Chapter VIII
Word Sense Disambiguation in Biomedical Applications:
A Machine Learning Approach

Torsten Schiemann
Humboldt-Universität zu Berlin, Germany

Ulf Leser
Humboldt-Universität zu Berlin, Germany

Jörg Hakenberg
Arizona State University, USA

ABSTRACT

Ambiguity is a common phenomenon in text, especially in the biomedical domain. For instance, it is frequently the case that a gene, a protein encoded by the gene, and a disease associated with the protein share the same name. Resolving this problem, that is, assigning to an ambiguous word in a given context its correct meaning is called word sense disambiguation (WSD). It is a pre-requisite for associating entities in text to external identifiers and thus to put the results from text mining into a larger knowledge framework. In this chapter, we introduce the WSD problem and sketch general approaches for solving it. The authors then describe in detail the results of a study in WSD using classification. For each sense of an ambiguous term, they collected a large number of exemplary texts automatically and used them to train an SVM-based classifier. This method reaches a median success rate of 97%. The authors also provide an analysis of potential sources and methods to obtain training examples, which proved to be the most difficult part of this study.

INTRODUCTION

Ambiguity, i.e., words with multiple possible meanings, is a common phenomenon in natural languages (Manning & Schütze 1999). Which of the different meanings of a word is actually meant in a concrete text depends on the context the word appears in and cannot be judged based only on the appearance of the word itself. For instance, the terms 'sin' and 'soul' both are common English words – but they are also names of proteins. If a person only sees one of these two words on a piece of paper, he cannot decide which of the two meanings (or senses) the paper tries to convey. However, given a phrase such as "Salvation from sins", humans immediately recognize the correct sense of the ambiguous word.

From a linguistics point of view, the term ambiguity in itself has different senses. The most common form is homonymity, that is, words that have multiple, possibly unrelated meanings. 'Sin' and 'soul' both are homonyms. However, there are also more complex forms of ambiguity, such as polysemy, which describes cases where a word has different yet closely related senses. Examples in the life sciences are identical names for a gene, the protein it encodes, and the mRNA in which it is transcribed.

Word sense disambiguation (WSD) is the problem of assigning to an ambiguous term in

Figure 1. Screenshot of the Ali Baba main window after searching PubMed for the term 'soul'. Colored boxes represent biological entities. One can see that various Meanings of 'soul' are intermixed in the display – mentioning of the immortal soul by the Greek poet Homer, results from a large international study called 'Heart and Soul', and facts describing the protein 'soul'.