



Department of Computer Science
Albert Einstein School Of Physical Sciences
ASSAM UNIVERSITY, SILCHAR
(A Central University constituted under Act XIII of 1989)
Silchar - 788 011, Assam, India

DECLARATION

I, Arindam Roy, bearing Registration Number PhD/731/2009 dated 22-11-2009, hereby declare that the subject matter of the thesis entitled “An Investigation of Machine Learning Techniques and Their Application to WSD” is the record of works done by me and that the contents of this thesis did not form the basis for award of any degree to me or to anybody else to the best of my knowledge. The thesis has not been submitted in any other University / Institute.

Place: Silchar

(Arindam Roy)

Date:

ACKNOWLEDGEMENTS

I address my sincere gratitude to the Almighty as whenever I faced any difficulty I used to pray to help me and He always was there helping me to complete this assignment.

I would like to express my sincere thanks to my supervisor Professor Bipul Syam Purkayastha for his continuous guidance and encouragement throughout the PhD journey. I am also thankful to Prof. Bipul Syam Purkayastha, Head, Department of Computer Science and Dean, Albert Einstein School of Physical Sciences, for his valuable support and suggestions.

I am thankful to Mrs. Sunita Sarkar, Assistant Professor, Department of Computer Science for her constant support and help during this research.

I also express my sincere thanks to all my colleagues, staff members and all of them who helped me directly or indirectly to accomplish my research work.

Last but not the least; I would love to thank my wife for her support and encouragement.

Date:

(Arindam Roy)

Table of Contents

List of Figures	ix
List of Tables	xi
Abstract	xiii
Chapter 1 Introduction	1
1.1 Motivation.....	3
1.2 Objectives	3
1.3 Methodology.....	4
1.4 Contributions	4
1.5 Thesis Outline	5
Chapter 2 Literature Review	7
2.1 Introduction	7
2.2 Multi Lingual Dictionary and WordNet	7
2.3 Approaches to WSD and related algorithms	9
2.4 Lemmatizers and sense tagged documents.....	14
Chapter 3 WordNet, IndoWordNet and Sense Marking Tool	20
3.1 Introduction	20
3.2 WordNet.....	21
3.2.1 Lexical Entries	22
3.3 Constituent Elements of WordNet.....	23
3.3.1 Synset and Concept	23
3.3.2 Relations in WordNet.....	24
3.3.2 .1 Semantic Relations	24
3.3.2 .2 Lexical Relations.....	27

3.4 Indo WordNet.....	27
3.5 Sense marking Tool.....	30
3.6 Chapter Summary	33
Chapter 4 Word Sense Disambiguation.....	34
4.1 Introduction	34
4.2 Approaches to WSD.....	35
4.2.1 Knowledge Based Approach.....	35
4.2.1.1 WSD Using Selectional Semantic Relations	36
4.2.1.2 Overlap Based Approaches.....	37
4.2.1.2.1 Lesk Algorithm.....	37
4.2.1.2.2 Walker Algorithm.....	38
4.2.1.2.3 WSD using Conceptual Density	39
4.2.1.3 Knowledge Based Approaches-Comparison.....	39
4.2.2 Machine Learning Based Approaches.....	40
4.2.2.1 Supervised Approaches	40
4.2.2.1.1 Naïve Bayes Algorithm	40
4.2.2.1.2 Decision List Algorithm.....	41
4.2.2.1.3 Exemplar based WSD (K-NN)	42
4.2.2.1.4 WSD Using Support Vector Machine.....	43
4.2.2.1.5 Supervised Approaches – Comparison of methodology	43
4.2.2.1.5 Supervised Approaches – Comparison of Performance	44
4.2.2.2 Semi Supervised Approach.....	45
4.2.2.2.1 Semi Supervised Decision List Algorithm.....	45
4.2.2.3 Unsupervised Approaches.....	46
4.2.2.3.1 Hyperlex	46

4.2.2.3.2 Yarowsky’s Algorithm (WSD Using Roget’s Thesaurus Categories)	47
4.2.2.3.3 Unsupervised WSD using Parallel Corpora.....	48
4.2.2.3.4 Dekang Lin’s approach.....	49
4.2.2.3.5 Unsupervised Approaches – Comparison of Performance	51
4.2.2.3.6 Unsupervised Approaches – Comparison of methodology.....	52
4.2.3 Hybrid Approach	52
4.2.3.1 Sense Learner.....	53
4.2.3.2 An Iterative Approach to Word Sense Disambiguation	53
4.2.3.3 Structural Semantic Interconnections(SSI).....	54
4.2.3.4 Comparison of Performance among the various Hybrid approaches....	54
4.3 Performance Metric of a WSD system	55
4.4 Chapter Summary	55
Chapter 5 Lemmatization Using Trie Data Structure and Hopfield Network.....	56
5.1 Introduction	56
5.2 Approaches for Lemmatization	57
5.2.1 Levenshtein Distance Dictionary Based Approach:	57
5.2.2 Morphological Analyzer Based Approach	59
5.2.3 Affix Lemmatizer	60
5.2.4 Fixed Length Truncation	60
5.2.5 Trie Based Approach.....	61
5.3 Basic Operations In Trie	62
5.2.1 Searching In a Trie	62
5.2.2 Insertion Into a Trie	62
5.4 Algorithm for Lemmatization	63
5.5 Hopfield Network.....	66

5.5.1. Attraction and repulsion between neurons	68
5.6 Energy of a Hopfield Network.....	68
5.7 Chapter Summary	69
Chapter 6 A Proposed Hybrid Approach for WSD	70
6.1 Introduction	70
6.1.1 Domain/Corpus Specific Sense Distribution	70
6.1.2 Dominant concepts within a domain.....	71
6.1.3 Conceptual Distance.....	71
6.1.4 Semantic Graph Distance	73
6.2 Sense information from Overlap Approach	74
6.2.1 Implementation of Overlap Approach (Overlap_App)	76
6.3 A Proposed Hybrid Algorithm for WSD (Hy_WSD).....	76
6.4 Experimental Setup	79
6.5 Experimental Results and Discussions	80
6.6 Observations from experiments	85
6.7 Chapter Summary	86
Chapter 7 Conclusions	87
7.1 Summary of Works.....	87
7.2 Summary of Contributions	89
7.3 Future Directions	89
Bibliography	90
Associated Publications.....	97

List of Figures

3.1 Illustration of aligned synset members for the concept id 4265: a youthful male person	21
3.2 Showing Hypernymy/Hyponymy relations	26
3.3 Meronymy/Holonymy relations	26
3.4 Linked Structure of IndoWordNet	28
3.5 Home page of IndoWordNet	29
3.6(a) A search page showing the result of search of the concept ‘ফল’	29
3.6(b) A search page showing the result of search of the concept ‘শাম’	30
3.7 Layout of Sense Marking Tool	31
3.8 Sense Tagged Document	32
4.1 Example of Decision List Algorithm.....	42
5.1 Lemmatization Process.....	57
5.2 Trie for a language consisting of words a , abase , abate and bat.....	61
5.3 Lemmatizer showing the output for some Bengali inflected words.....	65
5.4 Lemmatizer showing the output for Bengali inflected word ক্রীতদাসদের.....	65
5.5 Lemmatizer showing the output for Bengali inflected word তস্বাবধানে.....	66
5.6 Layout of Hopfield Network.....	67
6.1 Hypernymy-hyponymy graph for contextual distance example	72
6.2 Example of Semantic Graph Distance	74

6.3 A snapshot showing implementation of the Overlap_App	77
6.4 Snapshot of implementation of the proposed algorithm	79
6.5 Supervised training from Mixed to Tourism(Bengali).....	81
6.6 Supervised Training from Mixed to Health(Bengali)	81
6.7 Supervised Training from Tourism to Health (Bengali)	82
6.8 Supervised training from Health to Tourism (Bengali).....	82
6.9Supervised training from Tourism to Health (Nepali)	83
6.10Supervised training from Health to Tourism (Nepali)	83

List of Tables

2. 1 Corpora Tagged with Senses	16
3.1 Multi Dictionary Model	21
3.2 Lexical Matrix	22
3.3 Semantic Relations in WordNet	25
3.4 Examples of Synonyms in Bengali WordNet	25
4.1 Assigning scores in Walker's algorithm	38
4.2 showing comparisons among KB approaches	39
4.3 Performance comparison among Supervised Approaches	44
4.4 Example list showing a run of Yarowsky's algorithm	48
4.5 Subjects of "employ" in a 25-million-word Wall Street Journal corpus	50
4.6 Comparison among various Unsupervised Approaches	51
4.7 Comparison of Performance among the various Hybrid approaches	54
6.1 Dominant concepts for the Tourism and Health domains	71
6.2 Total number of Polysemous and monosemous words in 2 domains of Bengali corpus	79
6.3 Total number of Polysemous and monosemous words in 2 domains of Nepali corpus	80

6.4 Comparing the performance of WordNet first sense(wnfs) baseline , PPR with proposed algorithm(Hy_WSD) on Tourism and Health domain of Bengali corpus 84

6.5 Comparing the performance of WordNet first sense(wnfs) baseline , PPR with proposed algorithm(Hy_WSD) on Tourism and Health domain of Nepali corpus . 85