseline set. This shows that, in most of the texts, the subtopics are organized in paragraphs. Although the sentence baseline has the best R, it has the worst D. This is due to the fact that not every sentence is a subtopic, and to segment all of them becomes a problem when we are looking for major groups of subtopics. TextTiling is the algorithm that deviates the least from the reference segmentation. This happens because it is very conservative and detects only a few segments, sometimes only one (the end of the text), causing it to have a good deviation score, but penalizing R.

Algorithm	R	P	F	D
TextTiling	0.405	0.773	0.497	0.042
Paragraph	0.989	0.471	0.613	0.453
Sentence	1.000	0.270	0.415	1.000
Randomly	0.674	0.340	0.416	0.539
Simple Cosine	0.549	0.271	0.345	0.545
Cosine Nuclei	0.631	0.290	0.379	0.556
Cosine Depth	0.873	0.364	0.489	0.577
Nuclei Depth	0.899	0.370	0.495	0.586
Relation_Depth	0.901	0.507	0.616	0.335
Nuclei_Depth Relation	0.908	0.353	0.484	0.626
Topline	0.807	0.799	0.767	0.304

Table 1. Evaluation of algorithms

In the case of the algorithms based on RST, we may notice that they produced the best results in terms of R, P, and F, with acceptable D values. We note too that every time the salient units were used, R and P increase, except for Nuclei_Depth_Relation. Examining the measures, we notice that the best algorithm was Relation_Depth. Although its F is close to the one of the Paragraph baseline, the Relation_Depth algorithm shows a much better D value. One may see that the traditional TextTiling was also outperformed by Relation_Depth.

As expected, the Topline (the human, therefore) has the best F with acceptable D. Its F value is probably the best that an automatic method may expect to achieve. It is 25% better than our best method (Relation_Depth). There is, therefore, room for improvements, possibly using other discourse features.

We have run t-tests for pairs of algorithms for which we wanted to check the statistical difference. As expected, the F difference is not significant for Relation_Depth and the Paragraph algorithms, but it was significant with 95% confidence for the comparison of Relation_Depth with Nuclei_Depth and TextTiling (also regarding the F values). Finally, the difference between Relation_Depth and the Topline was also significant.

6 Conclusions and future work

In this paper we show that discourse structures mirror, in some level, the topic boundaries in the text. Our results demonstrate that discourse knowledge may significantly help to find boundaries in a text. In particular, the relation type and the level of the discourse structure in which the relation happens are important features. To the best of our knowledge, this is the first attempt to correlate RST structures with topic boundaries, which we believe is an important theoretical advance.

At this stage, we opted for a manually annotated corpus, because we believe an automatic RST analysis would surely decrease the correspondence that was found. However, better discourse parsers have arisen and this may not be a problem anymore in the future.

Acknowledgments

The authors are grateful to FAPESP, CAPES, CNPq and Natural Sciences and Engineering Research Council of Canada (Discovery Grant 261104-2008) for supporting this work.

References

Paula C.F. Cardoso, Erick G. Maziero, Maria L.R. Castro Jorge, Eloize M.R. Seno, Ariani Di Fellipo, Lúcia H.M. Rino, Maria G.V. Nunes, Thiago A.S. Pardo. 2011. CSTNews – A discourse-annotated corpus for single and multidocument summarization of texts in Brazilian Portuguese. In: *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105.

Paula C.F. Cardoso, Maite Taboada, Thiago A.S. Pardo. 2013. Subtopics annotation in a corpus of news texts: steps towards automatic subtopic segmentation. In: *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

T-H Chang and C-H Lee. 2003. Topic segmentation for short texts. In: *Proceedings of the 17th Pacific Asia Conference Language*, pp. 159-165.

Marti Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics* 23(1), pp. 33-64.

Leonhard Hennig. 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. In: *Recent Advances in Natural Language Processing*, pp. 144-149.

Eduard Hovy and C-Y Lin. 1998. Automated Text Summarization and the SUMMARIST system. In: *Proceedings of TIPSTER*, pp. 197-214.

Eduard Hovy. 2009. Text Summarization. In: Ruslan Mitkov. *The Oxford Handbook of Computational Linguistics*, pp. 583-598. United States: Oxford University.

Anna Kazantseva and Stan Szpakowicz. 2012. Topical Segmentation: a study of human performance and a new measure of quality. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 211-220.

Willian C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Hyo-Jung Oh, Sung Hyon Myaeng and Myung-Gil Jang. 2007. Semantic passage on sentence topics for question answering. *Information Sciences* 177(18), pp. 3696-3717.

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics* 23(1), pp. 103-109.

Violaine Prince and Alexandre Labadié. 2007. Text segmentation based on document understanding for information retrieval. In: *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems*, pp. 295-304.

Maite Taboada and William C. Mann. 2006. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies* 8(3), pp.423-459.

Xiaojun Wan. 2008. An exploration of document impact on graph-based multi-document summarization. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 755-762.