Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

Joint semantic discourse models for automatic multi-document summarization

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

Núcleo Interinstitucional de Linguística Computacional (NILC)

[1]Departamento de Ciência da Computação, Universidade Federal de Lavras (UFLA)
Caixa Postal: 3037 – CEP: 37200-000 – Lavras/MG

[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Caixa Postal: 668 – CEP: 13566-970 – São Carlos/SP

paulastm@gmail.com; taspardo@icmc.usp.br

Abstract. *Automatic multi-document summarization aims at selecting the essential content of related documents and presenting it in a summary. In this paper, we propose some methods for automatic summarization based on Rhetorical Structure Theory and Cross-document Structure Theory. They are chosen in order to properly address the relevance of information, multi-document phenomena and subtopical distribution in the source texts. The results show that using semantic discourse knowledge in strategies for content selection produces summaries that are more informative.*

Resumo. *Sumarização automática multidocumento visa à seleção das informações mais importantes de um conjunto de documentos para produzir um sumário. Neste artigo, propõem-se métodos para sumarização automática baseando-se em conhecimento semântico-discursivo das teorias Rhetorical Structure Theory e Cross-document Structure Theory. Tais teorias foram escolhidas para tratar adequadamente a relevância das informações, os fenômenos multidocumento e a distribuição de subtópicos dos documentos. Os resultados mostram que o uso de conhecimento semântico-discursivo para selecionar conteúdo produz sumários mais informativos.*

1. Introduction

Automatic Multi-Document Summarization (MDS) aims at selecting the relevant information from multiple documents on the same topic to produce a summary (Mani, 2001). It has seen increasing attention because it can be useful in a variety of areas, mainly due to help coping with information overload.

Two main approaches are generally considered in MDS. The *superficial approach* uses statistical or some limited linguistic information to build a summary, usually has low cost and is more robust (Haghighi and Vanderwende, 2009; Ribaldo, 2013; Castro Jorge, 2015). The *deep approach* uses linguistically motivated assumptions and demands high-cost resources, but it produces summaries of higher quality in terms of information, coherence and cohesion (Marcu, 1997; Afantenos et al., 2007; Uzêda et al., 2010; Castro Jorge and Pardo, 2010). However, studies based on superficial or deep knowledge do not deal jointly with relevance of different sentences in a source text, multi-document phenomena and subtopics.

In a source text, some sentences are more important than others because of their position in the text or in a rhetorical structure, thus, they cannot be treated uniformly (Wan, 2008). In the case of news texts, it is known that the first or leading paragraph usually expresses the main fact reported in the news. Therefore, selecting sentences from the beginning of the text could be a good summary (Saggion and Poibeau, 2013). More sophisticated techniques use analysis of the discourse structure of texts for determining the most important sentences (Marcu, 1997; O'Donnell, 1997; Uzêda et al., 2010).

In order to deal with multi-document phenomena such as redundant, contradictory and complementary information, that occur in a collection of texts, approaches that achieve good results use multi-document semantic discourse models (Radev, 2000; Zhang et al., 2002; Castro Jorge and Pardo, 2010; Kumar et al., 2014). However, those works are not concerned about the relevance of sentences in each text together with multi-document phenomena as a human does when writing a summary.

Another feature is that each text of a collection develops the main topic, exposing different subtopics as well. A topic is a particular subject that we write about or discuss, and subtopics are represented in pieces of text that cover different aspects of the main topic (Hearst, 1997; Salton et al., 1997; Hennig, 2009). For example, a set of news texts related to an earthquake typically contains information about the magnitude of the earthquake, its location, casualties and rescue efforts (Bollegala et al., 2010). There are some proposals that combine the subtopical structure and multi-document relationship (Salton et al., 1997; Wan, 2008; Harabagiu and Lacatusu, 2010) to find important information, but without treating the salience of a sentence in its text.

We may say that current strategies for MDS have separately used each of the three criteria of relevance of information, multi-document phenomena and subtopical distribution, resulting in summaries that are not representative of the subtopics and less informative than they could be. However, human summarization behaviour looks at (i) the subtopics and rhetorical structure of texts to select content (Jaidka et al., 2010) and considers that (ii) the redundant information (that is repeated across texts) tends to be important (Mani, 2001). Therefore, we need effective summarization methods to analyze the information from different texts and produce informative summaries.

As an example, Figure 1 shows an automatic multi-document summary produced from two texts organized in four subtopics related to the health of Maradona, the famous Argentine soccer player: the history of Maradona's disease, current state of health, messages of support and Maradona's relapse. The summary has repeated content (highlighted in bold) and sentences are only from two subtopics: *current state of health* (S1 and S3) and *Maradona's relapse* (S2). The summary would be better if the three criteria for summary production had been used.

In this paper, we propose to model the process of MDS using semantic discourse theories, in order to properly address the three cited criteria. To do that, we choose the theories RST (Rhetorical Structure Theory) (Mann e Thompson, 1987) and CST (Cross-document Structure Theory) (Radev, 2000) due to their importance for automatic summarization described in many works (O'Donnell, 1997; Marcu, 1997; Zhang et al., 2002; Castro Jorge and Pardo, 2010; Castro Jorge, 2015). The RST model details major aspects of the organization of a text and indicates relevant discourse units. The CST

model, in turn, describes semantically related textual units from topically related texts. We present some methods for content selection, aiming at producing more informative and representative summaries from the source texts. For this purpose, we use a multi-document corpus manually annotated with RST and CST. The methods produce satisfactory results, improve the state of the art and indicate that the use of semantic discourse knowledge positively affects the production of informative extracts. To the best of our knowledge, this is the first time RST and CST are combined in methods for MDS. Both theories' relations are domain-independent.

[S1] "Maradona had a relapse in acute hepatitis. Now he is stable**.** Despite he had got better on Sunday, he should continue hospitalized", said Cahe to the news La Nación.
[S2] Hospitalized in Buenos Aires, **he had a relapse** and felt pain again **due to acute hepatitis**, according to his personal doctor, Alfredo Cahe.
[S3] Cahe said that Maradona had not started to drink alcoholic beverages again, and that the causes of the relapse are being investigated.

Figure 1: Example of multi-document summary (Castro Jorge and Pardo, 2010)

The remainder of this paper is organized as follows: Section 2 gives a brief background about the semantic discourse models RST and CST; Section 3 presents some related work; Section 4 shows the developed methods for MDS; the corpus is described in Section 5; Section 6 presents some results; Section 7 presents some final remarks.

2. Discourse knowledge

RST (Mann and Thompson, 1987) is a descriptive theory of major aspects of the organization of a text. It represents relations among propositions in a text and discriminates nuclear (i.e., important propositions) and satellite (i.e., additional information). Each sentence may be formed by one or more propositions. Relations composed of one nucleus and one satellite are named mononuclear relations. On the other hand, in multinuclear relations, two or more units participate and are equally important. The relationships are traditionally structured in a tree-like form (where larger units – composed of more than one proposition – are also related in the higher levels of the tree). RST is probably the most used discourse model in computational linguistics and has influenced works in all language processing fields. Particularly for automatic summarization, it takes advantage of the fact that text segments are classified according to their importance: nuclei are more informative than satellites.

Inspired by RST and other researches, CST appears as a theory for relating text passages from different texts on the same topic (Radev, 2000). It is composed by a set of relations that detect similarities and differences among related texts. Differently from RST, CST was devised mainly for dealing with multi-document organization. The relations are commonly identified between pairs of sentences, coming from different sources, which are related by a lexical similarity significantly higher than random. The result of annotating a group of texts is a graph, which is probably disconnected, since not all segments present relations with other segments. CST was applied in MDS studies for English (Zhang et al., 2002; Kumar et al., 2014) and Portuguese texts (Castro Jorge and Pardo, 2010). These researchers take advantage of the fact that CST relationships indicate relevant information between sources and facilitate the processing of multi-document phenomena.

3. Related work

There are several works based on semantic discourse knowledge for MDS. Zhang et al. (2003) replace low-salience sentences with sentences that maximize the total number of CST relations in the summary. Afantenos et al. (2007) propose a summarization method based on pre-defined templates and ontologies. Kumar et al. (2014) take into account the generic components of a news story within a specific domain, such as *who*, *what* and *when*, to provide contextual information coverage and use CST to identify the most important sentences. Castro Jorge (2015) incorporates features given by RST to generative modelling approaches.

For news texts in Brazilian Portuguese, the state of the art consists in two different summarization approaches of Castro Jorge and Pardo (2010) and Ribaldo (2013). Based on deep knowledge, Castro Jorge and Pardo developed the *CSTSumm* system that employs CST relations to produce preference-based summaries. Sentences are ranked according to the number of CST relationship they hold. Ribaldo, in turn, took advantage of superficial knowledge and developed a multi-document system, called *RSumm*, which segments texts into subtopics using TextTiling (an adapted version for Portuguese, described in Cardoso et al., 2013) and group the subtopics using measures of similarity. After clustering, a relationship map is created and the relevant content is selected by the segmented bushy path (Salton et al., 1997). In the segmented bushy path, at least one sentence of each subtopic is selected to compose the summary.

As we can see, those works do not combine semantic discourse knowledge such as RST and CST for content selection. In this study, we argue that the semantic discourse knowledge improves the process of MDS.

4. The CSTNews corpus

Our main resource is the CSTNews[1] corpus (Cardoso et al., 2011), composed of 50 clusters of news articles written in Brazilian Portuguese, collected from several sections of mainstream news agencies: Politics, Sports, World, Daily News, Money, and Science. The corpus contains 140 texts altogether, amounting to 2,088 sentences and 47,240 words. On average, the corpus conveys in each cluster 2.8 texts, 41.76 sentences and 944.8 words. Besides the original texts, each cluster conveys single-document manual summaries and multi-document manual and automatic summaries.

The size of each summary corresponds to 30% of the size of the biggest text in the cluster (considering that the size is given in terms of the number of words). All the texts in the corpus were manually annotated with RST and CST structures in a systematic way, with satisfactory annotation agreement values.

5. Methods for MDS

In this section, we describe how RST, CST and subtopics may be used together in some strategies for content selection. This investigation was organized in three groups: (1) methods based solely on RST, (2) methods that combine RST and CST, and (3)

[1] http://www.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html

methods that combine RST, CST and subtopics. It is considered that the texts are segmented and clustered in subtopics, and annotated with CST and RST.

The **first group** is based on the literature for single document summarization using RST, specifically on Marcu's work (1997), which associates a score for each node in the RST tree depending on its nuclearity and the depth of the tree where it occurs. The salient units associated with the leaves are the leaves themselves. The salient units (promotion set) of each internal node is the union of the promotion sets of its nuclear children. Textual units that are in the promotion sets of the top nodes of a discourse tree are more important than units that are salient in the nodes found at the bottom. For scoring each segment, the method attributes to the root of the tree a score corresponding to the number of levels in the tree and, then, traverses the tree towards 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. Following the same idea, we proposed a strategy (which we refer to as *RST-1*) to compute a score for each sentence as the sum of its nodes' scores (propositions), given by Marcu's method (1997). It does this for all texts of a collection and, then, a multi-document rank of sentences is organized. From the rank, the next step is to select only nuclear units of the best sentences.

As an example, consider that there are 3 sentences in part A of Figure 2: sentence 1 is formed by proposition 1; sentence 2, by 2; sentence 3, by 3 to 5. The symbols N and S indicate the nucleus and satellite of each rhetorical relation. Applying *RST-1* method, the score (in bold) of sentences 1 and 2 is 4, and for sentence 3 is 6. Whereas sentence 3 has the higher score, its nuclei are selected to compose a summary. Since RST relations do not indicate if there is redundancy between nodes, we control it using cosine measure (Salton, 1989).

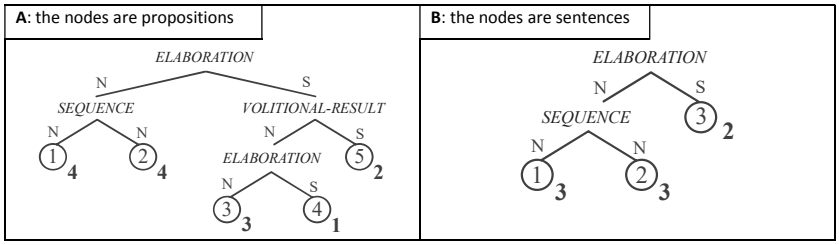

Figure 2: Example of a discourse tree using RST

Because all these scores depend on the length of the document (Louis et al., 2010) and on the number of propositions in a sentence, a rank based on the sum of propositions' scores may insert discrepancies in the method and does not mirror the relevance of sentences in a multi-document scenario. More than this, as we work on news texts, it is expected that first sentences are more relevant, differently from Figure 2 (part A), where the last sentence was more important than the former. As a solution, we proposed to compute the score for sentences, not for propositions, and to normalize each score by the height of the tree, resulting in a number ranged from 0 to 1. In Figure 2 (part B), each node represents a sentence; the bold numbers are sentences' scores before normalization. From this new sentence rank, we create two possibilities of content

selection: only nuclear units (propositions) of sentences (we refer to as *RST-2*) or full sentences (*RST-3*).

The **second group** of strategies combines **R**ST and **C**ST. We assume that the relevance of a sentence is influenced by its salience given by RST and its correlation with multi-document phenomena, indicated by CST model. We know that the more repeated and elaborated sentences between sources are, more relevant they are, and likely contain more CST relations (Zhang et al., 2002; Castro Jorge and Pardo, 2010; Kumar et al., 2014). If we find the relevant sentences in a set of related documents, we may use RST to eliminate their satellites and make room for more information. In this and the following groups of methods, redundancy is controlled by means of CST relationships. For example, if there is an EQUIVALENCE relation between two sentences, only one must be selected to the summary.

Based on that, we propose two strategy variations. In the first one (we refer to as *RC-1*), the rank of sentences is organized according to the number of CST relationships one sentence has. The more relevant a sentence is, the higher in the rank it is. The best sentence is selected and, if it has satellites, they are eliminated. This method is a variation of CSTSumm (Castro Jorge and Pardo, 2010). We tested two more variations for RC-1, which were not described in this work because they did not produce satisfactory results (for more details, see Cardoso, 2014).

The second strategy (we refer to as *RC-4*) is a combination of the number of CST relationships and RST-3 strategy (where the RST score of a sentence is normalized by its tree's height), constituting a score that represents the salience of the sentence and its relevance for a collection. In other words, RST and CST scores are added to form the final score of a sentence. In contrast to RC-1, RC-4 selects full sentences.

To illustrate RC-1 and RC-4 methods, consider Figure 3, where there are two discourse trees representing two texts (D1 and D2); D1 is upside down for better visualization; each node is a sentence with its RST score normalized in bold; dashed lines between texts are CST relationship.

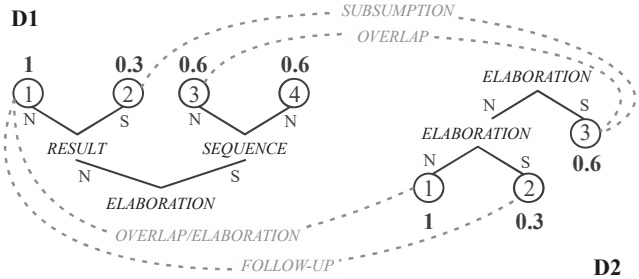

Figure 3: Example of RST and CST relationships for two texts

By applying RC-4, the rank according to the number of CST relationships is D1_1 > {D2_1, D2_3} > {D1_2, D1_3, D2_2} > D1_4. Using RC-4 strategy, the rank is organized as follows: D1_1 > D2_1 > D2_3 > D1_3 > {D1_2, D2_2} > D1_4.

The **third group,** composed of four strategies, combines RST, CST and subtopics, and is based on lessons learned from the previous methods. Texts are

segmented in subtopics (by a method described in Cardoso et al., 2013) and similar subtopics are clustered (by a method described in Ribaldo et al., 2013). We assume that a subtopic discussed in several documents is more significant than one that was discussed in only one (Ercan and Cicekli, 2008), thus, sentences of repeated subtopics are relevant. With that in mind, to benefit those subtopics during content selection, their sentences receive an extra score. One strategy of this group, called *RCT-1*, considers that the score of a sentence by RCT-1 method is the sum of its RST score by Marcu's algorithm (1997), applied to sentences, with its number of CST relationships and the relevance of subtopic to which it belongs. From the rank of sentences, content is selected without satellite propositions. Using the same rank, we propose a variation called *RCT-2,* which selects full sentences. Two other variations are the *RCT-3* and the *RCT-4* methods. For these strategies, the total score for each sentence is similar to the first two, with the difference that the RST score is normalized by the size (height) of its discourse tree. RCT-1 and RCT-3 only select nuclear propositions of the best sentences, while RCT-2 and RCT-4 pick out full sentences.

6. Results and discussion

This section presents comparisons of the results over the reference corpus using ROUGE (Lin, 2004), a standard evaluation metric used in text summarization, which produces scores that often correlate quite well with human judgments for ranking systems. This metric computes n-gram overlapping between a human reference and an automatic summary. The methods are compared to CSTSumm (Castro Jorge and Pardo, 2010) and RSumm systems (Ribaldo, 2013), that have used the same corpus as here.

In Table 1, it is observed that, in the **RST group** (lines 9-11), RST-3 method, that selects full sentences, has the best ROUGE evaluation. Since RST-1 and RST-2 select only nuclei, they produce summaries with many problems related to linguistic quality; sometimes it is impossible to get the gist.

Table 1: ROUGE evaluation

		ROUGE-1		
	Method	**Recall**	**Precision**	**F-measure**
1	**RC-4**	**0.4374**	**0.4511**	**0.4419**
2	RC-1	0.4270	0.4557	0.4391
3	**RCT-4**	**0.4279**	**0.4454**	**0.4346**
4	RCT-3	0.4151	0.4446	0.4274
5	RCT-2	0.4199	0.4399	0.4269
6	RSumm	0.3517	0.5472	0.4190
7	RCT-1	0.3987	0.4313	0.4128
8	CSTSumm	0.3557	0.4472	0.3864
9	RST-3	0.3874	0.3728	0.3781
10	RST-2	0.3579	0.3809	0.3671
11	RST-1	0.3198	0.3238	0.3206

In the **RC group**, RC-4 is slightly better in F-measure compared to RC-1. It reinforces that selecting full sentences produces more informative summaries. RC-4 was also considered better than all other methods; it indicates that considering the relevance of sentences between texts and for their source texts produces good summaries.

In the evaluation of methods that combine **three** knowledge types (RST, CST and subtopics), RCT-4 had better performance. However, RC-4 is slightly better than

RCT-4. Several factors may contribute to this: (1) the segmentation and clustering of subtopics may not be as good as expected; (2) the way to deal with relevant subtopics may not be appropriate; or (3) it may not be advantageous to invest in subtopics.

All methods of RC and RCT groups were better than those that used the models in isolation (RST group and CSTSumm) in terms of recall and F-measure. With the exception of RCT-1, those methods also outperform RSumm in terms of F-measure. This shows that the combination of semantic discourse knowledge positively affects the production of summaries. At this time of analysis, it is known other advantages of the methods: (1) to use RST to assign scores to full sentences (and not to parts of sentences) and normalized by the height of the tree is a good strategy; and (2) to maintain full sentences generate more informative summaries.

If we only consider F-measure, the three methods with better performance are: RC-4, RC-1 and RCT-4, in this order. If we manually judge them, RC-1 produces summaries with many problems of linguistic quality due to the elimination of satellites. We run t-tests for pair of methods for which we wanted to check the statistical difference. The F-measure difference is not significant when comparing RC-4 and RCT-4 with RSumm (with 95% confidence), but is for CSTSumm. When comparing RC-4 to RCT-4, there is not statistical difference.

7. Final remarks

We have introduced some new methods for MDS that combine different knowledge: RST, CST and subtopics. As far as we know, this is the first time RST is applied for MDS. From its isolated study, it was possible to find clues on how RST associated with a multi-document model could contribute to content selection. The results are more informative summaries than previous approaches. The information on subtopics and how to use it needs more investigation; summaries produced using subtopics are similar to the ones based only on RST and CST.

Acknowledges

The authors are grateful to FAPESP and CAPES for supporting this work.

References

Afantenos, S.D.; Karkaletsis, V; Stamatopoulos, P; Halatsis, C. (2007). Using synchronic and diachronic relations for summarization multiple documents describing evolving events. *Journal of Intelligent Information Systems*, Vol. 30, N. 3, pp. 183-226.

Bollegala, D.; Okazaki, N.; Ishizuka, M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management*, Vol. 46, N. 1, pp. 89-109.

Cardoso, P.C.F. (2014). *Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, November, 182p.

Cardoso, P.C.F.; Maziero, E.G.; Castro Jorge, M.L.R.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105. Cuiabá/MT, Brazil.

Cardoso, P.C.F.; Taboada, M.; Pardo, T.A.S. (2013). On the contribution of discourse to topic segmentation. In: *Proceedings of the 14th Annual SIGDial Meeting on Discourse and Dialogue*, pp. 92-96. Metz, France.

Castro Jorge, M.L.R. (2015). *Modelagem gerativa para sumarização automática multidocumento*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, November, 151p.

Castro Jorge, M.L.R.; Pardo, T.A.S. (2010). Formalizing CST-based Content Selection Operations. In: *Proceedings of the 9th International Conference on Computational Processing of Portuguese Language* - PROPOR, pp. 25-29. April 27-30, Porto Alegre/RS, Brazil.

Ercan, G.; Cicekli, I. (2008). Lexical cohesion based topic modeling for summarization. In: *Computational Linguistics and Intelligent Text Processing*, pp. 582-592. Springer Berlin Heidelberg.

Haghighi, A.; Vanderwende, L. (2009). Exploring content models for multi-document summarization. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* - NAACL, pp. 362-370. Boulder/Colorado.

Harabagiu, S.; Lacatusu, F. (2010). Using topic themes for multi-document summarization. *ACM Transactions on Information Systems*, Vol. 28, N. 3, pp. 13-45.

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

Hennig, L. (2009). Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis. In: *Proceedings of the Recent Advances in Natural Language Processing*, pp. 144-149.

Kumar, Y.J.; Salim, N.; Abuobieda, A.; Albaham, A.T. (2014). Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing*, Vol. 21, pp. 265-279.

Jaidka, K.; Khoo, C.; Na, J-C. (2010). Imitating human literature review writing: an approach to multi-document summarization. In: *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pp. 116-119.

Lin, C-Y. (2004). ROUGE: a package for Automatic Evaluation of Summaries. In: *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74-81. Barcelona, Spain.

Louis, A.; Joshi, A.; Nenkova, A. (2010). Discourse indicators for content selection in summarization. In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 147-156. Association for Computational Linguistics.

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

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

Marcu, D. (1997). From discourse structures to text summaries. In: *Proceedings of the ACL*, Vol. 97, pp. 82-88.

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

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document Structure. In: *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, pp. 74-83. Hong Kong-China.

Ribaldo, R. (2013). *Investigação de Mapas de Relacionamento para Sumarização Multidocumento*. Monografia de Conclusão de Curso. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, November, 61p.

Ribaldo, R.; Cardoso, P.C.F.; Pardo, T.A.S. (2013). Investigação de Métodos de Segmentação e Agrupamento de Subtópicos para Sumarização Multidocumento. In: *Proceedings of 3rd Workshop on Information and Human Technology* - TILic, pp. 25-27. October 21-23, Fortaleza/Brazil.

Saggion, H.; Poibeau, T. (2013). Automatic text summarization: Past, present and future. *Multisource, Multilingual Information Extraction and Summarization*. Springer Berlin Heidelberg, pp. 3-21

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

Salton, G.; Singhal A.; Mitra, M; Buckley, C. (1997). Automatic text Structuring and summarization. *Information Processing & Management*, Vol. 33, N. 2, pp. 193-207.

Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2010). A Comprehensive Comparative Evaluation of RST-Based Summarization Methods. *ACM Transactions on Speech and Language Processing*, Vol. 6, N. 4, pp. 1-20.

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

Zhang, Z.; Goldenshon, S.B.; Radev, D.R. (2002). Towards CST-Enhanced Summarization. In: *Proceedings of the 18th National Conference on Artificial Intelligence*, pp. 439-446. Edmonton/Canada.