Word Sense Disambiguation using WordNet Relations and Parallel Corpora

[1]S. G. Kolte, [2]S. G. Bhirud

[1]Bharati Vidyapeeth Deemed University, College of Engineering,
Pune-43, India
Veermata Jijabai Technological Institute, Matunga,
Mumbai – 400 019, India
iamksopan@hotmail.com, sgbhirud@yahoo.com

Abstract: *Word Sense Disambiguation is one of the most important challenge in computational linguistics. It is often described as "AI-complete" and is the most critical issue in natural language processing. Although it has been addressed by many researchers, no satisfactory results are reported. Rule based systems alone cannot handle this issue due to ambiguous nature of the natural language. Knowledge-based systems are therefore essential to find the intended sense of a word form. Machine readable dictionaries have been widely used in word sense disambiguation. The problem with this approach is that the dictionary entries for the target words are* *very short. WordNet is the most developed and widely used lexical database for English. The entries are always updated and many tools are available to access the database on all sorts of platforms. The WordNet database can be conveted in MySQL format and we have modified it as per our requirement. Sense's definitions of the specific word, "Synset" definitions, the "Hypernymy" relation, and definitions of the context features (words in the same sentence) are retrieved from the WordNet database and used as an input of our Disambiguation algorithm.*

Keywords: Word sense Disambiguation, Machine Readable Dictionary, WordNet, Ability link, Capability link, Function link, SemCor

1. 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. 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. From the pre-processed documents five words preceding to the word to be disambiguated and five words following it were extracted. These words included nouns, verbs or adjectives. Every sense of the word to be disambiguated was compared to each sense of surrounding words. Each combination was assigned a score which is based on number of overlapping words. This approach however suffers from the fact that large number of fine senses in WordNet is not distinguishable from each other. In this paper, we propose an improvement of the Lesk' s method. We try to improve the performance of the disambiguation task by using additional definitions based on the "Hypernymy / Hyponymy" relation to enrich further the bags of words. Only the definitions related to the "hypernyms" of the nouns and verbs found in the context words and the senses' definitions were used. In this paper we explain how the additional link can be used for WSD. To our knowledge no research work on use of these additional links is published.

2. Previous Work

Disambiguation methods based on WordNet definitions are poor performers. WordNet hyponymy/hypernymy relations are therefore used to improve word sense disambiguation. Resnik[5] used semantic similarity between two words and disambiguated noun senses. Other approaches used WordNet taxonomy. Lee et al. [6] and Leacock and Chodorow [7] proposed a measure of the semantic similarity by calculating the length of the path between the two nodes in the hierarchy. Agirre and Rigau [8] 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[9] handled the problem of the lack of "containment" of clear divisions in the WordNet hierarchy and defined some categories. Sussna [10] used a disambiguation procedure based on the use of a semantic distance between topics in WordNet.

Kostos Fragos et.al.[11] reported an improvement in accuracy by using the additional definitions based on hyponymy/hypernymy relations and accuracy of 49.95% was estimated. They used WordNet glosses. After preprocessing features were extracted and enclosed in bags. Each word within the bag was assigned a weight, depending on the depth of its related synset's position in the WordNet taxonomy. Bags of words related either with a sense or the context were prepared. A feature could be inserted only once into the same bag. To disambiguate a word, two types of bags were used: A bag of words related to every sense of the word and a bag of words related to the context. The Lesk's approach is based on the count of the common words between the bags related to each sense and the bag related to the context. The sense having the maximum overlapping with the context bag is chosen as the correct one.

3. The WordNet Taxonomy

The WordNet [12] is an electronic lexical database created at Princeton University in 1990. The WordNet organizes the lexical information in meanings (senses) and synsets (set of words – sentence(s) - describing the meaning of the word in a specific context). What makes WordNet remarkable is the existence of various relations between the word forms (e.g. lexical relations, like synonymy and antonymy) and the synsets (meaning to meaning or semantic relations e.g. hyponymy/hypernymy relation, meronymy relation).

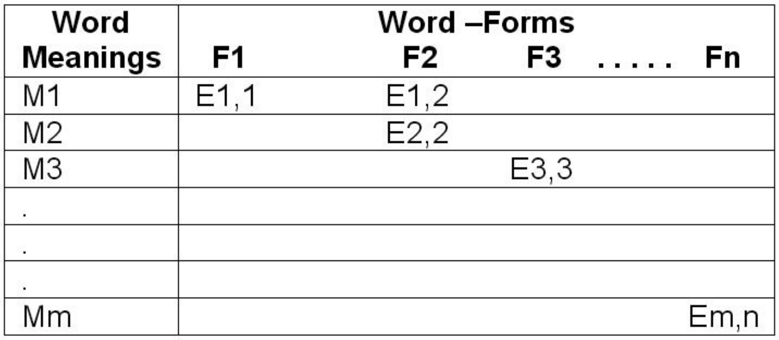
3.1 WordNet Database

For each syntactic category, two files represent the WordNet database : index.pos and data.pos, where pos is either noun, verb, adj or adv. The database is in an ASCII format that is human- and machine-readable, and is easily accessible to those who wish to use it with their own applications. The index and data files are interrelated. The WordNet morphological processing functions, morphy(), handle a wide range of morphological transformations.

3.2 Lexical Matrix

Table1 is called as Lexical Matrix- is an abstract representation of the organization of lexical information. Word–forms are imagined to be listed as headings for the columns and word meanings as headings for the rows. Rows express synonymy while columns express polysemy. Word forms are imagined to be listed as headings for the columns; word meanings as headings for the rows. An entry in a cell of the matrix implies that the form in that column can be used (in an appropriate context) to express the meaning in that row. Thus, entry E1,1 implies that word form F1 can be used to express word meaning M1. If there are two entries in the same column, the word form is polysemous; if there are two entries in the same row, the two word forms are synonyms (relative to a context).

Table 1. Illustrating the Concept of a Lexical Matrix



Word	Word –Forms				
Meanings	F1	F2	F3	. . . . .	Fn
M1	E1,1	E1,2			
M2		E2,2			
M3			E3,3		
.					
.					
.					
Mm					Em,n

A lexical matrix can be represented for theoretical purposes by a mapping between written words and synsets. Synonymy is, a lexical relation between word forms, but because it is assigned this central role in WordNet, a notational distinction is made between words related by synonymy, which are enclosed in curly brackets, '{' and '}', and other lexical relations, which will be enclosed in square brackets, '[' and ']'. Semantic relations are indicated by pointers.

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. An important class of lexical relations are the morphological relations between word forms. WordNet is a lexical inheritance system; a systematic effort has been made to connect hyponyms with their superordinates.

The power of WordNet lies in its set of domain-independent lexical relations. Table 2 shows a subset of relations associated with each of the three databases. Figure1 shows how synsets are linked to form a semantic network.

Table 2. Noun Relations 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

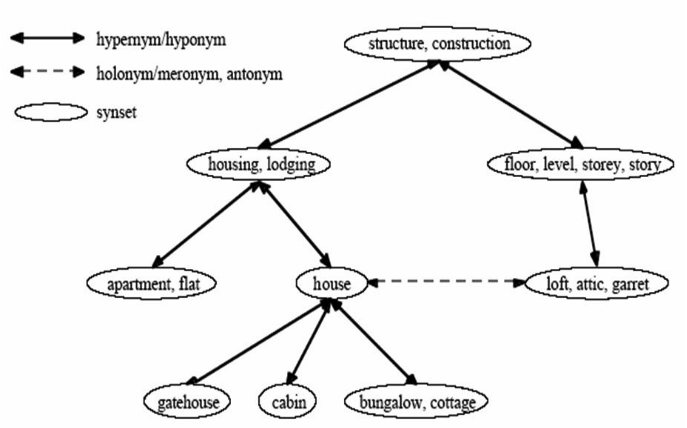


Figure 1. Linking of Synsets

4. Methodology: Our Approach to 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 W_n has n word senses in WordNet. In the given sentence we are going to decide the intended meaning of a noun W_n in verb context W_v. In addition to existing relations we investigated the use of additional links like ability, capability and function.

4.1 Using hypernym (is kind of) relationship

Consider the sentence:

He ate many dates.

This sentence is processed using the pos tagger and the output is

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

Here the word date has 8 senses with is-a-kind of relations like day, day, meeting, point, time, companion, calendar day, edible fruit.

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

4.2 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 4 is a part of relations viz. tree, body, luggage compartment, elephant.

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

4.3 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 Sense #4. crane — (lifts and moves heavy objects; lifting tackle is suspended from a pivoted boom that rotates around a vertical axis)

Sense #5. crane — (large long-necked wading bird of marshes and plains in many parts of the world)

Further the hypernym relationships for the two senses are

sense #4 lifting device — (a device for lifting heavy loads)

sense #5 bird — (warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)

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

4.4 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 viz. seat, position, person, instrument of execution. Since a person has capability to ask, we can say that the intended sense is sense #3.

4.5 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 — (a set of related records (either written or electronic) kept together)

2. (1) file, single file, Indian file — (a line of persons or things ranged one behind the other)
3. (1) file, file cabinet, filing cabinet — (office furniture consisting of a container for keeping papers in order)
4. (1) file — (a steel hand tool with small sharp teeth on some or all of its surfaces; used for smoothing wood or metal)

A careful observation shows that as the sense #3 is the correct sense in the given context.

WSD Algorithm: Finding the word's Correct Sense

Input:A simple sentence S, the target word wt

Output: Correct sense number of the target word(noun) from WordNet 2.1

1. For a polysemous word(target word) w_t needing disambiguation, a set of context words i.e. the content words from the sentence is collected. Let this collection be b_c.
2. The bag b_t contains noun senses with corrosponding
   (I) Synonyms
   (II) Glosses
   (III) Example Sentences
   (IV) Hypernyms
   (V) Glosses of Hypernyms
   (VII) Hyponyms
   (VIII) Glosses of Hypernyms
   (IX) Example Sentences of Hypernyms
   (X) Meronyms
   (XI) Glosses of Meronyms
   (XII) Example Sentences of Meronyms
   (XIII) Ablilty link
   (XIV) Capability link
   (XV) Function link

3. Measure the *overlap* between b_c and b_t using the relation and additional link and he intersection similarity measure.

4. Output the sense number of the target word as the most probable sense which has the *maximum overlap*.

4.6 **Use parallel Corpora for WSD**

The knowledge useful for WSD can be learned from corpora. However, supervised learning methods suffer from the high cost of manually tagging the sense onto each instance of a polysemous word in a training corpus. Bilingual parallel corpora, in which the senses of words in the text of one language are indicated by their counterparts in the text of another language, have also been used in order to avoid manually sense-tagging training data[13].We propose an unsupervised method for word sense disambiguation using a bilingual comparable corpus (English-Hindi). First, we can extract statistically significant pairs of related words from the corpus of each language. Then, aligning pairs of related words translingually, we can calculate the correlation between the senses of a first-language polysemous word and the words related to the polysemous word, which can be regarded as clues for determining the most suitable sense. Finally, for each instance of the polysemous word, the system can select the sense that maximizes the score, i.e., the sum of the correlations between each sense and the clues appearing in the context of the instance. The problem we can see is the availability of parallel corpora. However we can create parallel corpora out of the web online [14]. This involves the following three steps

- locating domains, sites or pages that might have parallel translations
- generation of candidate pairs from such data
- filtering candidate pairs with structural or content-based criteria.

Alternatively the Wikipedia database is also available for download, which can be used as knowledge-base.

5. Evaluation

Our disambiguation method was evaluated using the Semcor2.1 files [15]. The Semcor files are manually disambiguated text corpora using senses of WordNet 2.1

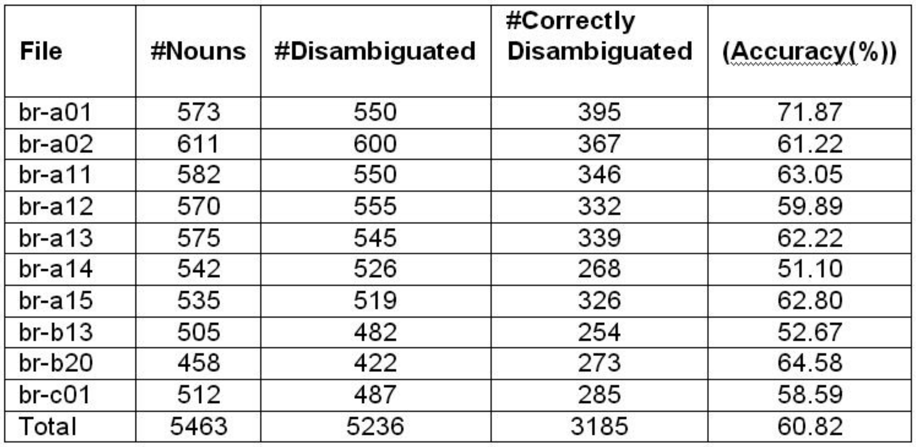
The evaluation procedure contains following steps

- Get the i unannotated sentence of k file of Semcor 2.1
- Begin disambiguation of sentence i.
- Compare system output with the Semcor 2.1 i annotated sentence of k file.
- Repeat the above steps for:

  i = 1 . . .NumberOfSentences(File k), k = 1 . . . 10

These files contained 5463 nouns. Out of which we could disambiguate 5236 nouns. The accuracy of our approach was 60.82%, which means that our system disambiguated correctly 3185 out of 5236 nouns. Table 3 shows the results of our system and Figure 2 shows the histogram of performance of our system for the ten brown corpus files.

Table 3. Results from the first 10 files of Brown 1 Corpus



File	#Nouns	#Disambiguated	#Correctly Disambiguated	(Accuracy(%))
br-a01	573	550	395	71.87
br-a02	611	600	367	61.22
br-a11	582	550	346	63.05
br-a12	570	555	332	59.89
br-a13	575	545	339	62.22
br-a14	542	526	268	51.10
br-a15	535	519	326	62.80
br-b13	505	482	254	52.67
br-b20	458	422	273	64.58
br-c01	512	487	285	58.59
Total	5463	5236	3185	60.82

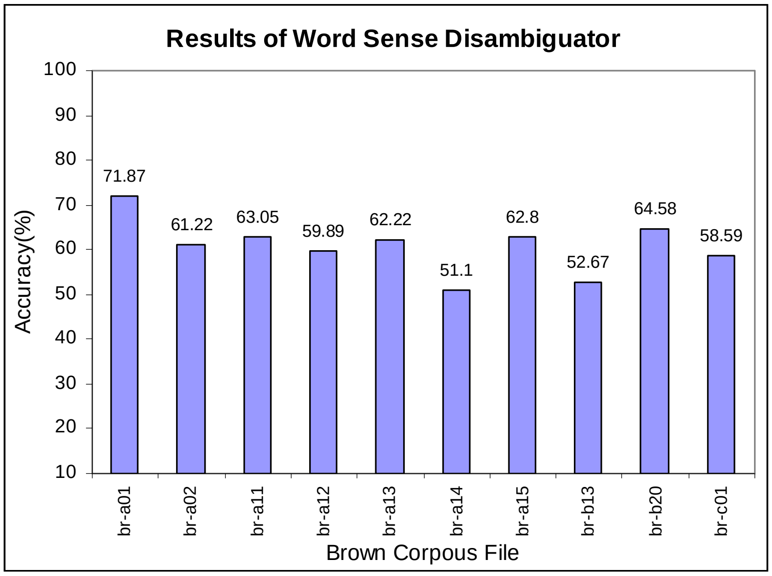


Figure 2. Results of Word Sense Disambiguator

6. Conclusion

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 when the sentence

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

when tagged, 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_% 50/0 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).**

The WSD system we have developed can be used in many NLP applications such as information retrieval, machine translation etc.

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. Resnik P.: Disambiguating Noun Groupings with Respect to WordNet Senses. *Proceedings 3rd Workshop on Very Large Corpora.* Cambridge, MA (1995) 54-68

6. Lee et al. 93] 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): 188-207

7. Leacock C., Chodorow M. (1998). Combining Local Context and WordNet5 Similarity for Word Sense Disambiguation. In *Wordnet: An Electronic Lexical Database,* pages 265-283. MIT Press, Cambridge MA, 1998.

8. Agirre E., Rigau G.: Word Sense Disambiguation Using Conceptual Density. *Proceedings16th International Conference on COLING*. Copenhagen (1996)

9. Voorhees, E.: Using WordNet to Disambiguate Word Senses for Text Retrieval. *Proceedings ACM SIGIR Conference*, (1993)

10. Sussna M.: Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. *Proceedings 2nd International Conference on Information and Knowledge Management(CIKM).* Arlington, Virginia, (1993)

11. Fragos, K., Maistros, Y., & Skourlas., C. (2003). Word sense disambiguation using WordNet relations. FirstBalkan *conference in Informatics*, Thessaloniki.

12. Fellbaum, C., *WordNet. An Electronic Lexical Database.* MIT Press. 1998.

13. Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the ACL,* pages 264-270.

14. Resnik, P. and Smith, N. (2002). The Web as a Parallel Corpus, University of Maryland technical report UMIACS-TR-2002.

15. Landes S., Leacock C., Tengi R.: Building Semantic Concordances. In: *WordNet, an Electronic Lexical Database,* MIT Press, Cambridge MA (1998) 199-216

* * * * *