Chapter 2

Comparing Distributional Semantics Models for Identifying Groups of Semantically Related Words

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó
University of Barcelona

Published at
Procesamiento del Lenguaje Natural, 2016
vol. 57, pp.: 109-116

Abstract Distributional Semantic Models (DSM) are growing in popularity in Computational Linguistics. DSM use corpora of language use to automatically induce formal representations of word meaning. This article focuses on one of the applications of DSM: identifying groups of semantically related words. We compare two models for obtaining formal representations: a well known approach (CLUTO) and a more recently introduced one (Word2Vec). We compare the two models with respect to the PoS coherence and the semantic relatedness of the words within the obtained groups. We also proposed a way to improve the results obtained by Word2Vec through corpus preprocessing. The results show that: a) CLUTO outperforms Word2Vec in both criteria for corpora of medium size; b) The preprocessing largely improves the results for Word2Vec with respect to both criteria.

Keywords DSM, Word2Vec, CLUTO, semantic grouping

2.1 Introduction

In recent years, the availability of large corpora and the constantly increasing computational power of the modern computers have led to a growing interest in linguistic approaches that are automated and data-driven [Arppe et al., 2010]. Distributional semantic models (DSM) [Turney and Pantel, 2010, Baroni and Lenci, 2010] and the vector representations (VR) they generate fit very well within this framework: the process of extracting vector representations is mostly automated and the content of the representations is data-driven.

The format of the vector is suitable for carrying out different mathematical manipulations. Vectors can be compared directly through an objective mathematical function. They can also be used as a dataset for various Machine Learning algorithms. VR are more often used on tasks related to lexical similarity and relational similarity [Turney and Pantel, 2010]. In such tasks, the emphasis is on pairwise comparisons between vectors.

This article focuses on another use of the Vector Representations: the grouping of vectors, based on their similarity in the Distributional space. This grouping can be used, among other things, as a methodology for identifying groups of semantically related words. High quality groupings can serve for many purposes: they are a semantic resource on their own, but can also be applied for syntactic disambiguation or pattern identification and generation [Martí et al., 2019], for example.

We compare two different methodologies for obtaining groupings of semantically related words in English - a well known approach (CLUTO) and a more recently introduced one (Word2Vec). The two methodologies are evaluated in terms of the quality of the obtained groups. We consider two criteria: 1) the semantic relatedness between the words in the group; and 2) the PoS coherence of the group. We evaluate the role of the corpus size with both methodologies and in the case of Word2Vec, the role of the linguistic preprocessing (lemmatization and PoS tagging).

The rest of this paper is organized as follows: Section 2.2 presents the general framework and related work. Section 2.3 describes the available data and tools. Section 2.4 presents the experiments and the results obtained. Finally Section 2.5 gives conclusions and identifies directions for future work.

2.2 Related Work

Distributional Semantics Models (DSM) are based on the Distributional Hypothesis, which states that the meaning of a word can be represented in terms of the contexts in which it appears [Harris, 1954, Firth, 1957]. As opposed to semantic approaches based on primitives [Boleda and Erk, 2015], approaches based on distributional semantics can obtain formal representations of word meaning from actual linguistic productions. Additionally, this data-driven process for semantic representation can mostly be automated.

Within the framework of DSM, one of the most common ways to formalize the word meaning is a vector in a multi-dimensional distributional space [Lenci, 2008]. For this purpose, a matrix with size **m** by **n** is extracted from the corpus, representing the distribution of **m** words over **n** contexts. The format of a vector allows for direct quantitative comparison between words using the apparatus of linear algebra. At the same time it is a format preferred by many Machine Learning algorithms.

The choice of the matrix is central for the implementation of a particular DSM. Turney and Pantel [2010] suggest a classification of the DSM based on the matrix used. They analyze three different matrices: term-document, word-context, and pair-pattern. The different matrices represent different types of relations in the corpus and the choice of the matrix depends on the goals of the particular research.

Baroni and Lenci [2010] present a different, sophisticated approach for extracting information from the corpus. They organize the information as a third order tensor, with the dimensions representing <'word', 'link', 'word' >. This third order tensor can then be used to generate different matrices, without the need of going back to the original corpus.

In this paper we focus on one of the classical vector representations - the one based on word-context relation. It measures what Turney and Pantel [2010] call "attributional similarity". In particular, we are interested in the possibility to group vectors together, based on their relations in the distributional space.

Erk [2012] offers a survey of possible applications of different DSM. She lists clustering as an approach that can be used with vectors, for word sense disambiguation. Moisl [2015] presents a theoretical analysis on the usage of clustering in computational linguistics and identifies key aspects of the mathematical and linguistic argumentation behind it.

Here we analyze and compare two approaches that induce vector representations from a corpus and apply algorithms to identify sets of semantically related words. We are interested in the quality of the obtained groups, as we believe that they can be a useful, empirical, linguistic resource.

Martí et al. [2019] present a methodology named DISCOveR for identifying candidates to be constructions from a corpus. As part of this methodology they use CLUTO [Karypis, 2002] for clustering words based on their vector representations. Their approach uses a word-context matrix where the context is defined by combining a syntactic dependency with a lemma. After all the vectors are extracted, CLUTO is used in order to obtain clusters of semantically related words. Later on these clusters are used to generate a list of the candidates to be construc-