Abstract: Now days, Word Sense Disambiguation (WSD) is a vital area which is very useful in today’s world. Many WSD algorithms are available in literature; we have chosen to an optimal and portable WSD algorithm. We are discussed the supervised, unsupervised, and knowledge-based approaches for WSD. In this paper we are discuses that association rules, Knowledge-based WSD, Corpus-based WSD.

Keywords: Association rules, Supervised, Unsupervised, Knowledge-based, WSD.
1. Introduction:

A. Word Sense Disambiguation:

In natural language processing (NLP), word sense disambiguation (WSD) is defined as the task of assigning the appropriate meaning (sense) to a given word in a text or discourse. A word can have more than one sense. The sense in which the word is used can be determined, most of the times, by the context in which the word occurs. Consider the widely used example of the word bank. Bank has several senses out of which bank as a financial institution and bank used as a sloping land bordering a river can be easily distinguished from the context. But distinguishing the word bank as a financial institution and bank as a building housing such an institution is more difficult. The processes of identifying the correct sense of words in context are called word sense disambiguation (WSD).

Word Sense Disambiguation is the process of differentiating among the senses of words. Machine translation is one of the most former and on growing research computational linguistic. In 1940’s WSD was developed as discrete field in computational linguistic due to fast research in of machine translation. In 1950’s Weaver acknowledged that context is crucial and recognized the basic statistical character of the problem in proposing that statistical semantic studies should be undertaken as a necessary primary step. The automatic disambiguation of word senses has been an interest and concerned since the earliest days of computer treatment of languages in the 1950’s. Then identifying work in estimating the degree of ambiguity in texts and bilingual dictionaries and applying simple statistical models. Sense disambiguation is an intermediate task which is not an end in itself, but rather is necessary at one level or another to accomplish most NLP task [1].

B. Graph-based WSD:

The graph based methods make the most of the semantic model they employ, thus trying to close the gap between the unsupervised and supervised approaches.

It describes the latest state-of the art methods for unsupervised graph-based word sense disambiguation. Next, this paper presents several comparative evaluations carried on the sensual data sets using the same semantic representation [3].

C. Some Background on WSD:

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.

The problem of word sense disambiguation (WSD) is defined by Sinha et al.[2] as the task of automatically assigning the most appropriate meaning to a polysemous word within a given context.

WSD methods are critical for solving natural language processing tasks like machine translation and speech processing, but also boost the performance of other tasks like text retrieval, document classification and document clustering. Approaches found in the bibliography face the tradeoff between unsupervised and supervised methods: the first one has fast execution time, but low accuracy and the second one require training in a large amount of manually annotated data.

D. Knowledge-based WSD:

Work on WSD reached a turning point in the 1980s and 1990s 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. Since Lesk (1986), 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. Agirre and Martinez (2001b) distinguish ten different types of information that can be useful for WSD. 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.

E. Corpus-based WSD:

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 reader can find an excellent introduction to ML, and its relation to NLP, in the articles by Mitchell (1997), Manning and Schütze (1999), and
Cardie and Mooney (1999), respectively. 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 from examples, in which statistical or ML classification models are induced from semantically annotated corpora. Generally, supervised systems have obtained better results than unsupervised ones, a conclusion that is based on experimental work and international competitions2. 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). Many standard ML techniques have been tried, including Bayesian learning (Bruce & Wiebe, 1994), Maximum Entropy (Suárez & Palomar, 2002a), exemplar-based learning (Ng, 1997; Hoste et al., 2002), decision lists (Yarowsky, 1994; Agirre & Martinez, 2001a), neural networks (Towell & Voorhees, 1998), and, recently, margin-based classifiers like boosting (Escudero et al., 2000) and support vector machines (Cabezas et al., 2001). Corpus-based methods are called “supervised” when they learn from previously sense-annotated data, and therefore they usually require a large amount of human intervention to annotate the training data (Ng, 1997). Although several attempts have been made (e.g., Leacock et al., 1998; Mihalcea & Moldovan, 1999; Cuadros et al., 2004), the knowledge acquisition bottleneck (too many languages, too many words, too many senses, too many examples per sense) is still an open problem that poses serious challenges to the supervised learning approach for 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 (Schütze 1998). 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.

2. Association Rules:
To mine the association rules is to discover the important relevance between the terms in a transaction database. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. In addition to the antecedent (the "if" part) and the consequent (the "then" part), an association rule has two numbers that express the degree of uncertainty about the rule. In association analysis the antecedent and consequent are sets of items (called item sets) that are disjoint (do not have any items in common). The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. (The support is sometimes expressed as a percentage of the total number of records in the database.)

The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

The task of mining association rules is to discover all of the strong association rules. The support degree of the rules is no less than the threshold of support degree and the confidence degree is no less than the threshold of the confidence degree. From formula (2), the support degree reflects the frequency of the relation X=>Y occurring in the transaction database D. The item set, which frequency is more than the threshold, is a frequent item sets. So we think the production of the frequent item sets is because there is some correlation between the corresponding items.
As text is an unstructured source of information, to make it a suitable input to an automatic method it is usually transformed into a structured format. To this end, a preprocessing of the input text is usually performed, which typically (but not necessarily) includes the following steps:

— *tokenization*, a normalization step, which splits up the text into a set of tokens (usually words);

— *part-of-speech tagging*, consisting in the assignment of a grammatical category to each word (e.g., “the/DT bar/NN was/VBD crowded/JJ,” where DT, NN, VBD and JJ are tags for determiners, nouns, verbs, and adjectives, respectively);

— *lemmatization*, that is, the reduction of morphological variants to their base form (e.g. was → be, bars → bar);

— *chunking*, which consists of dividing a text in syntactically correlated parts (e.g., [the bar] NP [was crowded] VP, respectively the noun phrase and the verb phrase of the example).

— *parsing*, whose aim is to identify the syntactic structure of a sentence (usually involving the generation of a parse tree of the sentence structure).

We report an example of the processing flow in Figure. As a result of the preprocessing phase of a portion of text (e.g., a sentence, a paragraph, a full document, etc.), each word can be represented as a vector of features of different kinds or in more structured ways, for example, as a tree or a graph of the relations between words. The representation of a word in context is the main support, together with additional knowledge resources, for allowing automatic methods to choose the appropriate sense from a reference inventory.
Figure 1: An example of preprocessing steps of text.

A set of features is chosen to represent the context. These include (but are not limited to) information resulting from the above-mentioned preprocessing steps, such as part of speech tags, grammatical relations, lemmas, etc. We can group these features as follows:

—local features, which represent the local context of a word usage, that is, features of a small number of words surrounding the target word, including part-of-speech tags, word forms, positions with respect to the target word, etc.;

—topical features, which—in contrast to local features—define the general topic of a text or discourse, thus representing more general contexts (e.g., a window of words, a sentence, a phrase, a paragraph, etc.), usually as bags of words;

—syntactic features, representing syntactic cues and argument-head relations between the target word and other words within the same sentence (note that these words might be outside the local context);

—semantic features, representing semantic information, such as previously established senses of words in context, domain indicators, etc.

3. 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 (NBWSD)

WSD is:

✓ Describe to be service of society.
✓ To get intellectual joy of work of doing some creative work.
✓ Accessible to anyone with an interest in NLP.
✓ Persuade we to work on word sense disambiguation
✓ It’s an interesting problem
✓ Lots of good work already done, still more to do

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.

4. **Problem Statement:**

Based on all of the previous research, we propose an approach to create the to construct an association rules – based database for WSD using association rules, which be used to mine the rules to find the sense of the ambiguous word.

5. **Literature Survey:**
Yong-le SUN and Ke-liang JIA[4] proposed a new WSD method based on the mining association rules, which can mine the association rules between the sense of the ambiguous word and its context, to construct an association rules – based database. At last the sense of the ambiguous word is determined by choosing the sense which the most association rules.

Sasi Kanth Ala and Narayana Murthy Kavi[5] proposed a method for doing This approach uses both lexical and syntactic information to do Word unrestricted WSD using association rules extracted from a sense tagged corpus. The lexical and syntactic features are extracted from within a sentence in which the target word lies. We show that high accuracy can be obtained by exploiting the accuracy coverage trade off.  We also show that there is a significant increase in performance when syntactic features are used in addition to lexical features.

Min Song, Il-Yeol Song, Xiaohua Hu and Robert Allen[6] presented a novel semantic query expansion technique that combines association rules with ontologies and information retrieval techniques. They proposed to use the association rule discovery to find good candidate terms to improve the retrieval performance. These candidate terms are automatically derived from collections and added to the original query.

Rion Snow Sushant Prakash, Daniel Jurafsky, Andrew Y. Ng[7] formulated sense merging as a supervised learning problem, exploiting human-labeled sense clusterings 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.

Andres Montoyo, Armando Suárez, German Rigau, Manuel Palomar[8] 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 an ambiguity requires several knowledge sources in order to solve the semantic ambiguity of the words.

Dan Klein, Kristina Toutanova, H. Tolga Ilhan[9], discussed ensembles of simple but heterogeneous classifiers for word-sense disambiguation, Examining the Stanford-CS224N system entered in the SENSEVAL-2 English lexical sample task. First-order classifiers are combined by a second-order classifier, which variously uses majority voting, weighted voting, or a maximum entropy model. While individual first-order classifiers perform comparably to middle-scoring teams’ systems, the combination achieves high performance. They discuss trade-offs and empirical performance. Finally, they present an analysis of the combination, examining how ensemble performance depends on error independence and task difficulty.

Dinakar Jayarajan [10] 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[11] 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, Hiero. 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.

6. Research Methodology:
Researchers organize their research by formulating and defining a research problem. This helps them focus the research process so that they can draw conclusion reflecting the real world in the best possible way. Research methodology involves the research providing a research hypothesis that has to be proved. Time money, feasibility, ethics, and availability to measure the phenomenon correctly are examples of issue construing the research. Choosing the scientific measurement are also crucial for getting the correct conclusion. Test must be conducted to test a hypothesis. Drawing a conclusion is based on several factors of the
research, it has to be based on the validity and reliability of the measurement, and how good the measurement was to reflect the real world and what more could have affected the results.

7. Conclusion:

In this paper we the field of word sense disambiguation (WSD). WSD is a hard task as it deals with the full complexities of language and aims at identifying a semantic structure using association rules from apparently unstructured sources of text. The hardness of WSD strictly depends on the granularity of the sense distinctions taken into account. Supervised methods undoubtedly perform better than other approaches. However, relying on the availability of large training corpora for different domains, languages, and tasks is not a realistic assumption. This paper will also furnish an idea of few of the WSD algorithms and their performances, which compares and assess the need of the word sense disambiguates.

There is a general agreement that WSD needs to show its relevance in vivo, that is, in applications such as information retrieval or machine translation. On the one hand, the Community must not discontinue in vitro (i.e., stand-alone) evaluations of WSD, as there are still unclear points to be settled. On the other hand, full-fledged applications should be built including WSD either as an integrated or a pluggable component. Some of the tasks in the Semeval competition went in this direction. Also, theoretical experiments could be performed to determine more precisely which minimum WSD performance (90%, 95%, 100% accuracy?) is needed to enable which application.

ACKNOWLEDGMENT:

I would like to thank all faculty members of CSE Department for their kind attention and valuable guidance which proved as a constant source of inspiration at each stage and also provide the friendly association to encouragements of free and frank discussion.

Finally, I would like to thank my parents for their constant love and support and for providing me with the opportunity and the encouragement to pursue my goals.
References:

6. Min Song, Il-Yeol Song, Xiaohua Hu and Robert Allen “Semantic Query Expansion Combining Association Rules with Ontologies and Information Retrieval Techniques” 7th International Congress on Data Warehouse and Knowledge Discovery (DAWAK’05), Volume 63, Issue 1, October 2007, Pages 63-75
7. Rion Snow Sushant Prakash, Daniel Jurafsky, Andrew Y. Ng ,”Learning to Merge Word Senses”, Computer Science Department Stanford University
8. Andres Montoyo, Armando Su´arez, German Rigau,“Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods”, Journal of Artificial Intelligence Research,2005
11. Yee Seng Chan and Hwee Tou Ng, David Chiang,“Word Sense Disambiguation Improves Statistical Machine Translation”, Department of Computer Science National University of Singapore.,