# Modeling and Evaluating Summaries Using Complex Networks

Thiago Alexandre Salgueiro Pardo[1], Lucas Antiqueira[1], Maria das Graças Volpe Nunes[1], Osvaldo N. Oliveira Jr.[1,2], Luciano da Fontoura Costa[2]

[1] Núcleo Interinstitucional de Lingüística Computacional (NILC)
CP 668 – ICMC-USP, 13.560-970 São Carlos, SP, Brasil
http://www.nilc.icmc.usp.br

[2] Instituto de Física de São Carlos
CP 369 – IFSC-USP, 13.560-970 São Carlos, SP, Brasil
http://www.ifsc.usp.br/

{taspardo,lantiq}@gmail.com, gracan@icmc.usp.br,
{chu,luciano}@if.sc.usp.br

**Abstract.** This paper presents a summary evaluation method based on a complex network measure. We show how to model summaries as complex networks and establish a possible correlation between summary quality and the measure known as dynamics of the network growth. It is a generic and language independent method that enables easy and fast comparative evaluation of summaries. We evaluate our approach using manually produced summaries and automatic summaries produced by three automatic text summarizers for the Brazilian Portuguese language. The results are in agreement with human intuition and showed to be statistically significant.

## 1 Introduction

Automatic text summarization is the task of automatically producing a shorter version of a text (Mani, 2001), which should convey the essential meaning of the source text and attend the reader's goals. Nowadays, due to the increasing amount of available information, mainly on-line, and the necessity of retrieving such information with high accuracy and of understanding it faster than ever, automatic summarization is unquestionably an important task.

Summaries are present in a wide range of our daily activities. During scientific papers writing, we have to write abstracts; when reading these papers, abstracts help us to determine whether the paper is important or not for our purposes. In a bookshop, the decision of buying a book is usually based on its cover synthesis. Some internet search engines use summaries to identify documents main parts and to help users in choosing which documents to retrieve.

In spite of the extensive investigation into methods for automatic summarization, it is still hard to determine which method is better. Summary evaluation remains an unresolved issue. Various aspects in summaries require evaluation (Mani, 2001), including amount of information, coherence, cohesion, thematic progression, legibility, grammaticality and textuality. Some are hard to define, while some significantly overlap. Depending on the final use of a summary, be it for humans or computer applications, different criteria need to be matched: if humans are the

intended readers, coherence and cohesion may be necessary; if the summary is to be used in a computer application, sometimes only the depicted information may be enough. There are several summary evaluation metrics, whose computation may be carried out either by humans or computers: if humans perform the evaluation, it becomes expensive, time consuming and prone to errors and inconsistencies; if computers perform it, subjective aspects of the evaluation are lost and evaluation may not be complete. Given the importance of the task, international conferences have been devoted to the theme, with DUC (Document Understanding Conference) being the most prominent, driving research in this area for the past 7 years.

Concomitantly, recent trends in Natural Language Processing (NLP) show the use of graphs as a powerful technique for modeling and processing texts. Such interest in graphs is due to their generic applicability, often leading to elegant solutions to difficult problems. For text summarization purposes, graphs have been used for both summary production (see, e.g., Erkan and Radev, 2004; Mihalcea, 2005) and evaluation (see, e.g., Santos Jr. et al., 2004). In particular, a special kind of graphs, called complex networks, has received great attention over the last few years. They have been proven useful to model NLP and Linguistics problems, in addition to many other applications (see, e.g., Barabási, 2003). Complex networks have been used, for instance, in modeling lexical resources (Sigman and Cecchi, 2002), human-induced words association (Costa, 2004), language evolution modeling (Dorogovtsev and Mendes, 2002), syntactic relationship between words (Cancho et al., 2005) and text quality measurement (Antiqueira et al., 2005a, 2005b).

This paper presents a first approach to the use of complex networks in summary evaluation. Particularly, it builds on the work of Antiqueira et al. (2005a, 2005b), by describing a possible representation of summaries as complex networks and establishing a correlation between summary quality and one of the network properties, namely the dynamics of the network growth. We evaluate our approach using TeMário corpus (Pardo and Rino, 2003), comprising 100 texts in Brazilian Portuguese and the corresponding human-produced (manual) summaries, and automatic summaries produced by the systems GistSumm (Pardo et al., 2003), SuPor (Módolo, 2003) and GEI (Pardo and Rino, 2004).

In the next section, complex networks are introduced. Section 3 describes how summaries are modeled as complex networks. Experimental results with manual and automatic summaries for verifying the correlation of the dynamics of the network growth property and quality are shown in Section 4. Section 5 presents the conclusions and final remarks.

## 2 Complex networks: an overview

Complex networks are particularly complex types of graphs, i.e. structures that contain nodes and edges connecting them. They have received enormous attention in the last few years, but their study can be traced back to initial development in graph theory. However, in contrast to simple graphs, complex networks present connecting structures that tend to depart from being randomly uniform, i.e., their growth is usually not uniformly random (Barabási, 2003). Complex networks have been used to describe several world phenomena, from social networks to internet topology. Such

phenomena present properties that often conform to the complex network characteristics, which caused the complex networks to be studied in a wide range of sciences, mainly by mechanical statistics and physics. See Barabási (2003) and Newman (2003) for a comprehensive scenario of complex network uses.

Some properties that may be observable in complex networks are worth mentioning. Networks known as *small world networks* point to the fact that there is a relatively short path between most nodes in the networks. For instance, social networks are usually small worlds. The *clustering coefficient* indicates the tendency of the network nodes to form groups; in a social network, the friends of a person tend to be friends too. A network is said to be *scale free* if the probability of a node having k edges connecting it to other nodes follows a power law distribution, i.e., $P(k) \sim k^{-\gamma}$, where $\gamma$ is a constant value dependent on the network properties (topology and connectivity factors, for instance). Scale free networks contain *hubs*, which consist of highly connected nodes. In internet, for example, hubs are the pages receiving links from many other pages. These properties are also applicable to NLP related tasks. Sigman and Cecchi (2002) modeled WordNet (Miller, 1985) as a complex network, where nodes represent the word meanings and edges represent the semantic relations between them. They showed that this network is a small world and contains hubs, mainly because of polysemic words. Motter et al. (2002) modeled a thesaurus as a network, where nodes represent words and edges represent the synonym relations between them, and detected that this network was scale free. Antiqueira et al. (2005a, 2005b) modeled texts as complex networks, where nodes represent the words and edges connect adjacent words in a text. Among other things, they suggested that text quality is somewhat related to the clustering coefficient, with quality deteriorating with an increasing coefficient.

In the next section, we show how to model summaries as complex networks.

## 3 Representing summaries as complex networks

Our representation of summaries as complex networks follows the scheme proposed by Antiqueira et al. (2005a, 2005b). Firstly, pre-processing steps are carried out: the summary stopwords are removed and the remaining words are lemmatized. Removing stopwords eliminates irrelevant and very common words; using lemmas instead of words causes the processing to be more intelligent, since it is possible to identify words with related meaning. The pre-processed summary is then represented as a complex network. Each word corresponds to a node in the network and words associations are represented as directed edges. In the representation adopted, each association is determined by a simple adjacency relation: for each pair of adjacent words in the summary there is a directed edge in the network pointing from the node that represents the first word to the node representing the subsequent word in the summary. The edges are weighted with the number of times the adjacent words are found in the summary. Significantly, in this representation, sentence and paragraph boundaries are not taken into consideration. As an example, the sample summary of Figure 1 (in Portuguese) is represented by the network in Figure 2.

*Lula e Fernando Henrique Cardoso estão nitidamente à frente nas eleições. Nas sondagens anteriores, o segundo lugar de Fernando Henrique era disputado por mais quatro. Hoje, quem mais o ameaça, mesmo assim sem perigo, é Sarney.*

**Figure 1.** Sample summary

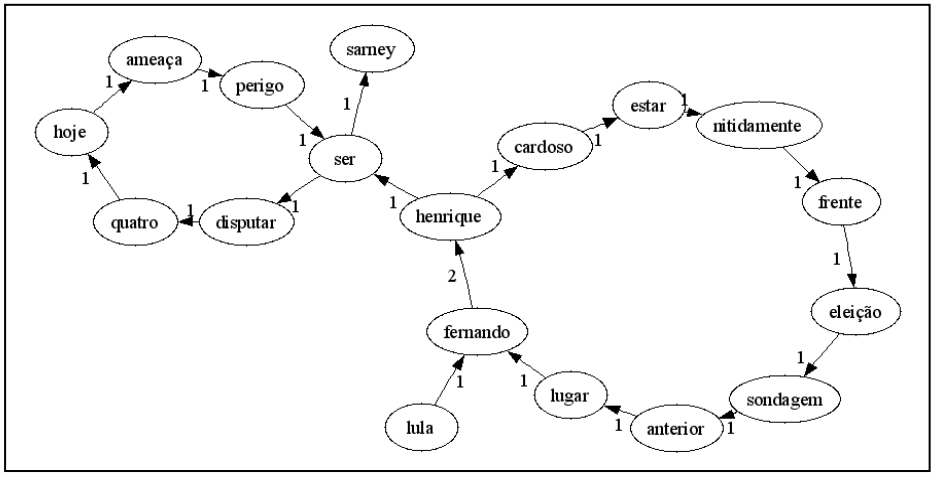


**Figure 2.** Complex network for the summary in Figure 1

## 4 Summary evaluation

Antiqueira et al. (2005a, 2005b) showed the existence of correlation between the dynamics of network growth and the quality of the text represented. The dynamics of a network growth is a temporal measure of how many connected components there are in the network as words associations are progressively incorporated into the network as it is constructed. Initially, in a time $t_0$, all N different words (nodes of the network) in the text under analysis are the components. In a subsequent time $t_1$, when an association is found between any two adjacent words $w_i$ and $w_j$ in the text, there are N-1 components, i.e., the component formed by $w_i$ and $w_j$ and the other N-2 words without any edge between them. This procedure is considered with each new word being added, until only one component representing the whole text is formed. For each text, Antiqueira et al. plot a graphic whose curve indicates the number of components in the network as new words associations are considered (which implies inserting a new edge, if it does not exist, or increasing the edge weight by 1 if it already exists). Considering a straight line in this graphic, which would indicate that there is a linear variation of the number of components as new words associations are considered, the authors showed that good-quality texts tend to be associated to a straight line in the dynamics plot. Moreover, text quality decreased with an increase in the deviation from the straight line.

The general deviation from the straight line for a text is quantified by following formula:

$$deviation = \frac{\sum_{M=1}^{A} |f(M) - g(M)| / N}{A}$$

where f(M) is the function that determines the number of components for M words associations and g(M) is the function that determines the linear variation of components for M words associations; N is the number of different words in the text and A is the total number of words associations found.

Figure 3 shows the plot for a longer version of the summary in Figure 1, which is a manual summary built by a professional abstractor. The straight dotted line is the one that assume linear variation of the number of components; the other line is the real curve for the summary. According to the above formula, the general deviation for the summary is 0.023. Figure 4 shows the plot for an automatic summary known to be worse, with same size and for the same source text of the summary of Figure 3. Its general deviation is 0.051. Note the larger deviation in the curve.

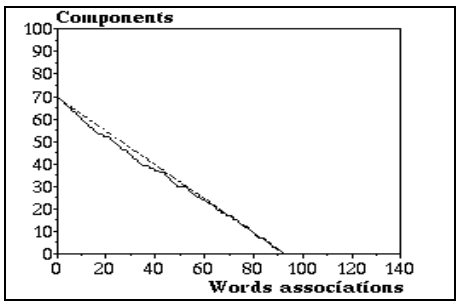


**Fig 3.** Plot for a manual summary

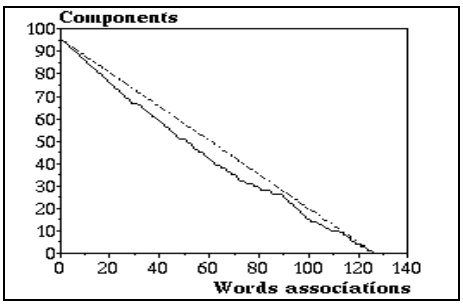


**Figure 4.** Plot for an automatic summary

Antiqueira et al. performed their experiments with news texts, supposed to be good, and students' essays, supposed to be worse than the news texts. In this paper, we evaluate the possibility of adopting such method in summary evaluation. In order to do so, we first assume, as most works on summary evaluation do, that a summary must display the same properties a text presents in order to be classified as text. Therefore, summaries, as texts, must be coherent and cohesive, legible, grammatical, and present good thematic progression.

In our evaluation, we used a corpus called TeMário (Pardo and Rino, 2003) for Brazilian Portuguese. TeMário consists of 100 news texts from the on-line newspaper *Folha de São Paulo* (containing texts from Sections Special, World, Opinion, International, and Politics) and their corresponding manual summaries written by a professional abstractor. To our knowledge, TeMário is the only available corpus for summarization purposes for the Brazilian Portuguese language.

We compared the manual summaries to automatic summaries produced by 3 systems, namely, GistSumm (GIST SUMMarizer) (Pardo et al., 2003), SuPor (SUmmarizer for PORtuguese) (Módolo, 2003) and GEI (*Gerador de Extratos Ideais*) (Pardo e Rino, 2004). We selected these systems for the following reasons: GistSumm is one of the first summarization systems publicly available for Portuguese; according to Rino et al. (2004), SuPor is the best summarization system

for Portuguese; GEI was used to produce the automatic summaries that also accompany TeMário distribution. In what follows, each system is briefly explained. Then, our experiment is described and the results discussed.

### 4.1. Systems description

The summarizers used in the evaluation are all extractive summarizers, i.e., they build the summary of a source text by juxtaposing complete sentences from the text, without modifying them. The summaries produced in this way are also called extracts.

GistSumm is an automatic summarizer based on a summarization method called gist-based method. It comprises three main processes: text segmentation, sentence ranking, and summary production. Sentence ranking is based on the keywords method (Luhn, 1958): it scores each sentence of the source text by summing up the frequency of its words and the gist sentence is chosen as the one with the highest score. Summary production focuses on selecting other sentences from the source text to include in the summary, based on: (a) gist correlation and (b) relevance to the overall content of the source text. Criterion (a) is fulfilled by simply verifying co-occurring words in the candidate sentences and the gist sentence, ensuring lexical cohesion. Criterion (b) is fulfilled by sentences whose score is above a threshold, computed as the average of all the sentence scores, to guarantee that only relevant sentences are chosen. All the selected sentences above the threshold are juxtaposed to compose the summary.

SuPor is a machine learning based summarization system and, therefore, has two distinct processes: training and extracting based on a Naïve-Bayes method, following Kupiec et al. (1995). It allows combining linguistic and non-linguistic features. In SuPor, relevant features for classification are (a) sentence length (minimum of 5 words); (b) words frequency; (c) signaling phrases; (d) sentence location in the texts; and (e) occurrence of nouns and proper nouns. As a result of training, a probabilistic distribution is produced, which entitles summarization in SuPor. In this paper, following Rino et al. (2004), we use the same features. SuPor works in the following way: firstly, the set of features of each sentence are extracted; secondly, for each of the sets, the Bayesian classifier provides the probability of the corresponding sentence being included in the summary. The most probable ones are selected to be in the summary.

Given a manual summary and its source text, GEI produces the corresponding ideal extract, i.e., a summary composed of complete sentences from the source text that correspond to the sentences content from the manual summary. This tool is based on the widely known vector space model and the cosine similarity measure (Salton and Buckley, 1988), and works as follows: 1) for each sentence in the manual summary, the most similar sentence in the source text is obtained through the cosine measure (based on word co-occurrence); 2) the most representative sentences are selected, yielding the corresponding ideal extract.

In general, ideal extracts are necessary to calculate automatically the amount of relevant information in automatic summaries produced by extractive methods. The automatic summaries are compared to the ideal extracts and two measures are usually computed: recall and precision. Recall is defined as the number of sentences from the

ideal extract included in the automatic summary over the number of sentences in the ideal extract; precision is defined as the number of sentences from the ideal extract included in the automatic summary over the number of sentences in the automatic summary. A third measure, called f-measure, is a combination of recall and precision, being a unique measure of a summarization system performance.

As described by Rino et al. (2004), GistSumm and Supor participated in a comparative evaluation. Recall, precision and f-measure were computed for TeMário corpus, using the ideal extracts produced by GEI. A 30% compression rate was used in producing the automatic summaries. The compression rate specifies the size of the summary to be produced in relation to the source text in terms of number of words. In this case, the 30% compression rate specifies that the summary must have at most 30% of the number of words in the source text. Recall, precision and f-measure for GistSumm and SuPor are shown in Table 1, which reproduces part of the evaluation that Rino et al presented.

**Table 1.** Systems performance (in %)

| Systems | Recall | Precision | F-measure |
|---|---|---|---|
| SuPor | 40.8 | 44.9 | 42.8 |
| GistSumm | 25.6 | 49.9 | 33.8 |

As can be noted, GistSumm had the highest precision, but the lowest recall. SuPor presented the best f-measure, being, therefore, the best system. These results will be commented upon in the next subsection, which describes the complex network experiment conducted in this paper.

### 4.2. Experiment

For running our experiment, we took the manual summaries and the ideal extracts (produced by GEI) that accompany TeMário and the corresponding automatic summaries produced by GistSumm and SuPor. As in Rino et al. (2004), we used a 30% compression rate. Based on our knowledge about the way the summaries were produced and on the evaluation that Rino et al. presented, we assume that the manual summaries are better than the ideal extracts, which are better than the automatic summaries. In terms of complex networks, the deviation from a straight line in the dynamics of network growth should be lowest for the manual summaries, and then increase for the ideal extracts and even more for the automatic summaries.

At this point, it is hard to predict how SuPor and GistSumm summaries will behave in relation to each other. Although SuPor is better than GistSumm in informativity evaluation (see Table 1), i.e., the amount of relevant information the summaries have, it is unlikely this will be reflected in the way we model summaries as complex networks. In fact, in the text quality experiment, Antiqueira et al. (2005a, 2005b) suggested that what is being captured by the complex network is the flow of new concepts introduced during the text: bad texts would introduce most of the concepts abruptly; good texts, on the other hand, would do it gradual and uniformly during the text development, resulting in a more understandable and readable text.

Table 2 shows the average deviation for each group of summaries and its increase in relation to the manual summaries deviation. For instance, for GistSumm (line 3 in

the table), the average of the summaries deviation is 0.03673, which is 20.62% larger than the average deviation for the manual summaries.

**Table 2.** Experiment results

| | **Avg. deviation** | **Over manual summaries (%)** |
|---|---|---|
| **Manual summaries** | 0.03045 | 0 |
| **GEI** | 0.03538 | 16.19 |
| **GistSumm** | 0.03673 | 20.62 |
| **SuPor** | 0.04373 | 43.61 |

Using t-student test (Casella and Berger, 2001) for comparing the average deviations of our data, with 99% confidence interval, the p-values are below 0.03, which indicates that the resulting numbers are not due to mere chance. In other words, the results are statistically significant. The only exception was the p-value for the comparison between GistSumm and GEI, which was around 0.60. This happened because of the short distance between the results of the two systems, as Table 2 illustrates.

Figure 5 shows the histograms for the summaries and their respective deviations, where the x-axis represents the deviation and the y-axis the number of texts. As the average deviation grows for each group of summaries, the Gaussian distribution has its peak (which corresponds to the mean) displaced to the right, i.e. there are more texts with higher deviations.

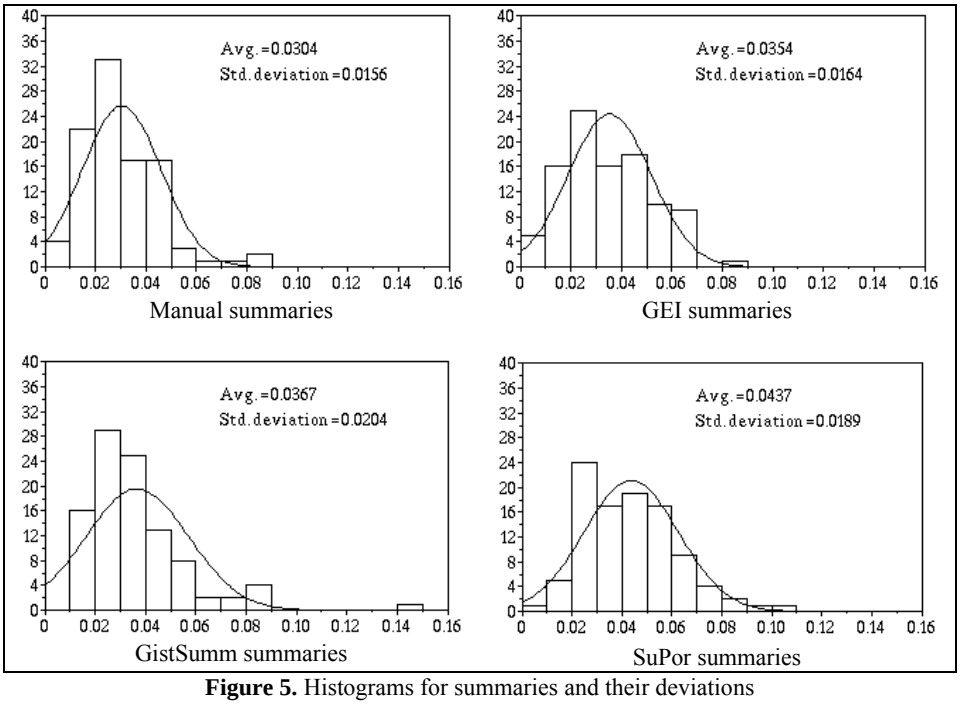


**Figure 5.** Histograms for summaries and their deviations

As expected, the results suggest that manual summaries are better than the ideal extracts, and that these are better than the automatic summaries. This observation positively answers our question about the possibility of using complex networks to evaluate summaries in a comparative fashion. We claim that it must be restricted to a comparative evaluation because it is difficult to judge the validity of a deviation number without any reference. The results also show that, in contrast to the informativity evaluation, GistSumm outperformed SuPor in this experiment, as mentioned above as a possible result. We believe the reason for this to be the summarization method used by GistSumm: to produce the summary, it selects sentences that correlate with the gist sentence, resulting in a summary with similar thematic elements across the sentences and, therefore, with a more natural flow of concepts. With GistSumm and SuPor numbers, it is also possible to conclude for the truth of the assumption that our modeling of summaries as complex networks probably does not capture summary informativity or that alternative complex networks measurements may be necessary.

## 5 Conclusions

This paper presented an application of the approach described by Antiqueira et al. (2005a, 2005b) to summary evaluation, which is considered a hard problem in NLP. By modeling summaries as complex networks and by exploring a network metric, we showed it to be possible to distinguish summaries according to their quality. The evaluation presented here can be used in association to other automatic evaluations, complementing the results obtained with the traditional informativity metrics – recall and precision – or new ones – ROUGE (Lin and Hovy, 2003), for instance. Because it is based on abstract representation of texts in terms of complex networks, the proposed solution looks elegant, generic and language independent.

In the future, we plan to apply such evaluation to other text genres, in addition to the news texts. We also aim at investigating other network properties and their usefulness for characterizing the several aspects of a summary that is worth modeling and evaluating, e.g., coherence and cohesion. Other ways of modeling summaries as complex networks may also be explored.

## Acknowledgments

The authors are grateful to CNPq and FAPESP.

## References

Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr., O.N.; Costa, L.F. (2005a). Modelando Textos como Redes Complexas. In *Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL. São Leopoldo-RS, Brazil. July 22-26.

Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr., O.N.; Costa, L.F. (2005b). *Complex networks in the assessment of text quality*. physics/0504033.

Barabási, A.L. (2003), *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume, New York.

Cancho, R. F.; Capocci, A.; Caldarelli, G. (2005). *Spectral methods cluster words of the same class in a syntactic dependency network*. cond-mat/0504165.

Casella, J. and Berger, R.L. (2001). *Statistical Inference*. Duxbury, Belmont, California.

Costa, L.F. (2004). What's in a name? *International Journal of Modern Physics C*, Vol. 15, pp. 371-379.

Dorogovtsev, S.N. and Mendes, J.F.F. (2002). Evolution of networks. *Advances in Physics*, Vol. 51, N. 4, pp. 1079-1187.

Erkan, G. and Radev, D.R. (2004). Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research* – JAIR, Vol. 22, pp. 457-479.

Kupiec, J.; Pedersen, J.; Chen, F. (1995). A trainable document summarizer. In the *Proceedings of the 18th ACM-SIGIR Conference on Research & Development in Information Retrieval*, pp. 68-73.

Lin, C-Y. and Hovy, E.H. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In the *Proceedings of Language Technology Conference* – HLT. Edmonton, Canada. May 27 - June 1.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.

Mani, I. (2001). *Automatic Summarization*. John Benjamin's Publishing Company.

Mihalcea, R. (2005). Language Independent Extractive Summarization. In the *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan.

Miller, G.A. (1985). Wordnet: a dictionary browser. In the *Proceedings of the First International Conference on Information in Data*. University of Waterloo.

Módolo, M. (2003). *SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português*. Master thesis. Departamento de Computação, UFSCar.

Motter, A.E.; Moura, A.P.S.; Lai, Y.C.; Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E*, Vol. 65, 065102.

Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Review*, Vol. 45, pp. 167-256.

Pardo, T.A.S. and Rino, L.H.M. (2003). *TeMário: Um Corpus para Sumarização Automática de Textos*. NILC technical report. NILC-TR-03-09. São Carlos-SP, October, 13p.

Pardo, T.A.S. and Rino, L.H.M. (2004). *Descrição do GEI - Gerador de Extratos Ideais para o Português do Brasil*. NILC technical report. NILC-TR-04-07. São Carlos-SP, August, 10p.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal. June 26-27.

Rino, L.H.M.; Pardo, T.A.S.; Silla Jr., C.N.; Kaestner, C.A.; Pombo, M. (2004). A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In the *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence* – SBIA (Lecture Notes in Artificial Intelligence 3171), pp. 235-244. São Luis-MA, Brazil. September, 29 - October, 1.

Santos Jr. E.; Mohamed, A.A.; Zhao Q. (2004). Automatic Evaluation of Summaries Using Document Graphs. In the *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 66-73.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, Vol. 24, pp. 513-523.

Sigman, M. and Cecchi, G.A. (2002). Global Organization of the Wordnet Lexicon. In the *Proceedings of the National Academy of Sciences*, Vol. 99, pp. 1742-1747.