

Word Sense Disambiguation

WordNet::SenseRelate::AllWords

Taavet Kikas, Margus Treumuth
April 29, 2007

1 The Task

The task was to evaluate and assess the usefulness of an open-source word sense disambiguator called WordNet::SenseRelate [SensRel]. A software developed by Ted Pedersen and his team.

The disambiguator comes in two flavors: WordNet::SenseRelate::AllWords, which can be used to disambiguate whole texts; and WordNet::SenseRelate::TargetWord, which is used for disambiguating a given word. The aim of this task was to evaluate first of them.

2 Word Sense Disambiguation

Word sense disambiguation is the process of finding the correct sense of a word depending on its context. For example word *bank* can mean financial institution, landform, supply etc. The actual meaning is determined by the context.

Word sense disambiguation techniques are often divided into two categories: supervised word sense disambiguation and unsupervised word sense disambiguation. Supervised word sense disambiguation relies on a sense-tagged corpus and uses information from the corpus to perform disambiguation. In contrast, unsupervised word sense disambiguation does not need sense-tagged corpus. It usually relies on a machine-readable dictionary [Ng & Zelle 1997].

The main advantage of unsupervised learning is that it does not need a sense-tagged corpus, which is time-consuming to construct. On the other hand it generally does not achieve as high accuracy as supervised word sense disambiguation [Ng & Zelle 1997].

WordNet::SenseRelate is also an example of a unsupervised word sense disambiguator. It uses a special machine-readable dictionary called WordNet [WordNet].

3 WordNet

WordNet [Fellbaum 1998] is a machine-readable semantic lexicon for the English language in which words and collocations with synonymous senses are grouped into synonym sets called *synsets*. For example a set {*world, human race, humanity, humankind, human beings, humans, mankind, man*} is a synset consisting of words and collocations defined as “all of the inhabitants of the earth”. Definitions like this are part of WordNet and besides being useful for humans; these definitions play an important role in certain disambiguation algorithms. Short definitions are also referred to as glosses.

A word in WordNet can belong into multiple synsets. For example, the word “man” has 11 meanings as a noun and 2 meaning as a verb resulting in 13 different synsets. WordNet

contains 4 word categories: nouns, verbs, adverbs and adjectives. Word senses are ordered in WordNet in the order of frequency they appear in SemCor corpus.

Synsets in WordNet are connected to other synsets by means of different relations:

- *hypernyms*: A is hypernym of B if B is a kind of A,
- *hyponym*: B is hyponym of A if B is a kind of A,
- *coordinate terms*: A is a coordinate term of B if A and B share a hypernym,
- *holonym*: A is a holonym of B if B is a part of A,
- *meronym*: B is a meronym of A if B is a part of A,
- *entailment*: the verb A is entailed by B if by doing B you must be doing A,
- *etc.*

These relations are used in various language technology application.

4 Semantic relatedness

One of the key concepts in WordNet::SenseRelate disambiguator is the measure of semantic relatedness. Semantic relatedness is a measure that describes how strongly words are semantically connected. For example, *Earth*, *planet* and *sun* are all semantically related, but *Earth* is related more to *planet* than to *sun*. This relatedness is of course somewhat subjective. There are several measures for describing semantic relatedness, some of which have been used in WordNet::SenseRelate. These measures include but are not limited to: adapted Lesk measure, context vectors and Hirst & St-Onge measure.

Two first methods calculate the measure of semantic relatedness by comparing sense definitions. The more similarities in the definitions, the more related the words are assumed to be. Hirst & St-Onge method is based on an assumption that a relatively short chain having a systematic direction connects related words. The method attributes different weights to different relation. Some of the other methods not listed above rely on information content concept. An overview of semantic relatedness measures can be found in [Budanitsky 1999],[Michelizzi 2005].

4.1 Adapted Lesk

One of the relatedness measures used in word sense disambiguation is Lesk measure [Lesk 1986]. The key idea behind this measure is that glosses of related senses are more similar than those of unrelated senses. The measure of relatedness is calculated from the length of gloss overlaps (longest sequences that occur in both glosses).

WordNet::SenseRelate::AllWords uses an adapted Lesk measure [Banerjee & Pedersen 2002], which takes uses of WordNet relations. The idea is not just to compare glosses of possible synsets but also the glosses of synsets that are related to that synset, e.g. hypernym, hyponym, holonym etc.

Experiments conducted by Michelizzi [Michelizzi 2005] show adapted Lesk being superior than other relatedness measures used in WordNet::SenseRelate::AllWords.

5 Disambiguation Algorithm

Word sense disambiguation is the process of finding the correct sense of a word in a given context. The word being disambiguated is also referred to as the *target word*. The words

around the target word are called *context words*. The latter form so called *context window*. See Figure 1.

TWO [TAX REVISION **BILLS** WERE PASSED]

Figure 1. An example of a target word *BILLS* with context words *TEXT*, *REVISION*, *WERE* and *PASSED* forming a context window of size 4.

The basic idea behind the WordNet::SenseRelate::AllWords disambiguation algorithm is to calculate a score for each sense of the target word, based on its context, and then choose the sense with the highest score [Michelizzi 2005]. The score of a sense s_i is calculated as

$$score_{s_i} = \sum_{j,k} \max_k relatedness(s_i, s_{jk})$$

where s_{jk} is the k -th sense of j -th context word. After calculating the scores for every sense, the sense with the highest score is chosen. The result depends on the choice of relatedness measure and on context window.

6 The Software

The installation of WordNet::SenseRelate::AllWords software turned out to be a real challenge that took about 3 hours to complete. The most important part of the installation was to get the right versions of all packages. If all the newest versions are used, the installation is most likely not going to work without seriously rewriting some Perl code.

To avoid these installation problems, one should obtain the compatible versions of all the packages.

These are the versions that are compatible with each other:

- 1) WordNet-SenseRelate-AllWords-0.06
- 2) Text-Similarity-0.02
- 3) WordNet-Similarity-0.08
- 4) WordNet-QueryData-1.37
- 5) WordNet 2.0

WordNet-QueryData-1.37 is not available on CPAN or on SourceForge, but it was possible to obtain the package with the help of a Google search.

If installing under Windows, *NMAKE* for Perl should be used instead on Perl *make*.

The installation (as well as the software itself) is quite slow even on fast computers and at some point no progress indicators are shown. The installation should not be interrupted at these points. It will seem hung, but its not. It will eventually finish.

We brought the installation topic out because we can see that the complexity of the installation can prevent from more wide use of the software.

7 Experiments

The goal of the experiments was to evaluate WordNet::SenseRelate::AllWords disambiguator by measuring its precision.

We used a subset of semantically annotated SemCor 2.0 corpora as the Gold Standard. SemCor corpus itself is a subset of Brown Corpus. Semantical tags in SemCor conform to WordNet, thus making the corpus well suitable for our testing purposes.

Still some preprocessing had to be done to unify SemCor format with the disambiguator output:

- 1) redundant XML notation was removed preserving only word senses and POS tags
- 2) all sentences from the Gold Standard that contained compound phrases like *primary_election* and *City_of_Atlanta* were removed as the word sense disambiguation software was not able to handle these and the automatic evaluation would have been disturbed by them

We created a web interface to reformat SemCor XML to match

Wordnet::SenseRelate::AllWords input/output. The implementation can be found at

http://www.dialogid.ee/courses/ngslt_semantics/xml_to_plain.php

The Gold Standard consists of 63 sentences (1022 words) of text – both raw and POS tagged. The small amount of the Gold Standard sentences can be seen as a problem, yet the word sense disambiguator was very slow – it spent ca 2 minutes per sentence - thus the small size of the test corpora.

Two different experiments were conducted: disambiguating raw text and text with POS tags. Adapted Lesk measure was chosen as the semantic relatedness measure for the experiments since it has shown good results in previous experiments [Michelizzi 2005]. Size of the context window was set to be 4. Both chosen options also correspond to WordNet::SenseRelate::AllWords defaults.

8 Results

We analyzed 63 sentences (1022 words) of text – both raw and POS tagged. An example of the disambiguation output is given on Figure 2. The complete results of the evaluation are given below.

8.1 Gold Standard and text without POS tags

We achieved 48.53% precision when disambiguating raw text. The precision in word category was higher: 76.12%. This means that the analyzer guessed correctly the word category (verb, noun, adjective, etc.) but then missed the sense. The precision in word category is still much lower than in good POS tagging systems.

8.2 Gold Standard and text with POS tags

We achieved 63.30% precision when disambiguating text with POS tags. This comes very close to what Michelizzi achieved in his experiment [Michelizzi 2005]. The precision in word

category was not relevant here (it would have been 100%) as the POS tags (directly derived from word categories) were taken straight from Gold Standard.

```
Current configuration:
context file : input.txt
format      : parsed
scheme     : normal
tagged text : no
measure    : WordNet::Similarity::lesk
window     : 4
contextScore : 0
pairScore  : 0
measure config: (none)
trace      : no
forcepos   : no
compound file : (none)
stoplist   : (none)
Loading WordNet... done.

The fisherman#n#1 jump#v#1 off#r#1 the bank#n#1 and into the water#n#1
The bank#n#1 down#a#1 the street#n#1 wa#n#1 rob#v#1
Back#n#5 in#n#3 the day#n#4 we have#v#2 an#n#1 entire#a#3 bank#n#1 of computer#n#1 devote#v#1 to this problem#n#3
The bank#n#1 in#n#3 that road#n#1 be#v#1 entirely#r#1 too#r#2 steep#v#2 and be#v#1 really#r#1 dangerous#a#1
The plane#n#1 take#v#17 a#n#5 bank#n#1 to the left#a#1 and then#r#2 head#v#5 off#r#1 towards the mountain#n#1
He#n#1 claim#v#2 to do#v#9 the good#a#1 chicken-fry in#n#3 Texas#n#1
Eve#n#4 be#v#1 blond#a#1
```

Figure 2. An example of disambiguator output. Words have been tagged for part of speech and sense. The sense is given as a number referring to a meaning in WordNet, e.g. *bank#n#1* refers to the first sense of noun *bank* in WordNet.

9 Conclusion

We achieved precision as high as 63.30%. The result shows that the method is competitive with other word unsupervised disambiguation algorithms [Ng & Zelle 1997][Ide & Véronis 1998].

It would be interesting to try the same approach with Estonian language and with the Estonian Wordnet as the disambiguator is open-source. Maybe it could be done with just a few adjustments. We might look into that in the further.

One shortcoming of the disambiguation method is its slow speed, which makes it impossible to use it in real-time applications. The other shortage that should be dealt with if possible is the complex installation process, which might prevent some users from utilizing this software.

References

[SensRel] WordNet::SenseRelate homepage:
<http://www.d.umn.edu/~7Etpederse/senserelate.html>

[Ng & Zelle 1997] Ng, Hwee Tou, & Zelle, John. Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing. *AI Magazine*, 18(4) (pp. 45 – 64), 1997

[Fellbaum 1998] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

- [Michelizzi 2005] J. Michelizzi. Semantic Relatedness Applied to All Words Sense Disambiguation - Master of Science Thesis, Department of Computer Science, University of Minnesota, Duluth, July, 2005.
- [Hirst and Onge 1998] Graeme Hirst & David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 13, pages 305–332. MIT Press, 1998.
- [Banerjee & Pedersen 2002] Satanjeev Banerjee & Ted Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, 2002
- [Lesk 1986] Lesk, Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. SIGDOC, New York, 1986
- [Budanitsky 1999] Alexander Budanitsky. Lexical Semantic Relatedness and Its Application in Natural Language Processing. 1999
- [WordNet] WordNet's homepage: <http://wordnet.princeton.edu/papers>
- [Ide & Véronis 1998] *Word Sense Disambiguation: The State of the Art*