Using Complex Networks for Language Processing: The Case of Summary Evaluation

Thiago Alexandre Salgueiro Pardo
Lucas Antiqueira
Maria das Graças Volpe Nunes
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
{taspardo@gmail.com, {lantiq,gracan}@icmc.usp.br

Osvaldo N. Oliveira Jr.
Luciano da Fontoura Costa
Instituto de Física de São Carlos
Universidade de São Paulo
{chu,luciano}@if.sc.usp.br

***Abstract* – The ability to access embedded knowledge makes complex networks extremely promising for natural language processing, which normally requires deep knowledge representation that is not accessible with first-order statistics. In this paper, we demonstrate that features of complex networks, which have been shown to correlate with text quality, can be used to evaluate summaries. The metrics are the average degree, cluster coefficient, and the extent to which the dynamics of network growth deviates from a straight line. They were found to be much smaller for the high-quality, manual summaries, and increased for automatic summaries, thus pointing to a loss of quality, as expected. We also discuss the comparative performance of automatic summarizers**

I. Introduction

Recent trends in Computer Science show the use of graphs as a powerful modeling technique, especially due to their wide applicability that often leads to elegant solutions for difficult problems. In this context, complex networks, which are special types of graphs, have received increased attention. Because the growth of such networks is governed by complex principles ([4], [18]), they have been used to describe several world phenomena, from social networks to internet topology. In fact, the wide applicability of complex networks was realized soon after the seminal work of [3], as it became clear that embedded knowledge in complex systems could emerge. Recently, complex networks concepts were applied to languages (see, e.g., [10], [25], [9], [7], [1], [2]). Perhaps the most important motivation for using complex networks in Linguistics and Natural Language Processing (NLP) is in the ability to access embedded knowledge. Of particular relevance is the possibility to analyze related concepts through metrics of complex networks, which is essential for deep knowledge representation in NLP. Complex networks offer further possibilities of text analysis in comparison to first-order statistics, which has already proven useful in NLP for decades (see, e.g., [16], [26], [12]).

We believe complex networks concepts can be applied in several NLP tasks, and reports have already appeared for some of these tasks, which will be briefly reviewed in Section II. In this paper, in particular, we exploit the strategies to model text as complex networks proposed by [1] and [2] and extend the work by [22] in modeling and evaluating summaries. The main results are presented in Section III, while Section IV brings conclusions and final remarks.

II. Complex Networks and natural language

Various examples exist of network properties applied to language processing tasks. For instance, WordNet [14] has been modeled by [25] as a complex network, in which word meanings were the nodes while the semantic relations between concepts represented the edges. Mainly because of polysemic words, WordNet was found to behave as a small-world network. A scale-free network could represent a thesaurus [17], with nodes representing words and edges representing relations such as synonyms and semantic field. A network is referred to as *scale free* if a power-law distribution describes the probability for a node with k edges, i.e., $P(k) \sim k^{-\gamma}$, where γ is a constant that depends on network properties. Scale-free networks can also represent word co-occurrence networks, with edges constructed according to the sequence of words in a text [5], the word association network, where words are interconnected if they share similar concepts [9], and the syntactic dependency network, in which edges represent syntactic relationships between words [6]. Other applications of complex networks in NLP included a method to group words according to their morphological classes [7] and a way to detect ambiguity in a text [11].

[1] and [2] modeled texts as complex networks, where nodes represented the words and edges connected adjacent words. They suggested that text quality is related to the clustering coefficient, the network degree (i.e., the average

number of nodes connections) and to the dynamics of network growth. Quality was shown to decrease with increasing cluster coefficients, degree and with the deviation from a straight line in the network dynamics. Following this work, in the next section, we show how to apply the complex networks concepts to summary evaluation.

III. Summary Evaluation

Summary evaluation is a hard and unsolved issue [13], which has been the subject of various international conferences, such as DUC (Document Understanding Conference). To our knowledge, complex networks were applied to summary evaluation for the first time in our earlier work [22]. There, we employed a Markov representation taking into account only the adjacent words, referred to as Markov-1 (see below), and showed that the quality of a summary could be correlated with the dynamics of the network growth. One limitation though was that quality was based on only one criterion, and that the context of only one word (Markov-1) may have not captured important features of the summaries.

In this paper, we extend the investigation into summary evaluation by employing 5 representations of complex networks. Furthermore, in addition to analyzing the network dynamics, we also measure cluster coefficients and node degrees. For all representations, the following pre-processing steps were carried out: the stopwords were removed and the remaining words were lemmatized, in order to consider as a single node words with related or similar meaning.

Our first representation follows the scheme proposed by [1] and [2]. Each node in the network corresponds to a word in the text and directed edges are established for every words association. 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 first word to the subsequent word in the summary. We do not consider sentence and paragraph boundaries. The edges of the network are weighted with the number of times the corresponding adjacent words are found in the summary. We refer to this representation as Markov-1 representation, i.e., a one-state Markov model, in which each word is related only to the immediate previous word. According to the theory, Markov models specify how the determination of a state depends on the observation of previous states. In our case, each state represents a summary word.

The other 4 representations, referred to as Markov-2, Markov-3, Markov-4 and Markov-5, are simple variations of Markov-1: they differ in the number of previous words that each word in the text is related to. In Markov-2, for a sequence of words w_1 w_2 w_3 in the text, edges are established from w_1 to w_3 and from w_2 to w_3, i.e., the word w_3 is related to the two previous words. Generalizing, in a Markov-K representation, each word in a text is related to the K previous words, i.e., edges are established from words w_1 … w_K to word w_{K+1}. Similarly, all the edges are weighted with the number of times the corresponding words happen to be related in the whole summary.

For each summary represented by a network three measures were taken: the average (out,in)-degree, the clustering coefficient and the deviation from a linear dynamics in the network growth (see below). The degrees were calculated for each node, and the average degree was computed for each network. Because averages were taken, the values for the in-degree and out-degree coincided, and we used only one of them, referred to here simply as *degree*. The clustering coefficient is the one used by [1] and [2], which takes into consideration the edges direction. The dynamics of a network growth is a temporal measure of how many connected components exist in the network as words associations are progressively incorporated into the network during its construction. Consider the Markov-1 representation: 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, 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; and so on. This procedure is considered with each new word being added, until only one component representing the whole text is formed. Similarly, in this paper, for a summary Markov-K representation (with $1 \leq K \leq 5$ in our case), in any time t in the network construction, the components are counted when all the K edges between the corresponding words are considered.

In order to analyze quantitatively the dynamics of the network growth, we followed the procedure by [1]. They plotted the number of components in the network vs. time as new word associations were considered (which implies inserting a new edge, if it does not exist, or increasing the edge weight by 1 if it already exists). It was found that good quality texts tend to be those for which the plot was a straight line, with text quality deteriorating with increasing deviations from the straight line. This deviation was quantified for a text as follows:

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

where f(M) is the function that determines the number of components for M word associations and g(M) is the function that determines the linear variation of components for M word associations; N is the number of different words in the text and A is the total number of word associations found. Figure 1 shows the plot for a manual summary written by a professional abstractor using the Markov-1 complex network representation. The linear variation for the number of components is represented by the dotted line; the other line is the real curve for the summary. Using the formula above, the overall deviation for the summary is 0.023. Figure 2 shows the plot for an automatic summary, with the same size and for the same source text of the manual summary of Figure 1. The automatic summary is known to be worse than

the manual summary. Its general deviation is 0.069. Note the larger deviation in the curve.

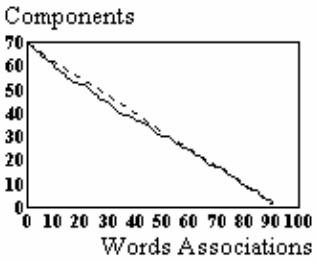

Figure 1. Plot for a manual summary

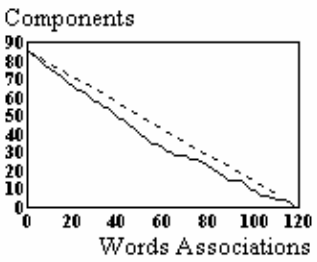

Figure 2. Plot for an automatic summary

The three measures were obtained for the manual summaries in Brazilian Portuguese from corpus TeMário [19] and for automatic summaries. TeMário consists of 100 news texts from the on-line newspaper *Folha de São Paulo* and their corresponding manual summaries written by a professional abstractor. We compared the manual summaries to automatic summaries produced by 3 automatic summarizers for Brazilian Portuguese, namely, GistSumm (GIST SUMMarizer) [21], SuPor (SUmmarizer for PORtuguese) [15] and GEI (*Gerador de Extratos Ideais*) [20], which are briefly described in the next subsection.

A. Systems description

The summarizers used are all extractive summarizers, i.e., they select complete sentences from a source text to compose the summary, which is also named extract. GistSumm is an automatic summarizer based on the gist-based method, comprising three main processes: text segmentation, sentence ranking, and summary production. Sentence ranking is obtained from word frequencies by assigning scores to each sentence of the source text by summing up the frequency of its words; 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.

SuPor is a machine learning-based summarization system and, therefore, has two distinct processes: training and extracting based on a Naïve-Bayes method. In SuPor, relevant features for classification are (a) sentence length, (b) words frequency; (c) signaling phrases, (d) sentence location in the texts and (e) occurrence of nouns and proper nouns. SuPor works as follows: firstly, the set of features of each sentence is extracted; then, 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 for 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 content of sentences from the manual summary. This tool is based on the vector space model and the cosine similarity measure [24], and works as follows: the most similar sentence in the source text for each sentence in the manual summary is obtained through the cosine measure (based on words co-occurrence); the selected sentences are juxtaposed to form the ideal extract. In general, ideal extracts are used to calculate automatically the amount of information in automatic summaries produced by extractive summarizers. 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, and is a general measure of how good an automatic system is.

[23] employed the TeMário corpus to produce extracts with a 30% compression rate, i.e. the summaries had 30% of words of the original summaries. As Table 1 (which reproduces their results) shows, SuPor outperforms GistSumm in terms of recall and f-measure, while the converse is true for precision. The values in the table were obtained with the ideal extracts produced by GEI as reference. We shall comment upon these results in the next subsection, which describes the complex network experiments conducted in this paper.

Table 1. Systems performance (in %)

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

B. Experiments

In the first experiment we obtained the deviation from a straight line in the network dynamics for all summaries from TeMário and the corresponding ones obtained with GistSumm, SuPor and GEI (with 30% compression rate), which were represented by Markov-1 through Markov-5 strategies. Table 2 shows the average deviation for each group of summaries, and as expected from [22] for Markov-1, the deviation is lowest for the manual summaries, and increases for the automatic summaries. Based on our knowledge about the way the summaries were produced, we expect the manual summaries to be better than the ideal extracts, which should be better than the automatic summaries. If we now assume – as indicated by the studies of [1] and [2] – that text quality decreases with increasing deviations, the trends in Table 2 are all consistent, with

perhaps one exception in the comparison between GistSumm and GEI. We shall return to this point later on.

Table 2. Deviation in the network dynamics

	Manual summaries	GEI	GistSumm	SuPor
Markov-1	0.03045	0.03538	0.03673	0.04373
Markov-2	0.03045	0.03538	0.03673	0.04374
Markov-3	0.03174	0.03657	0.03833	0.04489
Markov-4	0.03350	0.03807	0.04046	0.04643
Markov-5	0.03537	0.03977	0.04262	0.04808

As for the distinct representations, Table 2 indicates no difference between Markov-1 and Markov-2. For the other representations the deviation increased consistently, though the trends were preserved. It is therefore concluded that the context given by only one word, in Markov-1, is already sufficient to capture the important features of the complex network, and this confirms the findings by [2], who observed no significant effects in using Markov-1 or Markov-2 in the analysis of text quality.

It should be stressed that the differences in Table 2 are statistically significant, as we checked with the t-student test [8]. With 99% confidence interval, the p-values are below 0.06 for the average deviations of the data, which indicates that these differences in the network dynamics were not obtained simply by chance. The only exception was in the p-values for the comparison between GistSumm and GEI, which was around 0.60 in the worst case. This occurred due to the small differences in the results for the two systems, as Table 2 illustrates.

For the remaining experiments to obtain the node degrees and cluster coefficients, we employed only Markov-1 and Markov-2 representations. Table 3 shows for Markov-1 that again the manual summaries are much better than the automatic summaries, as both the cluster coefficient and the degree are considerably lower for the manual summaries. We recall that – according to [2] - text quality increases with decreasing values of these network metrics. Furthermore, consistent with the results from the network dynamics of Table 2, the performances of GEI and GistSumm are very similar. The reason why in the Markov-2 representation the cluster coefficient did not vary for the different sets of texts is unknown at the moment.

Table 3. Degree and clustering coefficient measures

	Markov-1		Markov-2	
	Degree	**C. Coef.**	**Degree**	**C. Coef.**
Manual summaries	1.23065	0.00267	2.44927	0.44933
GEI	1.28568	0.00395	2.56037	0.44594
GistSumm	1.27730	0.00447	2.54034	0.44846
SuPor	1.35283	0.00522	2.69500	0.44299

The t-student test was also performed for the data in Table 3. The p-values for the comparison between the mean degrees and clustering coefficient are below 0.005 and 0.07, respectively, with the exception (again) for the comparison between GistSumm and GEI. Therefore, apart from the differences between these two systems, the differences in Table 3 are statistically significant.

Taking together the data from Tables 2 and 3, we may conclude that the metrics of the complex networks assumed to be correlated with text quality can also be used to evaluate summaries. For with all the 3 metrics the manual summaries were considered better than the automatic ones. There are results, however, that are conflicting with previous studies or with our expectation. For instance, in our experiments GistSumm consistently outperformed SuPor, in contrast to the findings of [23] summarized in Table 1. Our main hypothesis to explain these results is that [23] evaluation is an informativity evaluation, i.e., it measures the amount of relevant information in the summaries, while it is unlikely that our representation of summaries as complex networks captures completely this aspect. On the other side, one may then ask whether the f-measure, i.e. combination of recall and precision, is the best parameter to assess the performance of a summarizer. In addition, could it be that precision – higher for GistSumm – is so much important than recall for a summarizer? Moreover, the values of precision and recall of Table 1 were obtained by comparing the automatic summaries with ideal extracts (not manual summaries). As the present analysis using complex networks appears to indicate, the ideal extracts may not be much better than automatic summaries. Definitive answers to the questions will require extensive studies of various summarizers (for other languages as well), whose comparative performances have been established beyond doubt. In fact, this will be the next step in our research.

Even more surprising was the result in which GistSumm displayed a similar performance with GEI summaries (see Tables 2 and 3). The latter, being obtained with sentences designed to match those of the manual summary, is considered an ideal extract. We have two possibilities to explain this unexpected result: (i) the procedure to calculate the similarity between sentences of the source text and manual summary to select those to be included in the extract may not yield good results; (ii) the performance of GistSumm is so good as to mimic what a system to obtain ideal extracts does. Similarly to the discussion about the differences between GistSumm and SuPor, further research is necessary to elucidate the reasons for the unexpected performance of GistSumm.

IV. Conclusions

The results with 3 metrics of complex networks confirmed that the correlations with text quality suggested by [1] and [2] may be extended to evaluate the quality of summaries. Significantly, the analysis with complex networks indicated much higher quality for the manual summaries, as one should expect. With regard to the outstanding performance of GistSumm, one may speculate – from a complex network perspective – about the consequences of the processes involved in GistSumm. In calculating the gist sentence, GistSumm considers statistics that may emulate the identification of hubs in a complex network. Furthermore, for the selection of the remaining

sentences to complete the summary, GistSumm picks those that are more statistically related to the gist sentence, and this again may simulate features of complex networks.

ACKOWLEDGMENTS

The authors are grateful to CNPq and FAPESP for the financial support and to Lucia Rino for her contributions to the ideas in the paper.

REFERENCES

[1] Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr., O.N.; Costa, L.F. 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. 2005.

[2] Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr., O.N.; Costa, L.F. *Strong correlations between text quality and complex networks features*. physics/0504033.v2. 2006.

[3] Barabási, A.L. and Albert, R. Emergence of scaling in random networks. *Science*, V. 286, pp. 509-512. 1999.

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

[5] Cancho, R.F. and Solé, R.V. The small world of human language. In the *Proceedings of The Royal Society of London*, Series B, V. 268. 2001.

[6] Cancho, R.F.; Solé, R.V.; Köhler, R. Patterns in syntactic dependency networks. *Physical Review E*, V. 69. 2004.

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

[8] Casella, J. and Berger, R.L. *Statistical Inference*. Duxbury, Belmont, California. 2001.

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

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

[11] Dorow, B.; Widdows, D.; Ling, K.; Eckmann, J.P.; Sergi, D.; Moses, E. Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. In the *Proceedings of the 2nd MEANING Workshop*. Trento, Italy. 2005.

[12] Gonçalves, L.L. and Gonçalves, L.B. *Fractal power law in literary english*. condmat/0501361. 2005.

[13] Mani, I. *Automatic Summarization*. John Benjamin's Publishing Company. 2001.

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

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

[16] Montemurro, M.A. and Zanette, D.H. Entropic analysis of the role of words in literary texts. *Advances in Complex. Systems*, Vol 5, N. 1. 2002.

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

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

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

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

[21] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. 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. 2003.

[22] Pardo, T.A.S.; Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr., O.N.; Costa, L.F. *Modeling and Evaluating Summaries Using Complex Networks*. To appear in the 7th Workshop on Computational Processing of Written and Spoken Portuguese. 2006.

[23] Rino, L.H.M.; Pardo, T.A.S.; Silla Jr., C.N.; Kaestner, C.A.; Pombo, M. 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. 2004.

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

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

[26] Zhou, H. and Slater, G.W. A Metric to Search for Relevant Words. *Physica A*, V. 329. 2003.