Raw Corpus Word Sense Disambiguation*

Ted Pedersen
Department of Computer Science & Engineering
Southern Methodist University
Dallas, TX 75275–0122
pedersen@seas.smu.edu

A wide range of approaches have been applied to word sense disambiguation. However, most require manually crafted knowledge such as annotated text, machine readable dictionaries or thesari, semantic networks, or aligned bilingual corpora. The reliance on these knowledge sources limits portability since they generally exist only for selected domains and languages. This poster presents a corpus–based approach where multiple usages of an ambiguous word are divided into a specified number of sense groups based strictly on features that are automatically obtained from the immediately surrounding raw text.

We are given N sentences, each of which contains a usage of a particular ambiguous word. Each sentence is converted into a feature vector $(F_1, F_2, \ldots, F_n, S)$ where $(F_1, \ldots, F_n)$ represent the observed contextual properties of the sentence and S represents the unobserved sense of the ambiguous word.

A probabilistic model is built from this data. First, a parametric form that describes the interactions among the observed contextual features and the unknown sense is specified. We use the form commonly known as Naive Bayes due to its favorable performance in previous studies of supervised disambiguation (e.g., Gale et. al., 1992, Mooney, 1996, Ng 1997).

The Naive Bayes model, when applied to disambiguation, implies that all contextual features are conditionally independent given the sense of the ambiguous word:

$$p(F_1, F_2, \ldots, F_n, S) = p(S) \prod_{i=1}^{n} p(F_i|S)$$

To complete the model the values of the parameters $p(F_i|S)$ must be estimated. However, since S is not observed in the text this can not be done directly. Instead, we use the Expectation Maximization (EM) algorithm and Gibbs Sampling, two popular methods for estimating parameters when data is missing.

Both algorithms iterate until convergence is detected. The EM algorithm imputes values for the missing data S and maximizes the parameter estimates $p(F_i|S)$ given those imputed values. While it is guaranteed to converge, the EM algorithm is susceptible to finding local maxima. We treat Gibbs Sampling as a stochastic version of the EM algorithm. It approximates the complete distribution of the parameters by repeatedly sampling from them rather than simply maximizing a point estimate. In so doing Gibbs Sampling finds the global maximum but also proves more difficult to monitor for convergence.

The evaluation of these methods is based on the degree to which the discovered sense groups agree with those created by a human judge. Both methods are used to disambiguate thirteen different words using three feature sets. Gibbs Sampling shows small but consistent improvments in accuracy over the EM algorithm. The comparable performance of these two methods is somewhat surprising given the tendency of the EM algorithm to converge at local maxima. However, in these experiments the EM algorithm often converges quite quickly, usually within 20 iterations, to a global maximum. These results suggest that some combination of the EM algorithm and Gibbs Sampling might be beneficial. A feature set using local context features, i.e., collocations that occur within ±2 positions of the ambiguous word, generally results in higher disambiguation accuracy than a feature set based on co–occurrences from a wider window of context. A more detailed discussion of these experimental results is found in (Pedersen & Bruce, this volume).

There are three areas of future work. First, we will use the convergence points of the EM algorithm as initial values for Gibbs Sampling in the hopes of speeding the convergence of Gibbs Sampling. Second, we will experiment with parametric forms based on expert knowledge rather than simply relying on Naive Bayes. Finally, we will identify additional local context features that increase disambiguation accuracy without significantly increasing the dimensionality of the problem.

Acknowledgments

This research was supported by the Office of Naval Research under grant number N00014-95-1-0776.