A Comparison of Co-occurrence and Similarity Measures as Simulations of Context

Stefan Bordag

Natural Language Processing Department, University of Leipzig
sbordag@informatik.uni-leipzig.de

Abstract. Observations of word co-occurrences and similarity computations are often used as a straightforward way to represent the global contexts of words and achieve a simulation of semantic word similarity for applications such as word or document clustering and collocation extraction. Despite the simplicity of the underlying model, it is necessary to select a proper significance, a similarity measure and a similarity computation algorithm. However, it is often unclear how the measures are related and additionally often dimensionality reduction is applied to enable the efficient computation of the word similarity. This work presents a linear time complexity approximative algorithm for computing word similarity without any dimensionality reduction. It then introduces a large-scale evaluation based on two languages and two knowledge sources and discusses the underlying reasons for the relative performance of each measure.

1 Introduction

One way to simulate associative and semantic relations between words is to view each word as a distinct entity. That entity may occur in a linear stream of sentences or other easily observable linguistic units. It is then possible to measure the statistical correlation between the common co-occurrence of such entities (i.e. words) within these units [1, 2]. If additional knowledge such as word classes or morphological relatedness is available, this model allows to construct a variety of applications that depend on knowledge about word relatedness, but do not necessarily need this knowledge to be precise. For example, it is sufficient to know the most significant co-occurring word pairs in a corpus to enable the creation of a helpful tool for extraction of collocations, idioms or multi-word-expressions [3–5]. Similarly, knowledge about contextual similarity modeled as co-occurrence vector comparisons helps to build thesaurus construction tools such as the Sketch engine [6] or to design specific semi-automatic algorithms that create approximations of a thesaurus [7–10].

Assuming the simple vector-space model where each word defines a new dimension, the question arises how exactly significant co-occurrence or word similarity is to be modeled. Several variations of the same underlying vector space model were proposed. One is to apply Latent Semantic Indexing (LSI) to the matrix containing the raw co-occurrence counts of words [11]. However, it is