Improvement in Word Sense Disambiguation by introducing enhancements in English WordNet Structure

Deepesh Kumar Kimtani
Computer Science and Engineering
IIIT Bhubaneswar
Bhubaneswar, India
deepesh.kimtani @gmail.com

Jyotirmayee Choudhury
Computer Science and Engineering
IIIT Bhubaneswar
Bhubaneswar, India
jyotichoudhury@gmail.com

Alok Chakrabarty
Computer Science and Engineering
IIIT Bhubaneswar
Bhubaneswar, India
alokmcs @gmail.com

Abstract— W**ord sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying the appropriate sense of a word (i.e. intended meaning) in a sentence, when the word has multiple meanings. In this paper we introduce a new WordNet database relation structure whose usage enhances the WSD efficiency of knowledge-based contextual overlap dependent WSD algorithms, such as the popular Lesk algorithm. The efficiency of WSD, on the usage of the proposed WordNet over existing WordNet as a knowledge-base, has been experimentally verified by using the Lesk algorithm on a rich collection of heterogeneous sentences. Use of the proposed WordNet for Lesk Algorithm highly increases the chances of contextual overlap, thereby resulting in high accuracy of proper sense or context identification of the words. The WSD results and accuracies, obtained using the proposed WordNet, have been compared with the results obtained using existing WordNet. Experimental results show that use of our proposed WordNet results in better accuracy of WSD than the existing WordNet. Thus its usage will help the users better, in doing Machine translation, which is one of the most difficult problems of natural language processing**

Keywords- **word sense disambiguation, lesk, wordnet, polysemous, knowledge-base,contextual overlap.**

I. Introduction

Word Sense Disambiguation (WSD) is defined as the task of finding the *correct* sense of the word in a context. The task needs large number of words and word knowledge. The aim of any Word Sense Disambiguation (WSD) system is to obtain the intended senses of a set of target words, or of all words of a given text against a sense repository using the context in which the word appears. The sense repository can be a machine readable dictionary, a thesaurus or a computational lexicon like a WordNet [1] [2].

Typically the relating of a sense to a word using a knowledge-based contextual overlap algorithm is done by finding the best overlap between:

(i) The environmental words amongst which the polysemous word, to be disambiguated, appears and
(ii) The information in a WordNet

The sense in a WordNet with gives maximum overlap is declared as the *winner sense*.

II. WORDNET AS KNOWLEDGE-BASE

In 1985, Princeton University started developing a semantic lexicon called WordNet for the English language [3] [4]. Since then the lexicon is continuously undergoing refinements from various aspects, for the increase of its usefulness as a very good knowledge-base for WSD.

For any given polysemous word, WordNet stores multiple unique entries for every distinct sense of the word. The principal component of every unique entry in a WordNet is a *synset*. A *synset* is a unique list of most popularly used synonymous words for a particular sense of a polysemous word. The first synonymous word that is kept in a *synset* is usually the word itself, other synonymous words appear in the order of their frequency of usage for that sense of the polysemous word. Presently most WordNets contain sense information for only nouns, verbs, adjectives and adverbs, the four open class categories or basic parts of speech. A desirable goal of all WordNet development projects is to construct rich knowledge-bases by identifying mechanisms to capture and store sense information for polysemous words that mimic the ways that human beings employ to process and store linguistic information for concepts and words of a particular language.

At present WordNet contains 203145 entries [3].

A. Wordnet Principle

Wordnet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [3] [4].

III. MOTIVATION

In [5] Alok et. al. have presented favourable justifications for the introduction of few new fields in the basic database relation structure of WordNets. The paper emphasizes on the different qualitative aspects about the writing of *suitable glosses* for more clear-cut and rich explanation of senses of polysemous words. Justifications about the introduction of proper nouns, morphological information, distributional constraints and highly expected words for better clarification of different senses of words, have also been presented. Based on those justifications 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. By conducting experiments we have verified that the introduction of these informative fields nicely enhances the efficiency of knowledge-based contextual overlap dependent WSD algorithms. We have introduced the following fields in the new WordNet database relation structure:

i. Field to store information related to the "frequently used" or "highly expected" words for that concept or sense of a word.

ii. Field to store information related to the "distributional constraints" for that sense of the word.

During data entry in the proposed WordNet database we have ensured that we keep multiple long glosses (sentences for explanations of the sense) made up of diverse but most frequently used terms that can be used to express the proper meaning of that sense of the word.

IV. DATABASE RELATION STRUCTURE OF WORDNET

A WordNet system consists of lexicographer files, code to convert these files into a database, and search routines and interfaces that display information from the database [1]. The lexicographer files organize *nouns*, *verbs*, *adjectives* and *adverbs* into groups of synonyms, and describe relations between synonym groups [1].

In the existing database relation structure of a WordNet, the following informative fields are kept for each entry (one entry in a WordNet corresponds to one sense of a polysemous word) [5]:

i. An unsigned numeric value as sense identifier or sense ID for one sense of a word

ii. Category value i.e. *noun*, *verb, adjective* or *adverb*

iii. Gloss or explanation of the sense of the word

iv. Example sentence(s)

v. A list of synonymous words (first word in the list is the word itself)

In our WordNet database relation structure we keep two more informative fields regarding:

vi. Highly expected words

vii. Words related to distributional constraints, like storage of information regarding the relation between the senses of words like "cigarette" and "ash."

V. WSD USING LESK ALGORITHM

We have used Lesk algorithm [6] for doing WSD. The algorithm returns the sense identifiers for a word by looking up the entries corresponding to the different senses of the polysemous word in a WordNet. The working of the algorithm is presented below as a pseudocode function:

```
function SIMPLIFIED-LESK(word,sentence) returns best sense of word
    best-sense <- most frequent sense for word
    max-overlap <- 0
```

```
        context <- set of words in sentence
        for each sense in senses of word do
        signature <- set of words in the gloss and examples of sense
        overlap <- COMPUTEOVERLAP(signature,context)
        if overlap > max-overlap then
        max-overlap <- overlap
        best-sense <- sense
        end
    return (best-sense)
```

VI. EXPERIMENTs

We conducted experiments using the proposed WordNet knowledge-base for WSD of several words in different heterogeneous sentences. The results showcase the usefulness and effectiveness of the proposed enhancement of addition of informative fields.

TABLE I. Comparison of WSD results obtained for the two WordNets with sense disambiguation results obtained using Human Intelligence

Sentence	Word to disambiguate	Sense ID obtained by Lesk Algorithm using		Sense ID assigned by employing Human Intelligence
		Proposed WordNet	Existing WordNet	
They will capture his property for illegal use of it	Capture	421	422	421
Police will capture Abu Salem	Capture	422	422	422
I will capture these moments in my mind when missing something that belongs to this college	Capture	92	422	92
She captured all the men' s mind with her emotions	Capture	423	723	423
I have done good study of Computation via books	study	44	42	44
I have done good study of TOC	study	42	42	42
I have done good study of TOC subject	study	41	42	41
to get an admission into master degree become difficult to acquire	acquire	63	63	63
We have actuated the circuit by spark	actuated	327	327	327
We have actuated the circuit to process well	actuated	329	327	329
Democracy in India controls all other parties according to their members	democracy	359	359	359
Some areas in orissa have a good development in recent	development	47	47	47
My project development has shown efficiency growth in WSD	development	45	45	45
Software development is a process of achieving a task by model used	development	46	46	46
She have got a good education by qualified teachers at IIIT	education	1	1	1
Education is primary thing for a growing child	education	3	3	3
Knowledge comes by good education	education	1	5	1
He has a good teaching experience	experience	13	13	13
He has an experience of failure in exam	experience	13	13	36

Table 1 presents a comparison of WSD results obtained for the two WordNets with sense disambiguation results obtained using Human Intelligence. From the results it can be easily understood that the proposed WordNet results in better WSD.

VII. COnclusiON

In the present paper we 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.

We experimentally verified the usefulness of the proposed enhancement of addition of informative fields to the WordNet database structure. We used the Lesk Algorithm to do word sense disambiguation. Our results indicate that the WSD based on proposed Wordnet is better.

VIII. FUTURE WORK

For future research, we are focusing on further enrichment of WordNet by introducing proper nouns and morphological information related to the senses and then carry out many or all-word WSD using Lesk and Lesk-like algorithms.

References

[1] Manish Sinha, Mahesh Kumar, Prabhakar Pande, Lakshmi Kashyap and Pushpak Bhattacharyya. 2004. *Hindi Word Sense Disambiguation*. International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, November, 2004.

[2] Hindi Wordnet from Center for Indian Language Technology Solutions, IIT Bombay, Mumbai, India http://www.cfilt.iitb.ac.in/WordNet/webhwn/

[3] *WordNet: a lexical database for English Language*; Available at: http://wordnet.princeton.edu/index.shtml.

[4] Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*, MIT Press.

[5] Alok Chakrabarty, Bipul Syam Purkayastha, Lavya Gavshinde. 2010. *Ideas to Enhance Contextual Overlap for Knowledge-based Overlap Algorithms for Word Sense Disambiguation using Wordnet.* In 3rd IndoWordnet Workshop of the 8th International Conference on Natural Language Processing (ICON 2010), Kharagpur, India, December, 2010.

[6] 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.

AUTHORS PROFILE

Deepesh Kumar Kimtani received the B.Tech degree in 2006 from Uttar Pradesh Technical University, UP, India. Currently he is pursuing his M.Tech degree from Department of Computer Science and Engineering, International Institute of Informational Technology, Bhubaneswar, Odisha, India. His current research interests include Machine Learning, NLP and Theory of Computation.

E-mail: deepesh.kimtani@gmail.com

Jyotirmayee Choudhury received the B.Tech degree in 2008 from Biju Pattanaik University Of Technology, Odisha, India. Currently she is pursuing his M.Tech degree from Department of Computer Science and Engineering, International Institute of Informational Technology, Bhubaneswar, Odisha, India. Her current research interests include Data Mining,NLP and Software Engineering.

E-mail: jyotichoudhury@gmail.com

Dr. Alok Chakrabarty received the Master of Science degree in Computer Science in 2007 from Assam University, Silchar, Assam, India. Currently he is an Assistant Professor in the Department of Computer Science and Engineering of International Institute of Information Technology, Bhubaneswar, Odisha, India. His current research interests include Pattern Recognition and Machine Learning, Natural Language Processing, Wireless Sensor Networks and Data Mining.
E-mail: mcscalok@gmail.com, alok@iiit-bh.ac.in