A Semi-supervised Approach to Word Sense Disambiguation

CS446 Class Project

Haoruo Peng(hpeng7@illinois.edu) Shyam Upadhyay(upadhya3@illinois.edu)

December 1, 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. Word sense disambiguation can be viewed as a multi class classification problem, where each word admits several possible senses and the task is to identify the correct sense of a given word given its context. 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.

A naive baseline for word sense disambiguation will be to output the most frequent sense of the word In this project, we will investigate into the performance of semi-supervised learning algorithms for word sense disambiguation.

Background

A general survey is provided in [**?**]. WSD has been described as an AI-complete problem [**?**]. 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 [**?**] that use the knowledge encoded in lexical resources, to supervised machine learning methods [**?**] in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely unsupervised methods [**?**] that cluster occurrences of words, thereby inducing word senses.

Supervised learning approaches have been remarkably successful for performing word sense disambiguation. But the lack of sense-tagged data poses a severe bottleneck. To address this problem, researchers have resorted to semi-supervised learning algorithms. The Yarowsky algorithm [**?**] 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.

[**?**], which allows both labeled and unlabeled data. More recently, researchers have leveraged word alignment information from parallel corporas to aid in obtaining coarse grained senses. In word alignment tasks, a wor Unlike sense-tagged datasets, good quality parallel corpora are readily available.

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) [**?**]. 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.