A Semi-Complete Disambiguation Algorithm for Open Text

Rada Mihalcea
Department of Computer Science and Engineering
Southern Methodist University
Dallas, Texas, 75275-0122
rada@seas.smu.edu

Word Sense Disambiguation (WSD) is one of the most difficult areas of Natural Language Processing (NLP); the semantic comprehension of a text, and the possibility to expand a text with semantically related information, drastically depends on the availability of a highly accurate WSD algorithm. Solutions considered so far by researchers for the WSD problem, are making use of machine readable dictionaries (Leacock, Chodorow and Miller 1998), or the information gathered from raw or semantically disambiguated corpora (Yarowsky 1995). These methods are designed either to work with a few pre-selected words, in which case a high accuracy is obtained, or they are general methods which disambiguate, with lower precision, all the words in a text.

With the present work, we are trying to achieve a compromise between these two different directions. There are fields in NLP, like Information Retrieval and others, which could benefit from a method which performs a semi-complete disambiguation (i.e. it disambiguates only a certain percentage of the words in a text), but which is highly accurate.

The method described in this abstract uses information gathered from a MRD, namely WordNet (Fellbaum 1998), and from SemCor - a corpus in which all words are sense tagged based on WordNet definitions. It differs from previous approaches in that it uses an iterative approach: the algorithm has as input the set of nouns and verbs extracted from the input text, and incrementally builds a set of disambiguated words. This approach allows us to identify, with high precision, the semantic senses for a subset of the input words. About 55% of the nouns and verbs are disambiguated with a precision of 91%.

Below, we are going to briefly describe the various procedures used to identify the correct sense of a word. These procedures are iteratively invoked within the main algorithm.

PROCEDURE 1. This procedure uses a Named Entity (NE) component to recognize and identify person names, locations, company names and others. We add TPER (person), TORG(group) and TLOC(location) tags. The words or word collocations marked with such tags are replaced by their role (i.e. person, group, location) and marked as having sense #1.

PROCEDURE 2. Identify the words having only one sense in WordNet (*monosemous* words). Mark them with sense #1.

PROCEDURE 3. For a given word W_i, at position i in the text, form two pairs, with the word before W_i (pair W_{i-1}-W_i) and the word after W_i (pair W_i-W_{i+1}). Then, we extract all the occurrences of these pairs found within the semantic tagged corpus formed with the 179 texts from SemCor. If, in all the occurrences, the word W_i has only one sense #k, and the number of occurrences of this sense is larger than 3, then mark the word W_i as having sense #k.

PROCEDURE 4. Find words which are semantically connected to the already disambiguated words, and for which the connection distance is 0. The distance is computed based on the WordNet hierarchy; two words are semantically connected at a distance of 0 if they belong to the same synset.

PROCEDURE 5. Find words which are semantically connected with each other, and for which the connection distance is 0.

PROCEDURE 6. Find words which are semantically connected to the already disambiguated words, and for which the connection distance is maximum 1; two words are semantically connected at a maximum distance of 1 if they are *synonyms* or they belong to a *hypernymy/hyponymy* relation.

PROCEDURE 7. Find words which are semantically connected with each other, and for which the connection distance is maximum 1.

The text to be disambiguated is first tokenized and part of speech tagged using Brill's tagger. We also identify the concepts based on WordNet definitions. Two sets of words are maintained, a set of ambiguous words SAW and the set of disambiguated words SDW. The procedures presented above are applied iteratively, until no more words can be disambiguated. Initially, all the words from the text are included in the SAW set and SDW is initialized with the empty set. As words are disambiguated by one of the procedures, they are removed from SAW and added to SDW. This allows us to identify a set of nouns and verbs which can be disambiguated with high precision.

We performed several tests using 6 randomly selected files from SemCor. Each of these files has been divided into sets of 15 sentences; these sets are used as input to the algorithm. The results have shown that about 55% of the nouns and verbs are disambiguated with 91% accuracy.

The method described here is a continuation of our previous work in the WSD field (Mihalcea and Moldovan 1999), and it is part of the work we are currently doing in the field of semantic indexing.

References

Fellbaum, C. *WordNet, An Electronic Lexical Database*. The MIT Press, 1998.

Leacock, C.; Chodorow, M. and Miller, G.A. Using Corpus Statistics and WordNet Relations for Sense Identification, *Computational Linguistics vol.24 no.1*, pages 147-165, 1998.

Mihalcea, R. and Moldovan D. A method for Word Sense Disambiguation of unrestricted text *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 152-158, College Park, MD, 1999.

Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics (ACL-95)*, pages 189-196, Cambridge, MA, 1995.