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

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

December 1, 2013

Abstract

We investigate the task of word sense disambiguation using a semi-supervised approach. Word sense disambiguation is the task of identifying the correct sense for a word which can possibly admit multiple senses. We use the Semeval dataset to evaluate the performance of our system. We use a graph based approach to disambiguate senses; the underlying assumption is that the sense of the neighboring words can assist in inferring the sense of the target word. In this respect, the senses are learnt in a somewhat joint-manner. Using appropriate weights in the graph constructed for all possible senses, we can use the shortest path algorithm to obtain the correct senses.

1 Introduction

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 [5]:

- *pen : a writing implement with a point from which ink flows*
- *pen : an enclosure for confining livestock*

Word sense disambiguation(WSD) 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.* WSD 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 and knowledge sources. The inherent difficulty of WSD is evident from the fact that the target classes change for each word in the lexicon. In this respect, WSD involves training n different classifiers, one for each word in a lexicon of size n. WSD is a key component for natural language processing systems which involves semantic interpretation of text. Tasks such as machine translation, information retrieval, data mining, web data analysis can greatly benefit from text disambiguation tools.

2 Background

Summarize and discuss related work that you are building on: this requires you to find, read and cite a few research papers. This is also something you can get started on as soon as you have settled on a task.

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 [7] 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 [7] 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.

[1], 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. The reader is encouraged to peruse a more general survey provided in [6].

3 Task and Data

Now describe the task and data in more detail. 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.

3.1 The Task

Now, try to formalize your task as a classification/ranking/... problem. Introduce mathematical/formal notation as necessary. How do you evaluate models, or measure success?

3.2 The Data

Describe the data you use to train and evaluate your models. Describe where you got it from (include references/citations to published works, or URLs!). Describe and give examples for the features that you have access to.

4 The Models

4.1 Baseline Models

In order to know how difficult the task is and how well we are doing, we need to know how well a suitable baseline model would perform. Define a baseline model for your task. This may not necessarily be a learned model. A naive baseline for word sense disambiguation will be to output the most frequent sense of the word

4.2 Existing Models

If people have worked on this task before, summarize (and cite) some of the existing models

4.3 Proposed Model(s)

Your models and your procedure for learning them go here. Describe both in detail, even if the learning procedure is standard.

5 Experiments

5.1 Experimental Hypotheses

Summarize the hypotheses (research questions) your experiments are designed to test (address). (Note that some of these hypotheses may emerge as you keep working on a problem; you will not necessarily have come up with all the questions you wish to address before you have started building a models for the specific task.

5.2 Experimental setup

Define test/training/dev data splits, describe how you tuned performance. Describe and your evaluation metric, and define it mathematically. List the models you will evaluate. Cite any existing tools or software you use to perform your experiments; describe what you implemented yourself. Describe how you obtained the features used by each of the models.

5.3 Experimental results

Now give the actual experimental results (use figures/tables/graphs as appropriate), and discuss whether they verify or falsify your hypotheses. How important are the various features your models use (consider ablation studies). How robust are your results? (Look at learning curves, or the variance when you perform cross-validation). Can you perform an error analysis?

6 Conclusion

Summarize your findings, and discuss their implications, e.g. for future work, or for related tasks. Discuss also the shortcomings of your proposed approach. .

Bibliography

Don't forget to create your own .bib file. If you call it `mybib.bib` *and put it in the same directory as this* `.tex` *file, add* `\bibliography{mybib}` *before* `\end{document}`

Your current to-do list

This should be an updated version of your initial to-do list. Compare what you have done with what remains to be done. If you have a group project: who will do what? Set yourself deadlines. Here are a few items that might appear on your to-do list

Done

1. `[Insert ...]`
2. `[Insert ...]`

Left to do

1. [Insert ...do you have data?]
2. [Insert ...do you know related work? (have you got the references for your .bib file?)]
3. [Insert ...what algorithm will you use? do you need to implement this yourself, or will you use an off-the-shelf package?]
4. [Insert ...what experiments do you plan to run?]
5. [Insert ...and don't forget to allocate time for the writeup!]

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] George A. Miller. Wordnet: A lexical database for english. 1995.

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

[7] 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.