Information Sciences 179 (2009) 584–599

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A complex network approach to text summarization

Lucas Antiqueira [a,*], Osvaldo N. Oliveira Jr. [a], Luciano da Fontoura Costa [a], Maria das Graças Volpe Nunes [b]

[a] Instituto de Física de São Carlos, Universidade de São Paulo, P.O. Box 369, 13560-970 São Carlos, São Paulo, Brazil
[b] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, P.O. Box 668, 13560-970 São Carlos, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 20 September 2007
Received in revised form 14 September 2008
Accepted 24 October 2008

Keywords:
Automatic summarization
Complex networks
Network measurements
Sentence extraction
Summary informativeness

ABSTRACT

Automatic summarization of texts is now crucial for several information retrieval tasks owing to the huge amount of information available in digital media, which has increased the demand for simple, language-independent extractive summarization strategies. In this paper, we employ concepts and metrics of complex networks to select sentences for an extractive summary. The graph or network representing one piece of text consists of nodes corresponding to sentences, while edges connect sentences that share common meaningful nouns. Because various metrics could be used, we developed a set of 14 summarizers, generically referred to as CN-Summ, employing network concepts such as node degree, length of shortest paths, *d*-rings and *k*-cores. An additional summarizer was created which selects the highest ranked sentences in the 14 systems, as in a voting system. When applied to a corpus of Brazilian Portuguese texts, some CN-Summ versions performed better than summarizers that do not employ deep linguistic knowledge, with results comparable to state-of-the-art summarizers based on expensive linguistic resources. The use of complex networks to represent texts appears therefore as suitable for automatic summarization, consistent with the belief that the metrics of such networks may capture important text features.

1. Introduction

Automatic text summarization is a well-established subfield of natural language processing, which is relevant for a number of scenarios [39,66]. A summary can be seen as a condensed representation of a source text that maintains the important information of its original counterpart [65]. When a summary is constructed by selecting and juxtaposing source pieces, such as sentences, it is called an *extract*. The summary produced with changes in the surface form of the source text, resulting from generalizations or paraphrases, is referred to as an *abstract*. Therefore, humans usually produce abstracts. Systems for generating abstracts require sophisticated resources such as discourse and semantic analyzers to infer the meaning of the source text, as well as language generators to compose the summary [42]. Extractive summarizers, on the other hand, do not require linguistic knowledge to select the most relevant pieces of the source text for an extract. Various approaches exist for extractive summarization, including the use of word frequency [19,38], cue words or phrases [19,52], machine learning [30], lexical chains [7] and sentence compression through syntactical or statistical restrictions [67,70]. Less familiar approaches include the application of psychology models to extractive summarization [24] and the use of a hierarchical fractal-based model for representing and condensing texts [69]. Although extractive summarization can produce texts that have cohesion and coherence problems, many systems have been proven to yield summaries whose informative level is

* Corresponding author. Tel.: +55 16 3373 9882; fax: +55 16 3373 9879.
E-mail addresses: lantiq@gmail.com, lantiq@ursa.ifsc.usp.br (L. Antiqueira).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.10.032

satisfactory. This is particularly true when the extract is used as a component of another system (e.g. in information retrieval [61]) and is not directly used by humans. Nevertheless, despite the efforts to develop high-quality automatic summarizers, improvement is required in a number of issues [66].

A graph, or network, is a representation that may capture text structure in various ways, being therefore suitable for extractive summarization. Network nodes (vertices) might represent words, sentences or paragraphs. Network edges (links) connect nodes according to some task-dependent criteria. For instance, if the sentences are linked when sharing complementary content about the same topic, one could argue that a relevant node for summarization is one that shares various edges with other nodes. Therefore, a network algorithm may be used to assign numerical values (relevance scores or ranks) to nodes and to select a subset of them (i.e. pieces of text) to compose an extract. In this paper, we investigate a graph-based, language-independent approach to extractive text summarization inspired by recent developments in the area of complex networks.

Complex networks have attracted a lot of attention [1,9,48] since the small-world [68] and scale-free [6,22] properties were identified in many real-world networks about 10 years ago. The origins of this field can be traced back to the beginning of graph theory [17], passing through the development of random graph theory [20] and analysis of sociological experiments using graphs [44]. The recent discoveries contributed significantly to elucidate the structure and dynamics of diverse real-world entities such as the Internet, the brain and the natural languages. When represented as graphs, some of these entities contain a few highly connected nodes, called hubs, that coexist with a large number of much less connected nodes. These networks are said to be *scale-free*. Other networks reveal a relatively small distance between each pair of nodes, i.e. on average a small number of nodes separates them. This type of complex network, known as *small-world*, also shows high clustering, i.e. if two nodes are connected to the same node, the probability that they are connected to each other is high. These and other properties can be identified in a network through the computation of some graph measurements, such as the average node degree, the average clustering coefficient and the average length of shortest paths [16]. Another interesting feature of some complex networks is the existence of a community (or modular) structure, which allows a network to be partitioned into well-defined groups of highly interconnected nodes. With an increasing number of systems being treated as networks, the concepts of complex networks have been widely used in many disciplines (e.g. in computer science [26]). For linguistics, for example, word co-occurrence networks have been shown to be small-world and scale-free [23], a theory of language evolution was developed [18], and language development on an individual basis was assessed through network measurements [27]. Natural language processing has also benefited from complex networks, including authorship analysis [5] and quality assessment of high-school essays [4], summaries [53] and translations [3].

Here we address the design of extractive summarizers based on complex networks concepts. Our method uses a simple network of sentences that requires only surface text pre-processing, thus allowing us to assess extracts obtained with no sophisticated linguistic knowledge. Given a network representation of a source text, where each node corresponds to a source sentence and an edge links sentences that share common lemmatized nouns, the proposed method selects a subset of sentences (nodes) to compose an extract by ranking them according to some network measurement. Generic summaries, i.e. neither user-specific nor topic-oriented, were produced for newspaper articles in Brazilian Portuguese using network parameters such as shortest paths, *k*-cores and communities. Experiments to evaluate the informativeness level of the extracts were carried out, and the resulting scores for ROUGE-1 [37] and Precision/Recall [62] were compared with the scores of other summarizers previously evaluated within the same experimental setup.

The remainder of this paper is organized as follows. Section 2 contains a review of related work on network-based extractive summarization. Section 3 introduces the network representation of a source text, as well as the graph measurements chosen to generate extracts. The evaluation setup and comparative results regarding informativeness scores are presented and discussed in Section 4. Section 5 complements the evaluation with a correlation analysis of the proposed systems. Finally, Section 6 brings concluding remarks and prospects for future work.

2. Related work

Several developments of summarization techniques based on graphs are reported in the literature. Salton et al. [63], for instance, denote paragraphs as nodes, which are interconnected according to a similarity measure based on the number of words they share. Routing algorithms were proposed to select the most prominent paragraphs. In an evaluation with a corpus of 50 texts in English, the best algorithm have chosen 45.6% of the paragraphs selected by human summarizers. Although simple, the approach based on paragraphs is limited by the compression rate, since paragraphs cannot be broken to fit into the extract.

In another approach, Mani and Bloedorn [41] represent instances of terms as nodes, which are connected by cohesion relations such as proximity, repetition, synonymy and co-reference. The extract generated must satisfy a topic given as input. After selecting nodes corresponding to terms of the topic, a spreading activation algorithm gives a weight to each node, and the topmost nodes indicate which text sentences should be selected to compose the extract. In an experiment with five texts in English, this algorithm has outperformed the tf.idf measurement [62] and the node degree, which is defined in Section 3.1.

In Mihalcea's work [43], an extract is generated by selecting the sentences with the highest ranks given by recommendation algorithms developed to classify Web pages: Google's PageRank [51] and HITS [28]. In a network of sentences, there is an edge between two nodes whenever they share common terms, and the number of shared terms corresponds to the edge

weight. Differently from PageRank, HITS distinguishes the nodes with high indegree values (authorities – $HITS_A$) from the nodes with high outdegree values (hubs – $HITS_H$). Moreover, Mihalcea defines three types of networks: (i) Undirected, (ii) Forward, whose directed edges between sentences follow the text reading course, and (iii) Backward, with directed edges in the opposite course. The author has evaluated her proposal with the English corpus of DUC'2002 (Document Understanding Conference) [50] and with the Brazilian Portuguese corpus TeMário [54] according to the metric ROUGE-1 [37]. For the networks with forward and backward edges, HITS algorithms were superior to the high-scoring system of DUC'2002, while PageRank for backward edges performed a little worse than these systems. The best three strategies of Mihalcea for Brazilian Portuguese texts were also considered in the evaluation of the method proposed in this paper. Detailed results are reported in Section 4.2.

A similar technique was employed by Erkan and Radev [21] in multi-document summarization. Two variations of the PageRank algorithm and a measure of degree centrality were grouped in a linear combination, following Edmundson's approach [19], to compute node relevance in a sentence-based network linked by similarity. The sentence position and sentence length attributes were used in an evaluation with English texts of DUC'2003 and DUC'2004. Experimental results show that Erkan and Radev's approach is among the best competing systems of these DUC editions.

The methods outlined in this section use different linguistic knowledge to establish node adjacency in a network. Although sophisticated resources have become available for languages such as English, this is not true for other languages, including Portuguese. Therefore, the method proposed in this paper aims at producing extracts using only shallow linguistic knowledge: we have used a lemmatizer and a part-of-speech tagger to determine the connectivity between sentences represented as nodes.

3. CN-Summ: a complex network approach to extractive summarization

The technique presented here for automatically generating extracts is based on a set of network measurements typically applied to characterize complex networks [16]. The underlying assumption is that each measurement selected reflects some features of the source text that might be interesting for summarization. Complex networks concepts were considered potentially useful for the summarization task because they offer different, often complementary, views of a network, and thus can be used to highlight a subset of its nodes. For the sake of clarity and to focus the analysis mainly on the extractive algorithms, we employed a simple network that encodes one type of lexical cohesion: nodes represent sentences and there is an edge between two nodes if the corresponding sentences have at least one word in common (i.e. lexical repetition). Furthermore, only lemmatized nouns are considered, thus restricting the number of edges in the network, highly increased if other words are considered, which would make it more difficult to select a group of sentences. In subsidiary experiments, we observed that including verbs in our analysis did not improve the performance of the summarizers.

We also argue that if two sentences are connected in this network they probably convey complementary information about related topics, possibly about the same topic. We illustrate this idea with an example (see Fig. 1) taken from the corpus employed in our experiments (details about this corpus are given in Section 4). Note that the noun 'USP' (acronym for University of São Paulo) in Fig. 1 has been highlighted, which allowed us to verify that there must be at least one edge between every pair of nodes in the corresponding network of sentences (network not shown). Upon inspection one notes that each

(...)

For the majority of citizens from São Paulo that do not study or work there, the University of São Paulo (**USP**) appears as nothing else as an Ibirapuera Park in the west region of the city.

But **USP** integrates itself in the city via projects stemming from buses run on natural gas for the Metropolitan Company of Collective Transport to the selective collection of garbage, in addition to other 3 thousand specific services, such as veterinary surgeries and extension courses for executives.

(...)

"**USP** cannot fill holes in areas of social and health services", says Eduardo Siqueira Barbosa, director of projects in the Coordination for Cooperation and Special Activities (CECAE), from **USP**.

(...)

And the medical and psychological services, or even dental care, in the academic departments of **USP** are directed toward the needs of teaching and research – in summary, it is an exchange among the population, students and researchers.

(...)

Fig. 1. Newspaper article excerpt, originally written in Portuguese, taken from the corpus Temário [54]. This example illustrates the complementary relationship between sentences sharing the same noun (in this case, we have chosen the noun 'USP' as the main subject of our discussion).

sentence conveys a unique piece of information regarding the topic 'facilities offered by the university', thus complementing all other sentences. Nevertheless, two sentences may contain redundant information. But we argue that sentences complement one another in most cases; otherwise, texts would be extremely repetitive. We believe that this reflects our everyday experience on reading and writing: we rarely write two sentences expressing nearly the same content. As our goal is to construct informative extracts, the concept of complementary sentences is crucial for the development of our summarization techniques.

The proposed method, called *CN-Summ* (Complex Networks-based Summarization), consists of four steps. In Step 1, the source text is pre-processed in a straightforward manner: sentence boundaries are recognized and nouns are lemmatized. Nouns are identified by the part-of-speech tagger MXPost [2,58] and lemmatization is done by accessing the Brazilian Portuguese lexicon KLS [29,49]. In Step 2, the resulting text is mapped into a network representation according to the adjacency and weight matrices of order $N \times N$ (where N is the number of nodes/sentences) explained in Appendix A. These matrices, respectively called A and W, are used in Step 3, which computes one of a set of network measurements (defined in the following sections). Thus, the result of Step 3 is the association of a numerical value, or rank, to each network node. In the last stage (Step 4) the first n nodes (i.e. sentences) of the ranking are selected to compose the extract. The number of sentences in the extract, n, depends on the compression rate, and the way the sentences are ranked yields one summarization strategy. In what follows we introduce 7 network measurements leading to 14 different sentence rankings, i.e. 14 different summarization strategies. One further version is also proposed, which combines the previous strategies in what we call a voting system. Roughly speaking, this combination puts into an extract the sentences best ranked by most of the 14 strategies.

3.1. Degree strategies: CN-Degree and CN-Strength

In the first summarization strategy we try to identify informative nodes by using the number of sentences a node is connected to, i.e. its *degree*. The degree k_i of a node i is the number of other nodes connected to it. More specifically, k_i can be obtained from the adjacency matrix A as follows,

$$k_i = \sum_{j=1}^{N} a_{ij}. \tag{1}$$

If the elements of the matrix W are considered instead in Eq. (1), a slightly different type of degree, referred to as *strength*, is obtained,

$$s_i = \sum_{j=1}^{N} w_{ij}. \tag{2}$$

Therefore, the strength of a node i is the sum of the weights of the edges associated to that node (matrices A and W are defined in Appendix A). These two measurements, although simple and already known from graph theory, are frequently used to characterize complex networks (e.g. degree distribution and degree–degree correlations) [16]. Nodes with high k_i or s_i share a considerable number of connections with other nodes, and are called *hubs*. They play an important role in complex network theory, being the hallmark of scale-free networks [6].

Since the network encodes some kind of lexical cohesion between sentences, a sentence that shares a large number of links with other sentences probably conveys relevant information that complements many other sentences. The n sentences with the highest k_i are selected to build an extract in the first summarization strategy – called CN-Degree. Similarly, s_i can also be used to identify highly connected nodes, and the corresponding strategy – called CN-Strength – selects the sentences with the highest s_i to compose the extract.

3.2. Shortest path strategies: CN-SP, CN-SP^{wc} and CN − SP^{wi}

Measurements associated with distance usually take into account the global structure of a network [16]. One example of such measurements is the length of the *shortest path* between two nodes. A path is a sequence of non-repeating edges that leads one node to another, and the length of a path is the number of edges in the sequence. In contrast to the degree, this measurement considers not only the immediate neighbors of a node i, but also the nodes indirectly connected to it. These nodes are at a distance higher than 1 from node i, and this information may be important to identify indirect cohesive relations. Thus, the shorter these distances are, the stronger the corresponding sentence inter-relationships might be.

The length d_{ij} of the shortest path that connects nodes i and j can be calculated from the matrix A through various algorithms [11]. We take the mean length of the shortest paths that associate a node i to every other node in the network as a measurement of the overall accessibility of i. More formally, it is given by

$$sp_i = \frac{1}{N-1} \sum_{i \neq j} d_{ij}, \tag{3}$$

where $d_{ij} = N$ if a path between i and j does not exist (recall that N is the number of nodes in the network). The absence of paths linking some pair of nodes can be observed in disconnected networks. For weighted networks, the distance between

two nodes is usually taken as the sum of the edge weights contained in the shortest path linking them. In our case, it is necessary to change matrix W to put the most important edges, those with high weights, inside shortest paths. Thus, we have created weights inversely proportional to the original weights w_{ij}. This step is necessary because an edge with a high weight significantly increases the length of a path, and thus may be frequently avoided by a shortest path algorithm, even if it represents a strong relation between two sentences. We then define two matrices, W^{wc} and W^{wi}, that are both modified versions of W. The first one has elements $w_{ij}^{wc} = 0$ if $w_{ij} = 0$, or $w_{ij}^{wc} = w_{max} - w_{ij} + 1$ if $w_{ij} > 0$, i.e. its weights complement the values of W taking the highest element of W (w_{max}) as reference. When considering W^{wc} in the shortest path calculations, the measurement given by Eq. (3) is calculated taking $d_{ij} = N\langle w^{wc}\rangle$ when there is no path between i and j ($\langle w^{wc}\rangle$ is the average weight of W^{wc}). The other modified version of W, denoted by W^{wi}, has elements $w_{ij}^{wi} = 0$ if $w_{ij} = 0$, or $w_{ij}^{wi} = 1/w_{ij}$ if $w_{ij} > 0$, i.e. it takes the inverse of the weights of W. Eq. (3) can also be applied using matrix W^{wi}, this time defining $d_{ij} = N\langle w^{wi}\rangle$ when there is no path between two nodes i and j. Because a node with low sp_i is close to the other nodes on average, we now define a summarization strategy that selects for the extract the n nodes with the lowest values of sp_i. With the matrices A, W^{wc} and W^{wi} in the computation of sp_i, we obtain three different summarizers, identified, respectively, by CN-SP, CN-SP^{wc} and CN-SP^{wi}.

3.3. Locality index strategy: CN-LI

The *locality index* is a measurement that takes into account the pattern of connectivity in the neighborhood of a node [15]. If the neighbors of a node i share few connections with the remaining nodes of the network, they have a localized connectivity pattern and therefore are more strongly related with i than with other nodes. The locality index is useful for pointing out these central nodes of relatively isolated groups, which can represent important sentences that summarize the meaning of their neighbors.

The locality index compares the number of internal connections of the neighbors of a node i (denoted by N_i^{int}) with the number of their external connections (denoted by N_i^{ext}). More specifically, N_i^{int} is the number of edges between the k_i neighbors of i, plus the k_i edges that connect node i to its neighbors. N_i^{ext} sums up the connections that the k_i neighbors of i maintain with the remaining nodes of the network. Hence, the locality index associates N_i^{int} and N_i^{ext} in the following manner,

$$l_i = \frac{N_i^{int}}{N_i^{int} + N_i^{ext}}, \tag{4}$$

where $0 \leqslant l_i \leqslant 1$ (if $N_i^{int} = N_i^{ext} = 0$, then $l_i = 0$). If $N_i^{ext} = 0$ and $N_i^{int} > 0$, l_i takes its maximum value, i.e. the pattern of connections surrounding node i is local. Conversely, l_i approaches zero when N_i^{ext} is much higher than N_i^{int}, which indicates that the neighbors of i are more connected with other nodes than with each other.

Therefore, it is possible to think of a node i with high l_i as the center of a group of nodes that share few connections with the remaining nodes of the network. Since each of these groups is almost isolated from the rest, they may contain sentences of topics that are not considered by other sentences. More informative extracts might then be built if one sentence of each of those groups is selected, thus covering the topic structure of a text and also avoiding topic redundancy. Thus, another strategy for summarization, called CN-LI, gives priority to the central nodes of the groups mentioned, by selecting the top n sentences sorted in decreasing order of l_i.

3.4. d-Rings strategies: $CN-Rings^l$, $CN-Rings^k$ and $CN-Rings^{lk}$

A *d-ring* is a subgraph generated by the morphological operation dilation [14,25]. Dilation $\delta(g)$ of a subgraph g of a graph G is another subgraph that includes the nodes of g plus the nodes connected to g. When this operation is performed d times on g, we obtain its d-dilation,

$$\delta_d(g) = \underbrace{\delta(\delta(\ldots(g)\ldots))}_{d}, \tag{5}$$

where $\delta_0 = g$. Finally, the d-ring of g is a subgraph $R_d(g)$ whose nodes are

$$\mathcal{N}(\delta_d(g)) \setminus \mathcal{N}(\delta_{d-1}(g)), \tag{6}$$

where the symbol $\setminus$ denotes the set difference operation, $\mathcal{N}()$ is a function that gives the set of vertices of a graph and $R_0 = g$. The d-ring has the nodes of $\delta_d(g)$ that are not included in $\delta_{d-1}(g)$, and it is also called hierarchical level d of subgraph g. For our purposes, we consider that g always has only one single node, thus we denote a d-ring by $R_d(i)$ instead of $R_d(g)$.

The d-rings generalize the concept of node neighborhood (see also [12,13]). The first-level neighborhood of a node i, i.e. its k_i neighbors, corresponds to the 1-ring; its second-level neighborhood (the neighbors of its neighbors) corresponds to the 2-ring, and so forth. This notion of 'concentric' subgraphs is useful for selecting sentences that complement the central idea of a text. Thus we initially select for the extract an important node, the one with the highest degree k_i, called *hub*, and then select all nodes of its d-rings, from $d = 1$ to $d = d_{max}$, where d_{max} is defined according to the compression rate of the extract. This is the basis of the following three strategies: (i) the first, called CN-Rings^l, uses the sentence location to select those sentences that appear first in the text when it is not possible to entirely include in the extract the outermost $R_d(hub)$ (the one with $d = d_{max}$); (ii) the second, called CN-Rings^k, uses the degree to extract the sentences with the highest k_i when it is not

possible to completely include the outermost $R_d(hub)$; (iii) the third, called CN-Rings^{lk}, selects from every $R_d(hub)$ only the nodes with degree no lower than the average network degree, and also extracts the sentences that appear first in the source text when the outermost d-ring does not fit into the extract. Although d-rings are not exactly network measurements, they are also used to define the sentence ranking in Step 4 of the CN-Summ method.

It may be argued that selecting sentences related to the *hub* will probably produce a redundant and not broadly informative summary. An important assumption we have made in the design of CN-Summ, and that was explained previously in this paper (Section 3), is that if two sentences are connected in the network they probably complement each other, rather than providing the same meaning. Therefore, the ring-based summarizers, as well as the ones to be presented in the two following sections (based on k-cores and w-cuts), aim at complementing the core information of the text.

3.5. k-Cores strategies: CN − Cores^l and CN − Cores^k

A subgraph g of a graph G is a *k-core* if every node i of g has degree $k_i \geqslant k$. This subgraph, denoted by $core_k(G)$, must also be the greatest subgraph of G that has this property [8]. The assumption here is that a k-core represents a set of closely related sentences. To obtain a $core_k(G)$, it suffices to recursively exclude every node of G whose degree is lower than k. For the purposes of summarization, we also define a slightly different version of the k-core, denoted by $core'_k(G)$, which is the largest connected component of $core_k(G)$ (for every pair of nodes in a connected component there exists at least one path that connects them).

A non-empty $core'_k(G)$ with the maximum possible k, called the innermost k-core, is a subgraph that consists of densely connected nodes. Extending to subgraphs the interpretation given to the *hub* in Section 3.4, we now assume that the innermost k-core is relevant for summarization because it seems to be a nuclear group of sentences that express the main idea of the source text. We then define a new summarizing procedure that initially includes in the extract all the nodes that belong to the innermost k-core. Instead of using d-rings as in the previous section, we keep on calculating k-cores with sequentially decreasing k. The sentences of each new k-core are also included in the extract, except the ones that have already been inserted, until the compression rate is fulfilled. Notice that this algorithm gradually relaxes k-cores, thus allowing the main k-core to be complemented with other sentences. Two variations are proposed: (i) CN-Cores^l, that selects the sentences of a k-core that appear first in the source text when the last k-core is not allowed to be included completely in the extract because of compression rate restrictions, and (ii) CN-Cores^k, that instead selects the sentences with the highest degrees k_i.

3.6. w-Cuts strategies: CN − Cuts^l and CN − Cuts^k

Inspired by the idea behind k-cores, we defined another type of subgraph called *w-cut*. The k-core is used to find nuclear groups of nodes using only node degrees, while w-cut is defined to identify groups of closely related nodes using edge weights. We require that the w-cut of a graph G, identified by $cut_w(G)$, be a subgraph whose edge weights w_{ij} are not lower than w. Moreover, $cut_w(G)$ must be the greatest connected component of G after the exclusion of the edges that do not meet the weight criterion. The strategies based on w-cuts are analogous to the ones based on k-cores, from which two more strategies are obtained, namely CN-Cuts^l and CN-Cuts^k.

3.7. Communities strategy: CN-Communities

Another concept borrowed from the complex networks field is the notion of *communities* [16], which correspond to groups of nodes that are highly interconnected, while different groups are scarcely connected to each other. We argue that communities might correspond to the topics conveyed by the text. To make the definition of communities more accurate we use a quantity Q, called network modularity [10]. It is initially obtained from the following fraction:

$$\frac{\sum_{ij} a_{ij}\delta(c_i, c_j)}{\sum_{ij} a_{ij}} = \frac{1}{2M}\sum_{ij} a_{ij}\delta(c_i, c_j), \tag{7}$$

that measures the proportion of edges connecting intra-community nodes, where c_i is the community number that node i belongs to, $\delta(a, b)$ is 1 if $a = b$ or 0 if $a \neq b$, and M is the total number of edges in the network $\left(M = \frac{1}{2}\sum_{ij} a_{ij}\right)$. It is worth pointing out that this δ has nothing to do with the one in Section 3.4. The modularity Q of a network G is obtained by subtracting from Eq. (7) its expected value in a random network,

$$Q = \frac{1}{2M}\sum_{ij}\left[a_{ij} - \frac{k_i k_j}{2M}\right]\delta(c_i, c_j), \tag{8}$$

where k_i is the degree of i, k_j is the degree of j and $k_i k_j/2M$ is the probability that an edge (i, j) exists in a random network that preserves the degree of the nodes of G. When $Q > 0$ the modularity of G is greater than the expected for its random counterpart, and a value of Q greater than 0.3 indicates good modular structure in G. We adopted the greedy algorithm of Clauset et al. [10] to divide a network into communities. This procedure starts with N unitary communities (i.e. comprised of single nodes) and at each step it joins two communities to produce a new network partition with the highest possible

value of Q, until only one community remains. Finally, the division of a network into communities adopted is the one that shows the highest Q calculated by the algorithm.

Unlike k-cores and w-cuts, a community structure associates every node with a single community. Therefore, a community division is a partition of a network, which can be seen as a set of interconnected subnetworks. For the purpose of summarization, communities supposedly represent the topic structure of the source text. Although we did not verify experimentally this assumption, the corresponding strategy, CN-Communities, aims at covering the entire topic structure of a text, thus avoiding topic redundancy. It selects a number of sentences from each community, which is proportional to the community size, to satisfy the compression rate. Finally, when a community is to be only partially included in the extract, preference is given to nodes with highest degree k_i.

3.8. A voting strategy: CN-Voting

It is known that the combination of methods using a voting scheme can improve individual performances (e.g. combination of classifiers [31]). Thus, our last strategy, called CN-Voting, joins all previous strategies in an integrated voting approach, giving priority to the sentences that consistently appear at the top of the sentence rankings defined by each strategy. Therefore, CN-Voting needs to store all 14 sentence rankings. A ranking is the set of ordinal positions assigned to sentences by one strategy, i.e. the first sentence included in the extract has ordinal position 1, the second has ordinal position 2 and so forth, until the total of N sentences are selected and numbered. Therefore, the lower the sum $\Sigma(s)$ of all 14 positions of a given sentence s, the higher is the relevance attributed to it by most of the 14 voting strategies. The sentences selected by this voting approach should represent what the other strategies (or at least most of them) agree to be relevant for an extract. Since this is a position-based voting, the n sentences with the lowest $\Sigma(s)$ are selected to compose the extract, where n is defined by the compression rate.

4. Informativeness results and discussion

Two evaluation experiments were carried out using TeMário corpus [54], which comprises 100 newspaper articles in Brazilian Portuguese with an average article size of 613 words, or 29 sentences. For each text there is a pair of reference summaries: an abstract written by a human and an automatically generated extract, created by the tool GEI [55]. GEI builds an extract by selecting source sentences that are closer to the sentences of the manually written abstract. The proximity is quantified by the cosine similarity measurement, computed between word frequency vectors. This automatically generated extract is therefore guided by the contents of the human abstract, making it an acceptable reference for evaluations as an approximation of a human summary [40]. Our experiments compare CN-Summ with other extractive summarizers previously evaluated with the same corpus. The first experiment determines Precision/Recall scores (Section 4.1), whereas the second experiment employs a ROUGE metric (Section 4.2). Complementary experiments were also carried out to evaluate the effect of different compression rates on the performance of CN-Summ (Section 4.3). In order to illustrate the type of extracts obtained with CN-Summ, two examples taken from TeMário were included in a publicly available document.[1]

4.1. First experiment

The first experiment uses the reference extracts to compute Precision, Recall and F-measure [62] for the extracts automatically generated by CN-Summ strategies. We denote the set of sentences of the reference extract by E_r and the set of sentences of the automatic extract by E_a. The Precision of E_a is $P(E_a) = |E_r \cap E_a|/|E_a|$ and its Recall is $R(E_a) = |E_r \cap E_a|/|E_r|$, i.e. they relate the number of sentences that co-occur in both E_r and E_a to the total number of sentences of the automatic or reference extract, respectively. The F-measure adopted combines these two metrics in a balanced way, being defined by $F(E_a) = [2P(E_a)R(E_a)]/[P(E_a) + R(E_a)]$. Extracts were obtained by removing 70% of the source sentences (a compression rate of 30%), following previous summarization experiments that also use the TeMário corpus [35,60]. Hence, we are able to compare our results with published ones.

In this experiment, the 15 versions of CN-Summ are compared with two baselines and six other extractive systems. If n is the number of sentences of the extract, the first baseline system, called Top Baseline, includes in the extract the first n sentences of the source text, while the second, called Random Baseline, randomly selects n sentences of the source text. Other systems evaluated under the same experimental conditions are ClassSumm, NeuralSumm, GistSumm, TF-ISF-Summ, SuPor and its improved version SuPor-v2. ClassSumm [32] is an extractive summarizer that uses a machine learning algorithm to determine the most relevant segments of a text, similarly to the technique employed by Kupiec et al. [30]. Sentence features such as length, location, proper names, anaphor occurrence in the beginning of the sentence and cosine similarity to the title were used to induce a Naïve Bayes classifier [30,34]. NeuralSumm [57] uses a SOM (Self-Organizing Map) neural network to classify text sentences as essential, complementary or superfluous. Essential sentences are selected to compose an extract, while the superfluous are not. The inclusion of complementary sentences depends on the compression rate. GistSumm [56] is an extractive summarizer that assigns a score to each sentence of the text according to one of two methods: keywords or

[1] http://cyvision.ifsc.usp.br/~lantiq/download/CN-Summ-extracts.pdf.

average keywords. The first one scores each sentence according to the sum of the frequencies of its words. The second method normalizes the score of each sentence by its size, avoiding a bias for longer sentences in the summarization process. The highest scored sentence by any of the scoring methods is said to be the "gist sentence", i.e. the sentence in the text that best expresses its main idea. Additional sentences are selected if they obey two criteria: correlation with the main idea, by sharing at least one word with the gist sentence; and relevance, by having a score above a threshold, which is computed as the average of all sentences scores in the text. Another system, called TS-ISF-Summ [33], computes sentence relevance by using the tf.isf metric, which is the well-known tf.idf metric of information retrieval [62] applied to a collection of sentences instead of a collection of documents. Topics are also detected and the chosen sentences of each topic are those containing the most frequent words of the topic. SuPor [59] employs a Naïve Bayes classifier to assign weights to sentences (i.e. the probability of being included in the extract), and computes features such as sentence length, word frequency, sentence location, occurrence of proper nouns and lexical chaining (this last feature uses the WordNet lexicon [45] and an ontology to identify semantic relationships between nouns). The top-ranked sentences are included in the extract. Its improved version, SuPor-v2 [35], can generate more informative extracts by handling sentence features in a flexible way. Moreover, it uses a feature selection algorithm to shrink the feature space. All these systems were previously evaluated with the same TeMário corpus of 100 texts [35,60], and the average Precision, Recall and *F*-measure scores of each one are reproduced in Table 1. This table also includes the results of the two baseline systems, as well as the results of all CN-Summ strategies.

In what follows we discuss the average scores of all systems, while in the end of this subsection we analyze the statistical significance of the differences between average *F*-measures. Some of CN-Summ versions are among the best systems for Portuguese summarization, with *F*-measure around 42%. Considering only our systems, the best results were produced, as expected, by CN-Voting, which has higher average Precision, but lower average Recall, than the overall best system (SuPor-v2). CN-SP^{wc} also presents a reasonable *F*-measure (42.4%), highlighting the importance of shortest paths in this experiment (CN-SP and CN-SP^{wi} have a little lower *F*-measure). CN-Rings^{k} has a good *F*-measure (42.2%) as well, while other variations of ring-based strategies have lower scores. The degree seems to positively influence the performance of some systems, since the *d*-rings are obtained from the most connected node, and the systems that directly use this measurement (CN-Degree and CN-Strength) have good performance. Groups of highly interconnected nodes also appear to be relevant for summarization, as shown by the results of the strategies based on *w*-cuts and *k*-cores. Moreover, these systems are more positively influenced by the degree rather than by the sentence location, since their best variations are CN-Cuts^{k} and CN-Cores^{k}. The versions based on locality index (CN-LI) and communities (CN-Communities) achieved *F*-measures close to the score of the Top Baseline system. Perhaps the reason for this lower performance is the modest modularity of the networks (*Q*, defined in Section 3.7, is 0.26 on average), since these two strategies rely on the modular structure of a network.

When comparing our summarization approach with others, CN-Voting has higher Precision than ClassSumm, SuPor and SuPor-v2, and higher Recall than ClassSumm. All these other systems are based on machine learning, thus requiring a training phase, and use substantially more complex resources. Both versions of SuPor demand, for example, information from lexical chains (i.e. word semantics) and from importance of topics, while ClassSumm uses an argument structure of the source text. CN-Summ, on the other hand, does not employ any semantic information nor demands a training stage. Nevertheless,

Table 1
Average Precision ⟨*P*⟩, Recall ⟨*R*⟩ and *F*-measure ⟨*F*⟩, in percentages (%), obtained in the first experiment with TeMário corpus. The lines are placed in decreasing order of ⟨*F*⟩.

	Systems	⟨*P*⟩	⟨*R*⟩	⟨*F*⟩
1	SuPor-v2	47.4	43.9	45.6
2	CN-Voting	48.1	40.3	42.9
3	SuPor	44.9	40.8	42.8
4	CN-SP^{wc}	47.4	39.9	42.4
5	ClassSumm	45.6	39.7	42.4
6	CN-Rings^{k}	47.2	39.8	42.2
7	CN-Degree	47.0	39.7	42.1
8	CN-Strength	47.0	39.3	41.8
9	CN-Cuts^{k}	46.5	39.2	41.6
10	CN-SP^{wi}	46.6	38.8	41.4
11	CN-SP	46.4	39.0	41.4
12	CN-Cores^{k}	46.2	38.9	41.3
13	CN-Cuts^{l}	46.0	38.7	41.1
14	CN-Rings^{lk}	45.7	38.6	40.8
15	CN-Cores^{l}	44.6	37.1	39.6
16	CN-LI	44.6	37.0	39.6
17	CN-Communities	44.1	37.0	39.4
18	CN-Rings^{l}	44.3	37.0	39.3
19	Top Baseline	41.7	35.0	37.1
20	TF-ISF-Summ	39.6	34.3	36.8
21	GistSumm	49.9	25.6	33.8
22	NeuralSumm	36.0	29.5	32.4
23	Random Baseline	34.0	27.8	30.0

SuPor-v2 exhibited higher scores than the other systems, reinforcing the good performance of its previous version [60]. A remarkable result is that all CN-Summ versions outperformed TF-ISF-Summ, GistSumm and NeuralSumm, systems that also work with shallow linguistic resources. Also worth noticing is that these three systems have lower scores than the Top Baseline.

For a statistical significance analysis of these results, we included in Fig. 2 the p-values (in gray scale) of the corresponding t-tests [47,64] performed between all CN-Summ strategies. We chose the paired t-test between average F-measures with a null hypothesis of 'equal averages'. Thus, a low p-value (near zero, since $0 \leqslant p\text{-value} \leqslant 1$) indicates a significant difference between two average F-measures. We were unable to include some systems in this analysis, since we do not have access to the complete experimental data generated by other authors (e.g. for ClassSumm). Nevertheless, we obtained the full data for SuPor-v2, which allowed a direct comparison with our systems. Fig. 2 shows the results in a tabular form where each square indicates the p-value of the t-test between two systems. We consider that a statistically significant result should have $p\text{-value} \leqslant 0.05$. Such specific results are indicated in the following analysis, since Fig. 2 does not show them numerically.

Significant differences were obtained between the Random Baseline and all CN-Summ strategies. Nevertheless, the CN-Summ strategies CN-Cores^{l}, CN-LI, CN-Communities and CN-Rings^{l} cannot be regarded as statistically different from the Top Baseline. Moreover, these four systems are not statistically different from each other (see the dark group of squares in the lower right portion of Fig. 2). Significant differences existed between these four strategies and some top-scoring systems (SuPor-v2, CN-Voting, CN-SP^{wc}, CN-Rings^{k}, CN-Degree and CN-Strength). Other differences were found between CN-Voting and CN-Cores^{k}, and between CN-LI and CN-Cores^{k}. In summary, CN-Summ strategies may be classified into two groups, viz.: systems with average F-measure higher or lower than 40% (see Table 1). The lowest p-values are placed, in general, between these two groups, whereas the highest p-values are generally located inside those groups (Fig. 2 clearly shows two groups of high p-values). In other words, the main interpretation of these results is that systems inside these groups are not statistically different from each other. Furthermore, it is worth pointing out that, in this experiment, SuPor-v2 cannot be considered better than CN-Voting, CN-SP^{wc}, CN-Rings^{k}, CN-Degree and CN-SP^{wi}, because the corresponding p-values are all larger

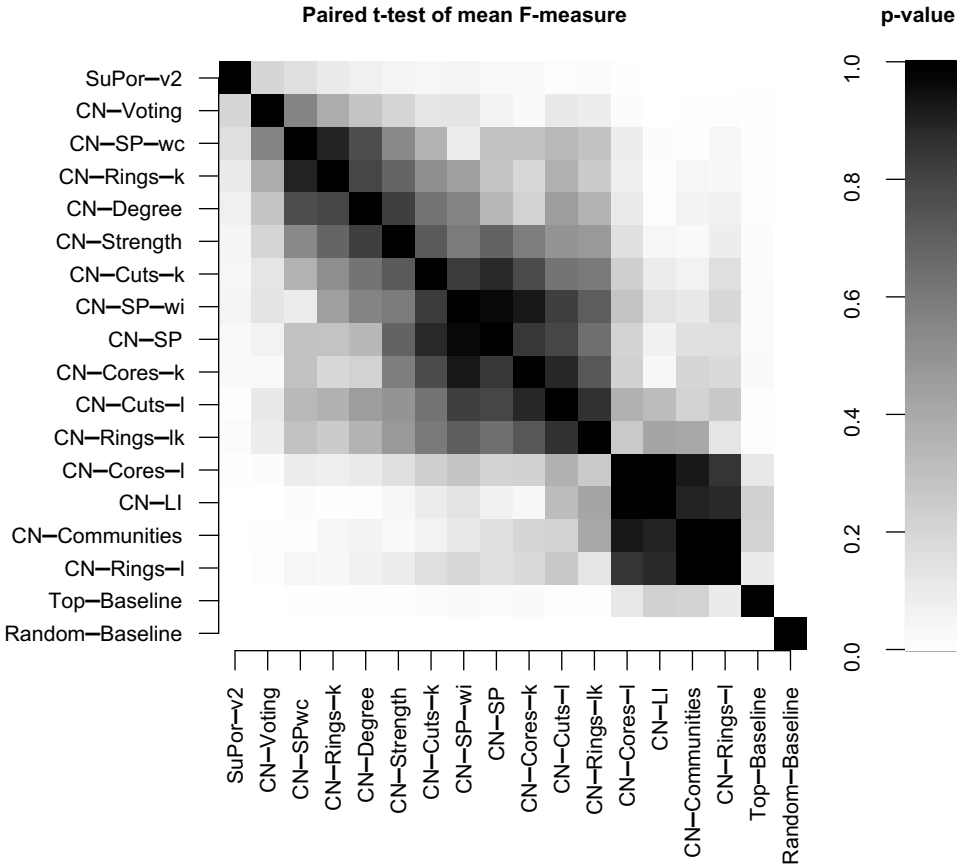

Fig. 2. p-Values of paired t-tests between summarizers. A low p-value (indicated in this figure by near white gray levels) indicates a statistically significant difference between two average F-measures. The names of the corresponding systems are also given in the figure, following the same ordering of systems presented in Table 1, i.e. in decreasing order of average F-measure.

than 0.05. This means that some of CN-Summ strategies perform as well as SuPor-v2, one of the best summarizers for Brazilian Portuguese texts reported so far.

4.2. Second experiment

In the first experiment we could not compare the performance of CN-Summ with some systems for Brazilian Portuguese [36,43], since the latter were evaluated in a different environment. Now, instead of using Precision/Recall, in this second experiment we applied a ROUGE metric to evaluate summary informativeness [37]. More specifically, we have chosen the unigram-based recall metric ROUGE-1, taking as reference the human-made abstracts of TeMário corpus. In order to have as meaningful a comparison as possible with published data, we did not remove stopwords from the extracts before computing ROUGE-1. This is also the reason why in this experiment we considered a different compression rate, which is based on the number of words of the reference abstracts of TeMário. This rate ensures that each automatic extract has approximately the same number of words of the reference abstract. Moreover, CN-Summ is not allowed to include a sentence in the extract when it exceeds the compression rate limit. That sentence is ignored and the selection process continues picking up sentences until one that fits into the extract is found. Therefore, no broken sentence is selected by a CN-Summ strategy.

Table 2 shows the average ROUGE-1 scores obtained with the CN-Summ versions, with the two baselines and six other extractive systems: SuPor-v2 [36], the best three variations of Mihalcea's method, presented in Section 2 [43], namely PageRank Backward, $HITS_A$ Backward and $HITS_H$ Forward, in addition to two modified versions of Mihalcea's PageRank Undirected, called TextRank + Thesaurus and TextRank + Stem + StopwordsRem [36]. For short, the last two are named here TextRank + T and TextRank + S + S, respectively. The former uses the entries of a thesaurus to define edges between sentences, while the latter applies a pre-processing step of stemming and stopwords removal.

Once again, some of CN-Summ strategies are close to the top-scoring systems, with ROUGE-1 ≈ 0.5. CN-Voting is again the best of our systems (on average), with a score of 0.5031. As in the previous experiment, the degree seems to play an important role on summarization, because CN-Strength and CN-Degree achieved good results, viz. 0.5020 and 0.5003, respectively. The degree is also relevant to CN-Rings^{lk}, a summarization strategy that completes the set of CN-Summ versions with ROUGE-1 score higher than 0.5. But now the best variation of k-cores is the one using location of sentences (CN-Cores^{l}) instead of node degree. Shortest paths show a poorer performance, closer to the Top Baseline. The versions based on w-cuts have relatively lower results than in the previous experiment, figuring now among the worst systems (near the Random Baseline), while CN-LI and CN-Communities have again low scores. In general, the variations of the strategies based on degrees, shortest paths, d-rings and k-cores consistently show good performances in both experiments. In other words, the number of connections (degree), the distance to other nodes (shortest paths), the distance to the *hub* (d-rings) and nuclear groups of nodes (k-cores) seem to be important to choose nodes (sentences) for the summarization of newspaper articles. Thus, our hypotheses proposed in Sections 3.1, 3.2, 3.4 and 3.5 appear to be valid.

Another important result of this experiment is the relative improvement of the Top Baseline performance when compared to the first experiment. The changes in the compression rate and in the evaluation metric put many systems below this

Table 2
Results of the second experiment using TeMário corpus, with average ROUGE-1 scores $\langle RG1 \rangle$ sorted in decreasing order. The 95% confidence level intervals for the average scores are also shown in this table (there is a dash for a system whose confidence interval we could not find in the literature). Both averages and confidence intervals were calculated with the ROUGE package using the bootstrap re-sampling method [47] with 1000 sampling points.

	Systems	$\langle RG1 \rangle$	Confidence interval
1	SuPor-v2	0.5839	–
2	TextRank + T	0.5603	–
3	TextRank + S + S	0.5426	–
4	PageRank Backward	0.5121	–
5	CN-Voting	0.5031	[0.4901, 0.5155]
6	CN-Strength	0.5020	[0.4886, 0.5144]
7	CN-Rings^{lk}	0.5019	[0.4877, 0.5156]
8	CN-Degree	0.5003	[0.4863, 0.5134]
9	$HITS_A$ Backward	0.5002	–
10	$HITS_H$ Forward	0.5002	–
11	CN-SP^{wi}	0.4995	[0.4861, 0.5124]
12	CN-Rings^{k}	0.4994	[0.4853, 0.5122]
13	CN-Cores^{l}	0.4992	[0.4861, 0.5124]
14	Top Baseline	0.4984	[0.4834, 0.5125]
15	CN-SP^{wc}	0.4982	[0.4853, 0.5108]
16	CN-Cores^{k}	0.4978	[0.4839, 0.5111]
17	CN-SP	0.4975	[0.4842, 0.5100]
18	CN-Rings^{l}	0.4968	[0.4824, 0.5102]
19	CN-Communities	0.4959	[0.4821, 0.5090]
20	CN-Cuts^{l}	0.4940	[0.4802, 0.5069]
21	CN-LI	0.4935	[0.4801, 0.5060]
22	CN-Cuts^{k}	0.4889	[0.4755, 0.5021]
23	Random Baseline	0.4765	[0.4634, 0.4897]

baseline. It can also be noticed that SuPor-v2, PageRank and TextRank appear at the top of Table 2, while HITS systems are at a little lower position. SuPor-v2 has consistently the best scores in both experiments, probably because it uses deep linguistic knowledge, as already mentioned. PageRank Backward, which does not require deep linguistic resources, shows the generality of the PageRank algorithm, which was created to rank Web pages. Moreover, the addition of simple resources like a stemmer and a stoplist in PageRank (which is what TextRank + S + S does), increased its score from 0.5121 to 0.5426. Thus it is likely that the performance of CN-Summ may be improved with addition of a thesaurus, analogous to what has been done in TextRank + T. Nevertheless, this experiment shows that CN-Summ is already competitive when compared with the best systems based on linguistically shallow resources (TextRank + S + S, PageRank Backward and HITS variations). Indeed, in the first experiment some CN-Summ strategies performed better, on average, than TF-ISF-Summ, GistSumm and NeuralSumm, which are other linguistically shallow systems.

The significance of the results presented in this section are shown in Fig. 3 using the same approach discussed in Section 4.1, i.e. with a paired *t*-test to assess the statistical difference between average ROUGE-1 scores. For lack of the full data for some systems, e.g. the scores per text generated by PageRank Backward, we could not include them all in the analysis. Fig. 3 is complemented by the confidence intervals generated by the software ROUGE, which are shown in Table 2. An average ROUGE-1 is inside the confidence interval with a 95% confidence level.

Statistically significant results (p-value $\leqslant 0.05$) are scarcer in this experiment than in the first. This means that the majority of summarizers included in Fig. 3 are not significantly different. This result is confirmed by the confidence intervals of Table 2, where they almost always overlap one another. Significant differences were found between the Random Baseline and all other systems, but not between the Top Baseline and CN-Summ strategies. Thus, as observed earlier in this section, the Top Baseline performs similarly to our systems in this experiment. Low p-values were found between CN-Cuts^{l} and CN-Voting, between CN-LI and the four top-scoring systems, and between CN-Cuts^{k} and the seven top-scoring systems. These results show that significant differences were only found between a few top-scoring and a few low-scoring systems. Ultimately, Fig. 3 demonstrates that almost all CN-Summ strategies, from CN-Voting to CN-Communities, are not significantly different from each other: every p-value in this case is greater than 0.05. These results are quite different from the ones obtained in the first experiment, where we were able to identify many significant differences among summarizers.

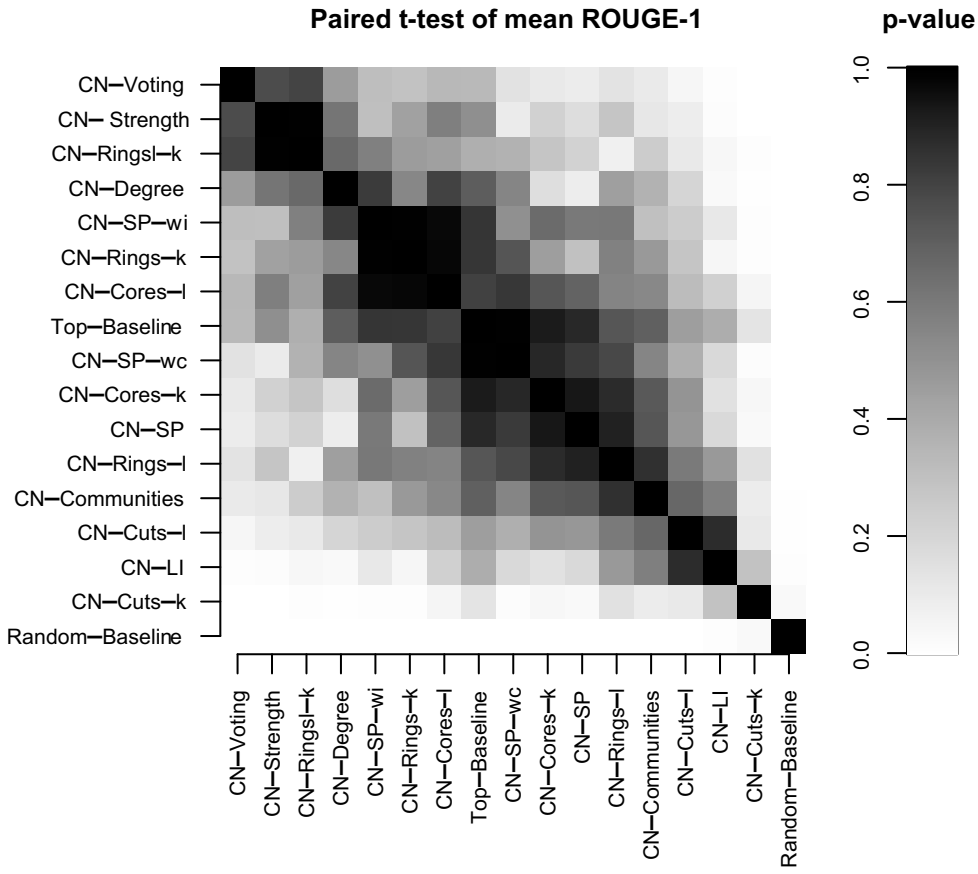

Fig. 3. p-Values of paired t-tests between summarizers. Please, refer to Fig. 2 for a more detailed description of the figure format. The names of the corresponding systems follow the same ordering of systems presented in Table 2, i.e. in decreasing order of average ROUGE-1.

4.3. Complementary experiments

In this subsection we report additional experiments carried out to investigate the effect of the compression rate on the performance of all CN-Summ strategies. We assessed the quality of shorter extracts employing the same framework as the experiments discussed in Sections 4.1 and 4.2, thus allowing the identification of which systems perform better in more restricted environments.

In Section 4.1 a compression rate of 30% was employed, i.e. 70% of source sentences were removed to generate extracts. Furthermore, co-selection scores (i.e. Precision, Recall and *F*-measure) were applied to assess the informativeness of extracts. We now complement this experiment by using three different compression rates: 20%, 10% and 5%. It is worth pointing out that, in these new experiments, we were only able to compare CN-Summ with the two baseline systems, namely Top Baseline and Random Baseline. Moreover, paired *t*-tests were again performed between all average *F*-measures, therefore all results reported in this subsection refer to statistically significant differences between systems (i.e. p-value $\leqslant 0.05$). When considering the new compression rate of 20%, we could notice that CN-Voting continues to be the highest scoring system, with *F*-measure equal to 34.6%. The Top Baseline and the Random Baseline are the systems with the lowest *F*-measures, 30.8% and 23.8%, respectively, and all systems are able to (statistically) improve on the Random Baseline. Nevertheless, the Top Baseline is now closer to CN-Summ systems, since only four of our summarizers are significantly better than this baseline. In general, almost all CN-Summ strategies are equivalent to each other in this experiment.

More restrictive compression rates continue to increase the performance of the Top Baseline. For instance, when using a 10% compression rate, the second best system is the Top Baseline with an *F*-measure equal to 22.1%, close to (and not statistically different than) the score of the best system (CN-Voting, with 22.8%). Nevertheless, paired *t*-tests also show that CN-Voting achieves a significant improvement on some low-scoring systems (such as CN-Cores^{k}, CN-Communities and CN-LI), a feature not shared with the Top Baseline. The Random Baseline continues with the lowest score (*F*-measure = 13.1%), again being statistically outperformed by all other systems. The same behavior was observed when reducing the compression rate once more (i.e. to 5%), where the *F*-measure for the Random Baseline is 8.6%. The Top Baseline maintains its top performance with a score equal to 15.3%, the same of CN-Cores^{l}. Indeed, both systems are the only ones to show significant improvements on other systems, such as CN-SP^{wi}, CN-Communities and CN-LI. All in all, these three complementary experiments show that CN-Summ strategies tend to be equivalent to each other when shorter summaries are generated, according to the statistical tests performed. Furthermore, their performances are likely to be equivalent to the observed for the Top Baseline, suggesting that our approach does not make an improvement over the Top Baseline for very short summaries. Nevertheless, CN-Summ consistently shows statistically significant improvements on the Random Baseline.

Another set of complementary experiments was carried out, this time changing the compression rate of the experiment reported in Section 4.2, where ROUGE-1 scores were employed. In this case, the compression rate is based on the size of the reference abstract according to its number of words. In Section 4.2, we have restricted the automatically generated extract to contain the same number of words as the reference abstract. Thus, we now reduce the size of extracts to 80%, 40% and 20% of the size of the reference summary, as opposed to the 100% (i.e. same sizes) employed earlier in this paper. Once more, we could only compare CN-Summ with the Top and Random Baseline systems. When considering the 80% reduction of the compression rate, we observed that CN-Voting maintains its highest performance with a ROUGE-1 score of 0.4334. Indeed, it statistically outperforms many other CN-Summ strategies, such as CN-SP^{wc}, CN-LI and CN-Cuts^{k}. The Random Baseline is at the opposite side of the ranking, with a score of 0.3994, and all systems are able to significantly improve on it (again, in a statistical sense), a feature consistently observed in every experiment reported in this subsection. Moreover, the other systems, excluding the high- and low-scoring ones, are all equivalent to each other.

When employing a 40% reduction in the compression rate, the four top-scoring systems (i.e. CN-Strength, CN-Cores^{l}, CN-SP^{wc} and CN-Voting, with scores ranging from 0.2441 to 0.2424) are able to significantly improve on four other systems, including the Random Baseline (with score 0.2115). In this experiment, CN-Cores^{l} outperforms the Top Baseline, which has a score of 0.2366. In no other ROUGE-based experiment previously reported in this paper a CN-Summ strategy was able to statistically outperform the Top Baseline. When shrinking extracts to 20% of the original compression rate, CN-Cores^{l} again improves on the Top Baseline (ROUGE-1 scores are 0.1263 and 0.1200, respectively). Indeed, CN-Cores^{l} has the best average score in this experiment, and the lowest average score was generated by the Random Baseline, with ROUGE-1 equal to 0.1046. In summary, these experiments show that all CN-Summ strategies are able to outperform the Random Baseline, consistent with the experiments based on *F*-measure scores. Furthermore, almost all CN-Summ strategies are statistically equivalent to each other when considering average scores. Nevertheless, as shorter summaries are generated, CN-Cores^{l} starts to outperform the Top Baseline, an improvement not observed earlier in ROUGE-based experiments.

5. Correlations between CN-Summ strategies

We conclude the evaluation of CN-Summ with an analysis of correlation between its different strategies. This type of analysis depends only on the corpus employed, not on a fixed compression rate or specific informativeness score. To accomplish this, we calculated the Spearman rank correlation coefficient r_s [47] between every pair of CN-Summ strategies, whose results are shown in Fig. 4. Recall that each CN-Summ strategy ranks the sentences from the source text (see Section 3.8), i.e. the first selected sentence by a strategy has rank 1, the second selected sentence has rank 2, and so on. Thus, for a given

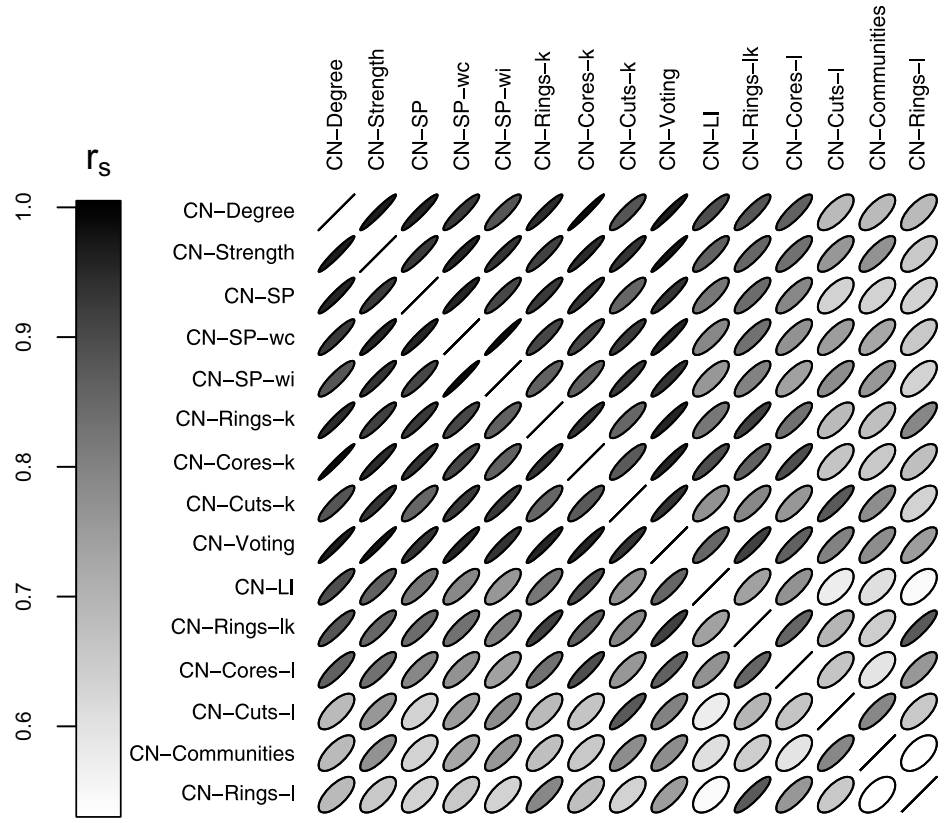

Fig. 4. Matrix of Spearman rank correlation coefficients r_s computed between CN-Summ strategies, where r_s is in the interval [0.54, 1.00]. The coefficients are denoted both by different shapes of an ellipse and by different gray levels. The size of the ellipse and its gray level color are inversely proportional to the corresponding correlation [46]. Thus, thin black ellipses refer to the strongest correlations. The inclination of an ellipse also indicates the sign of the correlation (in this case, there are only positive correlations). The lines and columns of the correlation matrix were permuted to improve visualization.

source text, it is possible to evaluate the correlation between two CN-Summ strategies (or between the rankings defined by them) using the Spearman correlation coefficient. The Spearman coefficient r_s is a quantity between -1 and 1, where strong correlations result in $|r_s|$ approaching 1, and the sign of r_s gives the inclination of the correlation, i.e. positive or negative. Since we have these correlations for every source text of the TeMário corpus, Fig. 4 shows the *average* correlation between all CN-Summ strategies, including CN-Voting. For the remainder of this section, we consider that a high correlation between summarizers has $r_s \geqslant 0.9$ on average. Notice that negative correlations do not exist in Fig. 4. It is worth pointing out that every high average r_s has low standard deviation (lower than 0.085) in this case, thus indicating that high average correlations do not vary considerably. Moreover, high correlations are all statistically significant in this experiment, since the corresponding p-values are lower than 2.3×10^{-4}.

High correlations were deliberately placed in the top left portion of Fig. 4 to enhance visualization. The group of strategies starting at CN-Degree and ending at CN-Voting (please, refer to the figure) contain most of the high correlations found. Indeed, 27 out of 29 high correlations occur between these strategies (these numbers were computed disregarding the symmetry of the correlation matrix). This set of strategies includes other correlations lower than 0.9 (e.g. between CN-SP and CN-Cuts^k); however, they are almost as strong as the other correlations of this group (all higher than 0.85), allowing us to refer to it as a set of strongly correlated strategies. Interestingly, this set of strategies consistently shows the highest average *F*-measures in the first experiment (Table 1), although in the second experiment the performance of our approach is not so distinctive. The observed behavior in the first experiment may be caused by the important role played by the degree in these strategies. For instance, in addition to the strategies based solely on degrees (CN-Degree and CN-Strength), the ones based on shortest paths are strongly influenced by degrees, since highly connected nodes tend to be closer to other nodes. Moreover, the strategies CN-Rings^k, CN-Cores^k and CN-Cuts^k are all based on highly connected nodes, although in different ways. Probably these strategies have biased CN-Voting, as they are the majority of CN-Summ strategies and would "vote" for the same sentences to be included in the extract. This idea is reinforced by the fact that CN-Voting is highly correlated with these strategies as well, and that it is not significantly better than the majority of them (see the significance tests in Section 4).

Two other strong correlations occur between CN-Rings^{lk} and CN-Rings^k, and between CN-Rings^{lk} and CN-Voting. Nevertheless, in general CN-Rings^{lk} is not highly correlated with other CN-Summ strategies. Similar behaviors were found for strategies CN-LI, CN-Cores^l, CN-Cuts^l, CN-Communities and CN-Rings^l. This means that these strategies can be used to complement one of the strategies of the group of highly correlated ones. For instance, information about community

structure could be used as a complementary parameter in CN-SP^{wc}, thus assisting the selection of sentences based on shortest paths. The weakest correlations can be found in the lower right portion of Fig. 4. This indicates that the corresponding strategies (from CN-LI to CN-Rings^{l}) tend to generate rather different extracts. As the results of Section 4 show, they also tend to have low informativeness scores. Therefore, although each one was based on different network features, they shared the common feature of not generating good extracts for the source texts of TeMário.

6. Final remarks

We have described the use of complex networks concepts for extractive summarization, as well as its evaluation through standard informativeness scores and significance tests. A simple network representation of texts was defined, which requires only shallow text pre-processing. Thus, the potential of our approach could be assessed by maintaining the focus on the summarization algorithms rather than on the construction of the networks. The systems proposed already show reasonable results for the summarization of newspaper articles in Brazilian Portuguese. Using automatic evaluation metrics, it was possible to identify the most promising strategies: the ones based on degrees, shortest paths, *d*-rings and *k*-cores. A voting summarizer was also created, which encompasses all the 14 strategies proposed. Some of the CN-Summ versions performed as well as the best summarizers of Portuguese texts reported in the literature, with evaluation being made within the same experimental setting. An analysis of correlation between the proposed methods was also performed, allowing us to identify similar or complementary strategies.

The definition of the network is extremely important, with possible large impact on the performance of network-based summarization strategies. Such a definition may be improved in future research using more language resources. For example, if we stick to the idea of using a network of sentences, the strengthening or fine-tuning of sentence interconnections may be performed by employing: (i) anaphor resolution, which can be used to identify a link between anaphors and the corresponding antecedents in different sentences, thus creating edges currently ignored; (ii) recognition of multiword expressions, since we only identify single words in our pre-processing step and compound nouns are incorrectly treated as separate nouns; and (iii) a thesaurus or lexical chains, thus being able to detect semantic, lexical relationships such as synonyms/antonyms and hyponyms/hypernyms, which also allows the assignment of distinct edge weights for different types of lexical links.

Further improvements also include joining all CN-Summ strategies in a machine learning approach, possibly using the correlation analysis of Section 5 for feature selection. Another choice would be integrating two or more non-correlated summarization strategies into a new summarizer, aiming at complementing one another. For example, since some strategies do not try to cover the topic structure of the source text (e.g. shortest path and degree strategies), it would be useful for complementing them with grouping (topic) information given by the communities and locality index strategies, thus trying to avoid topic redundancy in the former strategies. Evaluations using other corpora and different languages may also be carried out to assess the generality of our approach. To some extent, the hypothesis of this work has been proven by the results: network measurements, which are neither language nor domain dependent, can be used for extractive summarization, and can lead to informativeness scores close to the more linguistically complex and computationally costly systems. This should perhaps be expected, as the measurements of complex networks have already been shown to capture important features of texts [3–5,53].

Acknowledgements

The authors thank the scientific Brazilian agencies CNPq and FAPESP for supporting this work. L. Antiqueira is grateful for grants 132154/06-4 (CNPq) and 05/03361-8 (FAPESP). L. da F. Costa in recipient of grants 301303/06-1 (CNPq) and 05/00587-5 (FAPESP). We also thank D.S. Leite and L.H.M. Rino for providing detailed data about the evaluation of SuPor-v2 with *F*-measure scores.

Appendix A. Adjacency and weight matrices

An undirected graph or network G of N nodes and M edges can be represented by the *adjacency matrix* A, symmetric and of order $N \times N$, whose elements a_{ij} and a_{ji} are equal to 1 if there is an edge between nodes i and j, or equal to 0 otherwise. As explained in the beginning of Section 3, an edge exists if there is a co-occurrence of the same lemmatized noun between two sentences. If an edge has a numeric label, it is said to have a *weight*. Thus, another matrix can be employed, called the *weight matrix* W, which in our case is defined as follows. Let $P_i = \{p_1, p_2, \ldots, p_{n_i}\}$ be the set of n_i lemmatized nouns of the ith sentence of the source text. The weight w_{ij} (or w_{ji}) of the edge that links sentences i and j is the number of noun co-occurrences between them, i.e. $w_{ij} = w_{ji} = |P_i \cap P_j|$. If $w_{ij} = 0$, no edge exists between nodes i and j. The weights w_{ij} are therefore elements of the symmetric matrix W of order $N \times N$, which represents an undirected weighted network of a given source text.

References

[1] R. Albert, A.L. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (2002) 47–97.
[2] S.M. Aluisio, R.V. Aires, Corpus tagging and construction of a Portuguese tagger, Technical Report NILC-TR-00-2, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos, Brazil, 2000, 18pp. (text in Portuguese).

[3] D.R. Amancio, L. Antiqueira, T.A.S. Pardo, L. da F. Costa, O.N. Oliveira Jr., M.G.V. Nunes, Complex networks analysis of manual and machine translations, International Journal of Modern Physics C 19 (4) (2008) 583–598.
[4] L. Antiqueira, M.G.V. Nunes, O.N. Oliveira Jr., L. da F. Costa, Strong correlations between text quality and complex networks features, Physica A 373 (2007) 811–820.
[5] L. Antiqueira, T.A.S. Pardo, M.G.V. Nunes, O.N. Oliveira Jr., L. da F. Costa, Some issues on complex networks for author characterization, in: Proceedings of the Fourth Workshop in Information and Human Language Technology (TIL'06), 2006.
[6] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[7] R. Barzilay, M. Elhadad, Using lexical chains for text summarization, in: Advances in Automatic Text Summarization, MIT Press, 1999, pp. 111–121.
[8] V. Batagelj, M. Zaversnik, Partitioning approach to visualization of large networks, in: Proceedings of the Graph Drawing: Seventh International Symposium (GD'99), vol. 1731 of LNCS, 1999, pp. 90–98.
[9] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: structure and dynamics, Physics Reports 424 (4–5) (2006) 175–308.
[10] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Physical Review E 70 (2004) 066111.
[11] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, The MIT Press, New York, 2001.
[12] L. da F. Costa, The hierarchical backbone of complex networks, Physical Review Letters 93 (2004) 098702.
[13] L. da F. Costa, R.F.S. Andrade, What are the best concentric descriptors for complex networks?, New Journal of Physics 9 (2007) 311.
[14] L. da F. Costa, L.E.C. da Rocha, A generalized approach to complex networks, European Physical Journal B 50 (2006) 237–242.
[15] L. da F. Costa, M. Kaiser, C.C. Hilgetag, Beyond the average: detecting global singular nodes from local features in complex networks, 2006. arXiv:physics/0607272.
[16] L. da F. Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Characterization of complex networks: a survey of measurements, Advances in Physics 56 (1) (2007) 167–242.
[17] R. Diestel, Graph Theory, second ed., Springer-Verlag, New York, 2000.
[18] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, Proceedings of the Royal Society of London B 268 (2001) 2603–2606.
[19] H.P. Edmundson, New methods in automatic abstracting, Journal of the Association for Computing Machinery 16 (2) (1969) 264–285.
[20] P. Erdős, A. Rényi, On random graphs I, Publicationes Mathematicae Debrecen 6 (1959) 290–297.
[21] G. Erkan, D.R. Radev, LexRank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research 22 (2004) 457–479.
[22] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the Internet topology, in: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 1999, pp. 251–262.
[23] R. Ferrer i Cancho, R.V. Solé, The small world of human language, Proceedings of the Royal Society of London B 268 (2001) 2261–2265.
[24] Y. Guo, G. Stylios, An intelligent summarization system based on cognitive psychology, Information Sciences 174 (1–2) (2005) 1–36.
[25] H.J.A.M. Heijmans, P. Nacken, A. Toet, L. Vincent, Graph morphology, Journal of Visual Communication and Image Representation 3 (1) (1992) 24–38.
[26] S. Jenkins, S.R. Kirk, Software architecture graphs as complex networks: a novel partitioning scheme to measure stability and evolution, Information Sciences 177 (12) (2007) 2587–2601.
[27] J. Ke, Y. Yao, Analysing language development from a network approach, Journal of Quantitative Linguistics 15 (1) (2008) 70–99.
[28] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5) (1999) 604–632.
[29] T. Kowaltowski, C.L. Lucchesi, J. Stolfi, Finite automata and efficient lexicon implementation, Technical Report IC-98-2, Universidade Estadual de Campinas, Campinas-Brazil, 1998, 12pp.
[30] J. Kupiec, J. Pedersen, F. Chen, A trainable document summarizer, in: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 68–73.
[31] L. Lam, C. Suen, A theoretical analysis of the application of majority voting to pattern recognition, Pattern Recognition 2 (1994) 418–420.
[32] J. Larocca Neto, A.A. Freitas, C.A.A. Kaestner, Automatic text summarization using a machine learning approach, in: Proceedings of the 16th Brazilian Symposium on Artificial Intelligence (SBIA), vol. 2507 of LNAI, 2002, pp. 205–215.
[33] J. Larocca Neto, A.D. Santos, C.A.A. Kaestner, A.A. Freitas, Document clustering and text summarization, in: Proceedings of the Fourth International Conference Practical Applications of Knowledge Discovery and Data Mining (PADD-2000), 2000, pp. 41–55.
[34] S. Lee, Y.J. Son, J. Jin, Decision field theory extensions for behavior modeling in dynamic environment using Bayesian belief network, Information Sciences 178 (10) (2008) 2297–2314.
[35] D.S. Leite, L.H.M. Rino, Selecting a feature set to summarize texts in Brazilian Portuguese, in: Proceedings of the International Joint Conference IBERAMIA-SBIA 2006, vol. 4140 of LNAI, 2006, pp. 462–471.
[36] D.S. Leite, L.H.M. Rino, T.A.S. Pardo, M.G.V. Nunes, Extractive automatic summarization: does more linguistic knowledge make a difference? in: Proceedings of the TextGraphs-2 HLT/NAACL Workshop, 2007.
[37] C.Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Proceedings of the Workshop on Text Summarization Branches Out (WAS), 2004.
[38] H.P. Luhn, The automatic creation of literature abstracts, IBM Journal of Research and Development 2 (1958) 159–165.
[39] I. Mani, Automatic Summarization, John Benjamins Publishing Co., 2001.
[40] I. Mani, E. Bloedorn, Machine learning of generic and user-focused summarization, in: Proceedings of the 15th National Conference on Artificial intelligence, 1998, pp. 821–826.
[41] I. Mani, E. Bloedorn, Summarizing similarities and differences among related documents, Information Retrieval 1 (1–2) (1999) 35–67.
[42] D. Marcu, Improving summarization through rhetorical parsing tuning, in: Proceedings of the Sixth Workshop on Very Large Corpora, 1998, pp. 206–215.
[43] R. Mihalcea, Language independent extractive summarization, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, 2005, pp. 49–52.
[44] S. Milgram, The small world problem, Psychology Today 2 (1967) 60–67.
[45] G.A. Miller, WordNet: a lexical database for English, Communications of the ACM 38 (11) (1995) 39–41.
[46] D.J. Murdoch, E.D. Chow, A graphical display of large correlation matrices, The American Statistician 50 (2) (1996) 178–180.
[47] J.L. Myers, A.D. Well, Research Design and Statistical Analysis, Lawrence Erlbaum, Mahwah, NJ, 2003.
[48] M.E.J. Newman, The structure and function of complex networks, SIAM Review 45 (2003) 167–256.
[49] M.G.V. Nunes, F.M.C. Vieira, C. Zavaglia, C.R.C. Sossolote, J. Hernandez, The construction of a lexicon for Brazilian Portuguese: learned lessons and perspectives, in: Proceedings of the Second Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR), 1996, pp. 61–70 (text in Portuguese).
[50] P. Over, W. Liggett, Introduction to DUC: an intrinsic evaluation of generic news text summarization systems, 2002. http://www-nlpir.nist.gov/projects/duc/pubs/2002slides/overview.02.pdf.
[51] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: bringing order to the Web, Technical Report, Stanford Digital Library Technologies Project, 1998, 17pp.
[52] C.D. Paice, The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases, in: Proceedings of the Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1981, pp. 172–191.
[53] T.A.S. Pardo, L. Antiqueira, M.G.V. Nunes, O.N. Oliveira Jr., L. da F. Costa, Modeling and evaluating summaries using complex networks, in: Proceedings of the Seventh Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR), vol. 3960 of LNAI, 2006, pp. 1–10.
[54] T.A.S. Pardo, L.H.M. Rino, TeMário: a corpus for automatic text summarization, Technical Report NILC-TR-03-09, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos, Brazil, 2003, 11pp. (text in Portuguese).

[55] T.A.S. Pardo, L.H.M. Rino, Description of GEI – generator of ideal extracts for Brazilian Portuguese, Technical Report NILC-TR-04-07, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos, Brazil, 2004, 8pp. (text in Portuguese).
[56] T.A.S. Pardo, L.H.M. Rino, M.G.V. Nunes, GistSumm: a summarization tool based on a new extractive method, in: Proceedings of the Sixth Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR), vol. 2721 of LNAI, 2003, pp. 210–218.
[57] T.A.S. Pardo, L.H.M. Rino, M.G.V. Nunes, NeuralSumm: a connexionist approach to automatic text summarization, in: Proceedings of the Fourth Brazilian Meeting on Artificial Intelligence (ENIA), 2003, pp. 1–10 (text in Portuguese).
[58] A. Ratnaparkhi, A maximum entropy model for part-of-speech tagging, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1996, pp. 133–142.
[59] L.H.M. Rino, M. Módolo, SuPor: an environment for AS of texts in Brazilian Portuguese, in: España for Natural Language Processing (EsTAL), 2004, pp. 419–430.
[60] L.H.M. Rino, T.A.S. Pardo, C.N. Silla Jr., C.A.A. Kaestner, M. Pombo, A comparison of automatic summarizers of texts in Brazilian Portuguese, in: Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA), 2004, pp. 235–244.
[61] T. Sakai, K. Spärck Jones, Generic summaries for indexing in information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2001, pp. 190–198.
[62] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.
[63] G. Salton, A. Singhal, M. Mitra, C. Buckley, Automatic text structuring and summarization, Information Processing and Management 33 (2) (1997) 193–207.
[64] M.D. Smucker, J. Allan, B. Carterette, A comparison of statistical significance tests for information retrieval evaluation, in: Proceedings of the 16th ACM Conference on Information and Knowledge Management, 2007, pp. 623–632.
[65] K. Spärck Jones, Automatic summarising: factors and directions, in: Advances in Automatic Text Summarization, MIT Press, 1999, pp. 1–12.
[66] K. Spärck Jones, Automatic summarising: the state of the art, Information Processing and Management 43 (6) (2007) 1449–1481.
[67] L. Vanderwende, H. Suzuki, C. Brockett, A. Nenkova, Beyond SumBasic: task-focused summarization with sentence simplification and lexical expansion, Information Processing & Management 43 (6) (2007) 1606–1618.
[68] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.
[69] C.C. Yang, F.L. Wang, Hierarchical summarization of large documents, Journal of the American Society for Information Science and Technology 59 (6) (2008) 887–902.
[70] D. Zajic, B.J. Dorr, J. Lin, R. Schwartz, Multi-candidate reduction: sentence compression as a tool for document summarization tasks, Information Processing & Management 43 (6) (2007) 1549–1570.