A Review of Literature on Word Sense Disambiguation

Rakesh Kumar[1], Ravinder Khanna[2], Vishal Goyal[3]

[1]Research Scholar, Punjab Technical University, Kapurthala (India) E-Mail: rakesh77kumar@yahoo.com
[2]Principal, Sachdeva Engg. College for Girls, Gharuan, Mohali (India), E-Mail: ravikh_2006@yahoo.com
[3]Dept. of Computer Science, Punjabi University, Patiala (India) , E-Mail: vishal.pup@gmail.com

Abstract
Artificial intelligence (AI) has been a major research area in the later quarter of 20th century and is likely to be even more so in the 21st century. A key part of AI is Word Sense Disambiguation (WSD) which deals with choosing the correct sense of a word in the given text. All human languages have words with multiple meaning and selecting the intended sense is important. This paper briefly describes various methods presently used for WSD and their relative effectiveness. WSD applications currently find application in Information Retrieval, Information Extraction, Automated Answering Machine, Speech Reorganization, Machine Translation among many others. WSD has promise for the future in taking AI to the next higher level.
Kywords: Natural Language Processing (NLP), Artificial Intelligence (AI), Word Sense Disambiguation (WSD), Knowledge Based Methods, Supervised/Unsupervised Methods.

1. Introduction

Word Sense Disambiguation (WSD) is one of the most challenging and active research areas of Natural Language Processing (NLP), also referred as Natural Language Engineering (NLE). Ambiguity is found in all natural (human) languages, wherein words have multiple meanings. For example, the English language word 'fair' can mean blonde or treating people equally without favouritism. Similarly Punjabi language word ਸੁਰ (sur) can mean ਧੁਨੀ (dhunī) or ਏਕਤਾ ēkatā or ਨਾਸ (nās) or ਪਤਾ (patā). WSD is the process to determine which meaning of a word is used in the given sentence. It automatically assigns the most appropriate meaning to the ambiguous word in a given context. WSD contributes to various applications in NLP for which it is potentially an issue i.e. for Machine Translation, Information Retrieval, Question Answering and Dialogues etc. [1,2]. Figure 1 shows the working principal of WSD.

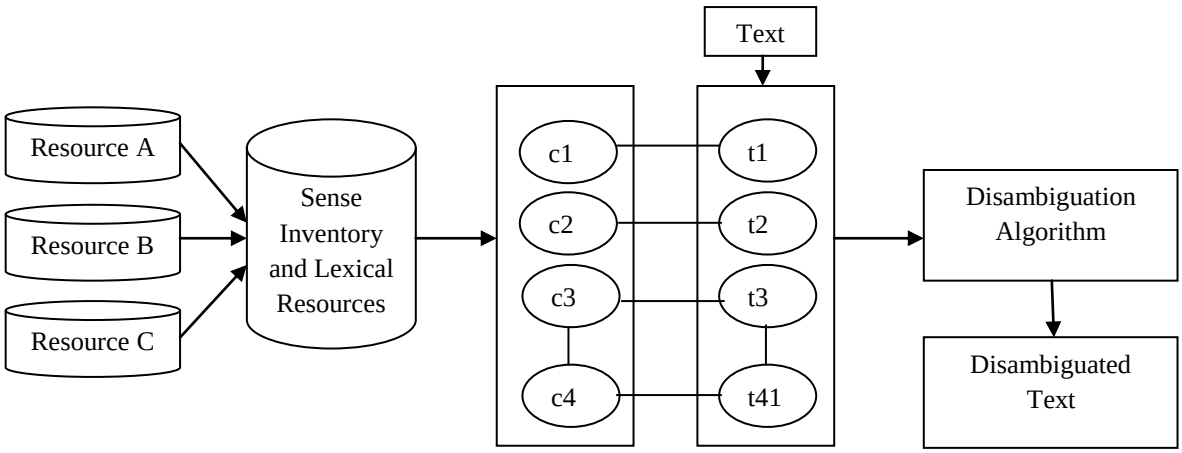

Figure 1: General Model of WSD [3]

A lot of research has been carried out in WSD in many foreign universities like Stanford University USA, National University Singapore etc. In India (at IIT Kanpur, IIT Bombay, IISc. Bengaluru, Punjabi University, Patiala etc.) the emphasis is on WSD in regional languages.

2. Literature Review

Research on WSD was started during the late 1940s [4]. Kaplan (1950) [5] determined that in a particular context two words on either side of an ambiguous word determine the sense of the word. Bar-Hillel (1964) [6] used WSD as a part of Machine Translation (MT). He pointed out that without a "Universal Encyclopedia", a machine would never be able to distinguish between the many meanings of a word. Madhu, Lytle (1965) [7] proposed that the figure of merit (sense) can be determined by a consideration of the probability of occurrence of a target word (ambiguous word) in the field (area of usage/specialization). The frequency with which a target equivalent occurs in one field is,

in general, different from that in another field. In other words, they considered the sense frequencies for different domains. They obtained probabilities of each sense given the context using Bayes' Rule. Wilks (1975) [8] developed a model on "preference semantics", where the selectional restrictions and a frame-based lexical semantics were used to find the exact sense of an ambiguous word. Lesk (1986) [9] proposed his algorithm based on overlaps between the glosses (Dictionary definitions) of the words in a sentence. The maximum number of overlaps represents the desired sense of the ambiguous word. In this approach the Oxford Advanced Learner's Dictionary of Current English (OALD) was used to obtain the dictionary definitions. This approach had shown the way to the other Dictionary-based WSD works. Miller (1990) [10] invented WordNet and brought a revolution in WSD because it was both programmatically accessible and hierarchically organized into word senses called synsets. Brown, Lai, Mercer (1991) [11] implemented corpus based Word Sense Disambiguation first time. Brown , Pietra (1992) [12] proposed a statistical model (trigram model n=3) to determine the sense of the word using a much larger text. Brown, Della Pietra S.A, Della Pietra V.J., and Mercer (1993) [13] developed five statistical models of translating (sentence to sentence) from one language to another and showed that it is possible to estimate their parameters automatically from a set of pairs of sentences. It was also shown that it is possible to align the words within pairs of sentences algorithmically. Their work mainly was related to English-French machine translation. However their algorithms had minimal linguistic content and thus can be used on any other pair of language. Lu, Z., Liu, Ting., Li, Sheng. (1996) [14] presented an input model of Neural Network that calculates the Mutual Information between contextual words and ambiguous word by using statistical method and taking the contextual words to certain number beside the ambiguous word according to (-M, +N). The experiment adopted triple-layer BP Neural Network model and proves how the size of training set and the value of M and N affect the performance of Neural Network model. The experimental objects are six pseudowords owning three word-senses constructed according to certain principles. SchUtze (1998) [15] presented context-group discrimination, a disambiguation algorithm based on clustering. Senses are interpreted as groups (or clusters) of similar contexts of the ambiguous word. Diab, Resnik (2002) [16] proposed using large bilingual corpora to improve performance on word sense disambiguation. The main idea is that knowing a French word may help determine the meaning of the corresponding English word. They apply this intuition to the SENSEVAL word disambiguation task by running off-the-shelf translators to produce translations which they then use for disambiguation. Michael (2003) [17] presented two approaches to German parsing (n-gram based machine learning and cascaded finite-state parsing), and evaluated them on the basis of a large amount of data. Montoyo, Suarez, Rigau, Palomar (2005) [18] explored some methods of collaboration between complementary knowledge-based and corpus-based WSD methods. Two complementary methods have been presented: specification marks (SM) and maximum entropy (ME). Mihalcea (2005) [19] introduced a graph-based algorithm for sequence data labeling, using random walks on graphs encoding label dependencies. The algorithm is illustrated and tested in the context of an unsupervised word sense disambiguation problem, and shown to significantly outperform the accuracy achieved through individual label assignment, as measured on standard sense annotated data sets. Vickrey, Biewald, Teyssier, Koller (2005) [20] considerd the related task of word translation, and found the correct translation of a word from context. Bilingual corpora of French-English such as the European Parliament proceedings was used to disambiguate word in given context. Sinha , Mihalcea (2007) [21] presented graph-based algorithm for unsupervised word sense disambiguation. The algorithm annotates all the words in a text by exploiting similarities identified among word senses, and using centrality algorithms applied on the graphs encoding these sense dependencies. They experimented with six knowledge-based measures of similarity and four graph centrality algorithms. Josan and Lehal (2008) [22] presented a approach to find out whether higher order n gram models improves the word sense disambiguation in Punjabi language and whether it has any relation with entropy of the models. In their experiments statistical analysis of n gram models (for n ranging from ±1 to ±6) is carried out. Authors also tried to explore the possibility of disambiguation by using future knowledge. It became clear that lower order n gram models (n=±1, ±2, ±3) are sufficient for word sense disambiguation and larger n gram model gives little improvement. Mishra, Yadav, Siddiqui (2009) [23] presented an unsupervised word sense disambiguation algorithm for Hindi. The algorithm uses a decision list using untagged instances. Some seed instances are provided manually. Stemming was applied and stop words were removed from the context. The list was then used for annotating an ambiguous word with its correct sense in a given context. Deepti Goyal, Deepika Goyal,. Singh (2010) [24] presented a hybrid approach for this problem based on the basic principle by Yarowsky's unsupervised algorithm for WSD. It also employed Naïve Baye's theorem to find the likelihood ratio of the sense in the given context. Broda, Mazur (2010) [25] focuseed on evaluation of a selected clustering algorithms (K-Means, K-Medoids, hierarchical agglomerative clustering, hierarchical divisive clustering, Growing Hierarchical Self Organising Maps, graph-partitioning based clustering) in task of Word Sense Disambiguation for Polish. Navigli, Lapata (2010) [26] introduced a graph-based WSD algorithm which has few parameters and does not require sense-annotated data for training. Using this algorithm, it also investigated several measures of graph connectivity with the

aim of identifying those best suited for WSD. It was examined how the chosen lexicon and its connectivity influences WSD performance. Nameh, Fakhrahmad, Jahromi (2011) [27] presented a supervised learning method for WSD, which was based on cosine similarity. It was based on inner product of vectors algorithm. Sense-tagged data was used to train the classifier. At the first step, extracts two set of features; the set of words that have co-occurred with the ambiguous word in the text frequently, and the set of words surrounding the ambiguous word.

3. Approaches to WSD

3.1. Dictionary and Knowledge Based Methods

These methods rely on knowledge resources of Machine Readable Dictionaries [MRDs] like WordNet and Thesaurus etc. They may use grammar rules and/or hand coded rules for disambiguation. In recent years, many dictionaries are made available in MRD format like Oxford English Dictionary (OED), Collins Dictionary (CD), Longman Dictionary of Ordinary Contemporary English (LDOCE) and Thesauruses which add synonymy information like Roget Thesaurus. MRD formats include a list of meanings, definitions (for all word meanings), and typical usage examples (for most word meanings), while a thesaurus adds an explicit synonymy between word meanings and the semantic network. 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 words in "pine cone", the definitions of the appropriate senses include the words evergreen and tree (at least in one dictionary) [1, 28, 29].

3.2 Supervised Methods

In these methods a labeled corpus based training data is used for WSD. Here, information is gained from training on some corpus. A corpus provides a set of samples that enables the system to develop some numerical models. Supervised methods are based on the assumption that the context can provide enough evidence on its own to disambiguate words (hence, world knowledge and reasoning are deemed unnecessary). These methods are subject to a new knowledge acquisition bottleneck since they rely on substantial amounts of manually sense-tagged corpora for training, which are laborious and expensive to create [30]. Generally, supervised approaches to WSD have found better results than other approaches. Some of the algorithms used in supervised methods are given below:

3.2.1. Naïve Bayesian Classifiers

A Naïve Bayesian classifier is well known method in machine learning community for good result and performance of word sense disambiguation. This algorithm is based on statistical methods and determines probabilistic parameters for disambiguation. It is based on the application of Baye's theorem in which joint probability of each sense Xi of a word W over the features defined (X1, X2,…Xn) in the given context is determined and the maximum value of joint probability is chosen for the correct sense of word the using the trained annotated corpora [1].

3.2.2 Decision Tree and Decision List Method

Decision tree and Decision list is one of the prominent methods for word sense disambiguation. It uses selective rules associated with each word sense. In this approach the system selects one or more rules which satisfies features and assign sense to the ambiguous word based on their prediction. It is a word specific classifier and a separate classifier needs to be trained for each word. This approach can be considered as weighted 'yes' or 'no' rules where the exceptional conditions appear at the root node of the list with high weight and the general condition of the list appear at bottom with low weights. Default condition of the list accepts all remaining. A scoring function calculates the weight which describe the association between the condition and the particular class and they are estimated from the trained corpus [1].

3.2.3 Support Vector Machine (SVM) Method

Support Vector Machine is a Kernal based techniques which represent a major development in machine learning algorithms and can be applied to classification or regression. This method introduced by Boser et al in 1992 [31] which is based on the idea of learning linear hyper plane from the training set that separates positive samples from the negative samples.

3.2.4 Neural network method

McCulloch and Pitts proposed neural network model in 1943 [32] which is an interconnected group of artificial neurons. It requires a large amount of hand written data for training and it is not clear that the same neural network model will be applicable to real world application .The main aim of this method is to make use of input features to partition the training contexts in non overlapping sets corresponding to the desired response [33, 34]. Inputs are provided with adjusted weight between neurons (nodes) so that the desire output is having larger activation than

other outputs. Major problems that occur in neural networks are: difficulties in interpreting the results, the need for a large quantity of training data and the tuning of parameters such as thresholds, decay, etc.

3.3 Unsupervised Methods

These are based on unlabeled corpus. The unlabeled corpus is required to train before using it on ambiguous words. Thus it avoids the use of labeled corpus which is a long drawn out and expensive process (this is called knowledge acquisition bottleneck). Unsupervised approach shows this problem by introducing concept that the sense of a particular word depends on its neighboring words. However, performance of this method has been lower than that of other methods [35].

Unsupervised approach divides the occurrence of specific word into number of classes in order to decide whether the occurrence of word have same sense or not. The important task of this approach is to identify sense clusters. Various methods used in unsupervised approach are: context clustering, word clustering and co-occurrence graph [21,36]. The limitations of such approach are: not suitable for large scale situation, the instances in training data may not assign the correct sense, formation of heterogeneous clusters and number of clusters may differ from the number of senses of the target word.

3.3.1. Context Clustering

In context clustering method, every occurrence of target word is represented as context vector in the corpus. These vectors are grouped into clusters for the identification of sense of the target word. Advantage of this approach is that large amount of manually annotated training data is not required, while the drawback of this approach is that the training data is required for each word that needs to be disambiguated.

3.3.2 Word Clustering

In this method, words that are semantically similar are clustered to form a specific meaning. The relationship between identification word and target word depends on the information content for single features, given by syntactic dependencies in the corpus (e.g., subject-verb, verb-object, adjective-noun, etc.).

3.3.3 Co-occurrence Graphs

Here a graph is created on the basis of grammatical relationship between words. In the graph, every word in the text is called a vertex and syntactic relationship is called an edge. Weights are assigned to the edge on the basis of relationships. An iterative algorithm is applied on the graph to find the word that have the highest degree node and at last minimum spanning tree is used to disambiguate the target word.

3.4 Semi-Supervised Methods

Semi-supervised methods have become an active research area in the field of Word Sense Disambiguation. It requires function estimation on untagged data together with few tagged data. This approach is inspired by the fact that tagged data is often expensive to generate, whereas untagged data is generally not. However, the big challenge is how to use mixed data (tagged/untagged) in this approach. This method also requires less human effort as untagged data is available in abundance to disambiguate polysemous word. A popular approach is boot strapping. In this we start with a small amount of seed data for each word which could be manually tagged training data or a small number of decision rules like ‘eat’ in the context of ‘health’ almost always indicates ‘food’ [37].

4. Applications of WSD [38,39]

4.1 Information Retrieval (IR)

It is a process to find unstructured data, usually text, that satisfies an information needed from large database, which may be unstructured. In other words it is used to perform more complex and accurate searches for various types of data. WSD is required to improve the accuracy of IR indexing and to get better result than original query.

4.2 Information Extraction (IE)

Unlike IR, IE is the process of creating structured information out of returned search data. It is a more recent application area. IE focuses on the recognition, tagging and extraction representation of certain key elements of information e.g. persons, companies, locations from large collection of text. WSD is required to fetch correct information using IE.

4.3 Automated Answering Machine

Sometimes we need an automated online assistant providing customer service on a web page. A user can ask a question in everyday language and receive an answer quickly with sufficient context to validate the answer. WSD is required to use to fetch accurate information if the words in the questions are ambigious.

4.4 Speech Recognition

Speech recognition technologies allow computers equipped with sources of sound input, such as a microphone, to interpret human speech. It is the process of converting an acoustic signal by microphone or telephone, to set of words. WSD is used in speech reorganization to achieve accurate results

4.5 Machine Translation

In automatic machine translation from one language to another (Hindi to Punjabi), words having more than one senses in one or the other (or both) languages cause inaccuracies in translation. The accuracy of translation can be improved with WSD by using the correct sense in either (or both) languages.

5. Resources for WSD

The biggest resource is the Machine Readable Dictionaries (MRDs), these are extensively used in Knowledge Based Methods. LDOCE (Longman Dictionary of Contemporary English) has been the most widely used MRD in context of WSD. Thesauri are also available in machine readable format. These provide relationship information of words like synonyms, antonyms. A WordNet compiled by Princeton University is most popular and useful for WSD. Hindi WordNet has been developed at IIT, Bombay. Wikipedia is a net based encyclopedia which gives a lot of information about the sense of words.

Individual researchers also developed/trained corpus as required by the particular domain. Untrained data of this corpus are collected from Magazine, Newspapers, Books and other sources of text from diverse in text area. Parallel corpus between two languages (one the source language and other the target language) is also available in limited quantity and can be most useful.

6. Conclusions and Scope for Future Work

Artificial intelligence is a sun rise and vital area of research in computer science. Its usefulness and applicability finds not only in computer science/ applications but in all areas of research from engineering to basic sciences to social sciences and even to literature. It promises to revolutionize the research process. WSD is a key area of AI. All human languages have words having multiple meanings/ senses and choosing the correct/intended meaning in a given sentence is important for proper understanding of the text. This is performed using WSD techniques. Currently, WSD finds applications in Information Retrieval, Information Extraction, Automated Answering Machine, Speech Reorganization, Machine Translation among many more. The broad methods employed in WSD are (a) Dictionary and Knowledge Based, (b) Supervised Methods (c) Unsupervised Methods (d) Semi-Supervised Methods.

Dictionary and Knowledge Based methods use machine readable dictionaries (MRDs) and Thesaurus. Lesk algorithm is used to determine the most likely sense of the word. Supervised methods employ a structured training data made from a large corpus. It is based on the assumption that the context can provide sufficient evidence on its own to disambiguate word. The construction of the training data is time consuming and expensive. However, it produces superior results. Unsupervised methods use unlabeled corpus for training which is easier to make and hence are less expensive. In this method nearby words are used to make sense clusters/ word clusters. Semi-supervised methods use a combination of the above methods. A limited training data is annotated while the remaining bulk of the data is un-annotated making the task much simpler. Future scope of WSD is limitless. Imagine a Japanese talking in his native language but we hear him in Hindi/Punjabi etc. Automatic machine translation of one language to another still requires lot of research, particularly among Indian languages like Hindi, Punjabi, Marathi, Tamil etc.

7. References

[1] Manning, C.D., Schiitze, H. (1999). Foundations of Statistical Natural Language Processing. *MIT Press, Cambridge, Massachusetts*.

[2] Stevenson, M., Wilks, Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

[3] Weaver, W. Translation. Mimeographed. Reprinted in Locke, W.N., Booth, A.D. (1955). Machine Translation of Languages. *John Wiley & Sons, New York, 15-23.*

[4] Agirre, E., Edmonds, P. Word Sense Disambiguation; Algorithms and Applications, Edited by, Springer, VOLUME 33.

[5] Kaplan, A. (1950). An experimental study of ambiguity and context", in Mimeographed, in November, pp18. Reprinted in Mechanical Translation,1955, vol: 2(2), pp: 39-46 "An experimental study of ambiguity and context".

[6] Bar-Hillel, Y. (1964). On Syntactic categories, Journal of Symbolic Logic 15.1-16 [Repr.Bar-Hillel 1964a, 19-37].

[7] Madhu, S., Lytle, D. W. (1965). A Figure of Merit Technique for the Resolution of Non grammatical Amiguity. *Mechanical Translation, 8(2):9–13*.
[8] Wilks, Yorick. (1975). A Preferential Pattern-Seeking Semantics for Natural Language Inference. *Artificial Intelligence, 6:53-74*.
[9] Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", *Proceedings of SIGDOC.*
[10] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J. (1990). WordNet An On-Line Lexical Database. *International Journal of Lexicography, 3(4): 235-244.*
[11] P.F. Brown, J.C. Lai, and R.L. Mercer. (1991). Aligning Sentences In Parallel Corpora. *In Proceedings of 29th ACL*, pages 169--176, Berkeley, California.
[12] Brown, P.F., Pietra, S.A.D. (1992). An Estimation of Upper Bound for the Entropy of English. *Association for Computational Linguistics, 18(1): 31-40.*
[13] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter estimation. *Computational linguistics 19(2), 263–311.*
[14] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.3476&rep=rep1&type=pdf
[15] SchUtze, H. (1998). Automatic Word Sense Discrimination. *Association for Computational Linguistics.*
[16] Diab, M., Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. *In Proc. of ACL02, Philadelphia*.
[17] Michael, S. (2003). Combining Deep and Shallow Approaches in Parsing German. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 112-119.*
[18] Montoyo, A., Suarez, A., Rigau, G., Palomar, M. (2005). Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods. *Journal of Artificial Intelligence Research, 299-330.*
[19] Mihalcea, R. (2005). Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP),* pp 411–418, Vancouver..
[20] Vickrey, D., Biewald, L., Teyssier, M., Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, 771–778.*
[21] Sinha , R., Mihalcea, R. (2007). Unsupervised Graph-Based Word Sense Disambiguation Using Measures of Word Semantic Similarity. *ICSC '07 Proceedings of the International Conference on Semantic Computing, IEEE Computer Society Washington, DC, USA, pp 363-369.*
[22] Josan, G. S., and Lehal, G. S. (2008). Size of N for Word Sense Disambiguation using N Gram Model for Punjabi Language. *International Journal of Translation, 20(1): 47-56*.
[23] Mishra, N., Yadav, S., Siddiqui, T.J. (2009). An Unsupervised Approach to Hindi Word Sense Disambiguation. *Proceedings of the First International Conference on Intelligent Human Computer Interaction pp 327-335.*
[24] Goyal, D., Goyal, D. Singh, S. (2010). A Hybrid Approach to Word Sense Disambiguation. *International Journal of Computer Science and Technology IJCST Vol. 1, Issue* .
[25] Broda, B., Mazur, W. (2010). Evaluation of Clustering Algorithms for Polish Word Sense Disambiguation. *Proceedings of the International Multiconference on Computer Science and Information Technology.25–32.*
[26] Navigli, Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *Roberto IEEE Transactions On Pattern Analysis And Machine Intelligence, VOL. 32.*
[27] Nameh, M., Fakhrahmad, S.M., Jahromi, M.Z. (2011). A New Approach to Word Sense Disambiguation Based on Context Similarity. *Proceedings of the World Congress on Engineering , Vol 1 WCE, London, U.K.*
[28] Michael, L. (1987). Automatic Sense Disambiguation: How to Tell A Pine Cone From an Ice Cream Cone. *SIGDOC '86 Proceedings of the 5th annual international conference on Systems documentation Conference, pp 24−26, Association for Computing Machinery, New York.*
[29] Vanderwende, L. (1990). Using an On-line Dictionary to Disambiguate Verbal Phrase Attachment. *Proceedings of the 2nd IBM conference on NLP*, pp 347-359.
[30] Lee, Y.K., Ng, H.T., Chia, T.K. (2004). Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources. *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, Association for Computational Linguistics.*
[31] Boser, H., Guyon, I. M., Vapnik, V. N. (1992). A Training Algorithm For Optimal Margin Classifiers. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory 144-152. Pittsburgh, PA: ACM Press.*

[32] McCulloch, W., Pitts, W.H. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophys. 5,115-133.*
[33] Gallant, S. (1991.) A Practical Approach for Representing Context and for Performing Word Sense Disambiguation Using Neural Networks. *Neural Computation, 3/3, pp. 293-309.*
[34] http://shodhganga.inflibnet.ac.in:8080/jspui/bitstream/10603/34324/12/12_chapter%203.pdf
[35] Chen, P., Bowes, C., Ding, W., Brown, D. (2009). A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge. *Human Language Technologies: Annual Conference of the North American Chapter of the ACL, pp.28–36, Boulder, Colorado.*
[36] Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics. pp.189-196.*
[37] Zhu, X., Goldberg, A.B. (2009). Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 3(1), pp. 1-130.
[38] Manning, C. D., Raghavan, P., Schütze H. (2008). *Introduction to Information Retrieval*, Cambridge University Press.
[39] Grishman, R. (2012). Information Extraction: Capabilities and Challenges. *Notes prepared for International Winter School in Language and Speech Technologies, Rovira i Virgili University Tarragona, Spain Ralph*