Word Sense Disambiguation with Lexical Attraction Models

Introduction

The word sense disambiguation (WSD) problem is a long recognized one in the field of Natural Language Processing. The notion of sense disambiguation first appeared within the Machine Translation research going on in 50's when it was observed that the meaning of the source word should be translated accordingly in the target language by choosing the translation among a reduced set of possible translations for the source word. Other uses for the WSD processing level of the language is to be found in the survey of the state of the art WSD given in Ide and Véronis (1998).

Most WSD algorithms fall in two distinct classes: supervised and unsupervised. Supervised WSD presupposes that an annotated corpus is available from which it can learn how to assign senses. Unsupervised WSD does not beneficiate from training data and thus is first concerned with the grouping of the occurrences of a given word in classes in which all occurrences have the same sense. What is important is that both methods use some kind of context representation which is one of the central issues of the WSD problem. Usually, contexts are represented as collections of attributes called '*features*'. Examples of features of contexts are: words surrounding the target word in a given window of $2k + 1$ (the so called '*bag of words*' model - for instance, in Yarowsky (1992), k equals 50); collocates of the target word (the '*one sense per collocation*' hypothesis of Yarowsky (1993)) and so on.

In what follows, we will briefly present an unsupervised WSD method that makes use of a pseudo-syntactic, dependency-like representation of the sentence called linkage. The linkage will serve as the context representation for the WSD algorithm.

Lexical Attraction Models

Lexical Attraction Models (LAMs) were first introduced by Yuret (1998). They describe methods by which an undirected, planar and connected graph with no cycles is generated for a given input sentence where the sentence's words are the vertices of the graph. This graph is in fact a tree and Yuret's general thesis is that lexical attraction relation is the likelihood of a syntactic/morphologic relation as defined by Meľčuk (1988). The tree obtained with a LAM is not a proper dependency-syntactic structure because the edges of the tree are not oriented and they have no labels such as *object*, *modifier* and so forth. We call this tree a linkage.

We have shown elsewhere (Nnnn Yyyy) that LAMs can be constrained with language dependent rules using part of speech (POS) tags to generate better linkages. The WSD algorithm presented here assumes that the input sentence is assigned a linkage by a specialized computer program we call linker. It uses the linkage in both its phases: training and sense labeling. It is clear that the WSD algorithm is very sensitive to the quality of the linkage because if the links are incorrect, arbitrary correspondences between the words' senses are estimated.

The Algorithm

The algorithm makes use of the assumption that a structured context is crucial to the task of disambiguating the senses of the words in a sentence. Not only the words surrounding the target word are important in determining the sense of the latter but also their relations to the target and their relations to themselves. This assumption comes from Meľčuk's Meaning Text Model (1988), where the semantic representation of the sentence derives from a deterministic projection of the syntactic one. The linkage is not a syntactic dependency structure but is the closest thing we can achieve with very little or no resources at all.

The algorithm uses the WordNet2.0 sense inventory. WordNet is a semantic dictionary and it is described in Fellbaum (1998). We make use of the 'hypernymy' and 'hyponymy' relations for nouns and verbs and of 'similar to' relation for adjectives in order to reduce the parameters for the estimation process presented below.

To be used, the algorithm needs training. The training process requires a corpus that is tokenized at sentence and word level, POS tagged, with every sentence parsed with the linker above. The main idea in training is to extract an association table between the senses of two, syntactically related, words. We use the EM algorithm from Brown et al. (1993) for the IBM1 translation model that is suited to our purposes. For *n* iterations, the training proceeds as follows:

```
1.   for every sentence s from the corpus
2.     for every link l from s
3.       L = extract and generalize senses of the left word of the link
4.       R = extract and generalize senses of the right word of the link
5.       add to the t-table IBM1(L,R)
6.     end for
7.   end for
```

The generalization at lines 3 and 4 aims at reducing the parameter space of the *t-table* by choosing the most general sense of a noun or verb such that the sense is not the common ancestor of two different senses of the same noun or verb. For adjectives, the 'similar to' relation is used and the center of the cluster replaces any of the cluster elements. The global *t-table* is created at line 5 by calling the estimation function on the left and right sets of senses (the reader may consult Brown et al. (1993) for the creation of *t-tables*).

The annotation phase of the WSD algorithm needs the *t-table* produced above and for a given linked sentence, it assigns the best combination of senses to the content words of it by maximizing the association score between senses of words participating in a link. To achieve this, it uses a Viterbi style algorithm to maximize the assignment.

Conclusions and Further Work

We have briefly described an unsupervised WSD algorithm that uses linkages as context representation.

First of all, we need to test the algorithm on the SemCor data. We will experiment with other sense inventories (such as SUMO concepts which are mapped onto WordNet2.0 synsets) and evaluate the accuracy. Furthermore, it is interesting to apply this algorithm on sentences parsed with a fully-fledged dependency parser and to include the relation orientation and labeling in the training/annotating mechanisms.

References

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). *The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, vol. 19, pp. 263-311, June 1993.

Fellbaum, C. (1998). *WordNet, an electronic lexical database.* MIT Press, 1998.

Ide, N., and Véronis J. (1998). *Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art.* Computational Linguistics, 24(1):1-40, 1998.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice.* SUNY Press, New York.

Yarowsky, D. (1992). *Word-sense disambiguation using statistical models of Roget's categories trained on large corpora*. Proc. of the 33th Annual Meeting of the Association for Computational Linguistics, pages 189-196.

Yarowsky, D. (1993). *One sense per collocation*. In the Proceedings of ARPA Human Language Technology Workshop, 1993.

Yuret, D. (1998). *Discovery of linguistic relations using lexical attraction.* Ph.D. diss., Department of Computer Science and Electrical Engineering, MIT.

(Nnnn Yyyy)