Physica A 391 (2012) 4406–4419

Contents lists available at SciVerse ScienceDirect

Physica A

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

Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts

Diego R. Amancio *, Osvaldo N. Oliveira Jr., Luciano da F. Costa

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

ARTICLE INFO

Article history:
Received 27 December 2011
Received in revised form 27 January 2012
Available online 27 April 2012

Keywords:
Similarity index
Complex networks
Machine translation evaluation
Topological analysis
Authorship recognition

ABSTRACT

The classification of texts has become a major endeavor with so much electronic material available, for it is an essential task in several applications, including search engines and information retrieval. There are different ways to define similarity for grouping similar texts into clusters, as the concept of similarity may depend on the purpose of the task. For instance, in topic extraction similar texts mean those within the same semantic field, whereas in author recognition stylistic features should be considered. In this study, we introduce ways to classify texts employing concepts of complex networks, which may be able to capture syntactic, semantic and even pragmatic features. The interplay between various metrics of the complex networks is analyzed with three applications, namely identification of machine translation (MT) systems, evaluation of quality of machine translated texts and authorship recognition. We shall show that topological features of the networks representing texts can enhance the ability to identify MT systems in particular cases. For evaluating the quality of MT texts, on the other hand, high correlation was obtained with methods capable of capturing the semantics. This was expected because the golden standards used are themselves based on word co-occurrence. Notwithstanding, the Katz similarity, which involves semantic and structure in the comparison of texts, achieved the highest correlation with the NIST measurement, indicating that in some cases the combination of both approaches can improve the ability to quantify quality in MT. In authorship recognition, again the topological features were relevant in some contexts, though for the books and authors analyzed good results were obtained with semantic features as well. Because hybrid approaches encompassing semantic and topological features have not been extensively used, we believe that the methodology proposed here may be useful to enhance text classification considerably, as it combines well-established strategies.

1. Introduction

The growing amount of text electronically available has placed Natural Language Processing (NLP) in the spotlight [1–3]. Examples of applications exploiting NLP include machine translation [4], automatic summarization [5], search engines [6,7], writing tools [8], text simplification [9], information retrieval [10], in addition to various resources such as thesaurus [11] and corpus [12]. In some of these applications, one needs to estimate the similarity between documents. Indeed, summarizers and translators are usually evaluated according to the similarity with a reference text produced by humans, while categorizers and clustering applications [13] employ similarity measures to establish clusters containing similar texts. Defining similarity is not straightforward, though.

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

0378-4371/
doi:10.1016/j.physa.2012.04.011

Due to the practical and even theoretical (since the computation of similarity involves understanding the cognitive processes) interests related to the estimation of pairwise similarities, a wide variety of indices have been developed. The vast majority are based on semantic similarity, which is calculated by counting the number of keywords or n-grams shared by two documents [14]. More sophisticated techniques based on semantic analysis [15] have also been used which go beyond counting the number of shared words in distinct texts. Although such methods can be considered efficient because they have a reasonable correlation with the human assessment [16], for some applications a semantic analysis may not suffice, since the textual structure plays a prominent role. In classifying different literary styles, for example, there may be a correlation between the theme of different styles, but textual structure is expected to be a key factor to characterize the styles [17]. Therefore, similarity indices based on text structure may be more useful in this type of application. Analogously, the structure-based similarity indices could be useful for clustering texts with the same quality of writing [18], quality of translation [19,20] and even texts endorsing the same point of view [21], since all these applications can be suitably characterized by the structural paradigm.

Since both the style and the semantics can be useful for estimating similarity, in this article we study the interplay between semantics and structure in 3 NLP applications: (i) evaluation of quality of translations, (ii) translation classification (i.e., identification of which machine translator generated a given translation) and (iii) authorship recognition. Using formalisms based on the representation of texts as complex networks we derived (dis)similarity indices based on semantics and structure to show that both types of indices are able to reveal patterns that would be hidden if only one of the paradigms were used.

2. Complex networks and natural language processing

Concepts and methods from complex networks have been employed to analyze many aspects of language [3,18–20,22], including analysis of syntactic networks [23,24], classification of languages through topological analysis [25–27] and investigation of phonetic aspects [28]. Even though the main aim in using complex networks has been to study linguistic phenomena, in the majority of the studies the semantics has been disregarded because the focus is normally on the topology of the network. On the other hand, typical applications of natural language processing [29] only consider the semantic relationship between documents, as is the case of methods to quantify pairwise similarity, such as the bag-of-words technique [29]. In the latter, the number of shared terms is used as an estimate of similarity, while the order in which words occur is neglected. Obviously, a more thorough analysis of linguistic phenomena should be expected if one is able to combine semantic with topological features.

Indeed, hybrid approaches have proven promising not only for text analysis but also for modeling networks. For example, Menczer [30] showed that models of citation networks are more accurate if in addition to topological characteristics (e.g., the degree distribution) they include features regarding content similarity. Likewise, Mehler [31] showed that the interplay between semantic and structural features emerging from social networks is essential for representing and classifying large complex networks. Along these lines, in this paper we propose methods to estimate text similarity taking into account both *topological and semantic* features. With topology one may capture stylistic features concerning authorship [32], complexity [33], quality [18] and aspects that depend on non-trivial relationships between textual concepts. As for the semantic investigation, our aim is to capture textual pragmatic features so that manuscripts are clustered together when they share a given topic. As we shall show, the combined use of both strategies leads to improved evaluation of machine translation systems because the vast majority of established quality indexes neglect long-range stylistic information.

3. Methodology

3.1. Modeling texts as networks

The model used in this work was adopted in many previous studies [18–21]. The process starts by eliminating words conveying low semantic content (the full list of stopwords is given in the Supplementary Information (SI)) for we are interested only in the relationship between content words. It is true that by removing stopwords one may miss out on very important linguistic information [34,35]. However, in the approaches we used, removing stopwords is entirely justified for two reasons: (1) the statistics of the metrics in the complex networks would be unduly affected by the highly frequent stopwords (such as articles and prepositions) that may be connected with any type of node; (2) the stopwords were removed to make the topological analysis consistent with the semantic analysis, where stopwords play no role on the prediction of pairwise similarities [29]. The remaining words are transformed to their canonical form, where verbs are converted to their infinitive form and nouns are converted to the singular form. After this step, each distinct word becomes a node of the network. Edges link two words if they appear as neighbors in the pre-processed text. Section 2 of the SI illustrates step-by-step the construction of the network. Mathematically, the network is represented by a weighted matrix W, where w_{ij} stores the number of times the word i appeared before the word j. Alternatively, we also use the unweighed and undirected representation of W, which is known as adjacency matrix A. If words i and j are neighbors in the text, then $a_{ij} = 1$. Otherwise, $a_{ij} = 0$.

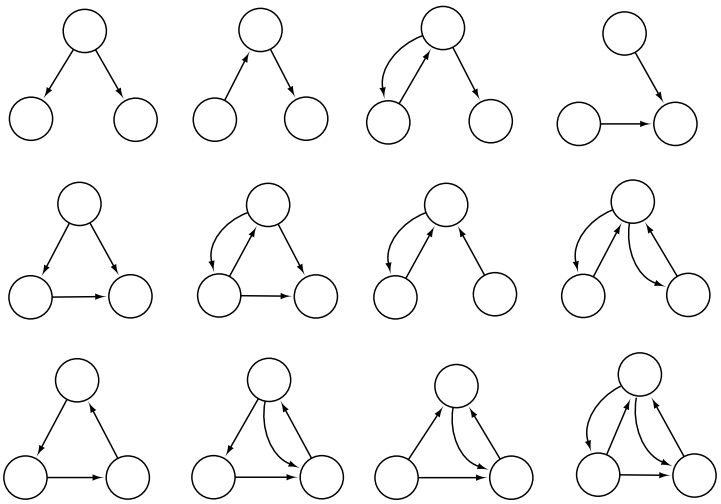

Fig. 1. Motifs employed to characterize the topology of the networks through the analysis of z-scores.

3.2. Topological network measurements

The topological measurements described in this section are associated with topological features of the network, with no concern for the semantics of the nodes. These metrics are: the degree k, the betweenness B, the average shortest path length l and the clustering coefficient C. All these measurements have been widely employed in the topological characterization of complex networks [36]. They will be employed in the derivation of the so-called topological similarity indices in Section 4.1.1. More details about these measurements are given in Ref. [36] and in Section 2 of the SI.

3.3. Topological network motifs

The network structure can also be characterized using motifs [37], which can be taken as small building blocks comprising nodes and edges. They may be seen as subgraphs that appear on the network more than expected just by chance. To verify if the frequency of a particular motif m is higher than that expected, the z-score measurement Z is used. To compute Z, let r_m and σ_m be, respectively, the average and the standard deviation of the frequency of m over 100 random equivalent networks (i.e., the random networks have the same number of vertices and edges as the original network). If n_m is the frequency of m in the network of interest, then Z is given by

$$Z(m) = \frac{n_m - r_m}{\sigma_m}. \tag{1}$$

It is known that some families of networks display high frequency of special motifs. Therefore, one may identify the function of the network by examining the frequency or significance profile of each motif [38]. In this work, we focused on motifs involving three vertices, as shown in Fig. 1. However, any other motif of particular interest could have been used.

4. Similarity and dissimilarity indices

In this section, we derive the indices used to evaluate and classify machine translations and recognize authorship of prose and poetry. The indices can be divided into three groups according to the use of topological or semantics features. If no topological measurements are used to define the index, then it is solely based on semantics. On the other hand, if no information on the label of the vertices is used, then the index is entirely topological. Finally, if both topological and semantic features are employed, the index is considered as a hybrid.

4.1. Topological dissimilarity indices

4.1.1. Dissimilarity index based on topological network features

This index is computed by obtaining the distance d_{st} of a given text T_s to a reference text T_t. Let $\overrightarrow{\mu}$ be a vector where each component represents the average for one measurement described in Section 3.2. The vectors $\overrightarrow{\mu_s}$ and $\overrightarrow{\mu_t}$, related to T_s and T_t, respectively, are compared, leading to the vector $\overrightarrow{\delta}$. The difference is computed component by component:

$$\delta(i) = \frac{\|\mu_t(i) - \mu_s(i)\|}{\mu_s(i)}. \tag{2}$$

The distance d_{st} is then obtained as the average of the differences:

$$d_{st} = \frac{1}{n}\sum_{i=1}^{n}\delta(i), \tag{3}$$

where each component i of $\overrightarrow{\delta}$ is the difference for a given measurement. Then, each component can be considered as a dissimilarity index itself.

4.1.2. Dissimilarity index based on motifs

Similarly to the previous one, this index is defined as the average over the differences in $\overrightarrow{\delta}$, with each component $\delta(i)$ being the z-score obtained for the motif associated with index i:

$$\delta(i) = \frac{\|Z_t(i) - Z_s(i)\|}{Z_s(i)}. \tag{4}$$

4.2. Semantic-based indices

In this section, we describe the indices based solely on the semantics. In other words, only the label and the information of the immediate neighborhood of the nodes are taken to quantify pairwise similarities. As in previous cases, all indices in this section consider that two nodes are similar if they share many neighbors.

When labels are used to compare networks of texts according to the number of shared neighbors, a problem arises if a node with a specific label does not appear in the other network. To obviate this problem, we adopted a strategy where a minimum similarity value is assigned to nodes that appear in only one of the networks. This can be obtained in the following manner. Let A^s and A^t be the networks being compared. If a given node belonging to A^s has label L and if A^t has no node with label L then a new node is created in A^t without any connections. Thus, since the node labeled with L in A^t has no neighbors, it will have no shared neighbors. Consequently, the similarity related to this node will be zero.

4.2.1. Cosine similarity

The number of sharing neighbors q_{ii} of two nodes v_i^s and v_i^t with the same label L_i in the networks A^s and A^t (which refer to T_s and T_t, respectively) is

$$q_{ii} = \sum_j A_{ij}^s A_{ij}^t. \tag{5}$$

While this measure captures the number of common neighbors, it is difficult to apply because one does not know whether q_{ii} is small or large [38]. In fact, a better approach should first normalize q_{ii} to confine its range within a strict interval. To perform such normalization, we divide q_{ii} by the geometrical mean between the degrees k_i^s and k_i^t of the nodes considered:

$$c_{ii} = \frac{\sum_k A_{ik}^s A_{ik}^t}{\sqrt{\sum_k A_{ik}^s}\sqrt{\sum_k A_{ik}^t}} = \frac{q_{ii}}{\sqrt{k_i^s k_i^t}}. \tag{6}$$

Note that c_{ii} can be interpreted as the cosine of the angle between the vectors of neighbors. Therefore, the higher c_{ii} the smaller the angle and consequently the larger the similarity is.

4.2.2. Similarity index based on the Pearson correlation coefficient

The cosine similarity above is an effective way to normalize q_{ii} in the sense that it limits its range in the interval between zero and 1, but other normalizations can be considered. For example, Ref. [38] suggests that q_{ii} should be compared to the expected number σ_{ii} of sharing neighbors, supposing that the choice of neighbors are made randomly. To quantify this expected value, let k_i^s and k_i^t be the degree of v_i^s and v_i^t being compared, respectively. If v_i^s randomly chooses its neighbors, the probability that it chooses a node that is also a neighbor of v_i^t is equal to k_i^t/n. Repeating the process for the remaining $k_i - 1$ neighbors of v_i^s, the expected number of sharing neighbors will be $\sigma_{ii} = k_i^s k_i^t/n$. Thus, the similarity Σ_{ii} can be computed as the difference between the actual number of shared neighbors q_{ii} and σ_{ii}, which leads to:

$$\begin{aligned}
\Sigma_{ii} &= q_{ii} - \sigma_{ii} = \sum_k A_{ik}^s A_{ik}^t - \frac{k_i^s k_i^t}{n} \\
&= \sum_k A_{ik}^s A_{ik}^t - n\overline{k}_i^s \overline{k}_i^t \\
&= \sum_k \left(A_{ik}^s A_{ik}^t - \overline{k}_i^s \overline{k}_i^t\right) \\
&= \sum_k \left(A_{ik}^s - \overline{k}_i^s\right)\left(A_{ik}^t - \overline{k}_i^t\right),
\end{aligned} \tag{7}$$

where the notation $\bar{k}$ represents the degree normalized by the number of nodes in the network:

$$\bar{k}_i = \frac{1}{n}\sum_k A_{ik}. \tag{8}$$

Eq. (7) can be seen as a covariance, i.e., a non-normalized correlation, which can be normalized by dividing by standard deviations of the vectors A^s_{ik} and A^t_{ik}, $k = 1\cdots n$. Performing such normalization, the covariance becomes the Pearson correlation measure ρ_{ii}:

$$\rho_{ii} = \frac{\sum_k \left(A^s_{ik} - \bar{k}^s_i\right)\left(A^t_{ik} - \bar{k}^t_i\right)}{\sqrt{\sum_k \left(A^r_{ik} - \bar{k}^s_i\right)^2}\sqrt{\sum_k \left(A^t_{ik} - \bar{k}^t_i\right)^2}}, \tag{9}$$

which ranges between -1 and 1. Just to keep the range of the similarity metrics in the interval between zero and 1, the following linear transformation was performed in ρ_{ii}, deriving ρ'_{ii}:

$$\rho'_{ii} = \frac{\rho_{ii} + 1}{2}. \tag{10}$$

Thus to interpret if ρ'_{ii} is high, it is sufficient to verify $\rho'_{ii} > 0.5$, since the threshold 0.5 corresponds to the similarity obtained when the number of shared neighbors is the same as expected by chance.

4.2.3. Leicht–Holme–Newman index

An alternative to quantify how the number of shared neighbors is greater than expected would be to check the ratio between the actual and expected values, instead of calculating the difference, as was done in the derivation of Σ in Eq. (7). In this case, the similarity coefficient, referred to as Leicht–Holme–Newman Index [39], is given by

$$\tau_{ii} = \frac{q_{ii}}{\sigma_{ii}} = n\frac{\sum_k A^s_{ik}A^t_{ik}}{\sum_k A^s_{ik}\sum_k A^t_{ik}}. \tag{11}$$

The threshold to be analyzed is 1. If τ_{ii} is above 1, the similarity is higher than expected. Otherwise, the value is less than 1 but always positive. It is still worth noting the resemblance of τ with c defined in Eq. (6), since while the former divides q_{ii} by $k^s_i k^t_i$, the latter divides q_{ii} by $\sqrt{k^s_i k^t_i}$. Even though in principle the difference between these measures is small, some authors suggest that τ is far more effective since it presents a well-defined threshold to interpret similarity [38,39].

4.2.4. Similarity based on the euclidean distance

This measure was derived using again the neighboring vectors A^s_{ik} and A^t_{ik}. The similarity between the two vectors is obtained by calculating the euclidean squared distance between them. The distance is then normalized by the maximum possible distance $k^s_i + k^t_i$, which occurs when there are no shared neighbors. Thus the resulting distance can be expressed as

$$d_{ii} = \frac{\sum_k \left(A^s_{ik} - A^t_{ik}\right)^2}{k^s_i + k^t_i} = \frac{\sum_k \left(A^s_{ik} + A^t_{ik} - 2A^s_{ik}A^t_{ik}\right)}{k^s_i + k^t_i} = 1 - 2\frac{q_{ii}}{k^s_i + k^t_i}. \tag{12}$$

Note that once again the index ranges from 0 to 1. Therefore, to obtain the corresponding similarity index κ_{ii}, the complement is taken:

$$\kappa_{ii} = 1 - d_{ii} = 2\frac{q_{ii}}{k^s_i + k^t_i}. \tag{13}$$

Interestingly, κ_{ii} can be seen as a variation of c_{ii}, since while the latter normalizes q_{ii} by the geometric mean, the former normalizes q_{ii} using the arithmetic mean.

4.3. Similarity indices based on both topological and semantic features

4.3.1. Katz similarity

While in Section 4.2 the number of shared neighbors played a prominent role in quantifying the similarity, the index derived here is based on the idea that two vertices are similar if the neighbors of one node are similar to the neighbors of the other node. In other words, two nodes need not share the same neighbors to be considered similar, they only need to have neighbors which are similar. To develop the measure, the similarity between all possible pairs of nodes of the network

is computed. Then, the correlation of such similarity is verified for the nodes with the same label (i.e., they refer to the same word) in the networks being compared. The description of these two steps is given below.

Initially, to store the similarity between the pairs of nodes i and j of one of the two networks being compared, the variable ς_{ij} is created. Assuming that ς_{ij} is proportional to the similarity of the corresponding neighbors k and l, then ς_{ij} can be recursively defined as

$$\varsigma_{ij} = \alpha \sum_k \sum_l A_{ik} A_{jl} \varsigma_{kl}. \tag{14}$$

Even though ς_{ij} seems to be a consistent measure, Ref. [38] highlights some drawbacks arising from this definition. For example, ς_{ij} does not necessarily take high values when the self-similarity ς_{ii} is computed. Consequently, nodes with many common neighbors can be overlooked. To solve this problem, Ref. [38] suggests adding an artificial term to ensure that ς_{ii} takes high values. The modification leads to a new definition of ς:

$$\varsigma_{ij} = \alpha \sum_k \sum_l A_{ik} A_{jl} \varsigma_{kl} + \delta_{ij}. \tag{15}$$

Isolating ς in Eq. (15), it can be written as a summation of the number of paths of even length connecting the nodes i and j. Obviously, there is no reason for using only paths of even lengths. For this reason, ς is redefined as

$$\varsigma_{ij} = \alpha \sum_k \sum_l A_{ik} \varsigma_{kj} + \delta_{ij}, \tag{16}$$

which leads to the following closed form:

$$\varsigma = (I - \alpha A)^{-1} = \sum_{i=0}^{\infty} (\alpha A)^i, \tag{17}$$

where I represents the identity matrix. Assuming that i and j are similar if i has k as neighbor and k is similar to j, then the closed solution also takes into account paths of odd lengths, since the summation is performed over the integers. With regard to the α parameter, to guarantee that the summation in Eq. (17) converges, it must lie in the interval $\alpha < \lambda_1^{-1}$, where λ_1 is the largest eigenvalue of A. In particular, we have chosen $\alpha = \lambda_1^{-1}/2$.

After calculating the similarity between all pairs of nodes of the networks under comparison, we compute the similarity using Hubert's coefficient:

$$\Gamma = \frac{1}{\Delta_s \Delta_t} \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\varsigma_{ij}^s - \mu_s)(\varsigma_{ij}^t - \mu_t), \tag{18}$$

where μ and Δ are given by

$$\mu = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \varsigma_{ij} \tag{19}$$

$$\Delta = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\varsigma_{ij} - \mu)^2. \tag{20}$$

Upon defining Γ as shown in Eq. (18), the networks are considered similar to each other in case the numbers of paths between every pair of nodes are correlated. In other words, Γ is high when strongly connected pairs of nodes in one of the networks tend to be also strongly connected in the other network. Analogously, Γ also takes high values when weakly connected pairs of nodes in the first network are weakly connected in the second network.

4.3.2. Similarity based on the ability to match nodes

One of the recent areas of research in complex networks encompasses the analysis of the interrelationship between networks, which contrasts with the study of isolated networks. For example, in communication networks, there exists a duality between online acquaintanceship networks and the network of phone contacts. In fact, this happens precisely because both networks are actually social networks [40–43]. In language networks, the same effect occurs, since a given word can display the same pattern in different languages, especially if the languages have a common origin [19]. Based on this interrelationship, we developed a similarity index which works in two steps. First, a heuristic is applied to perform the matching between nodes of the networks. Then, the quality of matching is evaluated by counting the number of accurate matching (i.e. the number of associations in which the associated labels correspond to the same word). In particular, we assume that the similarity is directly proportional to this accuracy rate, since similar texts are expected to share semantic as well as topologic properties.

The method employed to map nodes previously discussed can be separated into two steps.

(i) *Computation of similarity*: the similarity between two vertices of distinct networks can be computed using structural or semantic information. In the first case, referred to as topologic matching, the similarity is calculated using the local relative difference of topologic measurements, as defined in Eq. (2). In the second case, referred to as semantic matching, we used the cosine similarity defined in Eq. (6).

(ii) *Mapping*: representing similarities computed in (i) as a bipartite network where the weights of the links represent the similarities, we applied the KM algorithm [44] in order to find the pairs which maximize the sum of the matching links. Actually, this algorithm does not always find the best matching, since it is a heuristic to avoid evaluating all possibilities and enhance the efficiency in processing time. Nevertheless, more often than not, the matching found by the heuristic is very similar to the best matching [45].

4.3.3. Similarity indices based on the preservation of local measurements (slope)

Similarly to the previous measure, we used a mapping to evaluate the similarity between texts. However, while the previous one assesses the accuracy rate of the mapping from the information of the similarity between all pairs of nodes, this one evaluates the variation of the topological measurements knowing in advance the correct mapping. That is to say, the measurements for words with the same labels in the networks are compared. The procedure to perform this comparison begins by plotting the measurements extracted from both texts, node by node, so each distinct measurement leads to a scatter plot. Thus if a given measurement μ is computed for node v, μ_v^t represents the value of μ for v in one network and if μ_v^s represents μ for the same node in the other network, then the point (μ_v^t, μ_v^s) will belong to the scatter plot. Three descriptors are extracted from each scatter plot: the y-intercept, the slope and the Pearson product–moment correlation coefficient, obtained from the best straight line approximated by the least squares method. These descriptors are important because information about the preservation of metrics can be obtained. In particular, if the y-intercept is close to zero and the slope and Pearson are close to 1, then the texts are similar to each other. Although they cannot be considered as self-contained similarity indices (since they are mutually dependent), these indices are still useful to capture the ability to preserve local measurements. In fact, to illustrate the use of such measures and compare them with the other similarity indices, we computed the three coefficients for each of the topological measurement and used them as attributes of the texts.

5. Results and discussion

5.1. Evaluation of the quality of machine translation

Evaluating the quality of MT has been as important as it is a difficult task. For obvious reasons, human evaluation is the most reliable, but it is too costly for large scale use, in addition to the problems of lack of agreement among distinct evaluators. These difficulties have motivated the development of several parameters to assess the quality of MT systems, the most used of which have been BLEU [46] (Bilingual Evaluation Understudy) and NIST [47] (National Institute of Standards and Technology) indices. The latter quantify the quality of a translation according to the number of words appearing in both the translation and in one or more reference texts. Significantly, these parameters have been shown to correlate well with human judgment [46].

In our analysis we shall assume that high BLEU or NIST scores mean high quality of the translation, and therefore these scores will be a sort of golden standards. To verify the ability of the indices proposed to quantify similarity, we calculated the Pearson correlation between these indices and the two golden standards. The closer to 1 the Pearson coefficient the more appropriate is the parameter to quantify the translation quality.

In the experiments, we used a set of 100 pieces of text compiled manually from the online edition of the Brazilian magazine Pesquisa FAPESP [48]. The magazine is also available in Spanish [49] and English [50], and therefore these human translations were used as reference (referred to here as golden standards). The translations to be evaluated were generated with the following machine translators: Google [51] (for Portuguese–English and Portuguese–Spanish), Bing [52] (Portuguese–English and Portuguese–Spanish), Apertium [53] (Portuguese–Spanish) and InterTran [54] (Portuguese–English). Thus, for each pair of languages, we obtained 300 translations (target texts), 100 original texts (source texts) and 100 reference texts (golden standards).

The results for the pairs Spanish–Portuguese and English–Portuguese are shown in Tables 1 and 2, respectively. The indices based solely on co-occurrence of words, which basically amounts to capturing semantic features, give the highest correlations (above 0.9). This result was expected because the BLEU and NIST scores are themselves based on co-occurrence of words between the translated text and a reference text. With regard to the indices using topological measurements, the Katz parameter was the only one to provide high correlation with BLEU and NIST scores. It seems that somehow the number of paths between concepts is useful to evaluate the quality of translations in the same way that it has been useful to characterize the topology of various networks [55]. To further explore this ability to predict the semantic quality through structural analysis, we also calculated the correlation between the variation of each global measurement of the complex networks and the scores BLEU and NIST (results not shown). In both language pairs, we observed reasonable correlations for the variation of the standard deviation of betweenness (about 0.55). For the pair Spanish–Portuguese, we also observed

Table 1
Absolute values for the correlation between the (dis)similarity indices based on complex networks and the golden standards BLEU and NIST for the translations from Spanish to Portuguese. Semantic, topologic and hybrid measures are identified as (S), (T) and $(S+T)$, respectively. As expected, the predominantly semantic metrics correlate better with the golden standards. Interestingly, the Katz metric, which uses topological measurements to compare texts, also correlates strongly with BLEU and NIST.

(Dis)similarity Index	BLEU	NIST
Semantic matching (S)	0.95	0.61
Cosine (S)	0.94	0.70
Pearson (S)	0.93	0.68
Euclidean (S)	0.91	0.66
Katz $(S+T)$	0.81	0.78
Leicht–Holme–Newman (S)	0.70	0.44
Topologic matching $(S+T)$	0.61	0.26
Motifs (T)	0.36	0.26

Table 2
Absolute values of the correlation between the (dis)similarity indices based on complex networks modeling and the golden standards BLEU and NIST for the translations from Portuguese to English. Semantic, topologic and hybrid measures are identified as (S), (T) and $(S+T)$, respectively. As expected, the predominantly semantic metrics correlate better with the golden standards. Similarly to the result obtained for Spanish and English, the Katz similarity index also correlates strongly with BLEU and NIST.

(Dis)similarity index	BLEU	NIST
Cosine (S)	0.97	0.79
Pearson (S)	0.96	0.78
Euclidean (S)	0.95	0.76
Katz $(S+T)$	0.94	0.87
Semantic matching (S)	0.94	0.80
Leicht–Holme–Newman (S)	0.54	0.38
Topologic matching $(S+T)$	0.37	0.20
Motifs (T)	0.33	0.18

Table 3
Accuracy rate to distinguish machine translations from Google, Apertium and Bing using several similarity indices. The texts were translated from Spanish to Portuguese and the similarity indices were used to quantify the difference between the machine translation and the translation taken as a reference (human translation). Semantic, topologic and hybrid measures are identified as (S), (T) and $(S+T)$, respectively.

Similarity index	Accuracy rate (%)	ML algorithm
Pearson (S)	65	kNN-5
Slope $(S+T)$	63	Naive Bayes
Cosine (S)	60	Naive Bayes
Euclidean (S)	59	Ripper
Semantic matching (S)	59	C4.5
Topologic measures (T)	58	C4.5
Topologic matching $(S+T)$	53	C4.5
BLEU (S)	**51**	C4.5
NIST (S)	**50**	C4.5
Katz $(S+T)$	48	kNN-5
Motifs (T)	45	Ripper
Leicht–Holme–Newman (S)	37	kNN-1

a correlation between the standard deviation of out degree (about 0.52). This means that even without making use of any information about the nodes labels, it is still possible to predict with reasonable accuracy the semantic quality of translations, which depends on the information on labels.

One may therefore conclude that topology can be useful to quantify similarity in this type of application. Of special relevance is the Katz index, which besides being the only metric using topology that achieved a reasonably high correlation with the BLEU metric, it was also the index that best correlated with the NIST index (see Table 3). Thus, the hypothesis that stylistic factors combined with semantic factors may be useful for evaluating the quality of automatic translations (especially for the Katz index to predict NIST) is confirmed.

5.2. Classifying translations

For investigating how topology and semantics can be used to classify translations according to the source translator, we employed four translators: InterTran, Google, Bing and Apertium. The database is the same as the one employed in Section 5.1. To compare the ability to distinguish translations using the proposed and the traditional semantic-based indices, we also computed the accuracy rate when NIST and BLEU were used as attributes of machine learning (ML) algorithms [56].

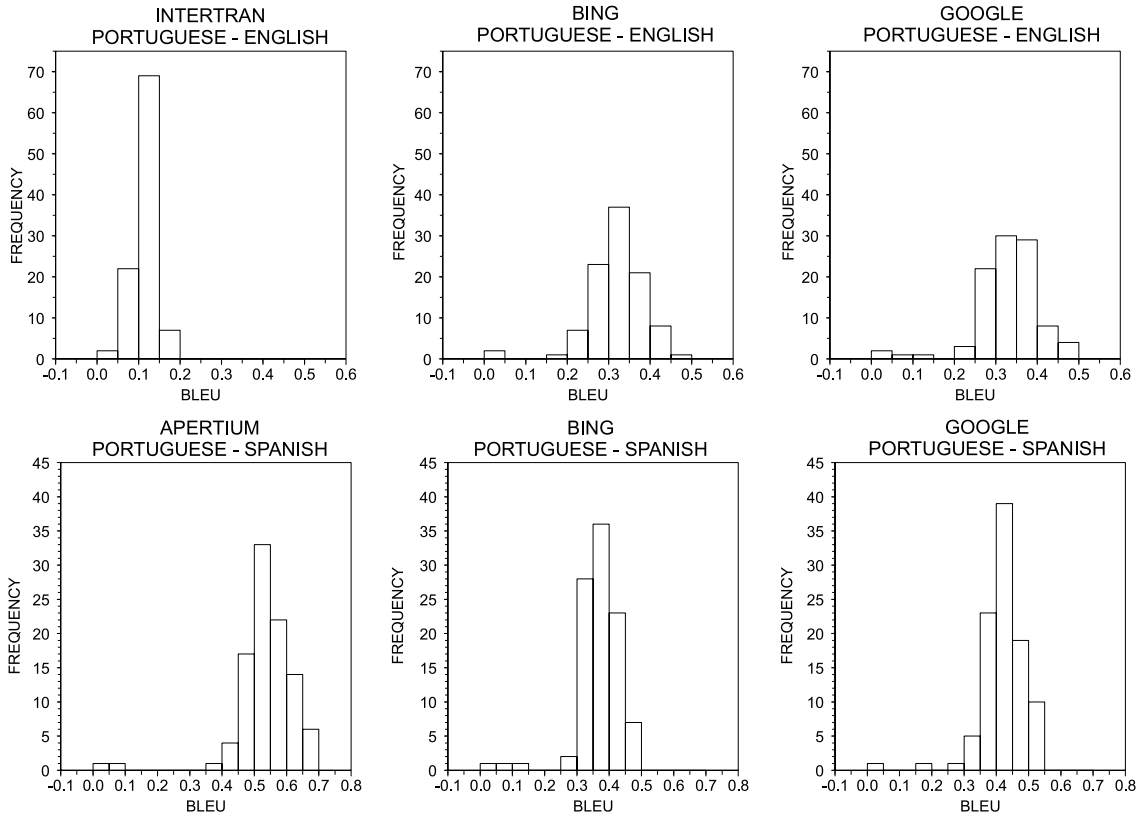

Fig. 2. Distribution of values of BLEU for Portuguese–English (top panel) and Portuguese–Spanish (bottom panel) in the corpus. Note that while Bing and Google show similar distributions, Apertium and Intertran display quite distinct distributions from Bing and Google.

The ML inductors used were: C4.5 [57], Naive–Bayes [58], RIPPER [59] and kNN[1] [60]. Specifically, those algorithms were used because they have been employed to discriminate texts in previous applications [20]. To evaluate the quality, we used the 10-fold-cross validation strategy [61], which continuously selects 9 folds from the training set (the choice is made randomly) to train the inductors and uses the remaining fold to evaluate the classifier generated. To illustrate how difficult it is to distinguish the translators, we analyzed the distribution of BLEU for each translator in Fig. 2. The conventional indices provide excellent distinction between Apertium/InterTran and Google/Intertran, but there is some overlap between Google and Bing, for both pairs of languages. Therefore, it is expected that the topological (dis)similarity indices could be useful to enhance the distinction among these translators.

The results concerning the Spanish–Portuguese pair are illustrated in Table 3, which reveals that the traditional metrics (BLEU and NIST) are outperformed by other metrics including topology. There is a considerable difference between the accuracy rate for the similarity based on the slope (which uses both semantic and topological information) and BLEU. Even the similarity based solely on topology (topological measurements) outperformed the BLEU metric. It appears that the topological analysis is able to provide useful information that was not possible to grasp by the traditional semantic analysis. Interestingly, the metric based on motifs, which is also completely based on topology (the label of the nodes is not employed to detect motifs), reached only 37% accuracy at best. This means that probably other motifs will need to be introduced in the analysis if the performance is to be improved.

Table 4 summarizes the best accuracy rates in the distinction between Apertium and Google. As expected, the rate increased substantially, since the quality of Google and Apertium are quite different, as shown in Fig. 2 (the difference can also be easily noticed by manually inspecting short translated extracts). As for the similarity indices, the measure based on the slope index again achieved the best results, although very close to the BLEU accuracy rate. This result confirms that, even comparing different quality of translations from the semantic point of view, it is still possible to improve the ability to characterize text through the addition of topology-based metrics. With regard to other similarity metrics, their ranking in Table 4 seems to have been maintained when compared with the ranking in Table 3.

A similar behavior was observed for the translations involving the English language. For example, Table 5, which illustrates the best accuracy rates in the classification between InterTran, Bing and Google, shows that the accuracy rates

[1] In the k-nearest neighbor algorithm we used k ranging from $k = 1$ (kNN-1) to $k = 5$ (kNN-5).

Table 4
Accuracy rate to distinguish machine translations from Google and Apertium using several similarity indices. The texts were translated from Spanish to Portuguese and the similarity indices were used to quantify the difference between the machine translation and the translation taken as a reference (human translation). Semantic, topologic and hybrid measures are identified as (S), (T) and ($S+T$), respectively.

Similarity index	Accuracy rate (%)	ML algorithm
Slope ($S+T$)	88	C4.5
BLEU (S)	**84**	Naive Bayes
Euclidean (S)	82	kNN-5
Cosine (S)	81	kNN-5
Pearson (S)	80	C4.5
Semantic matching (S)	77	C4.5
Topologic matching ($S+T$)	75	Ripper
Topologic measures (T)	74	kNN-5
Katz ($S+T$)	69	kNN-3
NIST (S)	**67**	C4.5
Leicht–Holme–Newman (S)	59	kNN-3
Motifs	59	kNN-5

Table 5
Accuracy rate to distinguish machine translations from Google, Intertran and Bing. The texts were translated from Portuguese to English and the similarity indices were used to quantify the difference between the machine translation and the translation taken as a reference (human translation). Semantic, topologic and hybrid measures are identified as (S), (T) and ($S+T$), respectively.

Similarity index	Accuracy rate (%)	ML algorithm
Semantic matching (S)	68	Ripper
Katz ($S+T$)	67	Ripper
BLEU (S)	**67**	kNN-5
Cosine (S)	65	Naive Bayes
Pearson (S)	65	Naive Bayes
Euclidean (S)	65	Naive Bayes
NIST (S)	**64**	Naive Bayes
Leicht–Holme–Newman (S)	62	Naive Bayes
Slope ($S+T$)	53	Naive Bayes
Topologic measures (T)	51	Naive Bayes
Topologic matching ($S+T$)	43	Naive Bayes
Motifs (T)	40	kNN-5

Table 6
Accuracy rate to distinguish machine translations from Google and Intertran. The texts were translated from Portuguese to English and the similarity indices were used to quantify the difference between the machine translation and the translation of reference (human translation).

Similarity index	Accuracy rate (%)	ML algorithm
Cosine (S)	96	kNN-5
Pearson (S)	96	kNN-5
Euclidean (S)	96	Ripper
Semantic matching (S)	95	Naive Bayes
Katz ($S+T$)	95	Naive Bayes
BLEU (S)	**95**	kNN-5
NIST (S)	**92**	kNN-5
Leicht–Holme–Newman (S)	90	Naive Bayes
Slope ($S+T$)	80	Naive Bayes
Topologic measures (T)	75	kNN-5
Topologic matching ($S+T$)	60	Naive Bayes
Motifs (T)	59	kNN-5

of Cosine, Euclidean and Pearson similarities remained quite close. Also, the metric based on motifs still led to the worst accuracy rates. On the other hand, the Katz and semantic matching similarity, which are based on both semantics and topology, displayed the highest accuracy rates, along with the BLEU measure. The same applies to Table 6 in the ability to distinguish between Google and InterTran.

In summary, the experiments with translation confirmed the hypothesis that topological measurements in conjunction with semantic features are able to improve the quantification of similarity in written texts, even to a small extent. Taken separately, the purely semantic metrics (such as the similarity based on the Pearson coefficient) outperformed the metrics based exclusively on topological features (such as topological measurements and motifs). This confirms that the nature of the problem studied is mainly semantic. In fact, it seems that the main factor in distinguishing quality is the use of correct words, and this is the probable reason why the semantic analysis has become the standard analysis [62]. To illustrate a scenario where both structure and semantics can be used for different purposes, we apply in the next section semantic and stylistic indices to detect authorship in poetry and prose.

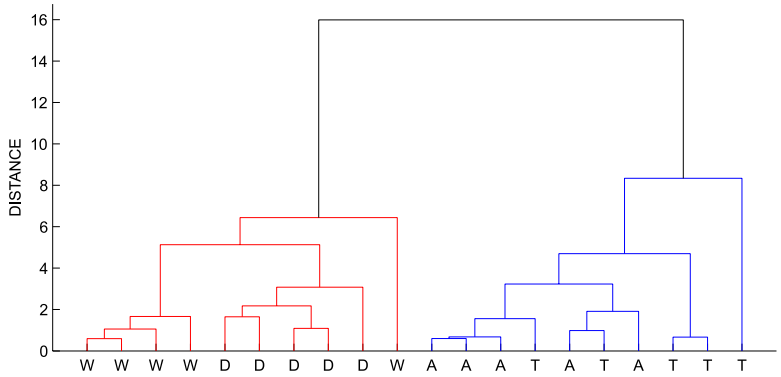

Fig. 3. Hierarchical clustering obtained using structural features to distinguish between Whitman (W), Thomas (T), Tennyson (A) and Dickinson (D).

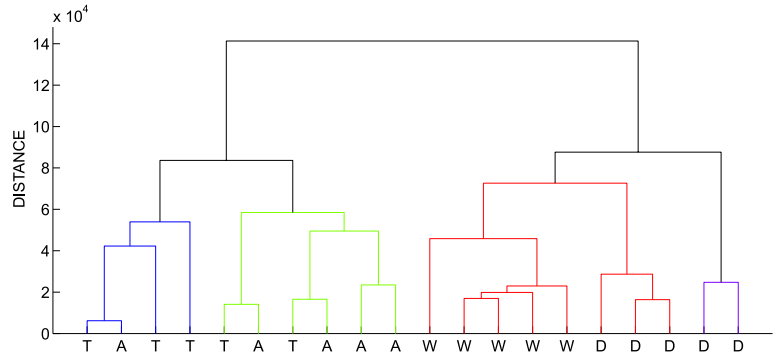

Fig. 4. Hierarchical clustering obtained using semantic features to distinguish between Whitman (W), Thomas (T), Tennyson (A) and Dickinson (D).

5.3. Topological similarity and applications related to the text style

As a third application, we examined how the structure and semantics are interrelated in the task of recognition of authorship. Two corpora were used in this experiment, one of poetry and another with prose. Several poems by Emily Dickinson, Alfred Tennyson, Dylan Thomas and Walt Whitman were obtained from an online repository [63]. Because they are generally short, in some cases poems by the same author were juxtaposed to obtain sufficiently long texts for the statistical analysis. As for the corpus of texts in the prose format, we collected 5 books from the Gutenberg Project [64] for each of the following authors: Arthur Conan Doyle, Charles Darwin, Thomas Hardy and Bram Stoker. More specifically, we used the first 18,000 words of each book to build the networks and compute (dis)similarity indices.

We applied the dissimilarity metric based on the Euclidean distance (see Eq. (12) in Section 4.2.4), which is based on semantic features, and the dissimilarity metric based on topological features (see Section 4.1.1) to recognize authorship in poems. The hierarchy obtained with the Ward method [65] using the topological and semantic dissimilarity metrics are illustrated in Figs. 3 and 4, respectively. The ability to distinguish authors seems to be equivalent. It turns out that for this literary style, both semantic and structure appear to be relevant factors to characterize authorship. Interestingly, these results are consistent with those from the evaluation of machine translation, once semantics and structure are roughly equivalent.

A second experiment on authorship recognition was carried out with a corpus on prose involving four authors of story books. The hierarchies obtained are illustrated in Figs. 5 and 6, which point to a high correlation between the semantic and topological paradigms, since the ability to distinguish among authors is quite similar. However, a more refined analysis indicates that different patterns do emerge. Consider, for example, Stoker and Darwin. While the topological analysis reveals that they display similar writing styles, the semantic contents in their texts are quite different. In other words, the writing style is shared, even though they write about completely different subjects.[2] Analogously, Stoker and Doylan shared semantic contents, but used different styles. Overall, the suitable methods to use in the classification depend on the purpose, whether one wishes to distinguish writing style or topics. Furthermore, semantic and topological features may be combined to identify authors when many authors are to be distinguished. For authors with the same style can eventually be

[2] Further analyzing the works of both authors, we confirmed that the themes developed by each one are different. For Stocker wrote story books, while Darwin compiled scientific manuscripts.

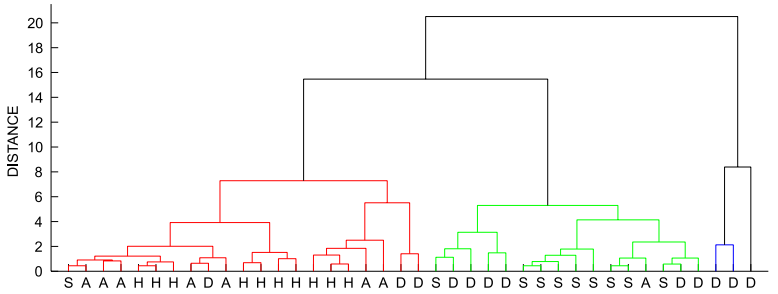

Fig. 5. Hierarchical clustering using structural features. The authors in the hierarchy are: Doyle (A), Darwin (D), Hardy (H) and Stoker (S).

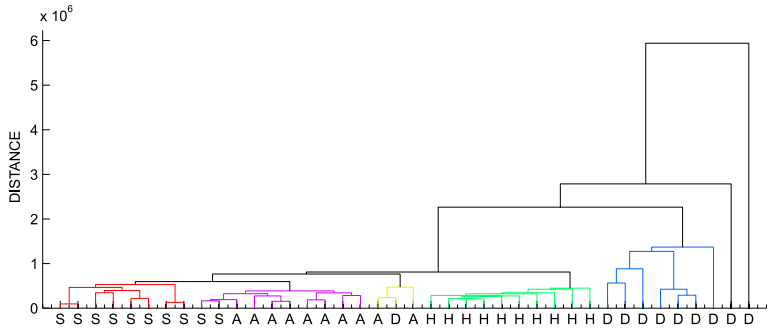

Fig. 6. Hierarchical clustering using semantic features. The authors in the hierarchy are: Doyle (A), Darwin (D), Hardy (H) and Stoker (S).

distinguished by the semantic contents, while authors who write about the same topic can be distinguished by the individual subtleties of style.

6. Conclusion

In this paper, we considered the problem of measuring similarity between pairs of texts, which is relevant in many situations of linguistic interest and have significant consequences for our understanding of textual phenomena. Many studies have been made for quantifying content similarity and classifying texts, but to the best of our knowledge this paper is the first to combine semantic features and topology of complex networks to enhance the performance of real applications. We performed a systematic evaluation for three natural language processing tasks, namely identification of machine translation systems, evaluation of quality of machine translated texts and authorship recognition. More specifically, applying the concepts and methodologies of complex networks to characterize texts according to the stylistic features, we proposed and evaluated several similarity and dissimilarity indices, some of which did not involve any kind of semantic information. Overall we showed that semantic contents are still the most important feature to define similarity. Nevertheless, for some applications the use of topological metrics may be beneficial, especially if combined with semantic evaluation. Of particular importance was the finding that the number of paths between concepts, the standard deviation of betweenness and out-degree seem to be good indicators of translation quality. Furthermore, in authorship recognition topological features may be the key for distinguishing styles. In fact, we have found that the relationship between authors' manuscripts mainly depends on the nature of the similarity index as semantically related authors may have developed completely different writing styles and vice-versa. It is hoped that the approach suggested here may lead to the development of more robust, efficient similarity indices, and boost research of other areas where topology has been shown to effectively characterize written texts.

Acknowledgments

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

Appendix. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.physa.2012.04.011.

References

[1] J.B. Michel, et al., Quantitative analysis of culture using millions of digitized books, Science 331 (2011) 176–182.
[2] A.P. Masucci, A. Kalampokis, V.M. Eguíluz, E. Hernández-García, Wikipedia information flow analysis reveals the scale-free architecture of the semantic space, PLoS ONE 6 (2011) e17333.
[3] F.N. Silva, M.P. Viana, B.A.N. Travençolo, L.F. Costa, Investigating relationships within and between category networks in Wikipedia, Journal of Informetrics 5 (2011) 431–438.
[4] D.J. Arnold, L. Balkan, S. Meijer, R.L. Humphreys, L. Sadler, Machine Translation: An Introductory Guide, Blackwells-NCC, 1993.
[5] K.S. Jones, Automatic summarising: the state of the art, Information Processing and Management 43 (6) (2007) 1449–1481.
[6] S. Lawrence, C.L. Giles, Accessibility of information on the web, Nature 400 (1999) 107.
[7] J. Mostafa, Seeking better Web searches, Scientific American 292 (2) (2005) 66–73.
[8] S.M. Aluísio, O.N. Oliveira Jr., A case-based approach for developing writing tools aimed at non-native english users, in: Proceedings of the First International Conference, in: Lecture Notes in Artificial Intelligence, vol. 1010, Springer-Verlag, 2005, pp. 121–132.
[9] S.M. Aluísio, C. Gasperin, Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts, in: Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, ACL 2010, vol. 1, 2010, pp. 46–53.
[10] L. Doyle, J. Becker, Information Retrieval and Processing, Melville, 1975.
[11] G.A. Miller, WordNet: a lexical database for english, Communications of the ACM 38 (11) (1995) 39–41.
[12] S.M. Aluísio, J.M. Pelizzoni, A.R. Marchi, L.H. Oliveira, R. Manenti, V. Marquivafável, An account of the challenge of tagging a reference corpus of Brazilian Portuguese, in: Proceedings of the International 6th Workshop PROPOR, 2003.
[13] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, Self organisation of a massive document collection, IEEE Transactions on Neural Networks 11 (3) (2000) 574–585.
[14] M. Damashek, Gauging similarity with *n*-grams: language-independent categorization of text, Science 267 (1995) 843–848.
[15] T.K. Landauer, S.T. Dumais, A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychological Review 104 (2) (1997) 211–240.
[16] M.D. Lee, B. Pincombe, M. Welsh, An empirical evaluation of models of text document similarity, in: Proceedings of the 27th Annual Conference of the Cognitive Science Society, 2005, pp. 1254–1259.
[17] Y. Yang, An evaluation of statistical approaches to text categorization, Information Retrieval 1 (1999) 69–90.
[18] L. Antiqueira, M.G.V. Nunes, O.N. Oliveira Jr., L.F. Costa, Strong correlations between text quality and complex networks features, Physica A 373 (2007) 811–820.
[19] D.R. Amancio, L. Antiqueira, T.A.S. Pardo, L.F. Costa, O.N. Oliveira Jr., M.G.V. Nunes, Complex networks analysis of manual and machine translations, International Journal of Modern Physics C 19 (4) (2008) 583–598.
[20] D.R. Amancio, M.G.V. Nunes, O.N. Oliveira Jr., T.A.S. Pardo, L. Antiqueira, L.F. Costa, Using metrics from complex networks to evaluate machine translation, Physica A 390 (1) (2011) 131–142.
[21] D.R. Amancio, R. Fabbri, O.N. Oliveira Jr., M.G.V. Nunes, L.F. Costa, Opinion discrimination using complex network features, in: 2nd Workshop on Complex Networks, 2011.
[22] D.R. Amancio, M.G.V. Nunes, O.N. Oliveira Jr., L.F. Costa, Extractive summarization using complex networks and syntactic dependency, Physica A 391 (2012) 1855–1864.
[23] H. Liu, The complexity of Chinese dependency syntactic networks, Physica A 387 (2008) 3048–3058.
[24] R. Ferrer i Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Physical Review E 69 (2004) 051915.
[25] H. Liu, W. Li, Language clusters based on linguistic complex networks, Chinese Science Bulletin 55 (2010) 3458–3465.
[26] O. Abramov, A. Mehler, Automatic language classification by means of syntactic dependency networks, Journal of Quantitative Linguistics 18 (2011) 291–336.
[27] H. Liu, C. Xu, Can syntactic networks indicate morphological complexity of a language? Europhysic Letters 93 (2011) 28005.
[28] S. Yu, H. Liu, C. Xu, Statistical properties of Chinese phonemic networks, Chinese Science Bulletin 54 (2009) 2781–2785.
[29] C.D. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, 1999.
[30] F. Menczer, Evolution of document networks, Proceedings of the National Academy of Sciences of the United States of America 101 (2004) 5261–5265.
[31] A. Mehler, Structural similarities of complex networks: a computational model by example of Wiki graphs, Applied Artificial Intelligence 22 (2008) 619–683.
[32] O. Uzuner, Identifying expression fingerprints using linguistic information, Ph.D. Thesis, 2005.
[33] D.R. Amancio, O.N. Oliveira Jr., L. da F. Costa, Complex network analysis of language complexity, New Journal of Physics (submitted for publication).
[34] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, J. Kleinberg, Echoes of power: language effects and power differences in social interaction, arXiv:1112.3670, 2011.
[35] H. Liu, Statistical properties of chinese semantic networks, Chinese Science Bulletin 54 (2009) 2781–2785.
[36] L.F. Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Characterization of complex networks: a survey of measurements, Advances in Physics 56 (2005) 167–242.
[37] R. Milo, S.S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (2002) 824–827.
[38] M.E.J. Newman, Networks: An Introduction, Oxford University Press, 2010.
[39] E.A. Leicht, P. Holme, M.E.J. Newman, Vertex similarity in networks, Physical Review E 73 (2006) 026120.
[40] D.J.S. Price, Networks of scientific papers, Science 149 (1965) 510–515.
[41] L.C. Freeman, Centrality in social networks: conceptual clarification, Social Networks 1 (1979) 215–239.
[42] S. Milgram, The small world problem, Psychology Today 2 (1967) 60–67.
[43] D.J. Watts, A twenty-first century science, Nature 445 (2007) 489.
[44] R.A. Pilgrim, Munkres' assignment algorithm modified for rectangular matrices. Available at http://csclab.murraystate.edu/bob.pilgrim/445/munkres.html.
[45] Q. Xuan, T.J. Wu, Matching between complex networks, Physical Review E 80 (2009) 026103.
[46] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[47] Automatic Evaluation of Machine Translation Quality Using *N*-gram Co-Occurrence Statistics, http://www.itl.nist.gov/iad/mig//tests/mt/doc/ngram-study.pdf.
[48] http://revistapesquisa.fapesp.br/.
[49] http://revistapesquisa.fapesp.br/?lg=es.
[50] http://revistapesquisa.fapesp.br/?lg=en.
[51] http://translate.google.com/.
[52] http://www.microsofttranslator.com/.
[53] http://www.apertium.org/.
[54] http://www.tranexp.com/win/itserver.htm.
[55] F.A. Rodrigues, L.F. Costa, Generalized connectivity between any two nodes in a complex network, Physical Review 81 (2010) 036113–036123.
[56] C.M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, New York, 2006.
[57] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[58] G.H. John, P. Langley, Estimating continuous distribution in bayesian classifiers, in: 11 Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338–345.
[59] W.W. Cohen, Fast effective rule induction, in: 12 International Converence on Machine Learning, 1995, pp. 115–223.
[60] D. Aha, D. Kibler, Instance based learning algorithms, Machine Learning 6 (1991) 37–66.
[61] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, 1995, pp. 1137–1143.
[62] J. Turian, L. Shen, I.D. Melamed, Evaluation of machine translation and its evaluation, in: Proceedings of the MT Summit IX, 2003, pp. 386–393.
[63] http://famouspoetsandpoems.com.
[64] http://www.gutenberg.org/.
[65] J.H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (301) (1963) 236–244.