Discovering Word Senses from Text Using Random Indexing

Niladri Chatterjee [1] and Shiwali Mohan [2]

[1] Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, India 110016
niladri@maths.iitd.ac.in
[2] Yahoo! Research and Development India, Bangalore, India 560 071
shiwali@yahoo-inc.com

Abstract. Random Indexing is a novel technique for dimensionality reduction while creating Word Space model from a given text. This paper explores the possible application of Random Indexing in discovering word senses from the text. The words appearing in the text are plotted onto a multi-dimensional Word Space using Random Indexing. The geometric distance between words is used as an indicative of their semantic similarity. Soft Clustering by Committee algorithm (CBC) has been used to constellate similar words. The present work shows that the Word Space model can be used effectively to determine the similarity index required for clustering. The approach does not require parsers, lexicons or any other resources which are traditionally used in sense disambiguation of words. The proposed approach has been applied to TASA corpus and encouraging results have been obtained.

1 Introduction

Automatic disambiguation of word senses has been an interesting challenge since the very beginning of computational linguistics in 1950s [1]. Various clustering techniques, such as bisecting K-means [2], Buckshot [3], UNICON [4], Chameleon [5], are being used to discover different senses of words. These techniques constellate words that have been used in similar contexts in the text. For example, when the word 'plant' is used in the living sense, it is clustered with words like 'tree', 'shrub', 'grass' etc. But when it is used in the non-living sense, it is clustered with 'factory', 'refinery' etc. Similarity between words is generally defined with the help of existing lexicons, such as WordNet [6], or parsers (e.g. Minipar [7]).

Word Space model [8] has long been in use for semantic indexing of text. The key idea of Word Space model is to assign vectors to the words in high dimensional vector spaces, whose relative directions are assumed to indicate semantic similarity. The Word Space model has several disadvantages: *sparseness* of the data and *high dimensionality* of the semantic space when dealing with real world applications and large size data sets. Random Indexing [9] is an approach developed to deal with the problem of high dimensionality in Word Space model.