Word Sense Disambiguation

K.R. Chowdhary, Associate Professor
Department of Computer Science and Engineering, Faculty of Engineering,
J.N.V. University, Jodhpur (INDIA).

Abstract

Word sense ambiguity is serious problem in many systems that deal with natural language texts. The systems like Information Retrieval (IR) and Information Extraction (IE) return the segment of the information relevant to the user's information need expressed by a query. Due to polysemy and synonymy, these systems return a portion of non-relevant information, and some times even important information is missed from retrieval. This paper presents a novel approach using Wordnet, which is improvement over the existing methods, and at the same time simple and more efficient. The results have been shown using illustrative examples.

Key words: Information Retrieval, Information Extraction, Synonymy, Polysemy, Disambiguation, Word context.

1. Introduction

The word-sense ambiguity due to *polysemy* is a major barrier for many systems that accept Natural Language (NL) input **[1]**. Due to this, a language translating system may translate two different senses of an English word into very different words in another language. Therefore, systems for machine translation must be able to determine the sense, which the author had in mind. In IR and IE, a query intended to elicit material relevant to one sense of a polysemous word may elicit unwanted material relevant to other senses of that word. For example, in computer-assisted instruction, a student asking the meaning of a word in a sentence should be supplied, it's meaning in the context of the sentence, and not a list of senses from which to pick the right one. Choosing among the alternative senses of a polysemous word is a matter of distinguishing between different sets of linguistic contexts in which the word form can be used to express the word sense. Human are quite skillful in making such distinctions. For instance, in the sentence "he nailed the boards across the windows", we do not notice that the words "board" and "nailed" are polysemous. Similarly, the queries

(i) Bat in his hands flies high,

(ii) Crane is in the field

are ambiguous. We as human we are able to resolve the ambiguities in these. However, to a machine in the first case it is not clear whether the *bat* stands for an instrument for playing a game or for the special kind of a bird. Similarly, in the second example it is not clear whether *crane* stands for a lifting machine or a bird with long neck. How human mind makes such distinctions is not very clear yet.

Different works in Word Sense Disambiguation are due to following. *Sanderson and Rijsbergen* **[2]**, who use artificially ambiguous words called pseudowords; *Krovetz* and *Croft* **[3]** attempt to resolve the lexical ambiguity using Longman Dictionary of Contemporary English (LDOCE); *Rilof* and *Lehnert* **[4]** have used training corpus for disambiguation in the application of automatic text classification; *Voorhees* **[5]** has used WordNet for disambiguation in the text retrieval applications making use of stem vectors; and *Roth* **[6]** has used statistics based

machine learning approaches for disambiguation. Further, Crovetz *and Croft* in [3] reports that "...little quantitative information is available about the extent of the problem or about the impact that the disambiguation has on information retrieval systems".

An algorithm for word sense identification must distinguish among the sets of linguistic contexts, raising the question of how much context is required. There are number of ways to define the linguistic contexts. In the method presented here, sentential context has been used. As per this, two words co-occur in the same sentence if their contexts are same. Therefore, sense identification is a matter of disambiguating among the sets of sentential contexts. Valid sense of an ambiguous word is found based on closeness of its sense to sentence context. Successful disambiguation of the words, like *crane* and *bat* would resolve the problem of retrieving the non-relevant documents, thus raising the *precision* of IR.

There are three basic approaches to Word Sense Disambiguation (WSD) [7]:

1. The WSD based on the synonyms information provided by the machine-readable dictionaries (MRDs) and thesaurus [8]
2. The WSD program learns the necessary disambiguation knowledge from a large sense-tagged corpus, in which word occurrences have been tagged manually with senses from some wide coverage dictionary, such as LDOCE or WordNet. After learning on sense tagged corpus, in which all the occurrences of a word has been correctly tagged, the WSD program assigns the correct sense to the word. This technique is called as *supervised learning*.
3. WSD uses the information gathered from raw corpora. This technique is called unsupervised *learning*.

The method presented in this paper uses the modified approach of first method above. The method is based on the following hypothesis: Representation of word senses in Network form shows the association, and thus the relations between different words. This can help in fast and easy navigation through the word senses, and can find the relation between different words through the *transitive* and *asymmetric relations*, thus resolving the ambiguity between the words.

A semantic network is one such representation technique for word relations [9]. The next sections present the methodology, illustrative example and their results.

2. Semantic Networks

Semantic networks are used to represent a graphical relationship between categories of objects. A semantic network consists: (i) *nodes*, denoting objects, (2) *links*, denoting the relations between the objects, and (3) *link labels*, which denote the particular relations. From the semantic perspective, the meaning of nodes and links depends on the application. The relations among the objects in this network are specified with the help of operators: *subset*, *member*, and *properties*. Following is a typical case of some objects and relations between them, which is graphically represented in **figure 1** using a semantic network.

$Mammals \subset Animals$
$Birds \subset Animals$
$Cat \subset Mammals$
$Bat \subset Mammals$
$Penguins \subset Birds$
$Mammals\ Has\_Legs\ 4$
$Birds\ Has\_Property\ Flies$
$Cheetah \in Cat$
$Pat \in Bat$
$Opus \in Penguin$

The network form of representations is helpful to deduce the relationship from one object to an another. For example,

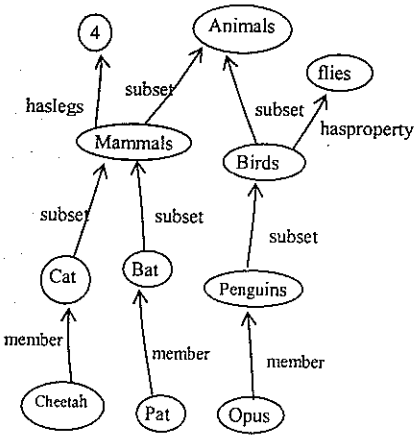

Figure 1. A Semantic Network

though it is explicitly not specified in the Knowledge-Base (KB) used for construction of this semantic network, still it can be deduced that "cat has 4 legs." In such a network, if there is any path leading from one object w_i to w_j via any number of object nodes, then w_i and w_j are related, otherwise not. The nodes in a continuous link form a common context, helping in resolving the ambiguity.

Due to the relational and inheritance properties of semantic networks, the lexical and semantics knowledge of dictionary words can be represented using these networks. The word forms represent nodes in the network, and various relations between them, like – Synonimity, category, and subcategory.

3. WordNet

WordNet is a manually constructed online lexical reference public domain dictionary in which Lexical objects are organized *semantically* with basic distinction between nouns, verbs, adjectives, and adverbs (table 1) [10] [11].

Table 1: Wordnet

Category	Unique forms	Number of senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

WordNet is in the form of a large network of words organized in synonyms sets, called *synsets*. The synonymy is a *symmetric relation*. There are four separate databases, one for each - noun, verb, adjective, and adverb. Each of these databases consists of a set of lexical entries corresponding to unique orthographic forms, accompanied by set of senses associated with each form.

The **figure 2** shows the senses for the entry *board*. It has nine senses, four of these are: {*board, committee*}, {*board, plank*}, {*board, control panel, instrument panel, panel*}, {*board, circuit board, circuit card*}. To disambiguate the word *board*, it needs to be established as which sense out of these nine is a valid sense.

The primary semantic relation defined in WordNet is the "*is-a*" relation. Each concept subsumes more specific concepts, called *hyponyms*, and is subsumed by more general concepts, called *hypernyms*. Thus, the synsets are organized in a hierarchy via super-class/sub-class relationship in the form of *hypernym/hyponym*. A word concept represented by the synset {x, $x\phi$,...} is said to be a hyponym for the concept represented by the synset {y, $y\phi$,...} if one accepts sentences constructed from this, such as, *An x is a (kind of) y*. This relation can be represented by pointers from x to y and reverse. **Figure 3 and 4** illustrate the examples of hypernym and hyponym, respectively, for the word form *board*.

The noun board has 9 senses(first nine from tagged texts)
1.board - - (a committee having supervisory powers; "the board has seven members")
2. board- -(the flat piece of material designed for a special purpose; "he nailed the boards across the windows")
....
9. dining table, board - - (a table at which meals are served; "he helped her clean the dining table"; "a feast was spread upon the board")

The verb board has 4 senses (first 2 from tagged texts)
1.board, get on - - (get on board of (trains, buses, aircraft, ships, etc.))

Figure 2 : Part of the entry for *board* in WordNet.

9 senses of board
sense 1
board - - (a committee having supervisory powers; "the board has seven members")
=> committee, commission - - (a special group delegated to consider some matter)
.....
=> social group - - (people sharing some social relation)

sense 2
board - - (a flat piece of material designated for a special purpose; "he nailed boards across the windows")

Figure 3: Hypernym (*board is a kind of* ...) from WordNet.

The Hyponymy is a *transitive* and *asymmetrical* relation, and since there is a single super-ordinate, it generates a hierarchical structure, in which a hyponym is said to be below its super ordinate. This forms a chain of relations. A hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that super-ordinate. For example, the concept {*Cat*} has hypernym {*mammal*}, and one of its hyponym is {*Cheetah*}. Thus, a lexical tree can be constructed by following trails of superordinate terms, like: *cheetah* @® *cat* @® *mammal* @® *animal*. Here '@®' is *transitive* and *asymmetric* semantic relation that can be read '*is-a*' or '*is a kind of* (*ako*)'. By convention '@® ' is said to point upward. This design creates a sequence of levels or hierarchy, going from many specific terms at the lower level to a few generic terms at the top. Hierarchies also provide a conceptual skeleton for nouns. Whenever it is the case that a noun v *@→ a noun* w, there is always an inverse relation, $w \sim\!\!\rightarrow v$. The inverse semantic relation goes from generic to specific, so it is a specialization. Thus, due to the existence of subordinate/superordinate relation the WordNet can be searched upwards as well as downward at equal speed.

9 senses of board
sense 1
board - - (a committee having supervisory powers; "the board has seven members")
=>appeal board, appeals board, board of appeals (a board of officials that are not judicial
...
sense 2
board - -(a flat piece of material designated for a special purpose; "he nailed boards across the windows")
=> aquaplane - - (a board that is pulled by a speedboat as person stand on it and skims over the top of water)

Figure 4. Hyponym (... is a kind of board) from WordNet.

The relations between synsets in WordNet builds a complex semantic network guides in navigation for searching the relations among the synsets, which in turn helps for disambiguation.

4. Disambiguation Model

The disambiguation principle used here is based on following hypothesis: the meaning of a sentence is result of the combined effect of semantics of words used in that sentence. Thus, there is a sort of dependency relationship between the senses of words in a sentence, and the sense of each word is affected by the sense carried by other words in the sentence. For example, the words in the phrases – "sales tax", "class teacher", "flood control scheme", the word "tax" indicates that it is the one which is due to the "sales", the word "teacher" stands for the one who is for teaching in the "class", and "control" stands for the one used for "flood". This dependency of meanings among the words in a sentence can be explored to eliminate the ambiguity in the meaning of words in a given sentence. Due this interdependency of semantics of words in a sentence, there will be some overlapping words between one of the sense definition (the one which corresponds to the valid sense as per the current sentence's context) of the word to be disambiguated, and the sense definitions

of rest of the context words in the sentence clubbed together.

For each sense of a word, all its hyponyms (subordinates), as well as its hypernyms (superordinates), are also semantically related to the sense of the word. Therefore, there should also be overlapping words between the hyponyms / hypernyms of valid sense of the word and the definitions of rest of the context words in the sentence. Thus, to establish which sense out of a number of senses of an ambiguous word in a given sentence is valid, it simply needs to find out overlapping words between – the sense definitions, their hyponyms and hypernyms for this word, taken one by one v/s the sense definitions of rest of the context words in the sentence taken together. The word sense, including its hyponym and hypernym, which has maximum overlap with the sense definitions of rest of the context words in the sentence, is correct sense in this context. In case no overlap is found, which may be due to a new word, or the context domain is rare. In such cases, the most frequently used sense of the word being disambiguated is taken as valid sense, i.e., position one in the list of senses of that word.

The algorithm in **figure 5** finds the valid sense of a single word in the query phrase or query sentence. Exactly same method can also be used for disambiguation of words, sentence by sentence, in the text to be searched for Information Retrieval and Information Extraction. **figure 6** illustrates the process of word sense disambiguation.

The problem of ambiguity is more serious for smaller size queries. In the case of longer queries, a single ambiguous word plays a small role in determining the sense of the query, as the sense of the query sentence is decided by the large number of remaining context words, until unless there is not a large number of ambiguous words in the query. Let there be a *query* sentence q, and it is required to disambiguate the word $w \hat{I} q$. Following are the terms used in the algorithm:

$q' = q - \{w\}$ are context words in the sentence $qU = \{s_1, s_2, \ldots, s_m\}$ are sense definitions of w, where each s_i is a synsetP_i is set of hyponyms (subordinates) for $s_i S_i$ is set of hypernyms (superordinates) for $s_i P = \{P_1, P_2, \ldots, P_m\}$, hyponyms sets for $s_1 \ldots s_m$ $S = \{S_1, S_2, \ldots, S_m\}$, hypernyms sets for $s_1 \ldots s_m$ $t_1, t_2, t_3, T_1 \ldots T_m$ = temporary storage

Algorithm - Disambiguate:
1. Parse the query q and find its noun, verb, adj, and adv
2. Eliminate stopwords in q
3. $q' = q - \{w\}$
4. C = NUL, // set of sense definitions of words in q'
5. for each context word $v \in q'$ do
a. $C = C \cup$ synset of v
6. for each s_i, $i=1,m$ do
i. $t_1 = |s_i \cap C|$ //context term overlapping with synset
ii. $t_2 = |P_i \cap C|$ // context term overlapping with hyponym of s_i
ii. $t_3 = |S_i \cap C|$ // context term overlapping with hypernym of s_i
iv. $T_i = t_1 + t_2 + t_3$
7. find the largest of $T_1 \ldots T_m$, let this be T_j
8. if $T_j \# 0$
output – "d_j is closest sense of w"
else
output sense No.1 (i.e., most frequently used sense for v)
9. end

Figure 5: Word Sense Disambiguation Algorithm.

5.Illustrative Example

Given two query phrases, it is required to disambiguate the word *v=board,* in both of the queries, using WordNet and context of the query phrase. Following are the two phrases:

1. Selection(n) board(n)
2. Domestic(a) wiring(n) board(n).

where 'a' stands for adjective form of the word, and 'n' for noun form of the word.

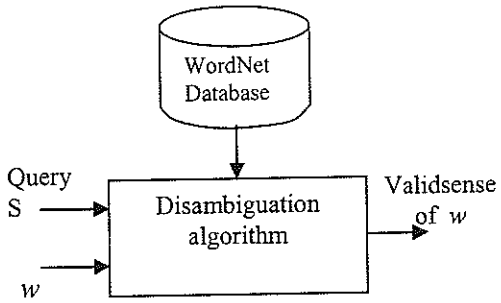

Figure 6. Word Sense Disambiguation

The different senses for *board*, as per WordNet, are shown in **table 2.** There are nine senses for noun *board*. Following are the sense definitions from WordNet for the context words in above query phrases.

Selection **Senses:**

Noun *selection* has five senses:

1. choice, selection, pick - -(the act of choosing, "your choice colours was unfortunate"; "you can take your pick")
2. selection - -(an assortment of things from which a choice can be made; 'the store carried a large selection of shoes")
3. choice, pick, selection - -(the person or thing chosen or selected; "he was my pick for Mayer")
4. survival, survival of the fittest, natural selection, selection - - (a natural process resulting in the evolution of organism best adapted to the environment)
5. excerpt, exact, selection - -(a passage selected from a larger work; "he presented excerpts from William James' philosophical writings").

Domestic **Senses:**

The adjective domestic has 5 senses:

1. domestic - -(of concern to or concerning the internal affairs of a nation; "domestic issues such as tax rates and highway construction")
2. domestic - -(of or relating to the home; "domestic servant"; "domestic science")
   domestic - - (of or involving the home or family' "domestic worries"; "domestic happiness"; "they share the domestic chores"; "everything sounded very peaceful and do 1. domestic"; "an author of blood-and-thunder novels yet quite domestic in his taste")
2. domestic, domesticated - -(converted or adapted to domestic use; "domestic animals"; "domesticated plants like maize")
3. domestic - -(produced in a particular country; "domestic wine"; "domestic oil")

Table 2: Overlap between the clubbed definitions of *selection* with *Synse,* and *hypernyms* of *board.*

Board Sense No.	Number of over-lap words in the synonym sets	Number of overlap words in the hypernym set	Total overlaps counts
1.	0	"select", 14 times	14
2.	0	0	0
3.	0	0	0
4.	0	0	0
5.	0	0	0
6.	0	"organism", 1 time	01
7.	0	0	0
8.	0	0	0
9.	0	0	0

Wiring **Senses:**

The noun wiring has 2 senses:

1. wiring - -(a circuit of wires for the distribution of electricity)
2. wiring - - (the work of installing the wires for an electrical system or device)

Disambiguating board in "selection board":

Let us first consider the first phrase above to disambiguate sense of *board.* Here, S="selection, board". To find out the correct sense of *board*, all synonyms in each sense of *board* are compared with the clubbed sense definitions of remaining context words, i.e., S-{*board*} = {*selection*}, in the sentence S. It has been found that the sense 1 in

board has maximum overlapping words with the clubbed sense definitions of *selection,* where word "select" appears total 14 times common in –synonyms, hyponyms, and hypernyms put together of sense number 1 of *board* and definitions of *selection*. The comparison count table for this is shown as **table 3.**

Table 3:Overlap between the clubbed definitions of *domestic* and *Wiring* with *synset* and *hypernyms* of *board.*

board Sense No.	Number of overlap words in the synonym sets	Number of overlap words in the hypernym set	Total over-laps count s
1.	0	0	0
2.	0	0	0
3.	0	0	0
4.	0	"device" 3 times	03
5.	0	0	0
6.	0	0	0
7.	"electrical" 2 times, "device" 1 time	"electrical" 2 times, "device" 5 times	10
8.	0	"electrical" 10 times, "device" 5 times	15
9.	0	0	0

Thus correct sense of *board* in the phrase "selection board" is*:*

1.board — (a committee having supervisory powers; "the board has seven members.")

Disambiguating board in "domestic wiring board":

Let us consider now phrase 2 above, where, again word form *board* is to be disambiguated. The context phrase is, S = "domestic wiring board". Just like previous case, the correct sense of *board* is to be found out, from its total nine senses to suit context present in the sentence carrying the word *board.* When synonyms sets of each sense of *board* are compared for overlapping words, with the clubbed sense definitions of all the remaining context words in the sense S, i.e., S-{*board*} ={*domestic wiring*}, the corresponding overlaps are shown in table 3. In both the tables, numbers of overlap words in the hyponym sets are zero for all the senses, hence not shown in the tables.

To disambiguate the word form *board* in phrase "domestic wiring board", the most close sense of *board* is number 8, as it has maximum overlap of 15 times of different words with the combined sense definitions of remaining context words in the given sentence carrying the word *board.* The *synsets* for the senses 7 and 8 of *board* are listed as follows from figure 2.

7. Control panel, instrument panel, control board, board, panel - - (an insulated panel containing switches and dials and meters for controlling electrical devices; "he checked the instrument panel"; "suddenly the board lit up like a Christmas tree")
8. Circuit board, circuit card, board, card - - (a printed circuit that can be inserted into expansion slots in a computer to increase the computer's capabilities).

These show that sense of *board* is – "*circuit board, circuit card, board, card"*. Among the many possible examples of "boards", it says that one type board can also be PCB (printed circuit board) used in computers and other electronics devices.

The other next possible sense of *board*, with number of overlaps 10, is – "*control panel, instrument panel, control board, panel"*. Which is also equally valid sense, but it is more suitable for control applications. Its sense fits better for a board in laboratory use, possibly electrical machines laboratory.

6. Discussion and Concluding Comments

The experiments have shown that semantic network based method using WordNet has resulted in 100% disambiguation. The reason is that for an ambiguous word, its synset, hyponyms, and hypernyms have been considered for matching with the sentential context of the ambiguous word. In addition, the WordNet synsets consists examples of sense tagged sentences based on the word being considered for disambiguation. Since these

sentences are collected from the tagged texts, thus suggesting the correct sense in the given context.

The presented method is superior than the learning based methods, like tagged texts making use of large text corpus. The tagged text corpora based method needs a very large. Making such a large text available is a difficult task. Even if such a text is made available, it will be in gigabytes, and finding the correct sense for a word out of that will be beyond the processing capabilities of current computers in real-time. The corpus is invariably domain specific, which many times do not suit the requirement of disambiguation in other domain contexts. Often it is difficult to get a corpus, which is general and covers all the kind of senses of all types of words.

The human beings reasoning for disambiguation are understood to be somewhat similar to the process presented here. In this approach, the semantic networks provide efficient navigation, and WordNet provides pointers for the purpose of navigation among the related words, which the disambiguation process to be is much faster than the corpus based methods. In addition, since all the variants of the word as well as all kinds of its relations, and in addition sentences from the tagged texts have been considered, disambiguation is bound to be correct and efficient.

The requirement for the new method is that the sentence or the query, in which a word is to be disambiguated, should be sufficiently large so that there are enough contexts available to help in resolving the ambiguity. The resolution will suffer if context words are limited.

The model of disambiguation system presented here is based on the principle that a document relevant to a query might contain either the words in the query or their synonyms. This implies that recall can be improved by considering the synonyms as part of the IR and IE queries. However, if all the possible synonyms of the words in the query are added as part of the query, then many irrelevant document are also likely to be retrieved, resulting to reduction in precision. Thus, to improve both the precision and recall, only the relevant synonyms should be used in the query. Finally, the problem rests in finding these relevant synonyms, which carry the valid sense of the word to be disambiguated in the query.

The model suggested here, makes use of word hierarchies through semantics networks, which find the words related to the word being disambiguated (i.e., the synonym set of the word), based on the relation of *transitivity* and *asymmetry*. If such a relation exists, and searched through the semantics network, the synset of the located word is correct sense of the word being disambiguated.

7. References

[1] Miller G.A., WordNet : A Lexical Database for English, *Communications of ACM*, Nov. 1995, Vol. 38, No.11, pp. 39-41.

[2] Sanderson M. and Rijsbergen C.J.V., The Impact on Retrieval Effectiveness of Skewed Frequency Distributions, *ACM Transactions on Information systems*, Vol. 17, No. 4, October 1999, Pages 440-465.

[3] Krovetz R. and Croft W.B., Lexical Ambiguity and Information Retrieval, *ACM Transactions on Information systems*, Vol. 10, No. 2, April 1992, Pages 115-141.

[4] Riloff E. and Lehnert W., Information Extraction as a Basis for High-Precision Text Classification, *ACM Transactions on Information systems*, Vol. 12, No. 3, July 1994, Pages 296-333.

[5] Voorhees E.M., Using WordNet™ to Disambiguate Word Senses for Text Retrieval, *ACM-SIGIR Conference*, June 93, Pittsburgh, PA, USA.

[6] Roth D., Learning to Resolve Natural Language ambiguities : A Unified approach, *Conference of American Association for Artificial Intelligence*, 1998.

[7] Mihalcea R., and Moldovan D.I., A Method for Word Sense Disambiguation of Unrestricted Text, *Department of Computer Science and Engineering, Southern Methodist University*, Dallas, Texas, USA.

[8] Miller G.A., Chodorow M, and Landes S., Using Semantic Concordance for Sense Identification, in *Proceedings of Coling* 1996.

[9] Winston P.H., Artificial Intelligence – 3rd Edition, *Addison-Wesley Pub. Co.*, 1993.

[10] Miller G.A. et al, Introduction to WordNet : An On-line Lexical Database, *Princeton University*, USA, August 1993.

[11] Jurafsky D. and Martin J.H., Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, *Pearson Education*, Inc., 2000.