Complex Networks and Extractive Summarization*,**

Lucas Antiqueira[1] and Maria das Graças Volpe Nunes[2]

[1] Instituto de Física de São Carlos, Universidade de São Paulo,
PO Box 369, Postal Code 13560-970, São Carlos, SP, Brazil
lantiq@ursa.ifsc.usp.br

[2] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo,
PO Box 668, Postal Code 13560-970, São Carlos, SP, Brazil
gracan@icmc.usp.br

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 work 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, as a well-established subfield of natural language processing, is relevant for a number of scenarios (Spärck Jones, 2007). A summary can be seen as a condensed representation of a source text that maintains the important information of its original counterpart. When a summary is constructed by selecting and juxtaposing source pieces, such as sentences, it is called an *extract*. Extractive summarizers usually do not require deep linguistic

* Authorized short version of (Antiqueira et al., 2009), published in the journal *Information Sciences* (Copyright Elsevier).

** Based on the MSc dissertation authored by L. Antiqueira under the supervision of M.G.V. Nunes (Antiqueira, 2007).

knowledge to select the most relevant pieces of the source text for an extract. Although extractive summarizers are likely to produce texts with cohesion and coherence problems, many systems have been proven to yield summaries whose informative level is satisfactory. This is particularly true when the extract is used as a component of another system – e.g. in information retrieval – and is not directly used by humans.

A graph, or network, is a representation that may capture text structure in various ways, being therefore suitable for extractive summarization. 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 work, 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 since the small-world and scale-free properties were identified in many real-world networks about 10 years ago (Albert and Barabási, 2002). The recent discoveries contributed significantly to elucidate the structure and dynamics of diverse real-world entities such as the natural languages (Ferrer i Cancho and Solé, 2001), including some applications to natural language processing (Antiqueira et al., 2007; Amancio et al., 2008).

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, the proposed method selects a subset of sentences (nodes) to compose an extract by ranking them according to some network measurement. This work can therefore be placed among others that employ graph algorithms for automatic summarization (e.g. Mihalcea, 2005). We produced generic summaries, i.e. neither user-specific nor topic-oriented, for newspaper articles in Brazilian Portuguese. Experiments to evaluate the informativeness level of the extracts were carried out, and the resulting scores for ROUGE-1 and Precision/Recall were compared with the scores of other summarizers previously evaluated within the same experimental setup.

2 The CN-Summ Framework

The technique presented here for automatically generating extracts is based on a set of network measurements typically applied to characterize complex networks (Costa et al., 2007). 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 lemmatized noun in common (i.e. lexical repetition). We also argue that if two sentences are connected in this network they probably convey complementary information. This reflects our everyday experience on reading and writing: we rarely create 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* (**C**omplex **N**etworks-based **Summ**arization), consists of four steps: (i) text pre-processing (sentence delimitation and lemmatization of nouns), (ii) network construction connecting sentences that share at least one lemmatized noun, (iii) node quantification through network measurements (sentence ranking) and (iv) selection of the best ranked sentences to compose the extract. In what follows we introduce the main ideas behind CN-Summ strategies.

1. **Degree Strategies**: 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. If edge weights are considered (i.e. frequency of noun co-occurrences), a slightly different type of degree, referred to as strength, is obtained by summing up all edge weights associated to a given node. A sentence that shares a large number of links with other sentences probably conveys relevant information that complements many other sentences. The sentences with the highest degree are selected to build an extract in the first summarization strategy - called CN-Degree. Similarly, a strategy using the strength - called CN-Strength - was also created.
2. **Shortest Path Strategies**: 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, a path considers not only the immediate neighbors of a node, but also the nodes indirectly connected to it. We take the mean length of the shortest paths that associate a given node to every other node in the network as a measurement of its overall accessibility. Since small path lengths indicate that a node is close to other nodes on average, we now define a summarization strategy that selects for the extract the n nodes with the lowest average path lengths. We developed three different variations, identified by CN-SP, CN-SP^{wc} and CN-SP^{wi} – for detailed information on these strategy variations, please refer to (Antiqueira et al., 2009).
3. **Locality Index Strategy**: The locality index is a measurement that takes into account the pattern of connectivity in the neighborhood of a node. The locality index is useful for pointing out central nodes of relatively isolated groups, which can represent important sentences that summarize the meaning of their neighbors. 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. Hence, another strategy for summarization, called CN-LI, gives priority to the central nodes of the groups mentioned, by selecting the top sentences sorted in decreasing order of locality index.
4. ***d*-Ring Strategies**: A d-ring of a node i is formed by the nodes distant d edges from i. It 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, and then select nodes of its nearby d-rings. This is the basis of the following three strategies: CN-Rings^{l}, CN-Rings^{k} and

CN-Rings^{lk} – please, refer to (Antiqueira et al., 2009) for more details on these variations.

5. ***k*-Core Strategies**: A subgraph g of a graph G is a k-core if every node i of g has degree at least equal to k. This subgraph must also be the greatest subgraph of G that has this property. Notice that a non empty k-core with the maximum possible k, called the innermost k-core, is a subgraph that consists of densely connected nodes. Therefore, we 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 nodes belonging to the innermost k-core, with further sentences being added to the extract by sequentially relaxing the k-core (i.e. by decreasing k). Two variations are proposed, CN-Cores^{l} and CN-Cores^{k} – see (Antiqueira et al., 2009) for detailed information.
6. ***w*-Cut Strategies**: 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 be a subgraph whose edge weights are not lower than w. 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}.
7. **Community Strategy**: Another concept borrowed from the complex networks field is the notion of communities, which correspond to groups of nodes highly interconnected, while different groups are scarcely connected to each other. 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.
8. **A Voting Strategy**: It is known that the combination of methods using a voting scheme can improve individual performances. 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. 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.

3 Informativeness Results and Discussion

Two evaluation experiments were carried out using TeMário corpus (Pardo and Rino, 2003), which comprises 100 newspaper articles in Brazilian Portuguese. For each text there is a pair of reference summaries: an abstract written by a human

Table 1. Average Precision $\langle P \rangle$, Recall $\langle R \rangle$ and F-measure $\langle F \rangle$ scores, in percentages (%).

Systems	$\langle P \rangle$	$\langle R \rangle$	$\langle F \rangle$
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

and an automatically generated extract (created using the contents of the human abstract). Our experiments also compared CN-Summ with other extractive summarizers previously evaluated with the same corpus. In order to illustrate the type of extracts obtained with CN-Summ, two examples were included in a publicly available document[1].

3.1 First Experiment

The first experiment uses the reference extracts to compute Precision, Recall and F-measure (Salton and McGill, 1983) (see Table 1). Extracts were obtained by removing 70% of the source sentences. In this experiment, the 15 versions of CN-Summ are compared with two baselines (the well-known Top and Random Baselines) and six other extractive systems (ClassSumm, NeuralSumm, GistSumm, TF-ISF-Summ, SuPor and SuPor-v2) – see Leite and Rino (2006) and Rino et al. (2004). Some of CN-Summ versions are among the best systems for Portuguese summarization, with F-measure around 42%. Considering only our systems, the

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

best results were produced, as expected, by CN-Voting. 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. A remarkable result is that all CN-Summ versions outperformed TF-ISF-Summ, GistSumm and NeuralSumm, systems that also work with shallow linguistic resources. For a statistical significance analysis of these average results, we performed paired t-tests between average F-measures. CN-Summ strategies can be classified into two groups of statistically similar systems, viz.: systems with average F-measure higher or lower than 40%. We were able to collect the full evaluation data for SuPor-v2, thus allowing its comparison with our summarizers. In this experiment, SuPor-v2 is statistically equivalent to CN-Voting, CN-SP^{wc}, CN-Rings^{k}, CN-Degree and CN-SP^{wi}, i.e. some of our strategies are among the best scoring systems in this evaluation.

3.2 Second Experiment

In the first experiment we could not compare the performance of CN-Summ with some other systems for Brazilian Portuguese (Leite et al., 2007; Mihalcea, 2005), since the latter were evaluated in a different environment. We also applied the unigram-based recall metric ROUGE-1 to evaluate summary informativeness (Lin, 2004), taking as reference the human-made abstracts of TeMário corpus. The compression rate now ensures that each automatic extract has approximately the same number of words of the reference abstract. Table 2 shows the average ROUGE-1 scores for CN-Summ, for the two baselines and six other extractive systems: SuPor-v2 (Leite et al., 2007), the best three variations of Mihalcea's method (Mihalcea, 2005), namely PageRank Backward, HITS$_{\mathrm{A}}$ Backward and HITS$_{\mathrm{H}}$ Forward, in addition to two modified versions of Mihalcea's PageRank Undirected, called TextRank+Thesaurus and TextRank+Stem+StopwordsRem (Leite et al., 2007). For short, the last two are henceforth called TextRank+T and TextRank+S+S, respectively.

Once again, some of CN-Summ strategies are close to the top-scoring systems. CN-Voting is the best of our systems (on average), with a score of 0.5031. In general, the variations of the strategies based on degrees, shortest paths, d-rings and k-cores consistently show good performances in both experiments. SuPor-v2 has the best scores in both experiments, probably because of its application of deep linguistic knowledge. 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). For lack of the full data for some systems, we could not include them all in the statistical significance analysis. The t-tests are complemented by the confidence intervals generated by the software ROUGE, which are shown in Table 2. The majority of summarizers are not significantly different, as the overlap of confidence intervals in Table 2 shows. Significant differences were only found between a few top-scoring and a few low-scoring systems. Ultimately, the t-tests demon-

Table 2. Average ROUGE-1 scores $\langle RG1 \rangle$ and 95% confidence level intervals where available.

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$_{\mathrm{A}}$ Backward	0.5002	-
10 HITS$_{\mathrm{H}}$ 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]

strate that in this experiment almost all CN-Summ strategies, from CN-Voting to CN-Communities, are not significantly different from each other.

4 Final Remarks

Although some of the CN-Summ versions performed as well as the best summarizers for Brazilian Portuguese, the definition of the network is extremely important, and could increase the performance of CN-Summ if improved. Other developments may consider joining all CN-Summ strategies in a machine learning approach. 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.

Acknowledgments

The authors wish to thank CNPq and FAPESP for financial support.

Bibliography

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

Amancio, D. R., Antiqueira, L., Pardo, T. A. S., Costa, L. da F., Oliveira Jr., O. N., and Nunes, M. G. V. (2008). Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, 19(4):583–598.

Antiqueira, L. (2007). Development of techniques based on complex networks for extractive text summarization (text in Portuguese). ICMC-USP, São Carlos, Brazil. `http://www.teses.usp.br/teses/disponiveis/55/55134/tde-26042007-145428/`. 108 pp.

Antiqueira, L., Nunes, M. G. V., Oliveira Jr., O. N., and Costa, L. da F. (2007). Strong correlations between text quality and complex networks features. *Physica A*, 373:811–820.

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

Costa, L. da F., Rodrigues, F. A., Travieso, G., and Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.

Ferrer i Cancho, R. and Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B*, 268:2261.

Leite, D. S. and Rino, L. H. M. (2006). Selecting a feature set to summarize texts in Brazilian Portuguese. In *Proceedings of the International Joint Conference IBERAMIA-SBIA 2006*, volume 4140 of *LNAI*, pages 462–471.

Leite, D. S., Rino, L. H. M., Pardo, T. A. S., and Nunes, M. G. V. (2007). Extractive automatic summarization: Does more linguistic knowledge make a difference? In *Proceedings of the TextGraphs-2 HLT/NAACL Workshop*.

Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS)*.

Mihalcea, R. (2005). Language independent extractive summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 49–52.

Pardo, T. A. S. and Rino, L. H. M. (2003). TeMário: a corpus for automatic text summarization (text in Portuguese). Technical Report NILC-TR-03-09, NILC-USP, São Carlos, Brazil. 11 pp.

Rino, L. H. M., Pardo, T. A. S., Silla Jr., C. N., Kaestner, C. A. A., and Pombo, M. (2004). A comparison of automatic summarizers of texts in Brazilian Portuguese. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA)*, pages 235–244.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Spärck Jones, K. (2007). Automatic summarising: the state of the art. *Information Processing and Management*, 43(6):1449–1481.