

Thesis Proposal

Varada Kolhatkar

November 20, 2007

Title: Optimizing the Selection of Context in All Words Sense Disambiguation

Student: Varada Kolhatkar

Advisor: Dr. Ted Pedersen

Proposal:

1. Problem Statement and Goals

The overall goal of this thesis is to advance the state of the art in performing all words sense disambiguation. This is the process of assigning every content word in a sentence a sense from a dictionary based on the surrounding context. In particular our work will start with the methods developed by [Mic05] and extend them in the following significant ways.

- (a) To determine the advantages of expanding and improving context window
- (b) Exploring how fixing the senses works once the sense is known and how local Vs global information helps in disambiguation
- (c) Knowing if part of speech tag or syntactic information is useful

2. Literature Review

In 1986, Lesk proposed a solution to word sense disambiguation based on semantic relatedness between words [Les86]. He suggests that given a specific word from a text, the sense of that word could be identified by counting the number of shared words (i.e. overlaps) in the definitions of that word and the definition of the word preceding or following that word. The sense with maximum matches (overlaps) would potentially be the intended sense. Lesk's experiments yielded accuracies of 50 – 70%. Lesk also raises important questions regarding the span of the word and settlement of results once a correct sense of a word is found. These questions have driven further research on word sense disambiguation problem.

A new measure of semantic relatedness based on extended gloss overlap was then introduced by [BP03], in 2003. The measure combines the advantages of Lesk's gloss overlap with the structure of a concept hierarchy to create an extended view of relatedness. Given two concepts, this

measure provides a quantitative measure for the relatedness between the concepts based on the overlapping words in their respective glosses, as well as the overlaps found in the glosses of concepts they are related to in a given concept hierarchy such as WordNet. The scoring mechanism is different than Lesk's in a way that, it doesn't differentiate between a single word and a phrasal overlap. A word sense disambiguation was carried out on SENSEVAL-2 lexical sample data. The disambiguation based on this measure proves to be most competent with other systems. This disambiguation method was extended to disambiguate all content words in a sentence by [Mic05], which will serve as the starting point for the work in this thesis.

The measures described above are domain independent and have been based on a hierarchical database such as WordNet. The article by [PPPC07] displays the effectiveness of adapting domain independent measures of semantic similarity and semantic relatedness to a specialized domain of biomedicine. This is based on the idea that some of the measures that have been found to be effective with WordNet can be adapted and extended to a growing number of ontologies in the biomedical domain such as SNOMED-CT and certain NLP tasks in the biomedical domain can be automated. The word sense disambiguation method in this thesis will be generic such that it can be applied to texts in specific domains like medicine without significant modification, assuming that measures of semantic similarity such as those described by [PPPC07] exist for that domain.

3. Resources

- Software Resources

Dictionary:

WordNet (<http://wordnet.princeton.edu>)

Applications:

WordNet::QueryData (<http://people.csail.mit.edu/jrennie/WordNet>)

WordNet::Similarity (<http://www.d.umn.edu/~tpederse/similarity.html>)

All these resources are freely available on the web.

- Hardware Resources

No special hardware requirement

- Evaluation Data Resources

SemCor (<http://lit.csci.unt.edu/~rada/downloads/semcor>)

SENSEVAL-2 (<http://lit.csci.unt.edu/~rada/downloads/senseval.semcor>)

SENSEVAL-3 (<http://lit.csci.unt.edu/~rada/downloads/senseval.semcor>)

4. Time Line and target dates

- (a) DECEMBER 2007 - Familiarizing with existing methods (Michellizzi's thesis work)

- (b) FEBRUARY 2008 - Replicate and understand Michelizzi's experiments and results
- (c) MARCH 2008 - Initial refinements to context window selection in place and evaluated write up, methods and results (short paper at *Annual Meeting of the Association for Computational Linguistics (ACL)*)
- (d) MARCH-APRIL 2008 - Web interface (public roll out). Demo at ACL.
- (e) MAY 2008 - Begin with goal (b), fixing the senses and local vs global information.
- (f) OCTOBER 2008 - Finish goal (b). Evaluate, write up, results and method
- (g) NOVEMBER 2008 - Start goal (c).
- (h) JANUARY 2009 - Finish goal (c). Evaluate, write up, results and method
- (i) FEBRUARY 2009 - Finalizing all experiments, getting final results, writing thesis
- (j) MAY 2009 - Colloquium/ Thesis Defending

5. Evaluation Method

We will compare our results to the manually tagged data such as SemCor, SENSEVAL-2 and SENSEVAL-3 that is used by the applications in resources section, and compare our results to those published in the literature that have used that same data. Evaluation data which is going to be used is created by the people who are independent of us. Hence it is a solid evaluation method.

6. The contribution to research that this work will make

Improving the ability to assign senses to all the words that appear in a written document will significantly advance the state of the art in Natural Language Processing, in particular methods that seek to understand documents so as to better summarize, categorize, or translate them. Overall, this research will contribute in better understanding of natural language by computers.

References

- [BP03] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, August 2003.

- [Les86] M.E. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM Press, 1986.
- [Mic05] J. Michelizzi. Semantic relatedness applied to all words sense disambiguation. Master’s thesis, University of Minnesota, Duluth, July 2005.
- [PPPC07] T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, June 2007.

Signatures:

Student

Advisor