Word sense disambiguation using WordNet and the Lesk algorithm

Jonas EKEDAHL
Engineering Physics, Lund Univ.
Tunav. 39 H537, 223 63 Lund, Sweden
f99je@efd.lth.se

Koraljka GOLUB
KnowLib, Dept. of IT, Lund Univ.
P.O. Box 118, 221 00 Lund, Sweden
koraljka.golub@it.lth.se

Abstract

Word sense disambiguation is the process of automatically clarifying the meaning of a word in its context. It has drawn much interest in the last decade and much improved results are being obtained.

In this paper we take the so-called Lesk approach. In our case, definitions of the senses of the words to be disambiguated, as well as of the ten surrounding nouns, adjectives and verbs, are derived and enriched using the WordNet lexical database.

Two possible implications of this project could be that the results are dependent on the characteristics of a test document and on the characteristics of glosses, which needs to be further investigated. The average precision performed worse (0.45) than baseline precision (0.60) which was based on always selecting the most frequent sense. However, the presented approach has several limitations: a small sample, and a big number of fine senses in WordNet, many of which are not that distinguishable from each other. The future work would include experimenting with different variations of the approach.

1 Introduction

Word sense disambiguation is the process of automatically clarifying the meaning of a word in its context. For example, the word *contact* can have nine different senses as a noun, and two different senses as a verb.

Word sense disambiguation has drawn much interest in the last decade and much improved results are being obtained (see, for example, (Senseval)). It can be important for a variety of applications, such as information retrieval or automated classification (for an example of the latter, see Jones, Cunliffe, Tudhope 2004).

Different approaches to word sense disambiguation have been taken. Many are based on different statistical techniques. Some require corpora that are tagged for senses and others employ unsupervised learning. In this paper we take the so-called Lesk approach (Lesk 1986), which involves looking for overlap between the words in given definitions with words from the text surrounding the word to be disambiguated. In our case, definitions of the senses of the words to be disambiguated, as well as of the ten surrounding nouns, adjectives and verbs, are derived and enriched using the WordNet lexical database (WordNet). The sense definition chosen as correct is the one that has the largest number of words in common with the definitions of the surrounding words. A version of Lesk algorithm in combination with WordNet has recently been reported for achieving good word sense disambiguation results (Ramakrishnan, Prithviraj, Bhattacharyya 2004).

In this paper we conduct a pilot experiment, which is a part of a larger project that employs word sense disambiguation for improving accuracy of automated classification.

In the following chapter (2 Methodology) the approach is described in detail. Results are presented and the third chapter (3 Results), and in the last chapter conclusions are given and the future work is suggested.

2 Methodology

2.1. Introduction

In the paper a pilot experiment is conducted, that is a part of a larger project in which this word sense disambiguation approach would be applied for improving accuracy of automated classification.

The Lesk algorithm has first been implemented in its simple form by M. Lesk (1986). It is based on the assumptions that when two words are used in close proximity in a sentence, they must be talking of a related topic and, if one sense can be used by each of the two words to refer to the same topic, then their dictionary definitions must use some common words (Banerjee 2002, p 1). This approach involves looking for overlap between the words in dictionary definitions with words from the text surrounding the word to be disambiguated. The problem of this approach is that dictionary definitions often do not have enough words for this algorithm to work well, which can be overcome by using the WordNet lexical database (WordNet) (ibid.), because it contains different types of relationships between words, such as, for example, syononymy and hyper/hyponymy.

2.2. Creation of glosses from WordNet

In the research conducted by G. Ramakrishnan, B. Prithviraj and P. Bhattacharyya (2004), different types of relationships in WordNet have been experimented with. It showed that the best results are obtained when concatenating the descriptions of word senses with the glosses of its first- and second-levels hypernyms (ibid., p. 218). We adopted their approach. For example, the word *contact* in WordNet has nine senses for the noun, and two senses for the verb:

The noun *contact* has 9 senses in WordNet:

1. contact -- (close interaction; "they kept in daily contact"; "they claimed that they had been in contact with extraterrestrial beings")
2. contact -- (the state or condition of touching or of being in immediate proximity; "litmus paper turns red on contact with an acid")
3. contact -- (the act of touching physically; "her fingers came in contact with the light switch")
4. contact, impinging, striking -- (the physical coming together of two or more things; "contact with the pier scraped paint from the hull")
5. contact, middleman -- (a person who is in a position to give you special assistance; "he used his business contacts to get an introduction to the governor")
6. liaison, link, contact, inter-group communication -- (a channel for communication between groups; "he provided a liaison with the guerrillas")
7. contact, tangency -- ((electronics) a junction where things (as two electrical conductors) touch or are in physical contact; "they forget to solder the contacts")
8. contact, touch -- (a communicative interaction; "the pilot made contact with the base"; "he got in touch with his colleagues")
9. contact, contact lens -- (a thin curved glass or plastic lens designed to fit over

the cornea in order to correct vision or to deliver medication)

The verb *contact* has 2 senses in WordNet:

1. reach, get through, get hold of, contact -- (be in or establish communication with; "Our advertisements reach millions"; "He never contacted his children after he emigrated to Australia")
2. touch, adjoin, meet, contact -- (be in direct physical contact with; make contact; "The two buildings touch"; "Their hands touched"; "The wire must not contact the metal cover"; "The surfaces contact at this point")

For each sense, we take the description given in the brackets, e.g. for the seventh noun sense it is: (electronics) a junction where things (as two electrical conductors) touch or are in physical contact; "they forget to solder the contacts."

Then we extract two nearest hypernym levels of the word. The resulting gloss for the seventh sense of the noun *contact* would be:

```
contact, tangency --
 ((electronics) a junction where things (as t
wo electrical conductors) touch or are in p
hysical contact; "they forget to solder the c
ontacts")
       => junction, conjunction --
 (something that joins or connects)
           => connection, connexion, connect
or, connecter, connective --
 (an instrumentality that connects; "he sold
ered the connection"; "he didn't have the ri
ght connector between the amplifier and th
e speakers")
```

Words in the form *bank_building* have been converted into their components, i.e. in this example into *bank building* for easier later comparison.

Finally, while comparing, all words containing three characters and less are left out. This was done in order to leave out frequent words such as articles or pronouns; when there were more than one occurrences of a word, only one was retained. The final gloss for the seventh sense of the word contact would be:

amplifier between conductors conjunction connecter connection connective connector connects connexion contact contacts didn't electrical electronics forget have instrumentality joins junction physical right solder soldered something speakers tangency that they things touch where

The glosses were prepared using Prolog, since WordNet is available in Prolog (Obtaining WordNet).

2.3. Pre-processing the documents

Fifteen documents were selected and downloaded from the World Wide Web. They had to be prepared for the algorithm. First, they were converted into .txt format. Then they were pre-processed into Penn Treebank (Penn Treebank project) tokens using a sed Unix script (Tokenizer.sed). The part-of-speech tagger was MXPOST (MXPOST). Finally, regular expressions were used to put one word per line.

2.4. Comparing for overlapping words

From the pre-processed document, words to be disambiguated were extracted, together with senses of surrounding words. The surrounding words were simply five nouns or adjectives or verbs preceding the word to be disambiguated, and five nouns or adjectives or verbs following it. If a noun/adjective/verb was not in the WordNet, the next closest one was chosen.

Every sense of the word to be disambiguated was compared to each sense of the surrounding words. A number of combinations was derived and scores were assigned to them, based on the number of the overlapping words. For example, if a word to be disambiguated had two senses, and it was surrounded by two words, one having three different senses, and the other having two different senses, the number of derived combinations was 12, out of which six were for the first sense of the word to be disambiguated, and the other six were for the second sense of the word to be disambiguated. The sense chosen was the one in which group of six there was the combination with the highest score out of all the 12 combinations.

The Lesk algorithm itself was implemented in Prolog.

2.5. Sample

Three words to be disambiguated have been selected: bank, contact, and m/Mercury. Although all of these words have more than two senses, the aim of this pilot experiment was to disambiguate between the two major senses:

bank:

1) depository financial institution (two documents in the sample)
2) sloping land, especially the slope beside a body of water (three documents in the sample)

contact:

1) close interaction between people (two documents in the sample)
2) a junction where things (as two electrical conductors) touch or are in physical contact (three documents in the sample)

m/Mercury:

1) mercury: Hg, metallic element (three documents in the sample)
2) Mercury: the planet. (two documents in the sample)

For each word five documents have been manually selected, out of which two of them had one main meaning, and three another.

3. Results

On our small sample, the average precision performed worse (0.45) than baseline precision (0.60) which was based on always selecting the most frequent sense. However, this result should not be taken for granted, since the sample of three words and 15 documents is too small for any trustworthy results.

Instead, we could use some qualitative analysis:

1) The word bank has 18 senses in WordNet. The precision for all the five documents was relatively bad: 0.25, 0.16, 0.27, 0.30, and 0.5. In all the documents the often assigned sense was that of a piggybank, which might have to do with the fact that its gloss contains a lot of frequent words, such as usually, with, that, from, some.
2) The word contact has 11 senses listed in WordNet. The precision for the five documents was the following: 0.08, 1, 0.6, 0.625, and 0.92. This good result is partly due to the fact that we merged together two rather closely related senses, that of contact as communicative interaction, and that of contact as close human interaction. We were able to do this since the main aim of the experiment was to distinguish

between two totally unrelated senses of contact (see 2.5). While in one example we obtained 23 correct senses out of 25 occurrences, in another only 3 out of 38 were correctly assigned and in this case the extracted senses were not related to the topic of electrical contact.

3) The word m/Mercury has four senses listed in WordNet. The precision for the five documents was the following: 0.82, 0.5, 0.66, 0, and 0.05. The three first numbers are quite good results and all refer to discovering the sense of mercury as a metallic element. Not-so-good results in one of the other two documents is due to the fact that the document was discussing the temperature of the planet of Mercury, which produces the third sense of the word mercury in WordNet, about temperature.

4. Conclusion

Two possible implications of this project could be that the results are dependent on the characteristics of a test document and on the characteristics of glosses, which needs to be further investigated. However, the presented approach has several limitations: a small sample, and a big number of fine senses in WordNet, many of which are not that distinguishable from each other.

In order to determine which solution is best, the future work would include conducting experiments with:

○ WordNet preparation and document pre-processing (create a collection-specific stop-word list, apply stemming, do part-of-speech tagging on WordNet glosses, exclude examples from glosses which are in quotation marks, replace the ten-surrounding-word frame with a paragraph/sentence frame; experiment with different combinations of WordNet relations);

○ modify algorithm (the role of *tfidf* in precision, taking into account the number of words per gloss, experiment with different similarity measures); and

○ utilize WordNet Domains (Domain Driven Disambiguation), a file that contains synsets annotated by domain labels, such as Medicine, Architecture and Sport.

References

Desire : Development of a European Service for Information on Research and Education. http://www.desire.org/.

Domain Driven Disambiguation. http://wndomains.itc.it/download.html

Ganesh Ramakrishnan, B. Prithviraj, Pushpak Bhattacharyya. A Gloss Centered Algorithm for Word Sense Disambiguation. Proceedings of the ACL SENSEVAL 2004, Barcelona, Spain. P. 217-221.

Jones I., Cunliffe D., Tudhope D. 2004. Natural Language Processing and Knowledge Organization Systems as an aid to Retrieval. Proceedings 8th International Society of Knowledge Organization Conference (ISKO 2004), UCL London. (Ed: Ia C. McIlwaine), Advanced in knowledge Organization, 9, Ergon Verlag. P. 351-356.

Lesk, Michael. 1986. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In Proceedings of the 1986 SIGDOC Conference, pages 24–26, New York. Association for Computing Machinery.

MXPOST : Maximum Entropy Part-Of-Speech Tagger, and MXPARSE: (local) Maximum Entropy Parser. http://www.cis.upenn.edu/~adwait/penntools.html#Tools

Obtaining WordNet. http://www.cogsci.princeton.edu/~wn/obtain.shtml

The Penn Treebank project. http://www.cis.upenn.edu/~treebank/

Satanjeev Banerjee. 2002. Adapting the Lesk algorithm for Word Sense Disambiguation to WordNet. Master's thesis. Dept. of Computer Science, University of Minnesota, USA. http://www.d.umn.edu/~tpederse/Pubs/banerjee.pdf

Senseval : evaluation exercises for Word Sense Disambiguation. http://www.senseval.org/

Tokenizer.sed. http://www.cis.upenn.edu/~treebank/tokenizer.sed

WordNet : a lexical database for the English language. http://www.cogsci.princeton.edu/~wn/