Co-referential chaining for coherent summaries through rhetorical and linguistic modeling

Eloize Rossi Marques Seno, Lucia Helena Machado Rino
Núcleo Interinstitucional de Lingüística Computacional – NILC/USFCAR
Rodovia Washington Luiz, km 235 – 13565-905 – São Carlos – SP – Brazil
{eloize@mail.fpte.br, lucia@dc.ufscar.br}

Key words: Automatic Summarization, Coherence, Co-referential chaining

Abstract

Co-referential chaining poses a difficult problem for Automatic Summarization (AS) when a text unit that embeds an anaphoric reference is chosen to compose a summary and its antecedent is not. Coherence, in this case, is often severely damaged and so may the degree of informativity be. This article presents an AS proposal that deals with rhetorical and linguistic knowledge to overcome that. A heuristics-based system is presented that addresses referentiality and rhetorical structuring to identify when an antecedent text unit must be included in a summary, to avoid coherence problems. The model has been assessed so far on both informativity and coherence for news articles written in Brazilian Portuguese. Only coherence is addressed in this paper.

1. Introduction

Significant knowledge-rich approaches to AS have been proposed that deal with the rhetorical structuring of source texts in order to produce coherent summaries. In general, they suggest that discourse segments potentially superfluous for exclusion may be recognized through their rhetorical relations (e.g., Sparck-Jones 1993; O'Donnell 1997), or that a relevance classification of elementary discourse units (hereafter, EDUs), or single propositions, may be drawn from the text rhetorical organization (Marcu 1997; 1999; 2000). In most cases, the Rhetorical Structure Theory, or RST (Mann & Thompson 1987), is used to model both, rhetorical/discourse organization and summarization.

In spite of focusing on discourse representation, none of the above referred work embeds a fine-grained semantic analysis aiming at disentangling simple content units and, thus, yielding a proper, propositional representation. Instead, propositions are directly delimited at the surface of the text and expressed as leaves of an RST tree. This results in rhetorically organized frozen text units. Whilst this approach makes knowledge-based AS less complicated, it prevents usual anaphora resolution, since anaphoric expressions are, in general, embedded in an EDU. As a result, non-sequiturs are very likely to occur when an EDU is chosen to compose a summary and its antecedent is not.

Still considering EDUs holding for text units in RST trees, the RHeSumaRST (Heuristic Rules for Summarizing RST trees) summarization system can use directly Marcu's algorithm of salience determination. Its singularity is that it also adds to salience determination a way of overcoming the non-sequitur problem using the Veins Theory, or VT (Cristea et al. 1998). This theory allows recognizing the "veins" of a discourse, by delimiting the domain of referential accessibility of its EDUs. In signaling the scope of the discourse in which anaphora antecedents may occur in RST structures, it may be used to determine co-referent discourse segments of a vein that may lead to a coherent discourse. So, VT combined to RST may help content selection and structuring on a better basis than RST, alone or combined with Marcu's salience model. This is exactly the methodology proposed in RHeSumaRST.

The system also provides the means to support discourse organization of multi-sentential summaries by focusing on preserving a satisfactory level of informativity, when compared to the source text.

RHeSumaRST does not address classical knowledge-rich techniques, such as abstracting or generalizing RST trees, to produce their corresponding summary RST trees. However, it brings forth a cooperative model that guides the

pruning of RST trees by means of heuristics. These combine decisions on rhetorical and co-referential chaining by dealing with rhetorical and linguistic knowledge to exclude superfluous information. Differently from excluding them at random, they are conditioned to verifying if EDUs that are candidate to exclusion do not damage the domain of referential accessibility of EDUs already chosen to compose a summary structure. So, whilst RST provides the means to tackle informativeness and does not necessarily guarantee coherence, VT does the opposite way, to keep the summary coherent.

In what follows firstly main features of RST and VT theories are outlined (Section 2), then the RHeSumaRST architecture is presented (Section 3). The rationale behind RHeSumaRST heuristics is presented in Section 4. RHeSumaRST assessment on coherence is, thus, presented in Section 5. Final remarks are presented in Section 6.

2. RST and Veins Theories

RST is an already well-known theory, widely explored in knowledge-rich NLP systems by Marcu and others (e.g., Ono et al. 1994; O'Donnell 1997). Nuclearity is its main features for AS. Once EDUs are delimited, they may be inter-related through mono- or multi-nuclear RST relations, yielding compound RST subtrees. These, in turn, may also be related to other RST subtrees. In the end, if the text under analysis is coherent, its corresponding RST tree is supposed to convey no dangling RST subtrees. Thus, in pruning an RST tree, the resulting summary RST tree must also obey the same principle: just fully interconnected summary RST trees may be produced.

Nuclearity, in RST, addresses relevance or, according to Marcu, the salience of the EDUs: RST nuclei (Ns) are more relevant than their satellites (Ss). So, mononuclear RST relations must be focused upon, in order to summarize a tree. Equally relevant EDUs are inter-related only by multinuclear RST relations. In this case, if one of the EDUs is chosen to compose the summary RST tree, all of them must also be chosen.

By focusing on nuclearity, RST-based approaches to AS aim at guaranteeing coherence and selecting the most relevant information to compose a summary. However, the only means to map relevance is through the position of Ns and Ss in the tree. Considering a highly condensing strategy, all the Ss could be considered superfluous and, thus, pruned from the source RST tree. This has already been proven unfruitful, due to incoherent summaries. Marcu tries to improve on that by considering, in his salience model, a thread of EDUs groups that are classified on their relevance. In doing so, coherence is improved, but inter-related EDUs do not necessarily mirror semantically and linguistically dependent content units, such as the co-referential chains under focus in RHeSumaRST.

The lack of a model that processes EDUs in a more fine-grained level is the reason for that: if an ideal discourse analyzer were used, co-referential chains of EDUs would all lead to the same, unique concept. In turn, there would be no non-resolved anaphors. Since the EDUs considered here are actually text units and no approach exists to properly deal with the posed problem, RHeSumaRST tries to improve on former RST-based methods to tackle coherence loss introduced by co-referential breaks. These may happen, for example, when an anaphoric EDU is an N and its antecedent is a S: if the drastic approach were taken, S would be excluded and a dangling anaphor would hold in the final structure.

The Veins Theory goal is to prevent the above to happen: it also addresses only nuclearity to build the "veins" of a discourse. A vein of an EDU delimits, thus, its domain of referential accessibility in an RST tree. Since the vein is defined as the set of discourse units that embeds the antecedent of an anaphor related to that EDU, VT is semantically and linguistically motivated, although it depicts the veins only on RST basis. This view gives rise to RHeSumaRST main premise: addressing co-referential chaining by verifying if a complete co-referential chain is embedded in a unique vein helps preventing summaries to be incoherent due to dangling anaphors.

Both theories that bring about the RHeSumaRST heuristics benefit from the previous mentioned work: Cristea et al.'s algorithm to compute the veins of the RST tree is first applied, yielding an RST annotated tree in which heads[1] and veins are marked. Then, Marcu's model is used to classify EDUs of the RST tree, as described in the next section.

[1] A head of an RST node N is the set of its most salient EDUs in the discourse segment which embeds N; its vein is drawn based on the head.

3. RHeSumaRST architecture

Figure 1 presents RHeSumaRST pipelined architecture: first, an input RST tree is annotated with its veins, by applying Cristea et al.'s algorithm (1998). Then, Marcu's model is used to classify EDUs of the RST tree and, finally, pruning takes place on the annotated RST tree through the application of candidate heuristics[2].

Following the same strategy as that adopted by Marcu (1997), input source RST trees are built using the RST Annotation Tool[3]. This provides only a graphic interface to support rhetorical annotation. Pruning is entirely based upon the application of the heuristics. These are defined on RST and VT bases as a result of corpus analyses of real texts in NL.

Focusing solely on the phenomenon of co-referentiality, i.e., on the occurrence of both anaphoric and its antecedent terms in a text, RHeSumaRST works in the following way: once a discourse segment that embeds an anaphor EDU is chosen to compose a summary, it looks after its antecedent EDU, in order to also include it and, thus, prevent a dangling anaphor to occur. Certainly, the problem does not exist for direct anaphors[4]. So, only anaphoric constructions that do not pose repetitions are considered.

More specifically, RHeSumaRST has been modeled analyzing only definite anaphors (Vieira et al. 2002), i.e., those signaled by nouns phrases. In Brazilian Portuguese, they are generally introduced by a definite article (e.g., 'o menino', or *the boy*) and, like in many other Romance languages, they are very often used as a stylistic resource to improve writing quality. Definite anaphors have been focused upon in RHeSumaRST because they occur very often in natural language texts and their potential to introduce coherence problems due to co-reference breaks in the intended summaries is high.

Pruning heuristics will be detailed in the following section.

[2]Both modules have been implemented in collaboration with Leandro M. Hanada.

[3]www.isi.edu/~marcu/discourse/AnnotationSoftware.html (march/2005).

[4]Those whose anaphoric expressions are the same as the ones in their antecedents.

4. Heuristics based on co-references for AS of RST trees

Pruning heuristics in RHeSumaRST address coherence and informativity by focusing on constraints that prevent coherence breaks introduced by particular choices of EDUs. Since RHeSumaRST does not resolve anaphors, the heuristics are driven towards including a complete vein, once its anaphor component is chosen.

Defining the set of heuristics has been corpus-driven: the corpus was composed of 30 newspaper articles from the TeMário corpus (Pardo & Rino 2003)[5] and its analysis aimed at (a) identifying those RST satellites that were indeed superfluous; (b) verifying the contexts of co-referentiation that could introduce coherence problems. The texts were pre-processed in three distinct phases, as follows: firstly, their RST trees were built with the RST Annotation Tool. Secondly, the veins of the resulting RST trees were automatically obtained. Finally, the occurring co-referential chains were annotated with the MMAX tool (Müller & Strube 2001) that provides only a graphic interface to support co-referential text annotation. So, the expertise of the user is still required.

To identify superfluous RST satellites (goal (a) above), each RST tree was compared with the corresponding manual summary[6]: we verified if each EDU in an RST tree had corresponding information units to those in the manual summary. The underlying hypothesis here was that, by defining heuristics based on information common to the manual summaries, the heuristics would be able to recognize content judged relevant in the source text under summarization. The comparison aimed, thus, at guaranteeing minimum informativity in the automatic summaries. This methodology implies that heuristics be based on those RST relations that signal more significantly the content of interest, for any source text (this paper does not discuss genre dependence).

Our analysis showed that most mononuclear RST relations (c.a. 97%) had their satellites

[5] Free download in http://www.linguateca.pt

[6]TeMário texts already come along with their manual summaries, built by a professional writer. So, we consider that they are the ideal summaries (Mani 2001).

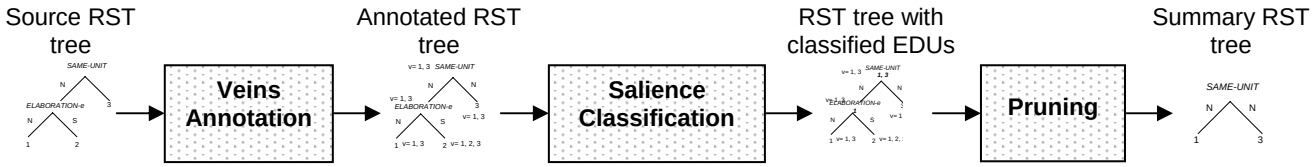

Figure 1: RHeSumaRST architecture

included in the manual summaries in 50% or less of the cases. Many of them had no satellite preserved at all, such as the CIRCUMSTANCE relation. Only EXPLANATION had more than 50% of its satellites present in the summaries (57%). However, this RST relation is meaningless in the corpus (only 0.5% occurrences). These results may indicate that satellites of RST trees are indeed non-relevant for AS and, thus, should be directly excluded, in pruning mononuclear RST relations. Multinuclear ones also appear in the corpus. However, they were not our focus, because if we decide to include in a summary RST tree one of the EDUs of those relations, all of them should be included. So, there are no pruning heuristics for them.

Concerning goal (b), i.e., verifying the contexts of co-referentiation that could introduce coherence problems, the corpus analysis helped identifying the domain of referential accessibility of definite anaphors occurring in the source text. The intention here was to verify its structural correspondence with its RST tree and derive proper heuristics to guarantee that a summary would not convey dangling anaphors. Then, we looked for its anaphoric and antecedent terms in its RST tree, to see if they were present in the same vein. The hypothesis here was that, if a complete chain were embedded in the same vein, heuristics should be based on the preservation of the full vein to guarantee the minimum of coherence of the summaries, concerning co-referential chaining.

The results showed that, for 80% of the co-referential chains in the corpus, both anaphor and antecedent occurred in the same vein. For the corresponding RST relations, heuristics were thus defined that were limited to excluding only those satellites that were not in the domain of referential accessibility of the EDUs already chosen to compose a summary. As a result of the corpus analysis, 30 pruning heuristics were defined, which compose the main module of the RHeSumaRST system, as described in (Seno & Rino 2005a). A heuristic involving the CIRCUMSTANCE RST relation is described below, together with an example extracted from the TeMário corpus (illustrated co-referential chains in bold).

> H1: Delete y from circumstance(x,y) if y $\notin$ vein V, for an V of any EDU z already included in the summary RST tree.
>
> Example-text[7]: [1] **The industry Produtos Pirata Indústria e Comércio Ltda.**, from Contagem [2] (metropolitan region of Belo Horizonte), [3] will register this year an increase in productivity in its commercial and industrial areas of 11% and 17%, respectively. [4] The gains are due by the board of **the industry** to the new philosophy that has being adopted in **the industry** since October last year, [5] when **the Pirata** was introduced in the Sebrae Program of Total Quality.

Assuming that a partial summary RST tree is under construction and that it already embeds EDUs 1, 3, and 4, H1 applies to the exclusion of EDU 5 from the original RST tree (Figure 2)[8], because the veins (v) of the nuclei 1, 3 and 4 do not include it. Summary RST tree (Figure 3) results from that. Satellite 2 could be excluded as well, under a heuristic that analogously concluded for its omission with no damage of the veins of the nuclei 1 and 3.

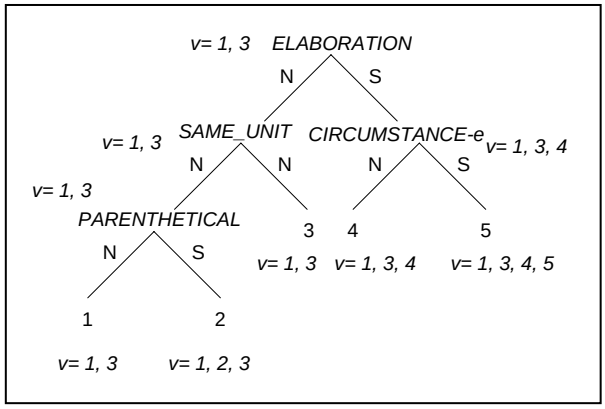

Figure 2: RST tree of example-text

[7] Translated into English for readability.

[8] The terminal nodes represent the EDUs and inner nodes the rhetorical relations.

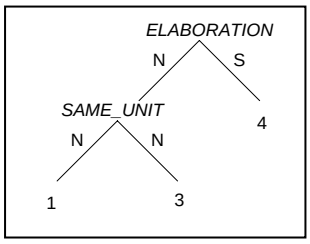

Figure 3: summary RST tree

5. Assessing RHeSumaRST

A first assessment of the system was carried out on informativity and coherence (Seno & Rino 2005b) on a test corpus of 10 texts from TeMário. For informativity assessment (using ROUGE), when compared to two other systems, namely, the Salience (Marcu 1997) and the Topline models, RHeSumaRST performed similarly to the others. The Topline model is a baseline which prunes every satellite of an RST tree leaving only its nuclei. It has been named so because pruning all the satellites and leaving all the nuclei of a source RST tree (thus, only central information, according to Mann & Thompson (1987)) is very likely to provide a highly informative summary.

With respect to coherence, which is the only focus of this article, RHeSumaRST performed better than the Topline and Salience models. However, the percentage of coherence loss due to co-referential chaining breaks was very low for all of them (5%, 8%, and 15 %, respectively). For this reason, another coherence assessment has been devised on a different test corpus.

The new corpus[9] amounts to 20 newspaper articles written in Brazilian Portuguese. Two RST specialists annotated all of them rhetorically, also using the RST Annotation Tool. In order to avoid annotation disagreements, rules of RST tagging (Carlson & Marcu 2001) were applied. Besides the RST annotation, the test corpus was also annotated with the co-referential chains.

To compute co-referential chaining breaks, all the summaries were generated and manually compared with their corresponding source texts. A 70% compression rate was used to prune source RST trees. The manual comparison was intended to verify if coherence problems found in the summaries were indeed due to co-referential chaining breaks. To confirm that, once identified a potential dangling anaphor, the corresponding source text was used to retrieve its context. If this did not introduce a complete co-referential chaining, the anaphor found in the summary was considered new and no coherence break was computed.

Differently from the first assessment, RHeSumaRST was compared only with the Topline Model, because this rated closer to it. The Salience Model was not considered because it would imply a much more struggling process, for the thorough manual comparison of each summary with its corresponding source needed. Table 1 shows the rates of co-referential chaining breaks (CRC breaks) in news summaries. CRCs stands for co-referential chains.

Table 1: CRC breaks in news summaries

System	# CRCs	# CRC breaks	% CRC breaks
RheSumaRST	45	2	4
Topline	45	8	18

Although the current test corpus is still small for a robust evaluation, compared to the first assessment RHeSumaRST performed better than Topline. Whilst there was a differing rate of c.a. 63% of co-referential chaining breaks between RHeSumaRST and Topline in the former experiment, with Rhetalho the differing rate fell to c.a. 22% breaks. In other words, RHeSumaRST rated still better than Topline, in comparison with the first experiment. This, indeed, shows a considerable improvement in its performance.

6. Final Remarks

The results reported so far bring about RHeSumaRST potentiality for summarizing news texts. However, more investigation is needed for scalability. Its improving on performance in the assessment reported in this paper also may be due to the way the test corpus was annotated: the Rhetalho corpus has been certified by two RST experts, under a clear formalization of the discourse analysis procedure.

Carrying on a deeper critique of the heuristics involves considering the specificities of both foundation models themselves: some RST relations are more likely to appear in texts of certain genres than others. It is still unclear if

[9] Corpus Rhetalho, available in: www.nilc.icmc.usp.br/~thiago/rhetalho.html (jun/2005).

genre-dependence influenced the results shown above.

Although RHeSumaRST demands discourse analyzes of the texts to be summarized, it adds to knowledge-rich approaches the advantages of both, the RST nuclearity and the Veins Theory. Clearly, for certain genres whose texts do not embed a significant amount of co-referential chains, the proposed model would be too sophisticated. However, in assuring that a vein will be completely reproduced in the summary RST tree, RHeSumaRST seems promising in improving the coherence of the generated summaries.

It is also noticeable that, although the system has been modeled focusing solely on definite anaphors, its proposed methodology is applicable to any linguistic expression of co-referential chains. This is due to its lacking of a mechanism to identify any of those reference phenomena.

RHeSumaRST still demands a knowledge-rich input, which has been so far handbuilt. This will be resolved in the near future by plugging to it DiZer, a discourse analyzer of texts written in Brazilian Portuguese (Pardo et al., 2004). In this way, a two-module automatic summarizer will be available and, thus, source texts will be given for producing their summary RST trees.

References

(Carlson & Marcu 2001) Carlson, L. and Marcu, D. *Discourse Tagging* Reference *Manual*. Technical Report ISI-TR-545, University of Southern Califórnia, 2001.

(Cristea et al. 1998) Cristea, D.; Ide, N.; Romary, L. *Veins Theory: A Model of Global Discourse Cohesion and Coherence*. In the Proc. of the COLING/ACL' 1998, pp.281-285. Montreal, Canadá, 1998.

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

(Marcu 1997) Marcu, D. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD Thesis, Department of Computer Science, University of Toronto, 1997.

(Marcu 1999) Marcu, D. Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization*, pp. 123-136, The MIT Press, 1999.

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

(Müller & Strube 2001) Müller C. and Strube, M. MMAX: A tool for the annotation of multi-modal corpora. *In the Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark, pp. 90-95, 2001.

(O'Donnell et al. 1998) O'Donnell, M. Variable-Length On-Line Document Generation. *In the Proc. of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duiburg, Germany, 1997.

(Ono et al. 1994) Ono, K.; Sumita, K.; Miike, S. Abstract Generation Based on Rhetorical Structure Extraction. *In the Proceedings of the International Conference on Computational Linguistic – Coling-94*, pp 344-348, Japan, 1994.

(Pardo & Rino 2003) Pardo, T.A.S. and Rino, L.H.M. *TeMário: A Corpus for Automatic Text Summarization*. Technical Report: NILC-TR-03-09, ICMC/USP, São Carlos, Brazil, 2003. (in Portuguese)

(Pardo et al. 2004) Pardo, T.A.S; Nunes, M.G.V.; Rino, L.H.M. *DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese*. XVII Brazilian Symposium on Artificial Intelligence - SBIA'04, São Luís – Maranhão, 2004.

(Seno & Rino 2005a) Seno, E.R.M.; Rino, L.H.M. *Summarization Heuristics for RST Structures*. Technical Report: NILC-TR-05-04, ICMC/USP, São Carlos, Brazil, 2005. (in Portuguese)

(Seno & Rino 2005b) Seno, E.R.M.; Rino, L.H.M. Summarizing RST trees focusing on referential chains: A case study. In *III Workshop em Tecnologia da Informação e da Linguagem Humana – TIL*, São Leopoldo, Brazil, 2005.

(Spark Jones 1993) Sparck Jones, K. *Discourse Modelling for Automatic Summarising*. Technical Report No. 290. University of Cambridge, February, 1993.

(Vieira et al. 2002) Vieira, R.; Salmon-Alt, S.; Schang, E. Multilingual Corpora Annotation for Processing Definite Descriptions. *In the Proc. of the Portugal for Natural Language Processing – PorTAL*, Faro, Portugal, 2002.