International Journal of Medical Informatics (2005) **74**, 573—585

ELSEVIER

International Journal of
Medical
Informatics

www.intl.elsevierhealth.com/journals/ijmi

Effects of information and machine learning algorithms on word sense disambiguation with small datasets

Gondy Leroy[a,*], **Thomas C. Rindflesch**[b]

[a] *School of Information Science, Claremont Graduate University, 130 E. Ninth Street, Claremont, CA 91711, USA*
[b] *National Library of Medicine, Bethesda, MD, USA*

Received 16 November 2004; received in revised form 17 February 2005; accepted 17 March 2005

KEYWORDS
Word sense disambiguation;
Machine learning;
Naïve Bayes;
Decision tree;
Neural network;
UMLS

Summary Current approaches to word sense disambiguation use (and often combine) various machine learning techniques. Most refer to characteristics of the ambiguity and its surrounding words and are based on thousands of examples. Unfortunately, developing large training sets is burdensome, and in response to this challenge, we investigate the use of symbolic knowledge for small datasets. A naïve Bayes classifier was trained for 15 words with 100 examples for each. Unified Medical Language System (UMLS) semantic types assigned to concepts found in the sentence and relationships between these semantic types form the knowledge base. The most frequent sense of a word served as the baseline. The effect of increasingly accurate symbolic knowledge was evaluated in nine experimental conditions. Performance was measured by accuracy based on 10-fold cross-validation. The best condition used only the semantic types of the words in the sentence. Accuracy was then on average 10% higher than the baseline; however, it varied from 8% deterioration to 29% improvement. To investigate this large variance, we performed several follow-up evaluations, testing additional algorithms (decision tree and neural network), and gold standards (per expert), but the results did not significantly differ. However, we noted a trend that the best disambiguation was found for words that were the least troublesome to the human evaluators. We conclude that neither algorithm nor individual human behavior cause these large differences, but that the structure of the UMLS Metathesaurus (used to represent senses of ambiguous words) contributes to inaccuracies in the gold standard, leading to varied performance of word sense disambiguation techniques.

© 2005 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Although many words we use in conversation and writing are ambiguous, we usually do not experi-

[*] Corresponding author. Tel.: +1 909 607 3270.
E-mail address: gondy.leroy@cgu.edu (G. Leroy).

1386-5056/$ — see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2005.03.013

ence problems with interpreting these words in context. People seem to take the context of a conversation effortlessly into account and assign the correct meanings to individual words. Such disambiguation, however, is not easily accomplished with automated methods. Since this is a problem for machine translation, information retrieval, thematic analysis, spelling correction, or any type of speech and text processing, researchers have devoted considerable effort to word sense disambiguation (WSD).

WSD techniques choose the correct sense for a word from a predefined set of available senses. Most existing techniques use the surrounding words and specific features of these to learn the correct sense of the ambiguous word. They are usually supervised machine learning algorithms based on large annotated datasets where the correct sense is indicated for each instance. Ide and Véronis [1] provide an overview of WSD from the early years (1950s) to the late 1990s.

We evaluated the effect of different types of symbolic information for terms in medical text by mapping sentences to the Unified Medical Language System (UMLS). We used small datasets to evaluate how much this knowledge base can contribute when few examples are available. For our first set of tests, we used a naïve Bayes classifier. We continued our study with the best condition by comparing with a neural network (feedforward/backpropagation) and a decision tree algorithm. Accuracy was similar for all three, but the variance between different words was very large. We then tried to discover why the variance was so high. We believe that it may be the different meanings available in the UMLS (a compilation of vocabularies not intended as a WSD resource) which led to the confusion in compiling the gold standard used for learning. Using individual expert's gold standards or specific gold standard characteristics could not explain the variance.

2. Word sense disambiguation

There exist many techniques that are used for word sense disambiguation. Which one is chosen depends on the final goal, the available information per word, and the number of available examples. In some cases, it is sufficient to distinguish between different meanings of words, without having to label the words. For example, a label may be unnecessary when clustering documents together that have similar topics. Schütze [2] labels this task as ''word sense discrimination''. He distinguishes this from ''sense labeling'' where each sense receives the correct label. This distinction often — but not always — coincides with unsupervised (discrimination) versus supervised (labeling) machine learning techniques.

2.1. Approaches to word sense disambiguation

2.1.1. Unsupervised learning techniques

Unsupervised learning algorithms learn patterns solely from input parameters without trying to match to pre-specified categories. In the case of word sense disambiguation, they learn to group words based on the information in the feature sets. But there is no label specified in advance for the group nor is the number of possible groups specified. Assigning a specific meaning can still be achieved by finding the common theme in the established clusters and mapping these to established meanings for the word in a dictionary or other knowledge source. This mapping can be done by a human or automatically based on similarity metrics.

Clustering techniques are especially useful for this type of disambiguation. For example, Pedersen and Bruce [3] tested three unsupervised learning algorithms: Wards and McQuitty's clustering and the EM algorithm. They mapped these clusters to dictionary senses so that there was maximal agreement.

2.1.2. Supervised learning techniques

Supervised learning is used more often for WSD. These techniques rely on outcome feedback provided to an algorithm so that it can take corrective action during its learning or training phase. The possible outcomes are known in advance and algorithms need to learn to combine a particular input with such an output. In the case of word sense disambiguation, the input usually consists of features of the ambiguous word and surrounding text. The output is the correct sense for the word. During the learning phase, supervised techniques learn to associate these feature sets with one particular sense of a limited list of provided senses. This happens by providing the techniques with feedback on its decision for every example. The supervised learning techniques rely on a training set comprised of example ambiguous words and their correct sense. Decision trees, such as ID3 or C4.5, artificial neural networks (ANN), such as the feedforward/backpropagation ANN, and probabilistic-based methods, such as naïve Bayes, are commonly used.

Mooney [4] tested seven such supervised learning methods with the word *line*. His work demonstrates the importance of a large dataset. The input

Download English Version:

https://daneshyari.com/en/article/517497

Download Persian Version:

https://daneshyari.com/article/517497

Daneshyari.com