CONSTRUCTION OF SEMANTIC RELATIONS FOR ENHANCING WORD SENSE DISAMBIGUATION IN QUESTION ANSWERING SYSTEMS

Mrs.C. Meenakshi
Asst.Professor, MCA Dept,
Vivekanandha Institute of Engineering
College for Women,
E-mail: meenasi.c@gmail.com

Dr.P. Thangaraj
Prof & Head, Dept. of CSE,
Bannari Amman Institute of Technology,
E-mail: ctpt@bitsathy.ac.in

Abstract

Word sense disambiguation is a significant problem at the lexical level of natural language processing. The philosophy is to determine the meaning of a word in a particular usage, by using sense similarity and syntactic context with corpus evidence as well as semantic relations from WordNet. A training set will be constructed for each word tag (using the corpus). Each training example is represented as a word and relationship label which is word#rel1#rel2#relN. In the testing phase, for each test sentence, the words are tokenized and relationship of the target word with other words is constructed through mapping with the training set. Established relation is taken to the wordnet for identifying the contextual sense which is selected as the correct sense.

***Keywords*:** Word Sense Disambiguation, Semantic relation, identifier and corpus.

Introduction

Word sense disambiguation is defined as the task of finding *the* sense of a word in a context. In the field of computational linguistics, ambiguity is one of the problems which pose a great challenge for computational linguists. In general, people are unaware of the ambiguities in the language they use because they are very good at resolving them using context and their knowledge of the world. However computer systems do not have this knowledge, and consequently do not do a good job of making use of the context. It is obvious that when a particular content posses more than one meaning and thereby understood in more than one possible way, ii becomes ambiguous. If the ambiguity is in a sentence or clause, it is called structural (syntactic) ambiguity. If it is in a single word, it is called lexical ambiguity.

The sentence "The man saw the girl with the telescope" belongs to structural ambiguity. This sentence is ambiguous since it can be interpreted in two ways: The man saw the girl who possessed the telescope or, the man saw the girl with the aid of the telescope. However, the sentence "The man saw the girl with a red hat" is not ambiguous for a human reader (people have the knowledge that a hat cannot be used to see), while it has the same ambiguity as the previous example for a computer.

Lexical semantic ambiguity occurs when a single word is associated with multiple senses. It is envisaged to focus on lexical semantic ambiguity. Examples of lexical ambiguity are everywhere. In fact, almost any word has more than one meaning.

For example, consider the noun *party*. It can refer to at least 5 different things as follows:

- an organization to gain political power
- an occasion on which people can assemble for social interaction and entertainment
- a band of people associated temporarily in some activity
- a group of people gathered together for pleasure
- a person involved in legal proceedings

The above said different senses can be subsumed in just one sense such as "group of people", however, for various applications, such as information retrieval or machine translation, it is important to be able to distinguish between the different senses of a word. In a machine translation application, the different senses of a word may be represented with different words in the target language. In order to correctly translate a text in one language to another, the prerequisite is to know the senses of the words and then find the best translation equivalent in the target language. Apart from these, for many other words there is no such general sense like the one for the noun *party*.

Lexical ambiguity can refer to both homonymy and polysemy. Homonyms are words that are written the same way, but are (historically or conceptually) really two different words with different meanings which seem unrelated. Examples are *suit* ("lawsuit" and "set of garments") and *bank* ("river bank" and "financial institution"). If a word's meanings are related, it is called a polyseme. The word *party* is polysemous because its senses can be generalized as "group of people", that is they are related.

The meaning of the noun *party* is considered in the following sentence:
Mr. Smith's party took 38% of the votes in the last elections.

It is clear to a human reader that the noun *party* is in the sense "an organization to gain political power" in the sentence above. Most people are not even aware of the ambiguity

contained in the sentence. Humans are so skilled at resolving potential ambiguities that they do not realize they are doing it. There is considerable focus on how people resolve ambiguities; however it is still not known how exactly humans do lexical disambiguation. Therefore, it is a difficult task to teach a computer to do the same thing. If there is more than one ambiguous word in a sentence, the number of potential interpretations of the sentence increases dramatically. The number of interpretations of a sentence is the product of all possible meanings of the words that construct that sentence. In the above sentence, the term *party* has 5, *take* has 42, *vote* has 5, *last* has 10, and *election* has 4 senses. Therefore there are 42000 possible interpretations for the example sentence. The most prominent way to disambiguate a word is examining its context. The context can be considered as the words surrounding the ambiguous word, which is the noun *party* in this case. The words such as *vote* and *election* might be a good clue for the sense of the noun *party*. But context is not the only information available for disambiguation. Syntactic classes of the words in the ambiguous word's context (whether they are noun, verb or adjective, etc.), be whether the ambiguous word plays the role of object or subject in the syntactic structure of the sentence may also be used in the disambiguation process.

WSD Approaches

WSD algorithms can be divided into three based on the way they acquire information. These approaches include the following:

- Corpus based approaches
- Knowledge based approaches

In *corpus based approaches*, information is gained from training on some corpus. A corpus provides a set of samples that enables the systems to develop some numerical models. It can further be classified into two subclasses based on the training corpus as follows:

- Supervised disambiguation
- Unsupervised disambiguation
-

In supervised WSD the training data is sense-tagged whereas in unsupervised WSD the training data is a raw corpora which are not semantically disambiguated. The aim in supervised disambiguation is to build a classifier which correctly classifies new cases based on their context of use. Machine learning algorithms such as Bayesian classifiers (Duda and Hat, 1973) [1], decision lists (Rivest, 1987)[2], decision trees (Quinlan, 1986)[3], k-nearest neighbor and neural networks (Rumelhart et al., 1986)[4] fall into this category. A major problem with supervised approaches is the need for a large sense-tagged training set. Despite the availability of large corpora, manually sense-tagging of a corpus is very difficult and very few sense-tagged data are available now. The two largest corpora that are available are the SemCor corpus (Landes et al., 1998) [5] and the SENSEVAL corpus (Kilgariff and Rosenzweig, 2000)[6].

There have been several efforts for finding a way to sense-tag corpora automatically. ***Bootstrapping*** is the most frequently used method for this purpose. Bootstrapping relies on a small number of instances of each sense for each lexeme of interest. These sense-tagged instances are used as seeds to train an initial classifier. This initial classifier is then used to extract a larger training set from the remaining untagged corpus. With each iteration of this process, the training corpus grows and the untagged corpus shrinks.

Another problem that supervised disambiguation methods face with is the data sparseness. Since the sense-tagged training corpus is finite and very few for WSD, some senses of polysemous words are very likely to be missing. The training data must ensure that all senses of a polysemous word are covered for a supervised algorithm to be successful.

Unsupervised Disambiguation

In unsupervised disambiguation, information is gathered from raw corpora which are not semantically disambiguated. Unsupervised methods correspond to clustering tasks rather than sense tagging tasks.

Indeed, completely unsupervised disambiguation is not possible for word senses since sense tagging requires characterization of the senses.

Infrequent senses and senses that have few collocations are hard to isolate in unsupervised disambiguation. In general, accuracy of unsupervised WSD systems are 5% to 10% lower than that of other algorithms since no lexical resources for training or defining senses are used.

Knowledge Based Approaches

The earlier methods require considerable amount of work to create a classifier for each entry in the lexicon. Because of this reason, they are able to report results on very few lexical items. With the availability of large-scale lexical resources, such as dictionaries, thesauri and corpora, work on WSD has focused on large-scale disambiguation. WSD based on machine readable dictionaries, thesauri and computational lexicons are briefly reviewed here.

Machine Readable Dictionaries

Machine Readable Dictionaries (MRD) provide a ready made information source of word senses. The first attempt to use MRD's came from Lesk (1986)[38]. He started from the simple idea that a word's dictionary definitions are likely to be

good indicators of the senses they define. The accuracy of the method is reported to be 50-70% which is a good result considering the fact that a fine set of sense distinctions are used.

In view of the fact that dictionaries are created for human use, not for computers, there are some inconsistencies. Although they provide detailed information at the lexical level, they lack pragmatic information used for sense determination. For instance, the relation between *ash* and *tobacco*, *cigarette* or *tray* is very indirect in a dictionary whereas the word *ash* co-occurs very frequently with these words in a corpus (Ide and Veronis, 1998) [7].

Thesauri

Thesauri provide information about relationships among words. Thesaurus based disambiguation makes use of the semantic categorization provided by a thesaurus or a dictionary with subject categories. The most frequently used thesaurus in WSD is Roget's International Thesaurus (Roget, 1946) which appeared in machine-tractable form in 1950's. The basic inference in thesaurus-based disambiguation is that semantic categories of the words in a context determine the semantic category of that context as a whole. This category then determines the correct senses that are used.

Similar to machine readable dictionaries, a thesaurus is a resource for humans, so there is not enough information about word relations. Therefore, Roget's or any other thesauri are not used extensively.

Computational Lexicons

The usefulness of lexical relations in linguistic, psycholinguistic and computational research has led to a number of efforts to create large electronic databases of such relations. Beginning from the mid-1980's, construction of semantic lexicons by hand has emerged. Some examples of these lexicons are WordNet (Fellbaum 1998; Miller et al. 1990), CyC (Lenat and Guha 1990), ACQUILEX (Briscoe 1991), and COMPLEX (Grishman, Macleod, and Meyers 1994).

Since WordNet is the most popular lexicon among the above and since it is used in the work done in this paper both for sense evaluation and for similarity measure, this section presents detailed *information about it.*

WordNet

WordNet is an online lexical reference system which originated at Princeton University under the direction of Professor George A. Miller. It combines many features used for WSD in one system. It includes definitions of word senses as in a dictionary; it defines "synsets" of synonymous words representing a single lexical concept like a thesaurus; and it includes word-to-word relations.

WordNet consists of three databases: noun database [9], verb database and one database for adjectives and adverbs. Each database consists of lexical entries corresponding to unique orthographic forms. Each form is associated with a set of senses.

Experiment

The construction of semantic relations is through improving Lin's algorithm by using semantic dependencies from the WordNet.

1. Training Corpus Construction

E.g. If "bank" is observed in the corpus then relationship among other words along with which it appears can be derived as "is a kind of building, river side". "has part transactions" "pertains to water", by using various relations.

- hypernymy (car#1 is a kind of vehicle#1) denoted by (kind-of)
- hyponymy (the inverse of hypernymy) denoted by (has-kind)
- meronymy (room#1 has-part wall#1) denoted by (has-part)
- holonymy (the inverse of meronymy) denoted by (part-of)
- pertainymy (dental#1 pertains-to tooth#1) denoted by (pert)
- attribute (dry#1 value-of wetness#1) denoted by (attr)
- similarity (beautiful#1 similar-to pretty#1) denoted by (sim)

2. Stop Word Removal

This is the process of linguistic normalisation, in which various forms of a word are reduced to a common form. For the sentence given as example:

"The crazy man said 'you are the funniest guy i know' to the man who stood near to him..."

(i) the tokenizer module would remove the punctuation and return an arrayList of words
(ii) the stop word remover would remove words like "the", "to", etc.
(iii) the stemmer would reduce each word the their 'root', for example 'funniest' would become funny.

3. Constructing Relations

First step is to arrive with tokens. Next relationship of the target word among other words is determined with the help of training corpus. Then the constructed relation is mapped against wordnet to identify the conceptual sense which is coined as the correct sense.

Evaluation of WSD

The test kit initially has coverage of 10 words. The precision and recall are computed as shown below. It's compared with the heuristic approach and our semantic relation construction algorithm gives higher precision.

	Cover.	Prec. %	Recall %
Semantic relation construction	W=10	72.1	64
Heuristic algorithm		68.3	64

Conclusion

In this paper, a method for constructing semantic relations using word net is proposed. The automatic method for the disambiguation presented in this paper is ready-usable in any general domain and on free-running text, given the relationship in the corpus. It does not need any training and uses word sense tags from WordNet, an extensively used lexical data base.

Another heuristic method, was also tried on the same texts, showing that our algorithm performs better. Results are promising, considering the difficulty of the task (free running text, large number of senses per word in WordNet), and the lack of any discourse structure of the texts. Two types of results can be obtained: the specific sense or an approximate level.

References

1. Michael Lesk. 1986. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone", in Proceedings of the 5th annual international conference on Systems documentation, Toronto, Ontario, Canada, 1986.
2. Walker D. and Amsler R. 1986. "The Use of Machine Readable Dictionaries in Sublanguage Analysis", in Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press, pp. 69-83, 1986.
3. Yarowsky, David. 1992. "Word sense disambiguation using statistical models of Roget's categories trained on large corpora", in Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, 454-460, 1992.
4. Yarowsky, David. 1994. "Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French", in Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL), Las Cruces, U.S.A., 88-95, 1994.
5. Yarowsky, David. 1995. "Unsupervised word sense disambiguation rivaling supervised methods", in Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), Cambridge, MA, 189-196, 1995.
6. Agirre, Eneko & German Rigau. 1996. "Word sense disambiguation using conceptual density", in Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, Denmark, 1996
7. Philip Resnik, 1999,"Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research, 1999.
8. Resnik P. 1995. Disambiguating Noun Groupings with Respect to WordNet Senses, in Proceedings of the Third Workshop on Very Large Corpora, MIT. Richardson R., Smeaton A.F. and Murphy J. 1994.
9. Using WordNet as a Konwledge Base for Measuring Semantic Similarity between Words, in Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland.
10. Rigau G. 1994. An experiment on Automatic Semantic Tagging of Dictionary Senses, WorkShop "The Future of Dictionary", Aix-les-Bains, France. published as Research Report LSI-95-31-R. Computer Science Department. UPC. Barcelona.
11. Rigau G. and Agirre E. 1996. Linking Bilingual Dictionaries to WordNet, in proceedings of the 7th Euralex International Congress on Lexcography (Euralex'96), Gothenburg, Sweden, 1996.
12. Sussna M. 1993. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia. Voorhees E. 1993.
13. Using WordNet to Disambiguate Word Senses for Text Retrival, in proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, pages 171-180, PA.
14. Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B. 1993. Providing Machine Tractablle Dictionary Tools, in Semantics and the Lexicon (Pustejovsky J. ed.), 341-401. Yarowsky, D. 1992.
15. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, in proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France.

Biography

C. Meenakshi is a research scholar in Computer Science Department, Mother Terasa University, Kodaikanal. She is currently working as Assistant Professor in the Department of MCA, Vivekanandha Institute of Engineering and Technology, Tiruchengode. Her current research interest include Natural Language Processing, Information Retrieval and Machine translation.

Dr. P. Thangaraj is Professor & Head, Computer Science and Engineering Department in Bannari Amman Institute in Technology, Sathyamangalam. His current research focuses on Natural languages, Fuzzy theory and computational problems.