Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation

Gaurav S Tomar, Manmeet Singh, Shishir Rai, Atul Kumar, Ratna Sanyal, Sudip Sanyal

Indian Institute of Information Technology
Allahabad, Uttar Pradesh, India

Abstract

This paper presents an unsupervised approach for disambiguating between various senses of a word to select the most appropriate sense, based on the context in the text. We have defined a Probabilistic Latent Semantic Analysis (PLSA) based Word Sense Disambiguation (WSD) system in which sense tagged annotated data is not required for training and the system is language independent giving 83% and 74% accuracy for English and Hindi languages respectively. Also, through word sense disambiguation experiments, we have shown that byapplying Word net in this algorithm, performance of our system can be further enhanced.

Keywords: *Word Sense Disambiguation, Probabilistic Latent Semantic Analysis, Word net, Algorithm*

1. Introduction

In the field of natural language processing, “Word Sense Disambiguation (WSD) is defined as the problem of computationally determining which "sense" of a word is activated by use of the word in a particular context” [1]. Sense disambiguation is used in NLP applications like Document Summarization, Document Categorization Information Retrieval,Information extraction and Text mining, Lexicography, Machine Translation, Document Similarity measurement, Document Classification, Question Answering systems and Cross Language Information Retrieval [1]. The paper provides a new technique for performing WSD that governs the process of identifying which sense of a word is used in a sentence, when the word can have multiple meanings (polysemy) [34]. For example, word “bank” can have up to 10 senses as per Word Net (the semantic lexicon for English Language). Consider the following two sentences: “He went to the bank to deposit money” and “The temple is situated on the bank of Ganga”. Here we have used the word “bank” in both the sentences but the sense in which the word “bank” comes is entirely different in the two sentences. A WSD system expects a sentence and it has to suggest the sense of each ambiguous word based on the context in which that word appears.

Various approaches have been proposed for WSD which can be categorized as Knowledge Based, Supervised, Semi supervised and Unsupervised. Knowledge based approaches depend on resources like dictionaries, ontologies and collocations. In this paper we present the induction of word sense using an approach based on word clustering which clusters semantically close words using a purely statistical method, Probabilistic Latent Semantic Analysis (PLSA) and second order co-occurrence which generates rich and informative clusters. PLSA, using tempered Expectation Maximization (EM), is then used to generate *‘k’* clusters (containing semantically similar words), each representing a certain concept/topic. These clusters are further expanded by enriching them with more semantically related words like synonyms, hypernyms and homonyms using Word Net which is a lexical database. The correct sense of a polysemic word, present in the test corpus, is then deduced by calculating the similarity between the test corpus having the target polysemic word and clusters generated earlier. The cluster with highest similarity score is attributed to be the most appropriate cluster representing the sense of the polysemic word.

Unsupervised approaches are highly robust, portable and do not require resources like concept hierarchies, dictionaries and hand crafted knowledge resources. Classical unsupervised approach for Word Sense disambiguation, Lesk Algorithm, was based on the hypothesis that words in a given neighbourhood will tend to share a common theme [35]. It compares the set of words enclosed in the dictionary definition and examples of a polysemic word with the words present in its neighborhood. Unfortunately, Lesk’s approach is perceptive to the exact wordings of the definitions in the dictionary. So, the existence / nonexistence of a certain word can drastically change the results. Further, the algorithm calculates overlaps only among the glosses and examples of the senses being measured [35]. This is a significant limitation of the dictionary that glosses tend to be fairly small and do not provide sufficient vocabulary (glossary) to communicate fine-grained sense distinctions [36]. That is why Lesk’s approach gives less accuracy due to its dependence on dictionary definitions and examples to build the set of words with which it compares the set of words present in the neighborhood of the ambiguous word.

Given the above limitations of the Lesk approach, one can think of three possible directions for improving it. These are:

(a) Improve the glosses

(b) Improve the method used for measuring similarity / distance

(c) Improve the context of the ambiguous word whose correct sense has to be identified.

The unsupervised approach proposed here considers the first possibility i.e. it tries to find a better set of words with which it measures the similarity of the words appearing in the context of the given ambiguous word. This different mechanism is based on a purely statistical method of Probabilistic Latent Semantic Analysis explained later in this paper which gives better set of words (clusters) to decide sense of a word in a given text. Set of words (clusters) obtained from PLSA are more rich and informative than set of words obtained using Lesk's approach.

In the next section we shall give a brief review of the existing approaches. The third section contains the details of our approach. In particular, this section contains a brief overview of the PLSA and the way it is used for obtaining word clusters that can be used for performing WSD. The fourth section describes our results. Since the method adopted is fairly generic, we have tested it on two different languages, English and Hindi. The fifth section contains the concluding remarks and suggestions for further improvement.

2. Related Work

Most of the unsupervised methods for Word Sense Disambiguation are based on similarity methods and graph based methods. Graph based methods have two steps [16][17]. In the first step, a graph is constructed from the lexical knowledge based on possible hidden meaning representation of all possible compilations of the word whose sense is being disambiguated. Basically, graph nodes represent possible senses and edges of graph correspond to relations between senses. This graph structure is then used to find the value of each node in the graph [19] [20]. In this way the graph is used for finding the most appropriate sense of the word. In a graph based approach proposed by Veronis [3], firstly a co-occurrence graph is formed in which the nodes are words appearing in the paragraph of the text corpora in which target word exists. An edge between a pair of words is added to the graph if they reappear in same paragraph. Finally, a minimum spanning tree is used to disambiguate specific examples of the target word. A graph based method that uses the degree of centrality for WSD has also been explored [6].

Recently Navigli [19][20] proposed a graph based algorithm for large scale WSD which does not require sense annotated data for the training but has to investigate the graph structure in its whole depth. This method [14] aims to capture word sequence dependency information in the given corpora. Similarity based methods use clustering which can be further categorized in two types of clustering - word clustering and context clustering. Similarity based approach assigns a best suited sense to a word with the help of its surrounding words. In the similarity based algorithm, sense is calculated for each word individually whereas in graph based approaches, the sense of a word is found along with all its neighboring words with the help of dependencies across senses. An approach based on context clustering depends on the concept of word space [26]. Word space method derives a vector for each word in the corpora using a co-occurrence matrix. But, this word space has a large dimension and so we need to apply Singular Value Decomposition for reducing the dimensionality of the word space. Many unsupervised approaches depend on context clustering.

Recent years have popularized unsupervised methods based on word clustering like Probabilistic Latent Semantic Analysis [9] and Latent Dirichlet Allocation [2] etc. for information retrieval, document similarity [11], Keyword spotting [32] and document categorization [7]. The proposed approach in this paper also uses word clustering. Agglomerative clustering of words [22] considers each word as a single cluster initially and then proceeds by including similar words into the same cluster until a predefined threshold is reached. This approach has been successfully used in biomedical domain [25]. But, the above approach needs a large amount of unlabeled training data for the construction of context vectors. A LDA based approach has been used for discovering concepts, which has subsequently been used for WSD and showed improvement over conventional methods [24]. However, the authors have considered domain specific documents for LDA as most words tend to have same sense in a specific domain. Similar idea was also explored for medical documents [28].

Clustering By Committee (CBC) takes a word type as input and finds clusters of words that represents each sense of the word [13]. Hyperspace Analogue to Language (HAL) is based on word by word co-occurrence statistics [4]. HAL does not include large units of context and it captures co-occurrence data for words by considering a window of 10 words. There is a fully unsupervised method that is to cope with WSD [12] using a mono-lingual corpus but uses a different clustering approach. Traupman and Wilensky performed experiments [29] to improve discrimination accuracy of Yarowsky's Classifier using an iterative approach to re-train the classifier, using part of speech knowledge and training using weighing of senses distributed to dictionary order. Yet another method [5] for unsupervised WSD combines multiple information sources, including semantic relations, large unlabeled corpora and cross lingual distributional statistics. There exists an unsupervised learning method using Expectation Maximization algorithm [21] [27] which perform an optimal number of iterations of CV EM and CV EM2. In this unsupervised approach only a dictionary and an un-annotated text are required as input. This proposed method overcomes problem of brittleness present in many existing methods.

3. Proposed Method

3.1 Probabilistic Latent Semantic Analysis

PLSA depends on latent class model (or aspect model) [8][9] which consists of latent class variables. These latent class variables represent aspects/topics (senses in our case) and the model uses Expectation Maximization (EM) Algorithm for maximum likelihood estimation. Tempered EM can be used for better results. Probabilistic Latent Semantic Analysis relates the latent variables $\{z_1, z_2, \ldots z_k\}$ with the documents $\{d_1, d_2 \ldots d_i\}$ in addition to the terms $\{w_1, w_2, \ldots w_j\}$. There are three variables related with this approach as defined below:

- Select a document d_i with probability $P(d_i)$.
- Pick a latent class or concept z_k with probability $P(z_k \mid d_i)$.
- Generate a word w_j with probability $P(w_j \mid z_k)$.

The distribution P(w,d) can be written as

$$P(d_i, w_j) = P(d_i)\, P(w_j \mid d_i) \quad (1)$$

Where

$$P(w_j \mid d_i) = \sum (P(w_j \mid z_k)\, P(z_k \mid d_i)) \quad (2)$$

Or, using Bayes' rule, we get

$$P(d_i, w_j) = \sum_k (P(z_k)\, P(w_j \mid z_k)\, P(d_i \mid z_k)) \quad (3)$$

The above distributions can be found by using Expectation Maximization Algorithm which consists of two steps namely the expectation step and the maximization step. In the expectation step the posterior probabilities for latent parameters are calculated based on the current estimates of the parameters. For the Expectation:

$$P(z_k \mid d_i, w_j) = \frac{P(z_k)(d_i \mid z_k) P(w_j \mid z_k)}{\sum_{r=1}^{K} P(z_r)(d_i \mid z_r)\, P(w_j \mid z_r)} \quad (4)$$

In the maximization step we update the parameters based on the maximized log likelihood probability, based on the values calculated in the expectation step. We get the following set of equations:

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_j, w_j) P(z_k \mid d_i, w_j)}{\sum_{s=1}^{M} \sum_{i=1}^{N} n(d_j, w_s) P(z_k \mid d_i, w_s)} \quad (5)$$

$$P(d_i \mid z_k) = \frac{\sum_{j=0}^{M} P(d_i \mid w_j) P(z_k \mid d_i, w_j)}{\sum_{q=1}^{N} \sum_{j=1}^{M} n(d_q, w_j) P(z_k \mid d_q, w_j)} \quad (6)$$

and

$$P(z_k) = \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i \mid w_j)} \sum_{i=1}^{N} \sum_{j=1}^{M} P(d_i \mid w_j)\, P(z_k \mid d_i, w_j) \quad (7)$$

In the above expressions, $P(w_j \mid z_k)$ represents the probability of observing a particular term or word in a given concept. $P(z_k \mid d_i)$represents the probability of a topic in a given document, $P(z_k)$ is the probability of a topic and $n(d_i, w_j)$ is number of times a term or word occurs in a particular document d. Now we iterate through these steps again and again till convergence. We have used the parameter $P(w_j \mid z_k)$ to create word cluster in PLSA because they essentially represent the probability of finding the word w_j in the cluster z_k. If the number of concepts is K, then we will have K clusters. The set of words belonging to a cluster will be those words that have a probability above a threshold value in that cluster. In this approach a word may belong to more than one cluster.

3.2The WSD System

In this paper we propose a WSD system based on word clusters obtained using PLSA. The system consists of two phases: the training phase and the testing phase. The training phase creates the word clusters that will be used for disambiguating and the testing phase performs the actual disambiguation. These phases are described below.

3.2.1The WSD System

In the training phase, the words that relate to a particular context are grouped together and each such grouping (cluster) represents one sense. In Figure1, we present the architecture of the training phase of the proposed system.

Step 1- Removal of stop words from training data:

Stop words are the high frequency words that have very low semantic value. These words comprise 30% of the whole training data and hence must be removed. Usually one maintains a list of such words. Any word in this list is not passed to the subsequent steps. For English language, {is, am, on, off, the, a, an, about} are some of the stop words.

IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, September 2013
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org

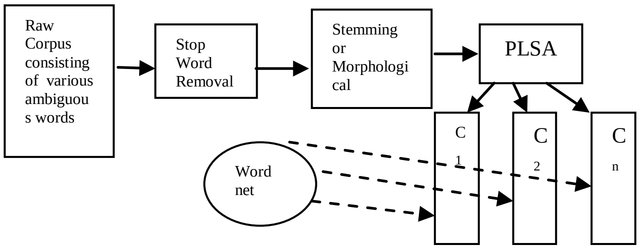

Fig. 1 Training Phase Architecture for the WSD System.

Step 2- Reduction of inflectional and derivational variants to their root form (Stemming):

To reduce the size of the training corpus to be processed further, we use a statistical stemmer for reducing inflectional or derived words to a reduced form that may or may not be the morphological root of the words. It is not necessary that the stemmed words should give the morphological root of the word. It is sufficient that similar words map to the same stem. For instance, the words "call", "caller", "calls" map to same stem "call" after stemming.

Step 3- Word clustering using PLSA**:**

PLSA helps in clustering similar words or the words related to a particular topic together by giving $P(w_j |z_k)$ as output for each term 'w' and topic 'z'. We take the list of probabilities $P(w_j |z_k)$ for each topic 'z_k' and sort the list so that among (w_1, w_2,…..w_j,..)**,** only the words which represent a topic/sense strongly come together. After sorting, for each topic words having low $P(w_j |z_k)$ value are removed as they do not represent the topic so strongly and perhaps belong to some other topic/sense. This step finally results in clusters/topics, each having words closely related to each other.

Step 4 - Expansion of clusters using a lexical database:

Although the clusters obtained from the previous step contain rich and informative words which strongly represent the topics, but the number of words in each cluster is quite small. Hence, disambiguation done with this set may not give best possible results. The list of words in each cluster, which we got as output of PLSA in previous step, are further expanded by including more semantically related words with the help of ontology in Word Net which is a lexical database consisting of the semantic networks of words, their synsets, homonyms, hypernyms and hyponyms. After the expansion of word list in each cluster, we have a significant number of words which define a particular sense in that cluster.

3.2.2The WSD System

After getting clusters of similar words representing topics (senses) from training phase, an ambiguous word present in test corpus can be classified into one of these clusters obtained from training corpus by computing the similarity score of test corpus containing the target ambiguous word with each cluster. The cluster with which the highest similarity score is obtained is attributed to be the most appropriate sense of the queried ambiguous word. The idea here is that word lists (clusters obtained from training phase) and expanded with the help of WordNet as explained in step 4 of training phase consists of all the related words which make up a sense. Overlap of these extended lists with test corpus can be calculated and then the cluster/word list having maximum overlap with the test corpus gives the best representation of the sense for the ambiguous word in the test corpus. The proposed WSD system uses cosine coefficient as a similarity score measure. Our experiments show that using other similarity score measures like Dice Coefficient, Jaccard coefficient etc. give similar results.

4. Result and Analysis

Presently we have evaluated our algorithm on two languages, English and Hindi. We have collected various categories of ambiguous words. In the first category, some words have two senses and some even have more than two senses. In the second category, the part of speech of the ambiguous word may be different for different senses. In the third category, the origin of different senses of the ambiguous word is same. For example, the word "bank" has originated from the word "safe". Thus, a financial bank means it is safe for money and a river bank means it saves from flowing water. We created a corpus of more than 150 documents containing instances of all the three categories. Within this corpus we earmarked 50 words which appeared in multiple senses. We also evaluated our WSD algorithm performance with, as well as without, using WordNet to expand the clusters. Our results for some words are summarized in Table 1. The first column shows the ambiguous word and its sense. The second column shows the number of occurrences of the ambiguous word in the test corpus. The third column gives the number of correct matches and the last column shows the accuracy of correct sense identification for a given ambiguous word. In Table 1 we have presented the results when the word clusters are expanded using WordNet.

An examination of Table1 shows that we get an average accuracy of 83.17%. As discussed earlier, we can put ambiguous words in three categories. We will now analyze the accuracy of disambiguation for each category separately. The first ambiguous word, "well" have senses belonging to different part of speech. The first sense of "well" is a noun while the second sense of "well" is an adverb. Our proposed approach has given similar accuracy (88%) for both types. We now consider the word "cricket" which can mean a sport or an insect. In this case both senses have the same part of speech - noun. In this case our proposed approach has given 100% accuracy.

Table 1: Accuracy of disambiguating several English ambiguous words

Ambiguous Word and Sense	**Number of Occurrencesin Corpus**	**Number of Correct matches**	**Accuracy percentage**
Well (Deep Hole)	16	14	88
Well (Good)	8	7	88
Fair (Gathering)	42	34	81
Fair (Good)	13	13	100
Fair (Travelling)	40	40	100
Book (Booking)	8	3	38
Book (Notebook)	15	15	100
………………….	……………..	…………	……….
………………….	……………..	…………	………..
Cricket (Game)	13	13	100
Cricket (Insect)	9	9	100

Average Accuracy = 83.17%

Now let us consider the word “fair”. In this corpus, word “fair” has three possible senses. Two of the senses are nouns while the third sense is an adverb. The results show that we get different accuracy for different senses. For example, in the training corpus, word “fair” as an adverb is present with word “fair” as a noun in the sense of gathering. Thus, the clusters corresponding to these two senses have considerable overlap and hence lead to a drop in accuracy. A similar effect occurs in the case of the ambiguous word “book”. One sense of “book” is a Verb (as in – booking a ticket) while the second one is a noun (as in – to read a book). We again see a drop in accuracy. This happens because in the training corpus we have sentences like “someone has left this book on this seat” and “this seat has been booked by someone”. As we can observe, these sentences have the same collocation words leading to an overlap in the corresponding word clusters. We also observed that WordNet has an important role in increasing the accuracy. On an average there is an increase of 2 to 5% in accuracy when we expand the clusters using WordNet.

Analysis of proposed approach shows that unsupervised approach discussed here is quite superior compared to the Lesk algorithm. Moreover, the technique is language independent since it is a purely statistical approach. It merely requires a suitable, untagged corpus to build the clusters. Availability of a Word Net type of resource would enhance the accuracy but it is not mandatory. We are presenting some examples of the results obtained with our WSD tool.

Example 1

Word: Date
Possible senses
Sense 1: Fruit (Noun)
Related cluster snapshot: date, fruit, desert, eat, market
Sense 2: related to Calendar (Noun)
Related cluster snapshot: date, month, year, birthday, current, time, number
Sense 2: related to Love (verb)
Related cluster snapshot: date, love, girl, escort, travel, single, email, internet, personal

Example 2

Word: Kite

Possible senses
Sense 1: Bird
Related cluster snapshot: kite, nest, bird, snail, raptor, wings, water, tree
Sense 2: Paper Toy
Related cluster snapshot: kite, fly, festival, people, art, competition, dor, cut, sky, makar, rajasthan, sakranti

The results obtained for Hindi are summarized in Table 2.

Example 3

Word: फल

Possible senses
Sense 1: Result
Related cluster snapshot: सफलता [success],द्वीप [island],फल [result], परिणाम [result],असफलता [failure],प्रतिफल [failure]
Sense 2: Fruit
Related cluster snapshot: आम [mango],फल [fruit],भारत [India], खेल [game], मोटर [automobile]

Example 4

Word: सोना

Possible senses
Sense 1: Gold

Related cluster snapshot: सोना [gold], पैदल [foot], धातु [metal], रंग [color], सिक्का [coin]
Sense 2: Sleep
Related cluster snapshot: सोना [sleep],नीद [sleep], सो [to sleep], रात [night], दौरान [during]

Table 2: Accuracy of disambiguating several Hindi ambiguous words

Ambiguous Word and Sense	Number of Occurrencesin Corpus	Number of Correct matches	Accuracy percentage
फल (Result)	8	6	75
फल (Fruit)	8	7	87
सोना (Gold)	4	5	80
सोना (Sleep)	4	7	57
कलम (Pen)	4	5	80
कलम(Kill)	4	6	66
आम (fruit)	34	30	88
आम (common)	5	4	80
…………………………	………………………	…………………	………………
…………………………	………………………	…………………	………………
लाल(red color)	24	21	88

Average Accuracy = 74.12%

Now when a test corpus having ambiguous word फल or सोना is given as input to test the system, our experiments showed that the cluster representing the related sense had the maximum overlap with the test corpus and hence was returned as output**.** For instance, if सोना is used in sense of 'sleep', it tends to have words like रात [night], सो [to sleep] etc. as its neighbouring words and hence, will have maximum overlap with cluster having these words thereby helping to identify correct sense.

In the above samples we see that the related cluster snapshots contain many words that are very relevant. Thus, we can say that PLSA, with the enhancements proposed in earlier sections, leads to very good clustering of words and thereby increasing the accuracy of the disambiguation process compared to the Lesk algorithm. Analysis of proposed approach also shows that unsupervised approach discussed here is language independent and contrary to standard supervised approach do not utilize manually tagged data in any way.

4. Conclusion and Future work

In this paper we employed word clustering based on Probabilistic Latent Semantic Analysis for developing an unsupervised and a relatively generic WSD algorithm. Our experiment shows that the proposed approach is language independent and obtained state of art performance on well managed evaluation data sets giving 83% and 74% accuracy for English and Hindi languages respectively. Adopting WordNet enriched clusters further improve the accuracy in the range of 2 to 5%. This shows that cluster based WSD algorithms perform better with more sense inventories as we get more clusters and more words in them. WordNet like reference lexicon exist for several languages. It is really an interesting future direction to establish to see how well our WSD algorithm performs with other such needful resources. Performance of the system proposed here can linearly increase according to size of training data. Our results focused basically on cluster measures and their improvement using word net sense inventories. More research problems need to be assessed to see whether our results can be extended to other NLP problems, other than WSD.

References

[1] E. Agirre, and P. Edmonds, Word Sense Disambiguation: Algorithms and Applications Springer, New York, NY, 2006.
[2] D. Blei, M. Jordan, and I. Michael, "Latent Dirichlet allocation". Journal of Machine Learning Research Vol.**3**: pp. 993-1022, 2003.
[3] R. Bruce, and J. Wiebe, "Word Sense Disambiguation using Decomposable model," In Proceedings of the 32nd Annual meeting of the association for Computational Linguistics (ACL, Las Cruces, NM).pp.139-145, 1994.
[4] B. Curt, and K. Lund, "Modeling parsing constraints with high dimensional context space", Language and Cognitive Processes, Vol.12 No.(2-3) pp. 177-210 ,1997.
[5] R. Florian, and R. Wincentowski, "Unsupervised Italian Word Sense Disambiguation using Wordnets and Unlabeled Corpora," In Proceedings of SigLEX'02, pp. 67-73,2002.
[6] E. Hessami, F. Mahmoudi, and A.H. Jadidinezad, "Unsupervised Graph based Word Sense Disambiguation using lexical relation of Wordnet", In International Journal of Computer Science Issues, Vol.8, Issue 6, No. 3, pp. 225-230, 2011.
[7] T. Hoffmann, "Tutorial on Matrix Decomposition Technique in Machine Learning and Information Retrieval" In: 5th Max-Planck Advanced Course on the foundation of computer science, Sept 6-10, 2004 Soarbriicken.
[8] T. Hofmann, "Probabilistic Latent Semantic Indexing" In Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in information Retrieval. Berkeley, California, United States: ACM, 1999, pp. 50 – 57, 1999.
[9] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis", In Journal of Machine Learning, Vol. 42, Numbers 1 – 2, pp. 177 – 196, 2001.

[10]V.N. Huynh, J.T. Ngugen, and C. Annie, "Adaptovely Entropy based Weighting classifiers in Combination using Dempster – Shaper Theory for Word Sense Disambiguation," In Journal of Computer Speech and Language Elsevier Volume 24 pp. 461 -473, 2010.

[11]A. Kumar, and S. Sanyal, "Effect of Pronoun Resolution on Document Similarity, in International Journal of Computer Applications" Volume 1, No. 16, pp. 60-64, 2010.

[12]D. Levinson, "Corpus-Based Method for Unsupervised Word Sense Disambiguation", www.stanford.edu/ dmitryle/acai99w1.ps, 1999

[13]D. Lin, and P. Pantel, "Concept discovery from text", In Proceedings of the 19th International conference on Computational Linguistics (COLING) pp.577-583, March 2002 Taipei, Taiwan.

[14]S. Lin, and K. Verspoor, "A Semantics-Enhanced Language Model for Unsupervised Word Sense Disambiguation" In CICLing 2008, LNCS 4919, pp. 287–298, 2008.

[15]R. Mihalcea, "Bootstrapping large sense tagged corpora," In Proceedings of the 3rd International Conference on Language Resources and Evaluations (LREC, Las Palmas, Spain) 2002a .

[16] R. Mihalcea, "Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-Based Algorithms for Sequence Data Labeling," Proc. Human Language Technology and Empirical Methods in Natural Language Processing, pp. 411-418, 2005.

[17]R. Navigli, and P. Velardi, "Structural Semantic Interconnections: A Knowledge- Based Approach to Word Sense Disambiguation," IEEE Transactions Pattern Analysis and Machine Intelligence, Vol. 27, no. 7, pp. 1075-1088, July 2005.

[18]R. Navigli, and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation" IEEE Transactions on Pattern Analysis and Machine Intelligence Vol 32, pp. 678-692, 2010.

[19]R. Navigli, and M. Lapata, , "Graph Connectivity Measures for Unsupervised Word Sense Disambiguation," In Proceedings of 20th International Joint Conference Artificial Intelligence, pp. 1683-1688, 2007.

[20]R. Navigli, "Word Sense Disambiguation: A Survey" In Journal Computing Surveys, Vol. 41, N0.2, pp. 1-69, 2009.

[21]K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM" In Machine Learning, 39(2/3): pp.103–134, 2000.

[22]T. Pedersen, and R. Bruce, "Distinguishing word senses in untagged text" In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP, Providence, RI) pp.197 – 207, 1997.

[23]T. Pedersen, "Learning Probabilistic Models for Word Sense Disambiguation" Ph.D. dissertation Southern Methodist University, Dallas, TX , 1998.

[24]J. Preiss, and M. Stevenson, "Unsupervised Domain tuning to improve Word Sense Disambiguation," In Proceedings of NAACL-HLT 2013, pages 680-684 Atlanta, Georgia, 9-14 June 2013.

[24] J. Preiss, and M.Stevenson, "Unsupervised Domain tuning to improve Word Sense Disambiguation," In Proceedings of NAACL-HLT 2013, pages 680-684 Atlanta, Georgia, 9-14 June 2013.

[25]G. Savova, T. Pedersen, A. Purandare, and Kulkarni, "A. Resolving ambiguities in biomedical text with unsupervised clustering approaches" Res. rep. UMSI 2005/80. University of Minnesota Supercomputing Institute, Minneapolis, MN, pp/ 494-503, 2005.

[26]H. Schutze, "Dimensions of meaning" In Proceedings of the 1992 ACM/IEEE Conference on Supercomputing. IEEE Computer Society Press, Los Alamitos, CA. pp.787-796 1992.

[27]H. Shinnou, and M. Sasaki, "Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in the EM algorithm" In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003.

[28]M. Stevenson, E. Agirre, and A. Soroa, "Exploiting domain information for Word Sense Disambiguation of Medical documents", In Journal of the American Medical Informatics Association, Mar-Apr Vol. 19 Issue 2 pp. 235-240, 2012.

[29]J. Traupman, and R. Wilesky, "Experiments in Improving Unsupervised Word Sense Disambiguation" In Report in UCB /CSD -03-1227 Berkley California, pp. 1-39, 2003.

[30]G. Tsatsaronis, M. Vazirgiannis, and I. Androoutsopoulos, "Word Sense Disambiguation with Spreading Activation Networks from Thesauri," In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI) pp. 1725-1730, 2005.

[31]J. Veronis, "Hyperlex: Lexical cartography for information retrieval" In Journal Computer Speech Language Vol. 18, No. 3, pp. 223–252, 2004.

[32]N. Yadav, Y. Gupta, M. Kumar, and R. Sanyal, "Semantic Document Classification and Keyword Spotting in Digital Repositories" In The International ACM Conference on Management of Emergent Digital EcoSystems (MEDES 09), Lyon/France, pp. 157-161, 2009.

[33]D. Yaroswsky, "Unsupervised Word Senses Disambiguation Rivaling Supervised Methods," In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (Cambridge, MA) pp. 189-196, 1995.

[34]M. Breithaupt, "A fuzzy Word Sense Disambiguation Algorithm using Wordnet" M S Thesis, Georgia State University GA US, 2010.

[35]M. Lesk, "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 (SIGDOC '86) New York, NY, USA pp 24-26, 1986.

[36]A. Kilgarriff, and J. Rosenzweig, "English SENSEVAL: Report and Results" In Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC, Athens, Greece, 2000.