Toward a Complex Networks Approach on Text Type Classification (Abstract)

Domagoj Margan, Ana Meštrović, Marina Ivašić-Kos, Sanda Martinčić-Ipšić
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{dmargan, amestrovic, marinai, smarti}@uniri.hr

Keywords. complex networks, linguistic co-occurrence networks, text type classification, document classification

The growing amount of text electronically available has placed text type classification among the most exciting issues in the field of exploratory data mining. This talk presents an preliminary approach to text type classification by features of linguistic co-occurrence networks. Text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links. The aim of our work-in-progress presented in this talk is to investigate the idea of replacing the standard natural language processing feature sets with linguistic network measures for the purpose of text type classification. This talk tackles the problem of binary classification of two different text types. Our dataset is consisted of 150 equal-sized Croatian texts divided in two classes: 75 literature texts and 75 blog texts. Literature texts represent segments from 7 different books written in or translated to Croatian language, while blog texts are collected from two very popular Croatian blogs. The trait which prompted us to do the classification of this particular text types is the linguistic distinction between book and blog. We constructed 150 different co-occurrence networks (one for each text in the dataset), all weighted and directed. Words are nodes linked if they are co-occurring as neighbors to each other in a sentence. The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a text. For each network we computed a set of 10 measures (number of components, average degree, average path length, clustering coefficient, transitivity, degree assortativity, density, reciprocity, average in-selectivity, average out-selectivity), which are used as feature set for classification. All features are rescaled to $[0-1]$ in order to make them independent of each other. We preformed a series of classification experiments using various types of classification algorithms and methods (support vector machine, classification trees, Naive Bayes, k-nearest neighbor, LDA, QDA). The performance of each classifier was evaluated with corresponding methods, such as misclassification error measures, confusion matrices and receiver operating characteristic curves. All classification experiments show very good classification accuracy, while the average in- and out- selectivity measures act as the most useful features in predicting the correct text type and reducing the misclassification rate. Precision and recall measures and ROC curves indicate that the node selectivity measures are the only measures from the feature set that can capture the structural differences between two classes of networks.