Chapter 24
A Comparative Study of an Unsupervised Word Sense Disambiguation Approach

Wei Xiong
New Jersey Institute of Technology, USA

Min Song
New Jersey Institute of Technology, USA

Lori Watrous deVersterre
New Jersey Institute of Technology, USA

ABSTRACT

Word sense disambiguation is the problem of selecting a sense for a word from a set of predefined possibilities. This is a significant problem in the biomedical domain where a single word may be used to describe a gene, protein, or abbreviation. In this paper, we evaluate SENSATIONAL, a novel unsupervised WSD technique, in comparison with two popular learning algorithms: support vector machines (SVM) and K-means. Based on the accuracy measure, our results show that SENSATIONAL outperforms SVM and K-means by 2% and 17%, respectively. In addition, we develop a polysemy-based search engine and an experimental visualization application that utilizes SENSATIONAL's clustering technique.

INTRODUCTION

Many English words have multiple meanings or senses. For example, the word *foot* in the sentence *The house is at the foot of the mountains* refers to the bottom part of the mountains, whereas in the sentence *One of his shoes felt too tight for his foot* it refers to the terminal part of the vertebrate leg upon which an individual stands. As we can see, the correct senses of the word *foot* can be inferred from its contextual words *house* and *mountains* in the first sentence and *shoes*, *felt*, and *tight* in the second sentence. The contextual words help in disambiguating and understanding the word that has multiple meanings. The ambiguity of an individual word or phrase that can be used in different contexts to express two or more different meanings is called polysemy.

DOI: 10.4018/978-1-60960-741-8.ch024

In general terms, Word Sense Disambiguation (WSD) involves the problem of determining the correct meaning an ambiguous word bears in a given context. This process relies to a great extent on the surrounding context of the word. It has been regarded as a crucial problem in many natural language processing (NLP) applications. For example, an information retrieval system could perform better if the ambiguities among queries were reduced. Other applied NLP applications that have benefited from WSD include information extraction (Stokoe, Oakes and Tait 2003), question answering (Pasca and Harabagiu 2001), and machine translation (Vickrey et al. 2005).

In the biomedical domain, WSD is a central problem. Many names of proteins and genes, abbreviations, and general biomedical terms have multiple meanings. For instance, *glucose* could be used to mean a biologically active substance which is a carbohydrate, or glucose measurement which is a laboratory procedure. These ambiguous words make it difficult for Natural Language Processing (NLP) applications and, in some cases, humans to correctly interpret the appropriate meaning.

SENSATIONAL is a novel unsupervised WSD technique proposed recently by (Duan, Song, and Yates 2009). Their original study presents impressive accuracy results. Our research contributes by benchmarking SENSEATIONAL against two well-received algorithms: Support Vector Machines (SVM) and K-means. Furthermore, we discuss SENSATIONAL's data preprocessing benefits related to reduced manual effort. These characteristics make SENSATIONAL's novel approach to WSD a very attractive application to a number of real-world problems in the area of search and data visualization that our research piloted for further exploration.

BACKGROUND

There are three types of WSD techniques (Ide and Veronis 1998): supervised learning, unsupervised learning and knowledge-based WSD. Supervised techniques need manually-labeled examples for each ambiguous term in the data set to predict the correct sense of the same word in a new context. This is referred to as training material which allows their corpus to build up a classification scheme based on the set of feature-encoded inputs and their appropriate sense label or category. The result of this training is a classifier that can be applied to future instances of the ambiguous word.

As a classification problem of machine learning, WSD has several characteristics that distinguish it from other traditional classification problems in NLP. Due to the difficulty of creating manually-labeled examples for the ambiguous terms, there is usually a small amount of training data available for WSD task. For example, the data set of ambiguous biomedical terms available from the National Library of Medicine (NLM) contains only 100 examples of each term being used in context. Also, the number of the senses of an ambiguous word for a WSD problem can be quite large. Take word *cold* as an example, there are more than 10 meanings of *cold* according to the Merriam-Webster Online Dictionary. Compared with WSD problem, for other classification problems in NLP, such as POS tagging (part-of-speech tagging which makes up the words in a text as corresponding to a particular part of speech), a word usually only has one or two POS's in a single language. Furthermore, the features, which are extracted from the context of a target word and used for classification, usually include lexical features. Without proper feature selection criteria, the amount of possible lexical features used by a machine learning algorithm can be very large, while the frequencies with which they occur in the data sets can be very low.

Knowledge-based WSD systems are similar to supervised learning because they use established, external knowledge, such as databases and dictionaries to disambiguate words. However, both of these approaches need extensive manual effort to create either training data or external resources. This can be time-consuming and expensive.