International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 4 Issue 4, April-2015, pp: (13-19), Impact Factor: 1.296, Available online at: www.erpublications.com

Natural Language Processing and Hindi Language

Sudesh Yadav
Department of Computer Science,
Govt. College, Ateli, Haryana, India

Abstract: Natural Language Processing (NLP) is the most important field of computer science which deals with human language and computers. In India, most of the people live in the rural areas, so they have the problem of understanding English language. To address this issue, we discussed here, about various factors which affect Hindi natural language processing. In this paper, a framework for information retrieval in Hindi is also proposed by authors. This framework is based on automatically inducing word sense using graph based method for all open class words present in the intended query. Till now researchers have worked for noun sense only, but here we expanded the framework for all open class words i.e. noun, verb, adverb and adjective.

Keywords: NLP (natural language processing), IR (information retrieval), Hindi WordNet, Search Engines.

Introduction

Natural language processing (NLP) [1] is the field of computer science and linguistics which concerned with the interaction between human language and computers. It is sometimes called as AI-complete problem, because Natural language recognition seems to require extensive knowledge about the outside world and ability to learn and manipulate it [2].

The main objective of this paper is to give a brief description of various factors which affect Hindi natural language processing and also to present importance of information retrieval process for Hindi.

Literature Review

The history of NLP(Natural language processing) started Over the 1970's and 1980's, much of the research in IR was focused on document retrieval, and the emphasis on this task in the Text Retrieval Conference (TREC) evaluations of the 1990's has further reinforced the view that IR is synonymous with document retrieval [12]. Web search engines are, of course, the most common example of this type of IR system. The enormous increase in the amount of online text available and the demand for access to different types of information have led to a renewed interest in a broad range of IR-related areas that go beyond simple document retrieval[12]. These areas include question answering systems, topic detection and tracking, summarization, multimedia retrieval (e.g., image, video and music), software engineering, chemical and biological informatics, text structuring, text mining, and genomics etc.

For information retrieval in English a lot of research has been done by researchers. (Krovetz and Croft 1992[15]) studied the sense matches between terms in query and the document collection. They concluded that the benefits of WSD in IR are not as expected because query words have skewed sense distribution and the collocation effect from other query terms already performs some disambiguation. Scholars (Liang-Yu Chen et.al 2007[16]) gives an idea for construction of automatic thesaurus for document retrieval. (Sanderson 1994[17]; 2000[18]) used pseudowords to introduce artificial word ambiguity in order to study the impact of sense ambiguity on IR. They concluded that the effectiveness of WSD can be negated by inaccurate WSD performance, high accuracy of WSD is an essential requirement to achieve improvement. Researchers (Bezdek et al. 1986[19]) extended the thesaurus to fuzzy thesaurus and defined fuzzy generalization and fuzzy synonymy relations for representing fuzzy thesaurus (Bezdek 1985[20]). They further used these two fuzzy relations in computing the value of six t-norms (Bandler & Kohout 1985[21]) for fuzzy information retrieval. Many studies investigated the effect of WordNet relations for construction of fuzzy thesaurus for web information retrieval. (Martine De Cock et al. 2005[22]) defined the composition of fuzzy document-term relations for WWW and further used it for building the fuzzy term-term relations for constructing fuzzy thesauri for and from WWW. Hang li et.al (May, 2006[23]), address the issue of search of definitions specifically for a given term. They used SVM as classification model and ranking SVM as the ordinal regression model. They categorized a definition into one of three levels: good definition, indifferent definition or bad definitions. Hang Li (2013[24]), proposed an approach to improve the state – of –the – art methods for relevance ranking in web search by query segmentation. They employed a (discriminative model) re-ranking approach in query segmentation to obtain segmentation results. Hang li et al.(2013[25]), introduced Regularized Latent Semantic Indexing (RLSI)---including a batch version and an online version, referred to as batch RLSI and online RLSI, respectively for topic modeling. In learning, batch RLSI processes all the documents in the collection as a whole, while online RLSI processes the documents in the collection one by one. Recently (Antonio Di Marco and Roberto Navigli 2011[26]) gave an approach to word sense induction for clustering of web search results using maximum spanning algorithm for acquiring meaning to the intended query. They showed that their method improves classical web search result in terms of both quality of clusters and degree of diversification. (David Hope and Bill Keller 2013[27]) introduce a linear time soft clustering algorithm for inducing word senses and clustering search results. They concluded that the method is comparable to existing state-of-art methods. In their paper they did not presented the methodology to assign the weights between edges of two nodes. Very recently (Antonio Di Marco and Roberto Navigli 2013[46]) presented a novel approach to Web search result clustering based on the automatic discovery of word senses from raw text, (also referred to as Word Sense Induction). Their approach is better than previous approaches because method proposed by them is parameter free and method improves web search results in terms of both clustering and degree of diversification. Moreover the method does not rely on the existing sense inventories.

In 2008, Prof. Battacharya dealt with the problem of ambiguity resolution in Hindi language for the first time. (Bhattacharyya et al. 2008[47]), proposed the Hindi word sense disambiguation(WSD) approach by comparing the linguistic context of the words in a sentence with the context constructed by Hindi WordNet using similarity based approach, which works for nouns only. (Tanveer J.Siddiqui 2009[28]), proposed an unsupervised approach to WSD by learning a decision list using untagged instances. In this work, firstly stemming has been applied and then stop words have been removed. After that the list is used for annotating an ambiguous word with its correct sense. This approach also works for nouns only. Rohan et.al. 2007[29], proposed an approach for resolving lexical ambiguity in Hindi language by making the comparisons between the different senses (nouns only) of the word in the sentence with the words present in synsets from the Hindi WordNet and the information related to these words in the form of parts-of speech. Avneet Kaur et.al. 2010[30], developed an approach for disambiguating ambiguous Hindi postposition by taking the problem with the case study of Hindi Punjabi machine translation. Rada Mihalea 2007[31], presents comparative evaluations on the basis of word semantic similarity based graph approach for unsupervised WSD. Siva Reddy 2009[32], proposed two unsupervised approaches namely Flat Semantic Category Labeler (FSCL) and Hierarchical Semantic Category Labeler (HSCR) for unsupervised semantic category labeling for Hindi WSD by using ontological categories defined in Hindi WordNet as sense inventory. The proposed approaches treat semantic categories as flat files and exploit the hierarchy among the semantic categories in a top down manner respectively. Researchers (S.K.Dwivedi 2011[7]), gave performance Comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language and the algorithm given by them based on Highest Sense Count and works well with Google. The objective of the paper was the comparative analysis of the WSD algorithm results on three Hindi language search engines- Google, Raftaar and Guruji and method was tested on a sample of 100 queries to check the performance of the WSD algorithm on various search engines. (Satyendr Singh and Tanveer J.Siddiqui 2012[33]), evaluates the effects of stemming, stop word removal and size of window on a manually created sense tagged corpus consisting of Hindi words (nouns). Romika yadav 2013[34], showed an improvement in the process of word sense disambiguation by using constraint solver i.e. Minion tool and results show that it correctly matches the context of word. Sabnam Kumari 2013[35], proposed a genetic algorithm based approach for Hindi WSD. It also works for nouns only. First time, the authors (Jain et al. 2013[13]) gave a method to disambiguate all open class words present in a sentence simultaneously for Hindi and method proposed by them is based on graph connectivity measures.

In Hindi web IR systems, only a very few research work is available. So there is an intense requirement to explore Hindi web IR and to build new approaches for improving the performance of Hindi IR. According to Sanderson (2008[38]) short queries are mostly benefitted from the ambiguity resolution. His study showed that disambiguation of queries improves performance of information retrieval. Lesk (1986[36]) proposed the algorithm for WSD, he also implemented his algorithm on the short text sample and found the good results. With the quite similar approach (Pushpak Bhattacharya 2008[37]) used his algorithm for the Hindi language WSD. His algorithm does not detect the ambiguity in the queries. First time, (S.K. Dwivedi et al. 2008[39]), proposed an entropy based approach for disambiguating queries on the basis of entropy threshold and concludes that detection of query disambiguation saves a lot of computational time in Hindi Language Information Retrieval. (Kumar Surabh et.al 2012[10]) showed the impact of English language on Hindi information retrieval and conclude that English language have left more impact on Hindi information retrieval. They also gave parameters for improving low recall problem in Hindi search processes. Kumar Surabh et.al., 2012[5], introduced a Hindi Query Optimization technique (design and development), which solved the problem of recall up to a great extent. Up to now the ambiguity in Hindi IR is resolved for noun phrases only using Hindi WordNet as resource.

Natural Language Processing for Hindi: Hindi is the mother tongue of India and is spoken by the major population of India mainly living in rural areas. About 5% of population living in these areas understands English as their second language. Hindi is spoken about 30% of the population[6]. Due to advancement in internet technologies a wide variety of Hindi Data and Literature is now available on web. The number of users who want the information in Hindi language is increasing [5]. Hindi is the language of dozens of major newspapers, magazines, radio and television stations and of other media. Major Hindi newspapers and TV channels have their websites in Hindi which are used by wide section of society [7]. A recent survey by a Delhi based research organization - Just Consult - says that 44 % of existing Internet users in India prefer Hindi over English, if made available. Similarly 25% existing Internet users prefer other regional languages. Many big companies like Google, Yahoo and Sify are also taking big steps in Hindi and other regional languages [8].

Hindi IR is still in a very nascent stage. As mentioned above problems like Phonetic nature of Hindi Language, morphology, word synonyms and ambiguous words affects the performance of the search engines in Hindi language information retrieval.

Factors which Affect Hindi Natural Language Processing

There are mainly four factors which affects Hindi IR:

1. **Morphological Factors:** Morphology is the branch of linguistics that studies patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. [9,10]
2. **Phonetic Nature of Hindi Language:** Various languages are spoken in India, each language being the mother tongue of tens of millions of people. While the languages and scripts are distinct from each other, the grammar and the alphabet are similar to a large extent. One common feature is that all the Indian languages are phonetic in nature [11,10]. For example; Following are the possible spelling variations for the Hindi word अ'ग्रेजी (angreji): (means English), अ'ग्रेजी, अंगरेजी, अन्गरेजी etc.
3. **Word Synonymy:**Because of different cultures, customs, religions India has rich diversity in languages, but the language structure and variation in dialects is making hindrances in the advantages of Information retrieval revolution in India[10]. For example: we know God is named as " भगवान" in Hindi but we can also call " भगवान" as "प्रभु" "इश्वर" or "देवता" and more. It is difficult to decide that which one is to choose?
4. **Ambiguous Words**: Many words are polysemous in nature[10]. Finding the correct sense of the words in a given context is an intricate task. One word has more than one meaning and meaning of word is depends on context of sentence. Example कर (Tax) having synonyms ब्याज , शुल्क, सूद , टैक्स in one context and in another context कर (Hand or arms) बॉह, हस्त, आच, शबर and कर करना (to do) in another context.

Now we discuss about various tasks of Hindi Natural Language Processing in detail:
WSD for Hindi: Word sense disambiguation (WSD)[1], the task of identifying the correct sense of words in given context is the growing research area in natural language processing. Sense disambiguation was considered as an essential fundamental task for many computational applications such as machine translation, information retrieval, question answering, text summarization, intelligent data retrieval and speech recognition [1].

Most of the people in India use Hindi as their primary language and the language spoken by these people is ambiguous i.e. many words can be interpreted in multiple ways depending upon the context. To disambiguate words, machines need to process unstructured textual information and transform them into data structures (tokenization) for analyzing intended meaning [2]. The automatic identification of meanings of words by machines is called word sense disambiguation.

The difficulty in analyzing the meanings in intended context by machines arise due to lack of intelligent resources, presence of different domain and different sense inventories Turing test [3].

Hindi Lexicon: Hindi WordNet[4] is the online lexicon reference system which is designed by centre for Indian language technology solutions, IIT Bombay, inspired by usability of Hindi language in all over the world. It contains noun, verb, adjectives and adverb sense of a word, represented by synonyms (synsets)

Here we discuss Hindi lexical database with an example- **फल** which have nine sense of noun, फल, तलवार, परिणाम etc.. These are represented as:-

a). 1. फल$_n^1$, प्रसून$_n^1$

2. फल$_n^2$, तीर$_n^2$

Each word in lexical database is associated with part of speech tagged, with a subscript: n stands for noun, v stands for verb, a for adjective and r for adverb and superscript denoting sense number (i.e. फल$_n^2$ has second sense of the verb फल$_n$). Sense numbering of a word is given by frequency of occurrence of that word in SemCor corpus. Each synset in WordNet is associated with a gloss: textual definition which explains its meaning.

Hindi WordNet lexical database have many relations: nominalization, hypernymy, pertainy, holonymy and hyponymy relations and so on.

Here in figure 1.1. , we show an excerpt of Hindi WordNet graph centered on the word फल$_n$ using Hindi WordNet as resource.

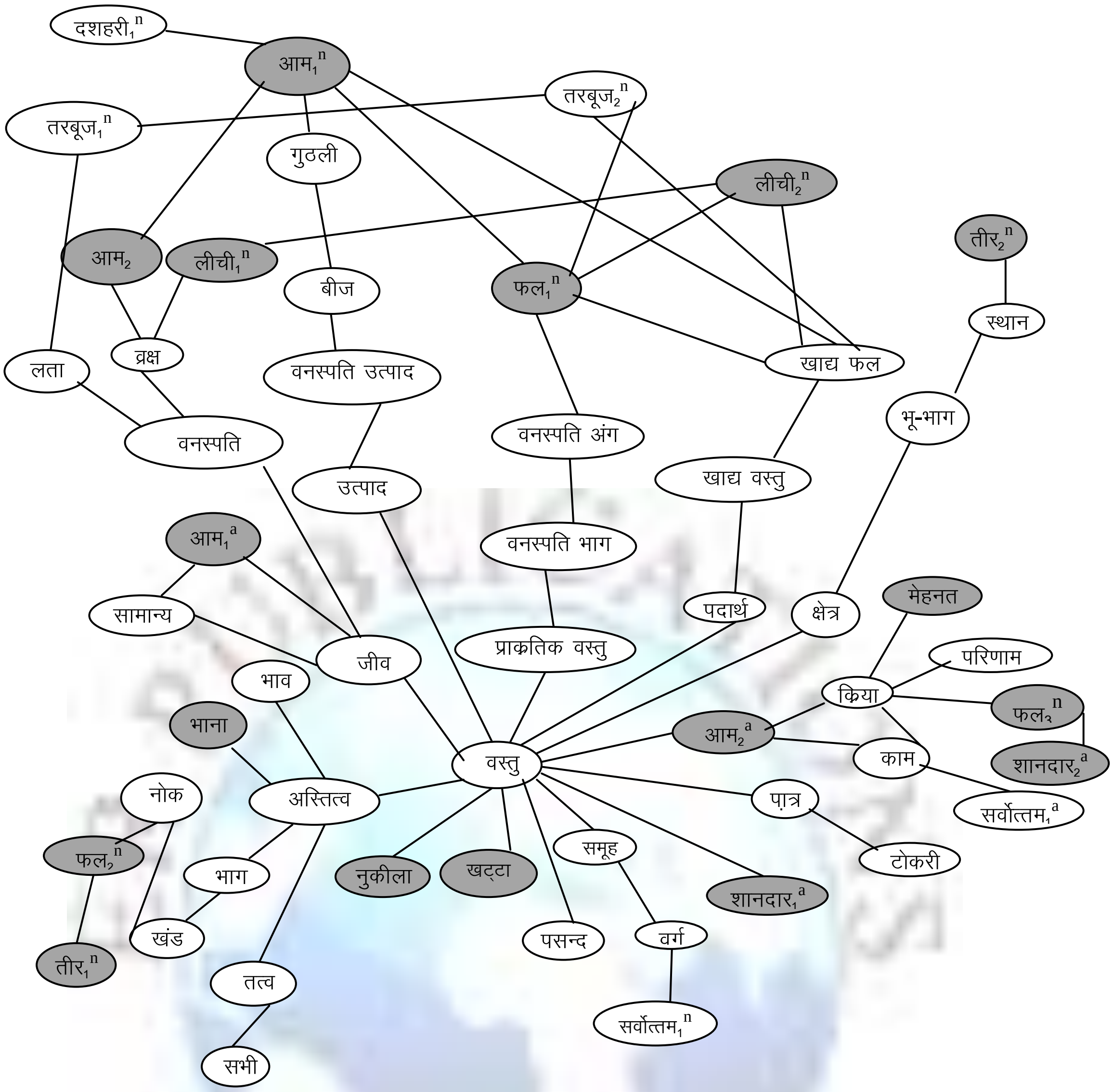


Fig. 1.1: An exert of Hindi WordNet Graph

INFORMATION RETRIEVAL IN HINDI

Information retrieval is the activity of obtaining information relevant to the exact information needed by a user from a collection of resources e.g. in web search a system has to search relevant data from billions of documents stored on millions of computers. In information retrieval word sense disambiguation plays an important role[1][5].

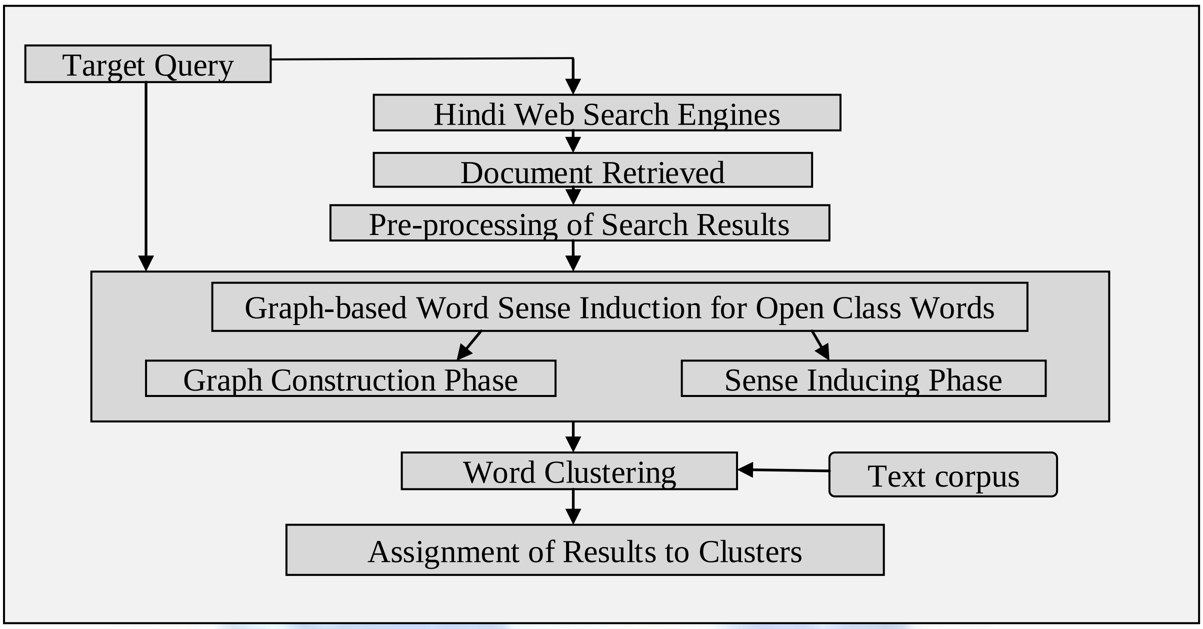
The ambiguity in natural language is considered as the major barrier in language processing applications, especially in information retrieval. Some query terms have a clear cut sense in their query. However some query terms hold ambiguity. The problem also persists with the Hindi language information retrieval as well. Hindi language information retrieval on the web is still in its nascent stage. It is the fact that to date Internet is vigorously used in India by the people who are comfortable in English language. The under development of web in Indian regional languages is one of the important reasons behind the limited growth of Internet in India. Indians use 22 official languages and 11 written script forms and among all the languages Hindi language is spoken by the major population of India. The number of users who want the information in Hindi language is increasing. This leads to the demand of the Hindi information retrieval on the web.

Various search engines are available on the internet as independent search engine sites in English. But very few like (Google, Raftaar and Webkhoj) Hindi language search engines are available. The search engines that support Hindi language search are not able to provide appropriate result for a user query. There are various problems that the search engines face with Hindi language information retrieval, sense ambiguity is one of the major problems in Information Retrieval on web in Hindi Language. Many words are polysemous in nature. Identifying the appropriate sense of the words in the given context is a difficult job for the search engines. Word sense disambiguation gives solution to the many natural language processing systems including information retrieval.

Up to now to disambiguates meanings for optimizing information retrieval researchers use Hindi WordNet as resource using different WSD approaches and then search engines for Hindi like Raftaar, Hinkhoj etc. are used optimize fuzzy information retrieval system for Hindi Language.

Proposed Framework for Hindi IR

In order to address ambiguity problem in Hindi Information Retrieval we proposed a graph based WSD framework for Hindi. In this framework, first of all we take target query. In our previous research, we focus on the noun words present in the target query, as noun contains the most sense of the query. Here we consider all the open class words present in the target query. Our proposed method also improves the performance of Hindi IR system by improving the process of search results in the form of clusters. The whole process of our proposed algorithm is given below:



1. Given a target query term, perform web search on Hindi search engines, which retrieve a list of documents i.e. ($T=T_1,T_2,T_{3,}\ldots\ldots T_n$).

2. Next, perform graph based Word Sense Induction for disambiguating ambiguous terms (all open class words i.e. noun, verb, adjective and adverb) and inducing semantics in the nodes of the graph.

3. Then we perform Max-Max soft clustering approach for assigning nodes present in the graph to more than one cluster on the basis of semantic similarity.

For accomplishing the above mentioned tasks, first of all we preprocess the web search results obtained by search engines. Then induce word semantics by means of Word Sense Induction algorithms [14]. Here sense induction is done for all open class words present in the target query. Here we use Max-Max soft clustering Word Sense induction algorithm for facilitation of information retrieval in the form of web clusters. After inducing word senses cluster sorting is done on the basis of various similarity notions.

Conclusion

In this paper, we discussed about natural language processing for Hindi language. The various factors which affect NLP for Hindi are also described here with a vast description of Hindi Lexicon database Hindi WordNet. To improve the performance of Hindi Search engine Information Retrieval, we proposed a framework for Hindi IR.

References

[1] Siddiqui Tanveer and Tiwari U.S.(2010) *Natural Language Processing and Information Retrivel*. Oxford.

[2] Robert Navigli (2009) *Word Sense Disambiguation: A Servey*. ACM Computing Surveys, Vol. 41, No. 2, Article 10.

[3] TURING, A. M. (1950) *Computing machinery and intelligence*. Mind 54, pp. 443-460.

[4] Hindi Wordnet from Center for Indian Language Technology Solutions, IIT Bombay India http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[5] Kumar Saurabh (October, 2012) *Query Optimization: A Solution for Low Recall Problem in Hindi Language Information Retrieval*. International Journal of Computer Applications (0975 – 8887) Volume 55– No.17.

[6] Burkhart, G.E., S.E. Goodman, A. Mehta and L. Press, (1998) *The internet in India: Better times ahead?*. Commun. ACM., 41: 21-26. http://portal.acm.org/citation.cfm?id=287835.

[7] Rastogi, P., & Dwivedi, S. K. (2011) *Performance comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language Supporting Search Engines*. International Journal of Computer Science Issues (IJCSI), 8(2).

[8] Srivastava Ranjan, Prabhat Khabar, (2006) *The Future Of Hindi On The Internet*. http://Www.Raftaar.In/Thehoot.Htm.

[9] Rathor Rajeev Master Of Engineering *Thesis Thapar University*, Patiala *Morphological POS Tagger For Hindi Language*.

[10] Kumar Saurabh (March,2012) *An Experimental Analysis on the Influence of English on Hindi Language Information Retrieval*. International Journal of Computer Applications (0975 – 8887) Volume 41– No.11.

[11] Ganpathiraju et al. (Sep., 2005) *Om: One Tool For Many (Indian) Languages*. Journal of Zhejiang University SCIENCE ISSN 1009-3095.

[12] "*Challenges in Information Retrieval and Language Modeling*", Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002.

[13] Jain A. , Yadav S. ,K Tayal D.(2013) *Measuring Context-Meaning for Open Class Words in Hindi Language*. In 6th International Conference on Contemporary Computing (IEEE), (IC3), pp. 118-123.

[14] Agrawal, Rakesh, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong (2009) *Diversifying search results*. In Proceedings of the 2nd International Conference on Web Search and Web Data Mining, pp. 5–14, Barcelona.

[15] Krovetz, R. and W.B. Croft,(1992)*Lexical Ambiguity and information retrieval. ACM* Trans. Inform. Syst., 10: pp.115-141.

[16] Liang-Yu Chen, Shyi-Ming,(2007) *New Approach for Automatic Thesaurus Construction and Query Expansion for Document Retrieval*. Information and Management Sciences, Volume 18, Number 4, pp. 299-315.

[17] Sanderson, M. (1994, August) *Word sense disambiguation and information retrieval*. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 142-151). Springer-Verlag New York, Inc. pp. 142-151 .

[18] Sanderson, M. (2000*) Retrieving with good sense*. Information retrieval, *2*(1), pp. 49-69.

[19] Bezdek(1986) *Transitive closures of fuzzy thesauri for information-retrieval systems*. InL J. Man-Machine Studies, pp. 343-356.

[20] Bezdek, James C., and Ian M. Anderson (1985) *An application of the c-varieties clustering algorithms to polygonal curve fitting*. Systems, Man and Cybernetics, IEEE Transactions on vol.5, pp. 637-641.

[21] Bandler, Wyllis, and Ladislav J. Kohout(1985) *Probabilistic versus fuzzy production rules in expert systems*. International journal of man-machine studies vol. 22, no. 3 pp. 347-353.

[22] De Cock Martine, Guadarrama Sergio, and Nikravesh Masoud(2005) *Fuzzy Thesauri for and from the WWW*.LNCL, Springer.

[23] Xu, J., Cao, Y. B., Li, H., Zhao, M., & Huang, Y. L. (2006) *A supervised learning approach to search of definitions*. Journal of Computer Science and Technology, 21(3), 439-449.

[24] Wu, H., Hu, Y., Li, H., & Chen, E. (2013) *Query Segmentation for Relevance Ranking in Web Search*. arXiv preprint arXiv:1312.0182.

[25] Wang, Q., Xu, J., Li, H., & Craswell, N. (2013) *Regularized latent semantic indexing: A new approach to large-scale topic modeling*. ACM Transactions on Information Systems (TOIS), 31(1), 5.

[26] Antonio Di Marco and Roberto Navigli (2011) *Clustering Web Search Results with Maximum Spanning Trees*. R. Pirrone and F. Sorbello (Eds.): AI*IA 2011,LNAI 6934, pp. 201–212,Springer-Verlag Berlin Heidelberg.

[27] Hope David and Keller Bill (2013) *Max-Max: A Graph-Based Soft Clustering Algorithm Applied to Word Sense Induction*. A. Gelbukh (Ed.): CICLing 2013, Part 1, LNCS 7816, pp. 368-381, Springer-Verlag Berlin Heidlberg.

[28] Mishra Neetu, Yadav Shashi and Siddiqui Tanveer J.(2009) *An Unsupervised Approach to Hindi Word Sense Disambiguation*. M.tech Thesis, Indian Institute of Information Technology,Allahabad UP,India.

[29] Rohan (2007) *Word Sense Disambiguation for Hindi Language*. M.Tech Thesis, Thapur University Patiyala, CSE Dept. India.

[30] Kaur Avneet (August 2010) *Development of an Approach for Disambiguating Ambiguous Hindi Postposition*. International Journal of Computer Applications(0975-8887), vol 5 , no.9.

[31] Sinha Ravi and Mihalcea Rada(September, 2007) *Unsupervised Graph-Based Word Sense Disambiguation Using Measures of Word Semantic Similarity*. IEEE International Conference on Semantic Computing, pp. 363-369.

[32] Reddy Siva, Abhilash, Sangal Rajeev, Paul Some (September, 2009) *All Words Unsupervised Semantic Category Labeling for Hindi*. Proceedings of International Conference RANLP, Borovets, Bulgaria, pp 365-369.

[33] Singh Satyender and Siddiqui Tanveer J. (2012) *Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation*. In International Conference on Information Retrieval & Knowledge Management,IEEE.

[34] Yadav Romika and Manker Rashmi(March, 2013) *Improvement of Word Sense Disambiguation Using MINION*. International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 3.

[35] Kumari Sabnam, Singh Paramjit (May, 2013) *Genetic Algorithm Based Hindi Word Sense Disambiguation*. International Journal of Computer Science and Mobile Computing, vol.2.Issue 5, May-2013, pp. 139-144.

[36] Lesk, M. (June, 1986) *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*. In Proceedings of the 5th annual international conference on Systems documentation. ACM, pp. 24-26.

[37] Sinha Manish , Bhattacharya R Pushpak (2008) *Hindi Word Sense Disambiguation*. Indian Institute of Information Technology, Department of Computer Science & Engineering Mumbai.

[38] Sanderson, Mark. (2008) *Ambiguous queries: Test collections need more sense*. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 499–506, Singapore.

[39] Dwivedi K. S.(2008) *An Entropy Based Method for Removing Web Query Ambiguity in Hindi Language*. Journal of Computer Science Vol. 4 No.9, pp.762-767,Science Publications.