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

1 Introduction

This should be a brief outline of the paper – use plain English, no math. Note that you should be able to write most of this section before you actually perform any experiments. First, define and motivate your task: what are you trying to learn, and why is this an important task? Second, define what kind of a machine learning problem this requires you to solve (binary/multiclass classification, ranking, ...). What is an appropriate baseline model for this task? What kind of model are you proposing? Briefly summarize the assumptions your model makes. Finally, describe the hypotheses you wish to test. These are typically statements of the form "we expect model/features A to perform better on this task than model/features B". Outline how your experiments will evaluate these hypotheses (comparisons of different models, ablation studies, learning curves, oracle experiments...).

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.

Word sense disambiguation is important for 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.

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.

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.

3 Task and Data

Now describe the task and data in more detail.

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.

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!]

Bibliography

If you need references for the background section, 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}`