The Word-Space Model

Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces

Academic dissertation
for the Degree of Doctor of Philosophy in Computational Linguistics at Stockholm University
to be publicly defended on Friday 2 June 2006 at 13:00
in hörsal 4, hus B, Universitetsvägen 10
by
Magnus Sahlgren

Department of Linguistics
Stockholm University
106 91 Stockholm

Abstract

Magnus Sahlgren: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces

The word-space model is a computational model of word meaning that utilizes the distributional patterns of words collected over large text data to represent semantic similarity between words in terms of spatial proximity. The model has been used for over a decade, and has demonstrated its mettle in numerous experiments and applications. It is now on the verge of moving from research environments to practical deployment in commercial systems. Although extensively used and intensively investigated, our theoretical understanding of the word-space model remains unclear. The question this dissertation attempts to answer is: what kind of semantic information does the word-space model acquire and represent?

The answer is derived through an identification and discussion of the three main theoretical cornerstones of the word-space model: the geometric metaphor of meaning, the distributional methodology, and the structuralist meaning theory. It is argued that the word-space model acquires and represents two different types of relations between words – syntagmatic and paradigmatic relations – depending on how the distributional patterns of words are used to accumulate word spaces. The difference between syntagmatic and paradigmatic word spaces is empirically demonstrated in a number of experiments, including comparisons with thesaurus entries, association norms, a synonym test, a list of antonym pairs, and a record of part-of-speech assignments.