Physica A 391 (2012) 1855–1864

Contents lists available at SciVerse ScienceDirect

Physica A

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

Extractive summarization using complex networks and syntactic dependency

Diego R. Amancio [a,*], Maria G.V. Nunes [b], Osvaldo N. Oliveira Jr. [a], Luciano da F. Costa [a]

[a] Instituto de Física de São Carlos, Universidade de São Paulo, CP 369, PO BOX 13560-970, São Carlos, SP, Brazil

[b] Instituto de Ciências Matemáticas e de Computação Universidade de São Paulo, CP 668, 13560-970, São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 13 May 2011
Received in revised form 22 September 2011
Available online 25 October 2011

Keywords:
Summarization
Complex networks
Diversity metrics
Entropy
Syntactical dependency

ABSTRACT

The realization that statistical physics methods can be applied to analyze written texts represented as complex networks has led to several developments in natural language processing, including automatic summarization and evaluation of machine translation. Most importantly, so far only a few metrics of complex networks have been used and therefore there is ample opportunity to enhance the statistics-based methods as new measures of network topology and dynamics are created. In this paper, we employ for the first time the metrics betweenness, vulnerability and diversity to analyze written texts in Brazilian Portuguese. Using strategies based on diversity metrics, a better performance in automatic summarization is achieved in comparison to previous work employing complex networks. With an optimized method the Rouge score (an automatic evaluation method used in summarization) was 0.5089, which is the best value ever achieved for an extractive summarizer with statistical methods based on complex networks for Brazilian Portuguese. Furthermore, the diversity metric can detect keywords with high precision, which is why we believe it is suitable to produce good summaries. It is also shown that incorporating linguistic knowledge through a syntactic parser does enhance the performance of the automatic summarizers, as expected, but the increase in the Rouge score is only minor. These results reinforce the suitability of complex network methods for improving automatic summarizers in particular, and treating text in general.

1. Introduction

The concepts and methodologies of complex networks have been used in a variety of areas [1], including the analysis of scientific collaboration [2,3], epidemic spreading [4,5], World Wide Web [6] and Internet [7]. Complex networks have also been used for the analysis of written texts as Natural Language Processing (NLP) tools [8] (examples of NLP tools are the machine translators [9], the grammar checkers [10] and the automatic summarizers [11]). The adequacy of complex networks in this type of analysis has been demonstrated in several instances, since a text can be represented by a scale-free network [12–15]. Of particular relevance owing to the vast amount of electronic information available today, automatic extractive summarizers have become a popular NLP tool which is aimed at producing a reduced piece of text by selecting fragments (usually sentences) from a longer text [16]. Hence, the source text is compressed while retaining the most relevant information. Similarly to other NLP applications, extractive summarization has also benefited from the use complex network concepts. For instance, Antiqueira et al. [17] showed that using networks can lead to good results of informativeness, despite the use of only very limited linguistic knowledge.

* Corresponding author.
E-mail addresses: diegoraphael@gmail.com, diego.amancio@usp.br (D.R. Amancio).

0378-4371
doi:10.1016/j.physa.2011.10.015

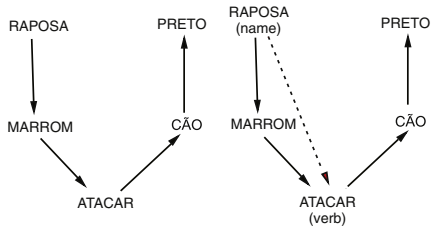

Fig. 1. Network created with the SUM-RC (left) and SUM-P (right) models. The dashed edge exemplifies the syntactic relationship between noun phrase and verb.

In this paper, we extend the study conducted in Ref. [17] proposing new metrics and new models to show that the recent measures designed to analyze networks can also identify relevant concepts in written texts. Additionally, we show the effect of adding deeper linguistic knowledge.

2. Methodology

2.1. Complex networks and texts

Complex networks are graphs whose structure follows complex principles of organization [18,19]. For instance, interesting patterns emerge when the World Wide Web [6] is modeled as a network, with the vertices represented by web pages and the edges being represented by hyperlinks. For the connections do not occur by chance [19], this and other systems such as the traffic in airports [20], roads [21] and electrical distribution [22] have peculiar properties, including a power law distribution, where many nodes have few connections while some nodes have many connections. These networks can also exhibit the small world property, characterized by shortest paths between any two nodes, and by the existence of communities [23], i.e. groups of highly interconnected nodes. Because the network formed with adjacent words in a text is also scale-free [12], the concepts and metrics of complex networks have been used in linguistics and Natural Language Processing [24]. For example, relying on works such as Ref. [13], who used networks to model lexical resources, Ref. [25] studied the association of words induced by humans. Similarly, Ref. [26] investigated the evolution of language by means of complex networks. A proposal for modeling texts as complex networks was presented in Ref. [27], which was then used for authorship attribution [28], assessment of the quality of essays [29] and of machine translation systems [30,31].

The model from Ref. [27] was adopted in this work under the name *SUM-RC*. In this model, the network corresponding to a text is built according to the adjacency between words. In a preprocessing stage, the text is lemmatized, i.e., the words with different inflected forms are grouped together so they can be examined as a single item. Moreover, the stopwords (articles and prepositions) are removed. Then, every distinct word is represented as a single node. The edges are obtained by joining nodes whose corresponding words are immediately adjacent (without considering limits of sentences or paragraphs). If an association is repeated, the weight of the corresponding edge is incremented by one. Thus, we obtain a weighted graph represented by a square matrix W where each W_{ji} is the number of associations $i \rightarrow j$ found in the text, provided that i and j are pairs of words immediately adjacent.

Because the incorporation of linguistic knowledge tends to improve the informativeness of summaries [32], we also employed a new model, named *SUM-P*. In particular, this adds to the *SUM-RC* model some edges related to syntactic dependency words, using the *PALAVRAS* automatic parser [33]. As an example of these models, Fig. 1 illustrates the network obtained from the *SUM-RC* (left) and *SUM-P* (right) models for the sentence "The brown fox attacks the black dog" ("A raposa marrom ataca o cão preto", in Portuguese), highlighting the syntactic edge. Besides the linguistic motivation, we used syntactic information because this type of construction also generates a complex network structure, as shown in Ref. [34]. Specifically, Ref. [34] proved that the syntactic dependence network in three fairly distant languages (Czech, German and Romanian) is complex, since it is scale-free and exhibits the small world property. Also, Ref. [34] has shown that these statistical patterns in the organization of the syntax seems to be language independent. Indeed, we could confirm that both properties also apply to Portuguese. Using a corpus collected from the "Zero Hora" Brazilian newspaper [35], which consists of $n = 7869$ distinct words (disregarding articles, prepositions and variations of number, gender and conjugation) we observed that the degree distribution of the corresponding network (see Fig. 2(a)) follows a power law with coefficient $\gamma = 2.48$.[1] The small world phenomenon was manifested in an average shortest path length $l = 3.81$ in the largest component (see distribution in Fig. 2(b)) being similar to the value expected for a random network $l_{rand} = \ln n / \ln\langle k\rangle = 5.48$. Furthermore, the clustering coefficient C (see distribution in Fig. 2(c)) of the network obtained for the texts in Portuguese is much larger than the $C_{rand} = \langle k\rangle / n(n-1) = 8.21 \times 10^{-4}$ for a corresponding random network, which confirms the presence of the small world feature.

[1] The fitting was performed with the cumulative distribution with logarithmic binning size, as suggested in Ref. [36].

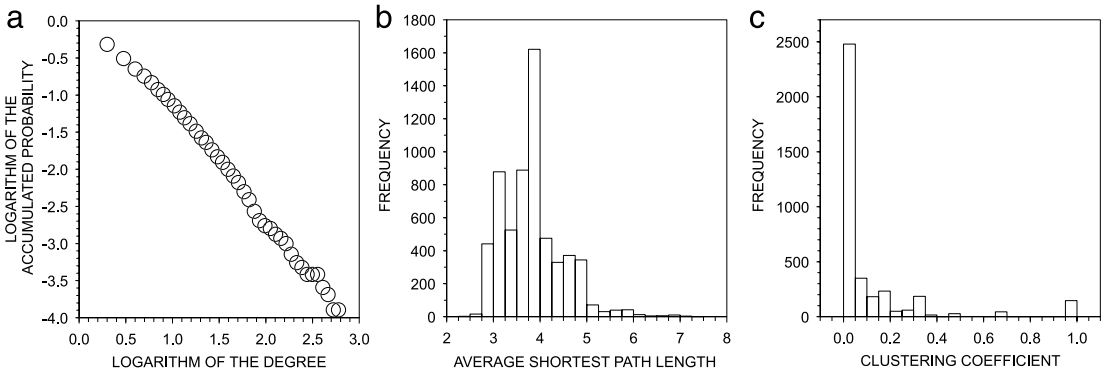

Fig. 2. (a) Accumulated probability distribution for the degree; (b) distribution of the average shortest path length and (c) distribution of clustering coefficient for the corpus in Portuguese.

2.2. Complex network measures

The topology of a complex network may be characterized by a number of metrics, and in this paper we applied some for the first time, including the diversity centrality [37], vulnerability [38] and betweenness [39]. The most used metric is the strength of a node, also known as connectivity in the Physics literature [1]. Although this metric has already been used in some summarization approaches [17], we decided to employ strength centrality in our methods to combine with the other new metrics. One recalls that there are two types of degree. The out-strength s_{out} corresponds to the sum of the edge weights leaving a node, while the in-strength of a node s_{in} is the sum of the edge weights that enter that node. If the average is taken considering all nodes, one obtains a global measure that differs from the local measure since the latter refers to a specific node. The global s_{out} and s_{in} are the same, since the sum of all edges that leave the nodes must be equal to the sum of all edges that enter the nodes. Using W, both the in- and out-strength are calculated as

$$s_{in}(i) = \sum_{j=1}^{N} W_{ij} \tag{1}$$

$$s_{out}(i) = \sum_{j=1}^{N} W_{ji}, \tag{2}$$

where N is the total number of nodes.

The centrality of each vertex is obtained by exploiting new ways to detect the border of a network. More specifically, the establishment of the so-called diversity metric quantifies how close a vertex is to the network borders, so that central and marginal vertices have high and low relevance, respectively. Simple, intuitive approaches to detect the border can quantify this property by means of algorithms that calculate the distance from the vertex analyzed to all leaves (leaves are vertices that satisfy the condition $s_{out} = 0$). With this definition, those nodes whose average distance to the leaves is relatively small are likely to be close to the border. However, this method is limited in that the average may have little meaning for distributions with large deviation. Novel methods address this problem by the use of entropy on self-avoiding random walks over the network, since a vertex close to the border will have little access to other vertices. An example of this methodology is illustrated in Fig. 3, where the variety of access expected for a fictitious network is shown. The inner vertices (white tones) have a greater homogeneity of possibilities to follow a random walk of a given length while the outer vertices (dark tones) have a greater diversity of opportunities to follow such paths, which features a low diversity. Thus, the definition of this metric indeed is able to detect central and marginal vertices through the analysis of random walks.

Formally, we define the diversity of a given vertex υ as the entropy of transition probabilities between neighboring vertices through random walks on the network. Let the diversity of the vertex under analysis be represented by δ_υ, i.e. the value that quantifies how different the access to this node is from random walks with length h (the path involves exactly h vertices), starting from the other $N - 1$ vertices. We then calculate δ_υ using Eq. (3), where NC_h represents the number of paths with length h starting at vertex υ and ending at vertex j. Eq. (4) gives the probability that the random walk follows the edge from vertex υ to vertex j.

$$\delta_\upsilon^h = -\frac{1}{\log(N-1)} \sum_{j=1}^{N} P_h(j,\upsilon) \cdot \log(P_h(j,\upsilon)) \tag{3}$$

$$P_h(j,\upsilon) = \sum_{c=1}^{NC_h} \prod_{(x,y)\in c} \frac{W_{yx}}{\sum_k W_{kx}}. \tag{4}$$

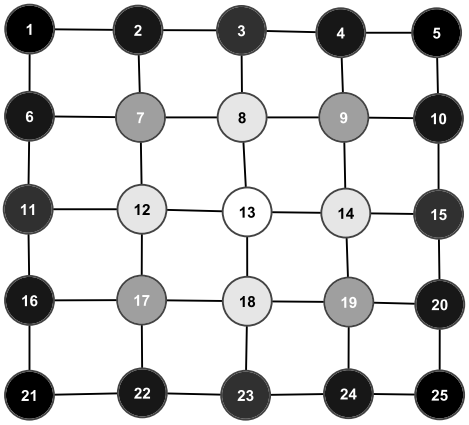

Fig. 3. Identification of borders using the metrics of diversity with $h = 3$. The darker the shade is, the lower the value of diversity.

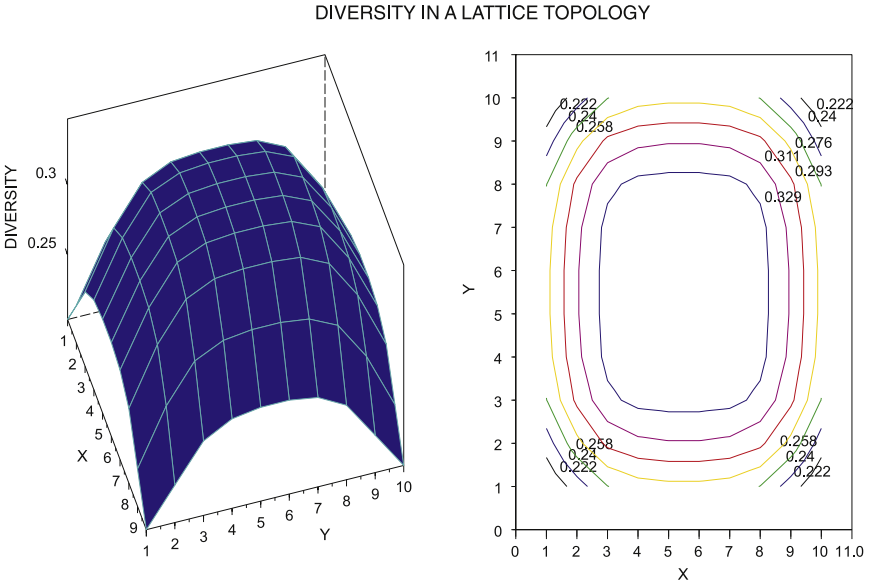

Fig. 4. For the diversity given as a function of the position of the vertex (left) and contour lines obtained from the diversity function (right), the internal vertices have the highest values.

In order to illustrate the ability of Eq. (3) to detect borders, we computed δ_v in a network comprising 100 vertices, arranged in a lattice topology (10×10 vertices), similar to the network topology illustrated in Fig. 3. The results are illustrated in Fig. 4 for $h = 6$. On the left, vertices are positioned on the xy-plane and the diversity of the vertices located at (x, y) is given in z-axis. In fact, the closer to the center (5, 5), the greater the diversity assigned to the vertex. The same conclusion can be inferred from the figure on the right, which corresponds to the contours obtained from the diversity function.

Another measure often used in network analysis is the global efficiency [40] GE, which is a geodesic metric. Although it is not used directly in the methods of summarization, we define here because it is an essential measure for the definition of vulnerability, as will be explained below. Specifically, GE can be defined as shown in Eq. (5). The interpretation of this measure is related to the velocity to exchange information between any two nodes, since a short distance d_{ij} contributes more significantly than a long distance. Note that the formula above prevents divergence; therefore, it is especially useful for networks with two or more components. The inverse of GE has also been used to characterize complex networks, under the name of *harmonic mean of geodesic distances*.

$$GE = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}. \tag{5}$$

As a centrality measure, we also computed the network vulnerability [38] for each vertex in the network. We use this measure as an additional strategy to find important words, since other connectivity measures may not be sufficient. For

instance, assuming a subgraph with topology close to a binary tree, the root of this tree would be a point of vulnerability (because it breaks the network into two components), even without being a hub (a highly connected vertex). To quantify the degree of relevance of a given vertex by means of vulnerability, we determine the variations in the network structure when the node is removed, i.e. when all the edges related to that vertex, besides the vertex, are removed. Here we consider the global efficiency as a performance measure to quantify the network variation so that the vulnerability for a node i is given in Eq. (6), where GE refers to the global efficiency of the network with all nodes and GE_i refers to the global efficiency of the network after removing vertex i. Unlike other measures, the global measure is not the average obtained from the vulnerability of each node. In this case, it is important to know the worst case, as Eq. (7) shows:

$$V_i = \frac{GE - GE_i}{GE} \tag{6}$$

$$V = \max_i V_i. \tag{7}$$

In addition to GE, we employed two other metrics related with geodesical distances: the shortest path and betweenness centrality [39]. A shortest path is simply a path connecting two nodes with minimum distance, taking into account the sum of edge weights traversed by this path. This shortest path, which is not necessarily unique, is calculated from a given fixed node to all other nodes. As a global measure, one defines the shortest path as the average of the lengths of all minimal paths between any two vertices. Locally, one defines the shortest path l from one vertex as in Eq. (8).

$$l_i = \frac{\sum_{j \neq i} d_{ij}}{N - 1}. \tag{8}$$

While l uses the length of paths, the betweenness uses the number of shortest paths. The betweenness centrality for vertex v is defined in Eq. (9), where the numerator brings the number of shortest paths passing through vertices i, v and j and the denominator is the number of shortest paths passing through vertices i and j. In other words, a node will have a high betweenness centrality if there are many shortest paths passing through it.

$$B_v = \sum_i \sum_j \frac{\sigma(i, v, j)}{\sigma(i, j)}. \tag{9}$$

2.3. Automatic extractive summarizers

Metrics of complex networks have been used to build automatic summarizers [17], which serves as the basis for the present work. We now use additional metrics, as discussed below.

2.3.1. Diversity

Four methods based on diversity were used, which differed by the length h of the self-avoiding random walk computed in the diversity metric. For h ranging from 1 to 3 we created the methods *DIV1*, *DIV2* and *DIV3*, respectively, in addition to the *DDM* method that takes into account the average h value. In the first step, the number of sentences of the summary is estimated according to the desired compression ratio (τ). Then, for each node (word), one calculates the diversity. Next, a weighting ω_s is assigned to each sentence s, according to the weight of each word ρ_i^s in s, as illustrated in Eq. (10), where η_s is the number of words in the sentence s.

$$\omega_s = \frac{1}{\eta_s} \sum_{i=1}^{\eta_s} \delta^x_{\rho_i^s}. \tag{10}$$

Next, the values of ω_s are sorted in descending order to determine the cutoff threshold as the value at position $(1 - \tau) \cdot \Theta$ in the sorted array, where Θ is the size of the original text (in number of sentences). Finally, the sentences of the original text are traversed one by one in the same order they appear in the original text. If a sentence has ω_s above the threshold, it will be added to the summary. Here one assumes that sentences with a high mean diversity also have a great chance to convey key concepts and probably provide a good level of informativeness to the summary. This seems reasonable since in our preliminary experiments (see Discussion and Results section) high levels of diversity appear to be correlated with the keywords of the text.

2.3.2. Diversity and strength

For this method, identified as *DDG*, both diversity and strength are calculated for each word individually. Basically, we attempt to take advantage of quantifying the centrality by using both measures. The same initial steps described for the diversity method (see last subsection) are used to obtain two sorted vectors according to the diversity and strength. Next, we verify the classification of each sentence in these vectors to select those appearing in the best positions of the two vectors, i.e., we choose the words with high values of strength and diversity centrality. Then, a new weighting ω'_s is assigned to each

sentence: if the sentence s is placed at position ρ_1 in the first vector (relative to the diversity) and at position ρ_2 in the second vector (relative to the strength), then ω_s' will be simply taken as described in Eq. (11). Finally, we select the sentences with the lowest values of ω_s', since the smaller the weighting, the greater the relevance (according to these two metrics used). For instance, if a sentence is placed in the first and third positions, while another is placed in the second and fourth positions, the former is considered more important because it is better classified in the two lists.

$$\omega_s' = \rho_1 + \rho_2. \tag{11}$$

2.3.3. Shortest paths

This method, identified as *SP*, is similar to the diversity method, but now we employ the shortest paths to quantify relevance. A vertex υ will be considered as relevant when a shortest path starting at υ is short, i.e., it is relatively easy to reach any other vertex in the network. This approach is justifiable from the standpoint that vertices relatively close to others can be *hubs* or at least can be close to *hubs*. Thus, it is likely that such vertices represent concepts with high informativeness.

2.3.4. Betweenness

As in the previous method, the method referred to as *BTW* uses a centrality measure (betweenness) to find keywords. Analogously to other traditional techniques of centrality, this metric is based on the calculation of shortest paths to classify relevance of vertices. In this case, a vertex is a candidate for a keyword if its betweenness is high; that is to say, if there are many shortest paths passing through it. Vertices with high betweenness values tend to be more accessible by efficient paths. Therefore, this method selects as keywords those words with a tendency to shorten the distance between concepts. Then, after measuring centrality, sentences are chosen similarly to the diversity method cited previously, using betweenness instead of diversity to compute ω_s, in Eq. (10).

2.3.5. Vulnerability

This method (*VUL*) uses the vulnerability measure locally in order to establish a relationship between relevance and vulnerability centrality. From the definition of vulnerability (see section on complex network measures), this method assumes that a word is important if removing the corresponding vertex reduces the global network efficiency. Therefore, vertices that maintain the network structure or are articulation points [41] will have high vulnerability scores and will be classified as keywords. Similarly to the Betweenness and Diversity methods, this method uses ω_s to detect the relevance of sentences, according to the vulnerability computed for each word in the sentence examined.

2.4. Assessing summary informativeness

Assessing the quality of extractive summaries is as difficult as the summarization process itself. Although several evaluation methods exist, there has been no consensus about what is best. One of the difficulties is the need to evaluate the summaries according to various criteria, including informativeness, coherence, cohesion, readability, grammaticality and textuality. Another approach for the assessment is to employ automatic metrics, which are advantageous for being more standardized. However, this method does not allow evaluation of several of the criteria mentioned. Among the most commonly used metrics are the recall Γ and precision ς [42], defined as the percentage of sentences belonging to the reference summary that appears in the summary generated and the percentage of sentences belonging to the generated summary in the reference summary. The so-called F-measure (or $F1$ score) is also very popular, primarily as a weighting factor between recall and precision:

$$\Phi = \frac{2 \cdot \Gamma \cdot \varsigma}{\Gamma + \varsigma}. \tag{12}$$

Note that, according to Eq. (12), the F-measure Φ is 1 at best and 0 at worst.

In this paper, the metrics used are precision, recall and F-measure available on the package for automatic evaluation ROUGE [42], based on the co-occurrence of units (n-grams) between automatically created summaries and reference summaries. We chose these metrics because they show strong correlation with human evaluation [42].

The summarizers were evaluated using the *TeMario* corpus. The texts comprising this corpus are divided into two sets, according to the purpose: source texts from which the summaries are built, and reference texts that are useful for evaluating the generated summary. The corpus was built especially to assist the automatic summarization task and contains over 60,000 words. The source texts comprise a set of 100 newspaper articles, 60 of which belong to the Brazilian newspaper Folha de São Paulo [43] (online version). The other 40 texts were extracted from the newspaper Jornal do Brasil (Brazilian Newspaper) [44]. The topics covered in the texts were varied, as they were selected from various sections: World News, Editorials and Special Reports (Folha de São Paulo), Politics and International News (Jornal do Brasil). The choice of this genre to compose the corpus is due to its linguistic heterogeneity, since these newspapers have a wide readership. Moreover, the journalistic genre is widely used in automatic summarization evaluation in international competitions, such as TAC (Text Analysis Conference) [45].

Table 1
ROUGE-1 values for methods based on the conventional model (SUM-RC).

Method	Recall	Precision	*F*-measure
DIV1	0.5085	0.3847	0.4305
DIV3	0.5058	0.3779	0.4245
DIV2	0.5046	0.3810	0.4265
DM3	0.5045	0.3796	0.4254
CLS	0.5032	0.3761	0.4228
DDG	0.4983	0.3942	0.4323
BTW	0.4954	0.3859	0.4268
VUL	0.4882	0.3766	0.4178

Table 2
ROUGE-1 values for methods based on the SUM-P model.

Method	Recall	Precision	*F*-measure
DIV1	0.5089	0.3865	0.4318
DIV3	0.5073	0.3815	0.4272
DIV2	0.5051	0.3722	0.4210
CLS	0.5047	0.3800	0.4253
DM3	0.5041	0.3816	0.4265
DDG	0.4989	0.3899	0.4304
BTW	0.4901	0.3834	0.4223
VUL	0.4887	0.3807	0.4202

3. Results and discussion

3.1. Assessing the quality of summaries

The extractive summarizers were built using both models (*SUM-RC* and *SUM-P*) and compared in terms of quality, first within each model. Additionally, we quantified the correlation between summarizers, in order to see similarities between summaries generated within a single model. Table 1 displays the ROUGE-1 values for the conventional model, while Table 2 shows the ROUGE-1 values for the syntactic model. Both tables show high scores for the methods based on the diversity metrics (*DIV1*, *DIV2* and *DIV3*), which indicates that this metric is not only suitable to detect borders in complex networks [46], but also to identify informative sentences in texts. Significantly, several of the methods based on diversity led to recall values for ROUGE-1 higher than the best value ever achieved [17] (the best recall value for summarizers based on complex networks found in Ref. [17] was 0.5031). In contrast, the summarizing methods employing vulnerability, closeness and betweenness led to poor results, even when syntactic information was used.

Also worth mentioning is the fact that the performance of the summarizers was enhanced when linguistic knowledge is incorporated, but this enhancement was very small. By adding only the syntactic information, the recall values were smaller than those obtained with summarizers employing deep linguistic knowledge (reaching 0.58 [32]). It is concluded that there is room for improvement on the statistical methods, for instance in using hybrid approaches, but significant improvement in performance may require sophisticated linguistic treatment.

In our experiments, we also examined the correlation between summarizers, to verify whether there is redundancy on the information provided by the different metrics. The correlation between the rank of sentences generated was calculated using the Spearman correlation coefficient [47]. This coefficient is maximum (equal to 1) when both methods rank the sentences to compose the summary equally, and is minimal (equal to -1) when the ranks resulting from the sorting process is reversed (the main sentence for one is the least important for the other and so on). The correlation was obtained: (i) between summarizers belonging to the *SUM-RC* model, (ii) between summarized belonging to the *SUM-P* model and (iii) between summarizers belonging to both models. The results are shown in Figs. 5–7. A visual inspection of the first two figures indicates a high correlation for the methods based on diversity (*DIV* 1, *DIV* 2, *DIV* 3, and *DM* 3) as well as those based on geodesic distances (*CLS*, *BTW* and *VUL*) are related to each other. The DDG method does not seem to correlate strongly with any other method, probably because it is a hybrid approach that considers diversity and geodesic metrics. Therefore, there are actually three types of summarizers: the geodesic-based version, the diversity-based version and the hybrid version. On the other hand, analyzing the correlation between summarizers in Fig. 7, one sees that the changes introduced by "syntactic edges" are indeed small, since all correlation values are above 0.90.

3.2. Identifying keywords through diversity centrality

Motivated by the excellent results obtained with diversity in the previous section, we applied this metric to determine keywords in texts on a single theme. Specifically, this experiment used a set of 300 essays produced by students in a national Brazilian exam, which were written based on the following question: "*The right to vote: how to make this a tool to promote*

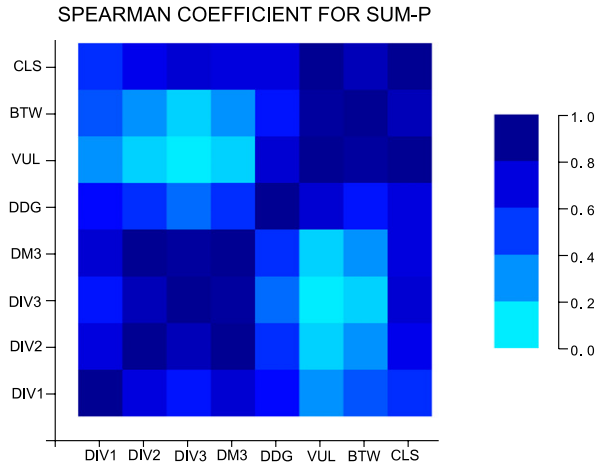

Fig. 5. Spearman coefficient for the SUM-P model.

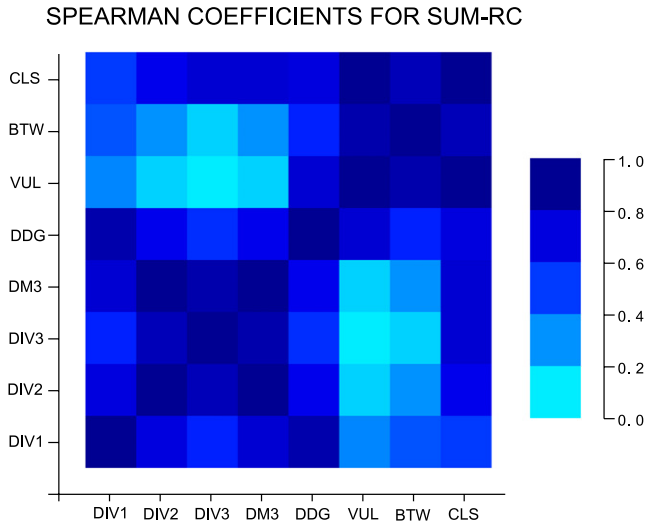

Fig. 6. Spearman coefficient for the SUM-RC model.

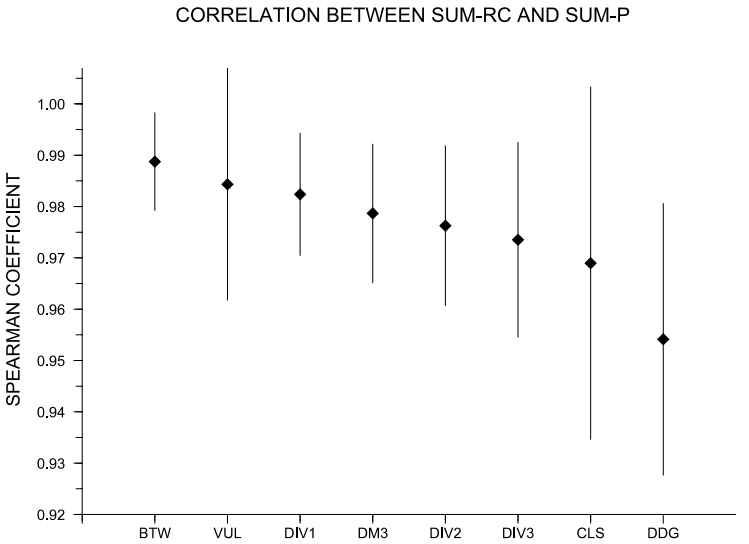

Fig. 7. Spearman coefficient for the SUM-RC and SUM-P models.

Table 3
The number of times a given word was considered among the 10% most important words when the metric of diversity is employed.

Word	Frequency	Word	Frequency
To be (ser)	732	To do	132
Not	447	Person	104
To vote	407	Candidate	104
To have	384	To be (estar)	90
Vote	379	To go	80
Power	255	Politico	84
Country	246	Best	63
Right	246	Brazilian	58
Duty	162	People	58
Brazil	162	To know	57

Table 4
The number of times a given word was considered among the 10% less important words when the diversity metric is employed.

Word	Frequency	Word	Frequency
Year	125	Now	62
Well	99	Before	61
Citizen	82	To analyze	61
To help	74	Conquest	60
To finish	72	Weapon	52
Good	72	To happen	52
Still	66	Through	50
To believe	64	Thing	45

the social changes that Brazil needs ?''. Thus, it is expected that the key concepts are semantically linked to the voting process, elections, candidates and so on.

The strategy developed here identifies for each text and for each h value (ranging from 1 to 3) the words – i.e. the vertices – with the highest (top 10%) and lowest (bottom 10%) values of diversity. For each text, a given word may appear at most h times within the 10% highest values, since this calculation is performed for each level. Therefore, the frequency of each word will lie in the range between 0 and 900, since the count is performed on the corpus of 300 essays.

The most frequent words in the list of high diversity are summarized in Table 3, which displays words related to the theme in addition to the high frequency words in the language, as the verbs ''to be'', ''to have'', ''to do'' and ''to go''. Table 4 displays the most common words in the list of lowest diversity values. It may be noted that most of the words are weakly correlated with the theme in the essays, which confirms that the vertices near the edge have little importance.

4. Conclusion

In this paper we have extended the use of complex networks for developing extractive summarizers. Upon testing several new metrics, we found that vulnerability, closeness and betweenness were not appropriate for selecting the sentences for the summarizers. In contrast, diversity metrics proved excellent for this selection, with which summarizers were obtained with the highest Rouge scores for extractive summarizers based on complex networks, and competitive with the best summarizers available. Another important contribution was related to the introduction of linguistic information for processing the text. As expected [32], adding syntactic information did enhance the summarizers performance, but the increase in the Rouge scores was only minor. Therefore, while there is room for improvement with the use of deep linguistic features, one needs to know that significant enhancement in performance may require sophisticated treatment (e.g. semantic information), especially as the performance of the summarizers based on diversity is already close to those using deep linguistic information. The good results obtained with the diversity metrics motivated us to use them in identifying keywords in a small corpus focusing on a single theme. Indeed, words with high diversity could be selected as keywords, and this opens new possibilities for the application of complex network concepts in natural language processing and other areas requiring the establishment of taxonomies and ontologies.

Acknowledgments

This work was supported by FAPESP and CNPq (Brazil).

References

[1] L.da F. Costa, O.N. Oliveira Jr., G. Travieso, F.A. Rodrigues, P.R. Villas Boas, L. Antiqueira, M.P. Viana, L.E.C. Rocha, Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications, Physics and Society, 2008.
[2] M.E.J. Newman, Scientific collaboration networks: network construction and fundamental results, Physical Review E 64 (2001).

[3] M.E.J. Newman, Scientific collaboration networks: shortest paths, weighted networks, and centrality, Physical Review E 64 (2001).
[4] M. Boguñá, R. Pastor-Satorras, Epidemic spreading in correlated complex networks, Physical Review E 66 (2002).
[5] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical Review Letters 86 (2001) 3200–3203.
[6] A.-L. Barabási, R. Albert, H. Jeong, Scale-free characteristics of random networks: the topology of the world wide web, Physica A 281 (2000) 69–77.
[7] A. Vázquez, R. Pastor-Satorras, A. Vespignani, Large-scale topological and dynamical properties of the internet, Physical Review E 65 (2002).
[8] M. Bates, Models of natural language understanding, Proceedings of the National Academy of Sciences of the United States of America 92 (1995) 9977–9982.
[9] W.J. Hutchins, H.L. Somers, An Introduction to Machine Translation, Academic Press, 1992.
[10] R.T. Martins, R. Hasegawa, M.G.V. Nunes, G. Montilha, O.N. Oliveira Jr., Linguistic issues in the development of ReGra: a grammar checker for Brazilian Portuguese, Natural Language Engineering 4 (1998) 287–307.
[11] D. Marcu, The theory and practice of discourse parsing and summarization, in: A Bradford Book, The MIT Press, 2000.
[12] R.F. Cancho, R.V. Solé, The small world of human language, Proceedings of The Royal Society of London, Series B, Biological Sciences 268 (2001) 2261–2265.
[13] M. Sigman, G.A Cecchi, Global organization of the wordnet Lexicon, Proceedings of the National Academy of Sciences 99 (2002) 1742–1747.
[14] G. AMiller, Wordnet: a dictionary browser, Proceedings of the First International Conference on Information in Data, University of Waterloo, 1985.
[15] A.E. Motter, A.P.S. Moura, Y.C. Lai, P. Dasgupta, Topology of the conceptual network of language, Physical Review E 65 (2002).
[16] K.J. Spärck, Automatic summarising: factors and directions, in: Advances in Automatic Text Summarization, MIT Press, 1999, pp. 1–12.
[17] L. Antiqueira, O.N. Oliveira Jr., L.da F. Costa, M.G.V. Nunes, A complex network approach to text summarization, Information Sciences 179 (2009) 584–599.
[18] A.L. Barabási, Linked: How Everything is Connected to Everything Else and What it Means for Business, Science, and Everyday Life, Plume, New York, 2003.
[19] M.E.J. Newman, The structure and function of complex networks, SIAM Review 45 (2003) 167–256.
[20] R. Guimerà, S. Mossa, A. Turtschi, L.A.N. Amaral, The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles, Proceedings of the National Academy of Science USA 102 (2005) 7794–7799.
[21] M. Rosvall, A. Trusina, P. Minnhagen, K. Sneppen, Networks and cities: an information perspective, Physical Review Letters 94 (2005).
[22] B.A Carreras, V.E. Lynch, I. Dobson, D.E. Newman, Complex dynamics of blackouts in power transmission systems, Chaos: An Interdisciplinary Journal of Nonlinear Science 14 (2004) 643–652.
[23] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Science of the United States of America 99 (2002) 7821–7826.
[24] D. Jurafsky, J.H. Martin, Speech and language processing: an introduction to natural language processing, in: Computational Linguistics and Speech Recognition, Prentice Hall, New Jersey, 2000.
[25] L.da F. Costa, What's in a name? International Journal of Modern Physics C 15 (2004) 371–379.
[26] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks, Advances in Physics 51 (2002) 1079–1187.
[27] L. Antiqueira, M.G.V. Nunes, O.N. Oliveira Jr., L.da F. Costa, Modelando textos como redes complexas. in: Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana, 2005, pp. 1–10.
[28] L. Antiqueira, T.A.S. Pardo, M.G.V. Nunes, O.N. Oliveira Jr., Some issues on complex networks for author characterization, Revista Iberoamericana de Inteligencia Artificial 11 (2007) 51–58.
[29] 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.
[30] 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 (2008) 583–598.
[31] D.R. Amancio, M.G.V. Nunes, O.N. Oliveira Jr., T.A.S.L. Antiqueira, L.da F. Costa, Using metrics from complex networks to evaluate machine translation, Physica A 390 (2011) 131–142.
[32] D. Leite, L. Rino, T.A.S. Pardo, M.G.V. Nunes, Extractive automatic summarization: does more linguistic knowledge make a difference? in: TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, 2007.
[33] E. Bick, The parsing system Palavras — automatic grammatical analysis of portuguese in a constraint grammar framework, Arhus, 2000, Conference HLT-NAACL-2003, Edmonton, Canada, 2003.
[34] R. Ferrer i Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Physical Review E 69 (1–8) (2004).
[35] http://zerohora.clicrbs.com.br.
[36] H. Bauke, Parameter estimation for power-law distributions by maximum likelihood methods, European Physical Journal B 58 (2007).
[37] M.P. Viana, B.A.N. Travençolo, E. Tanck, L.da F. Costa, Characterizing topological and dynamical properties of complex networks without border effects, Physica A 389 (2010) 1771–1778.
[38] V. Latora, M. Marchiori, Vulnerability and protection of critical infrastructures, Physical Review E 71 (2005) 015103R.
[39] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41.
[40] V. Latora, M. Marchiori, Efficient behavior of small-world networks, Physics Review Letters 87 (2001).
[41] R. Sedgewick, Algorithms in C, Part 5: Graph Algoritmhs, 3rd ed, 1998.
[42] C.Y. Lin, E. Hovy, Automatic evaluation of summaries using *n*-gram co-occurrence statistics, in: Proceedings of the 2003 Language Technology, 2003.
[43] http://www.folha.com.br.
[44] http://jbonline.terra.com.br.
[45] http://www.nist.gov/tac/.
[46] B.A.N. Travencolo, M.P. Viana, L.da F. Costa, Border detection in complex networks, New Journal of Physics 11 (2009).
[47] C. Spearman, The proof and measurement of association between two things, American Journal of Psychology 15 (1904) 72–101.