Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information

Rafael Ribaldo[1], Ademar Takeo Akabane[1],
Lucia Helena Machado Rino[2], Thiago Alexandre Salgueiro Pardo[1]

[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[2] Departamento de Computação, Universidade Federal de São Carlos

{ribaldo,takeusp}@grad.icmc.usp.br, lucia@dc.ufscar.br, taspardo@icmc.usp.br

Abstract. In this work we investigate the use of graphs for multi-document summarization. We adapt the traditional Relationship Map approach to the multi-document scenario and, in a hybrid approach, we consider adding CST (Cross-document Structure Theory) relations to this adapted model. We also investigate some measures derived from graphs and complex networks for sentence selection. We show that the superficial graph-based methods are promising for the task. More importantly, some of them perform almost as good as a deep approach.

Keywords: summarization, graphs, discourse

1 Introduction

Nowadays, with the huge and growing amount of information available in the Web and the sparse time to read and grasp it, manual analysis of the conveyed documents becomes almost impossible. According to a research conducted by the International Data Corporation [12], textual information in digital format currently amounts to c.a. 1.8 zettabytes, a number nine times bigger than five years ago. This makes evident that automatic treatment of texts is necessary, including information retrieval and extraction, topic detection and tracking, and text summarization, which is the focus of this paper.

Multi-Document Summarization (MDS) is defined as the task of automatically producing a unique summary of a set of documents on the same topic [17]. This task differs from single-document summarization in that it must detect and treat many phenomena that are typical of relating information on an inter-document basis. Important multi-document phenomena comprise dealing with redundant, complementary, and contradictory information, and both retrieving adequate links provided by referring expressions and ordering events and facts in time. More profound challenges refer to making writing styles uniform, fusing information, and balancing different perspectives of the same stories, among others. Practical applications for MDS include, e.g., making a

web search engine produce summaries of groups of news texts based on users queries, instead of just hits as normally do Google or Google News; indexing complex work, such as collections of scientific articles on a specific theme; producing biographies or synthesizing diverse opinions (being them in agreement or not) about a topic.

Traditional approaches to summarization comprise the well-known surface (shallow/superficial) and deep-based ones. The former uses relatively shallow linguistic information or no linguistic information at all, whilst the latter makes significant use of linguistic information during processing, usually including world knowledge and discourse models. In this paper, we address surface and hybrid approaches by dealing with graphs and discourse information processing.

Modeling texts as graphs implies normally having as their vertices (or nodes) text segments and as their links information on how these nodes relate to each other. Usually, text segments may be words, sentences, or paragraphs. At this level, having a surface-based or deep-based approach refers basically to how graph links depict information relationships. For summarization purposes, graph metrics signal the importance of a text segment. Based on those, the segment may be chosen to compose a summary or not. [3], [4], [13], [22] and [31] are examples of works that follow such approach. In particular, [3], [4] and [13] use a special type of graph, the complex networks. These differ from simple graphs in that they follow complex principles of organization and usually have a large number of nodes and show special topographic features [2].

In summarization, adding discourse features for deep processing implies having the system making decisions based on linguistic knowledge. In this case, it is possible to keep the granularity of the text segments similar to those used for a graph organization. However, weighting the importance of segments is now usually dependent on deeper text analysis. As a result, linguistically motivated relationships between those text spans may be depicted. The Cross-document Structure Theory (CST) [26] is an exemplar of the discourse models or theories that are used in MDS. It consists in a finite set of relations that are used to relate information from different texts, including relations for equivalent and contradictory segments, elaborated topics, citation of sources and authorship identification, etc. [1], [7], [9], [27], [28], [29] and [34] are good representatives of this research line. While some of these works explicitly use discourse models, others try to approach such relationships only indirectly.

In this paper we explore some graph-based summarization methods, aiming at (i) adapting to MDS the traditional Relationship Map [31], which has been originally proposed for single-document summarization, (ii) evaluating the impact of incorporating knowledge on the CST relations in the adapted model, and (iii) investigating how good content selection may be by taking into account some graph and complex network metrics. We focus on the production of generic and informative summaries. Experiments using a corpus of Brazilian Portuguese news texts were carried out to evaluate the contribution of each MDS method. Only summary informativeness was considered. Graph-based methods turned out to be well suited for the envisioned task. Actually, the results were close to the best ones obtained for Portuguese when a deep CST-based approach was considered. Since the latter is very effortful, MDS based on surface

methods seem more promising and scalable. We also show that adding CST relations to the graph metrics does not alter the quality of content selection. Instead, it only reinforces that graph-based methods are adequate for MDS so far.

In the next section, we briefly review the main initiatives related to the work presented in this paper. In Section 3, we present the graph-based methods that we explore. MDS assessment is described in Section 4, followed by some final remarks in Section 5.

2 Related Work

Graphs have shown to be applicable to many Natural Language Processing applications [21] and there are several graph-based approaches for both single and multi-document summarization (see, for example, [3], [4], [7], [11], [13], [18], [22], [31] and [32]). In this section we briefly introduce the ones that served as the basis for our work.

The work of Salton et al. [31] probably introduced the first widespread graph-based approach to single-document summarization. In the proposed relationship map method, the authors model a text as a graph/map in the following way: each paragraph is represented as a node, and weighted links are established only among paragraphs that have some lexical similarity. This may be pinpointed through lexical similarity metrics. The choice for representing paragraphs (and not words, clauses, or sentences) as nodes is due to the assumption that paragraphs provide more information surrounding their main topics and, thus, may be used for more coherent and cohesive summaries. For summarization purposes, only the highly weighted links are considered: given a graph with N nodes, only the 1.5 * N best links provide the means to select paragraphs to include in a summary. Once established such a threshold, three different ways of traversing a graph are proposed, namely, the Bushy Path, the Depth-first Path, and the Segmented Bushy Path. In the Bushy Path, the density, or *bushiness,* of a node is defined as the number of connections it has to other nodes in the graph. So, a highly linked node has a large overlapping vocabulary with several paragraphs, representing an important topic of the text. For this reason, it is a candidate for inclusion in the summary. Selection of highly connected nodes is done until compression rate is satisfied in the Bushy Path. In this way, the coverage of the main topics of the text is very likely to be good. However, the summary may be non-coherent, since relationships between every two nodes are not properly tackled. To overcame that, instead of simply selecting the most connected nodes, the Depth-first Path starts with some important node (usually the one weighted the highest) and continues the selection with the nodes (i) that are connected to the previous selected one and (ii) that come after it in the text, also considering selecting the most connected one among these, trying to avoid sudden topic changes. This procedure is followed until the summary is fully built. Its advantage over the Bushy Path is that more legible summaries may be built due to choosing sequential paragraphs. However, topic coverage may be damaged. The Segmented Bushy Path aims at overcoming the bottlenecks introduced by the other two methods: it tackles the topic representation

problem by first segmenting the graph in portions that may correspond to the topics of the text. Then, it reproduces the Bushy Path method in each subgraph. It is guaranteed that at least one paragraph of each topic will be selected to compose the summary. In their evaluation, Salton et al. show that the methods produce good results for a corpus of encyclopedic texts, with the Bushy Path being the best method.

Antiqueira et al. [3] [4] use complex networks to model texts for single-document summarization. In their networks, each sentence is represented as a node and links are established among sentences that share at least one noun. Once the network is built, sentence ranking is performed by using graph and complex network measures. The best-ranked sentences are thus selected to compose the summary, as usual. From several measures explored by Antiqueira et al., we selected only three for our work: degree, clustering coefficient and shortest path. The well-known degree measure indicates how many connections one node has to the others. It accounts for the density measure in Salton et al.'s model. It is assumed that the bigger the degree of a node, the more important the corresponding sentence is. Notice that this selection strategy is also similar to Salton et al.'s Bushy Path (although graph construction is different). The clustering coefficient measure signals how nodes tend to cluster together. It was introduced in [33] and, for summarization, it may indicate central topics to the summary. Also well-known, the shortest path measure indicates the length of the shortest path between 2 nodes. Antiqueira et al. use the average of the lengths of the shortest paths from a node to every other node in the network as an indication of its importance: the nearer a node is (in average) to the other nodes, the better the sentence is to compose the summary. The authors evaluated their work with news texts in Brazilian Portuguese using the TeMário corpus [23]. Good results were achieved, but they did not outperform the best single-document summarization system for Portuguese - SuPor [13], which is based on machine learning over a rich set of features.

Turning to MDS, Castro Jorge and Pardo [7] model several texts as just one graph, with nodes representing sentences and links representing discourse relations among the sentences. Discourse relations are CST relations [26], in particular, the ones refined by Maziero et al. [19]. Such relations pinpoint similar and different sentence content, as well as different writing styles and decisions among the texts. For sentence selection, sentences that have more relations/links to others are preferred, similarly to those approaches for single-document summarization. A further step here is verifying whether a selected sentence is redundant, i.e., if it embeds information that has already been conveyed by other previously selected sentences. By analyzing the CST relations, it is possible to detect redundancy of a candidate sentence. In this case, such sentence is ignored and the next candidate sentence in the graph is considered. Castro Jorge and Pardo evaluated their approach on the CSTNews corpus of news texts in Brazilian Portuguese [5]. This corpus is manually annotated with CST and it also conveys manual summaries. The results were the best produced so far for this language, for the MDS task.

3 Graph-based Methods for MDS

In this section we report on graph-based methods adapted for the MDS task and compare them with previous results for Portuguese. We considered three distinct models, namely: the classical Relationship Map one [31], now conveying a multi-document representation; its enrichment with CST relations [19] [26]; and a proposal using graphs and complex networks metrics, following the work of Antiqueira et al. [3] [4]. Annotating relationship maps with CST relations aimed at verifying their impact in the Salton et al.'s model. The hypothesis here was that since the best results for Portuguese were obtained by a CST-based approach, as reported by Castro Jorge and Pardo [7], an MDS model based upon relationship maps could benefit from that enrichment.

As far as we know, this is the first time that the classical Relationship Map approach is used for MDS purposes. Complex networks had already been explored in this context, providing the means for a machine learning solution [9] along with other surface and CST-based features. It is also interesting to notice that the approaches in this paper are complementary to current work in MDS for the Portuguese language, which has mainly focused on deep-based approaches (e.g., see [6], [7], [8], [9] and [10]).

For adapting the Relationship Map model, sentences (instead of paragraphs) are represented in nodes. To our view, this may allow for more refined summarizing decisions, although it may endanger coherence and cohesion of the final summaries. A multi-document graph is built in the following way: each sentence from a group of texts is represented by a node in the graph; links between nodes signal how similar the related sentences are. The cosine similarity measure [30] is used to compute lexical similarity among sentences, after removing stopwords and stemming the remaining words. A graph conveys as many nodes as sentences in the focused set of texts, and duplication is not verified beforehand. It turns out that even identical sentences, which are likely to occur in a set of texts on the same topic, are represented in the graph as different nodes. This is the reason for treating redundancy afterwards. As suggested by Salton et al., if there are N nodes in a graph, only the 1.5 * N best weighted links are considered for MDS. After modeling the texts in a graph, sentence selection proceeds as described in Section 2, for two paths only: the Bushy Path and the Depth-first Path. The Segmented Bushy Path was not adopted because it requires more sophisticated reasoning, which must be investigated in future work. Another modification took place in adapting the Depth-first Path for MDS: in the original single-document summarization model, paragraphs selection is subordinated to previous paragraphs choices, in order to observe paragraph ordering for cohesiveness. In MDS, it only makes sense to observe sentence ordering if sentences are selected from the same text. For this reason, this restriction is relaxed if the sentences under analysis come from different texts.

Finally, the summary is built with the most prominent sentences of all texts. As expected, there is a high degree of redundancy in texts, which is natural in multi-document analysis tasks. Such redundancy is frequently indicated by links with very high degree of similarity. Therefore, in order to avoid having redundant sentences in the

summary, we stipulated a redundancy limit that a new selected sentence may have in relation to any of the previously selected sentences. If this limit is reached, this new sentence is considered redundant and ignored, and the summarization process goes to the next candidate sentence; otherwise, the sentence is included in the summary. The limit was defined as the sum of the highest and the lowest cosine values in the graph divided by 2, resulting in an intermediate value between them. So, the redundancy limit is not fixed; it depends on the graph produced for the input texts. We believe that this flexibility allows a better redundancy treatment for varied domains and text types and genres.

The number of selected sentences to compose the summary is also limited by the specified compression rate, which gives the proportion between the length of the summary and the length of the texts. In this work, we consider the length of the longest text, in number of words. Therefore, a 70% compression rate indicates that the summary may be 30% long, in relation to the length of the longest text in the group.

To illustrate the process under focus here, Figures 1, 2 and 3 show 2 short source texts (from CSTNews corpus [5]) and the corresponding automatic summary produced by the previous method, using a 70% compression rate. The original language of the texts is Brazilian Portuguese, and so is the summary one. For clarity, English translation is also provided. Both Bushy and Depth-first Paths produced the same summary for the text set shown. One may see that the summary is good.

A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008.
A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico. Por isso, Jade participará do evento em Buenos Aires, na Argentina, única cidade da América do Sul a receber o símbolo dos Jogos.
O revezamento terminará em 8 de agosto, primeiro dia das Olimpíadas de Pequim.

=== English translation ===
The gymnast Jade Barbosa, who won three medals in the Pan-American Games in Rio in July, won an Internet poll and will be the Brazilian representative in the Olympic torch relay to Beijing in 2008.
The torch will pass through twenty countries, but Brazil is not in the Olympic route. Therefore, Jade will participate in the event in Buenos Aires, Argentina, the only city in South America to receive the symbol of the Games.
The relay will end on August 8, the first day of the Beijing Olympics.

Figure 1. First Text

Um dos destaques desta temporada do esporte brasileiro, a ginasta Jade Barbosa foi escolhida, na noite desta terça-feira, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim.
Em votação pela internet, a ginasta recebeu mais de 100 mil votos e superou o nadador Thiago Pereira, que ganhou seis ouros nos Jogos Pan-Americanos.
O Brasil não faz parte do trajeto da tocha olímpica. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.

Aos 16 anos, Jade conquistou três medalhas no Pan: ouro na disputa dos saltos, prata na apresentação por equipes e bronze no solo.
Ao todo, a chama olímpica percorrerá 20 países antes de chegar a Pequim para a abertura da competição, no dia 8 de agosto.

=== English translation ===
One of the highlights of this season in Brazilian sports, the gymnast Jade Barbosa was chosen Tuesday night to be the representative of Brazil in the torch relay in Beijing Olympic Games.
In an Internet poll, the gymnast has received over 100,000 votes and has beaten the swimmer Thiago Pereira, who won six gold medals in the Pan-American Games.
Brazil is not part of the Olympic torch route. In South America, the flame will pass through Buenos Aires, where Jade will participate in the relay on April 11.
At the age of 16, Jade won three medals in the Pan: gold in the Vault, silver in the Team competition, and bronze in the Floor.
In general, the Olympic flame will travel through 20 countries before arriving in Beijing for the competition opening on August 8.

Figure 2. Second Text

A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.

=== English translation ===
The gymnast Jade Barbosa, who won three medals in the Pan-American Games in Rio in July, won an Internet poll and will be the Brazilian representative in the Olympic torch relay to Beijing in 2008. In South America, the flame will pass through Buenos Aires, where Jade will participate in the relay on April 11.

Figure 3. Summary

In a variation of the above method, we add CST relations to the graph in order to verify the impact of discourse information in the summary informativeness. We expect that the use of such refined knowledge may improve sentence selection. Two strategies are followed in order to consider CST relations in the graph: (i) simply summing the number of relations that each sentence presents to its number of links, without considering the relation type itself, and (ii) assigning scores to every CST relation that appears in each sentence and, then, summing up those scores to the number of links of that sentence. Relations that account for content matters are considered more important and receive values from 0.5 to 1, with relations that indicate redundancy getting higher values, since redundancy usually indicates importance in MDS [17]; relations that deal mainly with different writing styles and decisions receive values below 0.5. After considering CST in the graph, sentence selection is performed in the same way it is before.

Finally, our last method consists in using some graph and complex network measures to rank the sentences in the graph, in order to select the candidates to a summary. A graph is built in the same way as before: sentences are represented in the nodes and links are

weighted according to their lexical similarity with other sentences through the cosine similarity measure. However, differently from Salton et al. method, in this approach we keep all the links, as Antiqueira et al. [3] [4] do. We use degree, clustering coefficient and shortest path measures to rank sentences, and the best-ranked ones are selected for inclusion in the summary. While degree and shortest path measures are also usual graph metrics, the clustering coefficient one is typical of complex networks.

Independently from the chosen measure to score the importance of the candidate sentences, redundancy is treated in the same way as that by Salton et al. adapted method, i.e., by applying a redundancy threshold to select sentences.

4 Evaluation of the proposed methods

The evaluation of the proposed MDS strategies was carried out over the CSTNews corpus, which amounts to 140 news texts in Brazilian Portuguese divided into 50 groups. Each group has from 2 to 3 texts on a same topic, having in average 49 sentences and 945 words. This corpus includes manual multi-document summaries (one per group) with 70% compression rate (in relation to the longest text). The texts are manually annotated with CST, with satisfactory agreement values among the annotators [5]. That indicates that the annotation is reliable and, for this reason, was used for evaluation in our work.

We used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [15], a tool created to enable direct comparison among an automatically generated summary and the corresponding human summaries. It provides precision, recall and f-measure values based on the number of common n-grams conveyed by the summaries. As its authors showed, ROUGE is as good as humans in ranking summaries according to their informativeness, being a good indicator of the quality of the summarization methods.

In this work, we report ROUGE results just for 1-gram comparisons, hence using the so-called ROUGE-1. This metric has been shown to be enough for comparing summary informativeness. We also produced ROUGE results for other n-gram comparisons, but do not show them here, since they corroborated the ROUGE-1 results.

We also evaluated summaries produced by other systems that can tackle texts on Brazilian Portuguese, namely: GistSumm [24] [25], Castro Jorge and Pardo's CSTSumm [7] [8], MEAD [28], and a baseline method that randomly select sentences. GistSumm was the first MDS system produced for Portuguese and follows a very naive approach, by simply juxtaposing all the texts and selecting the sentences according to the frequency of their words. CSTSumm has been graded so far the best MDS system for Portuguese and follows the purely CST-based method presented in Section 2. MEAD is one of the most famous multi-lingual MDS systems and it is based on centroids, sentence position and simple lexical features extracted from the sentences.

Table 1 shows the results of our assessment, ordered according to the systems f-measures. As expected, the best system is still CSTSumm. This is certainly due to its deep approach, which usually provides better results. Similarly to Antiqueira et al.'s findings

for single-document summarization, the degree measure yields very good results for MDS, following CSTSumm in the rank. Amongst Salton et al.'s methods, the Bushy Path was slightly better. This has also been reported by Salton et al. before. It is noticeable that MEAD f-measure is smaller than most of those provided by our proposed summarization strategies. It is also curious that some systems were worse than the random baseline. The MDS system based on the clustering coefficient measure was the worst one.

Table 1. Evaluation results

System/Method	Precision	Recall	F-measure
CSTSumm	0.5761	0.5065	0.5297
Degree	0.5328	0.5037	0.5155
Shortest Path	0.5306	0.5009	0.5131
Bushy Path	0.4844	0.5397	0.5083
Bushy Path with CST	0.4844	0.5397	0.5083
Depth-first Path	0.4811	0.5340	0.5040
Depth-first Path with CST	0.4811	0.5340	0.5040
MEAD	0.5242	0.4602	0.4869
Random Baseline	0.4494	0.4864	0.4652
GistSumm	0.3599	0.6643	0.4599
Clustering coefficient	0.4671	0.4476	0.4560

The system with the highest precision was also CSTSumm. Although being one of the worst systems, GistSumm obtained the highest recall value, much better than all other values. It is followed by Salton et al. methods.

While some results were expected, others were surprising. Except the clustering coefficient method, all the others were good, performing closely to CSTSumm, possibly indicating that it is worthy following a superficial approach, mainly if we consider that such methods are more scalable and do not depend on sophisticated resources or models based on discourse annotation. If we consider that the results of CSTSumm evaluation were obtained with a manually CST-annotated corpus and that an automatic CST annotation would produce worse summarization results (as it usually happens in the area), it would not be a surprise if some our proposed graph-based methods would outperform CSTSumm with automatic annotation. This is in line with [16], which had already indicated that discourse information could be replaced by superficial measures without great loss. However, to investigate that further, in future work we intend to use a CST parser available for Portuguese [20] in order to pursue automatic annotation of the same CSTNews corpus and reproduce the above evaluation setting.

It was also surprising that enriching Salton et al. paths with CST did not alter any of the results. In fact, CST was only useful to reinforce the results obtained by using the original path models. This observation also corroborates the conclusion that graphs are promising for MDS.

The machine learning solution to MDS presented in [9], which uses complex networks and other superficial and CST-based features to induce decision trees for classifying

important and non-important sentences for composing summaries, produced very good summaries for Portuguese. The CSTNews corpus was also used in their evaluation. However, their results may not be directly compared to ours, since they used another evaluation strategy, based on training and test sets (not using the full corpus for evaluation, therefore). In spite of that, according to the results in [9], which also evaluated CSTSumm and MEAD summaries, the machine learning approach ranked right after CSTSumm, which was the best system in the corresponding evaluation. Another interesting issue in that work is that the clustering coefficient feature showed to be a relevant feature when used together with other CST-based features. In this work such measure was the worst one when used alone, probably indicating that it is better used as complementary information rather than a unique indicator of sentence importance.

Based on the results in [9], we may expect that the same machine learning approach would rank after CSTSumm in this paper and would also suffer from using automatically CST-annotated texts, probably being outperformed by our graph-based methods.

5 Final Remarks

We showed in this paper that graph-based methods for MDS of texts in Brazilian Portuguese provide very good results, being close to the best system available for that language. Next steps consist in developing the Segmented Bushy Path to MDS, as well as to evaluate the methods not only for informativeness, but for other quality criteria, as coherence and cohesion. For such evaluation, it will also be important to incorporate in the methods a sentence ordering method, since the sentences are currently juxtaposed in a summary in the order they are selected. One of the sentence ordering strategies investigated in [14] might be used. Finally, it may also be worthy to explore other graph-based MDS strategies, as the ones proposed in [22].

Acknowledgements

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

References

1. Afantenos, S.D., Doura, I., Kapellou, E., Karkaletsis, V.: Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In: Proceedings of the 3rd Hellenic Conference on Artificial Intelligence, pp. 410-419. May 5-8, Samos Island, Greece (2004)
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74 (1), 47-97 (2002)

3. Antiqueira, L.: Desenvolvimento de Técnicas Baseadas em Redes Complexas para Sumarização Extrativa de Textos. MSc Dissertation. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. March, São Carlos/SP, Brazil. 124p (2007)
4. Antiqueira, L., Oliveira Jr., O.N., Costa, L.F., Nunes, M.G.V.: A Complex Network Approach to Text Summarization. Information Sciences 179 (5), 584-599 (2009)
5. 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.: 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. October 26, Cuiabá/MT, Brazil (2011)
6. Cardoso, P.C.F., Pardo, T.A.S., Nunes, M.G.V.: Métodos para Sumarização Automática Multidocumento Usando Modelos Semântico-Discursivos. In: Proceedings of the 3rd RST Brazilian Meeting, pp. 59-74. October 26, Cuiabá/MT, Brazil (2011)
7. Castro Jorge, M.L.R., Pardo, T.A.S.: Experiments with CST-based Multidocument Summarization. In: Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing, pp. 74-82. July 16, Uppsala, Sweden (2010)
8. Castro Jorge, M.L.R.: Sumarização automática multidocumento: seleção de conteúdo com base no Modelo CST (Cross-document Structure Theory). MSc Dissertation. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. April, São Carlos/SP, Brazil. 86p (2010)
9. Castro Jorge, M.L.R., Agostini, V., Pardo, T.A.S.: Multi-document Summarization Using Complex and Rich Features. In: Anais do VIII Encontro Nacional de Inteligência Artificial, pp. 1-12. July 19-22, Natal/RN, Brazil (2011)
10. Castro Jorge, M.L.R., Pardo, T.A.S.: A Generative Approach for Multi-Document Summarization using the Noisy Channel Model. In: Proceedings of the 3rd RST Brazilian Meeting, pp. 75-87. October 26, Cuiabá/MT, Brazil (2011)
11. Erkan, G., Radev, D.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22 (1), 457-479 (2004)
12. Gantz, J., Reinsel, D.: Extracting Values from Chaos. IDC IView. June (2011)
13. Leite, D.S.: Um Estudo Comparativo de Modelos Baseados em Estatísticas Textuais, Grafos e Aprendizado de Máquina para Sumarização Automática de Textos em Português. MSc Dissertation. Departamento de Computação, Universidade Federal de São Carlos. December, São Carlos/SP, Brazil. 231p (2010)
14. Lima, J.B.P., Pardo, T.A.S.: Ordenação de Sentenças em Sumários Multidocumento: Uma Abordagem Utilizando Relações CST. In: Proceedings of the 2nd STIL Student Workshop on Information and Human Language Technology, pp. 1-3. October 24-25, Cuiabá/MT, Brazil (2011)
15. Lin, C.Y., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 71-78. May 27 - June 1, Edmonton, Canada (2003)
16. Louis, A., Joshi, A., Nenkova, A.: Discourse indicators for content selection in summarization. In: Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialog, pp. 147-156. September 24-25, Tokyo, Japan (2010)
17. Mani, I.: Automatic Summarization. John Benjamins Publishing Co., Amsterdam (2001)
18. Mani, I., Bloedorn, E.: Summarizing Similarities and Differences Among Related Documents. Information Retrieval 1 (1-2), 35-67 (1997)

19. Maziero, E.G., Castro Jorge, M.L.R., Pardo, T.A.S.: Identifying Multidocument Relations. In: Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science, pp.60-69. June 8-12, Funchal/Madeira, Portugal (2010)
20. Maziero, E.G., Pardo, T.A.S.: Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pp. 1-10. October 24-26, Cuiabá/MT, Brazil (2011)
21. Mihalcea, R., Radev, D.: Graph-based Natural Language Processing and Information Retrieval. Cambridge University Press (2011)
22. Mihalcea, R., Tarau, P.: An Algorithm for Language Independent Single and Multiple Document Summarization. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing. October 11-13, Jeju Island, Korea (2005)
23. Pardo, T.A.S., Rino, L.H.M.: TeMário: Um Corpus para Sumarização Automática de Textos. Technical Report NILC-TR-03-09. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. October, São Carlos/SP, Brazil. 13p (2003)
24. Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V.: GistSumm: A Summarization Tool Based on a New Extractive Method. In: Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken, pp. 210-218. June 26-27, Faro, Portugal (2003)
25. Pardo, T.A.S.: GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades. Technical Report NILC-TR-05-05. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. February, São Carlos/SP, Brazil. 8p (2005)
26. Radev, D.R.: 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. October 7-8, Hong Kong, China (2000)
27. Radev, D.R., Jung, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In: Proceedings of the ANLP/NAACL Workshop on Automatic Summarization, pp. 21-30. April 30, Seattle, USA (2000)
28. Radev, D.R., Blair-Goldensohn, S., Zhang, Z.: Experiments in single and multidocument summarization using MEAD. In: Proceedings of the 1st DUC Workshop on Text Summarization. September 13-14, New Orleans, USA (2001)
29. Radev, D.R., Blair-Goldensohn, S., Zhang, Z., Raghavan, R.S.: NewsInEssence: A system for domain-independent, real-time news clustering and multi-document summarization. In: Proceedings of the 1st International Conference on Human Language Technology Research. March 18-21, San Diego, USA (2001)
30. Salton, G.: Automatic text processing. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1988)
31. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic Text Structuring And Summarization. Information Processing & Management 33 (2), 193-207 (1997)
32. Wan, X.: 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. October 25-27, Waikiki, USA (2008)
33. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440-442 (1998)
34. Zhang, Z., Blair-Goldensohn, S., Radev, D.R.: Towards CST-enhanced summarization. In: Proceedings of the 18th National Conference on Artificial Intelligence, pp. 439-446. July 28 - August 1, Edmonton, Canada (2002)