

Optimizing the Selection of Context In All Words Sense Disambiguation

Varada Kolhatkar
Department of Computer Science
University of Minnesota Duluth
kolha002@d.umn.edu

Self Introduction and Thesis Title

- Varada Kolhatkar
- Previous Degrees: Masters in Computer Science from University of Pune, India
- Work Experience: I worked with a software company, *Symantec Corporation* for a couple of years
- **Thesis Title: Optimizing the Selection of Context in All Words Sense Disambiguation**

Words often have multiple senses!

- Consider the following example

The basketball player shot the ball into the net.

- ball -> round object used in games
 - ball -> formal dance
- Fairly easy for human beings to understand the meaning according to the context
 - It will be useful in various applications if software could distinguish between different senses of a word
 - Word Sense Disambiguation (WSD) is the process of selecting the correct sense of a word in given context

All Words Disambiguation

- All words sense disambiguation is the process of disambiguating all words in the text separately.
- This is analogous to part of speech tagging. The correct sense of a word is decided according to the adjacent and related words in that sentence.

Why All Words?

- While searching documents we often give phrases instead of a single target word
- If we want to understand the overall content of the given text and give a title to it or categorize it, disambiguating all words in the text rather than a target word makes sense

Semantic Relatedness

- There are various approaches to All Word Sense Disambiguation. One of them is finding semantic relatedness between words. This is an unsupervised learning approach where no training data is used.
- Method that quantifies how similar two word senses are is called the measure of semantic relatedness

<i>ball</i>	<i>player</i>	<i>net</i>
round object	high	average
dance	low	low

SenseRelate::AllWords

- This is a Perl module which carries out WSD for all words using WordNet as the base dictionary for disambiguation
<http://search.cpan.org/dist/WordNet-SenseRelate-AllWords/>
- The concept of disambiguating all words is analogous to the concept of part of speech taggers
- In part of speech taggers, a tag is assigned to a word according to its definition and the context in which it is used in the sentence.
- Similarly in All Words sense disambiguation, a sense is assigned to a word according to its definition and the definitions of the words surrounding it.

SenseRelate::AllWords Example

- *The basketball player shot the ball into the net.*

SenseRelate::AllWords output

The basketball#n#1 player#n#1 shot#n#2 the ball#n#1 into the net#n#3

- ball#n#1 means first noun sense of *ball* in WordNet
S: (n) ball (round object that is hit or thrown or kicked in games) "the ball travelled 90 mph on his serve"; "the mayor threw out the first ball"; "the ball rolled into the corner pocket"

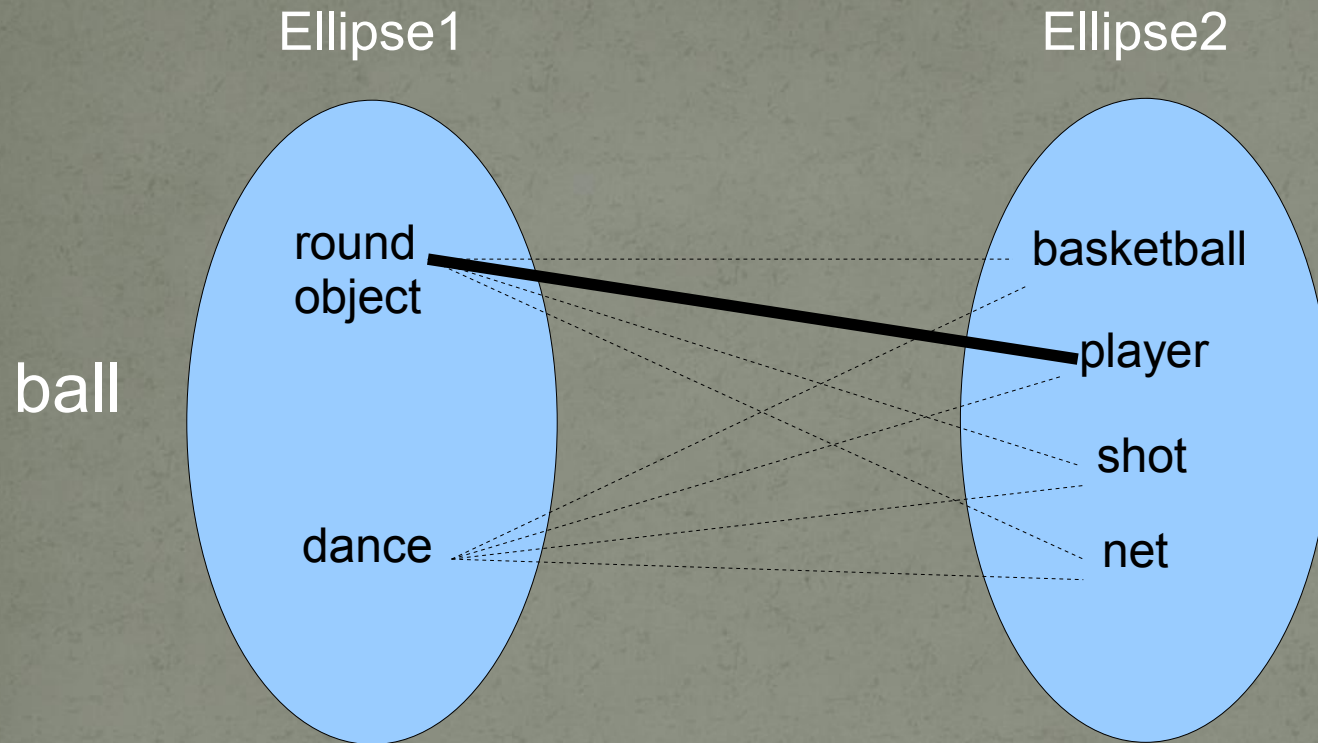
General Algorithm

The basketball player shot the ball into the net.

The algorithm works as below

- Remove the stop words like *the, into* from the sentence which do not significantly affect the disambiguation
- Start from the first word i.e. *basketball*, in this case, which is called the target word
- Compute the score for each sense of the target word by computing semantic similarity between the word and the words surrounding it
- The sense with the highest score is assigned as the most probable sense

Continued General Algorithm



Ellipse1 represents senses of the word *ball*. Ellipse2 represents all words with which all of these senses are compared. The sense 'round object', which has high semantic relatedness with the word *player* is represented with a bold line.

Goals of the thesis

- The overall goal of this thesis is to advance the state of the art in performing all words sense disambiguation.

In particular this can be used in the applications that seek to understand documents so as to better summarize, categorize, or translate them.

- This research will contribute in better understanding of natural language by computers.

Our work will start with the methods developed in SenseRelate:: AllWords and improve them in significant ways.

Goal 1

- To determine the advantages of expanding and improving context window
- Example:

The basketball player, who had been feeling very sick before, shot the ball into the net for the victory.

The word *player* which decides the sense of the word *ball* is far away from the word *ball*.

The situation is worse if a word related to the other sense of *ball* is present in the sentence!

Continued..

- As we saw in the previous example, choice of context window does matter in performance of disambiguation.
- Large Window Size
 - performs better if related words are far away from each other
- Smaller window size
 - faster as there are less comparisons
 - may result in best match

Current implementation allows user to specify context window size. However, deciding the context window according to the situation, certainly deserves more exploration which would help in better disambiguation.

Goal 2

- Exploring how fixing the senses work and how local Vs global information helps in disambiguation
- Does fixing sense help?
 - Yes, it might improve the performance by avoiding unnecessary comparisons
 - No, it might affect the disambiguation

Continued Goal 2..

- The current implementation gives an option to fix the sense, however, the questions below remain unanswered
 - in which situation we can fix the sense?
 - Will that affect the score of senses of other words

In some cases, if we assign a wrong sense to a word and fix it, it will propagate and the other words in the sentence might be assigned wrong senses

Moreover, if we fix sense depending upon global information, it may fail in certain local instances

Goal 3

- Knowing if part of speech tag or syntactic information is useful
- Example

Hand me the spoon.

Hand/VB me/PP the/DT spoon/NN

In this case *hand* has assigned a part of speech tag as a verb. This information would certainly be useful in disambiguation

Continued Goal 3

- Using part of speech tag can improve the performance as
 - there will be less comparisons
 - it restricts the senses of other words
- Current system can accept part of speech tagged text as an input
- It is necessary to find out in which case such information would really be useful and disambiguation could be improved

Evaluation Plan

For experiments and evaluation of the algorithm, a manually sense-tagged corpora will be used. A sense tagged corpora is a body text where human experts have assigned sense tag using a knowledge base such as dictionary.

The words in the manually sense-tagged data will be assigned senses using the improved algorithm, which will be compared with the senses of manually sense-tagged data

SemCor

- The SemCor corpus is created by the Princeton University
- It is the same group of people who developed WordNet
- SemCor is a subset of the English Brown corpus containing 360,000 words
- All the words are tagged by part of speech, and more than 200,000 content words are also sense-tagged according to Princeton WordNet 2.1

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

SENSEVAL

- The main goal of SENSEVAL is to evaluate strengths and weaknesses of various Word Sense Disambiguation systems
- These were competitions that were held in order to evaluate various WSD systems
- This data was created by the competition organizers in order to rank the performance of the participating systems.

SENSEVAL-2

- This was held in 2001 in France
- The data is small subset of Penn Treebank corpus
- Consists of 4873 words taken from Wall Street Journal articles
- 2473 are open class words most of which are found in WordNet 2.0

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

SENSEVAL-3

- This was held in 2004
- The data is small subset of Penn Treebank corpus
- Consists of 4883 words taken from Wall Street Journal articles and a work of fiction
- 2081 are open class words, out of which 1979 are found in WordNet 2.0

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

- SENSEVAL-2 and SENSEVAL-3 data will be used for the evaluation

Questions?

THANK YOU

