Explorando mapas de relacionamento com base em subtópicos para sumarização multidocumento

Exploring the subtopic-based relationship map strategy for multi-document summarization

Rafael Ribaldo [1]
Paula Christina Figueira Cardoso [2]
Thiago Alexandre Salgueiro Pardo [3]

Data de submissão: 04/10/2015, Data de aceite: 04/04/2016

Abstract: Neste artigo, foi adaptada e explorada uma estratégia de geração de sumários multidocumento com base na abordagem de mapas de relacionamento, a qual representa textos como grafos (mapas) de segmentos inter-relacionados e aplica técnicas de percurso em grafo para produzir os sumários. Em particular, utilizou-se o Caminho Denso-Segmentado, um método sofisticado que tenta representar em um sumário os principais subtópicos dos textos de origem, mantendo a informatividade do sumário. Além disso, também foram investigadas algumas técnicas de segmentação e agrupamento de subtópicos bem conhecidas, a fim de selecionar corretamente as informações mais relevantes para compor o sumário final. Mostra-se que este método de sumarização baseado em subtópicos supera outros métodos de sumarização multidocumento e atinge resultados do estado da arte, competindo com os métodos de sumarização profundos mais sofisticados da área.

***P*alavras-chave:** sumarização multidocumento, mapas de relacionamento, grafos, subtópicos, segmentação e agrupamento

[1] Núcleo Interinstitucional de Linguística Computacional (NILC), Departamento de Ciências de Computação, Instituto de Ciências Matemáticas e de Computação, USP,São Carlos/SP, Brasil. {ribaldorafael@gmail.com}
[2] Departamento de Ciência da Computação, Universidade Federal de Lavras, Avenida Doutor Sylvio Menicucci, 1001, Campus Universitário. CEP: 37200-000 - Lavras/MG, Brasil. {paula.cardoso@dcc.ufla.br}
[3] Núcleo Interinstitucional de Linguística Computacional (NILC), Departamento de Ciências de Computação, Instituto de Ciências Matemáticas e de Computação, USP, São Carlos/SP, Brasil. {taspardo@icmc.usp.br}

Abstract: In this paper we adapt and explore a strategy for generating multi-document summaries based on the relationship map approach, which represent texts as graphs (maps of interrelated segments and apply traversing techniques for producing the summaries. In particular, we work on the Segmented Bushy Path, a sophisticated method which tries to represent in a summary the main subtopics from the source texts while keeping its informativeness. In addition, we also investigate some well-known subtopic segmentation and clustering techniques in order to correctly select the most relevant information to compose the final summary. We show that this subtopic-based summarization method outperforms other methods for multi-document summarization and achieves state of the art results, competing with the most sophisticated deep summarization methods in the area.

Keywords: multi-document summarization, relationship maps, graphs, subtopics, segmentation and clustering

1 Introduction

In recent decades, many new technologies have emerged, bringing with it an increasing volume of information. Nowadays, many resources such as search engines, blogs and social networks make accessible an enormous amount of information and, therefore, processing all this becomes increasingly difficult. To have an idea, a study conducted by the International Data Corporation (IDC) showed that the volume of digital content would grow to 8 ZB (zettabytes) in the last year, driven by steady growth of internet users, social networks and smart devices, which enable new ways of working and communicating. In this context, multi-document summarization (MDS) appears as a tool that may assist people in acquiring relevant information in a short time.

Multi-document summarization aims at producing automatic summaries from a collection of source texts/documents about the same topic [1]. Some challenges for MDS are to deal with the multi-document phenomena, as redundant, complementary and contradictory information, to normalize varied writing styles (since the texts come from different authors), to temporally order events/facts (because the texts are written at different times), to balance different foci and perspectives, as well as to keep summary coherence and consistency. It also includes the traditional single document summarization challenges, as dealing with dangling anaphors and guaranteeing cohesion.

An equally important challenge for MDS is to tackle the topic/subtopic distribution in summaries. It is known that a text or a set of texts develop a main topic, exposing several 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 [2, 3, 4, 5]. A good summary should represent the main topics/subtopics of the source texts in order to be considered relevant and informative to the reader.

As an example, Figure 1 shows two texts (translated from the original language - Portuguese) and possible subtopic delimitations. Sentences are identified by numbers between square brackets. The identification of each subtopic is shown in angle brackets after the corresponding passage. As it is possible to see, the main topic is the health of Maradona, the famous Argentine soccer player (already retired). Notice that it is possible to find subtopics about the disease itself and the soccer match between Boca Juniors and River Plate teams.

Document 1	[S1]Maradona's personal doctor, Alfredo Cahe, revealed on Monday that a relapse of acute hepatitis that the ex-player suffers was the reason for his new hospitalization. [S2]Maradona had been discharged 11 days ago, but he was again hospitalized on Friday, and the medical reports did not specify what was wrong with the ex-player — Cahe ruled out ulcer or pancreatitis. <*subtopic: Maradona's relapse*> [S3]"Maradona had a relapse of acute hepatitis. Now he is stable. Despite the fact that he got better on Sunday, he should remain hospitalized", said Cahe to the newspaper La Nación. <*subtopic: Maradona's hepatitis*> [S4]Maradona, 46, developed toxic hepatitis due to excessive alcohol consumption, which had kept him hospitalized for 13 days before the most recent hospitalization. <*subtopic: history of Maradona's disease*> [S5]Cahe said that Maradona had not started to drink alcoholic beverages again, and that the causes of the relapse are being investigated. <*subtopic: Maradona's hepatitis*>
Document 2	[S1]Maradona had again health problems over the weekend. [S2]Hospitalized in Buenos Aires, he had a relapse and felt pain againg due to acute hepatitis, according to his personal doctor, Alfredo Cahe. <*subtopic: Maradona's relapse*> [S3]"Now his state of health is stable. Despite this improvement, he is still hospitalized", said the doctor, who has discarded the possibility that the ex-player has pancreatitis (inflammation of the pancreas, an organ located behind the stomach and that influences the digestion). [S4]Cahe emphasized that Maradona still has problems. [S5]"His liver values are not balanced and he is not well. But it is nothing serious", he said in an interview for the La Nación newspaper. <*subtopic: current state of health*> [S6]On Sunday, Maradona watched the 1-1 draw in the classic Boca Juniors and River Plate on television. [S7]Boca Juniors' fans, who turned out in large numbers to the stadium La Bombonera, led many banners and flags with messages of support for the Argentine idol. [S8]His daughter, Dalma, was in the stadium to watch the game. <*subtopic: support messages*>

Figure 1. Example of documents with diverse information

Considering the multi-document scenario, it is easy to find repeated subtopics in the texts, for example, <*subtopic: Maradona's relapse*>, and also unique subtopics that are not repeated and contain different details about the main topic, for example, <*subtopic: messages of support*>. In this case, before selecting content for a summary, it is necessary to find similar subtopics and clustering them according to a degree of similarity. Ideally, a multi-document summary should contain only once the key shared information among all the documents, plus other information unique to some individual documents that show to be relevant [6].

As an illustration of the necessity of subtopic treatment, we show in Figures 2 and 3 two summaries (also translated from Portuguese) automatically produced by GistSumm [7] and CSTSumm [8] systems, which use superficial and deep summarization methods, respectively. Superficial methods are those that make little or no use of linguistic knowledge, and are more scalable and robust in general. Deep methods, by contrast, make heavy use of linguistic knowledge, such as discourse models and semantic resources, being able to produce better results, but are more expensive and more narrowly applicable, typically. GistSumm system uses simple word frequency measures to identify the gist (the main idea) of the texts. CSTSumm system, in turn, uses discourse features for judging sentence relevance. We may notice that both summaries do not include all the subtopics present in the collection of texts, since the summarization strategies used by GistSumm and CSTSumm make uniform use of the sentences in different documents [9], i.e., the sentences are used without consideration of the subtopic distribution. We and also other authors [9, 10, 11] argue that sentences in the same collection may not be uniformly treated, because some sentences are more important than others, due to their different roles in the documents and the subtopics they belong to.

> "Maradona had a relapse of acute hepatitis. Now he is stable. Despite the fact that he got better on Sunday, he should remain hospitalized", said Cahe to the newspaper La Nación.
> "Now his state of health is stable. Despite this improvement, he is still hospitalized", said the doctor, who has discarded the possibility that the ex-player has pancreatitis (inflammation of the pancreas, an organ located behind the stomach and that influences the digestion).

Figure 2. Summary produced by GistSumm system

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

Figure 3. Summary produced by the CSTSumm system

In order to overcome the limitations of the current summarization strategies, in this paper we adapt and explore a classical summarization technique proposed by Salton et al. [3]. The authors have proposed single document summarization methods that are referred by "relationship maps", since a text is represented as a graph (a "map") of interrelated text segments and different traversing techniques are used to select the segments to compose a summary. We have already adapted two of the methods (called "bushy" and "depth-first" paths) for multi-document summarization (as reported in [12]). Here, we adapt and explore the most sophisticated method, the "Segmented Bushy Path", which addresses the subtopic issues for producing better summaries. As the Segmented Bushy Path was developed for single document summarization, it is necessary more than segmenting each text in subtopics for adapting it to MDS: it is also necessary to deal with subtopic correlations. For this reason, this paper also deals with strategies for subtopic segmentation and clustering.

We evaluate the adapted method in a benchmark collection for Portuguese language and show that it outperforms the other Salton et al. methods [3] and that it produces state of the art results, competing with the best deep method that we are aware of. We also comment on the delivering of our methods to the general public by incorporating them in an extension to a web browser, making the summarization system widely available.

The remainder of the paper is structured as follows. In the next section, we briefly review the main related work. In Section 3, we present the summarization algorithm and its steps. Experiments and evaluation results are reported in Section 4. In Section 5, we briefly describe our initiative to deliver to the final user our summarization methods. Finally, some final remarks are presented in Section 6.

2 Related Work

In what follows, we introduce the main related work and concepts that are the basis of this paper. We briefly review some related summarization methods already tested for Portuguese and mainly the ones of Salton et al. [3], which support our work, followed by a discussion of subtopic segmentation and clustering techniques. We conclude this section with a description of a discourse model that we test in this paper for subtopic clustering.

2.1 Text Summarization

Graphs have shown to be applicable to many Natural Language Processing applications [13] and there are several graph-based approaches for both single and multi-document summarization (see, for example, [3, 8, 9, 14, 15, 16, 17, 18, 19]).

Salton et al. [3] probably introduced the first widespread graph-based approach to single document summarization. In the proposed relationship map methods, 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 producing 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 the graph is built, three different ways of traversing the graph are proposed, namely, Bushy Path, Depth-first Path and 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 subtopic of the text. For this reason, it is a candidate for inclusion in the summary. Selection of highly connected nodes is done until the summary compression rate is satisfied in the Bushy Path. This way, the coverage of the main subtopics 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 overcome this, 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 subtopic 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, subtopic coverage is not guaranteed. The Segmented Bushy Path aims at overcoming the bottlenecks introduced by the other two methods. It tackles the subtopic representation problem by first segmenting the graph in portions that may correspond to the subtopics of the text. Then, it reproduces the Bushy Path method in each subgraph. This is done by selecting the most important paragraphs within a subtopic, and finally, uniting them with transitional paragraphs, which are chronologically prior to each first paragraph of each subtopic. It is guaranteed that at least one paragraph of each subtopic will be selected to compose the summary. In their evaluation, the authors showed that the methods produce good results for a corpus of encyclopedic texts, with the Bushy Path being the best one.

The first two methods adapted from [3] (Bushy Path and Depth-first Path) were already evaluated for MDS of texts written in Portuguese [12]. Using such methods for MDS, as we discuss later, implies in dealing with the multi-document phenomena, mainly with redundancy. The results, explained in more details in Section 4, are very promising, being close to the state of the art summarizers. It is important to note that, despite the use of paragraphs in [3] as the information unit to be selected, we chose to select sentences because of their more refined granularity, which allows for the construction of more informative summaries, as most of the works in the area usually do nowadays. The system for Portuguese that incorporates Salton et al. methods is referred by RSumm.

Antiqueira et al. [17, 18] use complex networks to model texts for single document summarization. In their networks/graphs, 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, as degree, clustering coefficient and shortest path, and the best ranked sentences are selected to compose the summary. Using some of the measures, such method was also adapted for MDS summarization for the Portuguese language [20, 12], producing good results. Such system was named RC-Summ.

Castro Jorge and Pardo [8], in a deep approach, model several texts as just one graph, with nodes representing sentences and links representing discourse relations among the sentences. Discourse relations are based on the ones predicted by the Cross-document Structure Theory [21]. 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. This method is the one used by CSTSumm, cited in the introductory section. Cardoso [22] improves this method, incorporating discourse relations that happen for each single document (following the Rhetorical Structure Theory [23]) and considering topic/subtopic distribution in the source texts, using the subtopic delimitation method that we report in this paper. Following the original work, we refer to this method by the RCT-4 acronym (which indicates that it was the 4th method variation investigated by Cardoso, using RST, CST and Topics - each word contributes with the first letter to the acronym).

It is also relevant to cite two more summarization approaches, that are not directly based on graphs, but that were also tested for Portuguese. MEAD, proposed by Radev et al. [24], is a very popular summarization system. The tool incorporates multiple strategies for selecting sentences for summarization, namely: 1) the position of sentences in their documents; 2) lexical distance of sentences in relation to the centroid, i.e., the central sentence of the texts; 3) the longest common subsequences among the sentences; and 4) presence of keywords in the sentences. It has been widely used in the area and produces good results. It was tested for Portuguese in [12]. To the best of our knowledge, GistSumm [7] was the first MDS system for Portuguese. It consists in a very simple approach, using word frequency to compute the most important sentence of the source texts and, then, using lexical similarity among this sentence and the other ones, it selects the best ranked ones to compose the final summary. It is considered a baseline system in the area.

The above works are some of the main ones for MDS in Portuguese that have been evaluated on the same corpus (which is a benchmark for MDS in Portuguese) with the same evaluation metrics and setup. For these reasons, they are the ones that we compare our approach to, as we will show in Section 4. It is important to say that there are other few initiatives in MDS for Portuguese (e.g., the work of Silveira and Branco [25], that introduces the SIMBA system, which applies clustering and text simplification for summarization), but that use different evaluation setups and make any direct comparison unfair.

2.2 Subtopic Segmentation and Clustering

There are several approaches for subtopic segmentation, for written and spoken language, using different features and techniques. We focus here on some of the main and most used ones for written language, since this is the case of this paper.

One well-known and heavily used approach for subtopic segmentation is TextTiling [2], which is based on lexical cohesion. In its strategy, it is assumed that a set of lexical items is used during the development of a subtopic in a text, and, when that subtopic changes, a significant proportion of vocabulary also changes. For identifying major subtopic shifts, adjacent text passages of a pre-defined size (blocks) are compared for overall similarity. The more words these blocks have in common, the higher the chance that they address the same subtopic. Subtopic boundaries are established in points in the text that show representative lexical gaps.

Choi [26] developed the algorithm called C99, also based on lexical cohesion. Starting from preprocessed sentences, C99 initially calculates the similarity between each pair of sentences and produces a similarity matrix. From the matrix, it produces a rank-similarity: the more similar the sentences are with their neighbors, the higher the score will be. The lower ranks in the classification matrix indicate subtopic boundaries.

Riedl and Biemann [27], based on TextTiling, proposed the TopicTiling algorithm that segments documents using the Latent Dirichlet Allocation (LDA) topic model [28]. The documents that are to be segmented have first to be annotated with topic IDs, obtained by the LDA inference method. The topic model must be trained on documents similar in content to the test documents. The IDs are used to calculate the cosine similarity between two adjacent sentence blocks, represented as two vectors, containing the frequency of each topic ID. Values close to 0 indicate marginal relatedness between two adjacent blocks, whereas values close to 1 denote connectivity.

Du et al. [29] presented a hierarchical Bayesian model for unsupervised topic segmentation. The model takes advantage of the high modeling accuracy of structured topic models to produce a topic segmentation based on the distribution of latent topics. The model consists of two steps: modeling topic boundary and modeling topic structure.

Hovy and Lin [30] have used various complementary techniques, including those based on text structure, cue words and high frequency indicative phrases for subtopic identification in a summarization system. They argue that discourse structure might help subtopic identification. Following in this line, Cardoso et al. [31] showed that in fact discourse structure mirrors the subtopic segmentation.

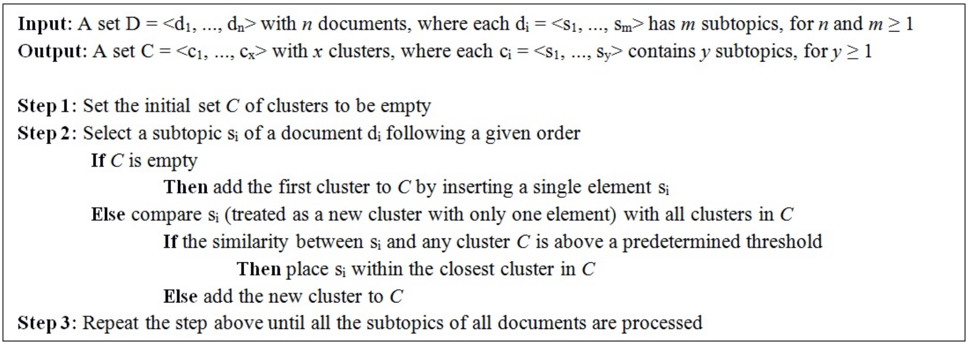
In this paper, we used an adaptation of TextTiling [31] for the characteristics of the texts that we used, which are news texts written in Portuguese (different from the original proposal of TextTiling, which was for expository texts in English). As the authors of its adaptation show, TextTiling is among the best methods for subtopic segmentation of news texts, being robust and scalable, which are goals that we follow in this work. The performance of the adapted version was 77% precision and 40% recall.

Since the subtopic segmentation technique applied by TextTiling is linearly made, i.e., on only one document by time, the following problem may happen: if two subtopics from different texts are found, there is no guarantee that they correspond to distinct information; this may even happen inside the same document if some subtopic is interrupted by another one and then resumed later. Thus, in order to identify similar subtopics within and across documents, we need to cluster the previously segmented subtopics.

Clustering is a concept that arises naturally in many fields, where there are heterogeneous set of objects. It is natural to search for such methods to group/cluster objects based on some measure of similarity. For example, to set the distance between objects, it may be considered that the closer they are, the more similar they are. Thus, clustering is centered around an intuitive, but vague goal: given a set of objects, one may partition them into a collection of clusters in which objects in the same group are close, while objects in different groups are distant from each other.

There are many works in the area. For Portuguese, there is a tool named SiSPI [32], which performs clustering on the basis of the Single-pass algorithm [33] for clustering similar sentences. In this paper, we have simply adopted this solution, adapting it for clustering subtopics with some varied similarity measures. It appears to be a sufficient and suitable solution, allowing us to focus on the summarization method.

In the context of the discovery of related subtopics, the Single-pass algorithm requires a single sequential pass over the set of subtopics to be clustered. It is an incremental clustering algorithm (groups are incrementally created by analyzing all other previously created groups). Figure 4 shows the Single-pass algorithm, already adapted to the case of subtopics.



```
Input: A set $D = <d_1, ..., d_n>$ with $n$ documents, where each $d_i = <s_1, ..., s_m>$ has $m$ subtopics, for $n$ and $m \geq 1$
Output: A set $C = <c_1, ..., c_x>$ with $x$ clusters, where each $c_i = <s_1, ..., s_y>$ contains $y$ subtopics, for $y \geq 1$

Step 1: Set the initial set $C$ of clusters to be empty
Step 2: Select a subtopic $s_i$ of a document $d_i$ following a given order
        If $C$ is empty
                Then add the first cluster to $C$ by inserting a single element $s_i$
        Else compare $s_i$ (treated as a new cluster with only one element) with all clusters in $C$
                If the similarity between $s_i$ and any cluster $C$ is above a predetermined threshold
                        Then place $s_i$ within the closest cluster in $C$
                Else add the new cluster to $C$
Step 3: Repeat the step above until all the subtopics of all documents are processed
```

Figure 4. Single-pass algorithm

Initially, the algorithm creates the first group by selecting the first subtopic of a collection of documents to be clustered. The algorithm iteratively decides whether a newly selected subtopic should be placed in a group already created or in a new one. This decision is made according to the specified similarity function, i.e., a predetermined similarity threshold. In this work, we use the cosine measure, a measure of lexical similarity [34], and also the occurrence of discourse relations among the subtopics (which we explain later in this paper). The higher the similarity value between two subtopics, the more similar they are. The threshold is chosen based on the calculation of the average similarity among all groups. We tested other similarity thresholds as well, and the average similarity produced good results and showed to be adaptable to different text types and genres.

2.3 Cross-document Structure Theory

The Cross-document Structure Theory (CST) [21] is used to describe discourse connections among topically related texts in any domain. The author proposed 24 CST relations, however, in this study we consider only 14 relations found in the corpus we use in our evaluation (see Section 4). Such relations are organized in a typology defined in [35][36]. This typology is shown in Figure 5.

The typology classifies CST relationships into two major groups: content and presentation/form. The content category refers to relations that indicate similarities and differences among contents in the texts. This category is divided into three subcategories: redundancy, complement, and contradiction. Redundancy includes relations that express a total or partial similarity among sentences. Complement relations link textual segments that elaborate, give continuity or background to some other information. The last subcategory for the content category only includes Contradiction. In the form category, all the relations that deal with

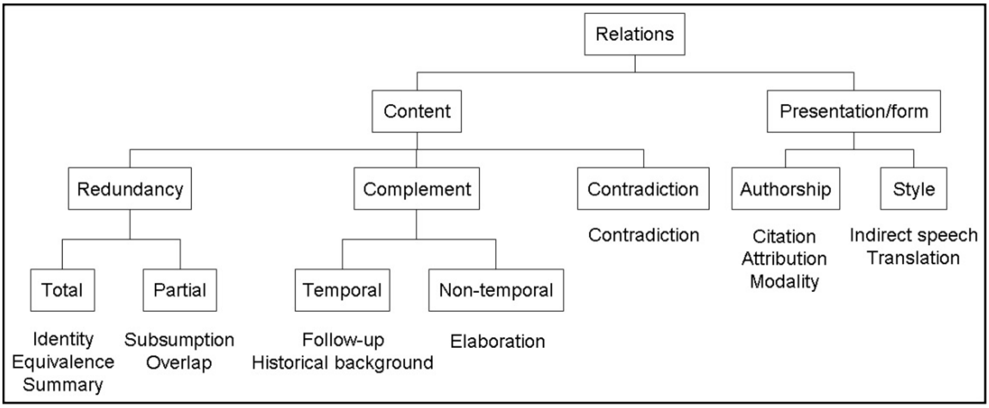


Figure 5. Typology of CST relations

superficial aspects of information are included.

Based on the meaning of the relationships defined in [35, 36], these may assist in the clustering phase, in order to obtain a better selection of similar subtopics. This is due to the fact that it was found in previous work that the discourse relations are closely related to the lexical similarity of textual segments, i.e., the closer such segments are, the more CST relationships they have with each other and, therefore, the greater the chance is that the segments belong to the same subtopic. Therefore, this model was used to investigate the issue of the relationship among subtopics.

The CST annotation of texts may be manually or automatically done. Manual annotation requires trained humans, making the process expensive and very time consuming. The automatic annotation, in turn, is performed by discourse parsers, which are softwares that automatically detect relationships among segments of texts and build graphs with the resulting annotation. For Portuguese, it is available a discourse parser for CST called CSTParser [35, 36], with a general accuracy of 68%. In this paper, we have used the manual annotation already available in our corpus, but, if one desires, the method may be scalable to new texts by using the available discourse parser.

In what follows, using the concepts and methods introduced before, we report our investigation and adaptation of the Salton et al. method under focus for MDS in Portuguese.

3 The Summarization Algorithm

The multi-document summarization method that we investigate in this paper - the Segmented Bushy Path - may be organized in a few steps. Firstly, the algorithm preprocesses the source texts and computes the lexical similarity among their sentences to build the map/graph. Then, it divides each text into subtopics using TextTiling (the adapted version to Portuguese news texts [31]). Once the texts are segmented, the next step is to identify and cluster common subtopics within and across the documents. With the resulting clusters, the relationship map is finally complete. The Segmented Bushy Path method may then be used to select the relevant information for the summary, performing the content selection. While the important information is being selected, the redundancy treatment is applied. In this way, it is guaranteed that the final summary does not have repeated information.

As mentioned before, the graph traversing was proposed by Salton et al.[3]. It is important to note that (1) we focus our investigation on the Segmented Bushy Path, due to the completion of the other two strategies in a previous effort [12], and (2) the chosen method was originally developed for the single document scenario, therefore, adaptations were made, in order that relevant information could be identified not only in one, but in several documents. We detail the main parts of the summarization method in the next subsections.

3.1 Preprocessing and Graph Building

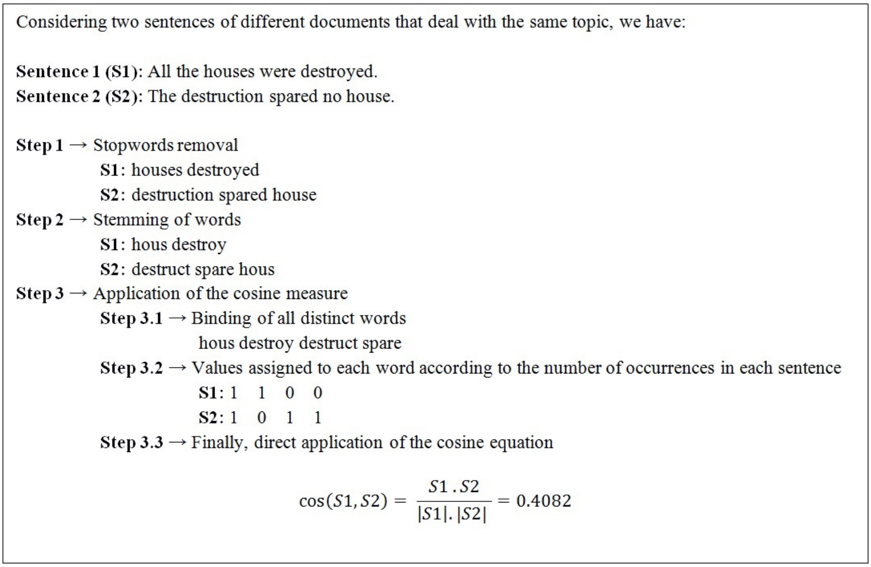
We often encounter similar words in the source texts, but in different forms, e.g., house and houses. Therefore, a treatment for these types of words, which may be done by normalization (either by lemmatization or stemming [37]), is required. This treatment aims to standardize the words of the sentences and make the necessary computation more meaningful. As in other approaches, the stopwords are also eliminated using a pre-defined list for Portuguese language. Their removal is essential for keeping only the relevant words that carry the main content.

The preprocessed sentences in the texts are then used to build a graph, where the vertices are the sentences and links have numeric values that indicate how lexically close the sentences are (using the cosine measure).

Figure 6 shows an example of the performed preprocessing steps for two sentences and the computation of the cosine measure for them.

3.2 Subtopic Segmentation and Clustering

After the preprocessing, the texts are segmented in subtopics using TextTiling, as illustrated in Figure 7. The horizontal lines indicate where to hypothetically segment each



Considering two sentences of different documents that deal with the same topic, we have:

Sentence 1 (S1): All the houses were destroyed.
Sentence 2 (S2): The destruction spared no house.

Step 1 → Stopwords removal
S1: houses destroyed
S2: destruction spared house
Step 2 → Stemming of words
S1: hous destroy
S2: destruct spare hous
Step 3 → Application of the cosine measure
Step 3.1 → Binding of all distinct words
hous destroy destruct spare
Step 3.2 → Values assigned to each word according to the number of occurrences in each sentence
S1: 1 1 0 0
S2: 1 0 1 1
Step 3.3 → Finally, direct application of the cosine equation

$$\cos(S1, S2) = \frac{S1 \, . \, S2}{|S1| . |S2|} = 0.4082$$

Figure 6. Preprocessing step and similarity computation

document.

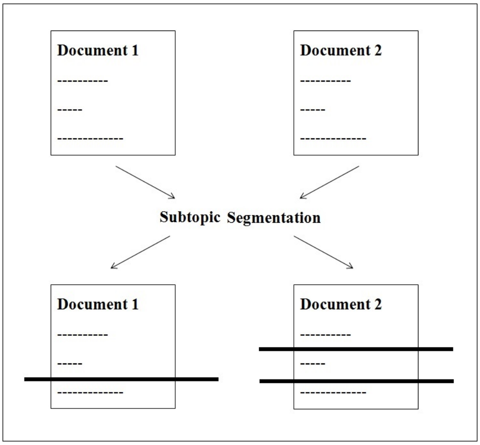


Figure 7. Subtopic segmentation by TextTiling

Since this technique was linearly applied for each document, as mentioned earlier, there is no correlation of the subtopics. Thus, it is necessary to perform subtopic clustering. For this, we have used the Single-pass algorithm [33] cited before. Figure 8 illustrates a possible (hypothetical) result for the clustering.

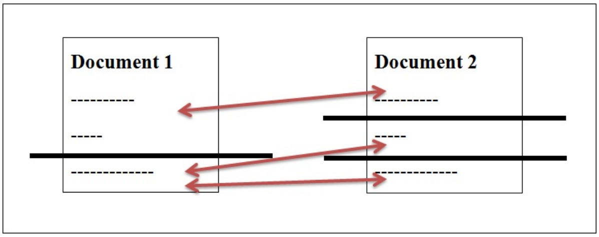


Figure 8. Clustering of subtopics

At this stage of subtopic correlation, 4 strategies for finding similar subtopics were considered: (1) Keywords - the most frequent words are discovered; then the cosine similarity is applied between the subtopics by analyzing only such words; (2) Subtopic similarity - all the words in the pair of subtopics are analyzed to determine whether they belong to the same cluster, using the cosine measure; (3) Unweighted CST - use of the number of CST relations among subtopics to investigate their correlation, i.e., the greater the number of connections between two subtopics, the greater the chance is of the subtopics being correlated; and (4) Weighted CST - use of numeric values for each CST relation between subtopics. These numerical values, in the range from 0 to 1, correspond to the level of similarity between each pair of subtopics. An Identity relation, for example, corresponds to the value 1 of similarity (since both segments are the same). For the Overlap relation, there is a value of 0.5 because of its aspect of partial redundancy: there may be a lot of information in common between sentences, but there may be also little redundancy. Thus, the greater the sum of the values of CST relationships between a pair of subtopics, the greater the chance is of grouping them.

As an illustration, Figure 9 shows the correlation of subtopics of two documents, after applying one of the above similarity techniques. It is a representation of subtopics in Figure 1. Sentences are indicated by *s* and subtopics by *sb*. Since sentences 1-2 of Document 1 and sentences 1-2 of Document 2 describe the same subtopic (Maradona's relapse), they must be grouped in a single cluster, identified by *sb1*. Sentences 3 and 5 of Document 1 are connected with sentences 3-5 of Document 2, forming subtopic *sb2*, since they describe details about Maradona's hepatitis. On the other hand, sentences 6-8 of Document 2 are not connected with any sentence of Document 1; in this case, they form a subtopic, identified by *sb4*. The same happens to sentence 4 of document 1. After the clustering, one may see that the set of texts has 4 subtopics.

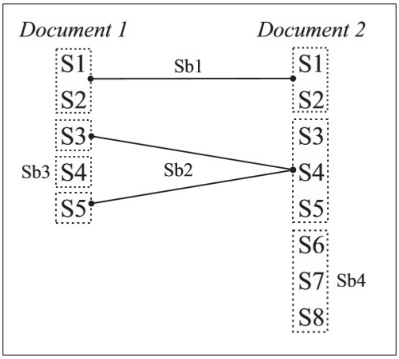


Figure 9. Correlation of subtopics

3.3 Content Selection for the Summary

After preprocessing, graph building, and subtopic segmentation and clustering, we have the content selection step, in which sentences are selected to compose the summary.

For a better overview of the method, Figure 10 shows the modeling of the source texts in a graph with the subtopic segmentation and clustering already performed. It is important to notice a few things about Figure 10: 1) the Si-Dj format indicates the location of a sentence (for instance, S1-D2 refers to the first sentence of the second document); 2) the lines between sentences must also indicate the cosine similarity they have; and 3) the strings above the segments indicate the keywords of the subtopics that they belong to (for illustrative purposes only, since our method do not need such topic signature words).

Having the graph, the application of Segmented Bushy Path is performed. As mentioned before, this path builds the final summary by selecting the most important sentences in each subtopic. For this, we choose the most connected sentence within a particular subtopic. For each new subtopic, it is also important to include a transitional sentence. This transitional sentence is always picked in a way that the chronology is maintained, i.e., the transitional sentence must come before the sentence of the next subtopic, being the one with the highest lexical similarity with the sentence of this new subtopic. This process of selecting sentences for each subtopic and then selecting transitional sentences occurs until the desired compression rate is reached.

It is noticeable that, using this method, many edges may possibly indicate a very high degree of similarity among the sentences (due to redundancy among texts). Therefore, it is necessary to calculate the limit of redundancy that two sentences may have with each other, establishing a threshold that enable to prune redundant sentences that might eventually be in-

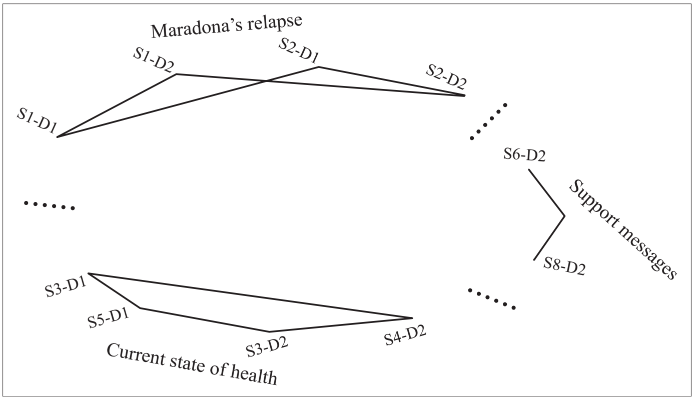


Figure 10. Relationship map

cluded in the summary. In other words, if a node (sentence) has a greater value of similarity (relative to the nodes that were already selected for the summary) than the established threshold, the sentence is not taken to the summary, as it is considered redundant. In this work, the threshold is computed as the average of the similarity values between every 2 sentences of every pair of documents. This way of establishing the threshold (instead of using a fixed value for any group of texts) allows the summarization method to be more generic and applicable to different types and collections of documents. We have also tested other threshold values and have empirically checked that the adopted strategy is a good solution.

Finally, the construction of the final summary should take into account its size, and, for this, we use a given compression rate (mentioned earlier), which limits the amount of information (counted in number of words, as usually done in the works in the area) that the summary will contain. We may notice that, just by applying the compression rate, relevant sentences may be left out of the summary in favor of transitional sentences. We opted to keep such an approach to, once again, preserve the summary coherence.

4 Experiments and Evaluation Results

In our experiments, we used the CSTNews corpus [38] for evaluating the MDS method ans also the subtopic segmentation and clustering steps. The CSTNews corpus is composed of 50 groups of news articles written in Brazilian Portuguese, collected from several sections (Politics, Sports, World, Daily News, Money, and Science) of mainstream online news agencies (Folha de São Paulo, Estadão, O Globo, Jornal do Brasil, and Gazeta do Povo). Specifically, each group contains 2 or 3 source texts on the same topic from different agencies (in a total of 140 documents) and a multi-document human abstract for each group.

The corpus is a benchmark for researches on MDS for Portuguese and has been used in several studies (e.g., in all the MDS researches cited in the related work section). It is freely available and has several linguistic annotation layers that were manually produced in a systematic and reliable way. For this paper, we have special interest on the following: all the texts in the corpus were manually annotated with subtopic boundaries and their clustering, with satisfactory annotation agreement values (more details may be found in [39]); the texts present their CST manual annotation with satisfactory agreement values (as described in [38, 40]), which we tested in the context of subtopic clustering. For such reasons, we haved adopted this corpus instead of others frequently used in the area (as the corpora of the Document Understand Conferences).

The size of the human summaries in the corpus corresponds to 30% of the size of the longest text in each group (considering that the size is given in terms of number of words), resulting in a compression rate of 70%, therefore. This is the compression rate we use for producing the automatic summaries, in order to the comparison with the human summaries to be fair.

As we adopted a ready-to-use solution for performing subtopic segmentation (TextTiling), we start the evaluation by the clustering techniques we employed over the automatically identified subtopics. The quality of the clustering method may be assessed by external measures of quality that indicate how close the automatically produced clusters are in relation to the reference clusters (i.e., the clusters produced by humans in the corpus). For this evaluation, measures of precision (see Eq. 1), coverage (see Eq. 2) and f-measure (see Eq. 3) [41, 42] were used. Precision indicates the proportion of correct segments there are inside each cluster; coverage shows the proportion of correct segments there are in each cluster in relation to what was predicted in the reference clusters; f-measure is a unique performance measure, combining precision and coverage values.

$$P(k_i, c_j) = \frac{n_{ij}}{|c_j|} \tag{1}$$

$$C(k_i, c_j) = \frac{n_{ij}}{|k_j|} \tag{2}$$

$$F(k_i, c_j) = \frac{2 * C(k_i, c_j) * P(k_i, c_j)}{C(k_i, c_j) + P(k_i, c_j)} \tag{3}$$

In Equations 1-3, k_i indicates each reference cluster; c_j indicates each one of the clusters that are automatically formed; and n_{ij} refers to the number of segments of the cluster k_i that are present in c_j. The f-measure for each cluster of the entire data set is based on the automatic cluster that best describes each reference cluster. Thus, the overall value of f-measure may be denoted by Equation 4.

$$Overall\ F - Measure = \sum\nolimits_{k_i \in K} \frac{|k_i|\ max\ c_j \in C\{F(k_i, c_j)\}}{N} \tag{4}$$

In this formula, N refers to the total number of segments to be grouped; K is the set of reference clusters; and C is the set of automatic clusters.

The four clustering methods (Keywords, Subtopical Similarity, Unweighted CST and Weighted CST) were evaluated against the reference clusters and the results are presented on Table 1, using the Equations 1, 2, 3 and 4.

Table 1. Clustering results over automatic subtopic segmentation

Clustering Technique	Precision	Recall	F-Measure
Keywords	0.7265	0.6374	0.6790
Subtopical Similarity	0.5823	0.4165	0.4856
Unweighted CST	0.5514	0.3829	0.4519
Weighted CST	0.5490	0.3777	0.4475

From Table 1, we may conclude that the best technique, taking into account the automatic method of subtopic segmentation (TextTiling), was the one that simply considered the most frequent words in each subtopic for computing the cosine measure: the Keywords clustering method. This possibly happens because the technique considers only the most relevant words of a subtopic, so clustering is more accurate. It is also interesting to notice that all techniques had better precision than recall.

We went one step further and also evaluated the impact of using the reference (manual) segmentation of the corpus instead of the one automatically performed by TextTiling. The clustering results over the reference segmentation are shown in Table 2. As it may be seen, the values surpass those achieved by the automatic subtopic segmentation. The overrun was expected since we deal this time with better data. In addition, it may be noticed that the values obtained by Keywords are the highest among all the methods.

Table 3 presents how much better the clustering results achieved with the reference segmentation are in relation to the ones obtained with the automatic segmentation. For instance, one may see that the f-measure for the Keywords clustering over the reference segmentation is 6.44% better than the f-measure for the same clustering method over the automatic segmentation. In general, we may conclude that the results for the automatic methods are not far from those obtained for reference data, which favors their use in actual fully automatic systems.

Table 2. Clustering results over reference subtopic segmentation

Clustering Technique	Precision	Recall	F-Measure
Keywords	0.7586	0.6901	0.7227
Subtopical Similarity	0.6184	0.4300	0.5072
Unweighted CST	0.5680	0.4241	0.4850
Weighted CST	0.5324	0.4352	0.4789

Table 3. Improvement of clustering results - reference over automatic segmentation

Clustering Technique	Precision	Recall	F-Measure
Keywords	0.0442	0.0827	0.0644
Subtopical Similarity	0.0620	0.0324	0.0445
Unweighted CST	0.0301	0.1076	0.0732
Weighted CST	0.0302	0.1522	0.0702

The last and most important step of the evaluation phase corresponds to the analysis of the automatic summaries over the manual ones, which relied on the use of ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [43], which is a suite of metrics for this purpose.

Basically, ROUGE was created to enable a direct comparison between an automatically generated summary and its human references. ROUGE calculates a score among 0 and 1 based on sets of words (e.g., the n-grams that may vary from 1 to 4) in common between human summaries and automatically generated summaries, producing precision, recall/coverage, and f-measure values. As already mentioned before, precision indicates the proportion of reference n-grams in the automatic summary; recall indicates the proportion of reference n-grams in the automatic summary in relation to the reference summary; f-measure is a unique measure of performance, combining precision and recall.

It is known that ROUGE evaluates the informativeness of a summary, i.e., the amount of information the summary contains, and its scores have been shown to correlate well with human judgment. As it may be automatically and fastly computed, and achieves reliable results, it has become a mandatory evaluation metric and almost all the works in the area adopt it for evaluation. In this work, following what most of the works do, we consider ROUGE evaluation for n-grams of size 1 only (which we refer to by ROUGE-1), which is considered enough for distinguishing summary informativeness.

We initially used ROUGE to evaluate the versions of Segmented Bushy Path method that use the 4 clustering techniques previously tested, indicated by the numbers that follow the method name (1: clustering with all the words in the subtopics; 2: clustering with keywords only; 3: clustering using only the number of CST relations; 4: clustering using the weights of the CST relations) and a version that uses the manually produced (reference) clusters in the corpus. We also included in the evaluation a random baseline, that randomly select sentences to compose the summary, for comparison purposes only. Table 4 shows the obtained results.

Table 4. ROUGE-1 results for segmented bushy path variations

Methods	**Precision**	**Recall**	**F-Measure**
Segmented Bushy Path^{Manual}	0.5803	0.3918	0.4407
Segmented Bushy Path^{1}	0.5472	0.3517	0.4190
Segmented Bushy Path^{2}	0.5507	0.3297	0.4023
Segmented Bushy Path^{3}	0.6079	0.2802	0.3571
Segmented Bushy Path^{4}	0.6033	0.2879	0.3637
Baseline	0.3015	0.2900	0.2948

As expected, the version with the reference clusters achieved the best results. Interestingly, looking at f-measure values for the automatic clustering versions, one may see that the best clustering technique (using the keywords) did not result in the best version of the summarization method. The best summarization method was the one that used all the subtopic words for clustering. This is not a surprise and probably happens due to variation in the summary content and due to the known fact in the area that several good summaries exist for the same group of texts. One may also see that all the method variations outperformed the baseline method.

In Table 5, there is a comparison with other summarizers for the Portuguese language, computed over the same corpus and with the same evaluation setup. The methods are ordered in the table by the f-measure values. The results correspond to the RSumm summarizers [12], RCT-4 [22], CSTSumm [8], MEAD [24], GistSumm [7] and RCSumm [20], making it possible to know the quality of the summarization method investigated in this paper in relation to other summarizers to Portuguese. The cited summarizers are the ones that were already introduced in the related work section. In addition, we included in the comparison existing versions of GistSumm and MEAD that were enriched with CST information (which help improving sentence ranking for performing sentence selection to compose the summary), which produce better results than their original versions.

Overall, one may see that the results were very satisfactory, with the investigated method overcoming not only most of the summarizers, but also the other two paths proposed in [3]. Our method only lost for the RCT-4 method, which is a discourse-based method, and, therefore, very expensive and not easily scalable.

The lower recall of our method in relation to RCT-4 may be explained by the fact that the priority was to select at least one sentence of each subtopic, and a transition sentence for new subtopics. Thus, compression rate was often hit up after including some transition

Table 5. Comparison among systems for Portuguese

Methods	Precision	Recall	F-Measure
RCT-4	0.4520	0.4416	0.4445
Segmented Bushy Path^1	0.5472	0.3517	0.4190
RCSumm	0.4218	0.4036	0.4102
Segmented Bushy Path^2	0.5507	0.3297	0.4023
MEAD with CST	0.4257	0.3876	0.4018
RSumm (Bushy Path)	0.4089	0.3704	0.3871
CSTSumm	0.4472	0.3557	0.3864
RSumm (Depth-First Path)	0.3977	0.3630	0.3795
Segmented Bushy Path^4	0.6033	0.2879	0.3637
MEAD	0.3691	0.3574	0.3616
GistSumm with CST	0.2800	0.5229	0.3583
GistSumm	0.3923	0.3343	0.3581
Segmented Bushy Path^3	0.6079	0.2802	0.3571
Baseline	0.3015	0.2900	0.2948

sentences, leaving out other important information (sentences of other subtopics) from the summary.

Another interesting measure we tested (but do not show here) is ROUGE-2. Its values for f-measure showed a different rank, in which the Segmented Bushy Path^1 obtained a higher f-measure value than the RCT-4 system (0.3434 vs. 0.2615, respectively), outperforming all the systems. Therefore, it is possible to claim that our method is competitive with the state of the art system, but at a lower cost, since it is a superficial method.

We have run t-tests for pairs of summarizers for which it was important to show that the differences in results were significant. The following pairs of comparisons were carried out: Segmented Bushy Path^1 RCT-4, Segmented Bushy Path^2 RCT-4, Segmented Bushy Path^1 RCSumm, Segmented Bushy Path^2 RCSumm, and Segmented Bushy Path^1 Segmented Bushy Path^2. The results indicated that there is statistical difference with 95% confidence.

As illustration, Figure 11 shows the automatic summary produced by the Segmented Bushy Path^1 method for the texts previously presented in the introductory section (Figure 1). It may be noticed that not only Maradona's relapse (first subtopic) was reported, but also Maradona's hepatitis (second subtopic). The third and forth subtopics (Current State of Heath and Support Messages, respectively) are not included due to the choice of the compression rate (70%).

Maradona had been discharged 11 days ago, but he was again hospitalized on Friday, and the medical reports did not specify what was wrong with the ex-player – Cahe ruled out ulcer or pancreatitis.
Cahe said that Maradona had not started to drink alcoholic beverages again, and that the causes of the relapse are being investigated.

Figure 11. Example of automatic summary produced by the investigated method

Undoubtedly, there are some issues regarding the completeness and coherence of the generated summary, e.g., Cahe is an entity that lacks a clear reference: a reader unfamiliar with the events reported in the source texts is not able to determine that Cahe is Marodona's doctor. Still, the Segmented Bushy Path [3] proves to be effective when subtopic segmentation and clustering techniques are performed together to obtain a summary capable of comprising the most relevant subtopics in a group of texts.

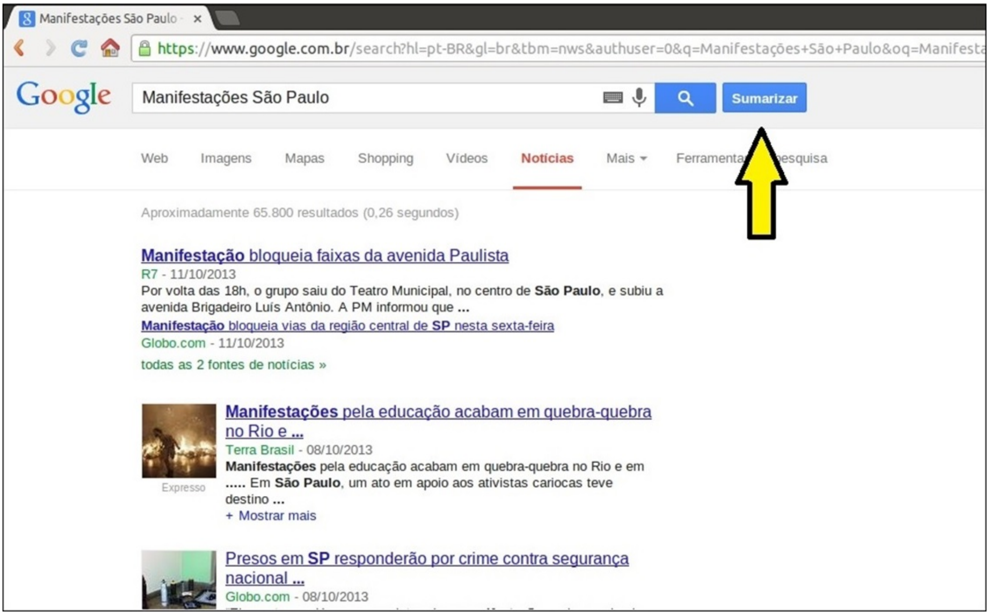
5 Google Chrome Extension

Given the need for a tool that handles the large amount of online information, mainly in the current multi-document scenario, in this work we developed an extension for Google Chrome browser, which summarizes the documents returned from a search with Google News website. The extension makes use of the following tools: 1) the Application Programming Interface of Google News (Google News API) for retrieval of documents; 2) NCleaner tool [44], trained for Portuguese, for removing the irrelevant content in the web pages (advertisements and links to other pages, for example); and 3) the summarization methods of Salton et al. [3].

Figure 12 shows the active extension with the search for the term “Manifestações São Paulo" (in English, “Protests São Paulo", regarding the public acts that happened in the city of São Paulo, in Brazil) in Google News website, followed by Figure 13, where the return of the search was summarized (in this case, the first eight most relevant texts were processed).

It may be noted that, in Figure 12, for the search results in Google News site, the “Sumarizar" (in English, “Summarize") button appears at the top right corner, giving the user the option to summarize the retrieved texts. It is important to explain that the system is currently customized for the Portuguese language, since the stoplist and stemmer are used for this language. However, this customization may be easily made to other languages too, once such resources are available.

Figure 13 shows a summary with a compression rate of 70% on the search conducted earlier. It may be noted that, on the left, there are the links related to the search, as well as links in each sentence of the summary (in the central panel) that connects to their corresponding source texts. Notice too that the users may add more texts to the process as well as increase or decrease the compression rate.



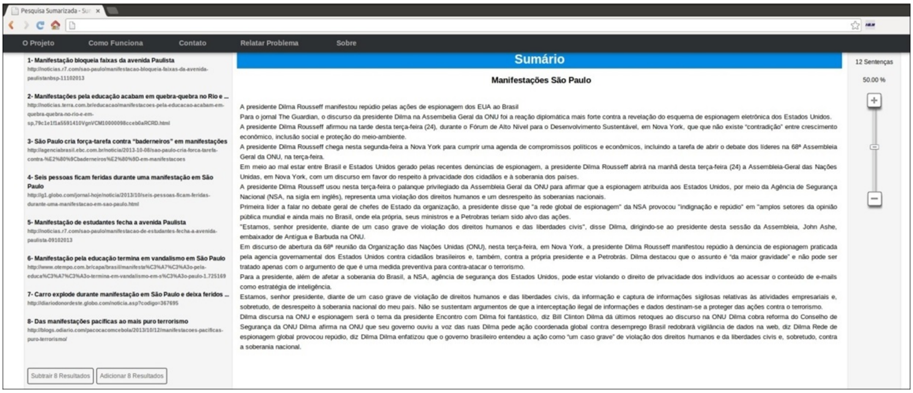
Aproximadamente 65.800 resultados (0,26 segundos)

Manifestação bloqueia faixas da avenida Paulista
R7 - 11/10/2013
Por volta das 18h, o grupo saiu do Teatro Municipal, no centro de **São Paulo**, e subiu a avenida Brigadeiro Luís Antônio. A PM informou que ...
Manifestação bloqueia vias da região central de **SP** nesta sexta-feira
Globo.com - 11/10/2013
todas as 2 fontes de notícias »

Manifestações pela educação acabam em quebra-quebra no Rio e ...
Terra Brasil - 08/10/2013
Manifestações pela educação acabam em quebra-quebra no Rio e em ..... Em **São Paulo**, um ato em apoio aos ativistas cariocas teve destino ...
+ Mostrar mais
Expresso

Presos em **SP** responderão por crime contra segurança nacional ...
Globo.com - 08/10/2013

Figure 12. Google Chrome extension



1- Manifestação bloqueia faixas da avenida Paulista
http://noticias.r7.com/sao-paulo/manifestacao-bloqueia-faixas-da-avenida-paulistanbsp-11102013

2- Manifestações pela educação acabam em quebra-quebra no Rio e ...
http://noticias.terra.com.br/educacao/manifestacoes-pela-educacao-acabam-em-quebra-quebra-no-rio-e-em-sp,79c1e1f1a5591410VgnVCM10000098cceb0aRCRD.html

3- São Paulo cria força-tarefa contra “baderneiros” em manifestações
http://agenciabrasil.ebc.com.br/noticia/2013-10-08/sao-paulo-cria-forca-tarefa-contra-%E2%80%9Cbaderneiros%E2%80%9D-em-manifestacoes

4- Seis pessoas ficam feridas durante uma manifestação em São Paulo
http://g1.globo.com/jornal-hoje/noticia/2013/10/seis-pessoas-ficam-feridas-durante-uma-manifestacao-em-sao-paulo.html

5- Manifestação de estudantes fecha a avenida Paulista
http://noticias.r7.com/sao-paulo/manifestacao-de-estudantes-fecha-a-avenida-paulista-09102013

6- Manifestação pela educação termina em vandalismo em São Paulo
http://www.otempo.com.br/capa/brasil/manifesta%C3%A7%C3%A3o-pela-educa%C3%A7%C3%A3o-termina-em-vandalismo-em-s%C3%A3o-paulo-1.725169

7- Carro explode durante manifestação em São Paulo e deixa feridos ...
http://diariodonordeste.globo.com/noticia.asp?codigo=367695

8- Das manifestações pacíficas ao mais puro terrorismo
http://blogs.odiario.com/pacocacomcebola/2013/10/12/manifestacoes-pacificas-puro-terrorismo/

Manifestações São Paulo

A presidente Dilma Rousseff manifestou repúdio pelas ações de espionagem dos EUA ao Brasil
Para o jornal The Guardian, o discurso da presidente Dilma na Assembelia Geral da ONU foi a reação diplomática mais forte contra a revelação do esquema de espionagem eletrônica dos Estados Unidos.
A presidente Dilma Rousseff afirmou na tarde desta terça-feira (24), durante o Fórum de Alto Nível para o Desenvolvimento Sustentável, em Nova York, que que não existe "contradição" entre crescimento econômico, inclusão social e proteção do meio-ambiente.
A presidente Dilma Rousseff chega nesta segunda-feira a Nova York para cumprir uma agenda de compromissos políticos e econômicos, incluindo a tarefa de abrir o debate dos líderes na 68ª Assembleia Geral da ONU, na terça-feira.
Em meio ao mal estar entre Brasil e Estados Unidos gerado pelas recentes denúncias de espionagem, a presidente Dilma Rousseff abrirá na manhã desta terça-feira (24) a Assembleia-Geral das Nações Unidas, em Nova York, com um discurso em favor do respeito à privacidade dos cidadãos e à soberania dos países.
A presidente Dilma Rousseff usou nesta terça-feira o palanque privilegiado da Assembleia Geral da ONU para afirmar que a espionagem atribuída aos Estados Unidos, por meio da Agência de Segurança Nacional (NSA, na sigla em inglês), representa uma violação dos direitos humanos e um desrespeito às soberanias nacionais.
Primeira líder a falar no debate geral de chefes de Estado da organização, a presidente disse que "a rede global de espionagem" da NSA provocou "indignação e repúdio" em "amplos setores da opinião pública mundial e ainda mais no Brasil, onde ela própria, seus ministros e a Petrobras teriam sido alvo das ações.
"Estamos, senhor presidente, diante de um caso grave de violação dos direitos humanos e das liberdades civis", disse Dilma, dirigindo-se ao presidente desta sessão da Assembleia, John Ashe, embaixador de Antígua e Barbuda na ONU.
Em discurso de abertura da 68ª reunião da Organização das Nações Unidas (ONU), nesta terça-feira, em Nova York, a presidente Dilma Rousseff manifestou repúdio à denúncia de espionagem praticada pela agencia governamental dos Estados Unidos contra cidadãos brasileiros e, também, contra a própria presidente e a Petrobrás. Dilma destacou que o assunto é "da maior gravidade" e não pode ser tratado apenas com o argumento de que é uma medida preventiva para contra-atacar o terrorismo.
Para a presidente, além de afetar a soberania do Brasil, a NSA, agência de segurança dos Estados Unidos, pode estar violando o direito de privacidade dos indivíduos ao acessar o conteúdo de e-mails como estratégia de inteligência.
Estamos, senhor presidente, diante de um caso grave de violação de direitos humanos e das liberdades civis, da informação e captura de informações sigilosas relativas às atividades empresariais e, sobretudo, de desrespeito à soberania nacional do meu país. Não se sustentam argumentos de que a interceptação ilegal de informações e dados destinam-se a proteger das ações contra o terrorismo.
Dilma discursa na ONU e espionagem será o tema da presidente Encontro com Dilma foi fantástico, diz Bill Clinton Dilma dá últimos retoques ao discurso na ONU Dilma cobra reforma do Conselho de Segurança da ONU Dilma afirma na ONU que seu governo ouviu a voz das ruas Dilma pede ação coordenada global contra desemprego Brasil redobrará vigilância de dados na web, diz Dilma Rede de espionagem global provocou repúdio, diz Dilma Dilma enfatizou que o governo brasileiro entendeu a ação como "um caso grave" de violação dos direitos humanos e da liberdades civis e, sobretudo, contra a soberania nacional.

Figure 13. Generated summary in the online interface

6 Final Remarks

We presented a new multi-document summarizer with the Segmented Bushy Path method adapted from Salton et al. [3], which may summarize news articles. This summarizer not only uses Salton et al.'s technique of selecting the most relevant information, but also explores techniques of subtopic segmentation and clustering. The results show that the method is competitive with the best systems for Portuguese. In fact, since it is a superficial method, it is more scalable and capable of on-line integration than the current best available systems.

There is still room for improvements, as dealing with dangling anaphors and producing update summaries. To handle anaphors, or, in more general terms, co-references, is a very difficult task in Natural Language Processing area and is still a problem without robust solution. If we consider the Portuguese language, current systems still present severe limitations. Update summaries, on the other side, appear to be a more reachable and promising line for future work. Considering that possible users of our summarization system aims at browsing and summarizing news on the web, it makes sense to present to these users only the most relevant unknown information, filtering out the information that the users may already know. There are several update summarization techniques proposed in the literature that might be combined with our summarization strategy at relative low cost.

Finally, it is important to say that most of the tools and resources cited in this paper are publicly available. They may be found at the webpage of SUCINTO project [4], which aims at investigating and developing summarization strategies and the associated linguistic-computational resources, tools and applications, mainly for the Portuguese language.

7 Acknowledgments

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

Contribuição dos autores:

Os autores contribuíram igualmente para o desenvolvimento da pesquisa relatada e para a escrita do artigo.

[4] http://www.icmc.usp.br/pessoas/taspardo/sucinto

References

[1] I. Mani, Automatic Summarization, John Benjamins Publishing, 2001.

[2] M. A. Hearst, TextTiling: Segmenting text into multi-paragraph subtopic passages, Computational Linguistics 23 (1) (1997) 33–64.

[3] G. Salton, A. Singhal, M. Mitra, C. Buckley, Automatic Text Structuring and Summarization, Information Processing & Management 33 (2) (1997) 193–207.

[4] E. Hovy, Text Summarization, in: R. A. Lewin (Ed.), The Oxford Handbook of Computational Linguistics, Oxford University Press, The United States, 2009, pp. 583–598.

[5] L. Hennig, Topic-based multi-document summarization with Probabilistic Latent Semantic Analysis, in: the Recent Advances in Natural Language Processing, 2009, pp. 144–149.

[6] X. Xu, A new sub-topics clustering method based on semi-supervised learning, Journal of Computers 7 (10) (2012) 2471–2478.

[7] T. A. S. Pardo, GistSumm-GIST SUMMarizer: Extensões e novas funcionalidades, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2005.

[8] M. L. R. Castro Jorge, T. A. S. Pardo, Experiments with CST-based multidocument summarization, in: the Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, 2010, pp. 74–82.

[9] X. Wan, An Exploration of Document Impact on Graph-based Multi-document Summarization, in: the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 755–762.

[10] S. Harabagiu, F. Lacatusu, Topic themes for multi-document summarization, in: Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2005, pp. 202–209.

[11] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: the Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval, ACM, 2008, pp. 299–306.

[12] R. Ribaldo, A. T. Akabane, L. H. M. Rino, T. A. S. Pardo, Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information, in: the Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243), 2012, pp. 260–271.

[13] R. Mihalcea, D. Radev, Graph-based Natural Language Processing and Information Retrieval, Cambridge University Press, 2011.

[14] I. Mani, E. Bloedorn, Summarizing Similarities and Differences Among Related Documents, Information Retrieval 1 (1-2) (1999) 35–67.

[15] G. Erkan, D. R. Radev, LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research (JAIR) 22 (1) (2004) 457–479.

[16] R. Mihalcea, P. Tarau, A Language Independent Algorithm for Single and Multiple Document Summarization, in: the Proceedings of the 2nd International Joint Conference on Natural Language Processing, 2005.

[17] L. Antiqueira, Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos, Ph.D. thesis, Universidade de São Paulo (2007).

[18] L. Antiqueira, O. N. Oliveira Jr, L. d. F. Costa, M. G. V. Nunes, A complex network approach to text summarization, Information Sciences 179 (5) (2009) 584–599.

[19] D. S. Leite, Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizado de máquina para sumarização automática de textos em Português, Master's thesis, Departamento de Computação, Universidade Federal de São Carlos. São Carlos/SP, Brazil (2010).

[20] A. T. Akabane, T. A. S. Pardo, L. H. M. Rino, Sumarização multidocumento com base em métricas de redes complexas, Anais do 19o Simpósio Internacional de Iniciação Científica da Universidade de São Paulo-SIICUSP (2011) 1–1.

[21] D. R. Radev, A common theory of information fusion from multiple text sources step one: cross-document structure, in: the Proceedings of the 1st SIGdial workshop on Discourse and Dialogue, Association for Computational Linguistics, 2000, pp. 74–83.

[22] P. C. F. Cardoso, Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo, Ph.D. thesis, Universidade de São Paulo (2014).

[23] W. C. Mann, S. A. Thompson, Rhetorical Structure Theory: A theory of text organization, in: University of Southern California, Information Sciences Institute, no. ISI/RS-87-190, 1987.

[24] D. R. Radev, H. Jing, M. Budzikowska, Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies, in: the Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, Association for Computational Linguistics, 2000, pp. 21–30.

[25] S. B. Silveira, A. Branco, Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries, in: IRI 2012: 14th International Conference on Artificial Intelligence, Las Vegas, USA, 2012.

[26] F. Y. Y. Choi, Advances in domain independent linear text segmentation, in: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000, Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, pp. 26–33.

[27] M. Riedl, C. Biemann, Topictiling: A text segmentation algorithm based on lda, in: Proceedings of ACL 2012 Student Research Workshop, ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 37–42.

[28] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[29] L. Du, W. Buntine, M. Johnson, Topic segmentation with a structured topic model, in: Proceedings of NAACL-HLT, 2013, pp. 190–200.

[30] E. Hovy, C.-Y. Lin, Automated text summarization and the SUMMARIST system, in: the Proceedings of a workshop on held at Baltimore, Maryland, Association for Computational Linguistics, 1998, pp. 197–214.

[31] P. C. F. Cardoso, M. Taboada, T. A. S. Pardo, On the contribution of discourse to topic segmentation, in: the Proceedings of the 14th Annual SIGDial Meeting on Discourse and Dialogue, 2013, pp. 92–96.

[32] E. R. M. Seno, M. G. V. Nunes, Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português, Linguamática 1 (1) (2009) 71–87.

[33] V. Rijsbergen, Information Retrieval, Butterworths, Massachusetts, 1979.

[34] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[35] E. G. Maziero, T. A. S. Pardo, Automatic Identification of Multi-document Relations, the Proceedings of the PROPOR (2012) 1–8.

[36] E. G. Maziero, M. L. R. Castro Jorge, T. A. S. Pardo, Revisiting Cross-document Structure Theory for multi-document discourse parsing, Information Processing & Management 50 (2) (2014) 297–314.

[37] M. Porter, Snowball: A language for stemming algorithms. Available at: http://www.snowball.tartarus.org/texts/introduction.html.

[38] P. C. F. Cardoso, E. G. Maziero, M. L. R. Castro Jorge, E. R. M. Seno, A. Di Felippo, L. H. Rino, M. G. V. Nunes, T. A. S. Pardo, CSTNews - A Discourse-Annotated Corpus for Single and Multi-document Summarization of News Texts in Brazilian Portuguese, in: the Proceedings of the 3rd RST Brazilian Meeting, 2011, pp. 88–105.

[39] P. C. F. Cardoso, M. Taboada, T. A. S. Pardo, Subtopic annotation in a corpus of news texts: Steps towards automatic subtopic segmentation, in: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, 2013, pp. 49–58.

[40] P. Aleixo, T. A. S. Pardo, CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory, Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2008.

[41] M. Steinbach, G. Karypis, V. Kumar, et al., A comparison of document clustering techniques, in: KDD workshop on text mining, Vol. 400, Boston, 2000, pp. 525–526.

[42] B. C. Fung, K. Wang, M. Ester, Hierarchical Document Clustering Using Frequent Itemsets., in: 3rd SIAM International Conference on Data Mining, Vol. 3, SIAM, 2003, pp. 59–70.

[43] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.

[44] S. Evert, A lightweight and efficient tool for cleaning web pages, in: the Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008, Citeseer, 2008.