Computational Linguistics and Chinese Language Processing
Vol. 4, No. 2, August 1999, pp. 1-20

A Model for Word Sense Disambiguation

Li Juanzi*, Huang Changning*

Abstract

Word sense disambiguation is one of the most difficult problems in natural language processing. This paper puts forward a model for mapping a structural semantic space from a thesaurus into a multi-dimensional, real-valued vector space and gives a word sense disambiguation method based on this mapping. The model, which uses an unsupervised learning method to acquire the disambiguation knowledge, not only saves extensive manual work, but also realizes the sense tagging of a large number of content words. Firstly, a Chinese thesaurus *Cilin* and a very large-scale corpus are used to construct the structure of the semantic space. Then, a dynamic disambiguation model is developed to disambiguate an ambiguous word according to the vectors of monosemous words in each of its possible categories. In order to resolve the problem of data sparseness, a method is proposed to make the model more robust. Testing results show that the model has relatively good performance and can also be used for other languages.

Key Words: natural language processing, word sense disambiguation, unsupervised learning, vector space, language modeling

1. Introduction

Word sense disambiguation, that is, identifying the correct sense of a word from all its senses as defined in a dictionary or a thesaurus, has long been one of the most difficult problems in natural language processing. In the 1990s, the research on this topic has entered a new phase with the availability of machine-readable dictionaries and very large corpora. Such research mainly falls into two classes: dictionary-based and corpus-based methods. The dictionary-based disambiguation methods, such as the models put forward by Lesk [1986] and Wilks [1990], do not perform well when the context of a word has little overlap with the text of its dictionary definition. Corpus-based methods, such as the

* The State Key Laboratory for Intelligent Technology and Systems, The Department of Computer Science and Technology, Tshinghua University, Beijing 100084.
e-mail: ljz@s1000e.cs.tsinghua.edu.cn