Physica A 390 (2011) 131–142

Contents lists available at ScienceDirect

Physica A

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

Using metrics from complex networks to evaluate machine translation

D.R. Amancio [a,*], M.G.V. Nunes [b], O.N. Oliveira Jr. [a], T.A.S. Pardo [b], L. Antiqueira [a], L. da F. Costa [a]

[a] Institute of Physics of São Carlos, University of São Paulo, P.O. Box 369, Postal Code 13560-970, São Carlos, São Paulo, Brazil

[b] Institute of Mathematics and Computer Science, University of São Paulo, P.O. Box 668, Postal Code 13560-970, São Carlos, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 23 March 2010
Received in revised form 17 August 2010
Available online 9 September 2010

Keywords:
Machine translation
Evaluation
Complex networks
Machine learning

ABSTRACT

Establishing metrics to assess machine translation (MT) systems automatically is now crucial owing to the widespread use of MT over the web. In this study we show that such evaluation can be done by modeling text as complex networks. Specifically, we extend our previous work by employing additional metrics of complex networks, whose results were used as input for machine learning methods and allowed MT texts of distinct qualities to be distinguished. Also shown is that the node-to-node mapping between source and target texts (English–Portuguese and Spanish–Portuguese pairs) can be improved by adding further hierarchical levels for the metrics out-degree, in-degree, hierarchical common degree, cluster coefficient, inter-ring degree, intra-ring degree and convergence ratio. The results presented here amount to a proof-of-principle that the possible capturing of a wider context with the hierarchical levels may be combined with machine learning methods to yield an approach for assessing the quality of MT systems.

1. Introduction

The discovery that many natural as well as artificial systems exhibit the topology of scale-free networks has boosted the area of complex networks (CN) [1], in which concepts of graph theory [2,3] and statistical physics [4,5] are merged. The many systems modeled with scale-free networks include the group of actors in Hollywood, the metabolism, and social networks [1]. This rich variety of systems reflects the rich-get-richer paradigm [6], which seems to be pervasive in Nature and human-created networks. The breadth and importance of such applications have also led to theoretical developments, e.g. in establishing requirements for connectivity and in creating novel metrics to assess distinct aspects of network topology and dynamics [7]. By way of illustration, there are more than a hundred metrics currently being used [8]. Of particular relevance for this work was the finding that texts can also be represented in scale-free networks [9]. This has allowed research into semantic networks [10,11], thesauri [12] and in natural language processing tasks. For example, metrics of CN have been used to assess the quality of written essays by high school students [13], in identification of authorship [14] and in building strategies for automatic summarization [15]. In this paper we extend our earlier work where machine translation texts were evaluated automatically by modeling the texts in the source and target languages as graphs [16].

Machine translation has experienced considerable changes with the adoption of statistical models as an alternative to the translation grammars manually produced or, in more recent approaches, automatically learned from a corpus [17]. Though using statistics — instead of making explicit the linguistic knowledge — has often been criticized, very good results were obtained even for distant languages (e.g., English and Chinese).[1] It made it possible to generate fast translators with the only requirement of a large aligned parallel corpus for the language pair. The statistical models proposed by IBM [17,18] and their

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

[1] http://www.itl.nist.gov/iad/mig/tests/mt.

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.08.052

recent evolutions are the state-of-the-art in the area. Because comparing different systems is important for consolidating new and old paradigms, automatic and fast evaluation has now an even more important role. One possibility for assessing quality is human evaluation, but this is too expensive, time-consuming, and may contain errors and inconsistencies that invalidate the conclusions. This is why automatic evaluation measures have been preferred, particularly when different systems are to be compared. The automatic metrics are objective, reproducible and yield results fast, which justify the use of methods akin to statistical physics (e.g. complex networks) in the analysis.

The motivation for employing a network-based method to analyze text stems from the fact that connection among concepts (nodes) is also taken into account in the statistics. Considering these connections one can use network statistics to study co-occurrence of concepts and even establish semantic domains. Here we shall show that stylistic features subjacent in the text may be retrieved through CN metrics for distinguishing between human and machine translations, as well as between machine translations of distinct qualities. Especially relevant will be the application of concentric (or hierarchical) measurements [19], which allows for a more comprehensive quantification of the network topology upon taking into account the wider context around each node. Before the results are presented and discussed in Section 3, we describe the methodology for representing texts as CNs, and the metrics for distinguishing the MT texts.

2. Methodology

2.1. Evaluation of MT systems

Any Natural Language Processing (NLP) application depends critically on evaluation, and Machine Translation (MT) is no different [20]. For situations requiring fine-grained judgments, human evaluations of MT texts are necessary, for which usual measures include: (i) adequacy and fluency [21] that measure the extent to which the meaning of the source text has been preserved and whether the target text is fluent, respectively; (ii) the word error rate, derived from the traditional distance edition measure of Levenshtein [22], computes the number of modifications that must be performed for the MT text to be ideal; (iii) the metrics computing the number of lexical, syntactical and semantic errors (see, e.g., Ref. [23]). The subjectivity in such evaluations is troublesome though. This is why automatic evaluation measures have been preferred. The idea of automatic evaluation is to use a weighted average of variable length phrase matches against a set of reference translations. According to different weighting values, this gives rise to a family of metrics which are objective, reproducible and yield results fast. Among the various metrics for automatic assessment, the BiLingual Evaluation Understudy [24] (BLEU) is the most representative, and has been used in international MT contests (e.g., the contests promoted by the National Institute of Standards and Technology (NIST)). There are several variations of BLEU, such as the NIST measure, METEOR [25] and the family of measures proposed by Ref. [26]. BLEU calculates precision — the proportion of the matched n-grams out of the total number of n-grams in the evaluated translation — for each n-gram order (from one up to 4-grams) and combine them via a geometric averaging. Although the information of recall — the proportion of the matched n-grams out of the total number of n-grams in the reference translation — is very important for assessing quality, it is not considered by BLEU since there is more than only one reference translation. To compensate for recall, BLEU uses a brevity penalty to penalize translations too short. The NIST metric is derived from BLEU and is similar to it but differs in considering the information gain from each n-gram: it gives more credit to a system that gets an n-gram match that is difficult [27]. The limitations and weaknesses of BLEU and NIST have been argued [28,25] and the questions they are based on include: (i) how valuable is the generation of correct words and how much credit for putting them in the correct order? (ii) Does the brevity penalty of BLEU adequately compensate for the lack of recall? (iii) Since N-gram matches do not require word-to-word matching, can this result in incorrect matches?

String-based metrics such as BLEU definitely contributed for the quality improvements of machine translation, but it is known that they are not able to discriminate efficiently the translation quality of more sophisticated MT systems. As consequence, metrics which incorporate more linguistically motivated resources have been proposed. Metrics which use synonym information, such as METEOR [25,29] and TERP [30], have obtained better correlation with human judgments. To capture the effect of syntactic information that today's systems incorporate, different resources have been used. Liu and Gildea [31], introduced the use of syntactic structure and dependency information for matching. Owczarzak et al. [32] used dependency graphs with labels from Lexical Functional Grammar, LFG [33] with better results for correlation with human judgment. Additional resources such as information from the dependency parser, textual entailment-based metrics [34] as well as penalties for discontinuous matches and tuning parameters [35] have improved the dependency-based metric's correlation with human judgment. Nevertheless, the BLEU score has been very convenient and precise in ranking translations and no alternative has substituted BLEU/NIST scores in the contests for ranking MT systems. BLEU can then be seen as a standard measure for MT and we shall use it for assessing the adequacy of the application of CN concepts to MT systems.

2.2. Representing texts as complex networks

The representation of texts as complex networks employed here is similar to the modeling described in previous studies [14–16]. Before building networks, stopwords are removed since they are highly frequent but usually irrelevant

Table 1
Adjacency list obtained from the sentence "The projection of the dataset into three dimensions is essential for very large datasets".

Before pre-processing	After pre-processing
The projection of the dataset into three dimensions is essential for very large datasets	Projection dataset three dimension be essential very large dataset

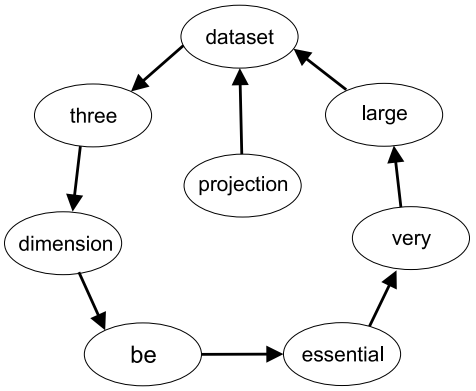

Fig. 1. Network obtained from the sentence "The projection of the dataset into three dimensions is essential for very large datasets".

for representing the concepts.[2] After this lexical reduction phase, the remaining words are lemmatized to combine concepts with the same canonical form, but with different inflections. Moreover, in order to resolve problems of ambiguity, all texts are labeled using the MXPost part-of-speech Tagger based on the Ratnaparki's model [36]. This is useful because, with the help of a computational lexicon [37], words with the same canonical form and the same meaning are grouped into a single node, while words that have the same canonical form but distinct meanings generate distinct nodes. Finally, to construct the network, each word in the pre-processed text is represented as a node and edges are established between words that are adjacent in the text. Additionally, the edges are weighted by the number of times that the corresponding adjacency appears in the text. To illustrate the process, Table 1 and Fig. 1 show respectively the pre-processed form and the subgraph obtained for the sentence "*The projection of the dataset into three dimensions is essential for very large datasets*".

The CN metrics commonly used to analyze textual characteristics are: out-degree (k_{out}), in-degree (k_{in}) and cluster-coefficient (C). The new metrics employed here are shortest path 1 (l_1), shortest path 2 (l_2), shortest path 3 (l_3), cyclic-coefficient (CY) and Search Information (S). We also used the hierarchical metrics [19]: hierarchical common degree, inter-ring degree, intra-ring degree and convergence ratio.

The node's output degree — referred to as out-degree — corresponds to the number of edges emanating from a particular node, also considering the weight. Analogously, the node's input degree — referred to as in-degree — is defined as the number of edges arriving at a given node. The network's k_{out} and k_{in} are evaluated by calculating the average among all the nodes. Note that for a global measure, k_{out} and k_{in} will always be equal. Regarding the adjacency matrix used to represent the network, for a given node i, its k_{out} and k_{in} are calculated using Eqs. (1) and (2), where N represents the number of distinct words in the pre-processed text:

$$k_{\text{out}}(i) = \sum_{j=1}^{N} W_{ji} \tag{1}$$

$$k_{\text{in}}(i) = \sum_{j=1}^{N} W_{ij}. \tag{2}$$

The cluster coefficient (C) is defined as follows. Let S be the set formed by nodes receiving edges of a given node i, and N_c is the cardinality of this set. If the nodes of this group form a completely connected set, then there are N_c ($N_c - 1$) edges in this sub graph. However, if there are only B edges, then the coefficient is given by Eq. (3):

$$C(i) = \frac{B}{N_c(N_c - 1)}. \tag{3}$$

If N_c is less than 1, then C is defined as zero. Note that this measure quantifies how the nodes connected to a specific node are linked to each other, with its value varying between zero and one.

[2] Examples of stopwords are prepositions and articles.

The shortest paths are calculated from all pairs of network nodes, except for cycles. If there is no shorter path between two pairs of vertices, the minimum path is calculated as $W_m \cdot N$, where N represents the number of vertices in the network and W_m is the arithmetic average of the weights of the edges. There are three types of shortest paths. The first one (l_1) takes into account all the edges with unity weight. The second type (l_2) is calculated with the complement of the weights. In other words, the weight of each edge is replaced by the subtraction of the maximum weight found in the network and the weight of the edge. Finally, the third type (l_3) considers for the calculation the edges with inverted weights: the weight P_i for each edge is replaced by P_i^{-1}.

We also used a new measure related to the degrees and entropy of the network, referred to as search information. Frequently, it is desired to find how difficult the task of seeking a vertex in a complex network is on a given path. This difficulty is determined considering the random paths of a network and the probability of a shortest path to be followed randomly. Let SP be a minimum path between the vertices i and b. A random path is chosen randomly, starting from the node i. If $p(i, b)$ is the probability of the minimum path to be followed, so it can be defined as shown in Eq. (4):

$$p(i, b) = \frac{1}{k_i} \prod_{j \in SP} \frac{1}{k_j - 1} \tag{4}$$

where the multiplication is performed along the whole path, except for the initial and final nodes. In a non-directed network k_i is chosen as the degree of the node. Because we use directed graphs, k_i is chosen as the out-degree of node i. S is related to this probability, in that the S of two vertices i and b is related to the entropy of all the minimum paths between these two vertices as shown in Eq. (5):

$$S(i, b) = -\log_2 \sum_{SPs} p(i, b). \tag{5}$$

In order to verify the number of cycles around the network and how this feature is present we use the cyclic coefficient. This is a local measure, so each vertex has a specific value. By definition, it is calculated as in Eq. (6):

$$CY_i = \frac{2}{k_i(k_i - 1)} \sum_j \sum_k \frac{a_{ij} a_{ik}}{T_{ijk}} \tag{6}$$

where the term T_{ijk} represents the minimum size of the cycle that has the nodes i, j and k. The smallest value that T_{ijk} can take is when these three points are connected in a cycle, forming a triangle. Hence, the minimum value for T_{ijk} is 3. The k_i term is the degree of the vertex (the sum of the out-degree with the in-degree). This local formula represents the average of the inverse of the smaller cycles between a node and its neighbor. Thus, this measure is directly proportional to the number of small length cycles.

For the experiments reported here, texts with ca. 500 nodes were used from the science dissemination magazine FAPESP[3], where the source language is Portuguese. These texts were randomly selected from a section referred to as "Laboratory". Translation into Spanish (target language) was carried out by human experts and with two machine translation systems, namely Apertium[4] and Intertran.[5] For English, translation was performed with the MT systems Free Translation[6] and Intertran, in addition to the manual translation. Text alignment of the two versions, i.e. in the source and target languages, was performed with the LIHLA system [38], which allowed a mapping of nodes to be made for testing the preservation of the network topologies for the source and target texts.

The node-to-node mapping between the source and target texts is used in the distinction. For each metric, viz. k_{in}, k_{out}, l_1, etc., graphs are built with the local measurements for the source text in the abscissa and those of the target text in the ordinate. The square minimum method is then used to obtain the best straight line, whose angular coefficient is one of the parameter (metrics) employed in the distinction. The other one is the Pearson coefficient, calculated from measurements of the networks representing the original and translated texts, which quantifies the relationship between the source and target texts. Therefore, for each piece of text and each complex network metric, one obtains the Pearson coefficient and the angular coefficient.

Three machine learning [39] methods, namely decision trees [40], rules [41,42] and clustering [43] were used to discriminate the translations, for which the input were also the angular coefficients and the Pearson coefficients for each metric as described above. The class represented the type of MT system, with a lower or higher quality, and the manual translation.

The quality of translation was also evaluated using the so-called hierarchical measurements [19], which are hierarchical common degree, hierarchical cluster-coefficient, Inter-ring degree, Intra-ring degree and convergence ratio. The hierarchical

[3] http://www.revistapesquisa.fapesp.br/index.php?lg=en.
[4] http://www.apertium.org.
[5] http://www.tranexp.com:2000/Translate/result.shtml.
[6] http://www.freetranslation.com.

metrics were calculated for each node of the source network (v_f) and of the target network (v_a), with hierarchical levels varying from 1 to N, where $N = \min\{v_a, v_f\}$. Each node in the source or target network is therefore characterized by the matrices Φ_v^f and Φ_v^a, respectively:

$$\mathbf{\Phi_v^f} = \begin{pmatrix} \mu_{f1}^A & \mu_{f1}^B & \mu_{f1}^C & \cdots & \mu_{f1}^X \\ \mu_{f2}^A & \mu_{f2}^B & \mu_{f2}^C & \cdots & \mu_{f2}^X \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{fN}^A & \mu_{fN}^B & \mu_{fN}^C & \cdots & \mu_{fN}^X \end{pmatrix}$$

$$\mathbf{\Phi_v^a} = \begin{pmatrix} \mu_{a1}^A & \mu_{a1}^B & \mu_{a1}^C & \cdots & \mu_{a1}^X \\ \mu_{a2}^A & \mu_{a2}^B & \mu_{a2}^C & \cdots & \mu_{a2}^X \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{aN}^A & \mu_{aN}^B & \mu_{aN}^C & \cdots & \mu_{aN}^X \end{pmatrix}.$$

Then, $\Phi_v^x = \{u_{ij}\}$, where x represents the source network (f) or the target network (a), i stands for the hierarchical level and j is the corresponding metric for the node v under analysis. Given a node i of the source network, the binary relationship $\Psi : i \to j$ is defined, with j belonging to the set of nodes in the source network (Eq. (7)), and ϵ_h (i, j) (Eq. (8)) is the minimum error among all j. The accumulated error (Eq. (9)) corresponding to the level h is also defined, which represents the calculated error for the metrics from level 1 to N, as shown in Eq. (9).

$$\Psi_h(i) = j \mid j \in \text{ target network} \tag{7}$$

$$\epsilon_h(i, j) = \sum_{k=1}^{N} \left(\Phi_i^f(h, k) - \Phi_j^a(h, k) \right)^2 \tag{8}$$

$$\epsilon_h^{AC}(i, j) = \sum_{h=1}^{N} \sum_{k=1}^{N} \left(\Phi_i^f(h, k) - \Phi_j^a(h, k) \right)^2. \tag{9}$$

In order to avoid that metrics with larger absolute values contribute unevenly to ϵ_h^{AC}, each component in the summation is normalized [8]. Upon defining ϵ_h^{AC} and $F(i)$, one defines the accuracy rate T_h for the node i of the source text: let A be the set of nodes obtained from $F_h(i)$ and B the set of nodes obtained from the alignment with LIHLA. For a given node i, the accuracy rate for level h is given by Eq. (10). Then, the accuracy rate T_{af} for the source–target pair is given by Eq. (11).

$$T_h(i) = \frac{|A \cap B|}{|B|} \tag{10}$$

$$T_{af} = \frac{1}{N} \sum_{i=1}^{N} T_h(i). \tag{11}$$

Note that the approach reported above is conceptually founded on the principle that the corresponding nodes in the original and translated networks will tend to have maximum similar contextual connectivity, as reflected by the measurements adopted. In this context, the hierarchical measurements are particularly interesting for they may provide a more comprehensive characterization of the connectivity patterns around each node [44].

In the experiments with machine learning methods, the input attribute was the average accuracy rate and its deviation for each hierarchical level and for each text. As for the classes, again use was made of the quality of translation in the MT systems. In order to verify the effects from the learning, the classifications were compared with an analysis of the confusion matrix [45]. In addition, the accuracy rates for the machine learning methods were calculated with 10-fold-cross validation analysis [46].

3. Results and discussion

3.1. Distinguishing different types of translation

The successful application of complex networks metrics, viz. indegree (k_{in}), outdegree (k_{out}) and clustering coefficient (C) [16], for distinguishing texts has motivated us to extend the study for additional metrics. Accordingly, l_1, l_2, l_3, S and CY have been used for distinguishing translated texts for the Spanish–Portuguese pair, with l_3 being the most efficient. As

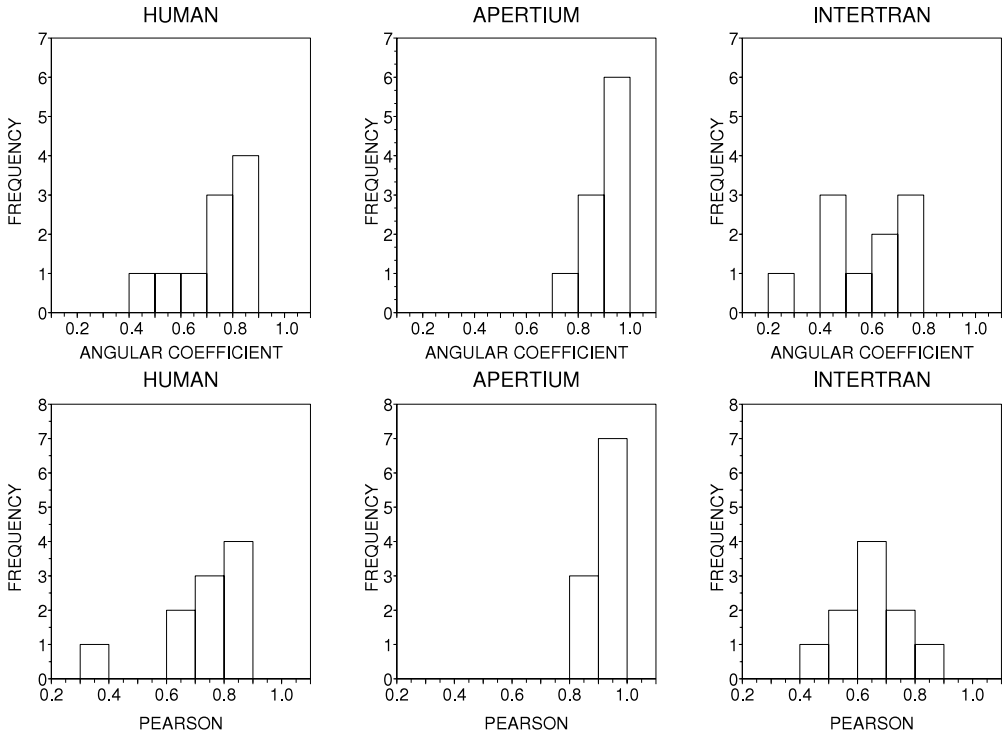

Fig. 2. Pearson coefficient and angular coefficient for the l_3 metric.

Table 2
Accuracy rates in distinguishing the type of translation.

Translation	Cluster 1 (%)	Cluster 2 (%)
Apertium	95	5
Human	80	20
Intertran	0	100

mentioned in the Methodology, for each complex network metric the source and target texts were compared by obtaining the Pearson coefficient and the angular coefficient of the fitted straight line for the metric of source vs. metric of the aligned target text. Fig. 2 shows a high tendency of node preservation in l_3 for the Apertium system, with both coefficients close to 1, in contrast to Intertran that has a low Pearson correlation and angular coefficient. The data for human translation lie in between the two machine translation systems, which may be explained by the fact that node preservation is not as high as in a quality MT system because humans tend to rephrase text while translating [16]. The values of CY were preserved in the source and target texts, regardless of the translation system. As for the S metric, there was little preservation of the node values. Therefore, these two local metrics on their own are not suitable for identifying the type of translation.

When all the data for the 5 metrics were put together in a multivariate analysis using the Projection Explorer (PEx) platform [47] (see Fig. 3), good distinction could be reached between the MT systems but not for the human translation, since the clusters generated using Ward's linkage [48] in Fig. 4 gather together Human (H) and Intertran (I) translations. In fact, if the degree (in and out) and cluster coefficient metrics are included, the distinction ability increases significantly, as illustrated in Figs. 5 and 6. In subsidiary experiments, we noted that if all the metrics were used there would be very little change in the ability of distinction (results not shown), as the two metrics chosen for Fig. 5 are among the best for the task. Then, the best metrics for distinction are k_{out}, k_{in}, l_3 and C.

With machine learning, an equally good distinction can be obtained. Table 2 shows that with the Expectation–maximization algorithm [43], the Human and the Apertium translations are mainly in cluster 1, while the translations with Intertran were entirely in cluster 2. Therefore, the use of these 8 metrics allowed one to distinguish the lower-quality from the high quality translation. Furthermore, an analysis with the J48 algorithm, using the 8 metrics, made it possible to identify the system of translation with 95% accuracy, as illustrated in the confusion matrix in Table 3.

For the English–Portuguese pair, the same approach was followed with English texts being obtained manually (human) or with two MT systems, viz. Free Translation and Intertran. The results summarized in Table 4 represent the confusion matrix obtained with the J48 implementation of the decision tree algorithm C4.5. Though less accurate than for the

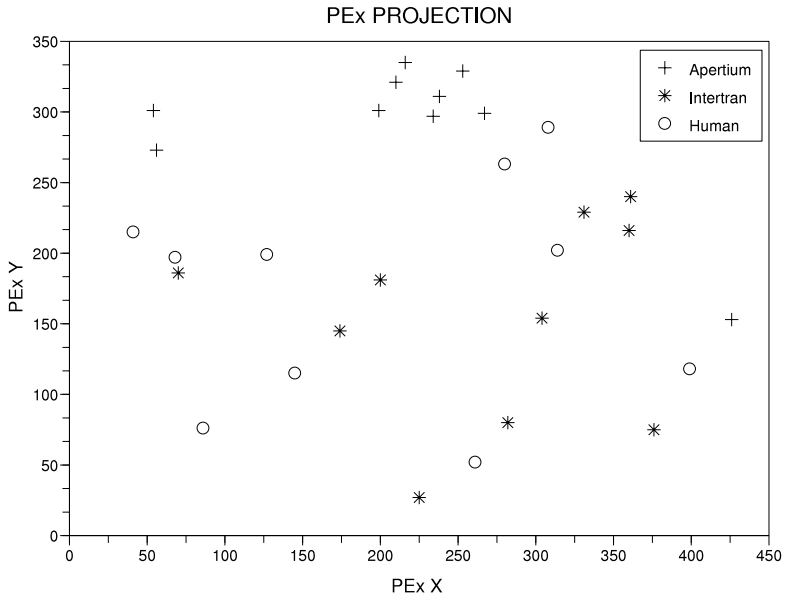

Fig. 3. Multivariate analysis using PEx for distinguishing texts translated from Portuguese into Spanish, in three ways: MT systems Apertium and Intertran, and human (manual) translation.

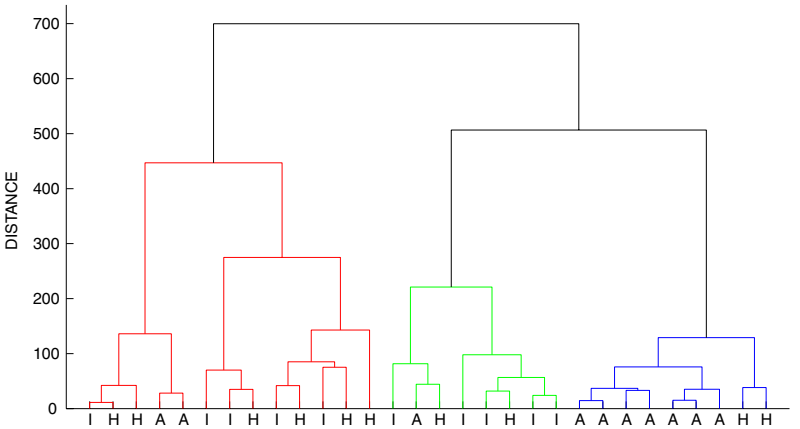

Fig. 4. Hierarchical cluster analysis using Ward's linkage for distinguishing texts translated from Portuguese into Spanish, in three ways: MT systems Apertium (A) and Intertran (I), and human (H) translation.

Table 3
Confusion matrix for the classification of the translated texts.

Apertium (%)	Human (%)	Intertran (%)	Classified as (←)
100	0	0	Apertium
5	90	5	Human
0	5	95	Intertran

Spanish–Portuguese pair, with an overall accuracy of 83%, it is still possible to distinguish the two MT systems as there was only an error of 5% in the classification as Intertran or Free Translation.

In conclusion, machine learning methods appear promising to be combined with complex network concepts in distinguishing different types of translation, and consequently in assessing the quality of translation systems. They may used in conjunction with the well-known metrics such as BLEU.

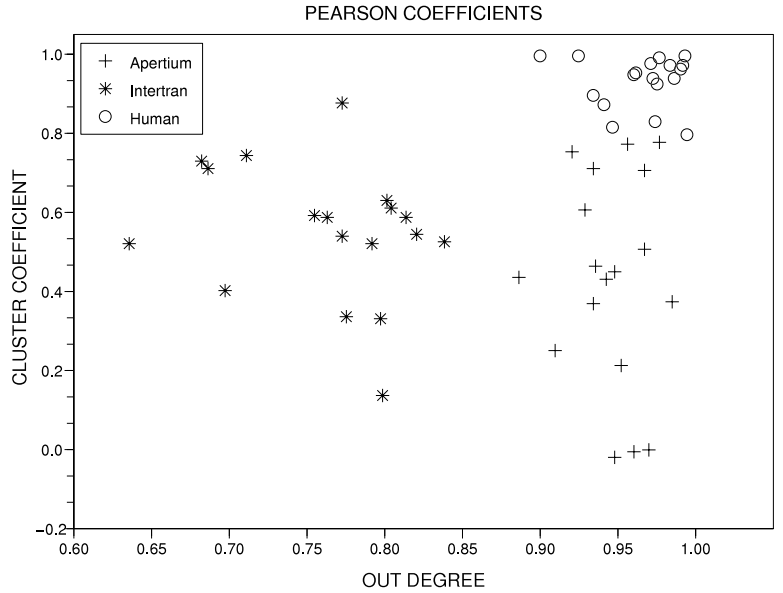

Fig. 5. Distinction obtained with the two best descriptors (k_{out} and C), among the 8 analyzed, for texts translated into Spanish with 2 MT systems and manual translation.

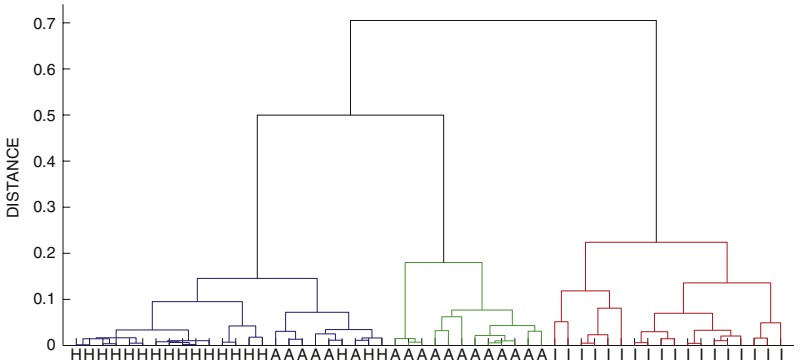

Fig. 6. Hierarchical cluster analysis using Ward's linkage for distinguishing texts translated from Portuguese into Spanish, in three ways: MT systems Apertium (A) and Intertran (I), and human (H) translation. In this case, only the two best descriptors (k_{out} and C) were employed among the 8 analyzed.

Table 4
Confusion matrix for the classification of texts translated into English.

Intertran (%)	Free Translation (%)	Human (%)	Classified as (←)
90	5	5	Intertran
5	90	5	Free Translation
15	15	70	Human

3.2. Using hierarchical metrics to evaluate MT

The motivation for using the accuracy rate with hierarchical metrics stemmed from the results reported in Ref. [44], where a source network was extracted from modeling the protein *S. Cerevisiae* [49] and the target network was obtained from a random arrangement of the nodes. The rationale behind this expectation is based on the ability of better representing a network by considering not only the features of local nodes but also along several layers of neighbors. In other words, if two related networks are to be compared, a higher accuracy in finding the corresponding nodes in the target network is expected if account is taken of the metrics for the nodes with increasing hierarchical levels. In the experiments performed for the Portuguese–Spanish pair, 50 pieces of text with ca. 500 nodes were analyzed, with Portuguese source texts and translations made by humans and through the MT systems Apertium and Intertran. As in the experiments discussed in the last section, the source and target networks were obtained through text aligning with LIHLA [38].

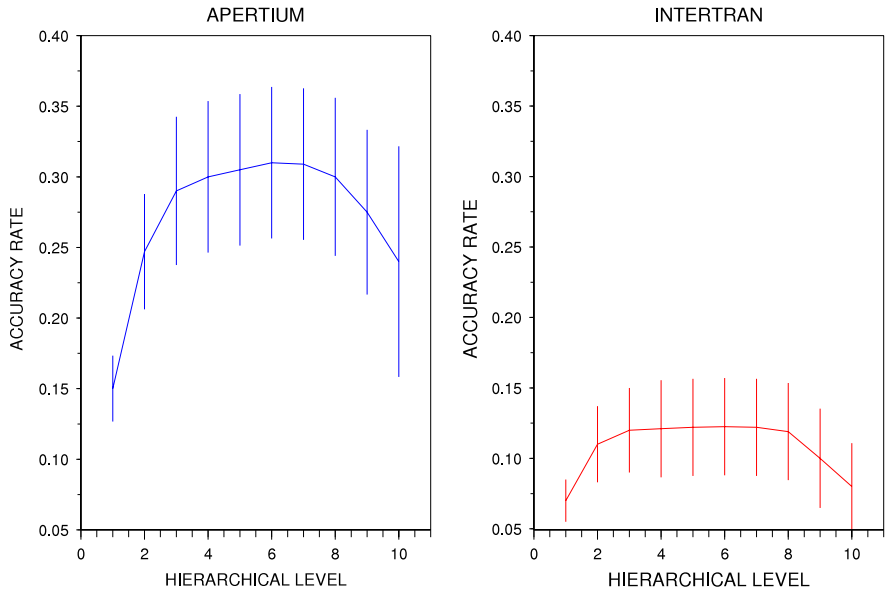

Fig. 7. Accuracy in assigning the correct node in the target text (Spanish) which corresponds to the node under analysis in the source text (Portuguese), as a function of the hierarchical levels considered for each metric (hierarchical common degree, hierarchical cluster-coefficient, Inter-ring degree, Intra-ring degree and convergence ratio). For each level the standard deviation is shown in the accuracy values. A higher accuracy is clear for Apertium in comparison to Intertran.

Table 5
Accuracy in determining the type of translation for the Spanish–Portuguese pair (Apertium or Intertran) and English–Portuguese pair (Google or Intertran) obtained with machine learning using the node-to-node mapping with hierarchical levels.

Algorithm	Spanish–Portuguese (%)	English–Portuguese (%)
Rules	99	87
Decision Trees	98	84

Table 6
Confusion matrices for the results obtained with the English–Portuguese pairs, with the Rules method.

Google (%)	Intertran (%)	Classified as (←)
90	10	Google
16	84	Intertran

Fig. 7 shows the accuracy in assigning the correct node in the target text which corresponds to a given node in the source text. The values are quite high considering that a random assignment would correspond to 1/500 (i.e. if one of the ca. 500 nodes were to be picked randomly). Such mapping was particularly good for the high-quality MT Apertium [50]. Moreover, analogously to the work for *S Cerevisiae*, the accuracy increased for the first few hierarchical levels, before it leveled off. When many hierarchical levels were considered, there was probably noise of uncorrelated information, and the accuracy dropped, as seen in Fig. 7.

An increased accuracy with increasing hierarchical levels in identifying the corresponding nodes in the target text was also observed for the Portuguese–English pair, in which the English target texts were obtained with the MT systems Google and Intertran. Data were taken with 50 texts with ca. 500 nodes each. The accuracy now is not as high as for the Spanish–Portuguese texts, again reflecting the larger structural differences between English and Portuguese [16], which hinders the ability of node preservation in the target network. Moreover, distinction between the MT systems is not easy, as it may be inferred by a visual inspection in Fig. 8.

The ability to distinguish between different types of translation was also assessed with machine learning methods employing the data with the hierarchical levels. Table 5 shows that very high accuracy is obtained for the Spanish–Portuguese pair, which was only to be expected based on the results above. For the English–Portuguese pair, the accuracy was also high, but not as much as for Spanish–Portuguese. The confusion matrices are then shown in Tables 6 and 7 to illustrate where errors occur.

Therefore, the use of only one feature (the accuracy for the node-to-node mapping) is already sufficient for distinguishing translations of variable quality. Even though the data were obtained with 10 hierarchical levels, the results from applying machine learning methods indicated that the first 3 or 4 levels are already sufficient to yield the accuracy achieved.

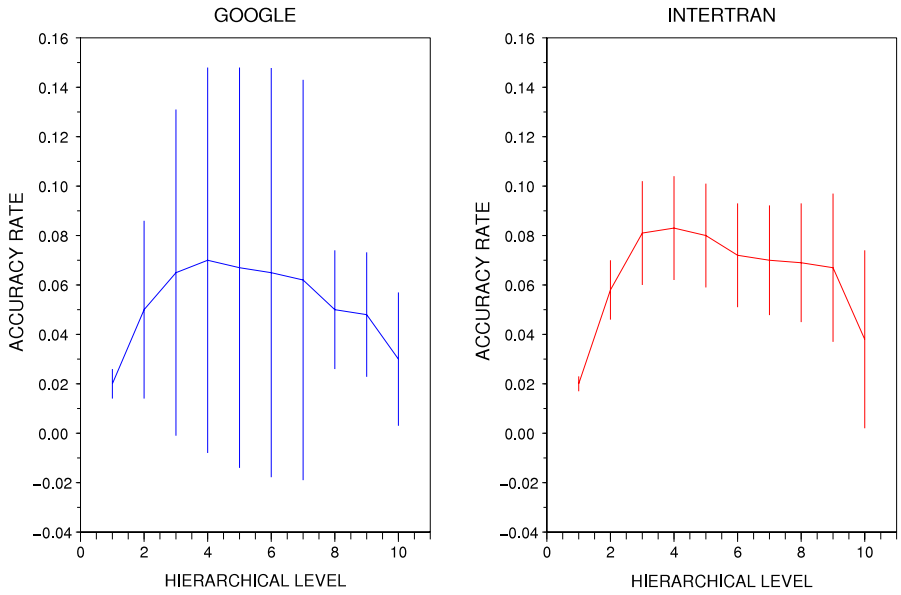

Fig. 8. Accuracy in assigning the correct node in the target text (English) which corresponds to the node under analysis in the source text (Portuguese), as a function of the hierarchical levels considered for each metric. For each level the standard deviation is shown in the accuracy values.

Table 7
Confusion matrices for the results obtained with the English–Portuguese pairs, with the Decision Tree method.

Google (%)	Intertran (%)	Classified as (←)
88	12	Google
20	80	Intertran

4. Conclusions and further work

We have exploited several metrics from complex networks to evaluate the quality of translations, where best distinctions were obtained with the out-degree (k_{out}), in-degree (k_{in}), minimum path (l_3) and cluster coefficient. The Cyclic Coefficient and the Search Information metrics, on the other hand, were not useful for assessing the quality of MT systems, because node preservation occurred regardless of the translation system for the former and there was little node preservation for the latter. The ability to distinguish between distinct types of translation was further enhanced by using machine learning methods. With the 8 metrics, the accuracy rate in distinguishing manual translations and those from 2 MT systems was very high for the Portuguese–Spanish pair, and a little lower for the Portuguese–English pair. Mention should be made of the work by Patel and Radev [51], in which manual and machine translations could be distinguished using lexical similarity. Upon employing traditional similarity measures, such as Levenshtein edit distance [22] and BLEU [24] they showed that machine translations can be distinguished from manual translations, since the former tend to create well defined clusters. In this sense, the results depicted in this paper can be understood as a complementation of Patel and Radev findings, since we showed that when topological (CN measures) and lexical (lexical alignment) information are combined, not only manual and machine translations can be distinguished from each other, but also machine translations with distinct quality.

Two important conclusions could be drawn from the use of hierarchical levels in analyzing the complex networks representing the texts. The first is that 3–4 higher levels are sufficient to enhance the ability of mapping the network of the target text into that of the source text. Second, using these metrics provided new data that were useful input for the machine learning methods.

With regard to further work to be performed, perhaps the most important implication from the results presented in this paper is associated with a general platform to assess MT systems. This platform consists basically of the combination of machine learning methods and various metrics of complex networks, including hierarchical levels. Indeed, the good results for assessing the quality of translation served as a proof-of-principle, which can now be extended to a much larger number of MT systems and languages, in addition to evaluation of human translations.

Acknowledgements

The authors are thankful to CNPq and FAPESP (2010/00927-9) for the financial support.

Appendix. Acronyms and symbols

The main acronyms and symbols employed in this paper are detailed below:

Acronym	Meaning
MT	Machine translation
CN	Complex networks
NLP	Natural language processing
METEOR	Metric for evaluation of translation with explicit ordering
BLEU	Bilingual evaluation understudy
NIST	National institute of standards and technology
MXPost	Maximum entropy part of speech tagger
LIHLA	Language-independent heuristics lexical aligner
MXPost	Maximum entropy part of speech tagger
k_{out}	Out degree
k_{in}	In degree
C	Clustering coefficient
l_1	Shortest path 1
l_2	Shortest path 2
l_3	Shortest path 3
CY	Cyclic coefficient
N	Number of vertices in the network
S	Search information
$p(i, b)$	Probability of a shortest path passing through nodes i and b to be followed randomly

References

[1] L.F. Costa, et al., Analyzing and modeling real-world phenomena with complex networks: a survey of applications, 2007. arXiv:0711.3199v1.
[2] M. Golumbic, Algorithmic graph theory and perfect graphs, Annals of Discrete Mathematics 57 (2004).
[3] R. Balakrishnan, K. Ranganathan, A Textbook of Graph Theory, 2000.
[4] C. Kittel, Elementary Statistical Physics, 1988.
[5] G.H. Wannier, Statistical Physics, 1987.
[6] A.L. Barabasi, E. Bonabeau, Scale-free networks, Scientific American (2003).
[7] M. Newman, A.L. Barabasi, D.J. Watts, The Structure and Dynamics of Networks, in: Princeton Studies in Complexity, 2006.
[8] 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.
[9] I. Ferrer, R. Cancho, R.V. Solé, The small world of human language, in: Proceedings of Biological Sciences, The Royal Society, vol. 268, 2001, pp. 2261–2265.
[10] M. Steyvers, J.B. Tenenbaum, The large-scale structure of semantic networks: statistical analyses and a model for semantic growth, Cognitive Sciences 29 (2005) 41–78.
[11] S.M.G. Caldeira, T.C. Petit Lobão, R.F.S. Andrade, A. Neme, J.G.V. Miranda, The network of concepts in written texts, The European Physical Journal B - Condensed Matter and Complex Systems 49 (2006) 523–529.
[12] O. Kinouchi, A.S. Martinez, G.F. Lima, G.M. Lourenço, S. Risau-Gusman, Deterministic walks in random networks: an application to thesaurus graphs, Physica A 315 (2002) 665–676.
[13] 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.
[14] L. Antiqueira, T.A.S. Pardo, M.G.V. Nunes, O.N. Oliveira Jr., Some issues on complex networks for author characterization, Revista Iberoamericana 11 (2007) 51–58.
[15] L. Antiqueira, O.N. Oliveira Jr., L.F. Costa, M.G.V. Nunes, A complex network approach to text summarization, Information Sciences 179 (2009) 584–599.
[16] 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.
[17] P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer, The mathematics of statistical machine translation: parameter estimation, Computational Linguistics 19 (2) (1990) 263–311.
[18] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, P. Roossin, A statistical approach to machine translation, Computational Linguistics 16 (2) (1990) 79–85.
[19] L.F. Costa, F.N. Silva, Hierarchical characterization of complex networks, Journal of Statistical Physics 125 (2006) 845–876.
[20] J.S. White, J.B. Doyon, S.W. Talbott, Task tolerance of MT output in integrated text processes, in: Proceedings of the ANLP/NAACL 2000: Embedded Machine Translation Systems Workshop, 2000, pp. 9–16.
[21] J.S. White, T.A. O'Connell, Evaluation of machine translation, in: Proceedings of the Human Language Technology Workshop, 1993, pp. 206–210.
[22] V.I. Levenshtein, Binary codes capable of correcting insertions and reversals, Soviet Physics-Doklady 10 (1966) 707–710.
[23] O.N. Oliveira Jr., A.R. Marchi, M.S. Martins, R.T. Martins, A critical analysis of the performance of English–Portuguese–English MT systems, in: Proceedings of V Encontro para o processamento computacional da Língua Portuguesa Escrita e Falada, 2000, pp. 85–92.
[24] 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.
[25] S. Banerjee, A. Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics, 2005.

[26] R. Soricut, E. Brill, A unified framework for automatic evaluation using *N*-gram co-occurrence statistics, in: Proceedings of the Association for Computational Linguistics Conference, 2004.
[27] NIST report, automatic evaluation of machine translation quality using *N*-gram co-occurrence statistics, 2002. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.
[28] Y. Zhang, S. Vogel, A. Waibel, Interpreting BLEU/NIST scores: how much improvement do we need to have a better system?, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC, 2004, pp. 2051–2054.
[29] A. Lavie, M.J. Denkowski, The meteor metric for automatic evaluation of machine translation, Machine Translation 23 (2009) 2–3.
[30] M. Snover, N. Madnani, B. Dorr, R. Schwartz, Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate, Machine Translation 23 (2010) 2–3.
[31] D. Liu, D. Gildea, Syntactic features for evaluation of machine translation, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 25–32.
[32] K. Owczarzak, J. van Genabith, A. Way, Labelled dependencies in machine translation evaluation, in: Proceedings of the Second Workshop on Statistical Machine Translation, 2007, pp. 104–111.
[33] R.M. Kaplan, J. Bresnan, Lexical functional grammar: a formal system for grammatical representation, in: The Mental Representation of Grammatical Relations, 1982, pp. 173–281.
[34] S. Pado, M. Galley, D. Jurafsky, C.D. Manning, Robust machine translation evaluation with entailment features, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, pp. 297–305.
[35] Y. He, J. Du, A. Way, J. van Genabith, The DCU dependency-based metric in WMT-MetricsMATR, in: Proceedings of the ACL HLT 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics, MATR, 2010.
[36] A. Ratnaparki, A maximum entropy part-of-speech tagger, in: The Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania, 1997.
[37] M.G.V. Nunes, F.M.C. Vieira, C. Zavaglia, C.R.C. Sossolote, J. Hernandez, O processo de construção de um léxico para o português do Brasil: lições aprendidas e perspectivas, in: II Encontro para o Processamento Computacional de Português Escrito e Falado, 1996, pp. 61–70.
[38] H.M. Caseli, M.G.V. Nunes, M.L. Forcada, LIHLA: a lexical aligner based on language-independent heuristics, in: Proceedings of the V Encontro Nacional de Inteligência Artificial, ENIA 2005, 2005, pp. 641–650.
[39] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kauffmann, 2005.
[40] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[41] W.W. Cohen, Fast effective rule induction, in: Proceedings of the 12th International Conference on Machine Learning, 1995.
[42] J. Furnkranz, G. Widmer, Incremental reduced error pruning, in: International Conference on Machine Learning, 1994.
[43] T. Hastie, R. Tibshirani, J. Friedman, 8.5 The EM Algorithm. The Elements of Statistical Learning, Springer, 2001, pp. 236–243.
[44] L.F. Costa, L.E.C. Rocha, A generalized approach to complex networks, The European Physical Journal B 50 (2006) 237–242.
[45] L.F. Costa, R.M. César Jr., Shape Analysis and Classification: Theory and Practice, CRC Press, 2009.
[46] 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, 1995, vol. 12, pp. 1137–1143.
[47] G.P. Telles, R. Minghim, F.V. Paulovich, Normalized compression distance for visual analysis of document collections, Computers & Graphics 31 (2007) 327.
[48] J.H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (301) (1963) 236–244.
[49] E. Sprinzak, H. Margalit, Correlated sequence-signatures as markers of protein–protein interaction, Journal of Molecular Biology 311 (2001) 681.
[50] C. Armentano-Oller, R.C. Carrasco, A.M. Corbí-Bellot, M.L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J.A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, M.A. Scalco, Open-source Portuguese–Spanish machine translation, in: Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006, in: Lecture Notes in Computer Science, vol. 3960, 2006, pp. 50–59.
[51] A. Patel, D. Radev, Lexical similarity can distinguish between automatic and manual translations, in: Proceeding of the Fifth International Conference on Language Resources and Evaluation, 2006.