A Graph Based Multilingual Word Sense Disambiguation

J. H. Patil, S. N. Patankar

Abstract*— Nowadays, the need of advanced free text filtering is increasing. Therefore, when searching for specific keywords, it is desirable to eliminate occurrences where the word or words are used in an inappropriate sense. This task could be exploited in internet browsers, and resource discovery systems, relational databases containing free text fields, electronic document management systems, data warehouse and data mining systems, etc. In order to resolve this problem in this work, we present joint approach to Word Sense Disambiguation (WSD). Our method exploits IndoWordNet, is a linked lexical knowledge base of word nets of 18 scheduled languages of India, a very large knowledge base, to perform graph based WSD across different languages in India, and brings together empirical evidence from these languages using ensemble methods. Therefore the results show that, by complementing the wide-coverage lexical knowledge with robust graph-based algorithms and combination methods, we can achieve the state of the art in WSD settings.

However, it does not require any sort of training process, no hand-coding of lexical entries, nor the hand-tagging of texts.

Index Terms—**Word Sense Disambiguation, IndowordNet, Graph Based Approach, Multilingual Information**

I. INTRODUCTION

Natural language can be ambiguous, such that many common words may have the same writing but indicate different meanings and multiple interpretations depending on the context in which these words occur. The fast growth of information and technologies on the Internet has increased the amount of unstructured data where this data is expressed in natural language. In addition, many knowledge resources available online such as blogs, surveys, articles, web pages, documents and corpora (collection of documents) are expressed in free (unstructured) texts written in natural language. This has increased the demand for software that analyses text of all forms to solve the ambiguity problem.

In Natural Language Processing (NLP) one of the main problems is the ambiguity of words. This problem affects different tasks such as: Information Retrieval, Information Extraction, Question Answering, etc. In order to deal with this problem an intermediate task called Word Sense Disambiguation (WSD) has been developed.

Revised Version Manuscript Received on January 27, 2018.

Jyotsna Harshal Patil, Computer Engineering, Datta Meghe College of Engineering, Airoli (Thane), India, (e-mail: patiljyotsna916@gmail.com).

Prof. S. N. Patankar, Associate Professor, Computer Engineering, Datta Meghe College of Engineering, Airoli(Thane), India, (e-mail: snp.cm.dmce@gmail.com).

Word Sense Disambiguation is a technique to find the exact sense of an ambiguous word in a particular context. For example, an English word ‘bank’ may have different senses as “financial institution”, “river side”, “reservoir” etc. Such words with multiple senses are called ambiguous words and the process of finding the exact sense of an ambiguous word for a particular context is called Word Sense Disambiguation. A normal human being has an inborn capability to differentiate the multiple senses of an ambiguous word in a particular context, but the machines run only according to the instructions. So, different rules are fed to the system to execute a particular task.

Word sense disambiguation (WSD plays a critical role as a classification task for automated translation of text, since in the late 1940s; it was considered as a major task for machine translation. WSD has been addressed by many researchers who have used the state-of-the-art techniques to identify the meanings of words in text. It has also been applied in some potential real applications such as machine translation, IR and IE. The task of WSD involves examining contextual information to find the correct sense of a target word in a sentence by assigning the most related one to this word depending on the context in which it occurs. The context is determined by the other words in the neighborhood in the same sentence, so that every sense of the target word to be disambiguated is compared to the senses of the surrounding words.

Nowadays the textual information needed by a user is available in very wide range of languages when user accesses the websites for content such as news reports, commentaries and encyclopedic knowledge. We know that English is the majority language of the web but nowadays other languages are also shares the web contents .For example, Chinese and Spanish languages ,and more languages are about to join them in the near future. This rapid development in language forces researchers to focus on the challenging problem of being able to analyze and understand text written in any language.

Here we address both the objectives i.e., disambiguating in an arbitrary language and using lexical and semantic knowledge from many languages in a joint way to improve the WSD task and propose a graph-based approach to joint Word Sense Disambiguation.

WSD approaches are categorized mainly into three types, Knowledge-based, Supervised and Unsupervised methods, which is described in detail later.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

II. BRIEF HISTORY OF RESEARCH ON WORD SENSE DISAMBIGUATION

WSD is one of the most challenging jobs in the research field of Natural Language Processing. Research work in this domain was started during the late 1940s. In 1949, (Zipf, George) [1] proposed his "Law of Meaning" theory. This theory states that there exists a power-law relationship between the more frequent words and the less frequent words. The more frequent words have more senses than the less frequent words. The relationship has been confirmed later for the British National Corpus. In 1950, Kaplan [2] determined that in a particular context two words on either side of an ambiguous word are equivalent to the whole sentence of the context. In 1975 (Wilks, Yorick) [3] 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. In 1980s there was a remarkable development in the field of WSD research as Large-scale lexical resources and corpora became available during this time. As a result, researchers started using different automatic knowledge extraction procedures (Wilks.1990) parallel with the handcrafting methodologies. In 1986, Lesk[4] 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. In 1991, (Guthrie) [5]. used the subject codes to disambiguate the exact sense using the Longman Dictionary of Contemporary English (LDOCE). In 1990s, three major developments occurred in the research fields of NLP: online dictionary Word Net became available, the statistical methodologies were introduced in this domain, and Senseval began. The invention of Word Net (Miller 1990) [6] brought a revolution in this research field because it was both programmatically accessible and hierarchically organized into word senses called synsets. Today, Word Net is used as an important online sense inventory in WSD research. Statistical and machine learning methods are also successfully used in the sense classification problems. Today, methods that are trained on manually sense-tagged corpora (i.e., supervised learning methods) have become the mainstream approach to WSD. Corpus based Word Sense Disambiguation was first implemented by (Brown ,1991) [7]. As the data sets, corpuses, online Dictionaries vary language to language all over the world, there was not any bench mark of performance measurement in this domain in the early age. Senseval brought all kind of research works in this domain under a single umbrella. The first Senseval was proposed in 1997 by Resnik andYarowsky. Preeti Yadav [8] proposes that Hindi Word Sense Disambiguation" that was the first attempt for an Indian language at automatic WSD. The use of the Wordnet for Hindi developed at IIT Bombay, which is a highly important lexical knowledge base for Hindi. The main idea is to compare the context of the word in a sentence with the contexts constructed from the Wordnet and chooses the winner. The output of the system is a particular synset number designating the sense of the word.

Banerjee, Pedersen (2002) [9] It provides an adaptation of Lesk‟s dictionary based word sense disambiguation algorithm. Lexical database is employed rather than is made using a standard dictionary as the source of glosses. Instead of using a standard dictionary as the source of glosses, the lexical database word net is employed. The Lesk algorithm is the prototypical approach and is based on detecting shared vocabulary between the definitions of words.

Mishra, Yadav, Siddiqui (2009) [10] 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) [11] 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) [12] focused on evaluation of a selected clustering algorithms (K-Means,K-Medoids, hierarchical agglomerative clustering, hierarchical divisive clustering, Growing Hierarchical Self Organizing Maps, graph-partitioning based clustering) in task of Word Sense Disambiguation for Polish.

Navigli, Lapata (2010) [15] 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 a proper choice of nodes and edge drawing criteria and weighing, graphs can be extremely useful for revealing regularities and patterns in the data, allowing us to bypass the bottleneck of data annotations.

Graphs' appeal is also enhanced by the fact that using them as a representation method can reveal characteristics and be useful for human inspection, and thus provide insights and ideas for automatic methods.

Graphs and graph-based algorithms are particularly relevant for unsupervised approaches to language tasks. Choosing what the vertices represent, what their features are, and how edges between them should be drawn and weighted, leads to uncovering salient regularities and structure in the language or corpora data represented. Transforming a graph representation allows different characteristics of the data to come into focus.

III. APPLICATIONS OF WORD SENSE DISAMBIGUATION

The main field of application of WSD is Machine Translation, but it is used in near about all kinds of linguistic researches.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

3.1 Machine translation (MT):

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.

3.2 Information retrieval (IR):

Resolving ambiguity in a query is the most vital issue in IR system. As for example, a word "depression" in a query may carry different meanings as illness, weather systems, or economics. So, finding the exact sense of an ambiguous word in a particular question before finding its answer is the most vital issue in this regard.

3.3 Information extraction (IE) and text mining:

WSD plays an important role for information extraction in different research works as Bioinformatics research, Named Entity recognition system, co-reference resolution etc.

3.4 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 ambiguous

IV. WSD APPROACHES

4.1 Unsupervised WSD

In this approach, no supervision is provided. It is divided into two type, type-based and token-based approach. The type-based approach is ambiguates by clustering instances of a target word while token-based approach disambiguates by clustering context of an ambiguous word. Main disadvantage of this approach is that senses are not well defined. This technique uses un-annotated corpus. Performance of unsupervised WSD has been lower than other methods.

4.2 Supervised WSD:

In this approach, there are large number of algorithms for WSD and it uses machine learning techniques ford disambiguation. It makes use of sense-annotated corpus. Disadvantage of this approach is that it requires large sense-annotated data. It is not suitable for resource scarce language. It is better than unsupervised and knowledge based approach.

4.3 Knowledge based:

This approach is based on machine-readable dictionaries or sense inventories. It can also be used with corpus-based methods. Word-net is used for knowledge-based approach. It assumes that words used together in text are related to each other and that the relation among them can be observed in the definitions of that 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.

V. GRAPH BASED APPROACHES TO WSD

One direction that modern research in WSD has taken is graph-based approaches. Work has been done in this area. Comparative studies performed by Mihalcea [14] indicate that graph-based methods often outperform the similarity-based ones by a significant margin. Navigli and Lapata [13] state that most unsupervised algorithms can be seen as either similarity-based approaches or graph-based approaches. The authors further note that these graph-based algorithms often have two stages - during the first stage a graph is constructed based on all the possible sequences of senses for the words to be disambiguated. During the next stage the structure of the graph is exploited to determine the most important nodes in the graph, which eventually leads to disambiguation of the polysemous words in the context. As mentioned in the above referenced paper, graph-based approaches attempt to assign senses to words collectively in a global manner, by exploiting dependencies across senses, while in contrast similarity-based approaches disambiguate each word individually without looking at the senses that are assigned to the words immediately before and after. Also, as noted in Navigli and Lapata [13], graph connectivity measures could be local or global. If the graph connectivity measure is local, then for each word to be disambiguated, the measure is able to select one sense which is most suitable in the given context. If the connectivity measure happens to be global, then the disambiguation methodology changes a little bit, in that instead of providing the disambiguated senses for each polysemous word, the algorithms score the overall interpretation of the sentence. Thus, if a sentence could have twenty possible interpretations, then a local graph connectivity measure provides us with the best scoring sense for each polysemous word in one iteration, whereas a global graph connectivity measure provides us with twenty overall possible assignments of senses, each assignment with a score. For their purposes, they view the WordNet graph as a set of nodes (comprising synsets) and edges (comprising the relationship between the synsets, e.g. synonyms, meronyms etc.)

For a sentence to be disambiguated, they start with a graph G = (V, E) where V consists of each synset in WordNet corresponding to all the words in the sentence. Next, they perform a depth-first search (DFS) of the WordNet graph. In order to do so, they start with any vertex u in V and keep searching the WordNet graph until they find another vertex v in it which was already present in V . Every time this happens, they add all the intermediate nodes on this path to G as undirected edges.

VI. METHODOLOGY

We present our methodology for WSD: we first introduce IndoWordNet, the resource used in our work (Section 6.1) and then present our algorithm for WSD, including its main components, namely graph-based WSD.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

6.1 IndoWordNet

IndoWordNet is a linked structure of WordNets of 19 different Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families (Bhattacharyya, 2010). [15]

IndoWordNet Dictionary or IWN Dictionary is an online interface to render multilingual IndoWord-Net information in the dictionary format. It allows user to view the results in multiple formats as per the need. Also, user can view the result in multiple languages simultaneously. The look and feel of the IWN Dictionary is kept same as a traditional dictionary keeping in mind the user adaptability. So far, it renders WordNet information of 19 Indian languages. These languages are: Assamese, Bodo, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Tamil, Telugu and Urdu. The WordNet information is also rendered in English.

6.2 A Graph Based Algorithm for Word Sense Disambiguation

Graphs are used to represent word senses(vertices) and their lexical and semantic connections (edges), as encoded by the reference knowledge resource. Next, graph-based algorithms are applied in order to perform WSD. These approaches have been shown to attain performance that is almost as good as supervised systems in domain-independent settings , and even to surpass them on specific domains.

6.2.1. Graph Representation

Given a sequence of words W = {w1, w2, ..., wn}, each word wi with corresponding possible labels Lwi = {l 1 wi , l2 wi , ..., lNwi wi }, we define a label graph G = (V,E) such that there is a vertex v ∈ V for every possible label l j wi , i = 1..n, j = 1..Nwi . Dependencies between pairs of labels are represented as undirected edges e ∈ E, defined over the set of vertex pairs V × V . Such label dependencies can be learned from annotated data, or derived by other means. In our case, these dependencies are learned from WordNet and measures of semantic similarity.

A property that makes these graph-based algorithms interesting is the fact that they take into account information drawn from the entire graph, capturing relationships among all the words in a sequence, and following this global technique they fall back to a local measure of graph centrality to assign labels. As will be seen ahead, this combination works very well.

6.2.2. Exploiting multilingual information in a knowledge-based WSD framework

We present a multilingual approach to WSD which exploits three main factors:

i) the fact that translations of a target word pro-vide complementary information on the range of its candidate senses in context;
ii) the wide-coverage, multilingual lexical knowledge stored in IndoWordNet;
iii) the support for disambiguation from different languages in a synergistic, unified way.

We call this approach multilingual joint WSD, since disambiguation is performed by exploiting different languages together at the same time. To this end, we first perform graph-based WSD using the target word in context as input, and then combine sense evidence from its translations using an ensemble method. The key idea of our joint approach is that sense evidence from different translations provides complementary views for the senses of a target word in context. Therefore, combining such evidence should produce more accurate sense predictions. We view WSD as a sense ranking problem.

Given a word sequence σ = (w1, . . . , wn), and given a target word w ∈ σ, we disambiguate w as follows.

We start by collecting the knowledge needed for disambiguation.

First, we collect the set S of Babel synsets corresponding to the different senses of the target word w. Next, we create the set T of multilingual lexicalizations of the target word w: to this end, we first include in T the word w itself, and then iterate through each synset s ∈ S to collect the translations of each of its senses into the languages of focus. Finally, we create a disambiguation context ctx by taking the word sequence σ and removing w from it.

Next, we calculate a probability distribution over the different synsets S of w for each term ti ∈ T . Each probability distribution quantifies the support for the different senses of the target word, determined using ti and the context ctx: we save this information in a |T| × |S| matrix LScore, where each cell lScorei,j quantifies the support for synset sj ∈ S, calculated using the term in ti ∈ T . We determine the scores as follows:

– We select an element ti from T at each step.
– Next, we create a multilingual context σ by combining ti with the words in ctx.
– We use σ' to build a graph Gi = (Vi,Ei) by computing the paths in IndoWordNet which connect the synsets of ti with those of the other words in σ . Note that by selecting a different element from T at each step we create a new graph where different sets of Indo synsets get activated by the context words in ctx.
– Finally, we compute the support from term ti for each synset sj ∈ S of the target word by applying a graph connectivity measure to Gi and store the result in lScorei,j .

By repeating the process for each term in T we compute all values in the matrix LScore. In the final phase we aggregate the scores associated with each term of T using an ensemble method M. For instance,M could simply consist of summing the scores associated with each sense over all distributions. As a result, the combined scoring distribution is returned.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

This sense distribution in turn can be used to select the best sense for the target word w ∈ σ.

6.2.3. An Example

Consider the task of assigning senses to the words in the text The church bells no longer rung on Sundays. For the purpose of illustration, we assume at most three senses for each word, which are shown in Figure 3.3. Word senses and definitions are obtained from the WordNet sense inventory. All word senses are added as vertices in the label graph, and weighted edges are drawn as dependencies among word senses, derived using the Lesk similarity measure (no edges are drawn between word senses with a similarity of zero). The resulting label graph is an undirected weighted graph, as shown in Figure 3.3. After running the PageRank graph centrality algorithm, scores are identified for each word-sense in the graph, indicated in Figure 3.4. Selecting for each word the sense with the largest score results in the following sense assignment: church#2, bell#1, ring#3, Sunday#1, which is correct according to annotations performed by professional lexicographers.

The church bells no longer rung on Sundays.

church

1: one of the groups of Christians who have their own beliefs and forms of worship

2: a place for public (especially Christian) worship

3: a service conducted in a church

bell

1: a hollow device made of metal that makes a ringing sound when struck

2: a push button at an outer door that gives a ringing or buzzing signal when pushed

3: the sound of a bell

ring

1: make a ringing sound

2: ring or echo with sound

3: make (bells) ring, often for the purposes of musical education

Sunday

1: first day of the week; observed as a day of rest and worship by most Christians

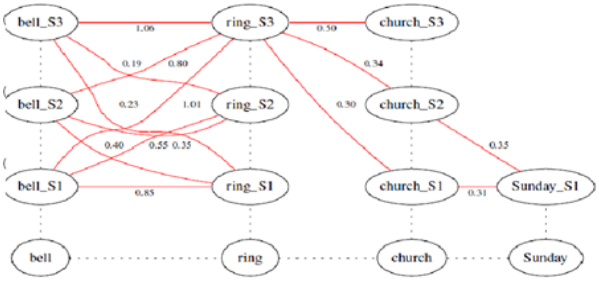

Figure 3.3. The graph for assigning senses to the words in sentence before PageRank is run.

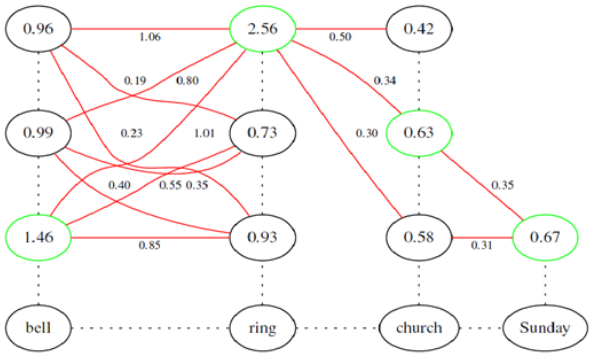

Figure 3.4. After PageRank has finished running, the nodes with lighter-colored borders are the senses assigned to the respective words.

VII. SUMMARY AND CONCLUSIONS

In this paper we presented a multilingual joint approach to WSD. The main key to our methodology is the effective use of a large-coverage multilingual knowledge base, i.e IndoWordNet. Here, we first perform graph-based WSD using the target word in context as input, and then combine sense evidence from its translations using an ensemble method. This is the first proposal to exploit structured multilingual information within a joint, knowledge-rich framework for WSD. The APIs to perform multilingual WSD using IndoWordNet are freely available for research purposes.

In the proposed approach, we achieve state-of-the-art performance on three different standards. In this approach we can not only achieves further advances by using multilingual lexical knowledge, but, more importantly, we combining multilingual sense evidence from different languages at the same time results in consistent improvements over a monolingual approach in monolingual lexical disambiguation . By using this multilingual knowledge base approach for WSD helps overcome the shortcomings of the underlying resource, thus results in performance boosts can come in the future.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

ACKNOWLEDGEMENT

I would like to express my gratitude to my project guide, Prof. Shreya Patankar, whose expertise and guidance added considerably to my graduate experience. I appreciate her Knowledge and her consistent assistance in completing this work, without whose motivation and encouragement, it would be difficult for me to move forward in my M.E Program. I would also like to thank the other sources of motivation.

REFERENCES

[1] Zipf, George Kingslay "Human Behaviour and the principal of least effort: An introduction to human ecology".Cambridge, MA: Addison-Wesley.Reprinted by New York: Hafner, 1972

[2] 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".

[3] Wilks, Yorick. (1975). A Preferential Pattern-Seeking Semantics for Natural Language Inference. Artificial Intelligence, 6:53-74.

[4] Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", Proceedings of SIGDOC.

[5] Guthrie, J., L. Guthrie, Y. Wilks, and H. Aidinejad. 1991. Subject-dependent co-occurrence and word sense disambiguation. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 146-152

[6] Miller, G. A., Ed. WordNet: An on-line lexical database. International Journal of Lexicography 3, 4 (Winter 1990), 235—312

[7] 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.

[8] International Journal For Research In Applied Science And Engineering Technology (IJRASET), Study of Hindi Word Sense Disambiguation Based on Hindi WorldNet, Preeti Yadav, Mohd. Shahid Husain ,Department of Computer Science, Lucknow, India

[9] Banerjee, S., Pedersen, T.,(2002) "An adapted Lesk algorithm for word sense disambiguation using WordNet", In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February.

[10] 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.

[11] 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 .

[12] 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.

[13] 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.

[14] Mihalcea et al., 2004] Mihalcea, R., Tarau, P., and Figa, E. (2004). Pagerank on semantic networks, with application to word sense disambiguation. In Proceedings of Coling 2004, pages 1126–1132, Geneva, Switzerland. COLING.

[15] Pushpak Bhattacharyya. 2010.IndoWordNet.In the Proceedings of Lexical Resources Engineering Conference (LREC), Malta.

[16] Roberto Navigli, "Word Sense Disambiguation: A survey," in ACM Comput. Surv. 41,2Article10,pages69,DOI=10.1145/1459352.1459355http://doi.acm.org/10.1145/1459532.145935 5, February 2009

[17] Nirali Patel1, Bhargesh Patel, Rajiv Parikh, Brijesh Bhatt, "A Survey: Word Sense Disambiguation", International Journal of Advance Foundation and Research in Computer (IJAFRC), January 2015.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.