Word Sense Disambiguation in Hindi Language Using Hyperspace Analogue to Language and Fuzzy C-Means Clustering

Devendra K. Tayal
Associate Professor
IGDTUW,
Delhi-110006

Leena Ahuja
Ex-Student
IGDTUW,
Delhi-110006

Shreya Chhabra
Ex-Student
IGDTUW,
Delhi-110006

Abstract

The problem of Word Sense Disambiguation (WSD) can be defined as the task of assigning the most appropriate sense to the polysemous word within a given context. Many supervised, unsupervised and semi-supervised approaches have been devised to deal with this problem, particularly, for the English language. However, this is not the case for Hindi language, where not much work has been done. In this paper, a new approach has been developed to perform disambiguation in Hindi language. For training the system, the text in Hindi language is converted into Hyperspace Analogue to Language (HAL) vectors, thereby, mapping each word into a high-dimensional space. We also deal with the fuzziness involved in disambiguation of words. We apply Fuzzy C-Means Clustering algorithm to form clusters denoting the various contexts in which the polysemous word may occur. The test data is then mapped into the high dimensional space created during the training phase. We test our approach on the corpus created using Hindi news articles and Wikipedia. We compare our approach with other significant approaches available in the literature and the experimental results indicate that our approach outperforms all the previous works done for Hindi Language.

1. Introduction:

Words in a language may carry more than one sense. Human beings can easily decipher the context in which the word is being used in a sentence. However, the same cannot be said for the machines. Various applications like speech processing, text processing, search engines, etc, in order to function properly, need to figure out the sense of the word. Thus there is a need for word sense disambiguation for correctly interpreting the meaning of a sentence written in natural language.

Given a word and its possible senses, as defined in a knowledge base, the problem of Word Sense Disambiguation (WSD) can be defined as the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings. It was first formulated as a distinct computational task during the early days of machine translation in the 1940s, making it one of the oldest problems in computational linguistics.

Warren Weaver (1949), in his famous 1949 memorandum on translation, first introduced the problem in a computational context. Early researchers understood the significance and difficulty of WSD well. Since then there have been various approaches for handling the problem of Word Sense Disambiguation. Majorly, WSD can be done using Knowledge Based Approaches (Navigli, 2009), Machine Based Approaches (Navigli, 2009) and Hybrid Approach (Navigli, 2009). Knowledge-based methods rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence. Lesk Algorithm (1986) is a classical approach based on Knowledge bases.

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 49–58,
Trivandrum, India. December 2015. ©2015 NLP Association of India (NLPAI)

Semi-supervised or minimally supervised methods make use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process, or a word-aligned bilingual corpus. Yarowsky Algorithm (1995) is based on this approach. Supervised methods make use of sense-annotated corpora to train from while unsupervised methods (Schütze, 1998) are based on the assumption that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context.

Although much work is available for WSD in English language, but for Hindi language very few research works have been contributed.

Sinha, Reddy and Bhattacharya (2012) did a statistical approach towards Word Sense Disambiguation in Hindi. In their work, a set of context words are selected using the surrounding window and for each sense *w* of a word, a semantic bag is created by referring to the Hindi WordNet® (hypernymy, hyponymy and meronymy). They claim that sense of the word that has the maximum overlap between the context bag and the semantic bag is the correct sense. But, their system does not detect the underlying similarity in presence of morphological variations. Kumari and Singh (2013) used genetic algorithm to perform word sense disambiguation on Hindi nouns. Genetic algorithm is a heuristic search algorithm used to find approximate solutions to optimization and search problems using techniques inspired by evolutionary biology. But in their work, the recall values associated with the algorithm depend upon the genetic parameters chosen for evaluation and therefore is not universally applicable. Yadav and Vishwakarma (2013) use association rules to first mine the itemsets depending upon the context of the ambiguous word and then mine the association rule corresponding to the most frequent itemset. Tomar *et al*. (2013) use the technique of PLSA for making '*k*' clusters representing the senses or the different contexts of the word. The clusters are then further enriched using the Hindi WordNet®. The cluster with which the maximum similarity score (cosine distance) is obtained gives the correct sense of the ambiguous word. The performance of their work varies linearly with the amount of training data used. In (2013), Jain, Yadav and Tayal used graph based approach for Word Sense Disambiguation in Hindi Text. Here, to construct a graph for the sentence each sense of the ambiguous word is taken as a source node and all the paths which connect the sense to other words present in the sentence are added. The importance of nodes in the constructed graph is identified using node neighbor based measures and graph clustering based measures. This method disambiguates all open class words and disambiguates all the words present in the sentence simultaneously.

In (2005) , Hao Chen, Tingting He, Donghong Ji and Changqin Quan used an unsupervised approach, for disambiguating words in Chinese Language, where contexts that include ambiguous words are converted into vectors by means of a second-order context method, and these context vectors are then clustered by the k-means clustering algorithm and lastly, the ambiguous words can be disambiguated after a similarity calculation process is completed. But, the K-means clustering limits the word to belong to only one cluster and hence also limits the accuracy of Word Sense Disambiguation. After going through the literature survey, we found that all the methodologies for WSD used till date do not take into account the fuzziness involved in disambiguation of the words.

In this paper, we develop an approach whereby we first train our system taking Hindi newspaper and Wikipedia articles as input and use Hyperspace Analogue to Language model to convert Hindi words into vectors, representing points in the high dimensional Hyperspace.

Fuzzy C-Means Clustering is then applied to get the clusters where each word may belong to more than one cluster with an associated membership value. The words belonging to one context are grouped together in a cluster. The polysemous words belong to more than one cluster, each cluster corresponding to the possible sense of that word. Once the training of the system is complete, the test data containing the polysemous word is processed using Hyperspace Analogue to Language (HAL) model to map the polysemous word of the test data as a point in the hyperspace created in the training phase. Then, Euclidean Distance is calculated between this point and all the cluster centers and the nearest cluster corresponds to the sense of the polysemous word used in the test data. And similar computation can be easily performed for each polysemous word of the test data in order to determine the correct sense of that word.

We tested our approach using Netbeans IDE to develop a Hyperspace Analogue to Language (HAL) Model and MATLAB for construction of fuzzy clusters and finding the nearest cluster. The results obtained were compared with all of the previous approaches and our approach shows the best results.

The remainder of the paper is organized as follows: Section 2 provides a review of Hyperspace Analogue to Language, Fuzzy Logic and Fuzzy C-Means Clustering Algorithm. Section 3 describes the specifics of the proposed algorithm to carry out word sense disambiguation. Section 4 illustrates the application of the proposed approach on a small data set. Section 5 describes the experimental results obtained for our approach and the compares it with already existing techniques for Word Sense Disambiguation for Hindi language. Finally, the last section concludes the paper, and makes some suggestions for future work.

2 Preliminaries:

2.1 Hyperspace Analogue to Language (HAL):

Hyperspace Analogue to Language is a numeric method developed by Kevin Lund and Curt Burgess (1996) for analyzing text. It does so by running a sliding window of fixed length across a text, calculating a matrix of word co-occurrence values along the way. The basic premise that the work relies on is that words with similar meanings repeatedly occur closely (also known as co-occurrence).

Given a N word vocabulary, the HAL space is a N × N matrix constructed by moving a window of length l over the corpus by one word increments. Given two words W1 and W2, whose distance within the window is d, the weight of association between them is computed by $l - d + 1$. After traversing the corpus, an accumulated co-occurrence matrix for all the words in a target vocabulary is produced. HAL is direction sensitive: the co-occurrence information for words preceding each word and co-occurrence information for words following each word are recorded separately by row and column vectors.

2.1.1 Illustration:

Consider the following piece of Hindi Text:

“भारत की राजधानी दिल्ली है। दिल्ली मे अनेक वर्ग के लोग है।”

HAL Matrix constructed is as follows:

	भारत	राजधानी	दिल्ली	वर्ग	लोग
भारत	-	-	-	-	-
राजधानी	4	-	-	-	-
दिल्ली	4	8	-	-	-
वर्ग	-	-	4	-	-
लोग	-	-	1	4	-

Taking a window of size 6, we consider all the words that are lying before and after the focus word but within the context window. For example, consider the word “““लोग”. The distance between “वर्ग” and “लोग” in the sentence given above is 3. So the value of the cell(लोग, वर्ग) = 6-3+1=4.

2.2 Fuzzy Logic:

Fuzzy logic (2005) is a form of many-valued logic; it deals with reasoning that is approximate rather than fixed and exact. Compared to traditional binary sets (where variables may take on true or false values), fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false (1999). The term fuzzy logic was introduced with the 1965 proposal of fuzzy set theory by Lotfi A. Zadeh (1965). Fuzzy logic has been applied to many fields, from control theory to artificial intelligence.

2.3 Fuzzy C-Means Clustering:

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. Some examples of measures that can be used as in clustering include distance, connectivity, and intensity.

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. Thus, points on the edge of a cluster, may be *in the cluster* to a lesser degree than points in the center of cluster. In (Ross, 2004), Fuzzy C-Means (FCM) is described as a method of clustering which allows one piece of data to belong to two or more clusters. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \,||\, \mathrm{x_i} - \mathrm{c_j}||^2 \qquad \text{………1}$$

where *m* is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i^{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{||\mathrm{x_i}-\mathrm{c_j}||}{||\mathrm{x_i}-\mathrm{c_k}||}\right)^{\frac{2}{m-1}}} \qquad \text{……......2}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \qquad \text{……......3}$$

3. Proposed Approach:

The disambiguation process can be divided into two phases- Training Phase and Testing Phase.

3.1 Training Phase

The Training Phase involves the use of Hyperspace Analogue to Language model and Fuzzy C-Means Clustering Algorithm to cluster the words in the Q-dimensional space where Q is the number of significant words in the training document. Each cluster denotes the context in which the ambiguous word may occur. Here, a word can belong to more than one cluster, and associated with each word is a set of membership levels. More the membership value, more the word belongs to that cluster. The flow chart of the Training Phase is shown below:

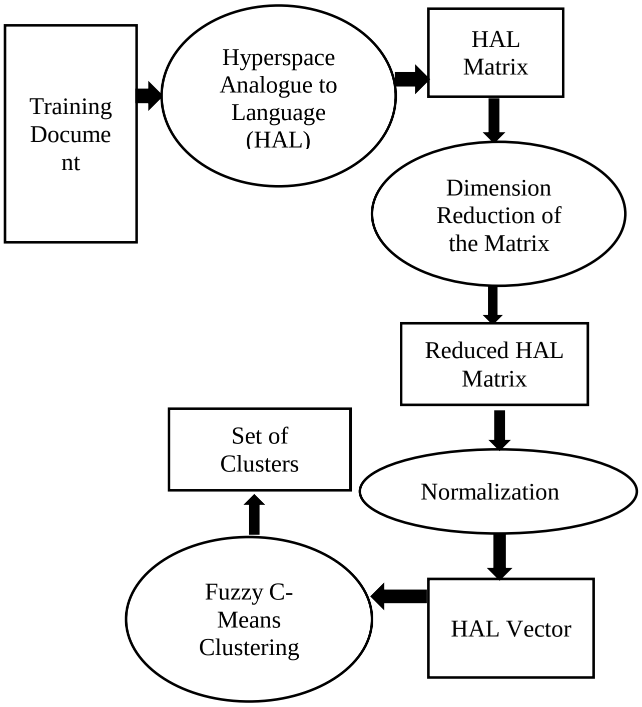


Figure 1: Training Phase

3.1.1 Training Document:

A set of documents were selected for the purpose of collecting words and then building the clusters of those words based on their co-occurrences.

Documents used for training included articles from Hindi newspapers, magazines and Wikipedia.

3.1.2 Hyperspace Analogue to Language:

In this phase, the training documents are processed as per the Hyperspace Analogue to Language Model. The algorithm for HAL (Lund and Burgess, 1998) calculates HAL matrix containing entries corresponding to all the unique words occurred during the training of the system. The matrix so generated is a N × N matrix where N is the total number of unique words in the training document set.

3.1.3 Dimension Reduction of the Matrix:

As mentioned previously, the training document is a collection of the words, taken from a source such as news articles, Wikipedia and magazines. Some words are significant during the process of disambiguation while the other words that are not significant with respect to disambiguation process are called as the stop words like की है, की,और,यहाँ,पर,को,था,यह,थे,का,जिसकी,हुई,थी etc. Since the words are not significant with respect to our disambiguation process, we can remove the rows and columns corresponding to these words from Hyperspace Analogue to Language (HAL) Matrix that helps in reducing the dimensionality of the Hyperspace Analogue to Language (HAL) space and further mathematical processing of the stop words. This reduction process reduces the N × N dimensional HAL matrix into Q × Q dimensional HAL matrix where Q is the number of significant words in the training document which are used for the disambiguation process.

3.1.4 Normalization:

The reduced HAL matrix is then normalized to get the HAL vector corresponding to each significant word.

In order to represent a Word as a Point in Q dimensional space, we consider each word as a concept:

Concept C = < W_{cp_1}, W_{cp_2}, W_{cp_3}, …… W_{cp_Q}>

where p_1, p_2... p_n are called dimensions of C, each corresponding to a unique word that forms a dimension of the hyperspace and W_{cp_i}, denotes the weight of p_i in the vector representation of C. In order to calculate weights, we normalize the values in the HAL matrix.

Normalization Formula:

$$W_{c_ip_j} = \frac{W_{c_ip_j}+W_{c_jp_i}}{\sqrt{\sum_{k=1}^{Q} W_{c_ip_k}{}^2+W_{c_kp_i}{}^2}} \quad ………4$$

Since we need to consider the correlation between two given words irrespective of the order in which those words appeared, we consider both $W_{c_ip_j}$ and$W_{c_jp_i}$.

Therefore, by constructing a HAL space from training document, concepts are represented as weighted vectors in the high dimensional space, whereby each word in the vocabulary of the corpus gives rise to an axis in the corresponding semantic space.

3.1.5 Fuzzy C-Means Clustering

The HAL vectors for Q words generated in the previous step represent Q points in the Q-dimensional space. These vectors are then fed in as the input to the module implementing Fuzzy C-Means Clustering algorithm. The output of the Fuzzy Clustering module is a set of fuzzy clusters each representing a context of the ambiguous word. As is the nature of fuzzy clusters, a word may belong to one or more clusters with different membership values.

3.2 Testing Phase

This phase describes the disambiguation of the senses of the target ambiguous word in the test data. The flow diagram shown below describes the process:

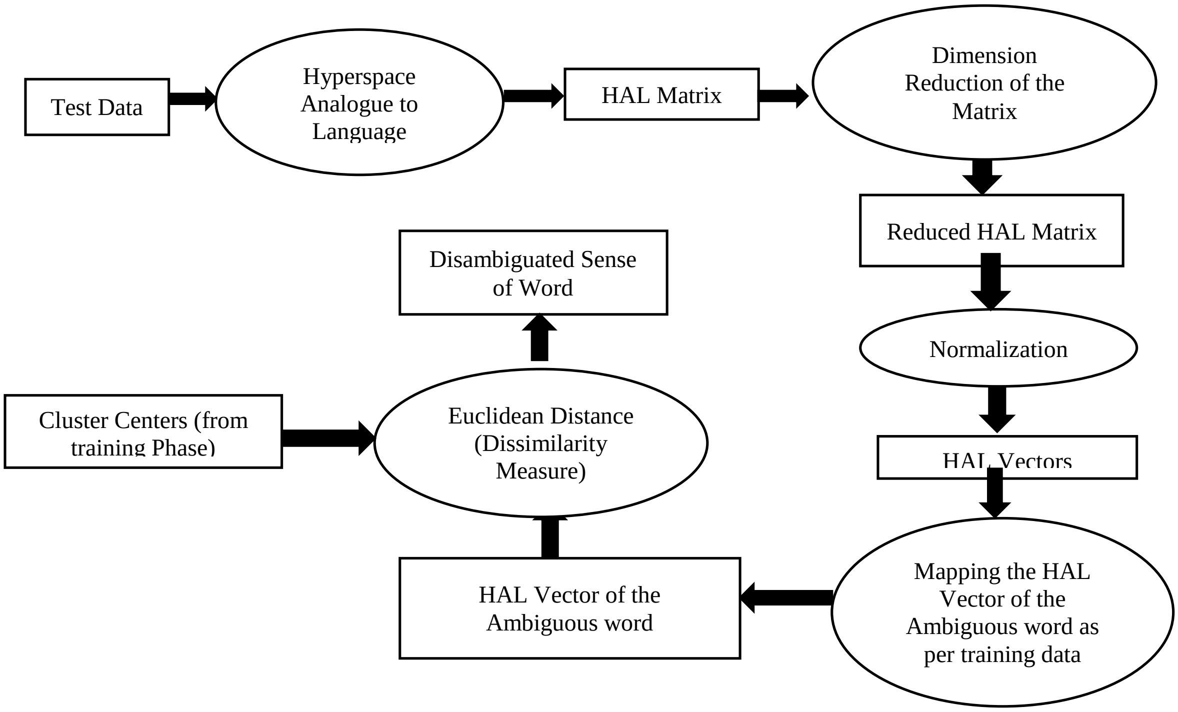

Test Data
Hyperspace Analogue to Language
HAL Matrix
Dimension Reduction of the Matrix
Reduced HAL Matrix
Normalization
HAL Vectors
Mapping the HAL Vector of the Ambiguous word as per training data
HAL Vector of the Ambiguous word
Euclidean Distance (Dissimilarity Measure)
Cluster Centers (from training Phase)
Disambiguated Sense of Word

Figure 2 Testing Phase

Given the test data, we apply HAL Model to get the M × M HAL Matrix where M is the number of unique words in the test data. The rows and columns in the HAL matrix correspond to the word in the test data. Dimension reduction and Normalization process is applied in the manner similar to the training phase to get the HAL vectors of the words in the test data.

In order to apply dissimilarity measure for disambiguation process in the Q-dimensional Hyperspace, we need to map the M-dimensional HAL vector of the target ambiguous word to the Q-dimensional vector in the Hyperspace generated in the training phase.

The mapping process involves initializing the Q-dimensional HAL vector of the ambiguous word to all zeros and then finding the common words in the training document and the test data. The weights corresponding to the common words are extracted from the M-dimensional HAL vector of the test data and then these weights are substituted in the Q-dimensional HAL vector of the test data with respect to the indexes of the words generated in the training phase. Hence, we get the Q-dimensional HAL vector of the target ambiguous word with respect to the test data containing weights for common words and zeros for the other dimensions.

In order to disambiguate the correct sense of the target ambiguous word, Euclidean distance between target word's HAL vector and centers of the clusters generated in the previous phase are calculated one by one as follows:

$$d(p,q) = d(q,p) \\ = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_Q - p_Q)^2} \\ = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

…………..5

The one with the minimum distance is then chosen to be the most related cluster and hence that corresponds to the most related sense.

4. Illustrative Example

The following example illustrates the proposed technique for disambiguation for the polysemous word '**तीर**'.

The word '**तीर**' can have different senses as follows:

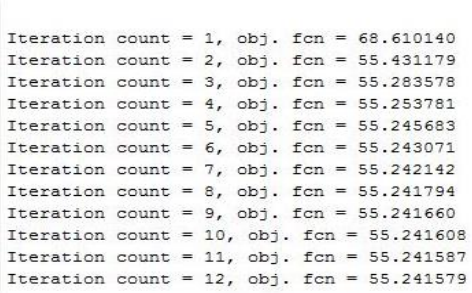
- धातु आदि का बना वह पतला लम्बा हथियार जो धनुष द्वारा चलाया जाता है (Arrow)
- नदी या जलाशय का किनारा (Bank of River)

The training data for the above mentioned ambiguous word can be found in the Appendix.

Training Data consists of total 91 words. After removing the stop words we get 63 base words:

धनुष,प्रयुक्त,वाला,अस्त्र,तीर,अग्र,भाग,नुकीला,सर्वप्रथम,उल्लेख
,ऋग्वेद,संहिता,मिलता,इषुकृत्,इषुकार,सिद्ध,दिनों,निर्माण-कार्य,व्यवस्थित,व्यवसाय,ऋग्वेदकालीन,लोहार,केवल,लोहे,काम
,तैयार,बनाता,शेष,
बाण,निर्मातानिकाय,भीषण,बाढ़,भूस्खलन,तबाही,सबक,उत्तरा
खंड,सरकार,राज्य,नदियों,नई,इमारतें,
बनाने,पूरी,रोक,लगा,दी,आपदा,प्रभावित,लोगों,वित्तीय,मदद,देने
,उबरने,अल्पकालिक,दीर्घकालिक,कदम,उठाने,वास्ते,
पुनर्निर्माण,पुनर्वास, प्राधिकरण, फैसला .

Using these base words, we form a HAL matrix of dimensions 63 × 63 using the window of size 10. Normalization is carried out over the HAL Matrix to get the HAL vectors corresponding to each unique word. This was developed in the code written in Java language using the NetBeans Platform. Fuzzy C-Means Clustering is then applied to the generated HAL Vectors. Two clusters have been generated using code written in the Matlab. The following snippet shows the minimization of the objective function (using Equation 1) in the process of generating the clusters.



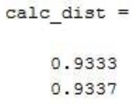
```
Iteration count = 1, obj. fcn = 68.610140
Iteration count = 2, obj. fcn = 55.431179
Iteration count = 3, obj. fcn = 55.283578
Iteration count = 4, obj. fcn = 55.253781
Iteration count = 5, obj. fcn = 55.245683
Iteration count = 6, obj. fcn = 55.243071
Iteration count = 7, obj. fcn = 55.242142
Iteration count = 8, obj. fcn = 55.241794
Iteration count = 9, obj. fcn = 55.241660
Iteration count = 10, obj. fcn = 55.241608
Iteration count = 11, obj. fcn = 55.241587
Iteration count = 12, obj. fcn = 55.241579
```

Figure 3: Minimization of the objective function in fuzzy clustering

Once we get the clusters, we consider the test data.

Refer to Appendix for test data used in this case.

The HAL vector (63 dimensional) is then obtained for the ambiguous word तीर. Euclidean distance between the Test Vector and the two cluster centers are calculated. The following snapshot shows the two values.



```
calc_dist =

    0.9333
    0.9337
```

Cluster 2 is hence obtained as the target sense of the ambiguous word which means that the word 'तीर' here is correctly disambiguated as 'Arrow'.

5. Results and Discussions

Since there is no standard corpus available for Hindi language, we created our own corpus by selecting relevant articles from Hindi Wikipedia as well as Hindi language newspapers viz Dainik Jagran, Nav Bharat Times etc.

The training data used consists of 3753 words in total. A collection of polysemous words was made and training data was collected depicting the different contexts in which the word was used. The formula for accuracy is given as follows:

$$Accuracy = \frac{Number\ of\ correctly\ disambiguated\ words}{Total\ number\ of\ ambiguous\ occurrences}$$

When the proposed technique for disambiguation of Hindi Text was applied on the corpus, we found an efficiency of nearly 79.16%. Our technique, therefore, performs better than all the previously used approaches used for performing word sense disambiguation in Hindi language. It can be noted that all the methodologies used till date do not take into account the fuzziness involved in disambiguation of the words. In this paper, we use the concept of Fuzzy C-Means Clustering to overcome this major drawback of the previous approaches.

The following table shows the comparative efficiency of our technique with the previous techniques available in the literature that have been used in Word Sense Disambiguation in the Hindi language with their comparative accuracies. Thus, we conclude that the results obtained from our approach are better than all the other approaches that currently exist in the Hindi language and this is what makes it more promising.

S.No.	Technique Used	Author	Year	Accuracy
1.	Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation	Gaurav S Tomar, Manmeet Singh, Shishir Rai, Atul Kumar, Ratna Sanyal, Sudip Sanyal[2]	2013	74.12%
2.	Lesk Algorithm	Manish Sinha ,Mahesh Kumar, Reddy .R ,Pushpak Bhattacharyya , Prabhakar Pandey ,Laxmi Kashyap [9]	2012	Varies from 40-70%
3.	Association Rules	Preeti Yadav, Sandeep Vishwakarma[13]	2013	72%
4.	A Graph Based Approach to Word Sense Disambiguation for Hindi Language	Sandeep Kumar Vishwakarma, Chanchal Kumar Vishwakarma[16]	2013	65.17%

6. Conclusion:

In this work, we have proposed a new approach for Hindi Word Sense Disambiguation which incorporates fuzzy measures. We found that unlike the previous unsupervised approaches which discard contexts into which an ambiguous word may fall, the current approach retains the fuzziness associated with the ambiguity, thereby, giving better results. Moreover, the technique proposed by us is not language specific; it can be extended to other languages as well. As is evident from the results obtained by the use of HAL that the context knowledge in context specific WSD is very important and that world knowledge from the surrounding words plays a very important role in revealing the actual sense of the word in the given context.

Thus, the above mentioned advantages make our approach particularly promising. This approach can be extended in the future to include semantic relations from sources like Wikipedia and Wordnet in the enrichment of clusters that will lead to better accuracy for disambiguating a polysemous word.

References:

1. Cao G, Song D and Bruza PD "Fuzzy C-means clustering on a high dimensional semantic space" In Proceedings of the 6th Asia Pacific Web Conference (APWeb'04),LNCS 3007, 2004, pp. 907-911.

2. Gaurav S Tomar, Manmeet Singh, Shishir Rai, Atul Kumar, Ratna Sanyal, Sudip Sanyal "Probabilistic Latent Semantic Analysis for Unsupervised Word Sense Disambiguation", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, September 2013

3. Hao Chen, Tingting He, Donghong Ji and Changqin Quan, "An Unsupervised Approach to Chinese Word Sense Disambigua-tion Based on Hownet",Computational Linguistics and Chinese Language Processing Vol. 10, No. 4, December 2005, pp. 473-482 473

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

5. Jain, A; Yadav, S. ; Tayal, D. , "Measuring context-meaning for open class words in Hindi language" Contemporary Computing (IC3), 2013 Sixth International Conference, 2013, 118 - 123

6. Kwang H. Lee , "First Course on Fuzzy Theory and Applications",Springer-Verlag Berlin Heidelberg 2005

7. Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, pages 24-26, New York, NY, USA. ACM.

8. Lund and Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence" Behavior Research Methods, Instruments, & Computers 1996, 28 (2), 203-208

9. Manish Sinha, Mahesh Kumar Reddy, R Pushpak Bhattacharyya ,Prabhakar Pandey, Laxmi Kashyap " Hindi Word Sense Disambiguation" International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 2,2012

10. Michael Sussna," Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network" Proceedings of the second conference on Information and knowledge management, New York, 1993, 67-74

11. Navigli, R..Word sense disambiguation: A survey. ACM Comput. Surv. 41, 2, Article 10,February 2009

12. Novák, V., Perfilieva, I. and Močkoř, J. "*Mathematical principles of fuzzy logic* "Dodrecht: Kluwer Academic, 1999

13. Preeti Yadav, Sandeep Vishwakarma "Mining Association Rules Based Approach to Word SenseDisambiguation for Hindi Language" International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 5, May 2013

14. Rada Mihalcea, Using Wikipedia for Automatic Word Sense Disambiguation, in Proceedings of the North American Chapter of

the Association for Computational Linguistics (NAACL 2007), Rochester, April 2007

15. Sabnam Kumari "Optimized Word Sense Disambiguation in Hindi using Genetic Algorithm" International Journal of Re-search in Computer andCommunication Technology, Vol 2, Issue 7, July-2013

16. Sandeep Kumar Vishwakarma, Chanchal Kumar Vishwakarma, "A Graph Based Approach to Word Sense Disambiguation for Hindi Language", International Journal of Scientific Research Engineering & Technology (IJSRET) Volume 1 Issue5, August 2012, 313-318.

17. Schütze, H. "Automatic word sense discrimination" Computational Linguistics, 1998, 97–123.

18. Shaul Markovitch, Ariel Raviv, "Concept-Based Approach to Word-Sense Disambiguation" In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 807-813 Toronto, Canada, 2012.

19. Song D and Bruza PD "Discovering information flow using a high dimensional conceptual space" Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01), 2001,327-333.

20. Timothy J.Ross, Fuzzy Logic with Engineering Applications, 2004, John Wiley & Sons Ltd PP:369-386

21. Weaver, Warren "Translation" Locke, W.N.; Booth, A.D. Machine Translation of Languages: Fourteen Essays. Cambridge, MA: MIT Press , 1949

22. Yarowsky "Unsupervised word sense disambiguation rivaling supervised methods" Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995

23. Zadeh, L.A. "Fuzzy sets", *Information and Control* Volume 8, Issue 3, 338–353, June 1965

Appendix:

Training Data: "धनुष के साथ प्रयुक्त होने वाला एक अस्त्र तीर है जिसका अग्र भाग नुकीला होता है| तीर का सर्वप्रथम उल्लेख ऋग्वेद संहिता में मिलता है| इषुकृत् और इषुकार शब्दों का प्रयोग सिद्ध करता है कि उन दिनों तीर निर्माण-कार्य व्यवस्थित व्यवसाय था ऋग्वेदकालीन लोहार केवल लोहे का काम ही नहीं करता था, तीर भी तैयार करता था| तीर का अग्र भाग लोहार बनाता था और शेष बाण-निर्मातानिकाय बनाता था| "

Translation: Arrow (तीर) is a pointed instrument used with the bow. Tracing the history, its first mention was in 'Rigved'.The use of words 'ishukrit' and 'ishukar' signifies that the people in those days were involved in the business of Arrow(तीर(making. The pointed front part of the arrow(तीर) was manufactured by Ironsmith while the rest was made by other specialized manufacturer.

"भीषण बाढ़ और भूस्खलन से तबाही से सबक लेते हुए उत्तराखंड सरकार ने राज्य में नदियों के तीर नई इमारतें बनाने पर पूरी तरह रोक लगा दी है| सरकार ने आपदा प्रभावित लोगों को वित्तीय मदद देने और आपदा से उबरने के लिए सभी अल्पकालिक एवं दीर्घकालिक कदम उठाने के वास्ते पुनर्निर्माण एवं पुनर्वास प्राधिकरण बनाने का भी फैसला किया है|

Translation: The Government of Uttarakhand has banned the construction of new buildings due to heavy floods and landslide in the region. The Government has also decided to take all kinds of steps that includes providing financial help to disaster prone people to overcome the big loss.

Test Data :अग्निपुराण में तीर के निर्माण का वर्णन है | यह लोहे या बाँस से बनता है | बाँस सोने के रंग का और उत्तम कोटि के रेशोंवाला होना चाहिए | तीर के पुच्छभाग पर पंख होते हैं | उसपर तेल लगा रहना चाहिए, ताकि उपयोग में सुविधा हो | इसकी नोक पर स्वर्ण भी जड़ा होता है | प्राचीन काल मे युद्ध के समय धनुष से तीर चलाकर शत्रु का वध किया जाता था

Translation: The art of arrow (तीर) making is described in 'Agnipuran'. It mentions that the arrow(तीर) is made up of either iron or the wood. The material used should be a good quality golden coloured wood. Feather like structure is attached to the end of the arrow(तीर). Proper oiling of the arrow should be done to enable ease of use. The pointed front end of the arrow also has small amount of gold attached to it. In past, Bow and arrow were used during war times to kill the enemy