ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 6.047

Volume 5, Issue 3, March 2017

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Implementation of Word Sense Disambiguation with scoring ability

Fatemeh Fathi Sedehi[1]
Department of Computer
University of Guilan
Rasht, Iran

Kobra Khoshraftar[2]
Department of computer
Islamic Azad University, Chalus Branch
Mazandaran, Iran

Ali Vahid Roudsari[3]
Department of computer
University of Guilan
Rasht, Iran

Abstract: ***Natural Language Processing is one of the important sub-branches in artificial intelligence. One of the first problems that we are facing in natural language processing system, the issue is the ambiguity of the meaning and structure of words. Word Sense Disambiguation note to the fact that word Whether used in a sentence and with what words are companion. Because a word gives a sense in one sentence, While the same Word in other sentence Takes another meaning. In this paper, we have reviewed four main methods of Word Sense Disambiguation, contains: Supervised Method, Unsupervised Method, Semi-Supervised Method and Knowledge base and Dictionary-based Method. According to a review of existing methods, Lesk methods were evaluated. This method is the simplest and most elementary Knowledge base and Dictionary-based Method. And by using the WordNet, we extract concepts of each of the words and Compare them with each other. And by scoring on each of the concepts of the ambiguous word, we chose the correct concept.***

Keywords: Word Sense Disambiguation, Supervised Method, Unsupervised Method, Semi-Supervised Method, Knowledge base and Dictionary-based Method, Lesk Algorithm.

I. INTRODUCTION

Natural language is used in everyday social interaction. There are different natural languages. Form of spoken and written language are different and independent from each other. Natural language is ambiguous. Find the correct sense of a word within a sentence, seemingly simple, but this act for a machine, is very difficult. Ambiguity may occur in different levels of natural language processing such as the lexical, syntactic and semantic [1]. Natural language processing or computational linguistics, Encompasses a wide range of computer-related research. Natural language processing, gives into the machine the ability to understand and read the language that person speaks. Natural language understanding is a problem of artificial intelligence. Because natural language recognition needs to understand the science such as computer science, statistics, science, language and so on. Today, Modern algorithms for natural language processing based on statistical machine learning [2].

In any language, there are a number of words have multiple meanings that changes according to the context in which they are used [3]. Word Sense Disambiguation, Plays an important role in the field of natural language processing. Always, ambiguous words influence in the development of information processing. Word Sense Disambiguation for determine the correct sense of an ambiguous word is used in a particular field. Word Sense Disambiguation mainly four types of unsupervised, semi-supervised, supervised and knowledge-based methods are divided. Supervised methods, Is effective for Word Sense Disambiguation. But this method requires a training data set. unsupervised Method, could be solve the problem of knowledge

acquisition. But, Result of disambiguation, is poor. In semi-supervised algorithm, a set of labeled and unlabeled used to teach a classification. Knowledge-based methods are based on lexical resources. These resources include the knowledge base, semantic networks, dictionaries and thesaurus [5] [4]. Several important applications of Word Sense Disambiguation including information retrieval, machine translation and information extraction [3].

The first step in word sense disambiguation, the phrase According to the syntax rules defined, transformed into a formal expression and then, based on a knowledge base Obtain the correct meaning of that word. In fact, the backbone of any disambiguation system has two parts: The first is a knowledge base and Second, an algorithm that can extract the proper sense of the knowledge base. Typically, word sense disambiguation as a matter of classification is shown. where in, a label to each word vague concept from a list of predefined concepts dedicated [7]. word sense disambiguation for vocabulary and word knowledge is important [8].

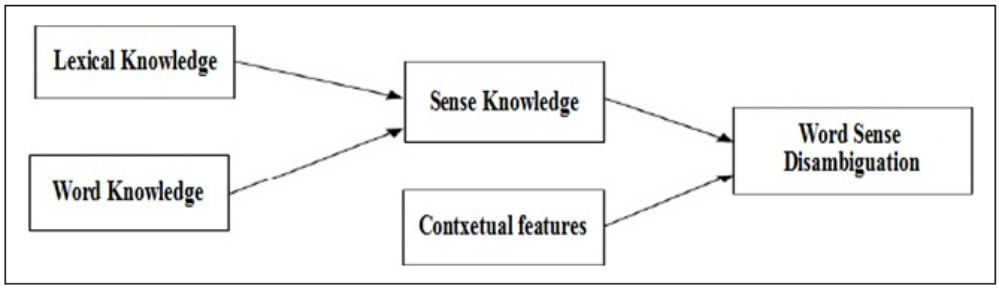

Figure 1. conceptual model for word sense disambiguation [8]

For example, consider the following phrase:

I can hear bass sounds.

They like grilled bass.

In these two phrase, the word bass is used with two different meanings: the first phrase the concept of the bass sound. And the second term is a kind of fish [9].

In the following sections, will be reviewed methods and techniques for word sense disambiguation. In the end, Will be discussed To Implementation, Conclusion and Suggestion for the future Work.

II. Word Sense Disambiguation Prepocrssing

Preprocessing disambiguation of words Include: tokenization, Part of speech tagging, lemmatization, chunking and Parsing. Is described below.

2.1. Tokenization:

Input phrase, Must, the constituent words (Token), Segmented.

2.2. Part of speech tagging:

Determine the role that each word in this phrase such as the: Nouns, verbs, adjectives or more.

2.3. lemmatization:

Any word in any position when placed in terms of grammar, Prefixes and suffixes stick to it.

2.4. Chunking:

According to the rules of linguistics, when two or more words, are placed together. Cause, the phrase special rule will be produced. Example: noun phrase or verb phrase.

2.5. Parsing:

Determining the sentence syntactic structure, Commonly, with the help draw the tree parser. where in, According to available data, Build tree. In order to observe. Is the structure of this statement is true or not?

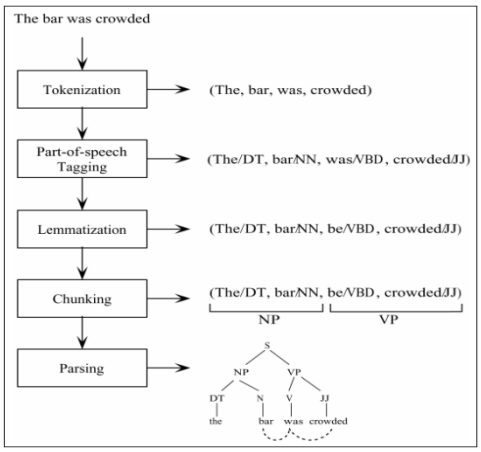

Figure 2. determine the correct sentence structure [10]

Figure 2, Represents the pre-processing workflow. So that we can examine the validity of such structures. As a result of pre-processing is done on the part of the text, the section can be a sentence, paragraph or a complete text. Every word includes a vector species is distinctive characteristics.

III. Learning Methods and Word Sense Diambiguation Techniques

In the following section, we will explain briefly four methods including the supervised method, unsupervised method, semi-supervised method and dictionary and knowledge-based methods.

3.1. Unsupervised method

This method, Is very powerful and portable. in fact, it is handheld. And does not require resources such as hierarchical concepts and handy glossary and sources of knowledge [11]. Words of similar meaning, In the vicinity of each other. With this method, cannot access to source of sense and Separating them based on the sense of words. Not used any label, ontology and dictionary for find the sense of words. and also, sense of words Clustering Occur in the input text. There is not knowledge base. Find the sense of words, Is possible. Based on its neighbor's word. The aim of this method, Determine the categories for each of the sense. Techniques of this approach include: Context clustring, Word clustring, Cooccurrence Graphs. an unsupervised word sense disambiguation algorithms into two general classes [14]:

A) Word sense disambiguation based on Token: In this approach, Find the sense of word by exploring similarities or the relationship between an ambiguous word and its context.

B) Word sense disambiguation based on the type: In this approach, a Word sense disambiguation, Simply, can be done by assigning all common examples for an ambiguous word.

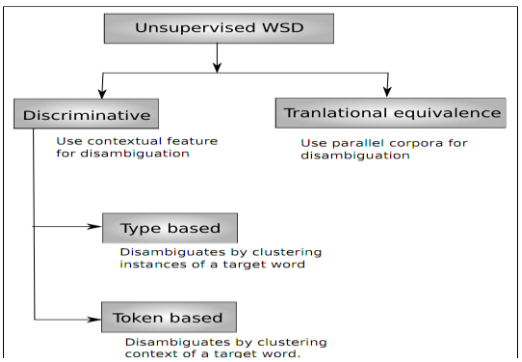

Figure 3. Different approaches to Unsupervised Word sense disambiguation [15]

3.2. Supervised method

Supervised Methods, mainly, Selects the context of words for disambiguation and includes Supervised process and training and testing phase. a disambiguation of words is available for training. In the training phase, a set of Sense-annotated training is needed. Training set consists of syntactic and semantic features. This method creates a classification using machine learning techniques, and based on training data, Appropriate decisions be considered. Categories of texts and sense of words, Occurs automatically rather than manually. a word sense disambiguation based on supervised learning methods are best performance in SenseEval and SemEval workshops [16]. The benefits of Supervised method, That, Focuses on word sense disambiguation with high performance [10]. This method is Inflexible and expensive Due to lack of information, and is difficult providing a training data set, that a majority of the concepts covered his words have [16]. Supervised method algorithms Includes Support vector machines, decision trees, neural networks, decision lists, Bayesian, k nearest neighbor and Boosting.

3.3. Semi-supervised method

In semi-supervised learning method, Data between supervised learning (with training data labeled) and unsupervised learning (without any training data with labels) is used. Many machine learning researchers, found that, unlabeled data, when used in conjunction with a small amount of labeled data, can make a significant improvement in the accuracy of learning. Obtain the labeled data for learning difficulties, often needs a skilled human factor and Expensive Process and time consuming. while, relatively inexpensive Access to Unlabeled Data and spends less time [12]. Many semi-supervised algorithms have been used for natural language processing tasks [13].

3.4. knowledge-based method

Knowledge-based methods, focuses on Word Sense Disambiguation with your text comparison with information from a source of predefined words such as WordNet [17]. The aim of Word Sense Disambiguation Based on the knowledge base and dictionary, Using the knowledge base resources such as dictionary, ontology, thesaurus, etc., for understand the meaning of words in a sentence. These methods are less efficient than supervised methods, but supervised methods cover the range of less than words. Therefore, is better Using the knowledge base covering all the words. Knowledge-based approach, focuses on the calculation of similarities between the two sentences based on semantic information collected from the knowledge base [18]. Typically, used some semantic similarity metric for calculating the relationship between concepts. In 1986 by Lesk and in 2004 was considered by mccarthy [19].

3.4.1. Lesk algorithm

The easiest method is based on the knowledge base for calculating the target words overlapping concepts. This method is known in terms of gloss overlap or Lesk algorithm.

This method operates as follows:

Two words (w1, w2) as input. And from knowledge base, Extracts Concepts of words.

$$(1) \quad S_1 \in \text{Sense}(w_1) \text{ and } S_2 \in \text{Sense}(w_2)$$

Then, do the analysis of whether these concepts are similar or not. In other words, Common points between the concepts acquires (3).

$$(2) \quad \text{score}_{\text{lesk}}(S_1, S_2) = |\text{gloss}(S_1) \cap \text{gloss}(S_2)|$$

And both a concept that has more subscriptions to each other, As the Concepts words chosen target (10).


```
for each sense i of W1
for each sense j of W2
compute Overlap (i,j), the number of words in common
between the definitions of sensei and sense j.
find i and j for which Overlap (i,j) is maximized.
assign sense i to W1 and sense j to W2.
```

Algorithm 1. Lesk algorithm based on a dictionary [20]

As specified that the above algorithm, First, Concepts are extracted from the database for each word. Then, each of the first word with the second word will review all the concepts. That is, count The number of words that are in common between these two concepts. And finally, they are a concept that has more in common, Considered as the meaning of any words. The main Lesk algorithm to Word Sense Disambiguation in short phrases deals. According to a word that is selected for disambiguation, a dictionary definition or set of concepts for words within a phrase considered. This algorithm is intended for new words and to concepts that have already been allocated, not used.

IV. IMPLEMENTATION

In discussing the implementation of the new method to be implemented. For this, we used Lesk methods. So that, Lesk, to select the most appropriate concept, Use the overlap words. That is, each of the concepts ambiguous words with other words Concepts compared and any concept that have more in common chooses. In order to share more, there are many common words in both words. For example, consider two words: pine and Cone. This Two can give up the Pine Cone. Concepts derived from each of these words. as follows [20]:

Pine

1 seven kinds of evergreen tree with needle-shaped leaves

2 waste away through sorrow or illness

3 pine for something, pine to do something

Cone

1 solid body which narrows to a point

2 something of this shape, whether solid or hollow

3 fruit of certain evergreen trees (fir, pine)

If the above algorithm run on the concepts presented, we see the first concept from pine and the latter from cone has three subscriptions and the Most shared between concepts. So, first of pine and third means are suitable Cone. In the new method, we tried, instead of words similar to our selection criteria, Near the two words in terms of concept, Give rating. And each concept was to score more points, As the most appropriate concept, For the ambiguous word is selected. Implementation is that, at first, a sentence as input is received, in which specified that the ambiguous word. Then, we separating the sentence into its constituent words. In the next step, we remove Stop Word. now, Should the notions of words remaining to be extracted. for this job, WordNet is used. In order to obtain the concepts of words from WordNet. WordNet external source in a way we implement, that receives a word. And that word can have all the roles that review. Such as nouns, verbs, etc. Then, it means That we define a role for returns. Word may have more than one meaning in any of the roles, WordNet external source gives us all of them. Below is an example of how to use WordNet to see. For example, the word "eat" in all roles that can be will check. Then it means the returns for all roles. It also means it can be extracted in a certain role.


```
word = "eat";
String pos = wordnet.getBestPos(word);
System.out.println("\n\nDefinitions for " + word + ":");
// Get an array of glosses for a word
String[] glosses = wordnet.getAllGlosses(word, pos);
// Display all definitions
for (int i = 0; i < glosses.length; i++) {
    System.out.println(glosses[i]);
}
```

Figure 4. is an example of how to use the word net

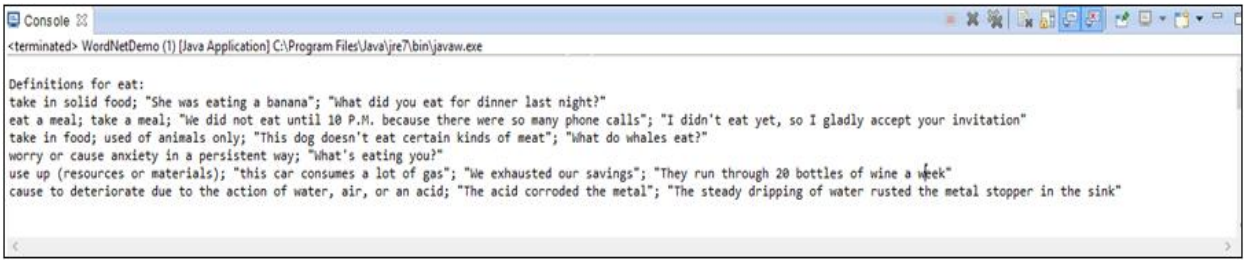

Figure 5. disambiguation is done on this text.

Disambiguation, is as follows that, If the word ambiguous, In WordNet external source, does not exist, Unfortunately, our algorithm cannot be disambiguation. Of course the problem here is that external source WordNet not contain that word. Given that currently the most comprehensive WordNet knowledge base is available, there are many meanings of words in WordNet, So, less confronted with this problem. If the word ambiguous, have Only One Sense in WordNet external source. Obviously, not in need of disambiguation. And get ready ambiguous sense of the word. If the word is ambiguous in WordNet external source, has more than one meaning, must be disambiguation as the most appropriate means to obtain it. So far, we have the word of the input sentence, That, stop words have been removed. now, the remaining words concepts must be extracted from a WordNet external source and put in One collection. For disambiguation the ambiguous word, should any of the concepts to be compared with this collection and every one that more score was to be chosen as the most suitable concept. Pseudo-code algorithms listed below:

```
for every word w[i] in the phrase
  let BEST_SCORE = 0
  let BEST_SENSE = null
  for every sense sense[j] of w[i]
    let SCORE = 0
    for every other word w[k] in the phrase, k != i
      for every sense sense[l] of w[k]
        SCORE = SCORE + number of words that occur in the gloss of
                        both sense[j] and sense[l]
      end for
    end for
    if SCORE > BEST_SCORE
      BEST_SCORE = SCORE
      BEST_SENSE = w[i]
    end if
  end for
  if BEST_SCORE > 0
    output BEST_SENSE
  else
    output "Could not disambiguate w[i]"
  end if
end for
```

Algorithm 2. Pseudo-code Lesk algorithm [6]

The algorithm works this way, for every word, all their meanings are extracted from a WordNet external source and put in a collection. Then each ambiguous word meanings with every word in this collection we compare and We score to it, and scores are added to the previous sum of scores. as a result, to each of the different meanings of ambiguous words, it will be given a score. The most appropriate means for ambiguous word will be chosen. In the following, each of meanings that have more score, has been selected as the most appropriate sense for the ambiguous word.

For example, the ambiguous word in the following sentence, is "accident". We want, disambiguation it. He crossed out the name Pool `It is, of course, no accident," he said out loud, testing to see if the words would come out on a printed page in a

bound volume, `that redundant theological speculation about the death of God should run parallel with an equally tedious literary preoccupation with the death of the novel".

The output is as follows:

An unfortunate mishap; especially one causing damage or injury

0: 3585.21074513431

Anything that happens suddenly or by chance without an apparent cause

1: 1751.588465461927

For the accident, there are two senses in the WordNet external source that you see. After running the algorithm to each is given a Score. The first Sense is that more Score and as the most appropriate Sense, for the ambiguous word "accident" is selected. Scoring algorithm is that two words are received and according to their proximity, One number of returns. That whatever is greater, indicates that the two words are closer together.

Typically for testing algorithms and to understand whether the algorithm, Correct doing Disambiguation, from xml Standard file is used. With the "instance" tag begins and text within the "context" tag and the ambiguous word in the text with the "head" tag is specified. "instance " Tag, included is one id. That, specifies the ambiguous word ID. The correct meaning of any ambiguous words in a separate file at our disposal, so that the output of the algorithm was tested with it. In this paper, we used sensual data sets. That More than 12,000 samples of text contained in it. In the following, for example, given a text from sensual data sets. The word accident is known as the ambiguous word.

<instance id="accident-n.700274">

<context>

He crossed out the name Pool .

`It is, of course, no <head>accident</head>," he said out loud, testing to see if the words would come out on a printed page in a bound volume, `that redundant theological speculation about the death of

God should run parallel with an equally tedious literary preoccupation with the death of the novel ".

</context>

</instance >

True meaning of the word "accident", A separate file at our disposal, But the word "accident", According to the text in it and can have different meanings. I'd give it a meaning so as to recognize the right text. In the figure below, True meanings of the word "accident", With, id text in it and is located, is given. In this text, True meaning of the word "accident", chance is intended. (In the figure id = 700274).

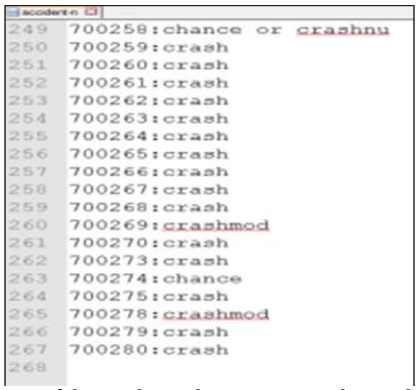

Figure 6. True meanings of the word "accident" is given with text id in it and is located.

First, this algorithm, the disambiguation word head that same word is ambiguous with regard to the text in it. So that, the output of the algorithm, has earned the highest score and has been selected as the most appropriate concept for the ambiguous word. To verify that the proposed algorithm, the word head, once again with true meaning of its concepts, that we had in separate files, We disambiguation. To do this all means, correct meaning ambiguous words are extracted from an outside source WordNet and put in a collection. Then do disambiguation. If the output is the same meaning that was selected in the previous step, means that the algorithm is working properly. Otherwise, the algorithm is not complete and cannot be trusted to verify validity. As previously observed for "accident" ambiguous word in the text below, the first meaning is more score and the "accident" ambiguous word is considered as the most appropriate means.

head: accident

context: He crossed out the name Pool It is of course no accident he said out loud testing to see if the words would come out on a printed page in a bound volume that redundant theological speculation about the death of God should run parallel with an equally tedious literary preoccupation with the death of the novel

Output:
an unfortunate mishap; especially one causing damage or injury

0: 3585.21074513431

, anything that happens suddenly or by chance without an apparent cause

1: 1751.5884654619276

For correctly accuracy of the algorithm, the word "head", Once again the true meaning of its concepts, we have seen is intended "chance", We disambiguation. To do this, all word meanings chance is extracted from a WordNet external source and put in a series, Then We disambiguation. The output is as follows:

An unfortunate mishap; especially one causing damage or injury

0: 271.7933494839498

anything that happens suddenly or by chance without an apparent cause

1: 138.0395552425189

As you see, this time, the first meaning, has more score. as a result, the algorithm works properly. Lesk algorithm complexity is high and its running time is relatively high. Besides that, in some of the words, run error.

V. Conclusion and Future Work

In this paper, we investigate the words sense disambiguation in natural language to find the correct meaning of a word in a sentence is applicable, will be discussed. The main routine disambiguation system has two parts: First, a knowledge base and Second, an algorithm that can extract the proper sense of the knowledge base. In the implementation, the typical Lesk algorithm that one of the branches of knowledge and dictionary-based method was used. In fact, this algorithm works on the basis of similarity and to enhance the accuracy of the results, the algorithm was combined with the algorithm Scoring. The scoring algorithm performance is that two words to get and according to their proximity returns a number that is much larger, suggesting that the two words are closer together. To extract text and data collection were used WordNet and the texts on which the operation was carried disambiguation for xml structure was introduced. In Figure 7, we introduced a new plan to implement in the future. According to the testing and evaluation, a number of algorithms widely considered and considering the strengths and weaknesses of them for new ideas in this field are required That the Lesk algorithm and one of these algorithms combined with observer methods to be used because of the advantage of high precision performance.

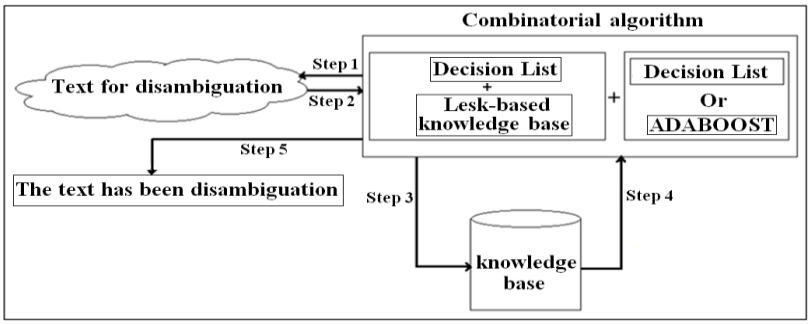

Figure 7. Combinatorial algorithm

References

1. J. Sarmah, Sh. Kr. Sarma (2016). "Decision Tree based Supervised Word Sense Disambiguation for Assamese".International Journal of Computer Applications (0975 – 8887), Volume 141 – No.1.
2. M. Kocaleva, D. Stojanov, I. Stojanovik, Z. Zdravev (2016). " Pattern Recognition and Natural Language Processing: State of the Art". TEM Journal 5(2), pp. 236–240.
3. M. Sawant, T. Sangoi, S. Nair (2016). "Supervised Word Sense Disambiguation". International Journal of Science and Research (IJSR). Volume 5 Issue 10, pp.1845-1848. ISSN (Online): 2319-7064.
4. Z. Chun-Xiang, S. Lu-Rong and G. Xue-Yao (2016). "Determine Word Sense Based on Semantic and Syntax Information". International Journal of Database and Theory and Application, Vol.9, No.2 (2016), pp.17-22.
5. I. Iacobacci, M. T. Pilehvar and R. Navigli (2016). "Embeddings for Word Sense Disambiguation: An Evaluation Study". Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.897–907.
6. S. Banerjee, T. Pedersen (2002). "Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet". Submitted in partial fulfillment of the requirements for the degree of Master of Science. Department of Computer Science, University of Minnesota, Duluth, Minnesota, U.S.A.
7. Z. Zhong, H. Tou, Y. Chan (2008). "Word Sense Disambiguation Using OntoNotes: An Empirical Study". Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 1002-1010, Honolulu.
8. P. Bala (2013). "Knowledge Based Approach for Word Sense Disambiguation using Hindi Wordnet". The International Journal Of Engineering And Science (IJES), Volume 2, Issue 4, Pages 36-41, ISSN(e): 2319-1813 ISSN(p): 2319-1805.
9. E. Palta, Prof. Om Damani (2007). "Word Sense Disambiguation". Submitted in partial fulfilment of the requirements of the degree of Master of Technology, Kanwal Rekhi School of Information Technology Indian Institute of Technology, Powai, Mumbai, pages 1-3.
10. R. Navigli (2009). "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. 41, No. 2.
11. G. Tomar, M. Singh, Sh. Rai, A. Kumar, R. Sanyal, S. Sanyal (2013). "Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation". IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2. ISSN (Print): 1694-0814, ISSN (Online): 1694-0784 www.IJCSI.org.
12. Ms. Ankita Sati (2013). "Review: Semi-Supervised Learning Methods for Word Sense Disambiguation". IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 4, PP. 63-68. www.iosrjournals.org.
13. A. Søgaard (2011). "Semisupervised condensed nearest neighbor for part-of-speech tagging". Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers, pages 48-52, Portland, Oregon.
14. S. Brody, R. Navigli, M. Lapata (2006). "Ensemble Methods for Unsupervised WSD". Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 97-104, Sydney.
15. D. Chaplot, Prof. Bhattacharyya (2014). "Literature Survey on UnsupervisedWord Sense Disambiguation". Department of Computer Science and Engineering Indian Institute of Technology, Bombay.
16. Z. Zhong, H. Tou Ng (2010). "It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text". Proceedings of the ACL 2010 System Demonstrations, pages 78-83, Uppsala, Sweden.
17. R. Cabrera, P. Rosso, M. Gómez, L. Pineda, and D. Avendaño (2009). "Semi-supervised Word Sense Disambiguation Using the Web as Corpus". A. Gelbukh (Ed.): CICLing 2009, LNCS 5449, pp. 256–265.
18. Ch. Ho, M. Murad, R. Kadir, Sh. Doraisamy (2010). "Word Sense Disambiguation-based Sentence Similarity". Coling 2010: Poster Volume, pp. 418–426, Beijing.
19. E. Agirre, A. Soroa (2009). "Personalizing PageRank for Word Sense Disambiguation". Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 33-41, Athens, Greece.
20. D. I. MORARIU, R. G. CRETULESCU, M. BREAZU (2012). "Word Sense Disambiguation for Text Mining", The third international conference in Romania of Information Science and Information Literacy, ISSN 2067-9882.

AUTHOR(S) PROFILE

Fatemeh Fathi Sedehi, she received Bachelor of Information Technology Engineering from Islamic Azad University of Zanjan. Also, She Graduated Master of Information Technology, Computer Networks Orientation in University of Guilan from Rasht in Iran. And her main research interest includes: e-commerce and RFID.

Kobra Khoshraftar, she received Bachelor of Software Engineering from Islamic Azad University of Roudsar and Amlash. Also, She Graduated Master of Software Engineering from Islamic azad University of Chaloos in iran. And now, She cooperation with young researcher club in Islamic Azad University of Lahijan. And her main research interest includes: software and wireless sensor networks.

Ali Vahid Roudsari, he received Bachelor of Software Engineering from Islamic azad University of Khalkhal. Also, he Graduated Master of Information Technology, Computer Networks Orientation in University of Guilan from Rasht in Iran. he Collaborated with young researcher club in Islamic Azad University of Roudsar and Amlash for 5years, his main research interest includes: Data mining in medical field and and e-commerce.

