Word Sense Disambiguation for Indian Language

Submitted By

Shivang B. Popat

14MCEN14

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Word Sense Disambiguation for Indian Language

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Shivang B. Popat

(14MCEN14)

Guided By

Prof. Tarjni Vyas

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2016

Certificate

This is to certify that the major project entitled **"Word Sense Disambiguation for Indian Language"** submitted by **Shivang B. Popat (Roll No: 14MCECN14)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Institute of Technology, Nirma University Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Tarjni Vyas
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. P. N. Tekwani
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Shivang B. Popat**, Roll. No. **14MCEN14**, give undertaking that the Major Project entitled "**Word Sense Disambiguation for Indian Language**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student
Date: 16^{th} May,2016
Place: Ahmedabad

Endorsed by
Tarjni Vyas
(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Tarjni Vyas**, Associate Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. P. N. Tekwani**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

See that you acknowledge each one who have helped you in the project directly or indirectly.

- Shivang B. Popat

14MCEN14

Abstract

Word sense disambiguation is a part of Natural Language Processing (NLP)and technique to computationally find the correct sense of the ambiguous words using the context in which they occur. It is used in many application like Machine translation, Information Retrieval, Information Extraction,etc. There are many approaches for WSD like supervised, unsupervised,etc. In this thesis all the approaches are discussed and Graph Based Unsupervised approach is applied with global measures for Indian Language using Hindi Word-net. Motive is to find more accuracy and Easy adaptability of WSD for Indian Language .

Contents

List of Figures

Chapter 1

Introduction

1.1 Natural language processing

Natural language process is a big step forward towards field of artificial intelligence and machine learning.Because of NLP he goal of more user-friendly machines and computers can be achieved since people can talk to their machines in a free manner. [2].

Natural Language processing is the ability to process human language. Humans communicate in the most natural ways through the language and since storage .transformation and analysis of documents which are written in natural language are done by machines lately,NLP is necessary[3][4].

As shown in figure,for NLP first,morphological processing of the documents is done. Later, this processed documents are parsed for lexical analysis which includes grammar and lexicon analysis.Then,in semantic Analysis semantic rules are applied to the documents to remove the ambiguity from the sentences. pragmatic analysis only done if needed and it uses the contextual information to find the accurate sense. At last documents represented in proper form[5][6].

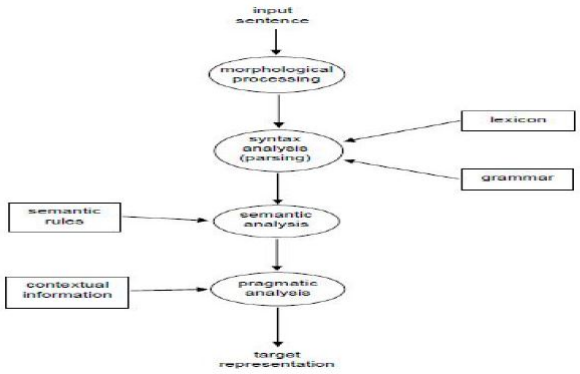

Figure 1.1: The logical steps in Natural language Processing

1.2 WordNet

Wordnet is Lexical Database.Wordnet is created and maintained at Princeton University and currently version 3.0 is available.it contains 155,000 words with approximate 117,000 synsets. example is Automobile with its synsets are (recall superscript and subscript denote the words sense identifier and part-of-speech tag, respectively) [2] : car, auto, automobile, machine, motorcar

Word-net has multiple synonym of words so they can be used to removed ambiguity in words. Verbs,nouns,adjectives all are in wordnet.

Wordnet is lexical resource which contains verbs,adverbs,adjectives into sets which are called synsets.this synsets are also interlinked to form a network which can be seen by a browser . Wordnets for different languages are freely and publicly available.wordnet is very much useful in WSD . Wordnet and thesaurus are almost similar.[2]

The code stucture of wordnwt is :

- 01 to 04 pos(part of speech noun / adjective / verb /adverb)
- 12 synsets(no of synsets)
- 0026 types of relation(hypernymy,holonymy,meronymy)

- 1102 hypernymy(ontology)(example, pigeon, crow, eagle and seagull are all hyponyms of bird (their hypernym))
- 1103 - hyponymy
- 1105 - attribute
- 1107 ability verb
- 1140 meronymy (example, "finger" is a meronym of "hand")
- 1150 Holonymy(example, "hand" is a holonym of "finger")

1.3 Introduction of WSD

Word Sense Disambiguation is used to identify correct meaning of a word.It is the main problem in NLP and it has may application like Information Retrieval, Information Extraction, Machine Learning.WSD is method to find the proper sense of the given word. Example, word bank has many meaning in English. These kind of words with multiple senses are call polysemous words.WSD is the process of finding out the exact sense of Polysemous word. [2].

1.4 Need of research

WSD finds its many application in different areas:

- Information Extraction (IE): In particular field Information Extraction is interesting to finding between specific instances of concepts. e.g : M. Smith likes fishing. But he doesnt like biking
- Machine Translation (MT):: Using word sense disambiguation roper translation of the word can be done which can be useful in machine translation.

- Word Processing: Spelling of word can be corrected using word sense disambiguation and also special characters can be added properly.

1.5 Objective of Study

- Analyse the approaches of English language and experiment those approaches for Indian word sense disambiguation system.
- Compare the accuracy and time taken by the algorithm for Indian language.
- Improve the accuracy of the algorithm for Indian Language.
- Develop an accurate and rapid algorithm to obtain the sense of target words or all words against a sense repository like Wordnet or thesauruses according to the context in which word appears for Hindi languages.

Chapter 2

Literature Survey

In human language,ambiguity in word sense is prime concern.To remove this ambiguity many researchers have chosen many different paths and proposed many solutions for different languages.

2.1 Approaches to WSD

There are many approaches in Word Sense Disambiguation. Mainly those approaches depends on the Word Net means data set which are available or not. Based on this mainly data set available or not.They are mainly Supervised and Unsupervised Disambiguation approach. Some other Disambiguation approaches also there but mainly are those.

2.1.1 Knowledge base Approach

In this technique mostly dictionary like wordnet or thesaurus are used in back end. The main difference between wordnet and thesaurus is: thesaurus contains only synsets which defines only one type of relation where WordNet contains more than 8 relationships among words[7][6].

2.1.2 Supervised Approach

Supervised approach uses classification methods to disambiguate the sense.classifier is focused on polysemous word and uses the different classification methods to disambiguate the particular word.For training of the classifier training data is used in which a set of examples with target word are manually tagged with senses.These senses are from the sense inventory of the word net or machine readable dictionary. [2].

Decision Trees

Decision Tree is the classification technique which uses rules as a tree to partition the training data.Each node except leaf nodes represents test on the feature value , each branch gives outcome of the test and results are given at the leaf nodes. [2][8][9].

Naive Bayes

A Naive Bayes model uses simple probabilistic classification technique. . conditional probability is calculated for each sense of the context word for finding out the accurate sense.[2][10].

For example, a fruit may be an apple if it is red, circular, and about 12 cm large. These features are considered independently in calculating the probability of fruit is an apple . Also, these features are considered independently without any possible co relation between them.

2.2 Unsupervised Approach

2.2.1 Similarity based methods

Similarity-based methods uses the comparison between different senses of ambiguous word and context words.the most similar sense is the winning sense.different methods uses different definition of senses as well as uses from sentence to whole corpus as

a context words. But experimental results shows that Graph Based approaches are more accurate then similarity based methods. [1][11][12].

2.2.2 Graph based methods

There is another unsupervised approach is based on graph. In this approach, after processing the sentence find the polysemous words and then creating virtual graph on vector. [13][14]These nodes are given labeled which are provide us to compute the similarity.Then based on measures the similarity value are computed and based on that sense label will be given to words.[15] WordNet as a graph whose nodes are synsets and edges lexical and semantic relations between synsets [1][16].

Working of Graph Based Algorithm

In this algorithm there are mainly three steps:

a. Graph is developing with nodes which indicates senses and joint means edges as relations between those nodes.

b. Between two nodes, weights are given using global measures of algorithm.

c. Select senses which have the highest weights are assigned to that respective words.

Above three steps understand using the below example She drank some milk.

In above sentence there are two polysemous words drinks and milk. As our algorithm find graph for there using wordnet. For the word "drink" the graph shown in the fig. For word "milk" the graph will be created same manner. According the steps one then graph will be create step by step. In wordnet for word drink an milk there are five and four seances accordingly.

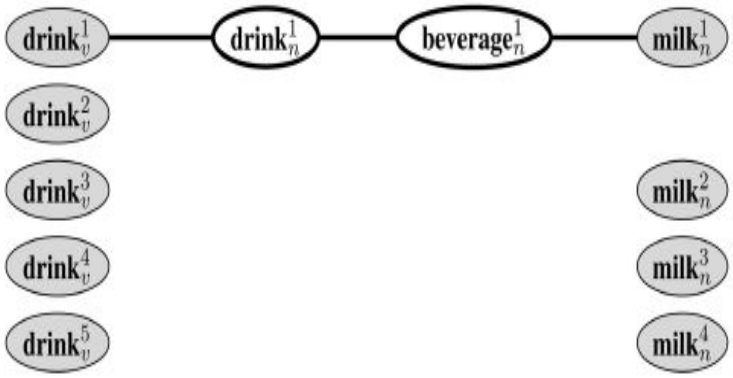

Figure 2.1: First Step Graph construction process for the sentence She drank some milk[1]

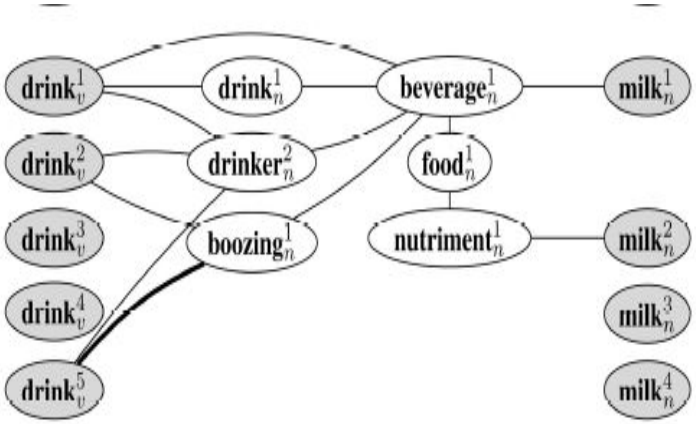

Figure 2.2: Last Step Graph construction process for the sentence She drank some milk[1]

After this, we need to find measures. For that there are mainly two connectivity measures which are [17][18]

- Local Measures
- Global Measures

Global Measures

- Compactness: When compactness is high, each vertex can be easily reached from other vertices. The measure is defined as: [1]

$$CO(G) = \frac{Max - \sum_{u \in V} \sum_{v \in V} d(u,v)}{Max - Min} \text{ [1]}$$

where $Max = K \times |v| \times (|v| - 1)$ is the maximum value the distance sum can assume (for a completely disconnected graph) and $Min = |v| \times (|v| - 1)$ the minimum value (for a fully connected graph)[1]. The compactness of the above graph is:

$$CO(G) = \frac{(5 \times 5 \times 4) - 28}{(5 \times 5 \times 4) - (5 \times 4)} = 0.90$$

Entropy: Entropy measures the amount of information (or, alternatively, uncertainty) in a random variable. In graph-theoretic terms, high entropy indicates that many vertices are equally important, whereas low entropy indicates that only a few vertices are relevant. We define a simple measure of graph entropy as

$$H(G) = -(\sum_{v \in V} p(v) \log(p(v))) \text{ [1]}$$

where the vertex probability p(v) is determined by the degree distribution:[1]

$$\left\{ \frac{deg(v)}{2 \times |E|} \right\}_{v \in V}$$

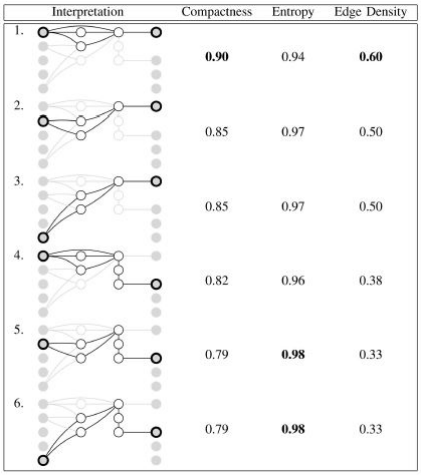

Figure 2.3: Result after processing

For above Graph the Entropy value is $P(v) = \left(\frac{3}{12}, \frac{2}{12}, \frac{4}{12}, \frac{1}{12}, \frac{2}{12}\right)$ and $H(G) = 0.94$

Edge Density: Density as a simple global connectivity measure. Edge density is calculated as the ratio of edges in a graph to the number of edges of a complete graph.

$$ED(G) = \frac{|E(G)|}{\binom{|V|}{2}} \text{ [1]}$$

For the above graph it is $ED(G) = \frac{6}{\binom{5}{2}} = 0.60$

After computing the score search which sense is for the given sentence and it is the final step of the algorithm. According sense the document will be retrieve from the number of document.

2.3 Working of Graph Based Algorithm in Block Diagram

As shown in figure, first we find the polysemous words from the documents then by using the word net we find all the senses of the wordnet .Then by applying DFS algorithm on wordnet the graph is created with nodes as senses and edges as sementic relation between them.After that, Graph based WSD algorithm is applied with either

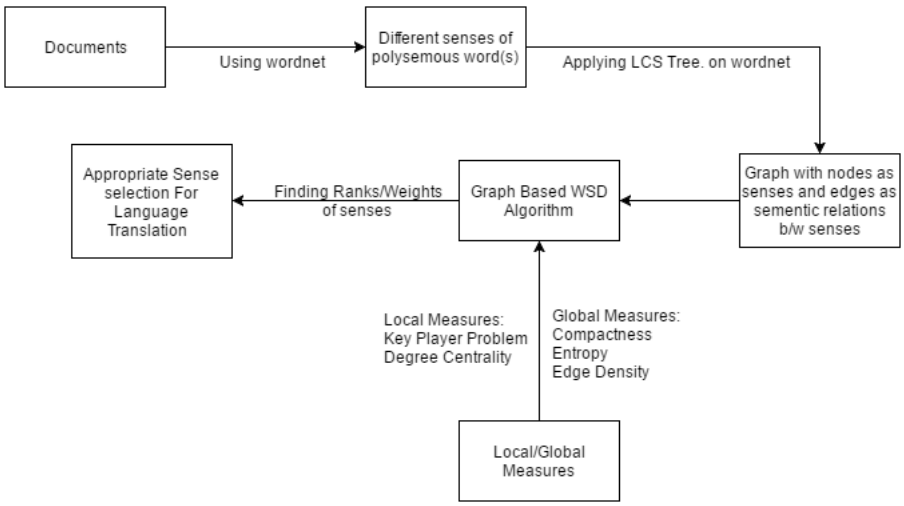

Figure 2.4: Working of Graph Based Algorithm in Block Diagram

local or global measures. At last, according to weights/ranks of senses appropriate sense is chosen for language translation.

Chapter 3

Implementation

3.1 Implementation unsupervised graph based approach in Hindi language

3.1.1 Implementation of words drink and milk in Hindi

This is the implementation of the Unsupervised Graph based approach on hindi word net.The implementation is done for the same example "she drank the milk" in Hindi. there are two polysemous words in the sentence milk and drink. As shown in screen shots,first the senses of both the words have found. Then, I have created tree for each sense of the word to its root word and compared these senses of one polysemous word with the other polysemous word. After that, I created the LCS tree by comparing different senses of word with other and then using that LCS tree we have created the graph which can be used to apply local or global measures for finding out the accurate sense of the polysemous word.

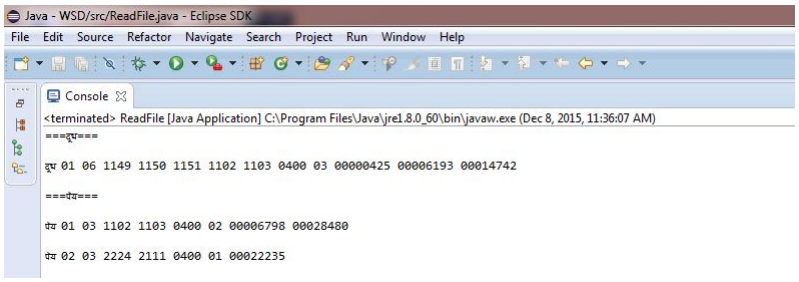

Figure 3.1: Screen shot of different relationships (hypernymy,hyponmy,antonymy,meronymy,type etc.) of both polyseomus words

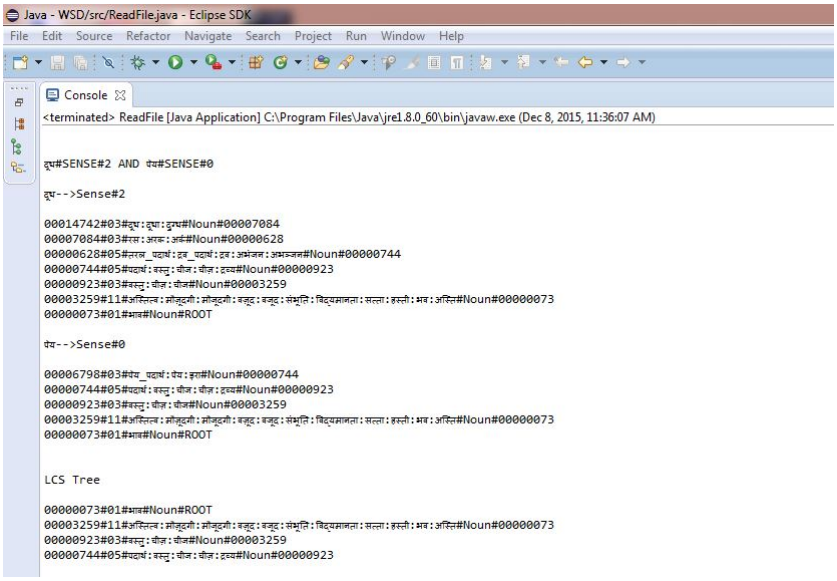

Figure 3.2: screen shot of the tree structure for sense2 of word milk and sense0 of word drink and generated LCS tree

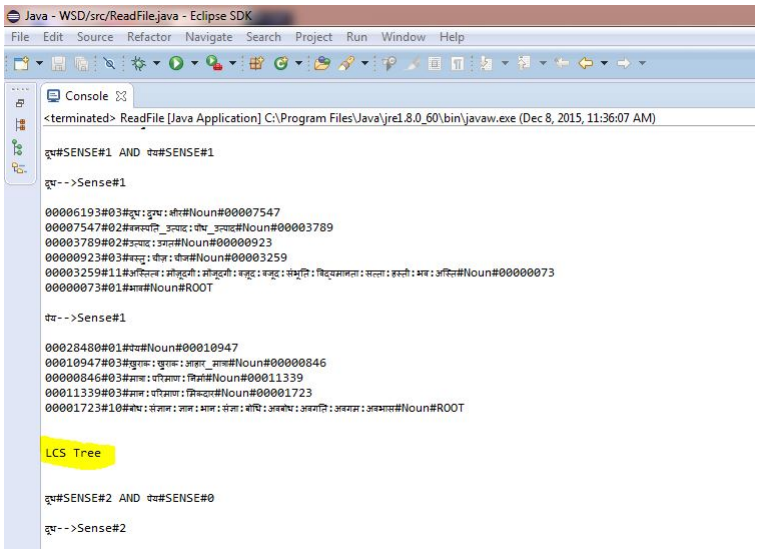

Figure 3.3: screen shot of the tree structure for sense1 of word milk and sense1 of word drink and since both senses are totally different,LCS tree is empty.

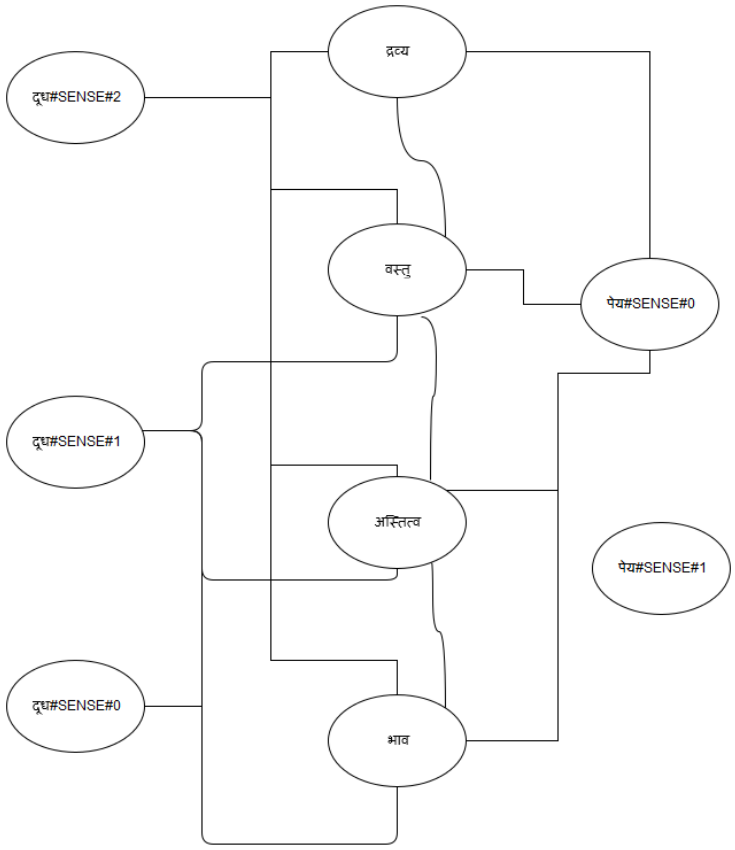

Figure 3.4: Graph for hindi words

As shown in figure, it is the Graph of words 'milk' and "drink' for hind language. Words are milk and drink with drink has 2 senses and milk has 3 senses.In graph, nodes are senses and edges are semantic relation between them. Also, drink sense2 is not connected to any other node because of empty LCS tree.

3.1.2 Graph Variation for Drink and Milk in hindi

Variations of Graphs are dependent on number of senses of Polyseomus words. There 3 sense of 'milk' and 1 sense of 'drink'. So, there are 3 different Graph variation for the given Graph.

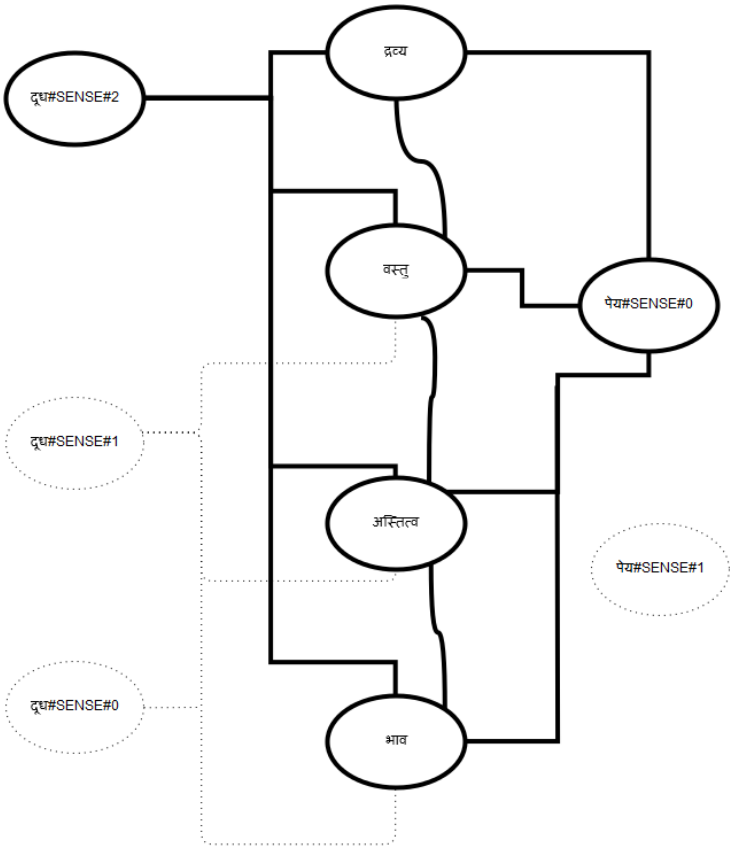

Figure 3.5: Sub graph for milk Sense2 and drink Sense0 with total 11 edges and 6 nodes.

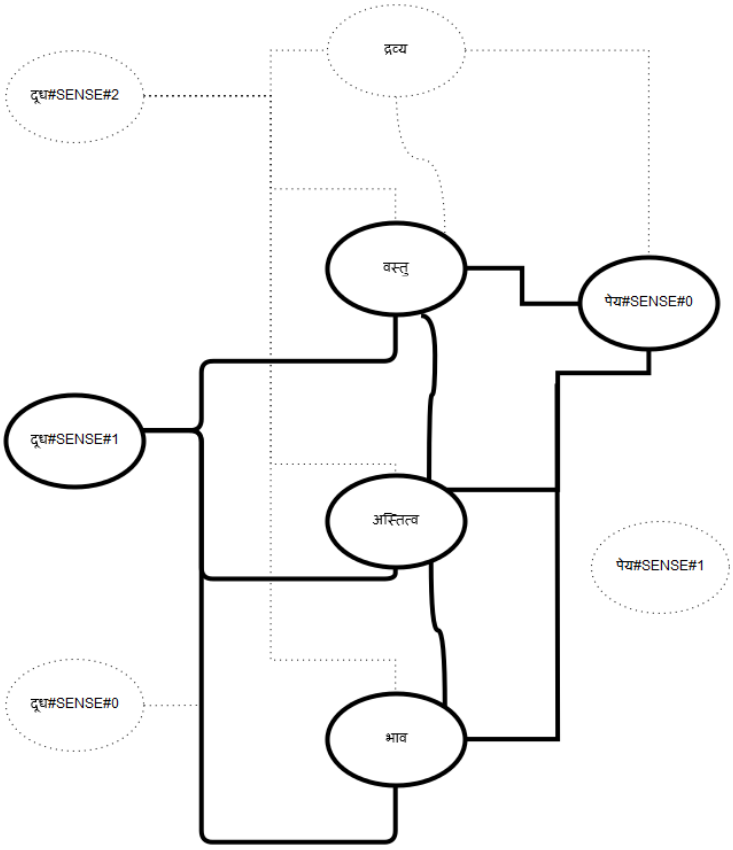

Figure 3.6: Sub graph 1 for milk Sense1 and drink Sense0 with total 9 edges and 5 nodes.

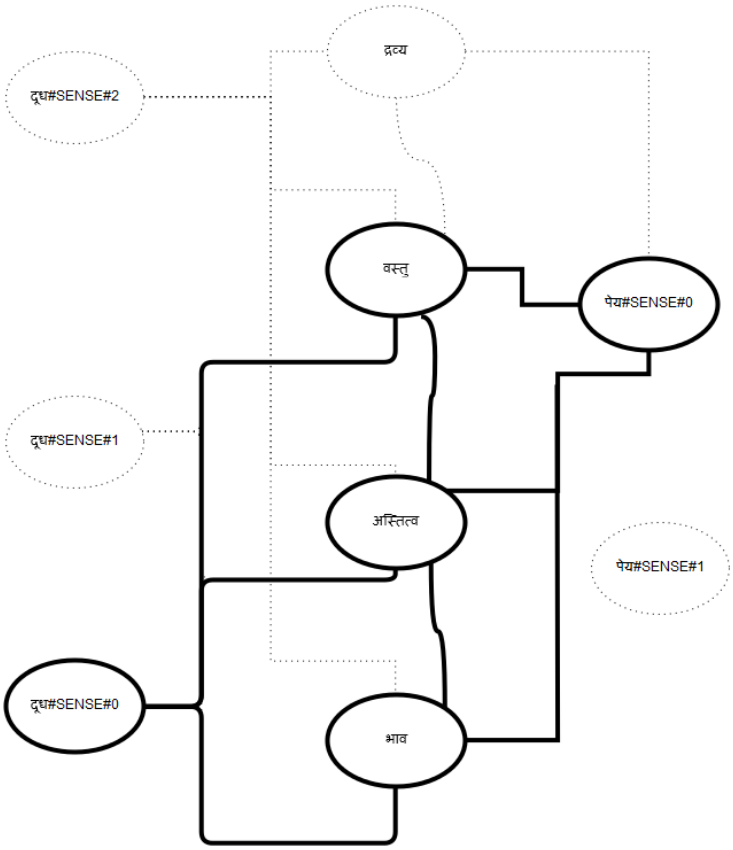

Figure 3.7: Sub graph 2 for milk Sense0 and drink Sense0 with total 9 edges and 5 nodes.

There are 3 different Global Measures which i have applied.

1.Edge Density

2.Compactness

3.Entropy

3.1.3 Implementation of Edge Density

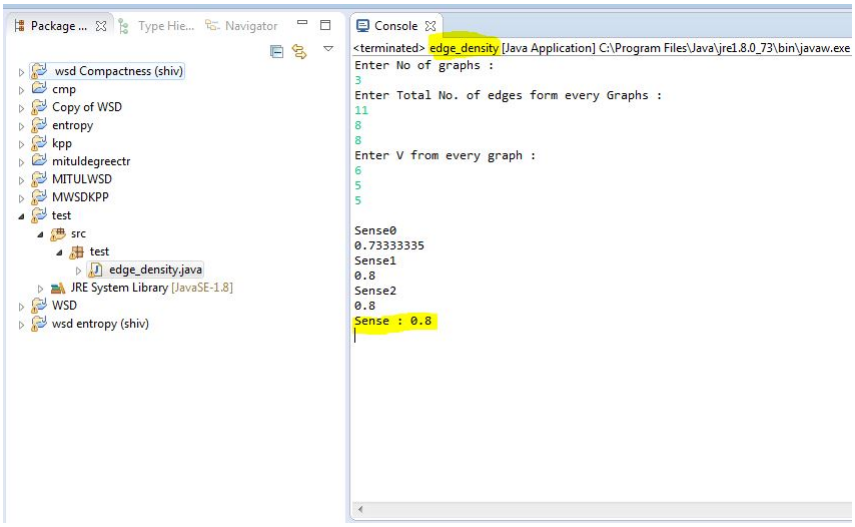

Figure 3.8: Implementation of Edge Density for Drink And Milk in Hindi with sense

3.1.4 Implementation of Compactness

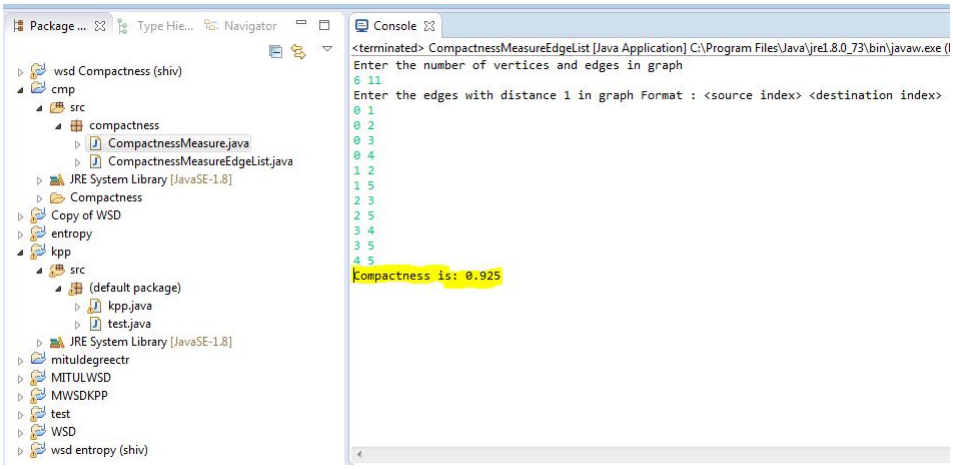

Figure 3.9: Implementation of compactness for Drink And Milk in Hindi for Sub-Graph1

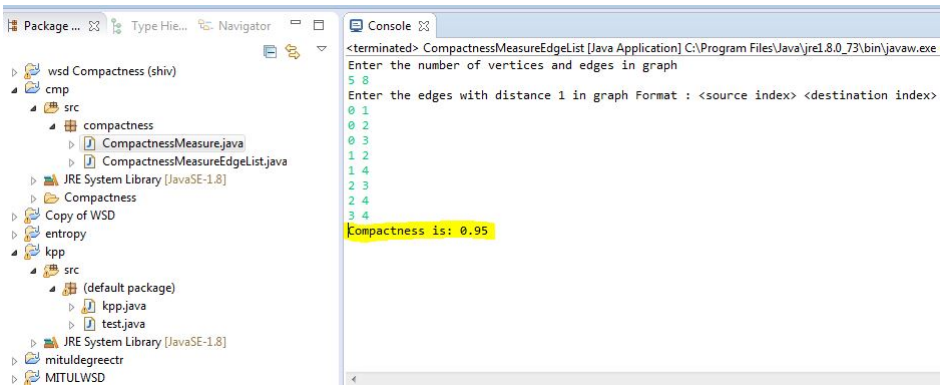

Figure 3.10: Implementation of compactness for Drink And Milk in Hindi for Sub-Graph2

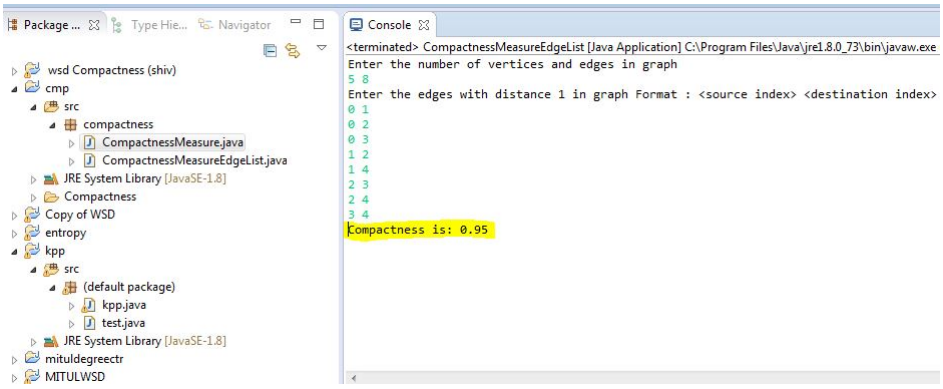

Figure 3.11: Implementation of compactness for Drink And Milk in Hindi for Sub-Graph3

3.1.5 Implementation of Entropy

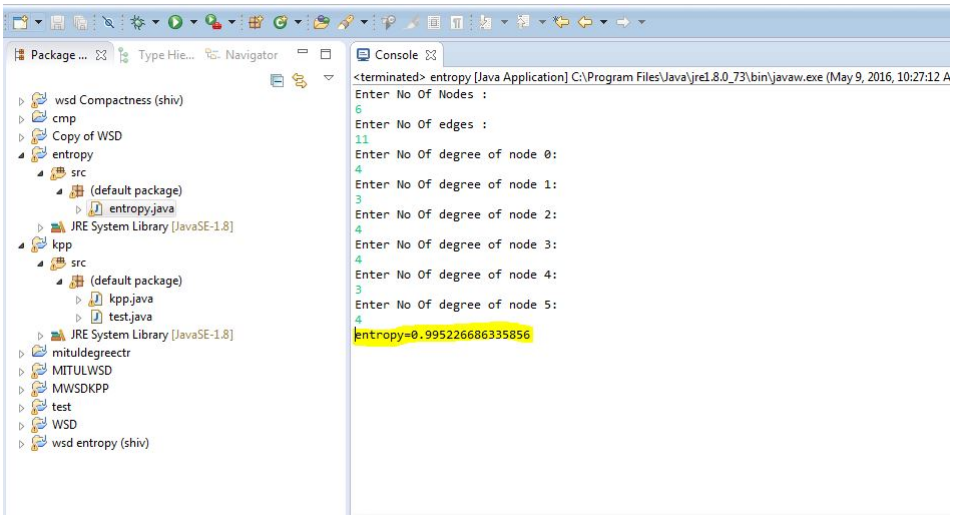

Figure 3.12: Implementation of Entropy for Drink And Milk in Hindi SubGraph1

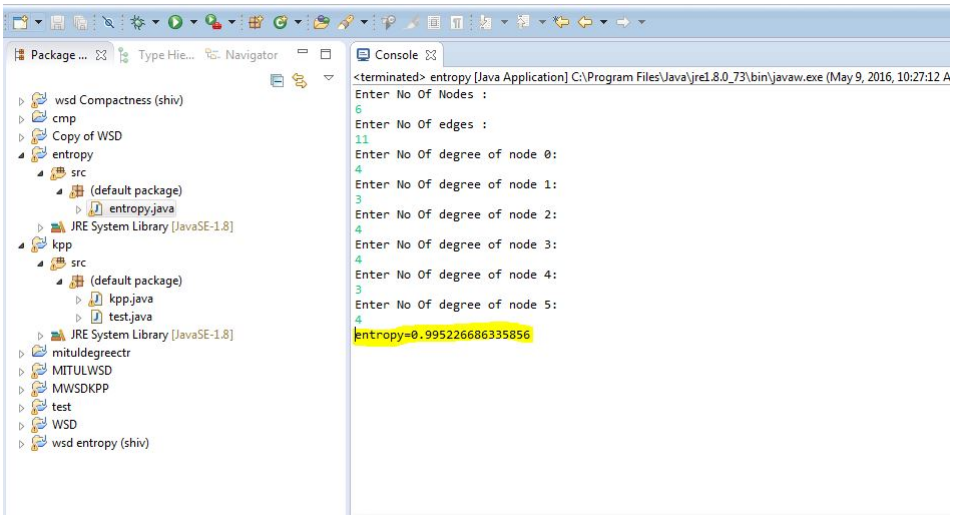

Figure 3.13: Implementation of Entropy for Drink And Milk in Hindi for SubGraph2

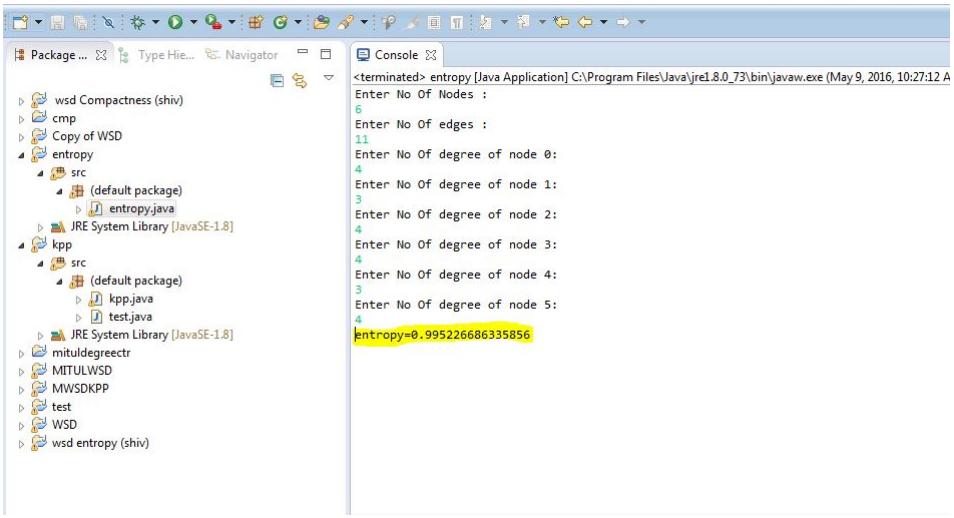

Figure 3.14: Implementation of Entropy for Drink And Milk in Hindi for SubGraph3

Chapter 4

Implementation Results

4.1 Correct Results for Global Measures

I have taken 300 different polysemous words to check the accuracy of the Global Measures. these are the correct results for every Global Measure.

4.1.1 Result for Edge Density

Global Measure: EDGE DENSITY

नौकर तानाशाह	G1(0.0)	G2(0.1)	G3(0.2)	G4(0.3)	G5(1.0)	G6(1.1)	G7(1.2)	G8(1.3)
Ed	0.55	**0.61**	**0.61**	0.47	0.55	**0.61**	0.60	0.47
वनस्पति वृक्ष	G1(0,0)	G2(1,0)						
Ed	**0.733**	0.666						
माँ पुत्र	G1(0.0)	G2(0.1)	G3(1.0)	G4(1.1)				
Ed	0.55	**0.61**	**0.61**	0.47				
गदा अस्त्र	G1(0.0)	G2(0.1)	G3(0,2)	G4(0.3)	G(0.5)			
Ed	0.60	0.66	0.66	**0.8**	0.66			
अब्ज अनुज	G1(0.0)	G2(0.1)	G3(0,2)	G4(1.0)	G(1.1)	G(1.2)		
Ed	0.47	0.61	0.61	0.41	0.73	**0.7**		
सूर्य सितारा	G1(0.0)	G2(0.1)	G3(0,2)					
Ed	**0.67**	**0.67**	0.6					
पाठशाला इमारत	G1(0.0)	G2(0.1)	G3(1.0)	G4(1.1)				
Ed	**0.666**	0.523	0.476	0.5				

Figure 4.1: Table for result

4.1.2 Result for Compactness

वनस्पति वृक्ष	G1(0,0)	G2(0,1)	
Cp	**0.925**	0.92	
पूजा तिलक	G1(0.0)	G2(0.1)	G3(0.2)
Cp	-	**1**	-
खेत फसल	G1(0.0)	G2(1.0)	G3(2.0)
Cp	0.95	**1**	0.925
दूध पैय	G1(0.0)	G2(0.1)	G3(0.2)
Cp	0.925	**0.95**	**0.95**
यंत्र कारखाना	G1(0,0)	G(0,1)	
Cp	0.95	**1**	

Figure 4.2: Table for result

4.1.3 Result for Entorpy

Global Measure: ENTROPY

तानाशाह नौकर	G1(0.0)	G2(0.1)	G3(0.2)	G4(0.3)	G5(1.0)	G6(1.1)	G7(1.2)	G8(1.3)
En	0.978	0.984	0.984	0.966	0.97	**0.984**	**0.984**	0.966
पुत्र पिता	G1(0.0)	G2(0.1)	G3(0.2)	G4(0.3)	G5(1.0)	G6(1.1)	G7(1.2)	G8(1.3)
En	0.97	0.97	0.97	0.97	0.98	**0.9847**	0.966	**0.9847**
महेश विष्णु	G1(0.0)	G2(0.1)	G3(0.2)	G4(0.3)	G5(1.0)	G6(1.1)	G7(1.2)	G8(1.3)
En	0.960	**0.984**	0.97	0.98	**0.984**	0.9511	**0.984**	0.966
वनस्पति वृक्ष	G1(0,0)	G2(1,0)						
En	**0.99522**	0.99072						
माँ पुत्र	G1(0.0)	G2(0.1)	G3(1.0)	G4(1.1)				
En	0.978	0.984	**0.984**	0.966				
देव मानव	G1(0.0)	G2(0.1)	G3(1.0)	G4(1.1)				
En	**0.995**	0.99	0.990	0.99				
गदा अस्त्र	G1(0.0)	G2(0.1)	G3(0,2)	G4(0.3)	G(0.5)			
En	0.984	0.990	0.990	**0.995**	0.990			
अग्रज अनुज	G1(0.0)	G2(0.1)	G3(0,2)	G4(1.0)	G(1.1)	G(1.2)		
En	0.966	0.984	0.984	0.955	**0.995**	0.995		

Figure 4.3: Table for result

4.2 Accuracy Comparison For Global Measures

As Shown in Graph,out of 150 tuples there are only 5 correct results for Compactness , 7 correct results for Edge Density and 8 correct results for Entropy. The accuracy is very poor as this is Unspervised approach.

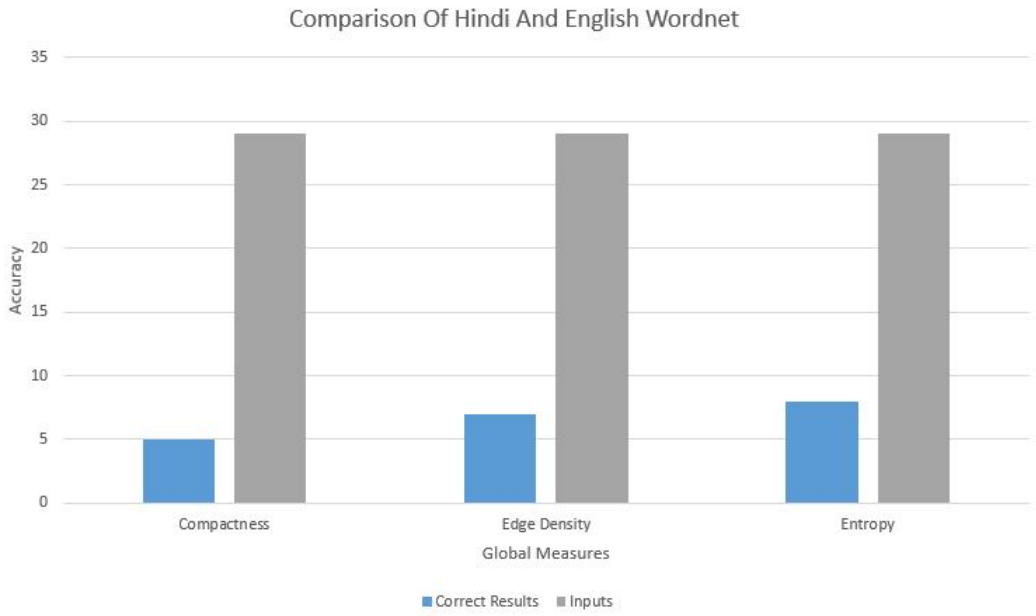

Figure 4.4: Comparison Graph Global Measure

Chapter 5

Conclusion and Future Scope

Word Sense Disambiguation using Graph based Unsupervised approach is done and global measures like Edge Density,Entropy,compactness are applied. Unsupervised algorithms doesn't require any training set however it will be generated from context word but each new polysemous training is required. Other limitation is poor accuracy and the techniques which have higher accuracy are very complex to implement.

Accuracy can be improved using more global measures and different ways for graph construction.Performance can be enhanced by automatic detection of polysemous words from documents.

References

[1] R. Navigli and M. Lapata, "An experimental study of graph connectivity for unsupervised word sense disambiguation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 678–692, April 2010.

[2] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 2, p. 10, 2009.

[3] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[4] M. E. B. Menai and W. Alsaeedan, "Genetic algorithm for arabic word sense disambiguation," in *Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on*, pp. 195–200, IEEE, 2012.

[5] M. Z. Merhbene Laroussi, Anis Zouaghi, "Unsupervised system for lexical disambiguation of arabic language using a vote procedure," *IEEE-ICoAC*.

[6] E. Agirre and P. G. Edmonds, *Word sense disambiguation: Algorithms and applications*, vol. 33. Springer Science & Business Media, 2007.

[7] S. P.Tamilselvi, "Optimal distance metric function with trigram features for case based word sense disambiguation using artificial neural network,"

[8] E. F. Kelly and P. J. Stone, *Computer recognition of English word senses*, vol. 13. North-Holland, 1975.

[9] E. Black, "An experiment in computational discrimination of english word senses," *IBM Journal of research and development*, vol. 32, no. 2, pp. 185–194, 1988.

[10] G. Escudero, L. Màrquez, G. Rigau, and J. G. Salgado, "On the portability and tuning of supervised word sense disambiguation systems," 2000.

[11] R. Sinha and R. Mihalcea, "Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity," in *null*, pp. 363–369, IEEE, 2007.

[12] R. Navigli and P. Velardi, "Structural semantic interconnections: a knowledge-based approach to word sense disambiguation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 7, pp. 1075–1086, 2005.

[13] R. Mihalcea, "Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 411–418, Association for Computational Linguistics, 2005.

[14] C. Fellbaum, "ed. wordnet: an electronic lexical database," *MIT Press, Cambridge MA*, vol. 1, p. 998, 1998.

[15] E. Agirre and A. Soroa, "Using the multilingual central repository for graph-based word sense disambiguation.," in *LREC*, Citeseer, 2008.

[16] M. Galley and K. McKeown, "Improving word sense disambiguation in lexical chaining," in *IJCAI*, vol. 3, pp. 1486–1488, 2003.

[17] D. Vickrey, L. Biewald, M. Teyssier, and D. Koller, "Word-sense disambiguation for machine translation," in *Proceedings of the conference on Human Language*

Technology and Empirical Methods in Natural Language Processing, pp. 771–778, Association for Computational Linguistics, 2005.

[18] S. Russell, P. Norvig, and A. Intelligence, "A modern approach," *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, p. 27, 1995.