“WORD SENSE DISAMBIGUATION”

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

In

Computer Science and Engineering

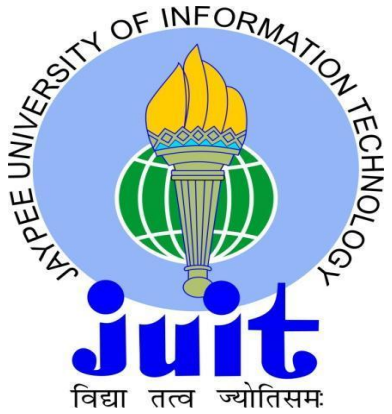
By

Taniya Chauhan (131287)

Under the supervision of

Ms. Ruhi Mahajan

To



Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

DECLARATION

I hereby declare that the work presented in this report entitled **" WORD SENSE DISAMBIGUATION"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2016 to May 2017 under the supervision of **Ms. Ruhi Mahajan** (Assistant Professor, Computer Science and Engineering Department). The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Taniya Chauhan (131287)……

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Ms. Ruhi Mahajan

Assistant Professor

Computer Science & Engineering Department

Dated:

ACKNOWLEDGEMENT

It gives me immense pleasure to recollect the whole time and effort utilized in making this project a success. First and foremost, I would like to thank Ms. Ruhi Mahajan, my project supervisor and Assistant Professor, Department of Computer Science and Engineering, Jaypee University of Information Technology Waknaghat, Solan, for her valuable time and for being always supportive to my work under her guidance and giving me the freedom of thought always. I wish to express my deep sense of gratitude and indebtedness to Prof. Dr. Satya Prakash Ghrera, Head, Department of Computer Science & Engineering and Information Technology, University of Information Technology Waknaghat, Solan, for providing me the opportunity to utilize the facilities of the Department of Computer Science & Engineering.

Table of Contents

List of Abbreviations

Acronym	Definition
NLP	Natural Language Processing
WSD	Word Sense Disambiguation
IR	Information Retrieval
IE	Information Extraction
MT	Machine Translation
MRD	Minimum Required Distance

List of Figures

List of Tables

List of Graphs

ABSTRACT

Natural language processing (NLP) is field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. Natural language generation systems convert information from computer databases into readable human language.

Word sense disambiguation (WSD) is still an open research area in natural language processing and computational linguistics. It is from both theoretical and practical point of view. WSD is considered an AIcomplete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence.

Here, the problem is to find the sense for word in given a context. It is a technique of natural language processing (NLP), which requires queries and documents in NLP or texts from Machine Translation (MT).

A method of word sense disambiguation that employs measures of gloss overlaps using English Wordnet for all works task is formulated. This system will assign a sense to every content word in a text that is found in the English Wordnet.

CHAPTER 1
INTRODUCTION

1.1 Natural Language Processing (NLP)

Natural language processing (**NLP**) involves creation of computational models of facets of processing of language used by humans.NLP involves natural language understanding, that is, you enable the computers to derive meaning when an input is put in human or natural language , and others involve natural language generation. It is in relevance with the area of human–computer interaction.

NLP application creation and development is quite demanding. The communication between the computers and humans has to be in some sort of programming language that is exact and set of rules, clear. On the other hand human language, however, is often full of multiple contexts and is not clear at all. Structure of human speech is ambiguous and consists of regional languages, lingo and is very informal. There are various methods to NLP present today. All of these approaches are based ona artificial intelligence that checks and improvises a program's analysis capacity by bringing data patterns in use .If someone wishes to research in this area of Natural Language Processing then he has to understand that whole of it will revolve around searching only.

By now most basic and fundamental NLP operations which are performed in software programs are:

- Breaking sentence into parts and parse methods
- Performing detailed analysis

- Fetching already named elements.
- Co-reference solutions.

1.2 Applications of NLP

There are 7 applications in NLP and these are as follow:

1) Machine Translation
2) Speech Recognition
3) Speech Synthesis
4) Information Retrieval (IR)
5) Information Extraction (IE)
6) Word Sense Disambiguation (WSD)
7) Parts-of-speech tagging

1.3 THE CHALLENGES OF NLP

There are a number of factors that make NLP difficult. These relate to the problem of representation and interpretation. Language computing requires precise representation of content. Given that natural languages are highly ambiguous and vague, achieving such representation can be difficult. The inability to capture all the required knowledge is another source of difficulty. It is almost impossible to embody all sources of knowledge that humans use to process language. Even if this were done, it is not possible to write procedures that imitate language processing as done by humans. Perhaps the greatest source of difficulty in natural language is identifying its semantics. The principle of compositional semantics considers the meaning of a sentence to be a composition of the meaning of words appearing in it. The ambiguity of natural language is another difficulty. These go unnoticed most of the times, yet are correctly interpreted. This is possible because we use explicit as well as implicit sources of knowledge.

Communication via language involves two brains not just one- the brain of the speaker/writer and that of the header/reader. Anything that is assumed to be known to the receiver is not explicitly encoded. The receiver possesses the necessary knowledge and fills in the gaps while making an interpretation. Our viewpoint in those words alone does not make a sentence. Instead, it is the words as well as their syntactic and semantic relation that give meaning to a sentence. As pointed out by Wittgenstein (1953); A word's meaning varies withe context in which it is used. With time the language keeps on changing. New words get introduced into the language constanly and even the existing words which are currently present in language also come up in some totally different new context.Example for this can be the reporting of World Trade Center terrorist attack that happened in 2004.Every newspaper reported it in different manner by making use of words in different context.They used the term 9/11 to draw reference to that incident. When we process written text or spoken utterances, we have access to underlying mental representation. The only way a machine can learn the meaning of a specific word in a message is by considering its context, unless some explicitly coded general world or domain knowledge is available. The English word "while" was initially used to mean "a short interval of time". But now it is more in use as a conjunction. Another example of cultural impact on language is the representation of different shades of white in the Eskimo world. It may be hard for a person living in plain to distinguish among various shades. Similarity, to Indian the word 'Taj' may mean a monument, a brand of tea, or a hotel, which may not be so for a non-Indian. As humans, we are aware of the context and current knowledge and also of the language and traditions and utilize these to process the meaning.

Quantifier-scoping is another problem. The scope of quantifier (the, each, etc) is often not clear and poses problem in automatic processing.

1.4 Word Sense Disambiguation (WSD)

Word sense disambiguation (WSD) is a crucial task in the area of NLP. WSD is a natural classification problem: If you are provided with a word and all of its possible meanings as present in the dictionary then classification of an presence of the word in context into one or more of its meaning classes. WSD has impact on many Natural Language Processing applications like Information Retrieval, Information Extraction and Machine Translation.

Word sense disambiguation (WSD) is a problem in which it is hard to determine which "sense"or which meaning of some word is initiatedwhen we make use of that word somewhere in particular with a predetermined context. One of the central and most widely investigated problems in NLP is word sense disambiguation. In WSD, given a sentence aligned from corpus and give different senses of words present in that sentence. The senses are taking from Wordnet such as indo-wordnet, somewhere these senses are noun, some where these are adjective, some where these are verb and adverb etc.

For example consider a word ‘knife’ which has two senses. One sense is “a tool” and another sense is “a weapon”. By using example this shown in bellow:-

(1) Chef cut the fruits with a kitchen knife.

(2) A woman was murdered with a knife.

A human sees it like that the first sentence is using the word "knife”as a tool to cut things and in the second sentence; the word "knife” is being

used as a weapon to kill someone. Destructive aspect of knife is coming into play in the second sentence.

The problem of word sense disambiguation is an AI complete problem. A problem which can be solved only by first resolving all the difficult problems in the artificial intelligence (AI) is known as AI complete. For example: representation of common sense encyclopaedic knowledge. It is an issue of choosing a meaning for a word from some already defined possibilities. The meaning database used here is taken from Wordnet.

Example: - If I take a word 'joker', it indicates three meaning. It shown bellow:-

1) A playing card.
2) A prankster.
3) Batman villain

In word sense disambiguation the ambiguity is a major problem for human in day to day communication because all most all natural has multiple meaning. It is more prominent problem for computer to understand ambiguous word. Human at least understand which sense was required of the target word with respect to the sentence after reading the sentence. But computer cannot understand which sense is required for the sentence. They got to learn it.

1.5 Objective and Motivation

- The survey of WSD using knowledge base approach has been going on in many years.
- The main aim of this survey is to identify the ambiguous word in a given context.

- Human language is not precise.It is full of ambiguity. A lot many words are there which can have different and multiple meanings on the basis of the context in which they are used in the sentence.

The primary objective of the documentation can be summarized as follows:

- To study WSD an open problem of NLP.
- Implementation of WSD algorithms using JAVA and Wordnet.
- To identify the most proper sense of ambiguous word(s).

1.6 Methodology

User needs to run the application. The user has to enter the context or the sentence and the target word for which the definition is to be projected in the given context. Given the concept, the algorithm searches for the best gloss of the target word in the already given frame of reference from the set of all the glosses of thatparticular word and that gloss is provided to the user as the answer.

Chapter 2

LITERARTURE SURVEY

2.1 Previous works on WSD

1) **"Manish Sinha, Mahesh Kumar Reddy .R, Pushpak Bhattacharyya , Prabhakar Pandey Laxmi Kashyap" [2]:**All belonging from Department of Computer Science and Engineering Indian Institute of Technology Bombay, Mumbai India, used the Hindi Wordnet for disambiguation of Hindi words.Till then there was never any work done for some Indian language, this being the first was a notable step towards Indian language processing. The working and efficiency can be definitelyrevamped if the analysis is managedcomprehensively. Presently computeris unable to detect the concealed similarity in existence of analyticalchanges. Since Indian languages are enriched in analysis thing, an exhaustive pre-refining for the analysis is a must thing to be done in the entire WSD process.

2) **Shallu, Vishal Gupta [4]:** From the University of Institute of Engineering & Technology, Punjab University, Chandigarh, India conducted "A Survey of Word-sense Disambiguation Effective Techniques and Methods for Indian Languages". Author has used the co-occurrence graphs. A tool was provided by author for domanin exploration. Web page repository was needed to examine this algo. Top30 contexts gotexamined for all60 uses which comprises 1300 contexts as whole..

3) **Arindam Roy1, Sunita Sarkar2 and Bipul Syam Purkayastha [5]:** This is clear from trial scanty outcome that execution of overlap based approach is poorer than blending reasonable separation and semantic diagram technique. It is normal since overlap based approach experience the ill effects of overlaping, particularly on account of nouns. Method introduced here gives better execution over overlap based method with machine decipherable lexicons in light of the fact that not only the gloss and cases of the objective and setting synsets are taken however it likewise considers the gloss and cases from their hypernyms. In the event that higher modifier precision is strived to be accomplished then second strategy ought to be contemplated as the semantic chart separate score has been considered in this technique.

4) **Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar, Kunal Sinha [8]:** Their approach has established better performance in enhanced WSD technique depending on specific learning sets. Since the datasets have been enriched by new data so disambiguation accuracy gets improved to large extent. By making use of extensive experimentation we can achieve more desirable precision value, recall value, and F-Measure.

5) **Andres Montoyo, Armando Su´arez, German Rigau, and Manuel Palomar [6]**: They proposed hypothetical work in WSD. Didtinct sources of knowledge were needed in this project. Distinct techniques were also needed. They aimed to delve into new approaches to combine two methods. One based on knowledge and other one based on collections.They came up with 3 distinct techniques as to exhibit their work. Both methods were combined in this scheme via sources of information.Ultimately they proved that methods based on

knowledge can encourage methods based on collections to produce better outcomes and it goes other way round too.

6) **Simone Paolo Ponzetto, Roberto Navigli [3]:**These people have represented a huge-scale method for automated enhancement of computational dictionary using encyclopedic relational sources. Large informational data put into Wordnet had high end quality.The experiments confirm that the huge amounts of knowledge injected into WordNet is of extremely high quality and, more importantly, Permits basic information based WSD frameworks to execute. Even allows maximum-executing supervised ones too to perform in coarse-grained context. Also beats them up in space particular content.

7) **Egoitz Laparra and German Rigau [9]:** SSI-Dijkstra is used in this approach to designate proper synset of wordnet to linguisticaly related logical units. Algorithm relies on bringing in use huge information data which is taken from wordnet and its extended versions. Original SSI-Dijkstra needs sets of words that are interpreted before. Develops algorithm to deal with polysemous words too.

2.2 DOES WSD IMPROVE INFORMATION RETRIEVAL?

A rudimentary form of semantic annotation is to label There are a number of factors that make NLP difficult. These relate to the problem of representation and interpretation. Language computing requires precise representation of content. Given that natural languages are highly ambiguous and vague, achieving such representation can be difficult. The inability to capture all the required knowledge is another source of difficulty. It is almost impossible to embody all sources of knowledge that humans use to process

language. Even if this were done, it is not possible to write procedures that imitate language processing as done by humans. Perhaps the greatest source of difficulty in natural language is identifying its semantics. The principle of compositional semantics considers the meaning of a sentence to be a composition of the meaning of words appearing in it. The ambiguity of natural language is another difficulty. These go unnoticed most of the times, yet are correctly interpreted. This is possible because we use explicit as well as implicit sources of knowledge.There are a number of factors that make NLP difficult. These relate to the problem of representation and interpretation. Language computing requires precise representation of content. Given that natural languages are highly ambiguous and vague, achieving such representation can be difficult. The inability to capture all the required knowledge is another source of difficulty. It is almost impossible to embody all sources of knowledge that humans use to process language. Even if this were done, it is not possible to write procedures that imitate language processing as done by humans. Perhaps the greatest source of difficulty in natural language is identifying its semantics. The principle of compositional semantics considers the meaning of a sentence to be a composition of the meaning of words appearing in it. The ambiguity of natural language is another difficulty. These go unnoticed most of the times, yet are correctly interpreted. This is possible because we use explicit as well as implicit sources of knowledge.

2.3 WSD APPROACHES

Word Sense Disambiguation Methods are categorizedmainly into 3 basicsections-

a) Knowledge based approach

b) Supervised approach

c) Unsupervised approach.

2.3.1 Knowledge-based WSD

Knowledge-based methods are established with respect to various learning sources like machine understandable word references. Wordnet (Miller 1995) to a great extent utilized machine understandable dictionary in the exploration area. Four information focused strategies are mainly brought into practice these days.

(1) LESK Algorithm

It is an essential machine lucid vocabulary based calculation created for WSD. The count depends upon the overlap of lexicon implications of words present in a sentence. Initial phase in this method is choice of a short expression containing the uncertain word from the sentence. After that vocabulary definitions for the various senses of the objective word and the other critical words appear in the expression are assembled from an online Lexicon. Next, each of the glossaries of objective word are diverged from glossaries of various words introduce in the sentence. Sensehaving most outrageous overlapis designated assense of objective word.

(2) Semantic Similarity

In situations where the words are connected with each other and along these lines share same or similar context , proper meaning or definition is picked via implications which are identified inside minimum semantic separation. In order to determine how much semantic relatedness is between two words, we have various similarity measures. In case there are more than two words, then this approach gets computationally cumbersome and intensive.

(3) Selectional Preferences

Itidentifies information regardingsimilar relations in word types.It announces common sense by making use of the source of information. For instance, Racing cars, Big tireshave a semantic relationship. In this method word-senses which are not appropriate gets ignored.Its made sure that senses havingeuphony with customary sense rules are kept.

Essential thought behind themethod is to check how number of times such word match happens in bulk with syntactic relation. Senses of words are thenlearned using the tally. Numerous other techniques are also available that claim to find similar relation among words by making use of conditional probability.

(4) Heuristic Method

Basic idea behind this approach is that you evaluate heuristicsfrom distinct language properties so to identify correctmeaning.WSD system is estimated using mainly these heuristics types:

1) Most Frequent Sense

2) One Sense per Discourse

3) One Sense per Collocation.

Most frequentsense works by discovering every single conceivable sense a word can have and is predominantly right, that one sense happens regularly than the others.

2.3.2 Supervised WSD

Supervised methodologies which are used with WSD frameworks utilize machine-learning techniques from written explanatory meaningful data which is created manually. Classifier will use the training set to learn and this training set comprise illustrations identified with target word. Dictionary helps in creating these tags manually. Primarily this providesbetter outcomes as compared to rest of approaches. Supervise WSD methods:

(1) Decision List

Decision list is arrangement of "if-then-else" rules. In case some specific features for a particular word need to be included then training sets are used. A few parameters are made by making utilization of those guidelines. Last rule order makes list and basis is diminishing scores. “When any word is considered, first its occurrence is calculated and its representation in terms of feature vector is used to create the decision list, from where the score is calculated. The maximum score for a vector represents the sense.”

(2) Decision Tree

This denotes the classification rules in some tree structure. Training data set is divided recursively that is it is divided over and over again. Decision tree contain internal nodes. These nodes denote tests. We apply this test to feature value then. Every branch of tree signifies an output.Once you get to a leaf node then word meaning gets represented.

(3) Neural Networks

Model processes the data by using connectionist method. Artificial neurons help in processing. If training context gets divided into sets which don't overlap then goal is achieved. Input to such computational models is given in pairs. Link weights and pairs are adjusted to improve initiation.These are adjusted very gradually. In such networks nodes are perceived as words. Words then initiate ideas with whom they are related in semantic manner. Intermediate layers let inputs inseminate from input layer to the output layer.Network allows inputs to inseminate easily so to reach output. If connections are spreading and loops have been generated then it is too cumbersome to remove outputs from network.

(4) Exemplar-Based or Instance-Based Learning

It is a supervised technique. It creates classification models with the help of examples.Current examples get accumulated and whatever new instances are arriving are considered for classification.These arriving examples slowly get added. An example of this technique is k-nearest neighbour.

A step sequence is followed. Right at the beginning all of the examples are taken into count at one place.Later Hamming distance is computed for those examples using algorithm such as k-NN.Distance further calculates closeness with examples which are stored up. If k >1 then maximum output sense resides with k-nearest acquaintances.

(5) Support Vector Machine

Main objective of this algorithm is to search hyperplane among two classes. Aim is to maximize the partition limit. Test example resides in hyperplane and its classification depends on its location.Input is mapped to space with an aim to decrease training cost.Kernel functions are used in the testing sequence. A regularization parameter with deafault value 1 is used in training examples which are not separable. Regularization parameter helps mediating between higher margins and less errors associated with training.

2.3.3 Unsupervised WSD

Unsupervised WSD methods are independentof any external knowledge sources. These algorithms for the most part don't designatesense to words rather separateword implications due to information that is found in un-namedtext collection. Two distributional methodologies associated with this technique;

- “Monolingual corpora”
- “Translational corpora”

When corpora is viewed in parallel.

They have subtypes:

- “Type-based”
- “Token-based”

First one collects occurences of mark word and then removes ambiguity. In second one context with the mark word is collected and then ambiguity is removed.Unsupervised methods are listed below:

(1) Context Clustering

First step is to generate context vectors. Soon they get clustered so to determine word sense. Whole of the context clustering technique is built on this foundation. There exists vector space instead of word space but dimensions are words. Every word present in corpus is considered vector and number of times it pops up is taken in account in context.A matrix is created administering semantic metrics. Later on some technique is applied with the aim of clustering.

(2) Word Clustering

It shares a lot many similarities with context clustering. Grouping of words is done in a way that they are are semantically un-differentiable from each other. Lin's strategy is taken up for performing such sort of grouping. Every un-differentiable word is perceived as mark word from whom ambiguity is to be removed. Compuattion is done on basis of commonalities they share. It is very easily taken up from corpus since words would be sharing similar kind of dependency.Later clustering is performed. If lsit is chosen then fisrt operation performed will be identification of similarity and then generating a similarity tree out of that. Only single node is there at start. Each word the list comprises of is treated with iteration. Lastly pruning is perfomed to the

tree and it leads to creation of sub-trees.Sub tree corresponding to root with initial word provides correct sense.

(3) Co-occurrence Graph

Co-occurrence graph is generated.

Vertex is V and edge is E. V denoting words in text and E summed up if syntactically similar words occur together in text.Graph is generated and corresponding adjacency matrix for that graph too. Weight for edge {m,n} is :

"wmn = 1- max{P(wm | wn), P(wn | wm)}"

Where P(wm|wn) is the freqmn/freqn and where freqmn is the co-event recurrence of words wm and wn, freqn is the event recurrence of wn. Word with high repeat is dispensed the weight 0, and the words which are occasionally co-happening, delegated the weight 1. Edges, whose weights outperform certain edge, are neglected. By then an iterative calculation is associated with chart and the center point having most surprising relative degree, is picked as focus. Calculation touches base at an end, when repeat of a word to its middle guide ranges toward underneath cutoff. Finally, entire focus point is shown as significance of the given target word.The centers of the objective word which have zero weight are connected up and the base traversing tree is produced out of thegraph. This spreading over tree is expected to unequivocal the genuine feeling of the objective word.

(4) Spanning tree based approach

"Word Sense Induction is the task of identifying the set of senses of an ambiguous word in an automated way. These methods find the word senses from a text with an idea that a given word carries a specific sense in a particular context when it co-occurs with the same

neighbouring words. In all of these approaches, foremost step is to create a co-occurrence graph (Gq) . Following that is a sequence of steps which are performed to find the most appropriate and exact sense of an equivocal word in a specific context:

a. All the nodes with degree is 1 are neglected out ofGq.
b. The maximum spanning tree (MST) TGq of the graph is derived.
c. After that, the minimum weight edge eεTGq is eliminated from the graph one by one, until the N connected components (i.e., word clusters) are formed or there remains no more edges to eliminate."

2.4 Comparison of WSD Approaches

Table 2.1 Contrastingmethods of WSD

Method	**Advantages**	**Disadvantages**
Knowledge-Based	Algorithms provide withincreasedPrecision.	Since these algorithms strongly rely on overlap, so in their case the problem of sparsity in overlap occurs and performance levelis quite determined by dictionary definitions.
Supervised	This algorithm type issuperior than other approaches keeping in mind	In case of resource scarce languages, unsatisfactory results are obtained from these

	the implementation view.	algorithms.
Unsupervised	No need of any sense stock and sense explained corpora in these methodologies.	Implementation is very cumbersome and performance is less appreciableif contrasted to other methods.

2.5 Work done on Algorithms

Table 2.2 WSD Algorithms

Type of algorithm	**Author/s**	**Language**	**Performance**	**Year**
Genetic Algorithm	Sabnam Kumari Prof. Paramjit Singh	Hindi	91.60%	2013
WordNet	Udaya Raj Dhungana and group	Nepali	88.06%	2014
Decision Tree based System	Sivaji Bandyopadhyay and group	Manipuri	71.75%	2014
Modified Lesk's Algorithm	Rakesh and Ravinder	Punjabi	Satisfactory	2011

Knowledge based Approach	Rosna P Haroon	Malayalam	Satisfactory	2010
Knowledge Based Approach using Hindi WordNet	Prity Bala	Hindi	62.50%	2013
WordNet	Manish Sinha and group	Hindi	40-70%	
Un-Supervised Graph-based Approach	Ayan Das, Sudeshna Sarkar	Bengali	60%	2013
Selectional Restriction	Prity Bala	Hindi	66.92%	2013
Semi-Supervised Approach	Neetu Mishra Tanveer J. Siddiqui	Hindi	61.70%	2012
Machine Readable Dictionary	S. Parameswarappa, V.N.Narayana	Kannada	Satisfactory	2011

Chapter 3

SYSTEM DEVELOPMENT

3.1 Software Requirements

The specifications associated with minimum software requirement used in the development of this project are as follows:

Operating System : Window 2000, XP

Presentation layer : Java, Swings

Database : WordNet 2.1

Database layer : jdbc.

Presentation : Power Point 2003

Documentation Tool : Ms Office

Java Platform : MyEclipse 8.6, jdk 1.7

3.2 Hardware Requirements

The specifications associated with minimum hardware requirement used in the development of this project are described below:

Processor : Pentium IV

RAM : 512MB RAM

Hard Disk : 10GB

Monitor : Standard Color Monitor

Keyboard : Standard Keyboard

Mouse : Standard Mouse

3.3 Data Used (WordNet)

In this project the English WordNet is used as data for Word Sense Disambiguation. The words are taken from sentences present in these corpora.

WordNet is basically a Lexical Knowledge Base in English Language. Professor George A. Miller developed it in Cognitive Science Laboratory at Princeton University under his supervision. Data of more than 130,000 words which are assembled in more than

89,000 synsets (ideas or equivalent word sets) is incorporated into this Wordnet database.They have many relations and the cheif out of all the relations happens to be the hyponymy relation . WordNet has quite similar characteristics as of a MRD, since it includes definitions of terms for individual senses exactly in a way it is present in dictionary. It characterizes sets of synonymous words that speak to a one of a kind lexical idea, and composes them in a calculated chain of command like a thesaurus.Various other relations such as meronymy, antonymy, etc.are also a part of WordNet that provide huge and richfreely available lexical resource. WordNet was designed with the thought of being used by programs; hence, it doesn't have a large portion of the MRD related issues.

- Synsets :-
  - Synset id :

The synset id gives the ids of the senses of a particular word.

  - Parts of speech

Parts of speech give the POS of the sense of a word. Somewhere this pos is noun, somewhere verb, somewhere adverb and some where adjective etc.

  - Synonyms

Synonyms give the similar names of a word at all senses.

  - Gloss

Gloss gives the definition of a word of that particular sense.

- Hypernymy:-

A word with a broad meaning consisting a category in which words with more specific meanings fall. Hypernymy is used for semantic and conceptual relations.

- Hyponymy:-

Hyponymy shows the more relations between the more general terms and more specific instances of it. A hyponym is a word or a phrase whose semantic field is more specific than its hypernymy.

3.4 Algorithms Implemented

For this particular project we have implemented three algorithms Lesk Algorithm, Modified Lesk Algorithm and Semantic Similarity Algorithm.

3.4.1 Lesk Algorithm

The **lesk algorithm** is a traditional and very classical algorithm for word sense disambiguation. It was presented by Michael E. Lesk in 1986. It recognizes word meanings in a particular aspect by bringing in use overlapping between definitionsproviding approximately 50-70% accuracy.

The original Lesk algorithm basically works on the principle of disambiguating a mark word by contrasting its meanings set with the meaning sets of neighboring words. That sense is assigned to the mark word whose glossary has the most covering words with the glossaries of its surrounding words.

As an example, the first three senses in WordNet of *key are:*

1)Metal gadget formed such that when it is embedded into the proper bolt the bolt's component can be turned

1) Anythingimportant for justifying; “the key to development is economic integration”

2) Parameter of voice; “he spoke in a low key”

And we have marked in italic the words which have same meaning with the input line given below:

- He put the *key* and window got bolted.

Sense 1has 3 overlaps

Rest senses have no overlaps hence the first one is chosen.

Algorithm calculates a value for all senses of the mark word. To findvalue for a meaning of mark word the target sense is contrastedwith senses of mark word. Algorithm finds meaning of each mark word which has maximum closeness to mark sense. Relatedness scores are calculated between the target sense and each most related context sense. Sum of these relatedness scores is computed and assigned to each target sense. Once the algorithm finds a value for all senses of mark word, the sense with greatest value is designated to mark word.

The greatest deficiency of straightforward lesk calculation is that lexicon definitions are regularly short and simply don't have enough words for this calculation to function admirably In order to curb this issuewe try attaching this algorithm to WordNet. Wordnet is semantically organized. Other than storage of words and their glosses like a typical lexicon, WordNet additionally "interfaces" related words together. Issue of short definitions is vanquished by spoting basic words not simply between the meanings of words which being disambiguated, additionally among meanings of words which have close relatedness with them in WordNet.

3.4.2 Modified Lesk Algorithm

Apply Lesk's fundamental method to make advantage of synonyms having strong interconnections between them offered by WordNet. While Lesk's calculation's examinations are confined just to the glossaries of the words being disambiguated, adapted lesk's approach is unrivaled since it is particularly ready to think about the glossaries of words that are sharing relatedness with the words to be disambiguated too.So, it beats the impediments:

- Get to a lexicon with senses organized in a progressive manner (WordNet). This broadened rendition utilizes the glossary/meaning of the synset as well as considers the importance of related words.
- In order to get more accuracy a refined technique can be implemented to beat the limitations of lesk algorithm.

Suppose we have to find the corect meaning of every word according to the context in a sentence of N words then each word whose meaning has to identified is taken as target word.Algorithm is summarized in steps described below:

1. Select a context: Choosing a context improves computational time so if in the event that N is long, we will characterize K context around the objective word (or k-closest neighbor) as the arrangement of words beginning K words to one side of the objective word and trailing K words to one side. This will limit the computational space that decreases the handling time. For instance: If k is four, there will be two words to one side of the objective word and two words to one side.

2. For each word in the chosen setting, we see and write down all conceivable senses of verb or noun.

3. For each sense of a word, we list theaccompanying relations

4. Now we club all conceivable glossary combines that are removed in the previously mentioned steps and compute the relatedness via hunting down overlapping. The general score is figured as the expansion of the scores for every connection match.

5. After every mix has been scored that sense is gotten which has the most astounding score to be the most suitable sense for the objective word in the picked setting space.Expectantly not only the most appropriate sense but also the associated part of speech for a word is provided by the output.

3.4.3 Semantic Similarity Algorithm

Semantic similarity is basically sometechnique to ascertain the relatedness or the similarity between two notionson the basis of some provided philosophy. Technically, this approach which is being brought into practice to identify the concepts which have common "characteristics" amongst them. Humans do not really know the established and orderlymeaning of relatedness amongnotions.Still they can figure out the commonalities. For instance, a human will figure out that “spoon” and “cutlery” are more closely relatedthan “spoon” and “plate”.

Semantic similarity methods are beingpracticed comprehensivelyin most applications areas such as insightful learning based and semantic data recovery frameworks to recognize the most appropriate match

between query and records. However, this algorithm is still less practical for real time applications and for a extensive large scale use.

Text similarity is also a part of semantic similarity and it play an ever increasing significant part in research work related to text. In operations, for example, recovery of data, grouping of content, report bunching, theme location, questions era, address replying, short answer scoring, machine interpretation and others content similitude is being utilized to a more prominent degree. Fundamental and most primary part of text similarity is finding similarity between the words .

A common character sequence characterizes logically similar words.Whereas if same thing is there or are used in same way or maybe are opposite of each other then they are semantically similar.

There are three types of measures in semantic similarity:

- String-Based measures:These measures perform on string sequences and character composition. If you text strings are given then string metric will calculate the commonalities or differences between those character arrays so to attain almost accurate string comparison.

- Corpus-Based measure: This is a similarity metric to decide commonalities between given character arrays in accordance with the information obtained from large collection or compilation.

- Knowledge-Based measure: It is a similarity metric that evaluates the amount of commonality among a set of given words using information obtained from semantic networks.

Chapter 4

PERFORMANCE ANALYSIS

4.1 Performance Measures

Parameters like “Precision”, “Recall”, and “F-measure” are typically the ones on whose basis efficiency is examined in fundamental WSD approaches.

1. “Precision” denoted by (P) :Proportion of “matched target words based on human decision” to “number of instances responded by the system based on the particular words”.
2. “Recall value” (R) :Proportion of “number of target words for which the answer matches with the human decided answer” to “total number of target words in the dataset”.
3. “F-Measure”: Computed as “(2*P*R / (P+R))” . It is dependent on the computation of rest of the performance parameters .

4.2 Result Analysis

Experiment has been carried out by taking numerous data sets, all of varied types so as to display the predominance of outline projectedby us.

Table 4.1 Lesk Algorithm Results

Test	No. Of sentences	Accuracy (in %)	Precision	Recall Value	F-Measure
1	25	48	0.68	0.41	0.51
2	30	36.67	0.7	0.15	0.25
3	35	40	0.75	0.3	0.43
4	40	45	0.65	0.48	0.55
5	45	48.88	0.7	0.45	0.55

Testing has been executed on large datasets among which a sample is contemplated for showing the comparison results between our proposed approaches. The results of Lesk Algorithm are shown in Table 4.1 and results of Modified Lesk Algorithm are shown in Table 4.2. Precision value is the most reliable parameter in this type of disambiguation tests. So we chiefly focus on the precision value since it is most accountable.

Table 4.2 Modified Lesk Algorithm Results

Test	No. Of sentences	Accuracy (in %)	Precision	Recall Value	F-Measure
1	25	72	0.7	0.48	0.57
2	30	76.67	0.71	0.46	0.56
3	35	74.28	0.74	0.49	0.59
4	40	75	0.71	0.48	0.57
5	45	75.55	0.83	0.48	0.61

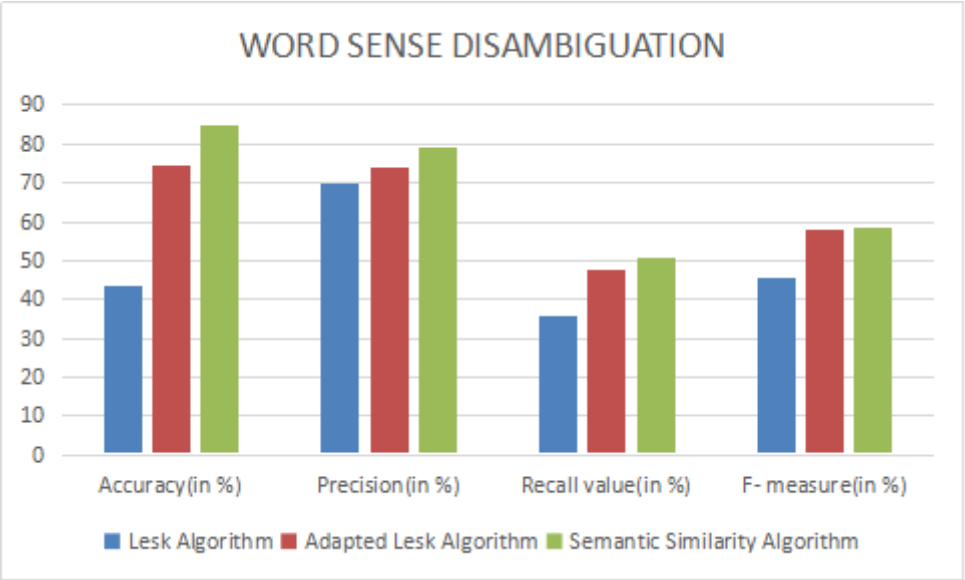
Table 4.3 Semantic Similarity Algorithm Results

Test	No. Of sentences	Accuracy (in %)	Precision	Recall Value	F-Measure
1	25	82	0.8	0.68	0.67
2	30	86.67	0.81	0.66	0.66
3	35	84.28	0.84	0.69	0.69
4	40	85	0.81	068	0.67
5	45	85.55	0.83	0.68	0.71

Average of 5 Test Cases:

Table 4.3 Comparison of Results

Algorithm	Accuracy (in %)	Precision (in %)	Recall Value (in %)	F-Measure (in %)
Lesk Algorithm	43.71	69.60	35.80	45.74
Modified Lesk Algorithm	74.7	73.80	47.80	57.97
Semantic Similarity Algorithm	84.7	79.20	50.6	58.67



Graph 4.1 Comparison of Algorithm

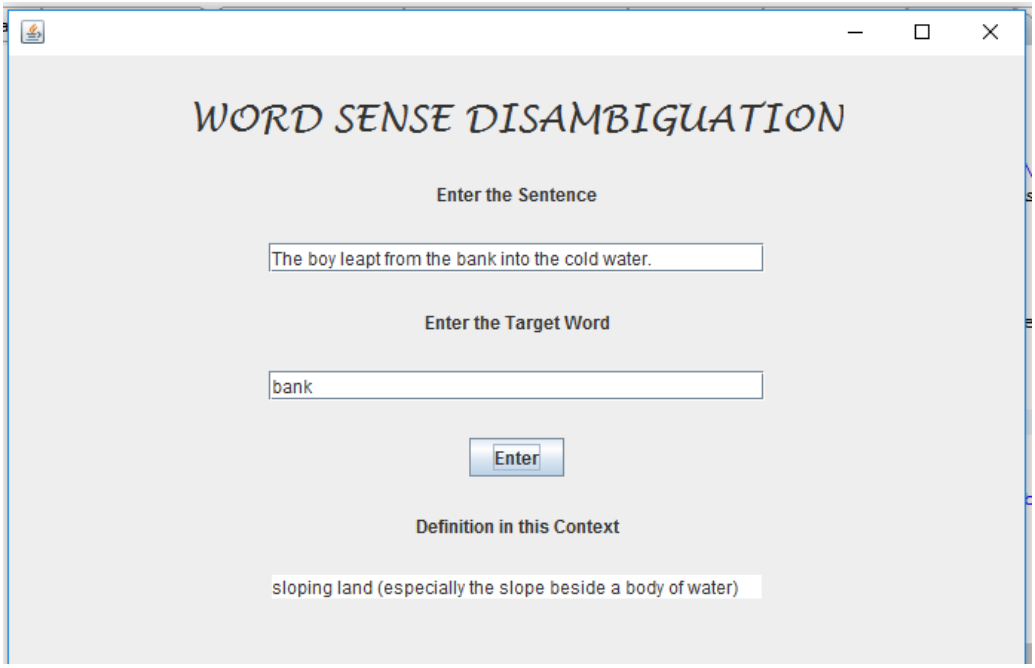
4.3 Output Screenshots

Case 1: Target Word – bank

a) The boy leapt from the bank into the cold water.

b) The bank cashed my check.

Screen 1



Screen 2

Case 2: Target Word - bar

c) The bar was crowded.

d) She washed clothes with soap bar.

Screen 3

Screen 4

Case 3: Target Word - key

e) Heput the key inside and the door got bolted.

f) Some students cheated by using answer key.

Screen 5

Screen 6

Chapter 5
CONCLUSION

5.1 Conclusion

In this project the knowledge based approach is used for identifying the ambiguated word in English Wordnet. Word sense disambiguation is one of the applications of NLP. A database acquired from indowordnet created of those words which have more than one meaning. With the help of Wordnet the meaning of the word can be identified and the id, synonyms, hypernymy and hyponymy, modifies noun can also be identified.

The WordNet arranges words in lexical database in order of their implications of the words rather than their forms as in lexicons. Things, verbs, modifiers and intensifiers are clubbed together into equivalent word sets, where each set is passing on an alternate idea. The words in an equivalent word set can be reciprocally utilized as a part of numerous specific situations. The primary relationship among the words in WordNet is the equivalent word. The explanation behind this could be summed up to the fact that WordNet is designed for broad utility in NLP errands however not engaged with WSD. WordNet is brought into use in diverse WSD approaches so as to draw the correct sense and meaning from a word which has got multiple sense to itself.

In this project we proposed the Lesk , Modified Lesk and semantic similarity methods to get the correct sense of a target word. The

algorithm calculatesvalue for all senses of mark word. To find the score for a sense of the target word the target sense is contrastedwith the senses of the context word. Algorithm finds the sense of each context word which has maximum similarity to the markword. Relatedness scores are calculated between the target sense and each most related context sense. Sum of these relatedness scores is computed and assigned to each target sense. Once the algorithm finds value for all senses of glossary of mark word, the meaning with maximumvalue is designated to mark word.

5.2 Future Scope

- We might want to examine and assess already available regulated, unsupervised and dictionary reference based methods to deal with weaknesses of the current frameworks.
- Change in prior proposed arrangements, if possible.
- We might likewise want to examine profound methodologies in detail and incorporate various types of data, i.e. combination of the nearby or syntactic components alongside heterogeneous data from learning bases.
- Scrutinize different methodologies.

5.3 Applications of Word Sense Disambiguation

- **Information Retrieval**

- This is worried with recognizing records important to client's question. NLP methods have discovered helpful applications in data recovery. Ordering , word sense disambiguation, question change, and learning bases have likewise been utilized as a part of IR framework to improve execution, e.g., by giving techniques for inquiry extension. WordNet, LDOCE (Longman Lexicon of Contemporary English) and Roget's Thesaurus are a portion of the valuable lexical assets of IR research.WSD helps in enhancing term ordering in data recovery. It has demonstrated that word senses enhance recovery execution if the senses are incorporated as record terms. Words alone ought not be the premise of positioning records, additionally word senses or possibly mix of word senses and words ought to be contemplated.

- **Machine Translation**

This alludes to programmed interpretation of content starting with one human dialect then onto the next. So as to complete this interpretation, it is important to have a comprehension of words and expressions, punctuations of the two dialects included, semantics of the dialects, and word information.It increases ability to to create more new sentences in new language and improvises the learning of some new language. Lexical choice is also altered depending upon the usage context. Example: Get “bill” converted from Dutch to Spanish

- A “borrow” or “hole”?

- **Speech Processing and Part of Speech tagging:**

Sometimes words have dissimilarspellingsfrom each other but then when a they get pronounced in human speech, sound exactly the same.

For example:

“base” , “bass”

- **Text Processing**

Words sometimes have same spelling but when they are pronounced the human then they have different meaning because the context in which they are being used is totally different from each other. Text processing requires WSD .

- **Question Answering:**

Given a query and an arrangement of reports, a query solving framework endeavors to locate the exact answer, or if nothing else the exact part of content in which an answer shows up. This is very different from an IR system, which comes up with an entire file quite in relevance with the query. A question answering system is entirely dissimilar since in this system we don’t know what content is to be retrieved. The outcome is not known that is not predicted. In non specific, a question noting framework assistants from having a data extraction framework to perceive substances in the content. Example: Which color is present over the dress?

- Yellow or Blue?

- **Knowledge Acquisition:**

- Out of these diverse areas of research information extraction and data mining is one of the most interesting and crucial fields. So a well done and exact analysis of textual data is a prerequisite for progressing in any sort of research. Taking up an example of this, the intelligence department of any country must not confuse between the spies and the undercover agents of their own country. There should be accurate judgement of identity of people residing in the country so as to

safeguard it against the outsider entities.Any new knowledge if acquired should be high on accuracy and free from any sort of ambiguity so as to exploit maximum advantage of it else it could lead to destruction also. WSD has juststartedbeingpracticed in these areas.

References

[1] "Natural language processing", En.wikipedia.org, 2016. [Online]. Available: http://en.wikipedia.org/wiki/Natural_language_processing. [Accessed: 14- Dec- 2016].

[2] Hindi Word Sense Disambiguation Manish Sinha Mahesh Kumar Reddy .R Pushpak Bhattacharyya Prabhakar Pandey Laxmi Kashyap

[3] ROBERTO NAVIGLI, Universit `a di Roma La SapienzaWord Sense Disambiguation: A Survey

[4] Shallu, Vishal Gupta, Journal of emerging technologies in web intelligence, Shallu, Vishal Gupta, University of Institute of Engineering & Technology, Punjab University, Chandigarh, India

[5] Arindam Roy1,Sunita Sarkar2and Bipul Syam Purkayastha3.International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2011 10.5121/ijnlc.2014.3305 51 Knowledge Based Approaches to Nepali WSD

[6] Andres Montoyo, Armando Su´arez, German Rigau, and Manuel Palomar, Journal of Artificial Intelligence Research 23 (2005) 299-330 Submitted 07/04; published 03/05: Combining Knowledge- and Corpus-based WSD Methods

[7] Alok Ranjan Pal, Diganta Saha: Word Sense Disambiguation: A Suevey, International Journal of Control Theory and Computer Modeling Vol.5, No.3, July 2015

[8] Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar, Kunal Sinha: An Approach To WSD Combining Modified Lesk And Bag-Of- Words

[9] Abhishek Fulmari1 , Manoj B. Chandak2, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013 Copyright to IJARCCE www.ijarcce.com 4667: A Survey on Supervised Learning for Word Sense Disambiguation

[10] Xiaobin Li " Stan Szpakowicz and Stan Matwin.A WordNet-based Algorithm for Word Sense Disambiguation .