ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
June 2013,
Vol. 6, No. (2):
Pgs.113-118

Enriching WordNet for Word Sense Disambiguation

SWATI RASTOGI and SANJEEV THAKUR

Department of Computer Sc. & Engineering , Amity School of Engineering & Technology, Noida, India.
Swatirastogi13@gmail.com, sthakur3@amity.edu

(Received: May 31, 2013; Accepted: June 10, 2013)

ABSTRACT

In computational linguistics, word-sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings. Research has progressed steadily to the point where WSD systems achieve 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. Among these, supervised learning approaches have been the most successful algorithms to date. The senses of a word are expressed by its WordNet synsets, arranged according to their relevance. The relevance of these senses are probabilistically determined through a Bayesian Belief Network. The main contribution of the work is a completely probabilistic framework for word-sense disambiguation with a semi-supervised learning technique utilising WordNet. Word sense disambiguation is a major problem in many tasks related to natural language processing. This paper aims to enriching wordnet for word sense disambiguation by adding some extra features to wordnet that may enhance the efficiency of knowledge-based contextual overlap WSD algorithms when they are used on wordnets.

Keywords: Wordnet, Word Sense Disambiguation, Natural Processing language

INTRODUCTION

Any of the machine-readable lexical databases modeled after the Princeton WordNet is called wordnet. WordNet is a lexicon - a database of over a hundred thousand words with meanings and a complex architecture of word links. In the field of computational linguistics, the problem is generally called word sense disambiguation (WSD), and is defined as the problem of computationally determining which "sense" of a word is activated by the use of the word in a particular context. WSD is essentially a task of classification: word senses are the classes, the context provides the evidence, and each occurrence of a word is assigned to one or more of its possible classes based on the evidence. This is the traditional and common characterization of WSD that sees it as an explicit process of disambiguation with respect to a fixed inventory of word senses. Words are assumed to have a finite and discrete set of senses from a dictionary, a lexical knowledge base, or an ontology (in the latter, senses correspond to concepts that a word

lexicalizes). Application-specific inventories can also be used. For instance, in a machine translation (MT) setting, one can treat word translations as word senses, an approach that is becoming increasingly feasible because of the availability of large multilingual parallel corpora that can serve as training data. Word sense ambiguity is a pervasive characteristic of natural language. For example, the word "cold" has several senses and may refer to a disease, a temperature sensation, or an environmental condition. The specific sense intended is determined by the textual context in which an instance of the ambiguous word appears. In "*I am taking aspirin for my cold*" the disease sense is intended, in "*Let's go inside, I'm cold*" the temperature sensation sense is meant, while "*It's cold today, only 2 degrees*", implies the environmental condition sense. To make word sense disambiguation more clear let's take another example-

Types Knowledge-Bases for Word Sense Disambiguation

English Dictionary: A reference book containing an alphabetical list of words, with information given for each word, usually including meaning, pronunciation etc.During the early 1980s, machine-readable dictionaries became a popular source of information for word sense disambiguation algorithms.

Thesauri

A word can appear in any number of different categories, although each of these categories is usually a distinct word sense.

FrameNet

The project has been in operation at the International Computer Science Institute in Berkeley since 1997. FrameNet is based on a theory of meaning called Frame Semantics. FrameNet has developed a visualization tool for viewing the relations between frames and their frame elements. The basic idea is straightforward: that the meanings of most words can best be understood on the basis of a semantic frame: a description of a type of event, relation, or entity and the participants in it. For example, the concept of cooking typically involves a person doing the cooking (Cook), the food that is to be cooked (Food), something to hold the food while cooking (Container) and a source of heat (Heating instrument).

WordNet

WordNet is a semantic lexicon for the English language created by Princeton University. WordNet groups English words into sets of synsets. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. The goal of WordNet was to develop a system that would be consistent with the knowledge acquired over the years about how human beings process language

Proposed WordNet Structure

We introduce a new WordNet database relation structure that keeps two more informative fields in addition to the five principal informative

fields which are found in the Princeton University's English WordNet database relation structure. The description of the fields is presented below:

The introduction of a "frequently used" or "highly expected" field in the synset structure of wordnets can scale-up the efficiency in determining winner sense of a polysemous word, as these highly related words will enrich the sense bag with more information, thereby enhancing the chances of appropriate overlap. For example the following list of terms may be considered as "highly expected" or "frequently used" with the concept of "computer":

1: Central Processing Unit
2: Keyboard
3: Mouse
4: Monitor
5: Universal Serial Bus
6.USB Stick

Thus putting the above list in the synset structure of the most appropriate sense of "computer" will result in attaining high degrees of overlap with sentences comprising of the word "computer" and several of the words from the above list.

we should try to capture the distributional constraints and other such relationships between different concepts of WordNet into account. For example the concepts of "cigarette" and "ash" have a relationship between them. If we keep information about this relationship in the WordNet then it will certainly be helpful in sense disambiguation of concepts of "ash" and "cigarette". If required, we may keep on enriching information related to such relationships between concepts in the WordNet. Incorporation of such relationships in the WordNet will increase the efficiency of knowledge-based contextual overlap WSD algorithms.

For each concept of the WordNet we must keep multiple glosses (explanations) for that sense. The phrases for the glosses must be made up of diverse vocabulary or words that are frequently used with that concept. This is because contextual overlap based WSD algorithms are usually very sensitive to the exact wording used in the definitions. So the absence of some words in the glosses for a sense *s* of a word *w* can radically change the overlap result, as glosses form the primary ingredient to the sense bag. The idea if implemented will increase the chances of overlap with the most appropriate sense of the concept in question using a Lesk like algorithm for WSD. However the decision on the choice of most frequently used words and/or phrases may be a tedious task for a lexicographer involved in designing the glosses for the concept. Introducing such additional fields may have will certainly enrich the sense bag with more information leading to high degrees of overlap for the most appropriate sense of a word.

Proposed approach has three additional fields which help in enriching the wordnet for word sense disambiguation.

CONCLUSION

In the present work I have presented a new WordNet database relation structure. The new database relation structure ensures enriching of the sense bag with more information leading to higher degrees of overlap for the most appropriate sense of a word in question, thereby achieving better quality word sense disambiguation of senses. By conducting experiments, I have verified that the introduction of these informative fields nicely enhances the efficiency of knowledge-based contextual overlap dependent WSD algorithms.I have used the Lesk Algorithm to do word sense disambiguation. My results indicate that the WSD based on proposed Wordnet is better.

REFERENCES

1. Satanjeev Banerjee and Ted Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet," *Lecture Notes In Computer Science*; **2276** , Pages: 136 - 145, (2002).
2. David Justin Craggs, "*An Analysis and Comparison of Predominant Word Sense Disambiguation Algorithms*," PhD

dissertation, Dept. of Computer Science, Edith Cowan University, (2011).

3. WordNet: a lexical database for English Language. [Online]; Available: http://wordnet.princeton.edu/index.shtml.
4. V. Balkova, A. Sukhonogov, S. Yablonsky, "Russian WordNet from UML-notation to Internet/Intranet Database Implementation", *GWC 2004,Proceedings*, Masaryk University,Brno, pp. 31-38, (2003).
5. B. Hettige, A. S. Karunananda, "Developing Lexicon Databases for English to Sinhala Machine Translation, Second International Conference on Industrial and Information Systems," *ICIIS* , **8** – 11 August , Sri Lanka, (2007).
6. S. R. Annam, M. Choudhury, S. Sarkar, A. Basu, "ABHIDHA: An extended wordNet for Indo Aryan Languages", *Journal of Research Issues in Data Engineering*, (2003).
7. D. Chakrabarti D. Narayan P. Pandey P. Bhattacharyya, "An Experience in Building the Indo-WordNet-A WordNet for Hindi," *GWC*- (2002).
8. Word sense disambiguation [Online]. Available: http://www.scholarpedia.org/article/Word_sense_disambiguation
9. http://en.wikipedia.org/wiki/Word-sense_disambiguation
10. http://www.cse.iitb.ac.in/~pb/papers/soft-wsd.pdf
11. http://wordnetweb.princeton.edu/perl/webwn?s=wordnet
12. http://opensource.ebswift.com/WordNet.Net/Introduction.aspx
13. http://www.wsdbook.org/chapter1.html
14. http://wsd.nlm.nih.gov/
15. http://kobus.ca/research/ugrad/
16. http://www.enggjournals.com/ijcse/doc/IJCSE12-04-07-062.pdf
17. Alok Chakrabarty, Bipul Syam Purkayastha, Lavya Gavshinde, "Ideas to Enhance Contextual Overlap for Knowledge-based Overlap Algorithms for Word Sense Disambiguation using Wordnet",(2010).
18. Manish Sinha, Mahesh Kumar, Prabhakar Pande, Lakshmi Kashyap and Pushpak Bhattacharyya, "Hindi Word Sense Disambiguation," *International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems*, Delhi, India, November, (2004).
19. Kilgarriff and J. Rosenzweig, "English SENSEVAL:Report and Results," *In Proceedings of the 2nd International Conference on Language Resourcesand Evaluation*, LREC, Athens, Greece. (2000).
20. Dongqiang Yang, David M.W. Powers, "Measuring semantic similarity in the taxonomy of WordNet," *In V. Estivill-Castro, editor, Proceedings of the 28th Australasian Computer Science Conference, Newcastle*, Australia, pp. 315-322, 2005).
21. Zakaria Elberrichi1, Abdelattif Rahmoun2, and Mohamed Amine Bentaalah, "Using WordNet for Text Categorization," *The International Arab Journal of Information Technology*, **5(1)**, (2008).
22. Y. Toshio, "The EDR electronic dictionary," *Communications of the ACM*, **38:11** , Pages: 42 - 44,(1995).
23. Aggire E. and Rigau G." Word Sense Disambiguation using Conceptual density*" In Proceeding of COLING96*.
24. Fellbaum, C. (ed.). WordNet: An Electronic Lexical Database, MIT Press (1998).
25. Michael Lesk. " 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* (SIGDOC 86), Virginia DeBuys (Ed.). ACM, New York, NY, USA, 24-26 (1986).
26. Hindi Wordnet from Center for Indian Language Technology Solutions, IIT Bombay, Mumbai, India [Online]. Available: http://www.cfilt.iitb.ac.in/WordNet/webhwn/
27. B. Hettige, A. S. Karunananda, "Developing Lexicon Databases for English to Sinhala Machine Translation," *Second International Conference on Industrial and Information Systems*, ICIIS , Sri Lanka, 8 – 11 (2007).
28. Hindi Corpora from Central Institute of Indian Languages, Mysore India. [Online]. Available: http://www.ciil.org
29. S. Jha D. Narayan P. Pande P. Bhattacharyya, "A Wordnet for Hindi," *Workshop on Lexical Resources in Natural Language Processing*,

India (2001).
30. Jaap Kamps, "Visualizing WordNet Structure"
31. D. Chakrabarti. P. Bhattacharyya, "Creation of English and Hindi Verb Heerarchies and their Application to Hindi Wordnet Building and English-Hindi MT", *GWC2004, Proceeding*, Masaryk University, Brno, pp 83-90, (2003).
32. S. Patanakul, P. Charnyote, "Construction of Thai WordNet Lexical database from Machine readable Dictionary", *Conference Proceedings: the tenth Machine Translation Summit*; pp.87-92 , Thailand, (2005).
33. Word sense disambiguation. [Online]. Available: http://www.scholarpedia.org/article/Word_sense_disambiguation
34. Yee Seng Chan, Hwee Tou Ng and David Chiang, "Word Sense Disambiguation Improves Statistical Machine Translation," *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, (2007).
35. Stanford Log-linear Part-Of-Speech Tagger [Online]. Available: http://nlp.stanford.edu/software/tagger.shtml
36. Java API for WordNet Searching (JAWS) [Online]. Available: http://lyle.smu.edu/~tspell/jaws/index.html
37. Word sense disambiguation [Online]. Available: http://www.scholarpedia.org/article/Word_sense_disambiguation
38. Lesk algorithm [Online]. Available: http://en.wikipedia.org/wiki/Lesk_algorithm
39. WordNet: a lexical database for English Language. [Online]; Available: http://wordnet.princeton.edu/index.shtml.
40. Manish Sinha, Mahesh Kumar, Prabhakar Pande, Lakshmi Kashyap and Pushpak Bhattacharyya, "Hindi Word Sense Disambiguation," *International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems*, Delhi, India, , (2004).
41. Kilgarriff and J. Rosenzweig, "English SENSEVAL:Report and Results," *In Proceedings of the 2nd International Conference on Language Resourcesand*
42. *Evaluation*, LREC, Athens, Greece. (2000).
43. Dongqiang Yang, David M.W. Powers, "Measuring semantic similarity in the taxonomy of WordNet," *In V. Estivill-Castro, editor, Proceedings of the 28th Australasian Computer Science Conference, Newcastle*, Australia, pp. 315-322, (2005).
44. Zakaria Elberrichi1, Abdelattif Rahmoun2, and Mohamed Amine Bentaalah, "Using WordNet for Text Categorization," *The International Arab Journal of Information Technology*, **5:1**, (2008).
45. Y. Toshio, "The EDR electronic dictionary," *Communications of the ACM*, **38:11** , Pages: 42 - 44,(1995).
46. Satanjeev Banerjee and Ted Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet," *Lecture Notes In Computer Science*; **2276**, Pages: 136 - 145, (2002).
47. David Justin Craggs, "*An Analysis and Comparison of Predominant Word Sense Disambiguation Algorithms*," PhD dissertation, Dept. of Computer Science, Edith Cowan University, (2011).
48. V. Balkova, A. Sukhonogov, S. Yablonsky, "Russian WordNet from UML-notation to Internet/Intranet Database Implementation", *GWC 2004,Proceedings*, Masaryk University,Brno, pp. 31-38, (2003).
49. B. Hettige, A. S. Karunananda, "Developing Lexicon Databases for English to Sinhala Machine Translation, Second International Conference on Industrial and Information Systems," *ICIIS* , 8 – 11 August , Sri Lanka, (2007).
50. S. R. Annam, M. Choudhury, S. Sarkar, A. Basu, "ABHIDHA: An extended wordNet for Indo Aryan Languages", *Journal of Research Issues in Data Engineering*, (2003).
51. Md. Abu Nuser Musud, Md. Muntusir Mamun Joarder, Md. Turiq-Ul-Azam, "A general approach to Natural Language Conversion,"
52. Aggire E. and Rigau G." Word Sense Disambiguation using Conceptual density*" In Proceeding of COLING96*.
53. Nakov, Preslav, & Ng, Hwee Tou , " Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-

Rich Languages," (2009).
54. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pp. 1358 – 1367, Singapore (2009).
55. Alok Chakrabarty, Bipul Syam Purkayastha, Lavya Gavshinde, "Ideas to Enhance
56. Contextual Overlap for Knowledge-based Overlap Algorithms for Word Sense Disambiguation using Wordnet",(2010).
57. Fellbaum, C. (ed.). WordNet: An Electronic Lexical Database, MIT Press. [17] 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* (SIGDOC 86), Virginia DeBuys (Ed.). ACM, New York, NY, USA, 24-26. (1998).
58. Hindi Wordnet from Center for Indian Language Technology Solutions, IIT Bombay, Mumbai, India[Online]. Available: http://www.cfilt.iitb.ac.in/WordNet/webhwn/
59. [19] B. Hettige, A. S. Karunananda, "Developing Lexicon Databases for English to Sinhala Machine Translation," *Second International Conference on Industrial and Information Systems*, ICIIS 2007, Sri Lanka, 8 – 11 August (2007).
60. [20] Hindi Corpora from Central Institute of Indian Languages, Mysore India. [Online]. Available: http://www.ciil.org
61. D. Chakrabarti D. Narayan P. Pandey P. Bhattacharyya, "An Experience in Building the Indo-WordNet-A WordNet for Hindi," *GWC*- (2002).
62. S. Jha D. Narayan P. Pande P. Bhattacharyya, "A Wordnet for Hindi," *Workshop on Lexical Resources in Natural Language Processing*, India (2001).
63. *Jaap Kamps,* "Visualizing WordNet Structure"
64. D. Chakrabarti. P. Bhattacharyya, "Creation of English and Hindi Verb Heerarchies and their Application to Hindi Wordnet Building and English-Hindi MT",
65. *GWC2004, Proceeding*, Masaryk University, Brno, pp 83-90, (2003).
66. [25] S. Patanakul, P. Charnyote, "Construction of Thai WordNet Lexical database from Machine readable Dictionary", *Conference Proceedings: the tenth Machine Translation Summit*; pp.87-92 , Thailand, (2005).
67. [26] Word sense disambiguation. [Online]. Available: http://www.scholarpedia.org/article/Word_sense_disambiguation
68. Yee Seng Chan, Hwee Tou Ng and David Chiang, "Word Sense Disambiguation Improves Statistical Machine Translation," *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, (2007).
69. Stanford Log-linear Part-Of-Speech Tagger [Online]. Available: http://nlp.stanford.edu/software/tagger.shtml
70. Java API for WordNet Searching (JAWS) [Online]. Available: http://lyle.smu.edu/~tspell/jaws/index.html
71. Word sense disambiguation [Online]. Available: http://www.scholarpedia.org/article/Word_sense_disambiguation
72. Lesk algorithm [Online]. Available: http://en.wikipedia.org/wiki/Lesk_algorithm
73. http://en.wikipedia.org/wiki/Word-sense_disambiguation