International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)
ISSN: 2150-7988 Vol.1 (2009), pp.188-196
http://www.mirlabs.org/ijcisim

A comprehensive summary informativeness evaluation for RST-based summarization methods

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
Av. Trabalhador São-carlense, 400 - Centro
Caixa Postal: 668 - CEP: 13560-970 - São Carlos - SP

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. Additionally, we test machine learning techniques trained on RST features. We also compare our results to superficial summarizers, which belong to a paradigm with severe limitations, and to hybrid methods, combining RST and superficial methods. Our results show that all RST methods have similar overall performance and that they outperform the superficial methods. In terms of precision, the method we propose is the best one, while it competes with other ones for coverage. Machine learning techniques achieved high accuracy in the classification of text segments worth of being in the summary, but were not able to produce more informative summaries than the regular RST methods.

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 (for an overview of the area, see, e.g., [1]).

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 [2]. 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 by [3], for instance.

Among the most interesting and investigated deep summarization methods, there are those based on Rhetorical Structure Theory (RST) [4]. RST is probably the most used discourse theory in Computational Linguistics and have influenced works in all language processing fields, as machine translation (e.g., [5]), anaphora resolution ([6]; [7]), essay scoring [8], 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.

Some RST parsers have arisen lately for supporting automatic summarization and other works that use RST representation of texts. 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 by [9]. Portuguese and Japanese languages also have similar parsers ([10] and [11], respectively).

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 in terms of summary informativeness. For this purpose, we use a well-established automatic evaluation measure in the area, namely, ROUGE [12]. We compare the RST methods to simple but efficient superficial summarizers 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, and test some machine learning techniques trained on RST features.

Most of our results are new in the area. They reveal very interesting conclusions. Our results show that all RST methods have similar overall performance, even the

simplest ones. In terms of precision, the method we proposed is the best one, while it competes with other methods for coverage. Mann and Thompson summarization assumption [13] (which will be introduced latter) does not produce better results. Machine learning techniques did not improve the results apparently, even though high accuracy was achieved in classification of text segments worth of being in the summary. On the other side, it was demonstrated that superficial methods are worse than the RST ones, which encourages following deep summarization approaches.

In the next section, we briefly introduce RST. The RST summarization methods and the superficial summarizers used in our evaluation are presented in Section 3. In Section 4, we describe our evaluation methodology and report the obtained results. After this, in Section 5, we discuss our experiments applying machine learning techniques for the summarization task. Finally, some final remarks are made in Section 6.

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

Figure 1. Text 1

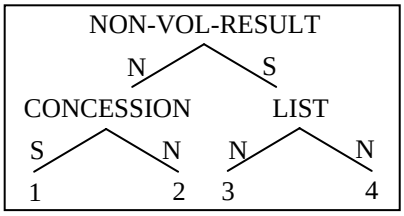

Figure 2. Text 1 rhetorical structure

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 about 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 another perspective, which segments to keep in the summary.

3. Summarization methods

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

3.1. Superficial methods

GistSumm (GIST SUMMarizer) [14] 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 [15]: 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 also tested the traditional relational method [16], which builds a graph from the text, where each node represents a sentence and arcs have weights computed as the number of common words among the nodes they connect (stopwords are removed). Each node/sentence is scored as the sum of the arcs weights that it is connected to. The highest scored sentences are selected to compose the summary.

The final superficial method we tested is the localization method [17], which selects for the summary the first sentences of a text. This method is particularly good for news texts, since the most important sentences appear in the beginning of such texts.

All the above methods, as well as the ones we will describe in the next section, are constrained by the compression rate used for building the summary. The compression rate specifies the size of the summary in relation to the size of the source text (in number of words).

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. By partial order we mean a rank in which some items (segments, in this case) may keep the same position (because they have the same score).

Ono [18] 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 at 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.

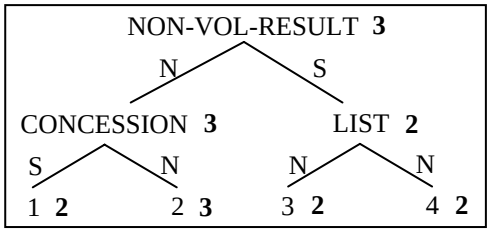

Figure 3. Example of Ono et al. method

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 [19] 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.

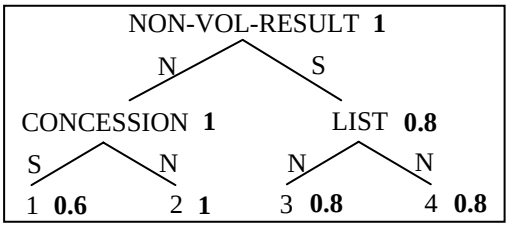

Figure 4. Example of O'Donnell method

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

Marcu ([20], [5]) 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: the promotion set of a leaf is composed of itself; each internal node in the tree includes in its promotion set the union of the promotion sets of its nuclear children. 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.

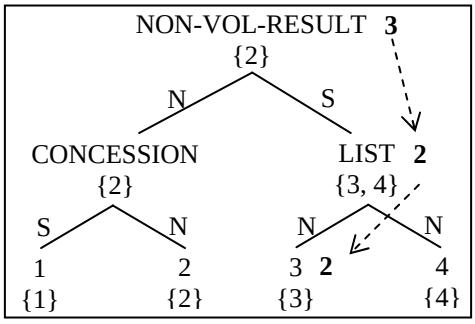

Figure 5. Example of Marcu method

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 [21] also proposes an improvement for his own method (based on a personal communication with Eduard Hovy, as he reports). He considers that it is also important to take into account how many times 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.

We also isolated the improvement proposed by Hovy as a single method (hereafter we refer to it as Hovy method). A variation of it (that we call Improved Hovy method) weights the score of a segment by the level of the promotion set in that the segment appears for the first time.

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.

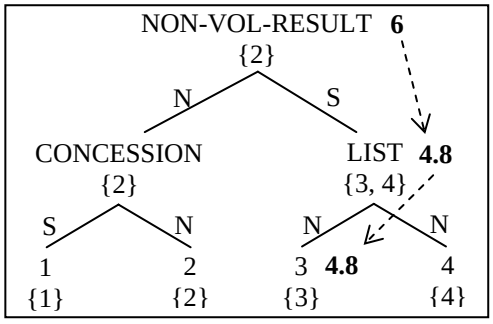

Figure 6. Example of our combined method

We also tested the strong assumption that Mann & Thompson [13] 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.

3.3. Hybrid methods

Besides the 7 original methods and the 7 variations with Mann and Thompson assumption, we still tried to incorporate some hybridism to all of these methods. The hybridism is simply achieved by summing the segment scores given by the superficial methods (more specifically, GistSumm with keywords and average-keywords methods, and the relational method) to the segment scores produced by the RST methods. Such new scores produce new partial orderings for summarization. We also allowed the combination of any of the above methods with the localization method: segment scores are multiplied by weights that decrease as the segments appear lower in the corresponding text (so that segments that appear first in

the text have some advantage over the segments below them). This method shows to be generic enough to be combined with any summarization strategy.

Considering all methods and variations, we end up with 128 methods to evaluate. We describe such evaluation in what follows. All methods were re-implemented for this work. The importance factors of the RST relations that we adopted in O'Donnell method are:

- relations of the types contrast, cause-effect, and sequence, which are considered the most important ones: importance factor of 0.8;
- relations with important complementary information (e.g., comparison, enablement, evaluation): 0.6;
- relations of text structuring (e.g., comment-topic and textual-organization) and that introduce additional information (e.g., summary and interpretation), which are considered less important: 0.4;
- relations of irrelevant complementary information (e.g., background, circumstance, and elaboration): 0.2.

[22] presents the details of such relations scores.

4. Evaluation

For evaluating the summarization methods, we used the RST Discourse Treebank [23], 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) [12], an informativeness 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 and 2 show the f-measure average results obtained for all methods, without and with the combination with the localization method, 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 Improved Marcu method and Improved Hovy method are labeled as IMarcu and IHovy for short. The first two lines represent the original superficial methods; the last two lines represent the method we proposed here. The four remaining columns are the results for the pure RST methods (without modifications), the results for the methods combined with the gist sentence computed by keywords method (referred as kw), the results combined with the gist sentence computed by average-keywords method (referred as avg-kw), and the results combined with the relational method (referred as relat), 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).

RST Methods	none	kw	avg-kw	relat
None	0.37443	0.36770	0.39006	0.39072
None-MT	0.39148	0.37749	0.39425	0.38764
Ono et al.	0.41794	0.37704	0.40222	0.38866
Ono et al.-MT	0.41997	0.37956	0.40387	0.38848
ODonnell	0.41226	0.37206	0.39330	0.39041
ODonnell-MT	0.41404	0.37990	0.39903	0.38846
Marcu	0.40703	0.37440	0.40112	0.38825
Marcu-MT	0.40510	0.38228	0.39441	0.38783
IMarcu	0.41215	0.37568	0.39831	0.38795
IMarcu-MT	0.40719	0.38152	0.39248	0.38957
Hovy	0.41377	0.37271	0.39776	0.38819
Hovy-MT	0.40705	0.38381	0.39509	0.38940
IHovy	0.41251	0.37412	0.39749	0.38889
IHovy-MT	0.40659	0.38158	0.39893	0.38897
Proposed	0.41616	0.37590	0.40318	0.38931
Proposed-MT	0.41365	0.37908	0.39967	0.38824

Table 1. F-Measure without localization method

RST Methods	none	kw	avg-kw	relat
None	0.37443	0.36770	0.39006	0.39072
None-MT	0.37443	0.37285	0.39484	0.36454
Ono et al.	0.40960	0.38451	0.41761	0.40929
Ono et al.-MT	0.41219	0.40016	0.42187	0.40983
ODonnell	0.41585	0.38781	0.41477	0.40430
ODonnell-MT	0.41689	0.39150	0.40827	0.39773
Marcu	0.41351	0.40266	0.42346	0.41309
Marcu-MT	0.41371	0.40069	0.41734	0.41027
IMarcu	0.41756	0.40365	0.42007	0.41766
IMarcu-MT	0.41689	0.40548	0.41067	0.41059

Hovy	0.36334	0.40490	0.41198	0.39102
Hovy-MT	0.38214	0.41107	0.40806	0.39632
IHovy	0.41892	0.39332	0.40921	0.39951
IHovy-MT	0.41008	0.39652	0.40114	0.40486
Proposed	0.42339	0.39759	0.42107	0.41095
Proposed-MT	0.42051	0.40614	0.41964	0.41486

Table 2. F-Measure with localization method

In general, one can see that:

- Mann and Thompson assumption does not significantly improve the general performance of RST methods (in fact, for some methods, the performance is slightly worse);
- the localization method makes the results slightly better;
- excepting the localization method, most of the hybrid methods have worse performance;
- RST methods outperform the superficial methods.

In general, one may say that all RST methods have comparable performance. Looking to the details, the method we proposed combined with the localization method and Marcu method combined with localization and average-keywords methods were the best ones. Without considering the localization method, Ono et al. method is surprisingly one of the best ones in despite of its simplicity. It is also curious that hybridism did not produce better results. We believe that different hybridism configurations could achieve it.

Now we focus on the precision and recall figures individually. Tables 3 and 4 show the average precision results. They must be read in the same way of Tables 1 and 2 before.

RST Methods	none	kw	avg-kw	relat
None	0.41745	0.39001	0.42153	0.41353
None-MT	0.43014	0.40211	0.42731	0.41277
Ono et al.	0.45637	0.39845	0.43414	0.41038
Ono et al.-MT	0.45522	0.40360	0.43569	0.41373
ODonnell	0.45273	0.39373	0.42454	0.41353
ODonnell-MT	0.44998	0.40382	0.43270	0.41367
Marcu	0.44014	0.39440	0.42749	0.40987
Marcu-MT	0.43692	0.40566	0.42611	0.41334
IMarcu	0.44197	0.39636	0.42348	0.40990
IMarcu-MT	0.43869	0.40598	0.42132	0.41359
Hovy	0.44626	0.39400	0.42762	0.41009
Hovy-MT	0.43834	0.41017	0.42549	0.41537
IHovy	0.44372	0.39413	0.42810	0.41064
IHovy-MT	0.43714	0.40639	0.43053	0.41489
Proposed	0.45170	0.39672	0.43082	0.41190
Proposed-MT	0.44773	0.40251	0.43105	0.41384

Table 3. Precision without localization method

RST Methods	none	kw	avg-kw	relat
None	0.41745	0.40857	0.43539	0.41353
None-MT	0.43014	0.41905	0.43540	0.41945
Ono et al.	0.45575	0.41561	0.45382	0.44188
Ono et al.-MT	0.45057	0.43096	0.45537	0.44014
ODonnell	0.45885	0.42011	0.44982	0.43532
ODonnell-MT	0.45238	0.42297	0.44357	0.42632
Marcu	0.45008	0.42961	0.45522	0.44409
Marcu-MT	0.44697	0.42624	0.44946	0.43703
IMarcu	0.44197	0.43128	0.45364	0.44645
IMarcu-MT	0.44837	0.43273	0.44340	0.43667
Hovy	0.40339	0.43898	0.45153	0.42944
Hovy-MT	0.42013	0.44139	0.44185	0.42871
IHovy	0.45127	0.42003	0.44307	0.42647
IHovy-MT	0.43973	0.42387	0.43350	0.43188
Proposed	0.46239	0.42569	0.45846	0.44343
Proposed-MT	0.45490	0.43392	0.45313	0.44547

Table 4. Precision with localization method

As it is possible to see, the localization method increased precision in general. On the other side, the hybridism did not improve the results. It is also interesting to notice that RST methods combined with the localization method and with Mann and Thompson assumption got better precision. The most precise method was the combination of the localization method with the method we proposed in this paper. Ono et al. method also presented very good precision, following our method.

Finally, Tables 5 and 6 show the coverage figures. Again, the tables must be read in the same way of the tables before.

RST Methods	none	kw	avg-kw	relat
None	0.37694	0.38881	0.40168	0.41245
None-MT	0.39582	0.39664	0.40588	0.40979
Ono et al.	0.42555	0.39756	0.41391	0.41400
Ono et al.-MT	0.42896	0.39833	0.41525	0.41050
ODonnell	0.41883	0.39301	0.40451	0.41198
ODonnell-MT	0.42085	0.39882	0.40856	0.41054
Marcu	0.41955	0.39583	0.41744	0.41364
Marcu-MT	0.41630	0.40134	0.40662	0.40974
IMarcu	0.42348	0.39760	0.41606	0.41356
IMarcu-MT	0.41639	0.39876	0.40642	0.41307
Hovy	0.41959	0.39225	0.41080	0.41291

Hovy-MT	0.41261	0.40018	0.40920	0.40959
IHovy	0.42123	0.39547	0.40845	0.41267
IHovy-MT	0.41477	0.39976	0.41124	0.40947
Proposed	0.42953	0.39708	0.41822	0.41366
Proposed-MT	0.42624	0.39796	0.41177	0.41005

Table 5. Recall without localization method

RST Methods	none	kw	avg-kw	relat
None	0.37694	0.38285	0.40094	0.36590
None-MT	0.39582	0.39752	0.40786	0.39239
Ono et al.	0.41024	0.40248	0.42924	0.42735
Ono et al.-MT	0.41764	0.41620	0.43567	0.42815
ODonnell	0.42583	0.40195	0.42692	0.42449
ODonnell-MT	0.42872	0.40692	0.41836	0.41721
Marcu	0.42270	0.42270	0.43879	0.42833
Marcu-MT	0.42435	0.42252	0.43140	0.42826
IMarcu	0.42442	0.42256	0.43015	0.43601
IMarcu-MT	0.42717	0.42575	0.42023	0.43005
Hovy	0.36721	0.41624	0.41630	0.39682
Hovy-MT	0.38708	0.42669	0.41781	0.40723
IHovy	0.42789	0.41572	0.41925	0.41788
IHovy-MT	0.41997	0.41556	0.41131	0.42392
Proposed	0.43259	0.41574	0.43091	0.42859
Proposed-MT	0.43120	0.42449	0.43056	0.43132

Table 6. Recall with localization method

The superficial methods present the worst results. For the majority of the cases, Mann and Thompson assumption caused the coverage to be slightly better. The localization method also increased the coverage. Surprisingly, the hybridism increased the coverage in relation to the RST methods with the localization method. The highest coverage was obtained by Marcu method with average-keywords and the localization methods, followed closely by the method we proposed combined with localization method.

Overall, the conclusions are:

- If a highly precise method is needed for some application (e.g., the question answering related tasks), the method we proposed combined with localization method must be chosen, since it is significantly better than the other ones;
- If a method with high coverage is needed for some application (e.g., biographical summarization), several options are available, namely, marcu method combined with localization and average-keywords method, O'Donnell method combined with localization method, or, still, the method we proposed combined with localization method.
- If general performance is important, the method we proposed combined with localization, O'Donnel method combined with localization, and Marcu method combined with localization and average-keywords methods are good candidates.

If simplicity is necessary (to time or resource limitations), Ono et al. method is a good one.

5. Experiments with Machine Learning

One last experiment we tried was to combine some RST features in machine learning techniques. For each segment in the RST tree, we extracted the following features:

- the segment level in the tree;
- the percentage of nuclei found in the path from the root to the segment;
- the percentage of satellites found in the path from the root to the segment;
- a number obtained by Huffman method of traversing the tree (assuming 1 to nucleus arcs and 0 to satellite arcs);
- a number obtained by the complement of the above method of traversing the tree;
- a number obtained by Huffman method of traversing the tree (assuming 1 to nucleus arcs and 0 to satellite arcs) assigning same size numbers to every segment (including right zeros in the shorter numbers);
- a number obtained by the complement of the above method of traversing the tree;

Besides these features, for each segment, we also used as features the scores produced by the pure summarization methods we tested in the previous experiments. Therefore, we end up with 19 features (i.e., the 7 features above plus 8 features for the RST methods plus 3 features for superficial methods plus 1 feature for the localization method).

The Huffman coding was used as a way of representing the path of nuclei and satellites to the segments.

The class used was "yes" for segments that should appear in the summary and "no" for the remaining segments. Such classes were obtained from the summaries that accompany the corpus that we used. If the segment was in the corresponding summary, then the class was "yes"; otherwise, the class was "no".

Our dataset had 4,054 learning instances. We replicated the "yes" learning instances in order to have a balanced dataset. In general, this has to be done in summarization experiments, since the "yes" class has considerable fewer examples than the "no" class, since we are dealing with short summaries.

We set 2/3 of the instances to train the machine learning techniques and 1/3 for testing the models built. We used decision trees (J48) and Naïve-Bayes methods in the Weka data mining environment [24].

In terms of general accuracy, the best result we obtained was for J48: 73%. For Naïve-Bayes, we got 71%

accuracy. The majority error is 50%, since our dataset is balanced. We also generated the summaries (for the test data) using the learned models and evaluated them using ROUGE (following exactly the same methodology from the previous section). The results are shown in Table 3. Naïve-Bayes produced better results, but still comparable to the results from the previous section for the pure RST methods in terms of summary informativeness. However, it is important to notice that direct comparisons are not fair, since only a subset of the corpus was used in this machine learning test.

Methods	F-measure
J48	0.31485
Naïve-bayes	0.41994

Table 3. F-Measure for the machine learning experiment

It is interesting to notice that summary informativeness did not increase with machine learning techniques, even though these techniques achieved good accuracy in segments classification.

6. 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.

Our results show that all RST methods have similar general performance, even the simplest ones. In terms of precision, the method we proposed was the best one. For coverage, our method competes with others. In general, Mann and Thompson summarization assumption does not produce better results as well. Machine learning techniques did not improve the results in terms of summary informativeness, but achieved high accuracy in classification of text segments worth of being in the summary. It was also demonstrated that superficial methods are worse than the RST ones, except the localization method, which generally makes the results better. However, for other text genres than the news one, the localization method may not help.

A point that remains open in this research is the effect of the different RST methods on other summary characteristics besides informativeness, like coherence and cohesion. Some works try to accomplish this by using the Veins Theory [6] over the RST trees (see, e.g., [25]). This is a theme for future investigation.

7. Acknowledgments

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

8. References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] Cristea, D.; Ide, N.; Romary, L. (1998). Veins Theory: A Model of Global Discourse Cohesion and Coherence. In the *Proceedings of Coling-ACL*, pp. 281-285.

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

[8] 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.

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

[10] Pardo, T.A.S. and Nunes, M.G.V. (2008). On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, Vol. 15, N. 2, pp. 43-64.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] Skorochodko, E.F. (1971). Adaptive Method of Automatic Abstracting and Indexing. *Information Processing*, Vol. 2, pp. 1179-1182.

[17] Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, Vol. 2, pp. 354-365.

[18] 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).

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

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

[21] 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.

[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.

[23] 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.

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

[25] Carbonel, T.I.; Seno, E.R.M.; Pardo, T.A.S.; Coelho, J.C.; Collovini, S.; Rino, L.H.M.; Vieira, R. (2006). A Two-Step Summarizer of Brazilian Portuguese Texts. In the *Proceedings of the 4th Workshop on Information and Human Language Technology* – TIL. Ribeirão Preto-SP, Brazil. October 27-28.