[03/11/2015 - 18:00 - Room Vitória]

Properties of semantic networks derived from automatic generated word embeddings, Marco Idiart, *Instituto de Física, UFRGS*, Aline Villavicencio, Rodrigo Wilkens, *Instituto de Informática, UFRGS* ■Recently Mikolov et al.[1] proposed a very successful neural network algorithm to automatically extract a vector representation of the words of a language (word embeddings) from a training set composed of large collections of representative texts. From the word embeddings with simple vector algebra it is possible to derive a series of language properties such as word similarity, plural and gender operations, compositionality of multiword expressions, etc. For instance, the similarity of two words in the schema is simply the cosine between their vector representations. Therefore the similarity matrix of the whole lexicon can be viewed as a semantic network, where words correspond to nodes and the distances between them are the weighted edges. In this work we characterized statistically the semantic network derived from Mikolov's vector embeddings for different spatial dimensions and compare it with studies performed in semantic networks from manually constructed thesaurus. In addition we show that the eigenvalues distribution of the word embeddings covariance matrix follows a power law. This indicates that the dimension of the word embeddings space cannot be well defined via representation reduction methods such as principal component analysis. We examine separately the networks of nouns and verbs, given their different characteristics, and to what extent these are reflected in the resulting network properties.

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.