A

PROJECT REPORT

ON

AN ADAPTED LESK ALGORITHM FOR WORD SENSE DISAMBIGUATION USING HINDI WORDNET

Submitted in partial fulfillment for the requirement of the award of
DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE
FROM ASSAM UNIVERSITY SILCHAR

SUBMITTED BY
ARPITA MITRA MAZUMDER
EXAM ROLL: 101611 NO.:02220107
REGN NO.: 22-110021738 OF 2011-12
SEMESTER: 10^{TH}, M.SC. (5 YRS)
SUB CODE: MSC 1005(C)

UNDER THE GUIDANCE OF
PROF. BIPUL SYAM PURKYASTHA
HEAD OF DEPARTMENT, PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE
ASSAM UNIVERSITY, SILCHAR

CERTIFICATE

This is to certify that Arpita Mitra Mazumder bearing Roll: 101611 No: 02220107 has carried out her work for the project entitled “An Adapted Lesk Algorithm for Word Sense Disambiguation Using Hindi Wordnet” under my supervision in partial fulfillment for the requirement of the award of degree of Master of Science in Computer Science of Assam University, Silchar. She has done sincerely her work for preparing this project. She has fulfilled all the requirements laid down in the regulations of the MSc (5 years) 10th Semester Examination (Paper MSC-1005) of the Department of Computer Science, Assam University, Silchar, for the session 2015-2016.

Date: **Signature of the Guide**

Place: **(PROF. BIPUL SYAM PURKAYASTHA)**
Supervisior, Professor
Department of Computer Science
Assam University, Silchar

CERTIFICATE

This is to certify that Arpita Mitra Mazumder bearing Roll: 101611 No: 02220107 has carried out her work for the project entitled "An Adapted Lesk Algorithm for Word Sense Disambiguation Using Hindi Wordnet" under my supervision in partial fulfillment for the requirement of the award of degree of Master of Science in Computer Science of Assam University, Silchar. She has done sincerely her work for preparing this project. She has fulfilled all the requirements laid down in the regulations of the MSc (5 years) 10th Semester Examination (Paper MSC-1005) of the Department of Computer Science, Assam University, Silchar, for the session 2015-2016.

Date: **Signature of the HOD**

Place: **(PROF. BIPUL SYAM PURKAYASTHA)**
HOD, Professor
Department of Computer Science
Assam University, Silchar

DECLARATION

I, Arpita Mitra Mazumder, student of 10th semester (MSc 5 years), Department of Computer Science do hereby solemnly declare that I have duly worked on my project entitled "An Adapted Lesk Algorithm for Word Sense Disambiguation Using Hindi Wordnet" under the supervision of Prof. Bipul Syam Purkayastha, Professor, Department of Computer Science, Assam University, Silchar.

Date: **Signature**

Place: **(ARPITA MITRA MAZUMDER)**
Roll: 101611 No.: 02220107
Regn. No.: 22-110021738 of 2011-12
Department of Computer Science
Assam University, Silchar

ACKNOWLWDGEMENT

At the very outset, I take the privilege to convey my gratitude to those persons whose co-operation, suggestions and heartfelt support helped us to accomplish the term paper successfully.

I take immense pleasure to express my sincere thanks and profound gratitude to my respected guide Prof. Bipul Shyam Purkayastha, Head of the Department of Computer Science, Assam University, Silchar, for his excellence and able guidance, valuable suggestions and encouragement he rendered for completing the term paper and also for his valuable suggestions.

I also indebted to my family members, friends and well wishers who encouraged me to do this work with vigor and seriousness.

Last but not the least I would like to acknowledge the cooperation I received from the entire staff of our department and thanks to all those who directly or indirectly extended their helpful hands and moral support while making this project.

(ARPITA MITRA MAZUMDER)

ABSTRACT

All human languages have words that can mean different things in different contexts. Natural Language Processing (NLP) is a field of computer science and linguistics concerned with the interactions between computer and human (natural) languages. Natural language generation systems convert information from computer databases into readable human language.

Word Sense Disambiguation (WSD) is a challenging technique in Natural Language Processing. Word sense disambiguation is the process of automatically figuring out the intended meaning of such a word when used in a sentence Since, there are some words in the natural languages which can cause ambiguity about the sense of the word.WSD is defined as the task of finding the correct sense of a word in a specific context. This is crucial for applications like Machine Translation and Information Extraction.

Michael Lesk’s 1986 algorithm is based on two assumptions. First, when two words are used in close proximity in a sentence, they must be talking of a related topic and second, if one sense each of the two words can be used to talk of the same topic, then their dictionary definitions must use some common words. Thus we can disambiguate neighbouring words in a sentence by comparing their definitions and picking those senses whose definitions have the most number of common words.

The biggest drawback of simple lesk algorithm is that dictionary definitions are often very short and just donot have enough words for this algorithm to work well. We deal with this problem by adapting this algorithm to the semantically organized lexical database called WordNet. Besides storing words and their meaning like a normal dictionary, WordNet also”connects” related words together. We overcome the problem of short definitions by looking for common words not only between the definitions of the words being disambiguated, but also between the definitions of words that are closely related to them in WordNet.

TABLE OF CONTENTS

Chapters Title Page no.

CHAPTER-1 INTRODUCTION

1.1. NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation, so it process information contained in natural language text.

Also known as Computational Linguistics (CL), Human Language Technology and Natural Language Engineering (NLE).There are a number of factors that make NLP difficult.

Natural language processing sometimes mistakenly termed as natural language understanding-originated from machine translation research. While natural language understanding involves only the interpretation of language, natural language processing includes both understanding (interpretation) and generation (production). The NLP also includes speech processing.

Computational linguistics is similar to theoretical and psycho linguistics, but uses different tools. Theoretical linguists mainly provide structural description of natural language and its semantics. They are not concerned with the actual processing of sentences or generation of sentences from structural description. They are in a quest for principals that remain common across languages and identify rules that capture linguistic generalization. For example, most languages have constructs like noun and verb phrases. Theoretical linguists identify rules that describe and restrict the structure of languages. Psycholinguistics explains how humans produce and comprehend natural language. Unlike theoretical linguists, they are interested in the representation of linguistic structures as well as in the process by which these structures are produced. They rely primarily on empirical investigations to back up their theories.

Computational linguistics is concerned with the study of language using computational model of linguistic phenomena. It deals with the application of linguistic theories and computational techniques for NLP. In computational linguistics, representing a language is a major problem; most knowledge representations tackle only a small part of knowledge. Representing the whole body of knowledge is almost impossible. The words knowledge and language should not be confused.

Computational models may be broadly classified under knowledge-driven and data-driven categories. Knowledge-driven systems rely on explicitly coded linguistic knowledge, often expressed as a set of handcrafted grammar rule. Acquiring and encoding such knowledge is difficult and is main bottleneck in the development of such system. They are, therefore, often constrained by the lack of sufficient coverage of domain knowledge. Data-driven approaches presume the existence of a large amount of data and usually employ some machine learning technique to learn syntactic patterns. The amount of human effort is less and the performance of theses system dependent on the quantity of the data. These systems are usually adaptive to noisy data.

The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system, in order to determine whether (or to what extent) the system answers the goals of its designers, or meets the needs of its users. Research in NLP evaluation has received considerable attention, because the definition of proper evaluation criteria is one way to specify precisely an NLP problem, going thus beyond the vagueness of tasks defined only as language understanding or language generation. A precise set of evaluation criteria, which includes mainly evaluation data and evaluation metrics, enables several teams to compare their solutions to a given NLP problem.

The first evaluation campaign on written texts seems to be a campaign dedicated to message understanding in 1987 (Pallet 1998). Then, the Parseval/GEIG project compared phrase-structure grammars (Black 1991). A series of campaigns within Tipster project were realized on tasks like summarization, translation and searching (Hirschman 1998). In 1994, in Germany, the Morpholympics compared German taggers. Then, the Senseval & Romanseval campaigns were conducted with the objectives of semantic disambiguation. In 1996, the Sparkle campaign compared syntactic parsers in four different languages (English, French, German and Italian). In France, the Grace project compared a set of 21 taggers for French in 1997 (Adda 1999). In 2004, during the Technolangue/Easy project, 13 parsers for French were compared. Large-scale evaluation of dependency parsers were performed in the context of the CoNLL shared tasks in 2006 and 2007. In Italy, the EVALITA campaign was conducted in 2007 and 2009 to compare various NLP and speech tools for Italian; the 2011 campaign is in full progress - EVALITA web site. In France, within the ANR-Passage project (end of 2007), 10 parsers for French were compared – passage web site.[1]

1.2. APPLICATIONS OF NLP

The following applications NLP are utilizing:

a) **Machine Translation:**

This refers to automatic translation of text from one human language to another. In order to carry out this translation, it is necessary to have an understanding of words and phrases, grammars of the two languages involved, semantics of the languages, and word knowledge.

b) **Speech recognition:**

Speech synthesis refers to automatic production of speech (utterance of natural language sentence). Such systems can read out your mail on telephone, or even read out a storybook for you. In order to generate utterances, text has to be processed. So, NLP remains an important component of any speech synthesis system.

c) **Parsing :**

Determine the parse tree of a given sentence. The grammar of natural languages is ambiguous and typical sentences have multiple possible analyses. In fact, perhaps surprisingly, for a typical sentence there may be thousands of potential parses.

d) **Speech Synthesis:**

Speech synthesis refers to automatic production of speech. Such systems can read out mails on telephone, or even read a storybook. In order to generate utterance, text has to be processed.

e) **Natural Language Interfaces to Database:**

Natural Language Interfaces allow querying a structured database using natural language sentences.

f) **Information Retrieval:**

This is concerned with identifying documents relevant to user’s query. NLP techniques have found useful applications in information retrieval. Indexing (stop word elimination, stemming, phrase extraction, etc.), word sense disambiguation, query modification, and knowledge bases have also been used in IR system to enhance performance, e.g., by providing methods of query expansion. WordNet, LDOCE (Longman Dictionary of Contemporary English) and Roget"s Thesaurus are some of the useful lexical resources of IR research.

g) **Information Extraction:**

An information extraction system captures and outputs factual information contained within a document. Similar to an information retrieval system, it responds to a users information need. However, unlike in an information retrieval system, the information need is not expressed as a keyword query. Instead it is specified as a predefined database schema or templates. Whereas an IR system identifies a subset of documents in a large repository of text database – e.g. in a library scenario a subset of resources in a library an information extraction system identifies a subset of information within a document that fits the pre defined template.

h) **Question Answering:**

Given a question and a set of documents, a question answering system attempts to find the precise answer, or at least the precise portion of text in which a answer appears. This is unlike an IR system, which returns whole document that seems relevant to the users query. A question answering system is different from an information extraction system in that the content that is to be extracted is unknown. In general, a question answering system benefits from having an information extraction system to identify entities in the text.

i) **Text Summarization:**

This deals with the creation of summaries of documents and involves syntactic, semantic, and discourse level processing of text.

j) **Part-of-speech tagging:**

Given a sentence, determine the part of speech for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb or adjective; and "out" can be any of at least five different parts of speech. Some languages have more such ambiguity than others.

k) **Relationship extraction:**

Given a chunk of text, identify the relationships among named entities (e.g. who is the wife of whom).

l) **Word segmentation:**

Separate a chunk of continuous text into separate words. For a language like English, this is fairly trivial, since words are usually separated by spaces. However, some written languages like Chinese, Japanese and Thai do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in the language.

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): „The meaning of a word is its use in the language.‟ A language keeps on evolving. New words are added continually and existing words are introduced in new context. For example, most newspapers and TV channels use 9/11 to refer to the terrorist act on the World Trade Centre in USA in 2004. 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

Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner. WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. [3]

In computational linguistics, word-sense disambiguation (WSD) is an open problem of natural language processing and ontology, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings, i.e. is the process of selecting a sense for a word from a set of predefined possibilities. [2]

i. Sense Inventory usually comes from a dictionary or thesaurus.

ii. Knowledge intensive methods, supervised learning, and

(sometimes) bootstrapping approaches.

The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, anaphora resolution, coherence, inference et cetera.

Research has progressed steadily to a point where WSD systems achieve sufficiently high levels of accuracy on a variety of word types and ambiguities. A rich variety of techniques have been researched, from dictionary-based methods that use the knowledge encoded in lexical resources, to supervised machine learning methods in which a classifier is trained for each distinct word on a corpus of manually sense-annotated examples, to completely unsupervised methods that cluster occurrences of words, thereby inducing word senses. Among these, supervised learning approaches have been the most successful algorithms to date, e. g,

What sorts of plants thrive in chalky soil?

Plant

Plant

Current accuracy is difficult to state without a host of caveats. In English, accuracy at the coarse-grained (homograph) level is routinely above 90%, with some methods on particular homographs achieving over 96%. On finer-grained sense distinctions, top accuracies from 59.1% to 69.0% have been reported in recent evaluation exercises (SemEval-2007, Senseval-2), where the baseline accuracy of the simplest possible algorithm of always choosing the most frequent sense was 51.4% and 57%, respectively.

A disambiguation process requires two strict things: a dictionary to specify the senses which are to be disambiguated and a corpus of language data to be disambiguated (in some methods, a training corpus of language examples is also required). WSD task has two variants: "lexical sample" and "all words" task. The former comprises disambiguating the occurrences of a small sample of target words which were previously selected, while in the latter all the words in a piece of running text need to be disambiguated. The latter is deemed a more realistic form of evaluation, but the corpus is more expensive to produce because human annotators have to read the definitions for each word in the sequence every time they need to make a tagging judgment, rather than once for a block of instances for the same target word.

To give a hint how all this works, consider two examples of the distinct senses that exist for the (written) word "bass":

1. A type of fish
2. Tones of low frequency

And the sentences:

1. I went fishing for some sea bass.
2. The bass line of the song is too weak.

To a human, it is obvious that the first sentence is using the word "bass (fish)", as in the former sense above and in the second sentence, the word "bass (instrument)" is being used as in the latter sense below. Developing algorithm to replicate this human ability can often be a difficult task.

1.5. APPLICATION OF WSD

a) Machine Translation:

This refers to automatic translation of text from one human language to another. In order to carry out this translation, it is necessary to have an understanding of words and phrases, grammars of the two languages involved, semantics of the languages, and word knowledge. It helps in better understanding of source language and generation of sentences in target language. It also affects lexical choice depending upon the usage context. Such as: Translate “bill” from English to Hindi

- Is it a “borrow” or a “hole”?
- Is it a bird jaw or an invoice?

b) Information Retrieval:

This is concerned with identifying documents relevant to user’s query. NLP techniques have found useful applications in information retrieval. Indexing (stop word elimination, stemming, phrase extraction, etc.), word sense disambiguation, query modification, and knowledge bases have also been used in IR system to enhance performance, e.g., by providing methods of query expansion. WordNet, LDOCE (Longman Dictionary of Contemporary English) and Roget’s Thesaurus are some of the useful lexical resources of IR research.WSD helps in improving term indexing in information retrieval. It has proved that word senses improve retrieval performance if the senses are included as index terms. Documents should not be ranked based on words alone, the documents should be ranked based on word senses, or based on a combination of word senses and words. Example: Find all Web Pages about “cricket”

- The sport or the insect?

c) Speech Processing and Part of Speech tagging:

Speech recognition i.e. when processing homophones words which are spelled differently but pronounced the same way. For example: “base” and “bass” or “sealing” and “ceiling”.

d) Text Processing:

Text to Speech translation i.e. when words are pronounced in more than one way depending on their meaning. For example: “lead” can be “in front of” or “type of metal”.

e) Question Answering:

Given a question and a set of documents, a question answering system attempts to find the precise answer, or at least the precise portion of text in which a answer appears. This is unlike an IR system, which returns whole document that seems relevant to the users query. A question answering system is different from an information extraction system in that the content that is to be extracted is unknown. In general, a question answering system benefits from having an information extraction system to identify entities in the text. Example: What is George Miller’s position on gun control?

- The psychologist or US congressman?

f) Knowledge Acquisition:

In information extraction and text mining, WSD is required for the accurate analysis of text in many applications. For instance, an intelligence gathering system might need to flag up references to, say, illegal drugs, rather than medical drugs. Bioinformatics research requires the relationships between genes and gene products to be catalogued from the vast scientific literature; however, genes and their proteins often have the same name. More generally, the Semantic Web requires automatic annotation of documents according to reference ontology. WSD is only beginning to be applied in these areas. Example: Add to KB: Herb Bergson is the mayor of Duluth

- Minnesota or Georgia

1.6. ORGANIZATION OF THE REPORT

The rest of the report is organized into chapters as follows:-

Chapter 2 provides problem statement that is, it states about ambiguity, motivation and objective for doing this project.

Chapter 3 provides a review of the prior work (Literature Survey) in process word sense disambiguation. We do not aim to give a comprehensive review of the related work. Such an attempt is extremely difficult due to the large number of publication in this area. So, we provide brief a review of history. Also gives overview of the approaches of WSD which include main and conventional approaches. These approaches are further classified.

Chapter 4 presents brief description of wordnet and an adapted algorithm which we have being using for WSD. The algorithm helps to disambiguate the senses of the word having different meaning.

Chapter 5 provides the implementation that is flowchart, peusdo code and the analysis of the results of adapted algorithms with its accuracy and also showing its graph.

Chapter 6 provides general conclusion, and outline several direction for future work.

CHAPTER-2
PROBLEM STATEMENT

2.1. AMBIGUITY

Ambiguity is an attribute of any concept, idea, and statement or claims whose meaning, intention or interpretation cannot be definitively resolved according to a rule or process consisting of a finite number of steps. It determines which sense of a word is used in a specific sentence as most words in natural languages have multiple possible meanings (homonymy / polysemy).

Example: (i) “pen” (noun)

- The dog is in the pen.
- The ink is in the pen.

(ii) “Take” (verb)

- Take one pill every morning.
- Take the first right past the stoplight.

Syntax helps distinguish meanings for different parts of speech of an ambiguous word. Example: “conduct” (noun or verb)

- John’s conduct in class is unacceptable.
- John will conduct the orchestra on Thursday

AMBIGUITY FOR HUMANS AND COMPUTERS

Computers versus Humans:

a) Polysemy – from Greek poly-, "many" and sêma, "sign". It is the capacity for a word has multiple meanings. It is thus usually regarded as distinct from homonymy.

b) Homonymy–In linguistics, a homonym is a word that has different meanings. In the strict sense, one of a group of words that share the same spelling and pronunciation but have different meanings

c) A computer program has no basis for knowing which one is appropriate, even if it is obvious to a human…

d) Ambiguity is rarely a problem for humans in their day to day communication, except in extreme cases…

AMBIGUITY FOR HUMANS - NEWSPAPER HEADLINES!

a) DRUNK GETS NINE YEARS IN VIOLIN CASE

b) FARMER BILL DIES IN HOUSE

c) STOLEN PAINTING FOUND BY TREE

d) INCLUDE CHILDREN WHEN BAKING COOKIES

e) MINERS REFUSE TO WORK AFTER DEATH

AMBIGUITY FOR A COMPUTER

a) The fisherman jumped off the bank and into the water.

b) The bank down the street was robbed!

c) Back in the day, we had an entire bank of computers devoted to this problem.

All Words WSD:

- Attempt to disambiguate all open-class words in a text
  - "He put his suit over the back of the chair"
- Use information from dictionaries
  - Definitions / Examples for each meaning
    - Find similarity between definitions and current context.
- Position in a semantic network
  - Find that "table" is closer to "chair/furniture" than to "chair/person"
- Use discourse properties
  - A word exhibits the same sense in a discourse / in a collocation.

Evaluating WSD:

- Difficulty in evaluation: Nature of the senses to distinguish has a huge impact on results.Coarse versus fine-grained sense distinction

  (i)chair = a seat for one person, with a support for the back; "he put his coat over the back of the chair and sat down"

  chair = the position of professor; "he was awarded an endowed chair in economics"

  (ii)bank = a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home"

  bank = a building in which commercial banking is transacted; "the bank is on the corner of Nassau and Witherspoon"

- Sense maps: Cluster similar senses and allow for both fine-grained and coarse-grained evaluation

2.2. MOTIVATION

It is one of the central challenges and extremely important problem in NLP, since human/natural language is ambiguous, so many words can be interpreted in multiple ways depending on the context in which they occur. Human beings can comprehend sentences and text inspite of multiple meanings of words and they can identify correct meaning form the context and from the way text is used but computers cannot. So, many tasks in natural language processing require disambiguation of ambiguous words:

1.1 Question answering
1.2 Information retrieval
1.3 Machine translation
1.4 Text mining
1.5 Part Of Speech tagging

1.6 Opinion mining/sentiment analysis

1.7 Phone help systems

While most of the time humans do not even think about the ambiguities of language, machines need to process unstructured textual information and transform them into data structures which must be analyzed in order to determine the underlying meaning.

Kernel Methods for Word Sense Disambiguation in the Medical Domain, while supervised machine learning methods represent the state-of-the-art in WSD, they face the problem of the so-called "Knowledge Acquisition Bottleneck." It also empirically establishes for WSD the fact that the availability of more training data for a learning algorithm leads to better accuracy on novel instances. This clearly shows the "data-hungry" nature of the supervised machine learning algorithms. In order to be able to build classifier models that are highly accurate as well as generalize well (i.e., have low error on unseen data instances), the training data size should be sufficiently large to allow the algorithms to acquire knowledge of the majority of the disambiguating aspects of a large number of ambiguous words and their senses.

With the advent of the World Wide Web and the exponential increase in digital storage of content such as newspaper articles, scholarly publications an medical records of patients, there is comparatively more data in a raw form than labeled data.

So, understanding how people disambiguate words is an interesting problem that can provide insight in psycholinguistics.

2.3. OBJECTIVE

The present study has been undertaken with the following objectives:

- To study Word Sense Disambiguation (WSD) an open problem of Natural Language Processing (NLP).
- To perform implementation Word Sense Disambiguation using the algorithm.
- To identify the most proper sense of the ambiguous word(s).

CHAPTER-3
LITERATURE SURVEY

3.1. BRIEF HISTORY

WSD was first formulated into as a distinct computational task during the early days of machine translation in the 1940s, making it one of the oldest problems in computational linguistics. Warren Weaver, in his famous 1949 memorandum on translation, first introduced the problem in a computational context, i.e. A word can often only be translated if you know the specific sense intended (A bill in English could be a pico or a cuenta in Spanish) .

In philosophy of language field, Wittgenstein (1958) gave meaning as use, i.e., “For a large class of cases-though not for all-in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language.”

In fact, Bar-Hillel (1960) used the above example to argue that WSD could not be solved by "electronic computer" because of the need in general to model all world knowledge. He posed the following:

a) Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.

b) Is “pen” a writing instrument or an enclosure where children play?

In the 1970s, WSD was a subtask of semantic interpretation systems developed within the field of artificial intelligence, starting with Wilk’s preference semantics. However, since WSD systems were at the time largely rule-based and hand-coded they were prone to a knowledge acquisition bottleneck.

By the 1980s large-scale lexical resources, such as the Oxford Advanced Learner's Dictionary of Current English (OALD), became available: hand-coding was replaced with knowledge automatically extracted from these resources, but disambiguation was still knowledge-based or dictionary-based.

In the 1990s, the statistical revolution swept through computational linguistics, and WSD became a paradigm problem on which to apply supervised machine learning techniques. Corpus based approaches and dependence on sense tagged text involved also (Ide and Veronis, 1998) overview history from early days to 1998.[12]

The 2000s saw supervised techniques reach a plateau in accuracy, and so attention has shifted to coarser-grained senses, domain adaptation, semi-supervised and unsupervised corpus-based systems, hybrid systems, and the return of knowledge-based systems via graph-based methods. Still, supervised systems continue to perform best. Minimizing or eliminating use of sense tagged text and taking advantage of the Web.

3.2 PREVIOUS WORKS ON WSD

Manish Sinha, Mahesh Kumar Reddy .R, Pushpak Bhattacharyya , Prabhakar Pandey Laxmi Kashyap [7]: From Department of Computer Science and Engineering Indian Institute of Technology Bombay, Mumbai India.They used the Hindi Wordnet for a fundamental NLP task, *viz.,* disambiguation of Hindi words. To our knowledge, this is the first attempt at automatic WSD for an Indian language and is a significant step towards Indian language processing. The performance can surely be improved if morphology is handled exhaustively. The system currently does not detect the underlying similarity in presence of morphological variations. Since Indian languages are rich in morphology, exhaustive pre-processing for morphology is crucial in the whole WSD process.

Shallu, Vishal Gupta [13]: 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. HyperLex given by the author provides a tool for domain and lexicon navigation. The results of the HyperLex algorithm were evaluated on the Web page corpus. The best 25 contexts were checked for each of the 50 uses which include 1245 contexts in all.

Arindam Roy1, Sunita Sarkar2 and Bipul Syam Purkayastha3[11] : From the experimental result it is seen that the performance of overlap based approach is less than the combination of conceptual distance and semantic graph method. It is expected because overlap based approach suffer from sparse overlap. especially in the case of nouns, the overlap based approach presented here gives better performance than the overlap based approach with machine readable dictionaries because not only the gloss and examples of the target and context synsets are taken but also the gloss and examples from their hypernyms have been taken into consideration. The adjective accuracy is more with the second method as the semantic graph distance score has been taken into account.

Alok Ranjan Pal, Anirban Kundu, Abhay Singh, Raj Shekhar,Kunal Sinha[21]: Their approach has established better performance in enhanced WSD technique depending on specific learning sets. The disambiguation accuracy is improved based on the enrichment of datasets having populated by new data. We have achieved better precision value, recall value, and F-Measure through extensive experimentation.

Andres Montoyo, Armando Su´arez, German Rigau, and Manuel Palomar[6]: They had done the hypothesis work in that WSD requires different kinds of knowledge sources (linguistic information, statistical information, structural information, etc.)& techniques. The aim of their work to explore some methods of collaboration between complementary knowledge-based and corpus-based WSD methods.

In order to demonstrate our hypothesis, three different schemes for combining both approaches have been presented. They presented different mechanisms of combining information sources around knowledge-based and corpus-based WSD methods. Finally, shown that a knowledge-based method can help a corpus-based method to better perform the disambiguation process, and vice versa.

Ehsan Hessami, Faribourz Mahmoudi and Amir Hossien Jadidinejad [22]: This paper has proposed a new method for word sense disambiguation. First builds a tree for some of the senses of ambiguity words which there are in the sentence and detects the best path. Then with these paths builds a Graph and uses the connectivity measure for choosing the best sense of word

Simone Paolo Ponzetto, Roberto Navigli[3]: On that paper, they have presented a large-scale method for the automatic enrichment of a computational lexicon with encyclopedic relational knowledge. The experiments show that the large amount of knowledge injected into WordNet is of high quality and, more importantly, it enables simple knowledge-based WSD systems to perform as well as the highest-performing supervised ones in a coarse-grained setting and to outperform them on domain-specific text.

Rigau, 2006; Agirre et al., 2009; Navigli and Lapata, 2010[17] and prove that knowledge-rich disambiguation is a competitive alternative to supervised systems, even when relying on a simple algorithm. They note, however, that the present contribution does not show which knowledge-rich algorithm performs best with WordNet++.

Egoitz Laparra and German Rigau[24]: The method uses a knowledge based Word Sense Disambiguation (WSD) algorithm called SSI-Dijkstra for assigning the appropriate synset of WordNet to the semantically related Lexical Units of a given frame from FrameNet. This algorithm relies on the use of a large knowledge base derived from WordNet and eXtended WordNet. Since the original SSI-Dijkstra requires a set of monosemous or already interpreted words, developed and empirically tested four different versions of this algorithm to deal with sets having only polysemous words. The resulting new algorithms obtain improved results over state-of-the-art.

Abhishek Fulmari[1], Manoj B. Chandak[2] [23]: This paper summarized the various approaches used for Word Sense Disambiguation. The hardness of WSD strictly depends on the granularity of the sense distinctions. Here that supervised methods perform well as compared to all other approaches. The reason behind that in supervised approach the training data is totally of domain specific and in unsupervised approach the training data is totally unannotated and to from cluster from that data set is quite difficult.

3.3. MAIN APPROACHES

Two main approaches which are used to WSD are---Deep approaches and shallow approaches. Deep approaches uses some kind of knowledge related to the word and shallow approaches see the context in which the word has been used [8]. The conceptual Model for Word Sense Disambiguation is given below:

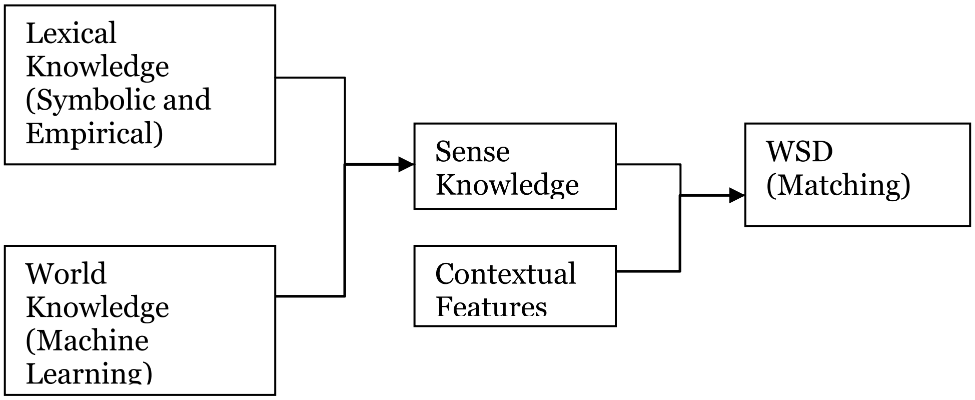

Lexical Knowledge (Symbolic and Empirical)
World Knowledge (Machine Learning)
Sense Knowledge
Contextual Features
WSD (Matching)

Fig: Conceptual model of WSD [5]

Deep approaches presume access to a comprehensive body of world knowledge. Knowledge, such as "you can go fishing for a type of fish, but not for low frequency sounds" and "songs have low frequency sounds as parts, but not types of fish", is then used to determine in which sense the word bass is used. These approaches are not very successful in practice, mainly because such a body of knowledge does not exist in a computer-readable format, outside very limited domains.[18]However, if such knowledge did exist, then deep approaches would be much more accurate than the shallow approaches. Also, there is a long tradition in computational linguistics, of trying such approaches in terms of coded knowledge and in some cases; it is hard to say clearly whether the knowledge involved is linguistic or world knowledge. The first attempt was that by Margaret Master man and her colleagues, at the Cambridge Language Research Unit in England, in the 1950s. This attempt used as data a punched-card version of Roget's Thesaurus and its numbered "heads", as an indicator of topics and looked for repetitions in text, using a set intersection algorithm. It was not very successful, but had strong relationships to later work, especially Yarowsky's machine learning optimization of a thesaurus method in the 1990s.

Shallow approaches don't try to understand the text. They just consider the surrounding words, using information such as "if bass has words sea or fishing nearby, it probably is in the fish sense; if bass has the words music or song nearby, it is probably in the music sense." These rules can be automatically derived by the computer, using a training corpus of words tagged with their word senses. This approach, while theoretically not as powerful as deep approaches, gives superior results in practice, due to the computer's limited world knowledge. However, it can be confused by sentences like The dogs bark at the tree which contains the word bark near both tree and dogs.

3.4. CONVENTIONAL APPROCHES

Major WSD approaches proposed till date can be broadly classified as Knowledge Based Approaches and Machine Learning Based Approaches.

Knowledge-Based or Dictionary Based Approach:

In this style of approach the dictionary provides both the means of constructing a sense tagger and target senses to be used. An attempt to perform large scale disambiguation has lead to the use of Machine Readable Dictionaries (MRD). In this approach, all the senses of a word that need to be disambiguated are retrieved from the dictionary. These senses are then compared to the dictionary definitions of all the remaining words in context. The sense with highest overlap with these context words is chosen as the correct sense. It makes use of external lexical resources such as dictionaries and thesauri; and also the discourse properties. A major drawback of Dictionary based approaches is the problem of scaling. It makes use of external lexical resources such as dictionaries and thesauri; and also the discourse properties [8] .

Task Definition:

- Knowledge-based WSD = class of WSD methods relying (mainly) on knowledge drawn from dictionaries and/or raw text

- Resources
    - Yes
        - Machine Readable Dictionaries
        - Raw corpora
    - No
        - Manually annotated corpora
- Scope
    - All open-class words

Knowledge-based WSD is based on lexical resources like dictionaries, thesauri, and corpora where Machine-readable dictionaries (MRDs) are the primary source of acquisition of data. There are various Knowledge based approaches such as WSD using Selectional Preferences, Lesk's algorithm, Walker's algorithm, WSD using conceptual density and WSD using Random Walk Algorithm.

A. Overlap Based Approach

Overlap based approach calls for the requirement of machine readable dictionary (MDR). It includes determination of the different features of the senses of words which are ambiguous along with features of the words in the context. The word sense having the maximum overlap is selected as the appropriate sense in the context. The commonly used algorithms used in overlap based approach are [10]:

1) WSD using conceptual density: The conceptual density is the measure of how the concept that the word represents is related to the concept of the words in its context. Conceptual density is related to conceptual distance inversely. The conceptual distance is determined from the wordnet.
2) Lesk's algorithm: The Lesk algorithm firstly introduced by the Michael E. Lesk . Lesk algorithm is based on the assumption of given neighbourhood, which gives the common topics. A version of Lesk algorithm is adapted from the WordNet, which

means the word having different senses, in a dictionary and should count the amount of words that are neighbourhood of disambiguated word. Finally the sense is chosen on the basis of highest count number. The Simplified Lesk algorithm, in which each word is taken individually by locating the sense between the numbers of dictionary definitions according to a given context. So, if W is a word creating disambiguation, C be the set of words in the context collection in the surrounding, S be the senses for W, B be the bag of words derived from glosses, synonyms, hyponyms, glosses of hyponyms, example sentences, hypernyms, glosses of hypernyms, meronyms, example sentence of meronyms, example sentence of hypernyms, glosses of meronyms then use the interaction similarity rule to measure the overlap and output the sense which is the most probable having the maximum overlap.

Merits: Provide a simple approach, which is easy to understand and do not need any trained data, Accuracy of the Lesk algorithm is 50 to 70 percent i.e. purely depend on the word.

Demerits: Lesk Algorithm is more sensitive towards the word definition, if word is absent so the result changes rapidly. It is based on the dictionary so if sometime dictionary do not provides the sufficient meaning to match with the fine-grained senses then it creates the problem.

3) Walker's approach: Walker's algorithm can be stated as each word is assigned to one or more categories of subjects in the thesaurus. Different subjects are assigned to different senses of the word.

B. Selectional Preferences

Selectional Preferences approach imposes restrictions on the possibility of the occurrence of number of meanings of the word in the context. The measure of semantic association is provided by the count of number of instances (W1, W2, Y) in the corpus given by pair of words W1 and W2 occurring in the relation Y. The senses that violate the constraint are omitted. The word sense imposes constraints on the semantic type of the words with which it usually combines grammatically. The semantic appropriateness of word to word can be estimated as the conditional probability of the word W1 given the word W2 as follows-

Y: P(W1/W2,Y) = count(W1,W2,Y)/count(W2,Y)

Semi-Supervised Approach:

In semi-supervised learning techniques, the information is present like in supervised but might be less information is given. Here only critic information is available, not the exact information. For example, the system may tell that only particular about of target output is correct and so. The semi-supervised or minimally supervised methods are gaining popularity because of their ability to get by with only a small amount of annotated reference data while often outperforming totally unsupervised methods on large data sets. There are a host of diverse methods and approaches, which learn important characteristics from auxiliary data and cluster or annotate data using the acquired information.

These make use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process, or a word-aligned bilingual corpus.

Bootstrap basically a Semi- supervised learning approach to select an appropriate meaning of a disambiguate word sense in a particular context. In

Bootstrapping Approach by taking a word and try to co-occur with the target word in the given sense then get the target word through the corpus and finally assume the target tag in the correct sense.

Merits: Eliminates the need of large trained data, Sense of the word is more clearly define with high accuracy.

Demerits: It is a repetitive process so the training corpus grows and the untagged instances are reduced, it requires trained data when it works on the labelled data.

Supervised Approach:

The learning here perform in supervision. Let us take the example of the learning process of a small child. The child doesn’t know how to read/write. He/she is being taught by the parents at home and then by their teachers in school. The children are trained and modules to recognize the alphabets, numerals, etc. Their each and every action is supervised by the teacher. Actually, a child works on the basis of the output that he/she has to produce. Similarly, a word sense disambiguation system is learned from a representative set of labelled instances drawn from same distribution as test set to be used. Input instances to these approaches are feature encoded along with their appropriate labels. The output of the system is a classifier system capable of assigning labels to new feature encoded inputs. System is informed precisely about what should be emitted as output. In supervised learning, it is assumed that the correct (target) output values are known for each Input. So, actual output is compared with the target output, if there is a difference, an error signal should be generated by the system. This error signal helps the system to learn and reach to the desired or target output. It is based on a labeled training set. The learning system has a training set of feature-encoded inputs and their appropriate sense label (category).

A. Decision Trees

A decision tree divides the training data in a recursive manner and represents the rules for classification in a tree structure. The internal nodes represent test on the features and each branch shows how the decision is being made and the leaf node refers to the outcome or prediction. It is often regarded as a prediction tool. Some popular algorithms for learning decision trees are ID3 and C4.5. On comparison with other machine learning algorithms it was found that several supervised approaches performed better than the decision tree obtained C4.5 algorithm. If the training data is small in size decision tree suffers from prediction unreliability and decision tree also suffers from sparseness of data when there are features with large number of values.

B. Neural Networks

Neural networks processes information based on computational model of connectionist approach. The input includes the input features and the target output. The training dataset is divided into sets which are non-overlapping based on desired responses. When the network encounters new input pairs the weights are adjusted so that the output unit giving the target output has the larger activation. The network can have weights both positive and negative corresponding to correct or wrong sense choice. Neural networks are trained unless the error between the computed and the target output is minimum. Learning in neural networks is eventually updating of weights.

C. Decision Lists

Decision lists contains ordered set of if-then-else rules for assigning category to the test data. Features are obtained from training data which includes rules in the form of (F, S, Score) where F represents feature value, S represents word sense. Rules are arranged in the list as per descending order of the score. For a word w represented in the form of feature vector, the winning sense for the word is the one whose feature has the maximum score in the decision list in matching the input vector.

D. Naive Bayes

Naive Bayes classifier is the classifier based on Bayes theorem and assumption that every feature is class conditionally independent of every other feature. The conditional probability of each sense of the word Si given the features in the context is calculated to make the final decision. This approach basically works on the trained data as it is the supervised learning method, which is assuming the features of independence. ArgmaxP means their senses over feature vector (senses|feature vector).P(s) defines prior of the sense. P (vj|s) is the conditional probability of any particular feature; both are come from the corpus with the encoded features.

Merits: Simple approach with the trained data, the accuracy of the approaches is 70 to 80 percent.

Demerits: Requires the trained data; Less applicable on the high dimensional data.

Unsupervised Approach:

In unsupervised learning technique, no supervision is provided. Let us consider an example of a tadpole. Learning is done by itself i.e. child fish learn to swim without any supervision. It is not taught by anyone. Thus its leaning process is independent and not supervised by a teacher. Unsupervised approaches to word sense disambiguation eschew the use of sense tagged data of any kind during the training. In this technique, feature vector representations of unlabeled instances are taken as input and are then grouped into clusters according to a similarity metric. These clusters are then labelled by hand with known word senses. Main disadvantage is that senses are not well defined.

It is based on unlabeled corpora. The learning system has a training set of feature-encoded inputs but not their appropriate sense label (category).

It uses Clustering method to select an appropriate sense on the basis of similarity in the context. The similar type of data become cluster and other types of data becomes another cluster.

Merits: Unsupervised Learning Do not required any trained data; similar senses are creating a group or cluster that is called homogenous data.

Demerits: Unsupervised Learning algorithms sometime do not identify the correct patterns for a specific problem because unsupervised is an unguided method.

Comparisons:

Approach	Advantage	Disadvantage
Knowledge-Based	These algorithms give higher Precision.	Performance depends on dictionary definitions.
Supervised	This type of algorithms is better than the two approaches w.r.t. implementation perspective.	These algorithms don’t give satisfactory result for resource scarce languages.
Unsupervised	There is no need of any sense inventory and sense annotated corpora in these approaches.	These algorithms are difficult to implement and performance is always inferior to that of other two approaches.

CHAPTER-4
KNOWLEDGE BASED APPROACH IN WSD

4.1. WORDNET

ENGLISH WORDNET

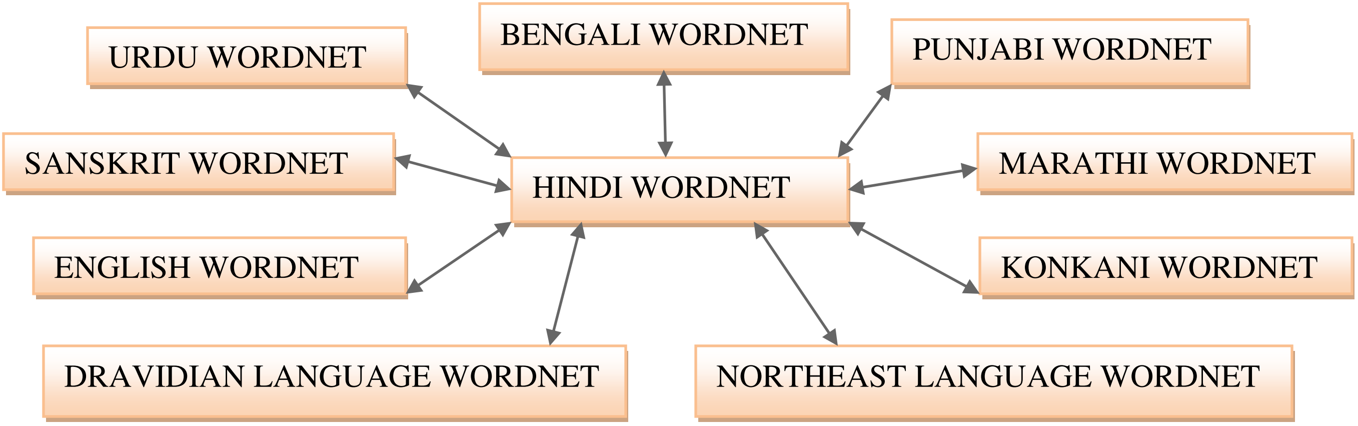
WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. While it is accessible to human users via a web browser, its primary use is in automatic text analysis and artificial intelligence applications. The database and software tools have been released under aBSD style license and are freely available for download from the WordNet website. Both the lexicographic data (lexicographer files) and the compiler (called grind) for producing the distributed database are available.

WordNet was created in the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George Armitage Miller starting in 1985 and has been directed in recent years by Christiane Fellbaum. The project received funding from government agencies including the National Science Foundation, DARPA, the Disruptive Technology Office (formerly the Advanced Research and Development Activity), and REFLEX. George Miller and Christiane Fellbaum were awarded the 2006 Antonio Zampolli Prize for their work with WordNet.

The initial goal of the WordNet project was to build a lexical database that would be consistent with theories of human semantic memory developed in the late 1960s. Psychological experiments indicated that speakers organized their knowledge of concepts in an economic, hierarchical fashion. Retrieval time required to access conceptual knowledge seemed to be directly related to the number of hierarchies the speaker needed to "traverse" to access the knowledge. Thus, speakers could more quickly verify thatcanaries can sing because a canary is a songbird ("sing" is a property stored on the same level as "canary"), but required slightly more time to verify that canaries can fly (where they had to access the concept "bird" on the superordinate level) and even more time to verify canaries have skin (requiring look-up across multiple levels of hyponymy, up to "animal"). While such experiments and the underlying theories have been subject to criticism, some of WordNet's organization is consistent with experimental evidence. For example, anomic aphasia, selectively affects speakers' ability to produce words from a specific semantic category, a WordNet hierarchy. Antonymous adjectives (WordNet's central adjectives in the dumbbell structure) are found to co-occur far more frequently than chance, a fact that has been found to hold for many languages.

WordNet has been used for a number of different purposes in information systems, including word sense disambiguation, information retrieval, automatic text classification,automatic text summarization, machine translation and even automatic crossword puzzle generation.

A common use of WordNet is to determine the similarity between words. Various algorithms have been proposed, and these include measuring the distance among the words and synsets in WordNet's graph structure, such as by counting the number of edges among synsets. The intuition is that the closer two words or synsets are, the closer their meaning. A number of WordNet-based word similarity algorithms are implemented in a Perl package called WordNet::Similarity, and in a Python package called NLTK. Other more sophisticated WordNet-based similarity techniques include ADW, whose implementation is available in Java. WordNet can also be used to inter-link other vocabularies.



INDOWORDNET

HINDI WORDNET

Hindi WordNet involves the representation of each Word Meaning as a set of word-forms known as synonym sets or synsets. These synsets are created for content words like Nouns, Verbs, Adverbs and Adjectives. It also defines a the following semantic relations to connect synsets.

i. Hypernym - Y is a hypernym of X if every X IS-A (KIND-OF) Y.
ii. Hyponym - Y is a hyponym of X if every Y IS-A (KIND-OF) X.
iii. Entailment - the verb Y is entailed by X if by doing X you must be doing Y.
iv. Troponym - the verb Y is a troponym of the verb X if the activity Y is doing X in some manner.
v. Meronym - Y is a meronym of X if Y is a part of X.
vi. Holonym - Y is a holonym of X if X is a part of Y.

As an example, the synset {घर, गृह} has the hypernym relation to {आवास, निवास}, a meronym relation to {बरामदा} and a hyponym relation to {झोपड़ी}.

EXAMPLE:

Synset:

{सूर्य,सूरज,भानु,दिवाकर,भास्कर,प्रभाकर,दिनकर,रवि:1,आदित्य:1,दिनेश,सविता,पुष्कर,मिहिर, अंशुमान,अंशुमाली,अग,अदित,अवि,अनड्वान्,आफ़ताब,आफताब,अफ़ताब,अफताब,अब्जबाँधव, अब्जहस्त,खगपति…………………}

Concept: हमारे सौर जगत का वह सबसे बड़ा और ज्वलंत पिंड जिससे सब ग्रहों को गरमी और प्रकाश मिलता है

Example: "सूर्य सौर ऊर्जा का एक बहुत बड़ा स्रोत है|/ पूरब से सूर्य को आते देख तिमिर दुम दबाकर भागने लगा ।"

4.2. KNOWLEDGE BASED APPROACH

Knowledge-based (dictionary-based) approaches utilize information from an explicit lexicon or knowledge base to disambiguate a word. The lexicon may be a machine-readable dictionary, a thesaurus or ontology and it is also uses to assign the correct sense to an ambiguous word. Hand coded knowledge may also be used. In general, WSD techniques using pre-existing structured lexical knowledge resources differ in:

- The lexical resource used (monolingual and/or bilingual MRDs, thesauri, lexical knowledge base, etc.);
- The information contained in this resource, exploited by the method; and
- The property used to relate words and senses.

Lesk (1986) proposes a method for guessing the correct word sense by counting word overlaps between dictionary definitions of the words in the context of the ambiguous word.

The Lesk algorithm is the seminal dictionary-based method. It is based on the hypothesis that words used together in text are related to each other and that the relation can be observed in the definitions of the words and their senses. Two (or more) words are disambiguated by finding the pair of dictionary senses with the greatest word overlap in their dictionary definitions.

For example, when disambiguating the words in "pine cone", the definitions of the appropriate senses both include the words evergreen and tree (at least in one dictionary).

Although knowledge-based systems have been proven to be ready-to-use and scalable tools for all-words WSD because they do not require sense-annotated data (Montoya et al.,302 Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods 2001), in general, supervised.

Other approaches measure the relatedness between words, taking as a reference a structured semantic net. Thus, Sussna (1993) employs the notion of conceptual distance between network nodes in order to improve precision during document indexing.

An alternative to the use of the definitions is to consider general word-sense relatedness and to compute the semantic similarity of each pair of word senses based on a given lexical knowledge base such as WordNet. Graph-based methods reminiscent of spreading activation research of the early days of AI research have been applied with some success. More complex graph-based approaches have been shown to perform almost as well as supervised methods or even outperforming them on specific domains. Recently, it has been reported that simple graph connectivity measures, such as degree, perform state-of-the-art WSD in the presence of a sufficiently rich lexical knowledge base. Also, automatically transferring knowledge in the form of semantic relations from Wikipedia to WordNet has been shown to boost simple knowledge-based methods, enabling them to rival the best supervised systems and even outperform them in a domain-specific setting.

The use of selection preferences (or selection restrictions) is also useful, for example, knowing that one typically cooks food, one can disambiguate the word bass in "I am cooking basses" (i.e., it's not a musical instrument).

LESK ALGORITHM

The **lesk algorithm** is a classical algorithm for word sense disambiguation introduced by Michael E. Lesk in 1986. It identify senses of words in context using definition overlap and gives 50-70% accuracy.

The original Lesk algorithm disambiguates a target word by comparing its gloss with those of its surrounding words. The target word is assigned the sense whose gloss has the most overlapping or shared words with its glosses of its neighboring words.Using the Oxford Advanced Learner’s Dictionary, demonstrates this algorithm on the words pine cone. It finds that the word pine has two senses:

Sense 1: kind of **evergreen tree** with needle–shaped leaves
Sense 2: waste away through sorrow or illness.

The word cone has three senses:

Sense 1: solid body which narrows to a point

Sense 2: something of this shapes whether solid or hollow

Sense 3: fruit of certain **evergreen tree**

As can be seen, the best intersection is Pine #1 ∩ Cone #3 = 2.

- There are two hypotheses that underlying this approach. The first is that words that appear together in a sentence can be disambiguated by assigning to them the senses that are most closely related to their neighboring words. This follows from the intuition that words that appear together in a sentence must be related in some way, since they are normally working together to communicate. The second hypothesis is that are related senses can be identified by finding overlapping words in their definition. The intuition here is equally reasonable, in that words that are related will often be defined using the same words, and in fact may refer to each other in their definitions.
- The main limitation to this approach is that dictionary glosses are often quit brief, and may not include sufficient vocabulary to identify related senses.

The algorithm computes a score for each sense of the target word. To find the score for a sense of the target word the target sense is compared to the senses of the context word. The algorithm finds the sense of each context word that is most related to the target sense. The score for each target sense is sum of the relatedness scores between the target sense and each most related context sense. Once the algorithm finds a score for each sense of the target word, the sense with the greatest score is assigned to the target word.

4.3. ALGORITHM

ADAPTED LESK ALGORITHM [14] :

Apply Lesk’s basic approach to take advantage of the highly inter-connected set of relations among synonyms that WordNet offers.

While Lesk’s algorithm restricts its comparisons to the glosses of the words being disambiguated adapted lesk’s approach is able to compare the glosses of words that are related to the words to be disambiguated. So, it overcomes the limitations:

- Access a dictionary with senses arranged in a hierarchical order (WordNet). This extended version uses not only the gloss/definition of the synset but also considers the meaning of related words.
- Apply a new scoring mechanism to measure gloss overlap that gives a more accurate score than the original Lesk bag of words counter

To disambiguate each word in a sentence that has N words, we call each word to be disambiguated as a target word. The algorithm is described in the following steps:

1. Select a context: optimizes computational time so if N is long, we will define K context around the target word (or k-nearest neighbour) as the sequence of words starting K words to the left of the target word and ending K words to the right. This will reduce the computational space that decreases the processing time. For example: If k is four, there will be two words to the left of the target word and two words to the right.
2. For each word in the selected context, we look up and list all the possible senses of noun and verb.
3. For each sense of a word (WordSense), we list the following relations (example of pine and cone):
   o Its own gloss/definition that includes example texts that WordNet provides to the glosses.
   o The gloss of the synsets that are connected to it through the hypernym relations. If there is more than one hypernym for a word sense, then the glosses for each hypernym are concatenated into a single gloss string (*).
   o The gloss of the synsets that are connected to it through the hyponym relations (*).
   (*) All of them are applied with the same rule.
4. Combine all possible gloss pairs that are archived in the previous steps and compute the relatedness by searching for overlap. The overall score is the sum of the scores for each relation pair.

When computing the relatedness between two synsets s1 and s2, the pair hype-hype means the gloss for the hypernym of s1 is compared to gloss for the hypernym of s2. The pair hype-hypo means that the gloss for the hypernym of s1 is compared to the gloss for the hyponym of s2.

OverallScore(s1, s2)= Score(hype(s1)-hypo(s2)) +

Score (gloss (s1)-hypo(s2)) + Score(hype(s1)-gloss(s2))...

(OverallScore (s1, s2) is also equivalent to OverallScore(s2, s1)).

In the example of "pine cone", there are 3 senses of pine and 6 senses of cone, so we can have a total of 18 possible combinations. One of them is the right one.

To score the overlap we use a new scoring mechanism that differentiates between N-single words and N-consecutive word overlaps and effectively treats each gloss as a bag of words. The shortest words are those which are used more often, the longest ones are used less often.

Measuring overlaps between two strings is reduced to solve the problem of finding the longest common sub-string with maximal consecutives. Each overlap which contains N consecutive words, contributes N2 to the score of the gloss sense combination. For example: an overlap "ABC" has a score of 3^2=9 and two single overlaps "AB" and "C" has a score of 2^2 + 1^1=5.

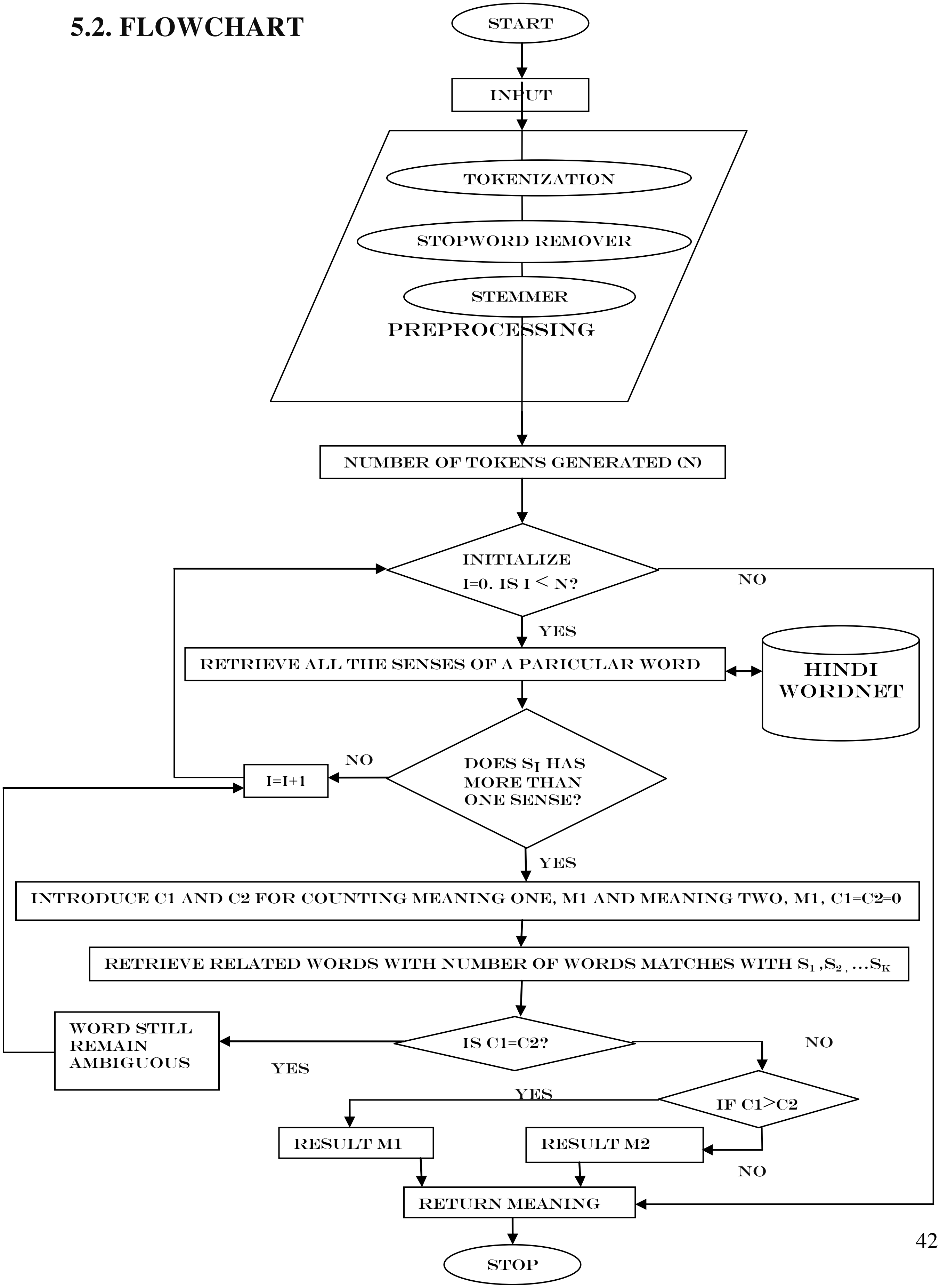
5. Once each combination has been scored, we pick up the sense that has the highest score to be the most appropriate sense for the target word in the selected context space. Hopefully the output not only gives us the most appropriate sense but also the associated part of speech for a word.

CHAPTER-5
IMPLEMENTATION AND RESULTS ANALYSIS

5.1. PSEUDOCODE

```
procedure disambiguate_all_words
  for all wt(word) in input do
          best_sense = disambiguate_single_word(wt)
          display best_sense
  end for
end procedure
function disambiguate_single_word(wt)
  for all st of target word wt do
  //sti is the ith sense of the target word wt
          score i = 0
          for i = t-c1 to t+cr do
                  if j = t then
                          next  j
                  end if
        for sjk of wj do //sjk is the kth sense of the target word wj
                          temp_score k = relatedness(st,sjk)
                  end for
          best_score = max temp_score
          if best_score  > threshold then
                  score i = score i + best_score
          end if
  end for
end for
Return sense i
// such that score i >score j for all j, where j = I
end function
```

5.2. FLOWCHART


START
INPUT
TOKENIZATION
STOPWORD REMOVER
STEMMER
PREPROCESSING
NUMBER OF TOKENS GENERATED (N)
INITIALIZE
I=0. IS I < N?
NO
YES
RETRIEVE ALL THE SENSES OF A PARICULAR WORD
HINDI WORDNET
NO
I=I+1
DOES SI HAS MORE THAN ONE SENSE?
YES
INTRODUCE C1 AND C2 FOR COUNTING MEANING ONE, M1 AND MEANING TWO, M1, C1=C2=0
RETRIEVE RELATED WORDS WITH NUMBER OF WORDS MATCHES WITH S1 ,S2 , …SK
WORD STILL REMAIN AMBIGUOUS
IS C1=C2?
NO
YES
YES
IF C1>C2
RESULT M1
RESULT M2
NO
RETURN MEANING
STOP

5.3. RESULT ANALYSIS

The running time complexity of implementing this algorithm is O (n^2).

$$\text{Accuracy} = \frac{\text{Number of ambiguous word correctly disambiguate by the system}}{\text{Total no. of ambiguous word}} \text{ X } 100\%$$

Table**:**

Test	No. of sentence	No. of ambiguous word	No. of word correctly disambiguate	No. of word incorrectly disambiguate	Accuracy (in %)
1	25	14	10	4	71.43
2	50	43	33	10	76.74
3	75	51	38	13	74.51
4	100	91	69	22	75.82
5	125	120	90	33	75

$$\text{Precision (P)} = \frac{\text{Number of matched target words based on human decision}}{\text{Number of instances responded by the system}} \text{ X } 100\%$$

$$\text{Recall (R)} = \frac{\text{Number of target words for which answer matches with human decided answer}}{\text{Total number of target words in the data}} \text{ X } 100\%$$

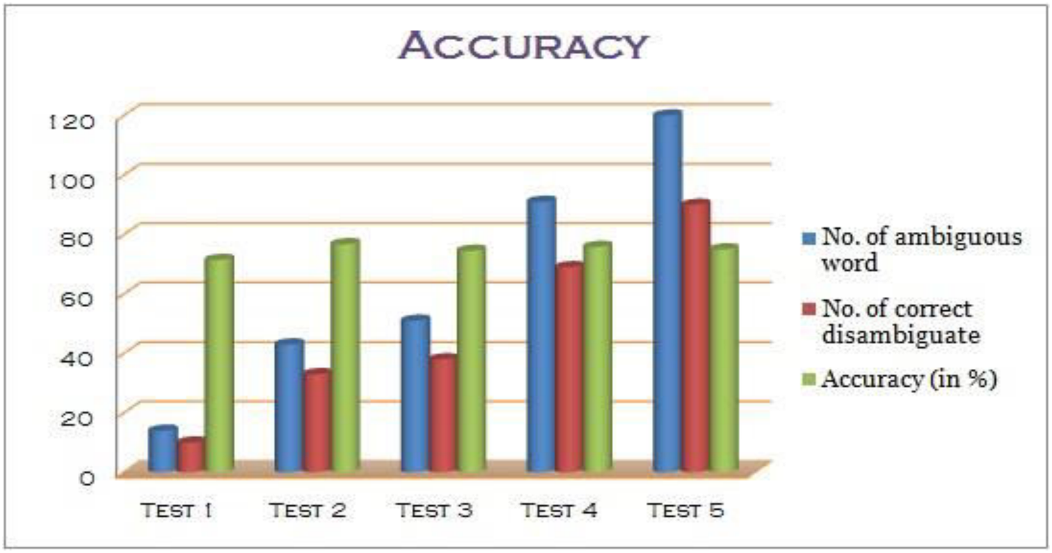
F-measure = (2*P*R/ (P+R)) X 100%

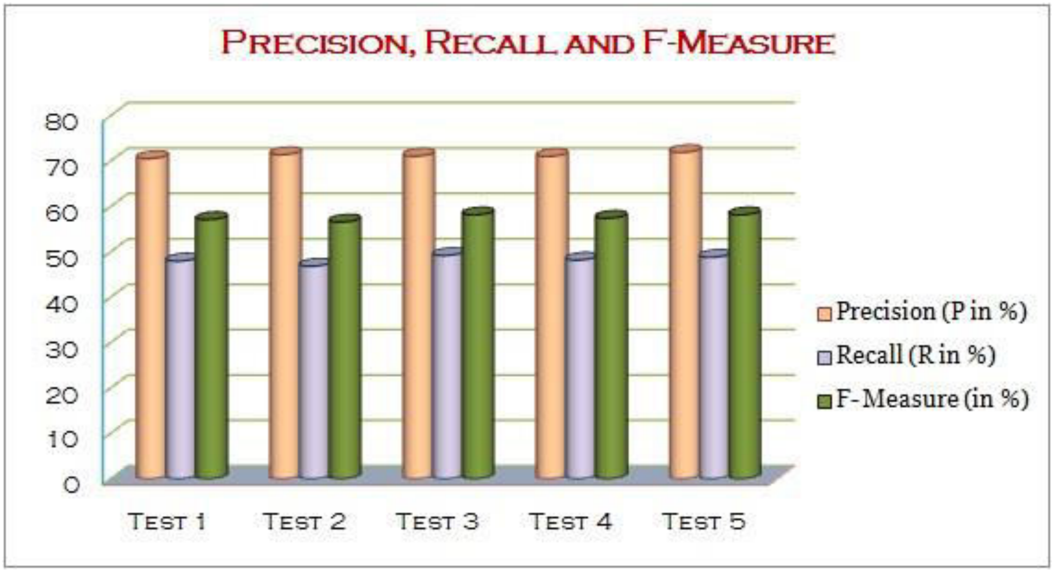
Table:

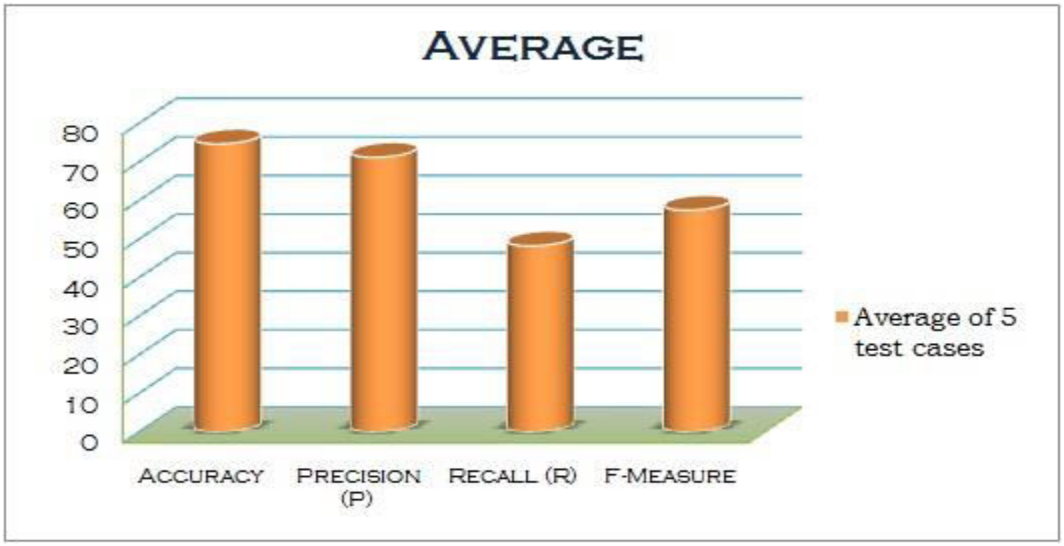
Test	No. of matched target words based on human	No. of instances responded by the system	No. of total target words in the data	Precision (P in %)	Recall (R in %)	F- Measure (in %)
1	12	17	25	70.58	48	57.14
2	15	21	32	71.43	46.87	56.60
3	32	45	65	71.11	49.23	58.18
4	27	38	46	71.05	48.21	57.44
5	42	57	86	71.93	48.84	58.18

Average: For 5 test cases

Accuracy	Precision (P)	Recall (R)	F-Measure
74.7	71.22	48.23	57.51


ACCURACY
120
100
80
60
40
20
0
TEST 1
TEST 2
TEST 3
TEST 4
TEST 5
No. of ambiguous word
No. of correct disambiguate
Accuracy (in %)


PRECISION, RECALL AND F-MEASURE
80
70
60
50
40
30
20
10
0
TEST 1
TEST 2
TEST 3
TEST 4
TEST 5
Precision (P in %)
Recall (R in %)
F-Measure (in %)


AVERAGE
80
70
60
50
40
30
20
10
0
ACCURACY
PRECISION (P)
RECALL (R)
F-MEASURE
Average of 5 test cases

CHAPTER-6
CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

In this project the knowledge based approach is used for identifying the disambiguate word. Word sense disambiguation is one of the applications of NLP. A database is created of those words which have more than one meaning. With the help of database the word of the meaning can be identified. We have found that many of the word contain different meaning.

In this section we proposed an algorithm which gives the desired output of the user's. The algorithm computes a score for each sense of the target word. To find the score for a sense of the target word the target sense is compared to the senses of the context word. The algorithm finds the sense of each context word that is most related to the target sense. The score for each target sense is sum of the relatedness scores between the target sense and each most related context sense. Once the algorithm finds a score for each sense of the target word, the sense with the greatest score is assigned to the target word.

6.2 FUTURE DIRECTION

The GUI can be making more attractive. In this paper the database that has been created has limited number of collection of words. In future one can modify the database and can insert more words which have more than one meaning. The different types of related terms like hypernyms, hyponyms, meronym, troponym, and etc. concept can be used.

Future work on WSD will focus on investigating the possibility of the involvement of more complex context in the WSD process and considering the effective combination between the current result with verb contexts and the possible future result with some other contexts.

Continue to build the knowledge base, enlarge the coverage and improve the system performance. The experiment result is clearly show that more word instances can improve the disambiguation accuracy and recall scores.

WSD is often an unconscious process for human beings. It is unlikely that a reader examines all surrounding words when determining the sense of word, which calls for a smarter and more selective matching strategy.

APPENDIX A: REFERENCES

[1] Natural_language_Processing",[Online].Available: http://en.wikipedia.org/wiki/Natural_language_processing

[2] Ted Pedersen (tpederse@d.umn.edu)http://www.d.umn.edu/~tpederse

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

[4] Rada Mihalcea, 2007, Word sense disambiguation, University of North Texas http://www.cs.unt.edu/~rada

[5] Francisco Tacoa, Danushka Bollegala and Mitsuru Ishizuka "A Context Expansion Method for SupervisedWord Sense Disambiguation," In IEEE Sixth International Conference on Semantic Computing, 2012

[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 Word-Sense-Disambiguation Methods

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

[8] Word-sense disambiguation - Wikipedia, the free encyclopedia.html

[9] Agirre, E.M.stevenson 2006 knowledge sources for WSD. In word sense disambiguation: algorithms and applications, E.Agirre and P.Edmonds, Eds. Springer, New York, NY.

[10] Lesk algorithm From Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Dictionary.

[11] Arindam Roy1,Sunita Sarkar2and Bipul Syam Purkayastha3.International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 201 10.5121/ijnlc.2014.3305 51 Knowledge Based Approaches to Nepali Word Sense Disambiguation

[12] Word sense disambiguation - Scholarpedia.html

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

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

[15] Agirre, E., Arregi, X., Artola, X., de Ilarraza, D., Sarasola, K., 1994. A methodology for the extraction of semantic knowledge from dictionaries using phrasal patterns.

[16] Chklovski, T. and Mihalcea, R. 2002. Building a sense tagged corpus with open mind word expert. In Proceedings of the Acl-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Morristown, NJ, 116-122.

[17] A. Eneko & G. Rigau, "Word sense using conceptual density" In Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, 1996 ndon, vol. A247, pp. 529–551, April 1996.

[18] Bhattacharya and Unny: Word sense disambiguation and text similarity measurement using wordnet. (2002).

[19] Prity Bala Apaji Institute, Banasthali Vidhyapith Newai, Rajasthan, India. Word Sense Disambiguation Using Selectional Restriction

[20] A. Moro, A. Raganato, R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. Transactions of the Association for Computational Linguistics (TACL), 2, pp. 231-244, 2014.

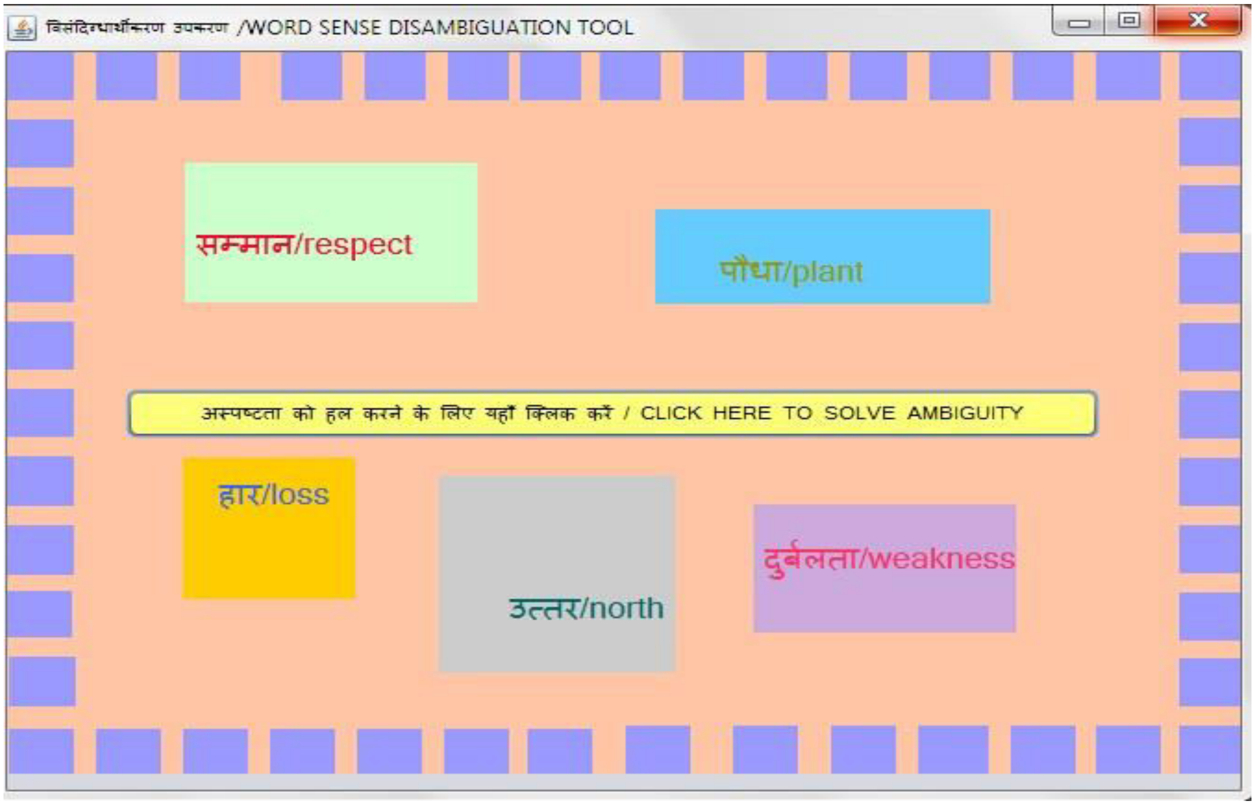
[21] Alok Ranjan Pal[1, 3], Anirban Kundu[2, 3], Abhay Singh[1] , Raj Shekhar[1] , Kunal Sinha[1]: An Approach To Word Sense Disambiguation Combining Modified Lesk And Bag-Of-Words

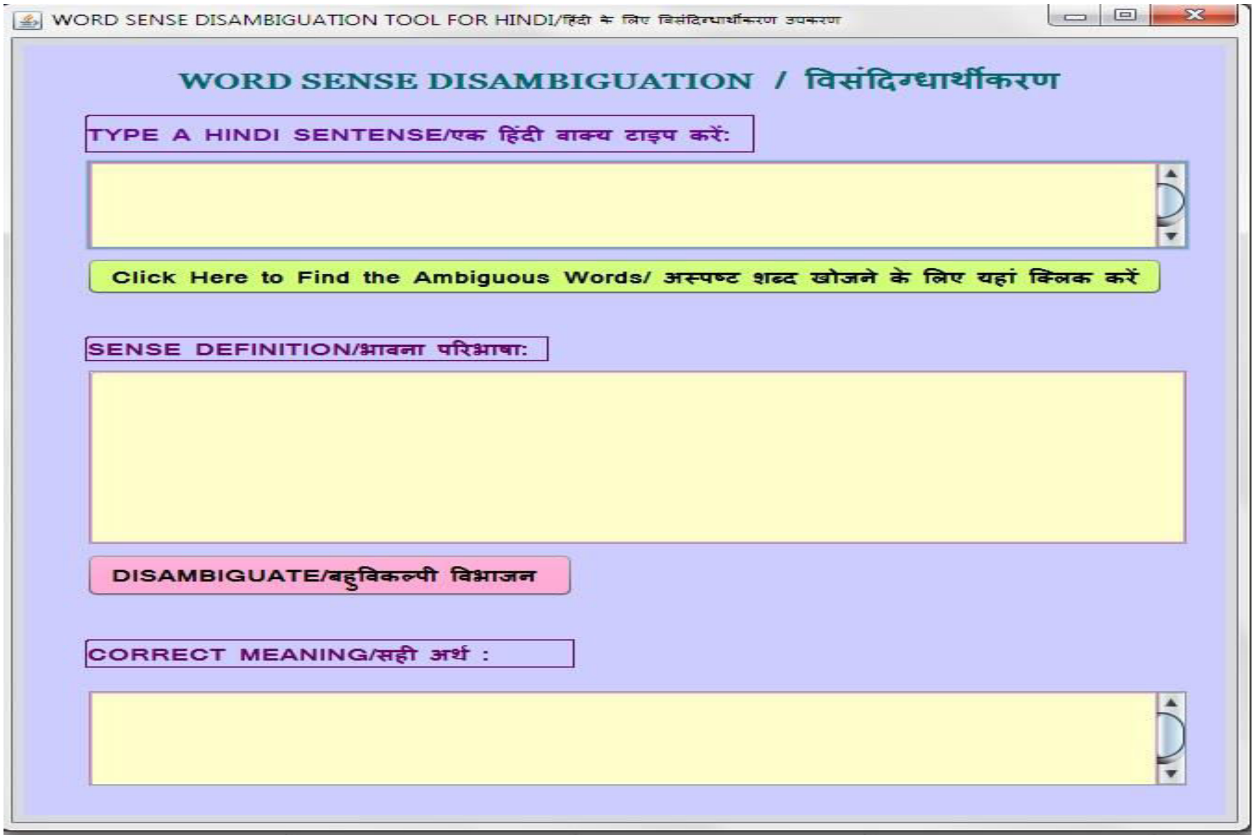
[22] Ehsan Hessami, Faribourz Mahmoudi, Amir Hossien Jadidinejad, 2011 World Congress on Information and Communication Technologies: Unsupervised weighted graph for Word Sense Disambiguation

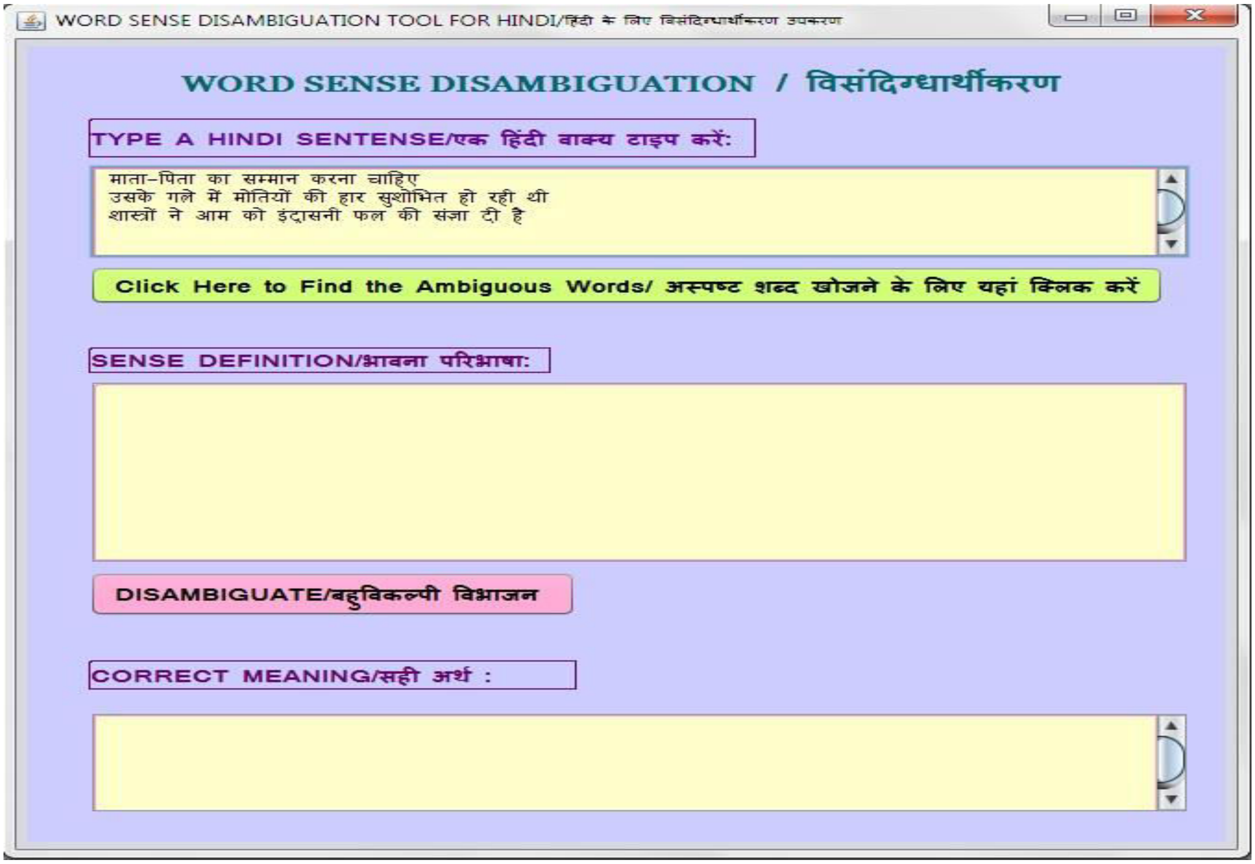
[23] Abhishek Fulmari1 , Manoj B. Chandak2, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013 Copyright to IJARCCE www.ijarcce.com 4667: A Survey on Supervised Learning for Word Sense Disambiguation

[24] Simone Paolo Ponzetto , Roberto Navigli: Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems

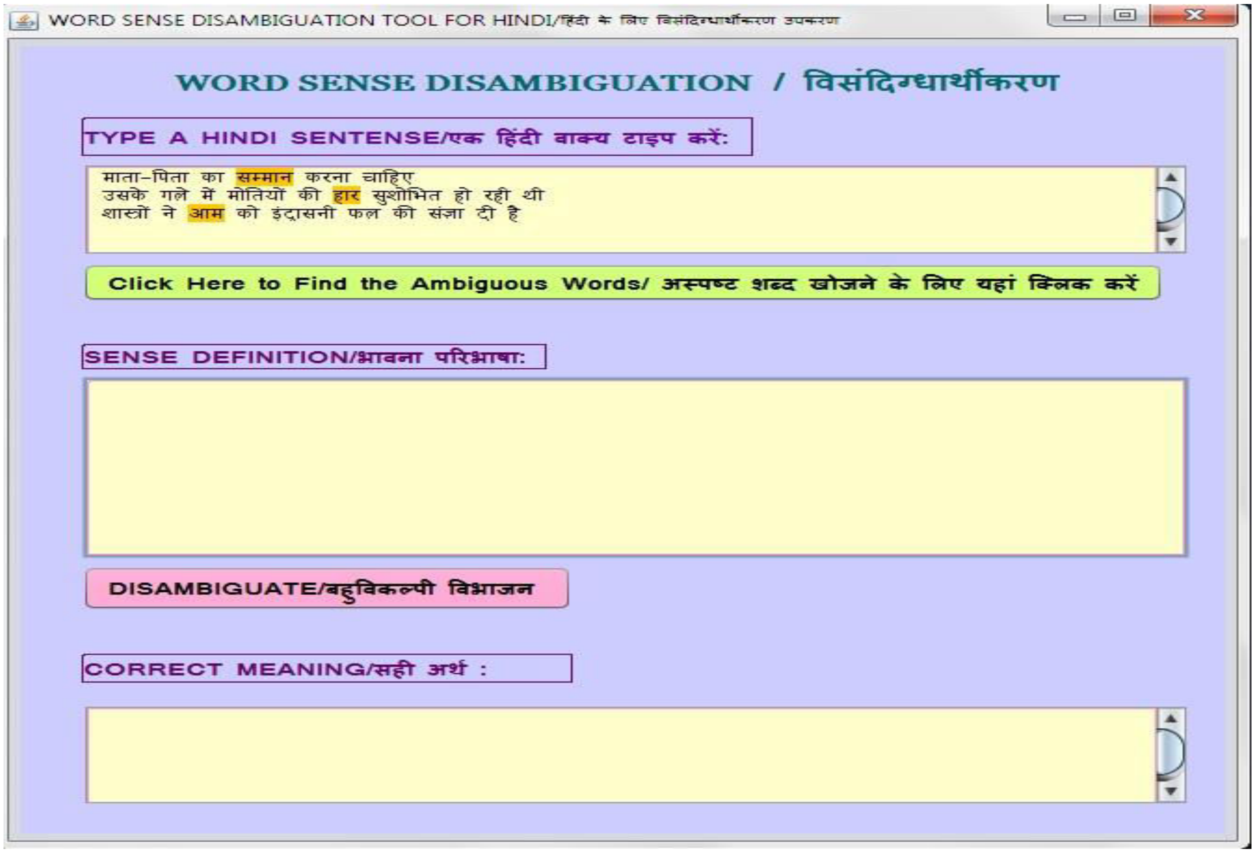
APPENDIX B: SNAPSHOTS


विसंदिग्धार्थीकरण उपकरण /WORD SENSE DISAMBIGUATION TOOL
सम्मान/respect
पौधा/plant
अस्पष्टता को हल करने के लिए यहाँ क्लिक करें / CLICK HERE TO SOLVE AMBIGUITY
हार/loss
उत्तर/north
दुर्बलता/weakness

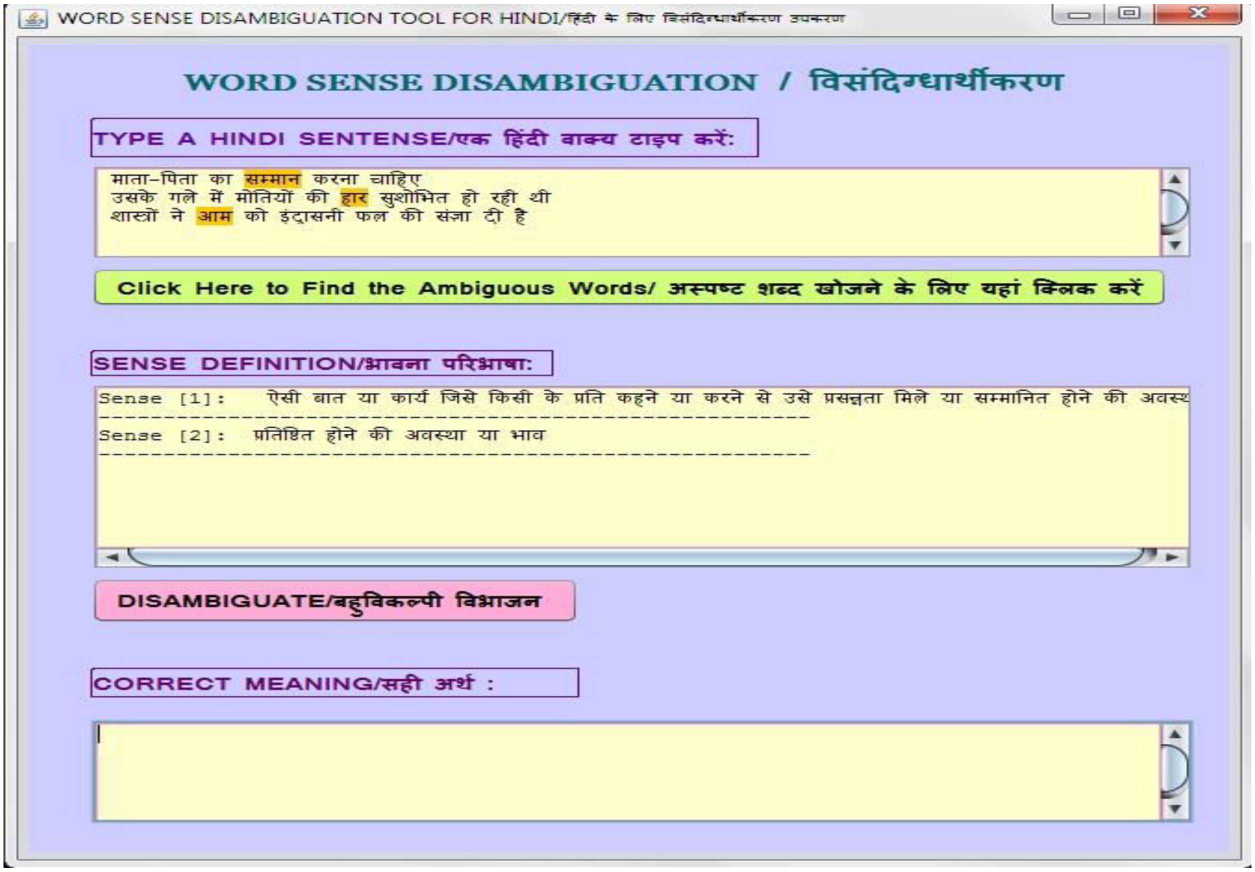

WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :


WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है

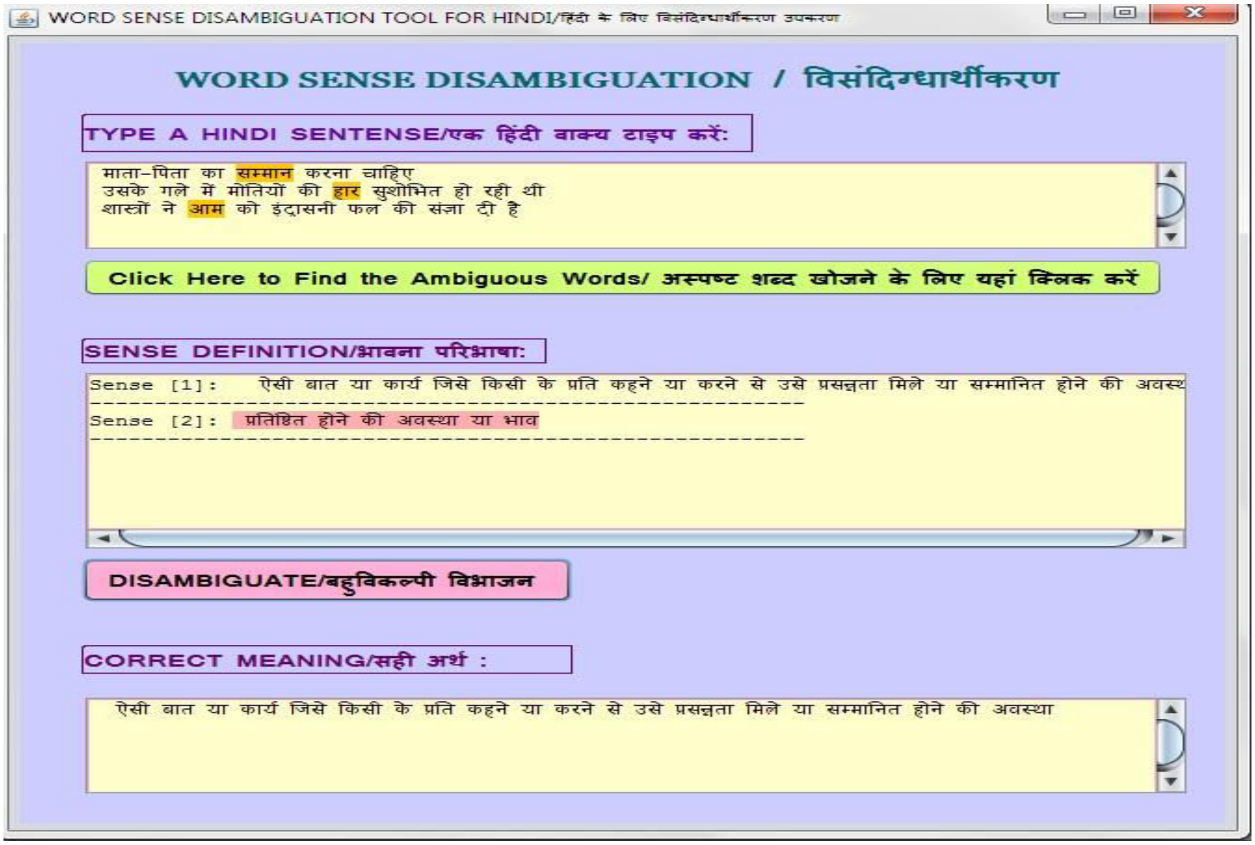

WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है


WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है

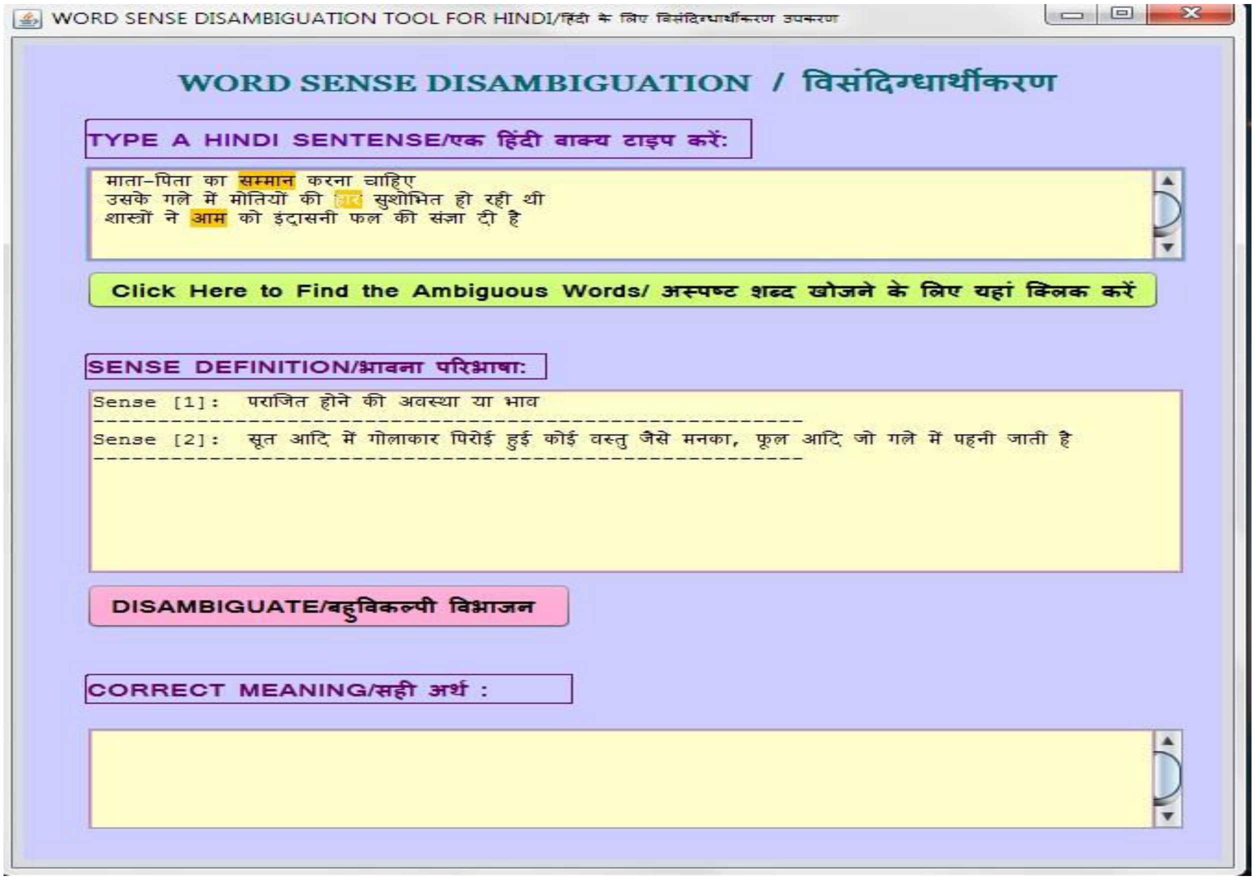
```
Sense [1]:   ऐसी बात या कार्य जिसे किसी के प्रति कहने या करने से उसे प्रसन्नता मिले या सम्मानित होने की अवस्
------------------------------------------------------
Sense [2]:  प्रतिष्ठित होने की अवस्था या भाव
------------------------------------------------------
```


WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है

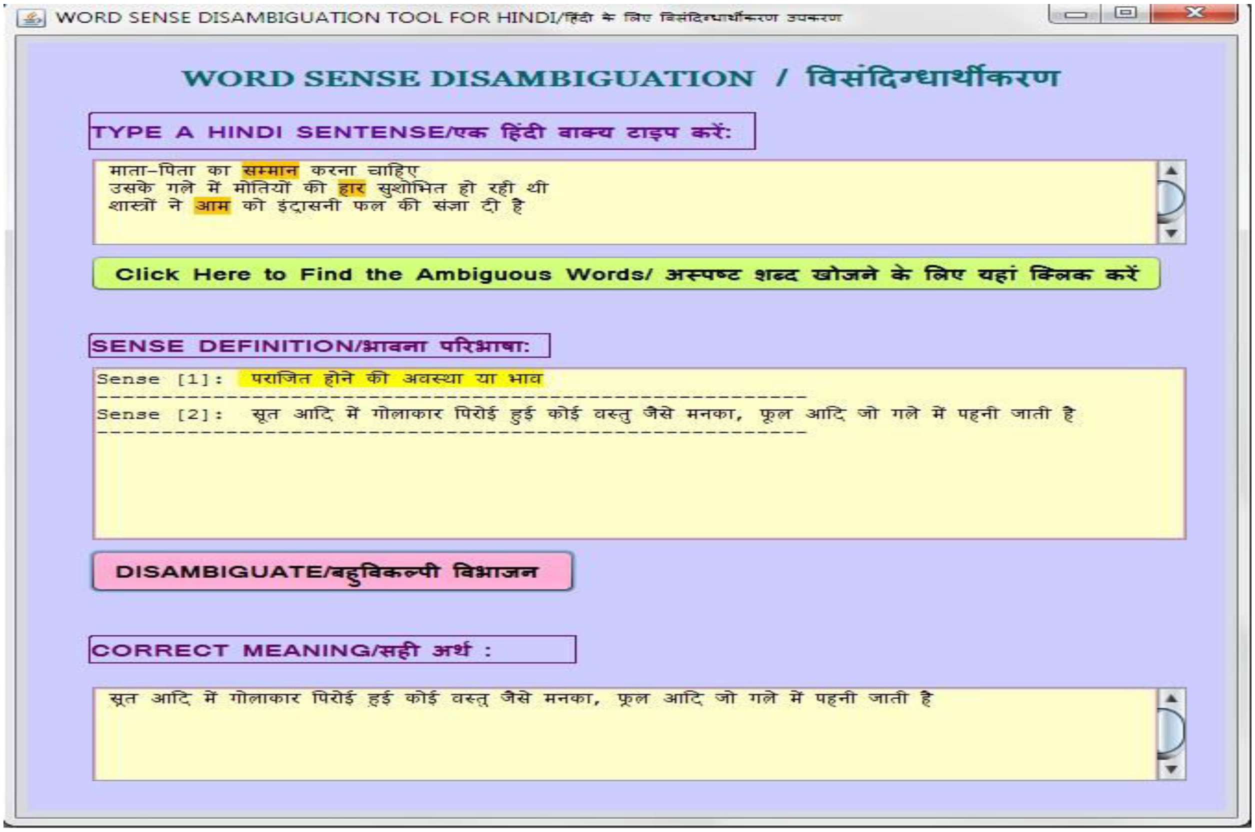
```
Sense [1]:   ऐसी बात या कार्य जिसे किसी के प्रति कहने या करने से उसे प्रसन्नता मिले या सम्मानित होने की अवस्
------------------------------------------------------
Sense [2]:  प्रतिष्ठित होने की अवस्था या भाव
------------------------------------------------------
```

ऐसी बात या कार्य जिसे किसी के प्रति कहने या करने से उसे प्रसन्नता मिले या सम्मानित होने की अवस्था


WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

```
माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है
```

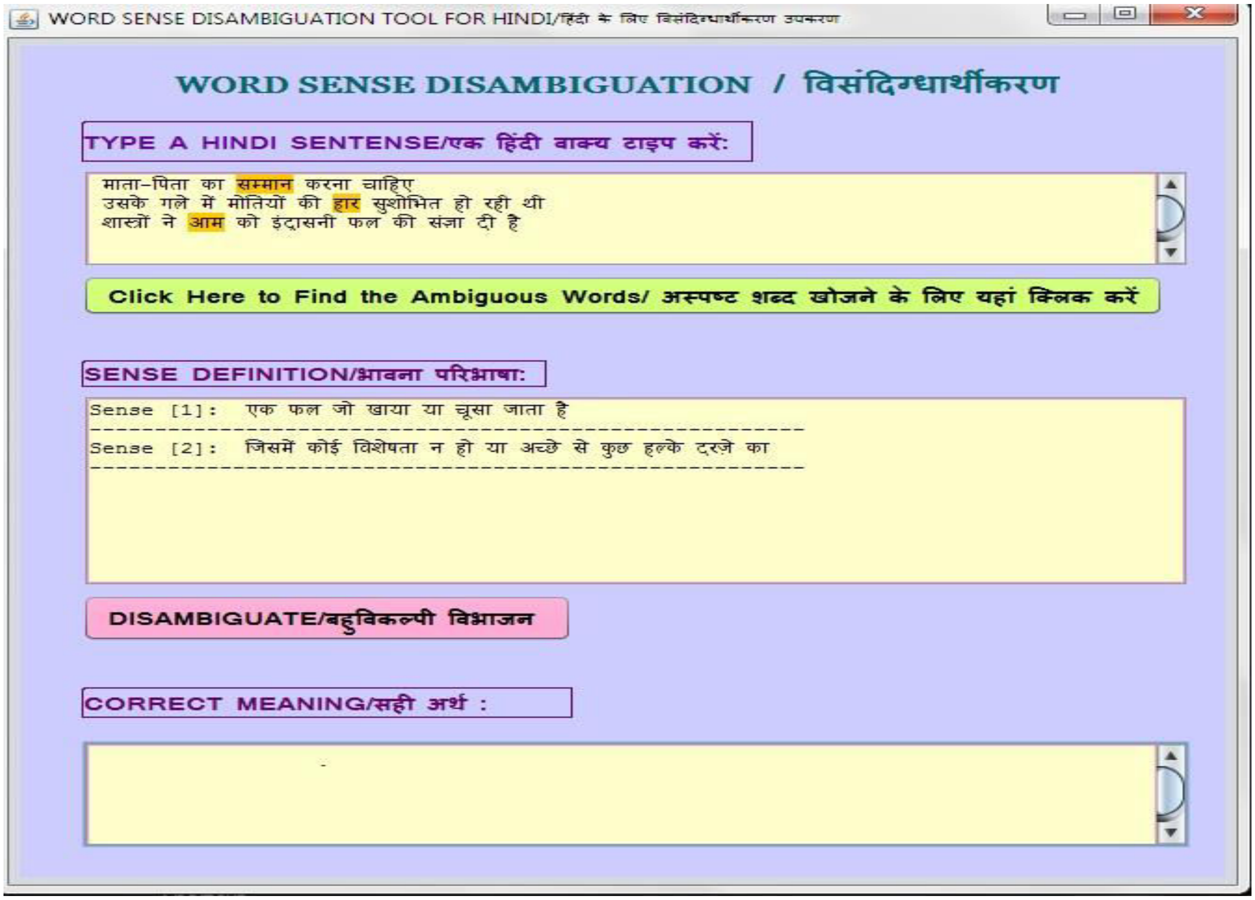
```
Sense [1]:  पराजित होने की अवस्था या भाव
----------------------------------------------------------
Sense [2]:  सूत आदि में गोलाकार पिरोई हुई कोई वस्तु जैसे मनका, फूल आदि जो गले में पहनी जाती है
----------------------------------------------------------
```


WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

```
माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है
```

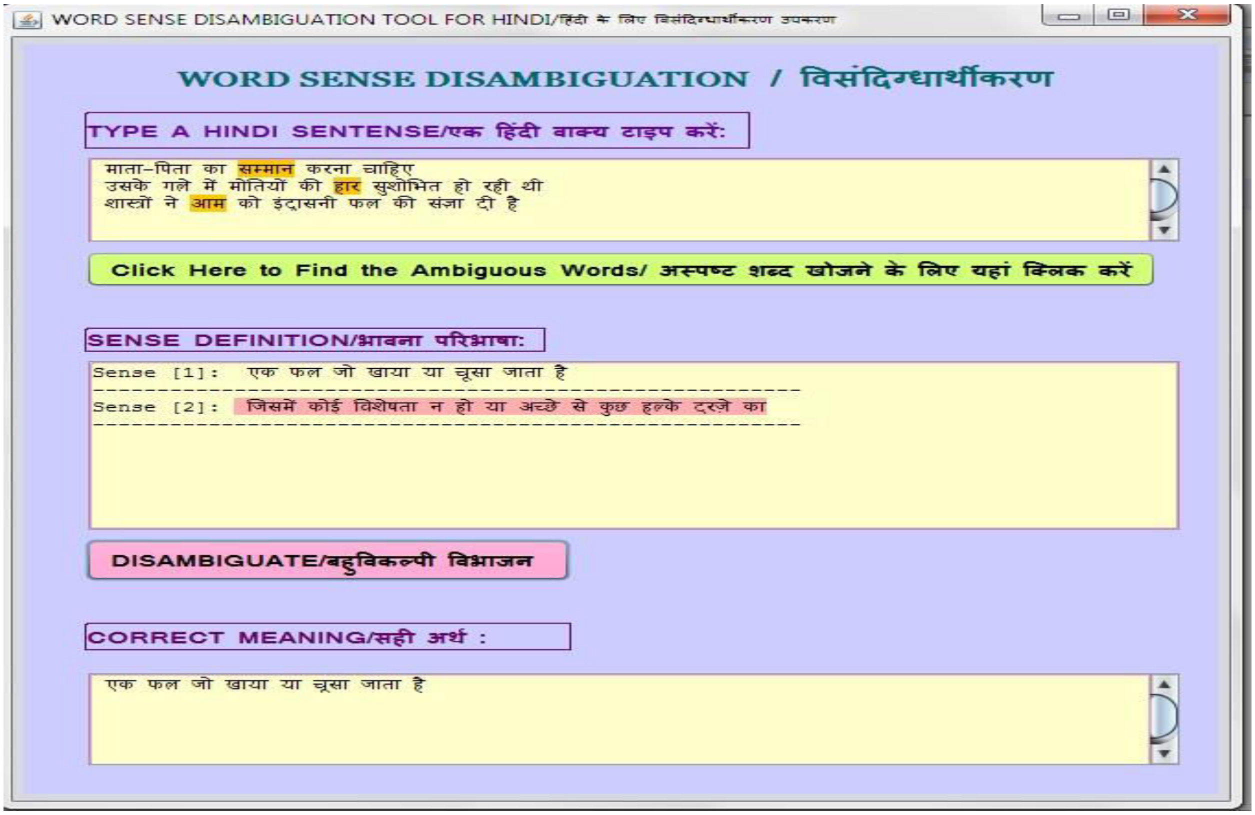
```
Sense [1]:  पराजित होने की अवस्था या भाव
----------------------------------------------------------
Sense [2]:  सूत आदि में गोलाकार पिरोई हुई कोई वस्तु जैसे मनका, फूल आदि जो गले में पहनी जाती है
----------------------------------------------------------
```

```
सूत आदि में गोलाकार पिरोई हुई कोई वस्तु जैसे मनका, फूल आदि जो गले में पहनी जाती है
```


WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है

```
Sense [1]:  एक फल जो खाया या चूसा जाता है
------------------------------------------------------------
Sense [2]:  जिसमें कोई विशेषता न हो या अच्छे से कुछ हल्के दरज़े का
------------------------------------------------------------
```


WORD SENSE DISAMBIGUATION TOOL FOR HINDI/हिंदी के लिए विसंदिग्धार्थीकरण उपकरण
WORD SENSE DISAMBIGUATION / विसंदिग्धार्थीकरण
TYPE A HINDI SENTENSE/एक हिंदी वाक्य टाइप करें:
Click Here to Find the Ambiguous Words/ अस्पष्ट शब्द खोजने के लिए यहां क्लिक करें
SENSE DEFINITION/भावना परिभाषा:
DISAMBIGUATE/बहुविकल्पी विभाजन
CORRECT MEANING/सही अर्थ :

माता–पिता का सम्मान करना चाहिए
उसके गले में मोतियों की हार सुशोभित हो रही थी
शास्त्रों ने आम को इंद्रासनी फल की संज्ञा दी है

```
Sense [1]:  एक फल जो खाया या चूसा जाता है
------------------------------------------------------------
Sense [2]:  जिसमें कोई विशेषता न हो या अच्छे से कुछ हल्के दरज़े का
------------------------------------------------------------
```

एक फल जो खाया या चूसा जाता है

APPENDIX C: DEVELOPMENT PLATFORM

The software requirements for implementing the system:

I. Operating System: Windows 7

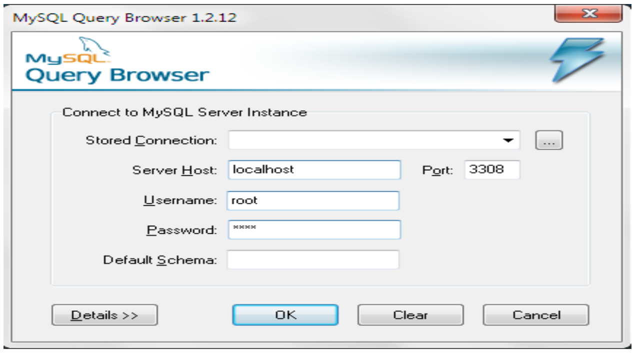
II. Platform Used:

(i) NetBeans IDE 8.0.2 (jdk-8u45-nb-8_0_2-windows-x64). It can be downloaded from https://netbeans.org/downloads

Reading module storage...done.

NetBeans IDE and NetBeans Platform are based on software from netbeans.org, which has been dual licensed under the Common Development and Distribution License (CDDL) and the GNU General Public License version 2 with Classpath exception. For more information, please visit www.netbeans.org.

(ii) MySQL Query Browser 1.2.12



Stored Connection:

Server Host: localhost Port: 3308

Username: root

Password: ****

Default Schema:

(iii) Notepad ++

The hardware requirements for developing and implementing the system:

A Core(TM) i3 based LAPTOP with minimum of

I. 4 GB RAM

II. 500GB Hard Disk Space

III. Intel(R) Core(TM) i3-2330 CPU @ 2.20 GHz processor

APPENDIX C: APPLICATION SETUP