Evaluation of Automatic Text Summarization Methods Based on Rhetorical Structure Theory

Vinícius Rodrigues Uzêda, Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes
Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
CP 668 – ICMC-USP, 13.560-970 São Carlos-SP, Brazil
http://www.nilc.icmc.usp.br
vruzeda@gmail.com,{taspardo,gracan}@icmc.usp.br

Abstract

Motivated by governmental, commercial and academic interests, automatic text summarization area has experienced an increasing number of researches and products, which led to a countless number of summarization methods. In this paper, we present a comprehensive comparative evaluation of the main automatic text summarization methods based on Rhetorical Structure Theory (RST), claimed to be among the best ones. We also propose new methods and compare our results to an extractive summarizer, which belongs to a summarization paradigm with severe limitations. To the best of our knowledge, most of our results are new in the area and reveal very interesting conclusions. The simplest RST-based method is among the best ones, although all of them present comparable results. We show that all RST-based methods overcome the extractive summarizer and that hybrid methods produce worse summaries. Finally, we verify that Mann and Thompson strong assumption in summarization and RST research area is not helpful in the way previously imagined.

1 Introduction

Motivated by governmental, commercial and academic interests, automatic text summarization area has experienced an increasing number of researches and products, which led to a countless number of summarization methods [20].

Some methods are based on statistics and empirical data and, for this reason, are said to be superficial methods. Other methods make use of linguistic knowledge of varied complexity, from syntax and semantics to discourse. These are usually called deep methods [7]. Research in the area has shown that deep methods are expensive because they need sophisticated knowledge resources and text interpretation techniques; on the other hand, it is a consensus that they may perform better than superficial methods, as it is demonstrated in [4], for instance.

Among the most interesting and investigated deep summarization methods, there are those based on Rhetorical Structure Theory (RST) [8]. RST is probably the most used discourse theory in Computational Linguistics and have influenced other works in all language processing fields, as machine translation [13], anaphora resolution [3][19], essay scoring [1], etc.

According to RST, a coherent text may be structured as a discourse tree, whose intermediate nodes are discourse relations and leaves are propositional units expressed by segments (usually clauses) in the text. As basic idea, summarization takes advantage of the fact that text segments in the tree are classified according to their importance.

Supporting automatic summarization and other works that use RST representation of texts, some RST parsers have arisen lately. Such parsers are able to automatically build good RST trees for texts, bridging the gap that existed until then, so that the text interpretation is not a big problem for this summarization research line. The most known RST parser for English is described in [12]. Portuguese [16] and Japanese [21] languages also have similar parsers.

Many RST summarization methods exist. To our knowledge, no comparative evaluation exists for all of them, so that it is hard to say which one is better or which one to choose for using. In this paper, we carry out a comprehensive comparative evaluation among the main methods using a well-established automatic evaluation measure in the area, namely, ROUGE [5]. We compare the RST methods to a simple but efficient extractive summarizer [17] in order to verify the actual benefits one has by using deep methods. We also propose new RST-based and hybrid methods based on the previous ones.

Most of our results are new in the area. They reveal very interesting conclusions. The simplest RST-based method is among the best ones, although all of them present comparable results. We show that all RST-based methods overcome the extractive summarizer and that hybrid methods produce worse summaries. Finally, we verify that a strong assumption in summarization and RST research area (more specifically, Mann and Thompson assumption [9], which will be introduced latter) is not helpful in the way previously imagined.

In the next section, we briefly introduce RST. The main RST summarization methods in the area, the extractive summarizer we included in our evaluation, and the methods we propose are reviewed in Section 3. In Section 4, we describe our evaluation methodology and report the obtained results. Finally, some final remarks are made in Section 5.

2 Rhetorical Structure Theory

According to RST, all propositional units in a text must be connected by rhetorical/discourse relations in some way for the text to be coherent. The connection of all the text propositional units produces its rhetorical/discourse structure. Rhetorical structures are usually represented by trees (not necessarily binary), with each relation connecting subtrees, which can be single propositional units (leaves in the tree) or other trees.

As an example of a rhetorical analysis of a text, consider Text 1 in Figure 1 (with segments that express basic propositional units numbered) and its rhetorical structure in Figure 2.

> [1] Although he is allergic to it, [2] he tried it. [3] Now, he has a headache and [4] his body is red.

Figure 1. Text 1

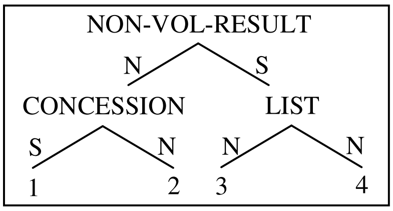

Figure 2. Text 1 rhetorical structure

The symbols N and S indicate the nucleus and satellite of each rhetorical relation: in RST, the nucleus indicates the most important information in the relation, while the satellite provides complementary information to the nucleus. In this structure, propositions 1 and 2 are in a CONCESSION relation, i.e., the fact of being allergic to something should avoid someone of trying it; propositions 3 and 4 RESULT (not volitionally) from propositions 1 and 2; propositions 3 and 4 present a LIST of allergy symptoms. In some cases, relations are multinuclear (e.g., LIST relation), that is, they have no satellites and the connected propositions have the same importance; otherwise, relations are mononuclear, with one nucleus and one satellite (e.g., CONCESSION and NON-VOL-RESULT relations). RST originally defines around 25 relations.

All RST summarization methods proposed in literature take advantage of the fact that the satellites in a rhetorical structure are secondary information. Besides this similarity, each method uses different criteria for selecting which satellites to eliminate or, viewed under other perspective, which segments to keep in the summary.

3 Summarization Methods

In what follows, we introduce the extractive summarizer we tested and the main RST summarization methods in the area, which are the ones we compare.

3.1 Extractive Summarizer

GistSumm (GIST SUMMarizer) [17] is an extractive summarizer, i.e., it produces a summary by juxtaposing frozen segments from the source text.

GistSumm comprises three main processes: text segmentation, sentence ranking, and extract production. Sentence ranking is based on the keywords method [6]: it scores each sentence of the source text by summing up the frequency of its words in the text. Optionally, GistSumm may normalize the score of a sentence by its size (in number of words). This normalized method is called average-keywords method. The highest scored sentence (by any of the two previous methods) is elected the gist sentence, i.e., the sentence that best expresses the text main idea. The extract production process focuses on selecting other sentences from the source text to include in the extract, based on: (a) gist correlation and (b) relevance to the overall content of the source text. Criterion (a) is fulfilled by simply verifying co-occurring words in the candidate sentences and the gist sentence, trying to ensure lexical cohesion. Criterion (b) is fulfilled by sentences whose score is above a threshold, computed as the average of all the sentences scores, trying to guarantee that only relevant sentences are chosen. All the selected sentences are then juxtaposed to compose the final extract.

According to its authors, GistSumm has already undergone several evaluations, the main one being DUC'2003 (Document Understanding Conference, the main summarization conference in the area, that recently changed its name to Text Analysis Conference). It showed to be very good in determining the gist sentence in news texts: in a range from 0 (the summary is useless) to 4 (the summary can substitute the source text) in a DUC evaluation performed by humans, GistSumm achieved an average result of 3.12.

We selected GistSumm for our evaluation for its good results and because it is freely available for use.

3.2 RST Methods

Based on the nuclearity of text segments in RST trees, many summarization methods were proposed. All of them produce partial orderings of segments, selecting for the summary the best ranked segments. The selection is restricted by the compression rate (i.e., the size of the summary in relation to the size of the source text – in number of words). By partial order we mean a ranking in which some items (segments, in this case) may keep the same position (because they have the same score).

Ono et al. [15] propose one of the first and simplest summarization methods. In the proposed procedure, the root of the RST tree has associated as score the number of levels that the tree has. Then, beginning in the root, the tree is traversed in depth-first mode, carrying through each level the score from the level above. Each time a satellite arc is found, the score in the following level is decreased by one. The partial ordering of the segments is given by ordering them according to their final score. As an illustration, Figure 3 shows the score of each node in the tree from Figure 2. The score is shown in bold in the right of the nodes. The partial ordering of segments for this structure would be 2 > {1, 3, 4}, where the > signal indicates left-priority for composing the corresponding summary. One of the characteristics of this method can be evidenced in the example: many segments have the same score.

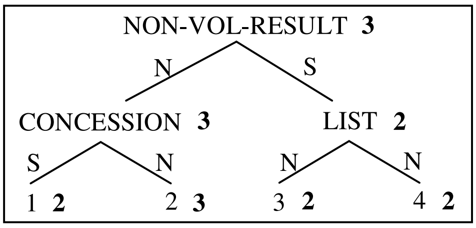

Figure 3. Example of Ono et al. method

The authors evaluated their method by verifying in the automatic summaries the number of preserved key sentences from the source text (indicated by humans) and whether the summaries included or not the most important sentence from the text. The evaluation used 72 news and technical texts in Japanese language. In the best case, the authors report that the summaries preserved 51% of the key sentences and included the most important sentence in 74% of the cases.

O'Donnell [14] adds to the scoring method the importance of the relations. The author assumes that each relation also has an associated score that indicates how important the segments/subtrees it relates are in the text. The method starts by associating the score 1 to the root of the tree and, then, traversing the tree in depth-first mode. Each time a satellite arc is found, the next node will have the corresponding score multiplied by the importance factor of the relation above it. O'Donnell empirically attributes importance factors to each relation (always between 0 and 1). Figure 4 shows the score of each node in the tree from Figure 2. For the example, we randomly assume that the relation NON-VOL-RESULT has the importance factor 0.8, CONCESSION has a factor of 0.6, and LIST a factor of 0.4. The partial ordering of the segments is 2 > {3, 4} > 1.

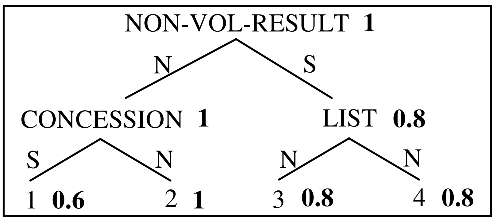

Figure 4. Example of O'Donnell method

O'Donnell does not evaluate its method. He only says that the results on a small-scale experiment showed that reasonable-quality texts can be produced.

Marcu [10][12] proposes the use of promotion sets to determine the most important segments in the tree. The promotion set for each node in a tree is built in a bottom-up way: each internal node in the tree includes in its promotion set the union of the promotion sets of its nuclear children. The promotion set of a leaf is composed of itself. For scoring each segment, the method attributes to the root of the tree a score correspondent to the number of levels in the tree and, then, traverses the tree toward the segment under evaluation: each time the segment is not in the promotion set of a node during the traversing, it has the score decreased by one. Figure 5 illustrates this process for segment 3. The promotion sets are shown under the nodes in the tree.

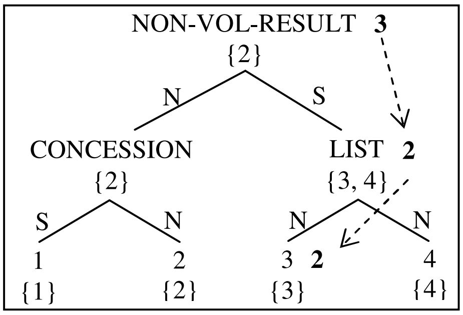

Figure 5. Example of Marcu method

In his experiments, Marcu verified that his method can account for determining the most important units in a text with near 70% recall and precision. He used only 5 texts from Scientific American magazine with the important units annotated by humans.

Marcu [11] also proposes an improvement for his own method. He considers that it is also important to take into account how long time each segment belonged to a promotion set. So, for the scoring obtained in the previous method for each segment, it is added the number of levels in the tree in which the segment was part of a promotion set. For segment 3 in Figure 5, for example, its final score (2) would be incremented by 2, since the segment appears in promotion sets of two levels. We will refer to this method as Improved Marcu method. Marcu arrives to this method by successively tuning and comparing Ono et al. method and his own previous method.

Based on the previous methods, we propose a new one that combines some features they present. The idea is to verify whether a richer method can produce better summaries. We take into consideration the nuclearity, the importance of relations, and the promotion sets concepts. Initially, we attribute to the root of the tree a score of two times the number of levels the tree has. For each segment, we traverse the tree toward the segment and do the following for each visited node: if the segment does not belong to the node promotion set, its score is decremented by the complement of the importance factor of the relation above the actual node (i.e., 1-relation importance factor), and, if a satellite arc is traversed, the score is also decremented by 1. Starting the root node with two times the number of levels in the tree guarantees that no negative score will be generated for any node. Figure 6 shows this scoring procedure for segment 3.

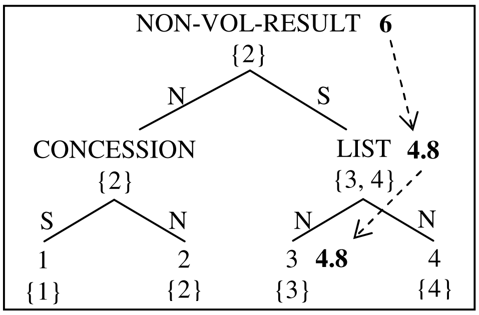

Figure 6. Example of our combined method

We also tested the strong assumption that Mann e Thompson [9] do. They say that a satellite can only be removed if it is not in the locus of effect of the corresponding rhetorical relation. The locus of effect of each relation specifies the segments (nuclei and/or satellites) that suffer the most the effects of the relation.

We incorporated such assumption in all previous methods: the original methods are normally applied; then, in the partial ordering produced, the segments that are in the locus of effect of relations in the tree have priority over the others, being selected first to compose the summary. If more segments are needed, the ones still not included in the summary are selected from the original ordering. Suppose, for instance, that we have the partial ordering 2 > {3, 4} > 1 given by O'Donnell method and that segments 3 and 4 are in the locus of effect of their relations. Segments 3 and 4 have preference for the selection; if more segments are needed, then segments 2 and 1 (in this order) are the remaining candidates.

Besides the 5 original methods and the 5 variations with Mann and Thompson assumption we have, we still tried to incorporate some hybridism to all of these methods. As we know GistSumm is good in selecting the gist sentences of texts, we combined this property to the partial ordering obtained with each RST method. Each segment in RST trees has its final score (given by some RST method) multiplied by the distance of this segment to the gist sentence. The distance is computed by the traditional cosine measure [18]: the closer the segment is to the gist sentence, the higher the cosine measure is. This way, the segments that share more content with the gist sentence are better scored in the end. We used gist sentences computed by keywords and average-keywords methods.

Considering all methods and variations, we end up with 32 methods to evaluate[1]. We describe such evaluation in what follows.

[1] All methods were re-implemented for this work. See [22] for details on our implementation of O'Donnell method.

4 Evaluation

For evaluating the summarization methods, we used the RST Discourse Treebank [2], a corpus composed of 385 Penn Treebank news texts annotated according to RST. This corpus is considered a reference corpus in the area for being annotated by more than one human annotator in order to have annotation agreement. In particular, we used a subset of 30 texts for which human summaries are available, which we use as reference summaries to which the automatic summaries are compared.

As evaluation metric, we applied ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [5], an informativity measure widely used in summarization area. It measures how much of the information in human summaries are reproduced in the corresponding automatic ones. Although it is a very simple measure (which basically computes the number of common n-grams in the human and automatic summaries), the authors show that it highly correlates with human judgment in summary ranking tasks. In fact, the metric is so well accepted that it become one of the main metrics in the Text Analysis Conferences.

ROUGE computes recall, precision and f-measure values, all between 0 (the worst) and 1 (the best), considering n-grams from 1 to 4. Recall indicates the percentage of the content in the human summary that is reproduced in the automatic summary. Precision indicates the percentage of the content in the automatic summary that is relevant. F-measure combines both measures, being a unique indication of the system performance. About the n-gram comparison ROUGE performs, the authors show that unigram comparison is enough to rank summaries very well. In this paper, we show the results obtained for unigram comparison only. The other n-gram comparisons showed similar results.

Tables 1, 2, and 3 show the recall, precision, and f-measure average results obtained for all methods, respectively. The first column in the tables lists the evaluated methods. The methods followed by the "MT" letters are the ones that incorporate Mann and Thompson assumption. The three remaining columns are the results for the methods without the use of the gist sentence, the results for the methods with gist sentence computed by keywords method (referred as kw), and the results with gist sentence computed by average-keywords method (referred as avg-kw), respectively. The compression rate used was the same used in the summaries in RST Discourse Treebank: about 75% (i.e., the summaries keep 25% of the original text size).

Table 1. Recall results

Methods	**---**	**kw**	**avg-kw**
GistSumm	---	0,3853	0,3863
Ono et al.	0,4166	0,3862	0,3908
Ono et al.-MT	0,4199	0,3888	0,3964
O'Donnell	0,4154	0,3856	0,3862
O'Donnell-MT	0,4182	0,3882	0,3910
Marcu	0,4122	0,3834	0,3979
Marcu-MT	0,4082	0,3928	0,4014
Improved Marcu	0,4173	0,3830	0,4012
Improved Marcu-MT	0,4106	0,3892	0,4005
Our method	0,4219	0,3844	0,3988
Our method-MT	0,4193	0,3917	0,4037

Table 2. Precision results

Methods	**---**	**kw**	**avg-kw**
GistSumm	---	0,4251	0,4154
Ono et al.	0,4621	0,4228	0,4245
Ono et al.-MT	0,4642	0,4299	0,4357
O'Donnell	0,4626	0,4254	0,4185
O'Donnell-MT	0,4618	0,4291	0,4318
Marcu	0,4538	0,4236	0,4247
Marcu-MT	0,4472	0,4309	0,4287
Improved Marcu	0,4485	0,4205	0,4304
Improved Marcu-MT	0,4435	0,4294	0,4340
Our method	0,4622	0,4235	0,4312
Our method-MT	0,4579	0,4317	0,4330

Table 3. F-measure results

Methods	**---**	**kw**	**avg-kw**
GistSumm	---	0,3839	0,3788
Ono et al.	0,4163	0,3826	0,3865
Ono et al.-MT	0,4189	0,3872	0,3951
O'Donnell	0,4154	0,3841	0,3804
O'Donnell-MT	0,4171	0,3865	0,3900
Marcu	0,4111	0,3821	0,3891
Marcu-MT	0,4055	0,3899	0,3923
Improved Marcu	0,4095	0,3804	0,3940
Improved Marcu-MT	0,4046	0,3877	0,3955
Our method	0,4188	0,3827	0,3920
Our method-MT	0,4152	0,3895	0,3963

In general, one can see that:

- Mann and Thompson assumption does not significantly improve the performance of RST methods (in fact, for some methods, the performance is slightly worse);
- RST methods have a worse performance when combined with the gist sentences;
- RST methods outperform GistSumm;

- Considering only f-measure values, Ono et al. with Mann and Thompson assumption and the method that we propose are better than the other RST methods, although the difference is small;
- In general, one may say that all RST methods have comparable performance.

Ono et al. method is surprisingly one of the best ones in despite of its simplicity. Following it, we can find the method we propose, which combines features from different methods. Marcu methods are the worst ones. It is also curious that hybridism did not produce better results. We believe that different hybridism configurations could achieve it.

5 Final Remarks

To our knowledge, the comprehensive comparative evaluation between RST methods we present in this paper, as well as most of its results, are new in the area.

A point that remains open in this research is the effect of the different RST methods on other summary characteristics besides informativity, like coherence and cohesion. This is a theme for future work.

6 Acknowledgments

The authors are grateful to FAPESP, CAPES, and CNPq.

7 References

[1] Burstein, J.; Marcu, D.; Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, pp. 32-39.

[2] Carlson, L.; Marcu, D.; Okurowski, M.E. (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.), *Current Directions in Discourse and Dialogue*, pp. 85-112. Kluwer Academic Publishers.

[3] Cristea, D.; Ide, N.; Romary, L. (1998). Veins Theory. An Approach to Global Cohesion and Coherence. In the *Proceedings of Coling/ACL*.

[4] Leite, D.S.; Rino, L.H.M.; Pardo, T.A.S.; Nunes, M.G.V. (2007). Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In the *Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pp.17-24. 26 April, Rochester, NY, USA.

[5] Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In the *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada.

[6] Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.

[7] Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co. Amsterdam.

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

[9] Mann, W.C. and Thompson, S.A. (1992). *Discourse Description: Diverse linguistic analyses of a fund-raising text*. Pragmatics & Beyond, New Series. John Benjamins.

[10] Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD Thesis. University of Toronto.

[11] Marcu, D. (1998). To build text summaries of high quality, nuclearity is not sufficient. In the *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*. Stanford, CA.

[12] Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

[13] Marcu, D.; Carlson, L.; Watanabe, M. (2000). The automatic translation of discourse structures. In the *1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, Vol. 1, pp. 9-17. Seattle, Washington.

[14] O'Donnell, M. (1997). Variable-Length On-Line Document Generation. In the *Proceedings of the 6th European Workshop on Natural Language Generation*. Duisburg, Germany.

[15] Ono, K.; Sumita, K.; Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In the *Proceedings of the International Conference on Computational Linguistics* (Coling-94).

[16] Pardo, T.A.S. and Nunes, M.G.V. (2006). Review and Evaluation of DiZer - An Automatic Discourse Analyzer for Brazilian Portuguese. In the *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese* (Lecture Notes in Computer Science 3960), pp. 180-189. Rio de Janeiro-RJ, Brazil. May 13-17.

[17] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal.

[18] Salton, G. (1989). *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.

[19] Schauer, H. (2000). Referential Structure and Coherence Structure. In the *Proceedings of TALN*. Lausanne, Switzerland.

[20] Spärck Jones, K. (2007). *Automatic summarising: a review and discussion of the state of the art*. Technical Report UCAM-CL-TR-679, University of Cambridge.

[21] Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; Amano, S. (1992). A discourse structure analyzer for Japonese text. In the *Proceedings of the International Conference on Fifth Generation Computer Systems*, V. 2, pp. 1133-1140. Tokyo, Japan.

[22] Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2007). *Estudo e Avaliação de Métodos de Sumarização Automática de Textos Baseados na RST*. ICMC-USP Technical Report. São Carlos-SP, August, 28p.