International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Optimized Combinatorial Approach of Learning Algorithm for Word Sense Disambiguation

Neetu Sharma[1], S. Niranjan[2]

[1]GITAM, Kablana, Jhajjar, Haryana, India

[2]Director,Ganga Technical Campus, , Soldha, Haryana, India

Abstract: ***Word sense disambiguation is the process to find best sense of ambiguous word from the existing senses to remove the ambiguity. This thesis work is an attempt to optimize the word sense disambiguation method. Most commonly supervised machine learning algorithms were used to solve this problem and improve the performance. Some attempts were made to use unsupervised machine learning algorithms also like K-means clustering algorithm. In this research work supervised learning algorithm Naïve Bayesian is combined with the unsupervised learning algorithm K-means Clustering and the performance is enhanced in getting best sense of ambiguous word. C# is used to create interface for getting input in the form of sentence containing ambiguous word and displaying the output as a best sense for that ambiguous word. SQL 2008 is used a s database to store the sentences entered and their corresponding meanings. WORDNET as a database for extracting senses of ambiguous word is used. Performance is evaluated on the basis of scores of precision, recall and F-score that how well this optimized algorithm works now to improve the accuracy.***

Keywords: Naïve Bayesian Algorithm, K-Means Clustering, WORDNET

1. Introduction

Word sense disambiguation (WSD) has been a very active area of research in computational linguistics field . Most of the work has been focused on English language. One of the factors that hampered WSD research for other languages has been the lack of appropriate resources, particularly in the form of sense-annotated corpus data. WSD is a fundamental problem in natural language processing. It can be potentially used as a component in many applications, such as machine translation (MT) and information retrieval (IR). Word sense disambiguation (WSD) is one of the most critical and widely studied Natural Language Processing tasks, which is used in order to increase the success rates of NLP applications like machine translation, information search and information extract, natural language understanding (such as man-machine conversation system, interrogator-responder system), text auto-proofreading, speech recognition, sound-character transformation, syntax structure recognition and the language study etc..The classical approach to WSD that relies on an underlying Naïve Bayes model represents an important theoretical approach in statistical language processing: Bayesian classification. The idea of the Bayes classifier (in the context of WSD) is that it looks at the words around an ambiguous word in a large context window. Each content word contributes potentially useful information about which sense of the ambiguous word is likely to be used with it. The classifier does no feature selection. Instead it combines the evidence from all features. The mentioned classifier is an instance of a particular kind of Bayes classifier, the Naïve Bayes classifier. Naïve Bayes is widely used due to its efficiency and its ability to combine evidence from a large number of features.The Naïve Bayes assumption is that the attributes used for description are all conditionally independent, an assumption having two main consequences. The first is that all the structure and linear ordering of words within the context are ignored, leading to a so-called “bag of words model”.

Three main approaches have been applied in the WSD field. These are knowledge-based approaches, corpus based approaches and hybrid approach. Knowledge based approaches use Machine Readable Dictionaries (MRD). It relies on information provided by MRD. Corpus based approaches can be divided into two types, supervised and unsupervised learning approaches. Supervised learning approaches use information gathered from training on a corpus that has sense-tagged for semantic disambiguation. The classification approach of WSD makes use of statistical approaches either referring lexicons or using corpus for training. Thesauri, lexicons and corpus are the main source of training in the supervised approach. Unsupervised leaning approaches determine the class membership of each object to be classified in a sample without using sense-tagged training examples. Hybrid approach combines aspects of aforementioned methodologies.

Word sense disambiguation a task of removing the ambiguity of word in context, is important for many NLP applications such as:

1) **Information Retrieval**
   WSD helps in improving term indexing in information retrieval word senses improve retrieval performance if the senses are included as index terms. Thus, 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. For example: Using different indexes for keyword "Java" as "programming language", as "type of coffee", and as "location" will improve accuracy of an IR system.
2) **Machine Translation**
   WSD is important for Machine translations. 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.
3) **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".

4) **Text Processing**
Text to Speech translation i.e., when words are pronounced in more than one way depending on their meaning. For example: "bank" can be "a side of river" or "financial instituitions".

2. Motivation

WSD is one of the most important open problems in the Natural Language Processing (NLP) field. Despite the wide range of approaches investigated and the large effort devoted to tackle this problem, it is a fact that to date no large scale broad coverage and highly accurate word sense disambiguation system has been built.
WSD is:

- Accessible to anyone with an interest in NLP.
- Persuade you to work on word sense disambiguation.
- It's an interesting problem.
- Lots of good work already done, still more to do.
- There is infrastructure to help you get started.

Persuade you to use word sense disambiguation in your text applications. Machine learning is a branch of artificial intelligence which studies mechanisms to mimic the ability of humans to learn. Machine learning strives to get the computer to learn tasks such as discriminating between objects, segregating similar objects from dissimilar ones and learning from experience.

Various Machine Learning (ML) approaches have been demonstrated to produce relatively successful Word Sense Disambiguation (WSD) systems. There are still unexplained differences among the performance measurements of different algorithms, hence it is warranted to deepen the investigation into which algorithm has the right 'bias' for this task. These tasks are formally known as supervised, unsupervised and reinforcement learning in the machine learning parlance. In supervised learning, the system is presented with a set of data which is labeled into various categories and involves learning a function which maps the data to the categories. This function is then used to map an unseen instance of the data to its corresponding category. Unsupervised learning on the other hand works on unlabelled data and involves grouping this data based on its characteristics, i.e., infer potential categories from unlabelled data. Reinforcement learning is a system which learns an effective way of doing a task from the experience of doing the task and feedback from the environment on the outcome.

3. Basic Approaches to WSD

Approaches to WSD are often classified according to the main source of knowledge used in sense differentiation. Methods that rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence, are termed dictionary-based or knowledge-based. Methods that eschew (almost) completely external information and work directly from raw unannotated corpora are termed unsupervised methods (adopting terminology from machine learning). Included in this category are methods that use word-aligned corpora to gather cross-linguistic evidence for sense discrimination. Finally, supervised and semi-supervised WSD make use of annotated corpora to train from, or as seed data in a bootstrapping process. Almost every approach to supervised learning has now been applied to WSD, including aggregative and discriminative algorithms and associated techniques such as feature selection, parameter optimization, and ensemble learning. Unsupervised learning methods have the potential to overcome the new knowledge acquisition bottleneck (manual sense-tagging) and have achieved good results. These methods are able to induce word senses from training text by clustering word occurrences and then classifying new occurrences into the induced clusters/senses and the Web. The objective of clustering is to take a set of instances that incorporate ideas from both. The algorithm acts as a search strategy that dictates how to proceed through the instances. The actual choice of which clusters to split or merge is decided by a criteria function represented as either a similarity matrix or context vectors and cluster together instances that are more like each other than they are to the instances that belong to other clusters. Clustering algorithms are classified into three main categories, hierarchical, partitional, and hybrid methods. Frequently, research in machine learning (ML) of natural language takes the form of comparative ML experiments, either to investigate the role of different information sources in learning a task, or to investigate whether the bias of some learning algorithm fits the properties of natural language processing tasks better than alternative learning algorithms.

A. Knowledge based WSD

Work on WSD reached a turning point when large-scale lexical resources such as dictionaries, thesauri, and corpora became widely available. The work done earlier on WSD was theoretically interesting but practical only in extremely limited domains. Many researchers have used machine-readable dictionaries (MRDs) as a structured source of lexical knowledge to deal with WSD. These approaches, by exploiting the knowledge contained in the dictionaries, mainly seek to avoid the need for large amounts of training material. Most of them can be located in MRDs, and include part of speech, semantic word associations, syntactic cues, selection preferences, and frequency of senses, among others.

B. AI-Based Approaches

In the 1960's and 1970's, there was a lot of growth in AI research, and consequently, most of the methods that tackled WSD during this period used AI approaches. These systems relied on a wealth of both language and world knowledge, to determine the meaning of a word in context. Majority of these systems were grounded in language understanding theories and attempted to model deep knowledge of linguistic theory, especially in the area of syntax and semantics. Consequently, these systems tried to produce a semantic representation for an entire sentence in an attempt to capture its meaning, and from which word ambiguity problems would be solved. However, due to the pervasive nature of both structural and lexical ambiguity in natural language, a sentence can have several possible interpretations. In order to determine the correct

interpretation, these systems adopted a strategy of combining syntactic, semantic and world knowledge and enforcement of constraint satisfaction, to produce syntactic and semantic representation of an entire sentence. The scheme adopted for world knowledge representation as well as the process used to integrate syntactic, semantic and world knowledge, serve as the main distinguishing factors amongst these systems. \

C. Dictionary-based Approaches

In the 1980's, there was a surge in computing machinery and a corresponding increase in the availability of electronic linguistic resources, popularly known as MRDs, as most publishers started to produce electronic versions of their products. This precipitated the shift from AI-based systems to the emergence of dictionary-based approaches. MRDs presented a viable solution to the knowledge acquisition bottleneck facing AI-based approaches since they provided comprehensive lexical coverage of natural language. This meant that systems no longer suffered vocabulary limitations, spurring interest in language processing of unrestricted text. One of the first attempts to utilize these resources for WSD was Lesk (1986). His work was based on the observation that the coherence of a sentence is dependent on the cohesion of the words in it, meaning that the choice of one sense in a text is a function of the senses of the words close to it. He devised an algorithm that chooses the correct sense of a word by calculating the word overlap between the context sentence and the dictionary definition of the word in question. Lesk's work influenced most of the subsequent work in knowledge-based WSD Other machine readable resources that have been used in knowledge-based WSD include thesauri such as ROGET's thesaurus that has been used severally by different researchers including Masterman (1957) and Yarowsky (1992), and lexicons. A major hindrance to dictionary-based techniques such as those based on Lesk's idea is their crucial dependence on similarity in wording between a text and the MRD. Dictionary definitions are usually too short to generate an overlap from which an adequate set of indicators can be obtained. Also, despite their well-structured information and increased vocabulary coverage, pre-coded knowledge sources suffer from limitations in domain-specific coverage and in coping with the introduction of new words.

D. Corpus based WSD

WSD is one of the most important open problems in the Natural Language Processing. In the last fifteen years, empirical and statistical approaches have had a significantly increased impact on NLP. Of increasing interest are algorithms and techniques that come from the machine-learning (ML) community since these have been applied to a large variety of NLP tasks with remarkable success. The types of NLP problems initially addressed by statistical and machine-learning techniques are those of language ambiguity resolution, in which the correct interpretation should be selected from among a set of alternatives in a particular context (e.g., word-choice selection in speech recognition or machine translation, part-of-speech tagging, word-sense disambiguation, co-reference resolution, etc.). These techniques are particularly adequate for NLP because they can be regarded as classification problems, which have been studied extensively in the ML community. Regarding automatic WSD, one of the most successful approaches in the last ten years is supervised learning. Generally, supervised systems show better results in comparison to unsupervised ones, a conclusion that is based on experimental work and international competitions. This approach uses semantically annotated corpora to train machine learning (ML) algorithms to decide which word sense to choose in which contexts. The words in such annotated corpora are tagged manually using semantic classes taken from a particular lexical semantic resource (most commonly WordNet).

4. Related Work

Since the 1950s, many approaches have been proposed for assigning senses to words in context, although early attempts only served as models for toy systems. Currently, there are two main methodological approaches in this area: knowledge-based and corpus-based methods. Knowledge-based methods use external knowledge resources, which define explicit sense distinctions for assigning the correct sense of a word in context. Corpus-based methods use machine-learning techniques to induce models of word usages from large collections of text examples. Both knowledge-based and corpus-based methods present different benefits and drawbacks. Common problems faced in natural language processing are data sparseness and inconsistency in vocabulary. When the number of features increases, the sparseness is unavoidable. Smoothing is really required to overcome the above problem for improving the performance.

A. Azzini, C. da Costa Pereira, M. Dragoni, and A. G. B. Tettamanz [1] proposed a supervised approach to word sense disambiguation based on neural networks combined with evolutionary algorithms. Large tagged datasets for every sense of a polysemous word are considered, and used to evolve an optimized neural network that correctly disambiguates the sense of the given word considering the context in which it occurs. The viability of the approach has been demonstrated through experiments carried out on a representative set of polysemous words.

Rion Snow Sushant Prakash, Daniel Jurafsky, Andrew Y. Ng [2] formulated sense merging as a supervised learning problem, exploiting human-labeled sense clustering as training data. They train a discriminative classifier over a wide variety of features derived from WordNet structure, corpus-based evidence, and evidence from other lexical resources. Their learned similarity measure outperforms previously proposed automatic methods for sense clustering on the task of predicting human sense merging judgments, yielding an absolute F-score improvement of 4.1% on nouns, 13.6% on verbs, and 4.0% on adjectives. Finally, they propose a model for clustering sense taxonomies using the outputs of our classifier, and they make automatically sense-clustered Word Nets of various sense granularities.

Yoong Keok Lee and Hwee Tou Ng and Tee Kiah Chia[3] participated in the SENSEVAL-3 English lexical sample task and multilingual lexical sample task. They adopted a supervised learning approach with Support Vector Machines, using only the official training data provided. No

other external resources were used. The knowledge sources used were part of speech of neighboring words, single words in the surrounding context, local collocations, and syntactic relations.

Gerard Escudero, Llu´ıs M`arquez and German Rigau[6] described an experimental comparison between two standard supervised learning methods, namely Naïve Bayes and Exemplar–based classification, on the Word Sense Disambiguation (WSD) problem. The aim of the work is twofold. Firstly, it attempts to contribute to clarify some confusing information about the comparison between both methods appearing in the related literature. In doing so, several directions have been explored, including: testing several modifications of the basic learning algorithms and varying the feature space.

Dinakar Jayarajan [9] presented a new representation for documents based on lexical chains. This representation addresses both the problems achieves a significant reduction in the dimensionality and captures some of the semantics present in the data. They represent an improved algorithm to compute lexical chains and generate feature vectors using these chains.

Yee Seng Chan and Hwee Tou Ng, David Chiang[10] presented conflicting evidence on whether word sense disambiguation WSD) systems can help to improve the performance of statistical machine translation (MT) systems. In this paper, we successfully integrate a state-of-the-art WSD system into a state-of-the-art hierarchical phrase-based MT system. They show for the first time that integrating a WSD system improves the performance of a state-of-the- art statistical MT system on an actual translation task. Furthermore, the improvement is statistically significant.

Andres Montoyo, Armando Su´arez, German Rigau, Manuel Palomar [11] concentrated on the resolution of the lexical ambiguity that arises when a given word has several different meanings. This specific task is commonly referred to as word sense disambiguation (WSD). The task of WSD consists of assigning the correct sense to words using an electronic dictionary as the source of word definitions. They present two WSD methods based on two main methodological approaches in this research area: a knowledge-based method and a corpus-based method. Their hypothesis is that word-sense is ambiguities require several knowledge sources in order to solve the semantic ambiguity of the words.

S.K.Jayanthi and S. Prema[13] prompted a number of investigations into the relationship between information retrieval (IR) and lexical ambiguity in web mining. The work is such an exploration. Starting with a review of previous research that attempted to improve the representation of documents in IR systems, this research is reassessed in the light of word sense ambiguity. The results of these experiments lead to the conclusions that query size plays an important role in the relationship between ambiguity and IR in web content mining. Word Sense Disambiguation (WSD) is tested and analyzed for some of the existing Information Retrieval engines like Google, Clusty, yahoo, Altavista and msn search using Brill's tagger, and the derived results for the IR systems recommends how to accommodate the sense information in the selected document collection.

Antonio J Jimeno-Yepes [14], Alan R Aronson found that the graph-based approach, using the structure of the Meta thesaurus network to estimate the relevance of the Meta thesaurus concepts, does not perform well compared to the first two methods. In addition, the combination of methods improves the performance over the individual approaches. On the other hand, the performance is still below statistical learning trained on manually produced data and below the maximum frequency sense baseline.

P.Tamilselvi, S.K.Srivatsa [15] implemented disambiguation system with three different set of features with three different distance measuring functions combined with three different classifiers for word sense disambiguation. Using Neural Networks with enormous number of features, accuracy measured from 33.93% to 97.40% for words with more than two senses and 75% of accuracy for words with two senses.

M. Nameh, S.M. Fakhrahmad, M. Zolghadri Jahromi [17] presented a supervised learning method for WSD, which is based on Cosine Similarity. As the first step, they extract two sets of features; the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word. Then they presented the results of evaluating the proposed schemes and illustrate the effect of weighting strategies proposed.

A.R.Rezapour, S. M. Fakhrahmad and M. H. Sadreddini[18] presented a supervised learning method for WSD, which is based on K-Nearest Neighbor algorithm. They extracted two sets of features; the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word. In order to improve the classification accuracy, they proposed a feature weighting strategy. The results are encouraging comparing to state of the art.

Arindam Chatterjee, Salil Joshii, Pushpak Bhattacharyya, Diptesh Kanojia and Akhlesh Meena [19] shows that in almost all disambiguation algorithms, the sense distribution parameter P(S/W), where P is the probability of the sense of a word W being S, plays the deciding role. The widely reported accuracy figure of around 60% for all-words-domain-independent WSD is contributed to mainly by P(S/W), as one ablation test after another re-veals.Their experience of working with hu-man annotators who mark with WordNet sense ids, general and domain specific corpora brings to light the interesting fact that producing sense ids without looking at the context is a heavy cognitive load. Sense annotators do form hypothesis in their minds about the possible sense of a word, but then look at the context for clues to accept or reject the hypothesis. Such clues are minimal, just one or two words, but are critical nonetheless. Without these clues the annotator is left in an indecisive state as to whether or not to put down the first sense coming to his mind.

5. Algorithms Used

5.1 Naïve Bayesian Algorithm

Naive Bayes text classification is a supervised and probabilistic learning method. It calculates the probability of a document d being in class c by the following formula. P( ) is the conditional probability of term occurring in a document of class c. P(c) is the prior probability of a document occurring in class c.

The goal of classification is to find the best class for the document. The best class in naive bayes classification is the most likely or maximum a posteriori(MAP) class C_{map}

$C_{map} = \text{argmax}_c \, P(c|d) = \text{argmax}_c \, P(t_k|c)$

1) Preprocessing
   a. Segment input sentence
   b. Remove stop words from input
2) Multi sense lookup
   Lookup possible sense meanings of the ambiguous word from the corpus
3) Calculating Probability
   for all senses si of W do
   for all words fi in the vocabulary do
   P(fi|si) = C(fi,si)/C(si)
   end
   end
   for all senses si of W do
   P(si) = C(si) / N
   end
   4) Disambiguation
   for all senses si of W do
   score(si) = log P(si)
   for all words fi in the context window c do
   score(si) = score(si) + log P(fi|si)
   end
   end
   Choose s' = arg max score(si)

5.2 K-Means Clustering Algorithm

Simple K-Means is one of the simplest clustering algorithms K-Means algorithm is a classical clustering method that group large datasets in to clusters. The procedure follows a simple way to classify a given data set through a certain number of clusters. It select k points as initial centroids and find K clusters by assigning data instances to nearest centroids. Distance measure used to find centroids is Euclidean distance.

- Initially, the number of clusters must be known, or chosen, to be K say.
- The initial step is the choose a set of K instances as centres of the clusters. Often chosen such that the points are mutually "farthest apart", in some way.
- Next, the algorithm considers each instance and assigns it to the cluster which is closest.
- The cluster centroids are recalculated either after each instance assignment, or after the whole cycle of re-assignments.

5.3 Database: Wordnet

WordNet is a manually-constructed lexical system developed by George Miller at the Cognitive Science Laboratory at Princeton University . It reflects how human beings organize their lexical memories. The basic building block of WordNet is synset consisting of all the words that express a given concept. Synsets, which senses are manually classified into, denote synonym sets. Within each synset, the senses, although from different keywords, denote the same meaning.

- A detailed database of semantic relationships between English words.
- Developed by famous cognitive psychologist George Miller and a team at Princeton University.
- About 144,000 English words.

DESCRIPTION

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Polysemy information

POS	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	101863	15935	44449
Verb	6277	5252	18770
Adjective	16503	4976	14399
Adverb	3748	733	1832
Totals	128391	26896	79450

6. Experimental Setup

6.1 Precision and Recall

Precision and **recall** are the basic measures used in evaluating search strategies. As shown in below diagram, these measures assume:

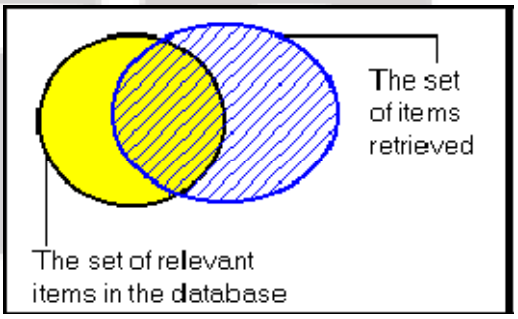

Figure 6.1

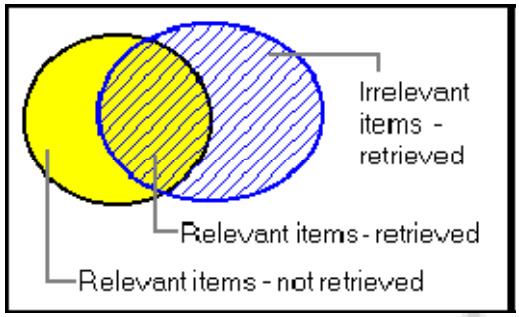

Figure 6.2: Representation of database and retrieve it.

There is a set of records in the database which is relevant to the search topic. Records are assumed to be either relevant or irrelevant *(these measures do not allow for degrees of relevancy)*. The actual retrieval set may not perfectly match the set of relevant records.

RECALL: is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

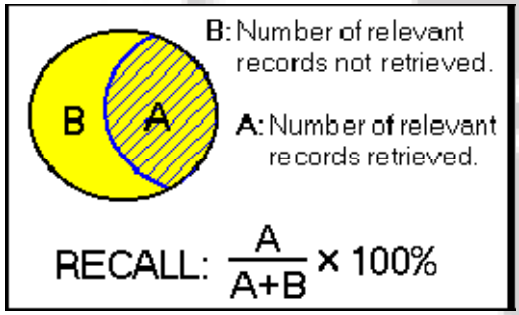

Fig-6.3 Recall Retrieved Record

PRECISION: is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

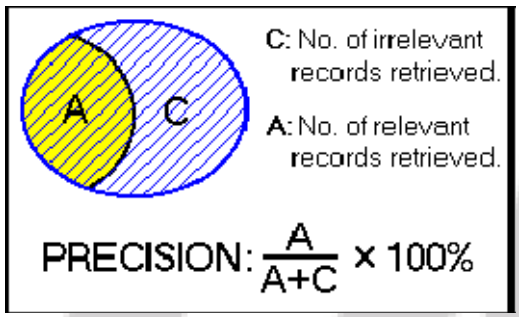

Fig-6.4 Precision Retrieved Record

6.2 Performance Measures

- **Precision** : *p* is the number of correct results divided by the number of all returned results
- **Recall:** *r* is the number of correct results divided by the number of results that should have been returned

Example 1:- Suppose a program for recognizing dogs in scenes identifies 7 dogs in a scene containing 9 dogs and some cats.
If 4 of the identifications are correct, but 3 are actually cats,
- the program's **precision** is 4/7
- While its **recall** is 4/9.

When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages,
precision is : 20/30 = 2/3
While its **recall** is: 20/60 = 1/3.

Example 2:- Assume the following:
- A database contains **80** records on a particular topic
- A search was conducted on that topic and **60** records were retrieved.
- Of the 60 records retrieved, **45** were relevant.

Calculate the **precision** and **recall** scores for the search.

Solution:

Using the designations above:
- A = the number of relevant records retrieved,
- B = the number of relevant records not retrieved, and
- C = the number of irrelevant records retrieved.

In this example A = 45, B = 35 *(80-45)* and C = 15 *(60-45)*.

- **Recall** = (45 / (45 + 35)) * 100% => 45/80 *100% = **56%**
- **Precision** = (45 / (45 + 15)) * 100% => 45/60 * 100% = **75%**
- The **F score** (also **F-measure**) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score :

$$F\ score = 2 * \frac{(Precision \cdot Recall)}{(Precision + Recall)}$$

7. Results and Discussions

We presented a creation of database that use the contextual words and the sense of ambiguous word. In this we have entered nearly about 20 sentence of word “bank” actual results are comes.
Noun
- S: (n) bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"
- S: (n) depository financial institution, bank, banking concern, banking company (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"
- S: (n) bank (a long ridge or pile) "a huge bank of earth"
- S: (n) bank (an arrangement of similar objects in a row or in tiers) "he operated a bank of switches"
- S: (n) bank (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) bank (the funds held by a gambling house or the dealer in some gambling games) "he tried to break the bank at Monte Carlo"
- S: (n) bank, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, bank (a container (usually with a slot in the top) for keeping money at home) "the coin bank was empty"

- S: (n) bank, bank building (a building in which the business of banking transacted) "the bank is on the corner of Nassau and Witherspoon"
- S: (n) bank (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) "the plane went into a steep bank"

Verb

- S: (v) bank (tip laterally) "the pilot had to bank the aircraft"
- S: (v) bank (enclose with a bank) "bank roads"
- S: (v) bank (do business with a bank or keep an account at a bank) "Where do you bank in this town?"
- S: (v) bank (act as the banker in a game or in gambling)
- S: (v) bank (be in the banking business)
- S: (v) deposit, bank (put into a bank account) "She deposits her paycheck every month"
- S: (v) bank (cover with ashes so to control the rate of burning) "bank a fire"
- S: (v) count, bet, depend, swear, rely, bank, look, calculate, reckon (have faith or confidence in) "you can count on me to help you any time"; "Look to your friends for support"; "You can bet on that!"; "Depend on your family in times of crisis". We created a database of different sense of the ambiguous word in different contextual word this data base can be further to improve the accuracy of the system. We have implement a scheme in this we have given a text it built a semantics representation of text. The semantic is generated so that it can associate with ambiguous word. Meanings of ambiguous word are search in database and we can filter out that search by matching the database with context word. It will result much more definite reputed of the word for text.

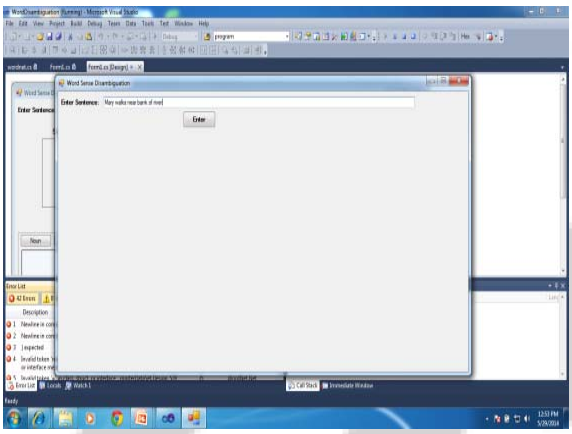

Figure 7.1

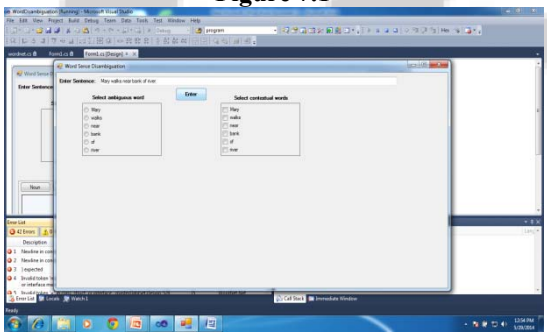

Figure 7.2

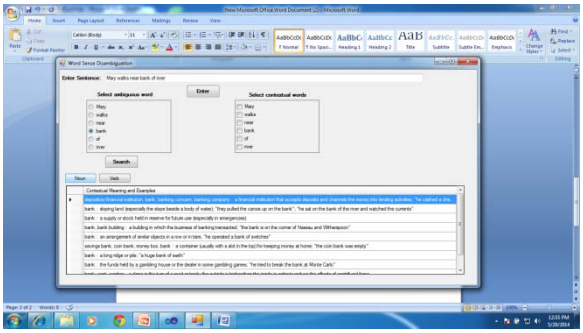

Figure 7.3

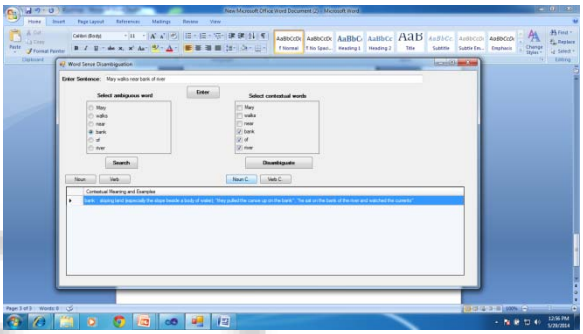

Figure 7.4

- This scheme is more suitable to the task of WSD where the ambiguous word plays a key role in representing the text of its sense as these text is ambiguous we use tokenize to construct more specific text representation.

Table 7.1

Words	Correctly Identified Sense (A)	Incorrectly Identified Sense (B)	Not identified (C)	Recall (A/ (A+B))*100 %	Precision (A/ (A+C))*100 %	F-score 2*(Precision *recall)/ (precision+ recall)
Bank	27	5	0	84	100	91
Bass	10	3	4	77	71	73
Coach	4	0	3	100	57	73
Cut	47	11	5	81	90	85
Play	19	2	0	90	100	95
Light	36	12	10	75	78	76
Free	21	9	3	70	88	78
Step	23	4	1	85	96	90
Field	7	3	2	70	78	74
Attack	16	0	1	100	94	97
Master	12	4	3	75	80	77

In this thesis, we have implemented a database for WSD. In this implementation process we have flow step by step process. Before implementation work we have study the approaches of WSD: Corpus-based and knowledge base approach for supervised, unsupervised and semi-supervised. In the database implementation work we have taking any ambiguous word sentence and we have gives the actual sense of that particular word. In this last we have discuss the result through precision, recall and F- score of our system with the help of graphically representation. In this we have use tools and technology of visual studio 2010 and SQL server 2008.

8. Future Scope

- This work can be extended for Hindi Language and other regional languages.
- Can be implemented using other advance machine learning algorithm.
- Other corpus based approach can be used to perform WSD tasks

References

[1] Azzini, C. da Costa Pereira, M. Dragoni, and A. G. B. Tettamanzi," Evolving Neural Networks for Word Sense Disambiguation", Eighth International Conference on Hybrid Intelligent Systems, 2008.
[2] Rion Snow Sushant Prakash, Daniel Jurafsky, Andrew Y. Ng ,"Learning to Merge Word Senses", Computer Science Department Stanford University,2007.
[3] Yoong Keok Lee and Hwee Tou Ng and Tee Kiah Chia," Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources", Department of Computer Science National University of Singapore, 2004.
[4] Dan Klein, Kristina Toutanova, H. Tolga Ilhan, Sepandar D. Kamvar and Christopher D. Manning," Combining Heterogeneous Classifiers forWord-Sense Disambiguation", Computer Science Department Stanford University, 2002.
[5] T. Theodosiou1, N. Darzentas,L. Angelis1 and C. A. Ouzounis," PuReD-MCL: a graph-based PubMed document clustering methodology", Vol. 24 no. 17, 2008.
[6] Gerard Escudero, Llu´ıs M`arquez and German Rigau," Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited", Proceedings of th 14th European Conference, 2000.
[7] David Martinez Iraolak," Supervised Word Sense Disambiguation: Facing Current Challenges", Informatikan Doktore titulua eskuratzeko aurkezturiko Tesia Donostia, 2004.
[8] Rada Mihalcea," Word Sense Disambiguation", the 18th European Summer School in Logic, Language and Information 31 July - 11 August, 2006.
[9] Dinakar Jayarajan," Using Semantics in Document Representation: A Lexical Chain Approach ", Department of Computer Science and Engineering Indian Institute of technology Madras, June 2009.
[10] Yee Seng Chan and Hwee Tou Ng, David Chiang,"Word Sense Disambiguation Improves Statistical Machine Translation", Department of Computer Science National University of Singapore, 2007.
[11] Andres Montoyo, Armando Su´arez, German Rigau,"Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods", Journal of Artificial Intelligence Research ,2005.
[12] Hwee Tou Ng," Exemplar-Based Word Sense Disambiguation: Some Recent Improvements", DSO National Laboratories 20 Science Park Drive Singapore, 1996.
[13] S.K.Jayanthi and S. Prema," Word Sense Disambiguation in Web Content Mining Using Brill's Tagger Technique", International Journal of Computer and Electrical Engineering, Vol. 3, June 2011.
[14] Antonio J Jimeno-Yepes, Alan R Aronson, "Knowledge-based biomedical word sense disambiguation: comparison of approaches", Jimeno-Yepes and Aronson BMC Bioinformatics 2010
[15] P.Tamilselvi , S.K.Srivatsa ," Case Based Word Sense Disambiguation Using Optimal Features", IPCSIT vol.16, IACSIT Press, Singapore,2011
[16] David Martinez, Oier Lopez de Lacalle, Eneko Agirre," On the Use of Automatically Acquired Examples for All-Nouns Word Sense Disambiguation", University of the Basque Country 20018, Journal of Artificial Intelligence Research 33,79-107,2008.
[17] M. Nameh, S.M. Fakhrahmad, M. Zolghadri Jahromi," A New Approach to Word Sense Disambiguation Based on Context Similarity", Proceedings of the World Congress on Engineering 2011 Vol I, July 6 - 8, 2011
[18] R. Rezapour, S. M. Fakhrahmad and M. H. Sadreddini ," Applying Weighted KNN to Word Sense Disambiguation", Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011, July 6 - 8, 2011.
[19] Arindam Chatterjee, Salil Joshii, Pushpak Bhattacharyya, Diptesh Kanojia and Akhlesh Meena, "A Study of the Sense Annotation Process: Man v/s Machine", International Conference on Global Wordnets , Matsue, Japan,, Jan, 2012.

Author Profile

Neetu Sharma, has completed BE(CSE) in 1996. and M. Tech (CSE) completed in 2006 from Panjab University. Presently she is doing Ph. D. in Computer Science and Engineering. Her total teaching experience is 15 years. She is presented 2 papers in National and three papers in International conferences sponsored by IEEE**.**

Prof.S. Niranjan, did his M.Tech (Computer Engg.) from IIT Kharagpur in 1987. Completed Ph.D.(CSE) in 2004 and Ph.D.(I&CT) in 2007, Have total 26 years of teaching experience. Presently working as Principal in PDM College of Engg. Bahadurgarh (Haryana). Having 6 publications in journals and 30 other papers published in conferences