Abstract

For the last decade, Distributional Semantic Models (DSMs) have been an active area of research to address the problem of understanding the semantics of words in natural language. The central assumption which functions as the core of DSMs is that the linguistic context surrounding a given word provides important information about its meaning, and hence word co-occurrence statistics can provide a natural basis for semantic representations.

The main idea behind this work is to explore ways to incorporate syntactic information within the distributional semantic models. This approach is known as structured DSMs, as opposed to unstructured DSMs which use a set of unordered words as linguistic context without any syntactic information.

There have been some works focusing on structured DSMs which use syntactic categories such as part-of-speech (POS) tags (*i.e.* nouns, verbs, adjectives, etc.) and dependency relations between words (*i.e.* subject, direct object, modifier, etc.), to enrich the basis elements of context vector used to represent the meaning of a word in DSMs. Although it has been shown that dependency paths are a good approximation to semantic relations between words, the lack of information about syntactic category of each word in the context might become the shortcoming to fully exploit syntactic information within DSMs.

To this end, we exploit Combinatory Categorial Grammar (CCG) categories that provide a transparent relation between syntactic category and semantic type of the linguistic expression, to enrich the linguistic context used to represent the word meaning in DSMs. CCG categories, which are regarded as supertags, are supposed to be better at defining semantics than the traditional less informative POS tags since they carry more rich syntactic information.

We use the standard framework to build DSMs, which includes defining linguistic context, building the co-occurrence matrix, and applying various weighting schemes to the produced co-occurrence matrix. However, in order to build CCG-based DSM, constructing the co-occurrence matrix requires the use of CCG parsed corpus to extract the CCG categories of each word, hence full CCG parsing is carried out beforehand.

The constructed CCG-based DSM is then evaluated on one of widely known semantics tasks, which is word categorization, by comparing it with other types of DSM. From the experiments we could conclude that CCG-based DSM is generally better than POS-based DSM, and most certainly outperforms unstructured DSM that do not provide any syntactic information in the linguistic context. It is also shown that CCG-based DSM performs better than dependency-based DSM in the case of verb categorization task, demonstrating the importance of syntactic categories in defining the meaning of verbs.

We also investigate the impact of employing different context window on the performance of CCG-based DSM. Finally, we explore the effect of including the function words, which are grammatical words such as determiner, preposition, pronoun, etc. belonged to the group of closed-class words, in the co-occurrence matrix since existing works covering this topic usually use only content words including nouns, verbs, adjectives, and most adverbs.

Key words: Distributional Semantic Models, Combinatory Categorial Grammar, Word Categorization