

WordNet: A Knowledge Source for Word Sense Disambiguation

S. G. Kolte¹ and S. G. Bhirud²

¹K. J. Somaiya Polytechnic, Mumbai-77, India

Email: iamksopan@hotmail.com

²Veermata Jijabai Technological Institute, Mumbai – 19, India

Email: sgbhirud@yahoo.com

Abstract— Word Sense Disambiguation is the most critical issue in natural language processing. Although it has been addressed by many researchers, no satisfactory results are reported. In this paper we present the methodology for Word Sense Disambiguation based on domain information and WordNet hierarchy. Domain is a set of words in which there is a strong semantic relation among the words. The words in the sentence contribute to determine the domain of the sentence. The availability of WordNet domains makes the domain-oriented text analysis possible. The domain of the target word can be fixed based on the domains of the content words in the local context. This approach can be effectively used to disambiguate nouns. We present the unsupervised approach to Word Sense Disambiguation using the WordNet domains. The model determines the domain of the target word and the sense corresponding to this domain is taken as the correct sense. We have used the WordNet domains 3.1.as lexical database. The model can be further improved by using the WordNet relation links.

Index Terms— Word sense Disambiguation, Machine Readable Dictionary, WordNet, WordNet Domains, Ability link, Capability link, Function link, SemCor2.1

I. INTRODUCTION

Word Sense Disambiguation (WSD) is the process of resolving the meaning of a word unambiguously in a given natural language context. Given a polysemous word in running text, the task of WSD involves examining contextual information to determine the intended sense from a set of predetermined candidates [1]. WSD is task of classification in which the senses are the classes, the context provides the evidence and each occurrence of the word is assigned to one or more of its possible classes based on evidence [2]. The problem is so difficult that it was one of the reasons why the Machine Translation systems were abandoned. However after 1980 large-scale lexical resources and corpora became available and WSD drew attention of researchers. At present WSD is well addressed issue and has occupied important stage in the Natural Language Processing (NLP).

The sense of a word in a text depends on the context in which it is used. The context is determined by the other words in the neighborhood in the sentence. Thus if the word file, hard disk or data appears near the word virus, we can say that it is the program and not the biological virus. This is called as local context or sentential context.

Real world knowledge is necessary to build the context. Knowledge sources are thus the fundamental components for WSD systems. These knowledge sources include corpora of text, thesauri or machine readable dictionaries. One of the first attempts to use dictionary-based approach was by Lesk [3]. He devised an algorithm that chooses the appropriate sense of a polysemous word by calculating the word overlap between the context sentence of the word in question and the word's definition in a Machine Readable Dictionary (MRD).

The Lesk algorithm can be effectively used with the WordNet lexical database. Such attempt is made at IITB [4] and the results are promising. Similar experiments are made by Jonas at Lund University. Magnini [5] introduced "Word Domain Disambiguation" (WDD) in which domain label was selected instead of the sense label. This algorithm however does not account for domain variation in long texts. To improve the domain calculations contexts were introduced.

II. PREVIOUS WORK

Domain is a set of words which exhibit a strong semantic relation among themselves. Semantic domains are considered as a list of related words describing a particular subject or area of interest. A common problem of many previous attempts to utilize semantic domains in WSD is that very frequent words have, in general, many senses belonging to different domains. The first attempt to introduce domains in WSD was by Cowie [6]. The use of WordNet domains in WSD is being tried at ITC-irst. The basic idea used is that the disambiguation of a word in its context is mainly the process of comparison between the domain of context and domains of word's senses. This system achieved the highest scores (precision) of 0.75 and 0.65 for all_words and lexical_sample respectively at the SENSEVAL-2 [7].

WordNet hyponymy/hypernymy relations are therefore used to improve word sense disambiguation. Resnik [8] used semantic similarity between two words and disambiguated noun senses. Other approaches used WordNet taxonomy. Lee [9] and Leacock [10] proposed a measure of the semantic similarity by calculating the length of the path between the two nodes in the hierarchy. Agirre [2] proposed a method based on the conceptual distance among the concepts in the hierarchy and provided and presented experimental results comparing

the above systems in a real-word spelling correction system. Voorhess [11] handled the problem of the lack of “containment” of clear divisions in the WordNet hierarchy and defined some categories. Sussna [12] used a disambiguation procedure based on the use of a semantic distance between topics in WordNet. Fragos [13] reported an improvement in accuracy by using the additional definitions based on hyponymy/hypernymy relations and accuracy of 49.95% was estimated.

III WORDNET

WordNet is a semantic lexicon for the English language. WordNet was created and is being maintained at the Cognitive Science Laboratory of Princeton University [14]. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. WordNet is organized as a network of lexicalized concepts, called synsets, that comprise set of synonyms. Each word meaning can be represented by a set of word-forms known as synonym sets or synsets. Synsets are created for content words, i.e., for Noun, Verb, Adjective and Adverb. For example the nouns {president, chairman, chair, chairperson } form a synset. WordNet Domains is an extension of WordNet where synonym set have been annotated with one or more subject domain labels [15]. The domain set used in WORDNET DOMAINS has been extracted from the Dewey Decimal Classification [16] and a mapping between the two taxonomies has been computed in order to ensure completeness.

IV. OUR APPROACH TO WSD

We describe here how the WordNet domains can be used for WSD. Since context is the only means to identify the meaning of polysemous word, we propose the context based approach. In this approach, context is considered as text consisting of content words in some small window surrounding to the target word. Such a context is called as *local context*. To disambiguate a word, three types of bags are used: The algorithm tags the text with part of speech tags using pos tagger. Bag b1 contains the content words. The domains for each word from b1 relating to pos tag sense are selected and inserted into bag b2. The domains corresponding to pos tag of

target word are inserted into bag b3. For each domain in b3, the domains of the other words are compared (domain factotum can match with any domain). The domain of target word which maximises the match with the domains of other words becomes the domain of the text. The sense belonging to this domain is the correct sense of the target word.

Let us assume that $(w_1, w_2, w_3...w_n)$ is the bag *b1* containing pos tagged content words. And b2 is the bag containing $(d_1, d_2, d_3...d_n)$, sets of all domains corresponding to the content words w.r.t their pos tags. Each set di contains all possible domains corresponding to the pos tag sense. The bag b3 contains the domains of target word.

A. Algorithm

1. Input the sentence
2. Perform POS Tagging.
3. From the POS tagged text separate the content words and insert into bag b1.
4. for each content word, insert set of domains corresponding to it’s pos tag into bag b2.
5. for the target word wt, insert domains corresponding to it’s pos tag into bag b3
6. compare each domain in b3 with set of domains of remaining content words.
7. The domain (except factotum) in b3 which maximises with domains of other content words is the domain of the text.
8. The sense belonging to domain obtained from step 7 is the correct sense.

B. Example

Let us consider the sentence

The *virus* infected all files on the hard disk.

This sentence is passed to the POS Tagger. The output is the tagged text as follows

The/DT virus/NN infected/VBD all/DT files/NNS on/IN hard_disk/NN ./.

Out of the above the system selects only the content words viz. virus, infected, files and disk.

Suppose we want to find the correct sense of the word *virus* in the above sentence.

The sense numbers and domains of target word and that of content words are shown in Fig. 1.

<table border="1" style="margin: auto;"> <tr><td>Virus (noun sense)</td></tr> <tr><td>Infected (verb sense)</td></tr> <tr><td>Files (noun sense)</td></tr> <tr><td>disk (noun sense)</td></tr> </table> <p>b1</p>	Virus (noun sense)	Infected (verb sense)	Files (noun sense)	disk (noun sense)	<table border="1" style="width: 100%;"> <tr> <td>virus (noun sense) Target word</td> <td>infected (verb sense)</td> <td>files (noun sense)</td> <td>hard_disk (noun sense)</td> </tr> <tr> <td>01254816- factotum</td> <td>00087224- medicine</td> <td>06106818- telecommunications</td> <td>03364489- computer science</td> </tr> <tr> <td>13209397- factotum</td> <td>00086241- medicine</td> <td>07917489- factotum</td> <td></td> </tr> <tr> <td>06179311- Computer_science</td> <td>02503346- factotum</td> <td>03215630- administration furniture</td> <td></td> </tr> <tr> <td></td> <td>00585683- psychological features</td> <td>03215329- building industry</td> <td></td> </tr> </table> <p style="text-align: center;">b2</p>	virus (noun sense) Target word	infected (verb sense)	files (noun sense)	hard_disk (noun sense)	01254816- factotum	00087224- medicine	06106818- telecommunications	03364489- computer science	13209397- factotum	00086241- medicine	07917489- factotum		06179311- Computer_science	02503346- factotum	03215630- administration furniture			00585683- psychological features	03215329- building industry		<table border="1" style="margin: auto;"> <tr><td>01254816- factotum</td></tr> <tr><td>13209397- factotum</td></tr> <tr><td>06179311- Computer_science</td></tr> </table> <p>b3</p>	01254816- factotum	13209397- factotum	06179311- Computer_science
	Virus (noun sense)																												
	Infected (verb sense)																												
	Files (noun sense)																												
disk (noun sense)																													
virus (noun sense) Target word	infected (verb sense)	files (noun sense)	hard_disk (noun sense)																										
01254816- factotum	00087224- medicine	06106818- telecommunications	03364489- computer science																										
13209397- factotum	00086241- medicine	07917489- factotum																											
06179311- Computer_science	02503346- factotum	03215630- administration furniture																											
	00585683- psychological features	03215329- building industry																											
01254816- factotum																													
13209397- factotum																													
06179311- Computer_science																													

Figure 1. Contents of bag b1, b2 and b3.

In the given example

b1= {virus, infected, files, disk}
 b2= {{factotum, factotum, computer science},{medicine, medicine, factotum, psychological features}, {telecommunication, factotum, administration, building industry}, {computer science}}
 wt= virus
 b3= {factotum, factotum, computer science}

Since the domain computer-science from bag b3 has maximum score (factotum, factotum, computer science), we fix computer-science as the domain of the text. The noun sense belonging to computer science domain is the selected as the correct sense.

V. EXPERIMENTAL SETUP

We used Stanford POS Tagger to tag the tokens. The output of this tagger is in the following form.

The/DT virus/NN infected/VBD all/DT files/NNS on/IN the/DT hard_disk/NN ./.

From these only the content words were selected. The domains corresponding to these words were obtained from the file wn-domains-3.2. The sense corresponding to the domain was obtained from the WordNet 2.1 database.

VI. WSD USING WORDNET HIERARCHY

It has been observed that most of the sunsets (36820) in WordNet belong to the domain factotum and do not contribute for any domain information. Further a word may have multiple senses for a particular domain. For example the word bank has sense#1, sense#6 and sense#8 belonging to the domain economy.

Thus, all methods based on simple frequency counting often turn out to be inadequate. The drawback of the algorithm is that it can disambiguate a word provided it has only one sense per domain.

WordNet hyponymy/hypernymy relations are therefore used to improve word sense disambiguation. WordNet is organized by semantic relations. Since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets. Two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made. The antonym of a word x is sometimes not-x, but not always. For example, rich and poor are antonyms. Antonymy is a lexical relation between word forms, not a semantic relation between word meanings. Hyponymy/hypernymy is a semantic relation between word meanings: e.g., {maple} is a hyponym of {tree}, and {tree} is a hyponym of {plant}. A concept represented by the synset {x, x', . . .} is a meronym of a concept represented by the synset {y, y', . . .} if sentences constructed from such frames as y has an x (as a part) or an x is a part of y. The power of WordNet lies in its set of domain-independent lexical relations. Table I shows a subset of relations associated with each of the three databases.

TABLE I. RELATIONS USED IN WORDNET

Relation	Definition	Example
Hypernym	From concepts to superordinates	Breakfast meal
Hyponym	From concepts to subtypes	Meal lunch
Has-Member	From groups to their members	Faculty professor
Has-Part	From wholes to parts	Table leg
Part- Of	From parts to wholes	Course meal
Antonym	Opposites	Leader follower

VII. METHODOLOGY USED FOR WSD

We describe here the technique used to assign the correct sense using the links in WordNet hierarchy. The task is to find the meaning of noun in the given verb context from all candidate word senses in WordNet. Suppose a noun Wn has n word senses in WordNet. In the given sentence we are going to decide the intended meaning of a noun Wn in verb context Wv. In addition to existing relations we investigated the use of additional links like ability, capability and function.

A. Using Hypernym (is_a_kind) Relationship

Consider the sentence

He ate many dates.

This sentence is processed using the POS tagger, the output is

He/ PRP ate/VBD many/ DT dates/ NNS Fp 1

Here the word date has 8 senses with is_a_kind of relations as follows

1. (503) date, day of the month
2. (119) date -- (a particular day specified as the time something happens)
3. (104) date, appointment, engagement
4. (55) date, particular date
5. (37) date -- (the present; "they are up to date")
6. (29) date, escort -- (a participant in a date)
7. (26) date -- (the particular day, month, or year)
8. (20) date -- (sweet edible fruit)

Since date is a kind of edible fruit, the intended sense is sense #8

B. Using Meronymy and Holonymy (Part-whole) Relationship

Consider the sentence

The trunk is the main structural member of a tree that supports the branches.

The output of POS tagger is

The/DT trunk/NN is/VBZ the/DT main/JJ structural_member/NN of/IN a/Z tree/NN that/WDT supports/ VBZ the/DT branches/NNS . . Fp 1

The word trunk has following senses

1. PART OF: {12934526} <noun.plant> tree#1
3. PART OF: {05154650} <noun.body> body#1,
4. PART OF: {02929975} <noun.artifact> car#1,
5. PART OF: {02480939} <noun.animal> elephant#1

- 6. PART OF: {02482174} <noun.animal>
mammoth#1

Since the context contains the noun tree; the algorithms detects the sense #1 as correct sense.

C. Using The Ability Link

This link specifies the inherited features of a nominal concept. This is a semantic relation.

Consider the sentence

A crane was flying across the river.

The output of POS tagger is

A/Z crane/NN was/VBD flying/VBG across/ IN the/
DT river/NN. . Fp 1

In the above sentence the word crane has following noun senses.

1. Crane, Stephen Crane -- (United States writer (1871-1900))
2. Crane, Hart Crane, Harold Hart Crane -- (United States poet (1899-1932))
3. Grus, Crane -- (a small constellation in the southern hemisphere near Phoenix)
4. crane -- (lifts and moves heavy objects)
5. crane -- (large long-necked wading bird)

The intended sense of the word crane is sense #5 in the context of verb flying.

D. Using The Capability Link

This link specifies the acquired features of a nominal concept. This is a semantic relation.

For example consider the sentence.

The chair asked members about their progress.

After pos tagging we get

The/DT chair/NN asked/VBD members/NNS
about/IN their/PRP\$ progress/NN. . Fp

The word chair has 4 noun senses.

1. (35) chair -- (a seat for one person)
2. (2) professorship, chair
3. president, chairman, chairwoman, chair, chairperson
4. electric chair, chair, death chair, hot seat

Since a person has capability to ask, we can say that the intended sense is sense #3.

E. Using the Function link

Consider the sentence

Please keep the papers in the file.

This is pos tagged as

Please/UH keep/VB the/DT papers/NNS in/IN the/DT
file/NN. . Fp 1

The noun file has 4 senses as shown below

1. (17) file, data file
2. (1) file, single file, Indian file
3. (1) file, file cabinet, filing cabinet
4. (1) file -- (a steel hand tool)

A careful observation shows that the sense #3 is the correct sense in the given context. The algorithm for the creation of the bags is as follows.

VIII. EVALUATION

Our disambiguation method was evaluated using the Semcor2.1 files [17]. The Semcor2.1 files are manually disambiguated text corpora using senses of WordNet 2.1

The evaluation procedure contains following steps

- Get the i^{th} unannotated sentence of k file of Semcor2.1
- Begin disambiguation of sentence i.
- Compare system output with the i^{th} annotated sentence of k file.
- Repeat the above steps for:
 $i = 1 \dots \text{NumberOfSentences}(\text{File}_k), k = 1 \dots 10$

These files contained 5463 nouns. Out of which we could disambiguate 5236 nouns. The accuracy of our approach was 63.92%, which means that our system disambiguated correctly 3492 out of 5463 nouns. Table II shows the results of our system and Fig. 2 is the histogram showing the accuracy across the Brown corpus

TABLE II.
RESULTS FROM THE FIRST 10 FILES OF BROWN I CORPUS

File	#Nouns	#Disamb- iguated	#Correctly Disamb- iguated	Accuracy (%)
br-a01	573	550	395	71.81
br-a02	611	600	367	61.16
br-a11	582	550	346	62.90
br-a12	570	555	332	59.81
br-a13	575	545	339	62.20
br-a14	542	526	268	50.95
br-a15	535	519	326	62.81
br-b13	505	482	254	52.69
br-b20	458	422	273	64.69
br-c01	512	487	285	58.52
Total	5463	5236	3185	60.82

files.

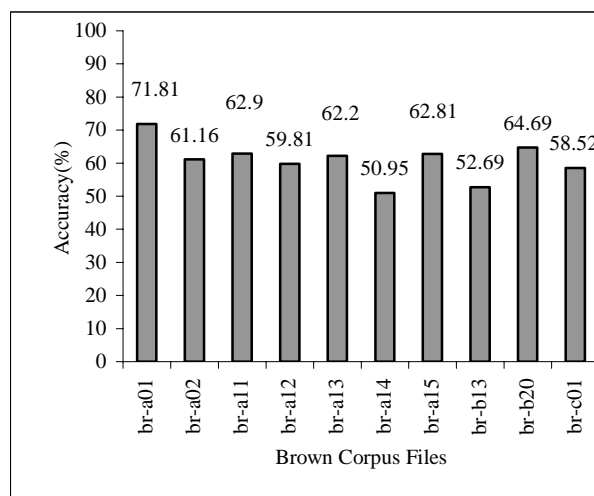


Figure 2. Histogram showing accuracy.

IX. CONCLUSION AND FUTURE WORK

We have tried to use existing relations available in WordNet hierarchy. The additional links available in Hindi WordNet motivated us to check possibilities of use of these links for word sense disambiguation. Although these links are not directly available in WordNet 2.1, we have added these links manually in the WordNet database tables available in MySQL format for limited words. To our knowledge no research work is sighted making the use of these additional links. The accuracy of WSD system highly depends on the POS tagger module. The efficiency of our system is limited due to the fact that it can not pos tag some words correctly. For example

Pycnogenol, a bark extract from the French maritime pine tree, reduces jet lag in passengers by nearly 50%, a study suggests.

When the above sentence is tagged, it gives the output as

Pycnogenol/NNP, , Fc a/Z bark/VB extract/NN from/IN the/DT French/JJ maritime/NN pine_tree/NN , , Fc reduces/VBZ jet_lag/NN in/IN passengers/NNS by/IN nearly/RB 50_% /Zp , , Fc a /Z study/NN suggests/VBZ . . Fp 1

Here we can notice that the word bark has been incorrectly tagged as VB(verb). Thus the accuracy of our model largely depends upon the accuracy of the POS tagger. We need to improve the POS tagger which can correctly POS tag the content words.

REFERENCES

- [1] Chen, J. and J. Chang 1998. Topical Clustering of MRD Senses Based on Information Retrieval techniques. Computational Linguistics. MIT Press, Cambridge, MA. Vol.24(1), pp. 61-95.
- [2] E. Agirre and G. raigu. 1996. Word Sense Disambiguation using Conceptual Density. In Proceeding of COLLING, pages 16-22.
- [3] M. Lesk, Vocabulary problems in retrieval systems, in Proc. 4th Annual Conference of the University of Waterloo Centre for the New OED. 1988.
- [4] 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.
- [5] B. Magnini and G. Cavagli_a. 2000. Integrating subject field codes into WordNet. In Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, June, 2000, pp. 1413-1418.
- [6] Cowie et al., 1992 J. Cowie, J. Guthrie, and L. Guthrie. Lexical disambiguation using simulated annealing. In Proc. of COLING-92, pages 359-365, Nantes, France, 1992.
- [7] Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A., July 2001. Using domain information for word sense disambiguation. In: Proceedings of SENSEVAL-2 Second

- International Workshop on Evaluating Word Sense Disambiguation System. Toulouse, France, pp. 111-114.
- [8] Resnik P 1995. Disambiguating Noun Groupings with Respect to WordNet Senses. Proceedings 3rd Workshop on Very Large Corpora. Cambridge, MA 54-68.
 - [9] Lee J. H., Kim H. and Lee Y. J. 1993. Information retrieval based on conceptual distance in IS-A hierarchies. In Journal of Documentation, 49(2). pp. 188-207.
 - [10] Leacock, C. and Chodorow, M. 1998. Combining Local Context and WordNet Similarity for Word Sense Disambiguation. In WordNet: An Electronic Lexical Database, MIT Press, Cambridge MA. pp. 265-283.
 - [11] Voorhees, Ellen. M. 1993. "Using WordNet to Disambiguate Word Senses for Text Retrieval", In Proceedings of SIGIR'93
 - [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 (CIKM-93), Arlington, Virginia
 - [13] Fragos, K., Maistros, Y., and Skourlas., C. 2003 Wordsense disambiguation using WordNet relations. FirstBalkan Conference in Informatics, Thessaloniki
 - [14] Fellbaum, C., WordNet. An Electronic Lexical Database. MIT Press. 1998.
 - [15] B. Magnini and G. Cavagli_a. 2000. *Integrating subject field codes into WordNet*. In Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, June, 2000, pp. 1413-1418.
 - [16] J. P. Comaroni, J. Beall, W. E. Matthews, and G. R. New, editors. 1989. Dewey Decimal Classification and Relative Index. Forest Press, Albany, New York, 20th edition.
 - [17] Landes S., Leacock C. and Tengi R. 1998. Building Semantic Concordances. In WordNet: an Electronic Lexical Database, MIT Press, Cambridge MA pp.199-216