CS446 Class Project: A Semi-supervised Approach to Word Sense Disambiguation

Haoruo Peng(hpeng7) Shyam Upadhyay(upadhya3)

October 18, 2013

Task description

In natural language, a word may be associated with possibly multiple meanings, depending on the context in which the word occurs. For instance, the word pen has the following senses according to Wordnet:

1. *pen : a writing implement with a point from which ink flows*

2. *pen : an enclosure for confining livestock*

Word sense disambiguation is the problem of determining the correct sense of a word in a given sentence. For example, determining the correct sense of the word "pen" in the following passage:

Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

In this project, we will investigate into the performance of semi-supervised learning algorithms for WSD.

Background

A general survey is provided in [5]. WSD has been described as an AI-complete problem [2]. Researchers have done much progress to WSD problem achieving sufficiently high levels of accuracy on a variety of word types and ambiguities. A rich variety of techniques have been researched, from dictionary-based methods [4] that use the knowledge encoded in lexical resources, to supervised machine learning methods [3] in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely unsupervised methods [6] that cluster occurrences of words, thereby inducing word senses.

Because of the lack of training data, many word sense disambiguation algorithms use semi-supervised learning [1], which allows both labeled and unlabeled data. The Yarowsky algorithm [6] was an early example of such an algorithm. It uses the 'One sense per collocation' and the 'One sense per discourse' properties of human languages for word sense disambiguation. From observation, words tend to exhibit only one sense in most given discourse and in a given collocation.

The bootstrapping approach starts from a small amount of seed data for each word: either manually tagged training examples or a small number of surefire decision rules (e.g., 'play' in the context of 'bass' almost always indicates the musical instrument). The seeds are used to train an initial classifier, using any supervised method. This classifier is then used on the untagged portion of the corpus to extract a larger training set, in which only the most confident classifications are included. The process repeats, each new classifier being trained on a successively larger training corpus, until the whole corpus is consumed, or until a given maximum number of iterations is reached.

Other semi-supervised techniques use large quantities of untagged corpora to provide co-occurrence information that supplements the tagged corpora. These techniques have the potential to help in the adaptation of supervised models to different domains.

Data and evaluation

For general tasks, we will use datasets from **SemEval** "http://www.senseval.org". It is an ongoing series of evaluations of computational semantic analysis systems; it evolved from the Senseval word sense evaluation series. The evaluations are intended to explore the nature of meaning in language. While meaning is intuitive to humans, transferring those intuitions to computational analysis has proved elusive.

We will also test our developed system on the medical disambiguation data "http://wsd.nlm.nih.gov". This test collection was constructed using a method that automatically extracts instances of ambiguous terms from **MEDLINE** without manual curation which also uses **MeSH** indexing of **MEDLINE** as a resource. The resulting data set contains both biomedical terms and abbreviations and is automatically created using the **UMLS Metathesaurus** and the manual **MeSH** indexing of **MEDLINE**.

Your approach

Semi-supervised approach for word sense disambiguation was first used in Yarowsky (1995) [6]. Our approach to the problem will be similar in that we will also use a small set of labelled examples for seeding our decision parameters. Once we have developed a prototype, we will make incremental improvements and try to achieve performance comparable to the state of the art.

Your to-do list

1. We have medical disambiguation data
2. We have mentioned the related work in this proposal.
3. We will use the semi-supervised approach of Yarowsky. We plan to implement it ourselves. (by week 2)
4. We will compare the performance of our system with the state of the art by comparing the F1 scores on standard datasets.
5. Week 6 - Writeup and Conclusion.

References

[1] Anh-Cuong Le, Akira Shimazu, Van-Nam Huynh, and Le-Minh Nguyen. Semi-supervised learning integrated with classifier combination for word sense disambiguation. *Computer Speech & Language*, 22(4):330–345, 2008.

[2] John C Mallery. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers. In *Master's thesis, MIT Political Science Department*. Citeseer, 1988.

[3] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[4] Rada Mihalcea. Using wikipedia for automatic word sense disambiguation. In *HLT-NAACL*, pages 196–203, 2007.

[5] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

[6] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.