Feature Selection for Chinese Character Sense Discrimination

Zheng-Yu Niu and Dong-Hong Ji

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613, Republic of Singapore
{zniu, dhji}@i2r.a-star.edu.sg

Abstract. Word sense discrimination is to group occurrences of a word into clusters based on unsupervised classification method, where each cluster consists of occurrences having same meaning. Feature extraction method has been used to reduce the dimension of context vector in English word sense discrimination task. But if original dimension has a real meaning to users and relevant features exist in original dimensions, feature selection is a better choice for finding relevant features. In this paper we apply two unsupervised feature selection schemes to Chinese character sense discrimination, which are entropy based feature filter and Minimum Description Length based feature wrapper. Using precision evaluation and known ground-truth classification result, our preliminary experiment results demonstrate that feature selection method performs better than feature extraction method on Chinese character sense discrimination task.

1 Introduction

Word sense discrimination is to group occurrences of a word into clusters based on unsupervised learning method, where the occurrences in same cluster have same meaning [15]. In contrast with word sense disambiguation, word sense discrimination determines only which occurrences have the same meaning, but not what the meaning is. Compared with supervised word sense disambiguation, the burden to provide lexicon, hand tagged corpus or thesaurus can be avoided in word sense discrimination.

In [15] the author presents context group discrimination algorithm to solve word sense discrimination problem. Singular Value Decomposition (SVD) technique is used to reduce the dimension of context vectors. Principle Component Analysis (or Karhunen Loeve transformation, Singular Value Decomposition) is a well-established unsupervised feature extraction technique [3]. But if original dimension has a real meaning to users and relevant features exist in original dimensions, feature extraction technique is not appropriate for finding relevant features. The reason is that (1) the combination of original dimensions is difficult to interpret, and (2) the irrelevant original dimension are not clearly removed because they are required to determine the new dimension, and this will deteriorate the performance of clustering algorithm. In Chinese language, an ambiguous