ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

A Hybrid Approach for Word Sense Disambiguation Using Data Mining Techniques

Samit Kumar[1], Dr. Mahesh Kumar[2]

[1]M.Tech Student, Computer Science &Engg., MRK Institute of Engineering and Technology, Rewari, Haryana
[2]Associate Professor, Department of Computer Science & Engineering, MRK Institute of Engineering and Technology, Rewari, Haryana

ABSTRACT:Natural Language Processing is a core component of artificial intelligence. In NLP, solving the ambiguities is an important topic. The reason for this when a sentence translated automatically understand the meaning of the sentence or correct senses of the words is required. The goal has been to discover interesting rules between context and senses of ambiguous word. WordNet was discovered as a potentially useful source for knowledge about terms appearing in textual documents. We use a hybrid approach to fulfil this aim. The work in paper has tried to improve the results by further enriching the semantic value of the terms extracted from the different texts.

KEYWORDS: Apriori, Data mining, Word Net

I. INTRODUCTION

One of the first problems that is encountered by any natural language processing system is that of lexical ambiguity, be it syntactic or semantic. The problem is that words often have more than one meaning, sometimes fairly similar and sometimes completely different. The meaning of a word in a particular usage can only be determined by examining its context. This is, in general, a trivial task for the human language processing system, for example consider the following two sentences, each with a different sense of the word bank:

1. The boy leapt from the bank into the cold water.
2. The van pulled up outside the bank and three masked men got out.

We immediately recognize that in the first sentence bank refers to the edge of a river and in the second to a building. In modern WSD systems, the senses of a word are typically taken from some specified dictionary. These days Word Net is the usual dictionary in question. WSD has been investigated in computational linguistics as a specific task for well over 40 years, though the acronym is newer.

A. Word Sense Disambiguation

In the field of computational linguistics, the problem is generally called word sense disambiguation (WSD), and is defined as the process of identifying which sense of a meaning is used in any given sentence, when the word has a number of distinct senses [1]. WSD is essentially a task of classification: word senses are the classes, the context provides the evidence, and each occurrence of a word is assigned to one or more of its possible classes based on the evidence.

Ambiguity is intrinsic to human language and it constitutes an important challenge for most computational applications in the field of Natural Language Processing (NLP). Ambiguity [26] is explained as "the problem that an utterance in a human language can have more than one possible meaning. Ambiguity expresses itself at different levels. There are two main types of approach for WSD in natural language processing called as deep approaches and shallow approaches.

Deep approaches: These approaches involve the intention to understand and create meaning from what is being learned, Interact vigorously with the content, make use of evidence, inquiry and evaluation, Take a broad view and relate ideas to one another and Relate concepts to every time experience. These approaches are not very successful in practice, mainly because such a body of knowledge does not exist in a computer readable format, outside of very

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

limited domains. There is a long tradition in computational linguistics, of trying such approaches in terms of coded knowledge and in some cases; it is hard to say clearly whether the knowledge involved is linguistic or world knowledge.

Shallow approaches: These approaches are not concerned of learning the text instead they deal with the surrounding words of the ambiguous word and try to identify only parts of interest for a particular application. They just consider the surrounding words, using a training corpus of words tagged with their word senses the rules can be automatically derived by the computer[14]. This approach, while theoretically not as powerful as deep approaches, gives superior results in practice, due to the computer's limited word knowledge.

B. Word Net

WordNet [28] is a lexical set database of words having more than one meaning or we can call them synonymous words. It has a large vocabulary of nouns, verbs, adjectives and adverbs. In Word Net, a form is represented by a string of ASCII characters, and a sense is represented by the set of (one or more) synonyms that have that sense. Word Net contains more than 1.18 lakhs of different word forms and more than 0.9 lakh different word senses, or more than 1.66 lakhs (a, s) pairs. Approximately 17% of the words in WordNet are polysemous; approximately 40% have one or more synonyms. WordNet respects the syntactic categories noun, verb, adjective, and adverb the so called open-class words. It is assumed that the closed-class categories of English some 300 prepositions, pronouns, and determiners play an important role in any parsing system; they are given no semantic explication in WordNet

C. 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. 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 of an association rule A) B is defined as the percentage of transactions in which both A and B appear [9]. That is, the probability of the union of the item sets A and B, P(A U B).

We state the problem of mining association rules as follows: I ={i1, i2, ... ,im} is a set of items, T={t1, t2,...,tn} is a set of transactions, each of which contains items of the item set I . Thus, each transaction ti is a set of items such that ti $\subseteq$ I. An association rule is an implication of the form: X$\rightarrow$ Y , where X $\subset$I , Y $\subset$ I and X $\cap$Y =θ. X (or Y) is a set of items, called item set.

D. Data Mining

Data mining, and in particular text mining, has attracted much attention in recent years due to the vast amounts of data available, and the rate of growth. Data mining tools can be used to uncover patterns or hidden relations in the available data, and can potentially contribute greatly to business strategy decisions, knowledge bases, and scientific and medical research. The emergence of data mining tools has come as a result of the natural evolution in the field of information technology. Data mining is the process of discovering interesting knowledge from large amount of data stored in database, data warehouse or other information repositories.

II. RELATED WORK

The history of WSD research is as old as that of MT. WSD was first formulated as a distinct computational task during the early days of machine translation in the late 1940s, making it one of the oldest problems in computational linguistics. The 1950s then saw much work in estimating the degree of ambiguity in texts and bilingual dictionaries, and applying simple statistical models.

Yong-le SUN And Ke-liangJIA[11] 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.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

SasiKanthAla and Narayana Murthy Kavi[12] 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. Sense Disambiguation. 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[13] 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[14] 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 Word Net 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 Su´arez, German Rigau, Manuel Palomar[15] 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 ambiguation requires several knowledge sources in order to solve the semantic ambiguity of the words.

Dan Klein, Kristina Toutanova, H. Tolga Ilhan[16], 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.

DinakarJayarajan [17] 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 HweeTou Ng, David Chiang[18] 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.

III.EXPERIMENTAL SETUP

Problem is to gain an understanding of the senses of the word with its relative position in the entence.The mining model consists three major steps: 1. Pre-processing—splits the text into tokens(tokenization), POS tagging, chunking and parsing, 2.Creating transactional database from the pre-processed files and finally applying Apriori algorithm to get association rules on transactional database.Pre-processing is a very important step because numeric data and punctuation marks increase the number of 1-itemsets that results in more higher order invalid frequent item sets and

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

mining of these frequent item sets is a wastage of precious resources. Conversion of text files to transactional database is a fundamental requirement of Apriori algorithm.

The interface is created in Java language using Net Beans to enter the sentence with ambiguous word and displaying the result as best sense out of all senses retrieved from the Word Net database. MySql is used to store the results after getting senses from the Word Net. The Apriori Algorithm is used to improve the performance of word sense disambiguation.

The process of WSD usually consists of two steps, given below.

1. Find all possible senses for all the relevant words in a text.

2. Assign each word its correct sense.

The first step is straightforward. As, Word Net contains a number of synsets for each word, where each synset is a possible sense of the given word. Therefore, the first step can be accomplished by retrieving the possible senses from Word Net. Word Net is only one option for finding the possible senses, any available machine-readable dictionary, or knowledge source, may be used. This report, will however focus on Word Net, since Java interfaces to the Word Net dictionary are readily available.

The second step of WSD is accomplished by relying on two major information sources [32]. The first is the context of the word to be disambiguated, this includes both information within the ext, and extra-linguistic information about the text, for example the situation. The second source is an external knowledge source such as lexical or encyclopaedic resources, or hand-devised knowledge sources. The task of these sources is to provide data useful for associating a word with a sense.

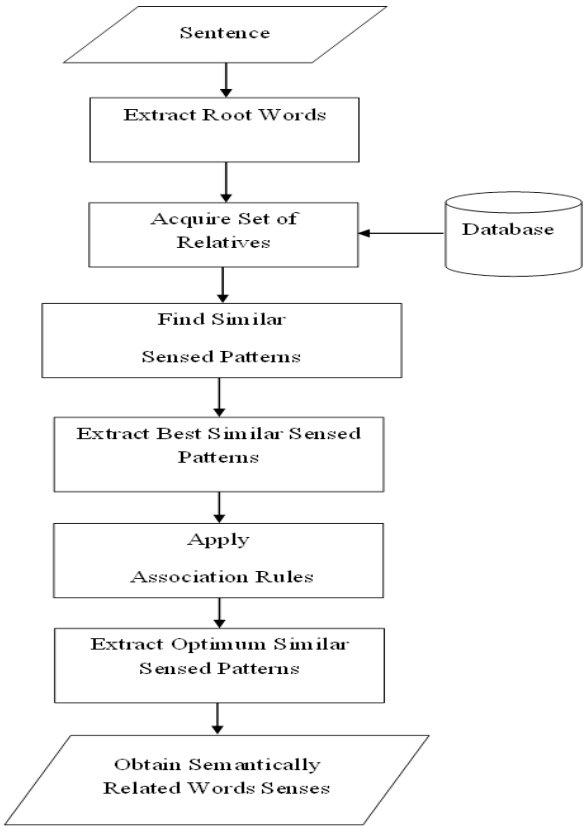

Fig. 1: WSD using Data Mining Techniques

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

When only ambiguous word is selected and search for the meaning in the Word Net database, then its returns all the possible senses of ambiguous word. When we select it context words then it returns the meaning of the word that is related to its context. In above example if search bank in Word Net database then records are given above.

IV. IMPLEMENTATION

The database used is a large lexical dataset called Word Net which has a large vocabulary of data set including nouns, verbs, adjectives and adverbs. The interface is created in Java language using NetBeans to enter the sentence with ambiguous word and displaying the result as best sense out of all senses retrieved from the Word Net database. MySql is used to store the results after getting senses from the Word Net. Since individual algorithms produce diverse results in terms of precision that complement each other well in terms of coverage. A hybrid approach outperforms score of best individual WSD approach. The data mining techniques are used to improve the performance of word sense disambiguation. At last performance measures are shown in tabular form to compare the results with previous techniques which are used to evaluate the performance. F-Score measure value increases as on applying optimized algorithm on input data.

We implement our work using followings steps:

1. We create a user interface for entering the sentence at run time in using Java NetBeans
2. We create a module for tokenization of the entered sentence for selection of ambiguous word and its context words.
3. We select an ambiguous word from the created tokens
4. We create a class to access the Word Net API database for displaying the all possible meaning of selected word.

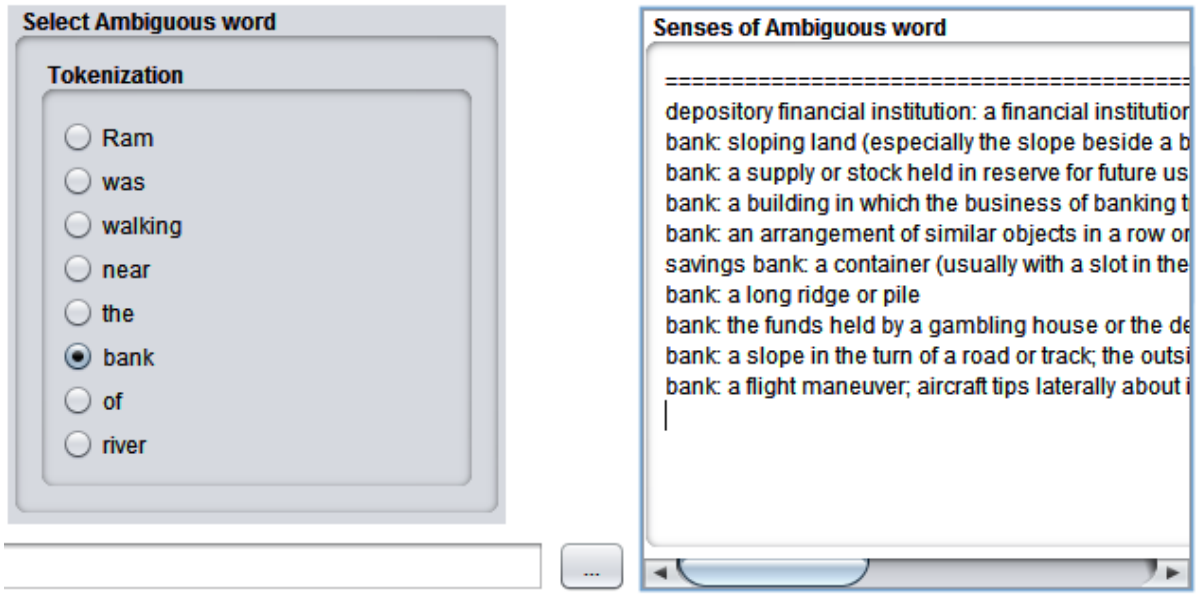

Fig. 2: Database of ambiguous word along with its senses and context word

5. We create Interface to select the context word
6. We create an API for adding words related to the context words from the corpus
7. We create a transactional database of the context word and senses of the ambiguous word using MySql.
8. Now we applied, the association rules using apriori algorithm and fuzzy association rules on the database created above. First, it generate the 1item frequent sets. Then using more itemset are generated using previous generated candidate sets. In the last, strong association rules are generated using the generated frequent sets. This rule will deduce the sense of ambiguous word.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

9. We find out the exact sense of the selected ambiguous word by mining the strong association rules.

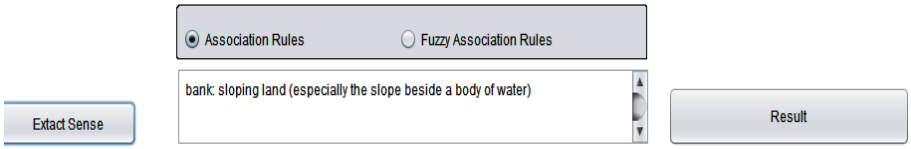

Fig. 3. Exact Sense of Ambiguous Word

10. We take different words to perform above tasks and noted the output.

V. RESULTS AND DISCUSSION

We Calculate the precision-recall, F-Score and Accuracy from above results. This above table shows the measurement of the system. In this table we have five cases these are given below:

- No of sentences
- Correctly identified
- Incorrectly identified
- Not identified

We have taken the followings performance measures to evaluate our work.

1. Precision : p is the number of correct results divided by the number of all returned results

$$\text{Precision} = \frac{t_p}{t_p + f_p}$$

wheret_p and f_p are the numbers of true positive and false positive predictions for the considered class.

2. Recall: Recall = Sensitivity

$$Recall = \frac{t_p}{t_p + f_n}$$

wheretp and fn are the numbers of true positive and false negative predictions for the considered class. tp + fn is the total number of test examples of the considered class.

3. F-Score=F-Measure

$$\text{F} - \text{Score/F} - \text{Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Accuracy:

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$$

After comparing the results it has been asserted that all metrics result in their improvement over earlier results.

We applied these two algorithms first on data without sampling. Then we apply the sampling on the data file and then association rules algorithms. Following Table shows the results of our experiments. By checking of all these cases for taking different examples like: bank, step, master, bass, etc. all these examples give the measurement of the system. By the help of above table we have also measure the performance of system.
For calculation of our system:

A = Correct identified sentence

B = incorrect identified sentence

C = Not identified sentence

Then:

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

$$\text{Recall} = \frac{A}{A+B} \times 100\ \% \text{Precision} = \frac{A}{A+C} \times 100\ \%$$

$$F - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Words	# Correctly Identified Sense	# Incorrectly Identified Sense	Not identified	Precision	Recall	F-Score
Bank	21	2	1	91	95	93
Bass	10	2	1	83	91	87
Coach	6	1	0	86	100	92
Put	13	2	0	87	100	93
Play	6	1	1	86	86	86
Deer	7	1	1	88	88	88
Free	8	1	1	89	89	89
Step	4	1	0	80	100	89
Save	7	1	0	88	100	93
Crack	13	2	2	87	87	87
Live	8	1	1	89	89	89

Table 1: Overall Results and Performance Measures

The resultant table below lists out the precision, recall and F-Score measure values for all ten ambiguous words taken for experimentation. From the table it is concluded that value of performance measure varies for different ambiguous words and it shows enhanced results while using optimized algorithm for disambiguation. Some words have shown 100% accuracy in finding the sense of ambiguous word.

Results are stored graphically as follows which shows that the value of recall varies from 80 to 100 and precision value varies from 85 to 100 when randomly taken the data of ten ambiguous words from English language.

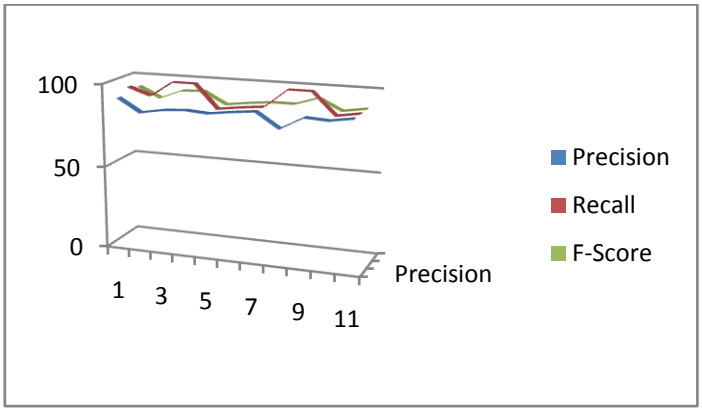

Fig 4. Graph between Recall, Precision and F-score

VI. CONCLUSION AND FUTURE SCOPE

In this paper, we use a hybrid approach to fulfil this aim. We used Word Net as knowledge based database and other corpora as dictionary approach to create the transactional database mining the association rules. In previous research only a single approach either dictionary based or knowledge based approach for mining the association rules was used to disambiguate the words. Ambiguous word, its context and its possible senses are stored in a database. This database

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

is presented to the algorithms for mining the decision rules. The sense which is inferred by most association rules will be the exact sense of that ambiguous word. Performance can be further improved by studying new algorithms for WSD, or to use another hierarchical knowledge structure where the senses are not so fine grained. The optimized approach can be used for some other applications of NLP like machine translation, information retrieval, parsing etc. Automatic creation of database is main challenge of this research. One can further work to create the transactional database automatically. More advance algorithm like genetic algorithm can be also applied to find out the sense.

REFERENCES

[1] Yarowsky, David. 2000. Word-sense disambiguation. Handbook of Natural Language Processing, ed. by Dale et al. 629–654. New York: Marcel Dekker.
[2] Zipf, George Kingsley. 1949. Human Behaviour and the Principle of Least Effort: An introduction to human ecology. Cambridge, MA: Addison-Wesley. Reprinted by New York: Hafner, 1972.
[3] Kaplan, Abraham. 1950. An experimental study of ambiguity and context. Mimeographed, 18pp, November 1950.
[4] Edmonds, Philip. 2005. Lexical disambiguation. The Elsevier Encyclopedia of Elsevier.Language and Linguistics, 2nd Ed., ed. by Keith Brown, 607–23. Oxford: Elsevier.
[5] Rieger, Chuck & Steven Small. 1979. Word expert parsing. Proceedings of the 6th International Joint Conference on Artificial Intelligence (IJCAI), 723–728.
[6] BAR-HILLEL, Y. 1960. The present status of automatic translation of languages.Advan.Comput. 1, 91–163.
[7] WILKS, Y. 1975. Preference semantics. In Formal Semantics of Natural Language, E. L. Keenan, Ed. Cambridge University Press, Cambridge, U.K., 329–348.
[8] WILKS, Y. A., FASS, D. C., GUO, C.-M., MCDONALD, J. E., PLATE, T., AND BRIAN, B. M. 1990. Providing machinetractable dictionary tools. Mach. Transl. 5, 99–154.
[9] IDE, N. AND V´ERONIS, J. 1998. Word sense disambiguation: The state of the art. computat. Ling. 24, 1, 1–40.
[10] Francis HEYLIGHEN, "Mining Associative Meanings from the Web: from word disambiguation to the global brain" Proceedings of Trends in Special Language and Language Technology, R. Temmerman (ed.),
[11] Yong-le SUN And Ke-liang JIA " Research of Word Sense Disambiguation Based on Mining Association Rules" Third International Symposium on Intelligent Information Technology Application Workshops, 2009
[12] SasiKanthAla and Narayana Murthy Kavi" Unrestricted Word Sense Disambiguation Using Association Rules" Tech. report LERC/UoH/2004/, University of Hyderabad Hyderabad, India, 2004
[13] 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
[14] Rion Snow Sushant Prakash, Daniel Jurafsky, Andrew Y. Ng ,"Learning to Merge Word Senses", Computer Science Department Stanford University
[15] Andres Montoyo, Armando Su´arez, German Rigau,"Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods", Journal of Artificial Intelligence Research ,2005
[16]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
[17]DinakarJayarajan, "Using Semantics in Document Representation: A Lexical Chain Approach" ,Department of Computer Science and Engineering Indian Institute of technology Madras ,June 2009.
[18]Yee Seng Chan and HweeTou Ng, David Chiang,"Word Sense Disambiguation Improves Statistical Machine Translation", Department of Computer Science National University of Singapore,
[19] Ganesh Ramakrishnan, B. Prithviraj, Pushpak Bhattacharyya. A Gloss Centered Algorithm for Word Sense Disambiguation. Proceedings of the ACL SENSEVAL 2004, Barcelona, Spain. P. 217-221.
[20] Zhang Yangsen. Study on the method of Chinese language word sense disambiguation and tagging for language resource constructing Post-doctoral research report of Peking University, 2006.12
[21] Lucia Specia, Maria das Gracas Volpe Nunes, Mark Stevenson "Mining Rules for Word Sense Disambiguation" Springer Berlin Heidelberg, 2012, volume 7053, PP 307-317
[22] Alberto J. Cañas, Alejandro Valerio, Juan Lalinde-Pulido, Marco Carvalho, Marco Arguedas, "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction" Springer , SPIRE 2003 – 10th International Symposium on String Processing and Information Retrieval, October 2003, Manaus, Brazil
[23] Ying Liu, Peter Scheuermann, Xingsen Li, and Xingquan Zhu, "Using WordNet to Disambiguate Word Senses for Text Classification" ICCS 2007, Part III, LNCS 4489, pp. 780–788, , Springer-Verlag Berlin Heidelberg 2007.
[24] Hirst, Graeme. 1987. Semantic Interpretation and the Resolution of Ambiguity. Cambridge, UK: Cambridge University Press.
[25] Coppin, B. (2004). Artificial Intelligence Illuminated., Sudbury, Massachusetts: Jones and Bartlett Publishers.
[26] Saad Ahmad, "Tutorial on Natural Language Processing", Springer 2007
[27] Navigli, Roberto, _Word Sense Disambiguation: a Survey_,ACM Computing Surveys, ACM Press, 2009.
[28] George A. Miller, Richard Beckwith, Christiane Fellbaum,Derek Gross, and Katherine Miller ,Introduction to WordNet: An On-line Lexical Database, August,1993
[29] M. H. Dunham. Data Mining: Introductory and Advanced Topics. Prentice Hall, 2003.
[30] H. Lu, L. Feng, and J. Han. Beyond Intratransaction Association Analysis: Mining Multidimensional Intertransaction Association Rules. ACM Trans. Inf. Syst., 18(4):423–454, 2000.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 5 , May 2015

[31] A. K. H. Tung, H. Lu, J. Han, and L. Feng. Efficient Mining of Intertransaction Association Rules. IEEE Transactions on Knowledge and Data Engineering, pages 43–56, 2003.
[32] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2006.