

DECLARATION

I, Arindam Dey, bearing Registration Number PhD/2051/2012 dated 12-09-2012, hereby declare that the subject matter of the thesis entitled “ Named Entity Recognition: A Computational Approach ” is the record of works done by me and that the contents of the thesis did not form the basis for award of any other degree to me or to anybody else to the best of my knowledge. The thesis has not been submitted in any other University / Institute. This thesis is being submitted to Assam University for the degree of Doctor of Philosophy in Computer Science.

Place:
Date:

(Arindam Dey)
Research Scholar

CONTENT

1 INTRODUCTION	11
1.1 Different approaches of NER	12
1.2 Current status in Named Entity Recognition (NER) for Indian Language(IL)	14
1.3 Issues with Indian Language	14
2 REVIEW OF LITERATURE	17
2.1 Named Entity Recognition and Classification (NER)	17
2.2 NER Applications	18
2.3 Approaches of NER	19
2.4 Challenges in NER	24
2.5 Related Works	25
3 INFLECTIONAL LANGUAGE WITH SPECIAL REFERENCE TO NEPALI	30
3.1 Inflectional Language	30
3.2 Nepali Language	30
3.3 Nepali as an Inflectional Language	44
4 STEMMING	38
4.1 Types of Stemming	38
4.2 System Description	38
4.3 Proposed Algorithm for STEMMER	40
4.4 Experimental Results and Discussion	44
5 CHUNKING USING HIDDEN MARKOV MODEL (uni-gram Technique)	47
5.1 Chunk Representation	48
5.2 Architecture of the CHUNKER	49
5.3 The Corpus	51
5.4 Test Results	51
6 NAMED ENTITY RECOGNITION USING GAZETTEER LIST AND HIDDEN MARKOV MODEL(n-gram technique)	53
6.1 Gazetteer Method	53
6.2 Advantages of Gazetteer Method	53
6.3 Hidden Markov Model	54
6.4 N-Gram Technique under HMM	55

6.5 System Design briefing	56
6.6 Result Analysis	58
7 CONCLUSION	61
8 REFERENCES	63
APPENDIX	69
Appendix A	69
Appendix B	72
Appendix C	73
Appendix D	74

LIST OF TABLES

Table 1 : Comparison of Rule based and Machine learning approach	13
Table 2 : The number and case suffixes of nouns	32
Table 3 : The number and case inflection of noun Book/किताब	32
Table 4 : The second and third person singular forms	34
Table 5 : Simple Present conjugation of the verb हुनु hunu	36
Table 6 : Test Cases	42
Table 7 : Accuracy of System on different Test Cases	44
Table 8 : Test Result for Nepali Based Phrase Chunker.	51
Table 9 : Different Approaches of NER and Their Accuracy.	56
Table 10: Total number of tags in the Corpus.	59
Table 11: Accuracy is 85.71% from 1000 sentences using n-gram and Gazetteer Method.	59

LIST OF FIGURES

Figure 1: Different Approaches of NER and Their Accuracy	28
Figure 2: Architecture of Stemmer	43
Figure 3: Work flow of the Chunking Process	50
Figure 4: Architecture of NER tool using Gazetteer Method.	58

ACKNOWLEDGMENT

First, I wish to thank my supervisor Professor Bipul Syam Purkayastha for giving his support for this thesis, and helping me set a direction for my work, for which I am extremely grateful. This thesis would also not have been possible without Mr.Abhijit Paul, who was always there with his insight whenever I needed it. Long conversations and discussions with him have shaped this thesis, and so he deserves a lot of the credit for the work presented here. I also want to thank Mr.Kh Raju Singha and Mrs.Krishnabati Singha for encouraging me. Whenever I was stuck, they boost me by giving me several good ideas in the process. I have to thank my parents without whom I would not be here. Their love has made me what I am. I also thank my friends and other co-scholars for there very valuable support.