Using complex networks in natural language processing tasks

Diego R. Amancio[1,2,3], Maria das Graças V. Nunes [2,3], Osvaldo N. Oliveira Jr[1,3], Luciano da F. Costa[1]

[1]São Carlos Institute of Physics, [2]University of São Paulo, Institute of Mathematics and Computer Science, University of São Paulo, Interinstitutional Center for Computational Linguistics (NILC)

diegoraphael@gmail.com, gracan@icmc.usp.br, chu@ifsc.usp.br, ldfcosta@gmail.com

Concepts from complex networks and other methods in statistical physics have been used in a variety of language-related applications, including in natural language processing (NLP). In this lecture, an overview will be provided of NLP tasks in which text is represented as a network with concepts being taken as nodes and edges established based on co-occurrence [1]. The topology and dynamics of the network are investigated with several metrics, including degree, strength, minimum paths and inbetweenness, whose values are taken as features for classification purposes. Machine learning methods are used in classification for various NLP tasks, such as authorship recognition, summarization, evaluation of machine translation, study of consistency in the use of words and categorization of books according to literary movements. In addition to these applications for written text, audio signals can be treated with the statistical and computational methods. The simplest way to do this is to take the whole digitalized data and extract features that may be used for classifying pieces of audio signal, such as distinguishing professional newscasters from ordinary men or women in reading the news. The signal may also be represented as a complex network where the nodes comprise narrow ranges of frequency, with edges being created with co-occurrence. Much in the same way as in written text, the topology and dynamics of the network can be used to classify speech or music passages, as in the identification of problems in speech therapy.

Examples of contributions from our research group over the last few years, such as cases in which texts represented as complex networks were used for classification purposes, will be provided. Indeed, extractive summarization strategies could be built upon reducing the networks while trying to preserve the gist of the original texts [2], with resulting methods being competitive with state-of-the-art summarization techniques. The relative importance of semantic and syntactic features in assessing similarity in texts has been exploited in three natural language processing tasks, viz. identification and evaluation of quality in machine translation systems and authorship recognition [3]. Author recognition can also be based on the way words are used in a text, which was quantified by measuring the extent of preservation of the node neighborhood in the network [4]. The words could be ranked according to a log-normal distribution when consistency of use was considered, rather than obeying Zipf´s law, which is the usual case for the frequency of word use. The literary movements to which a set of books published from 1590 to 1922 belong to could be predicted by classifying their corresponding networks, where multivariate techniques were used to generate six clusters of books [5]. It was possible to identify a trend over time toward increased syntactic complexity with larger average shortest path lengths. Finally, it is important to emphasize that the network representation and the classification methods used are entirely generic, being therefore applicable to many other types of natural language processing tasks as well as for analyzing features of languages and cultures.

Acknowledgments

The authors acknowledge the financial support from FAPESP, CNPq and Pró-Reitoria de Pesquisa da USP (Brazil).

References

[1] Costa, L.D., Oliveira Jr., O.N., Travieso, G., Rodrigues, F.A., Boas, P.R.V., Antiqueira, L., Viana, M.P., Rocha; L.E.C. 2011. Analyzing and modeling real-world phenomena with complex networks: a survey of applications, *Adv. Phys*. 60, 329-412.

[2] Amancio, D.R., Nunes, M.G.V., Oliveira Jr., O.N., Costa, L.D., 2012 Extractive summarization using complex networks and syntactic dependency, *Physica* A 391, 1855–1864.

[3] Amancio, D.R., Oliveira Jr., O.N., Costa, L.D., 2012 Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts, *Physica* A 391, 4406–4419.

[4] Amancio, D.R., Oliveira Jr., O.N., Costa, L.D., 2012 Using complex networks to quantify consistency in the use of words, *J. Stat. Mech*. P01004.

[5] Amancio, D.R., Oliveira Jr., O.N., Costa, L.D., 2012 Identification of literary movements using complex networks to represent texts, *New J. Phys*., 14, P 043029.

***