

# DECLARATION

I, Prajadhup Sinha, Registration No.: Ph.D/1780/11 Dated 22/09/2011 hereby declare that the subject matter of the thesis entitled “**Part-of-Speech tagging with Computational Approach: A Special Reference to Nepali Language**” 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:**

*(Prajadhup Sinha)*

**Date:**

Research Scholar

# ACKNOWLEDGEMENT

First of all, I am very grateful to God for His blessings and inspirations which have enabled me to complete this thesis. I also express my heartfelt gratitude and unforgettable indebtedness to my guide **Prof. Bipul Syam Purkayastha**, Head of the Department of Computer Science, Assam University, for his expert guidance, valuable suggestions, constant encouragement and above all the understanding and wholehearted cooperation throughout the entire period.

I am immensely grateful to Asst Prof Miss Nuzotalu M. Veyie for her constant support for successful completion of this work. I also thank Mr Mathew and my faculty members of Computer Science, Kohima Science College, for providing the necessary software and also for the help rendered at various phases of the project work.

I am very much thankful to Poireiton, Bablu, Bhairab, Abhijit and all the research scholars of the Department of Computer Science (Assam University) for their direct or indirect help in various forms during my research work.

I convey my sincere gratitude to Asst Prof Dr. Shanta Gurung, Department of Hindi, Kohima Science Department, for her help to gather the required knowledge of Nepali language and its typological features.

I also would like to express my gratefulness to all my family members Dipak,

Upasana, Dulal, Runa, Suma, Partha, Pratap, Uday and Prachi for their continuous support, care and encouragement.

Finally I express my special appreciation and acknowledgement to my wife Barnali(Pinky), for her constant support, co-operation and sacrifice throughout my research work. Last but not the least; I thank all my well-wishers who directly or indirectly contributed for the completion of this thesis.

**Date:**

*Prajadhip Sinha*

# DEDICATION

*To my Parents*

*Mr. & Mrs Krishnadas and Pramila Sinha*

*And*

*To my Wife and Daughter*

*Barnali & Prachi Sinha*

*This humble work is a sign of my love to*

*YOU!*

# Content

List of Figures-----	ix
List of Tables-----	x
List of Abbreviation -----	xi
Abstract-----	xii

## CHAPTER 1

<b>INTRODUCTION .....</b>	<b>1</b>
1.1 Part of speech tagging.....	3
1.2 Problems of POS tagging .....	5
1.3 Motivation .....	6
1.4 Objectives .....	7
1.5 Methodology .....	8
1.6 Thesis Organization.....	8

## CHAPTER 2

<b>LITERATURE SURVEY.....</b>	<b>10</b>
2.1 A review of POS Tagging Approaches.. .....	10
2.2 Indian Language Taggers.....	15
2.3 Nepali Language Taggers.....	19
2.4 Corpus Review.....	19
2.5 Nepali Corpus.....	20

## CHAPTER 3

<b>PART OF SPEECH TAGGING TECHNOLOGIES.....</b>	<b>23</b>
3.1 Different approaches for POS tagging .....	23
3.2 Supervised and Unsupervised Part of Speech Tagging .....	24

3.3	Rule based Approach .....	24.
3.4	Stochastic Approach.....	25
3.4.1	Hidden Markov Model.....	26
3.4.2	Conditional Random Fields.....	28
3.4.3	Maximum Entropy Model.....	28
3.4.4	Memory Based Learning .....	30
3.5	Hybrid Part of Speech tagging .....	31

## **CHAPTER 4**

### **FOUNDATIONAL CONSIDERATIONS.....32**

4.1.	Nepali Language overview.....	32
4.2	Computational Morphology.....	33
4.2.1	Noun and lexical Morphology.....	33
4.2.2	Adjective and Lexical morphology.....	35
4.2.3	Verb and Lexical morphology.....	37
4.2.4	Affixes.....	38
	4.2.4.1 Prefix.....	39
	4.2.4.2 Suffix.....	39
4.3	Nepali Tagset Review.....	41
4.3	Description of Nepali Tagset .....	42

## **CHAPTER 5**

### **PROPOSED POS TAGGER.....57**

5.1	POS Tagger Architecture .....	57
5.2	Description of the proposed architecture .....	57

5.3	The Proposed Algorithm.....	65
5.4	Website & GUI Tool.....	67
5.4.1	Website.....	67
5.5.2	Graphical User Interface (GUI) Tool.....	68

## CHAPTER 6

<b>EXPERIMENTAL RESULTS AND EVALUATION.....</b>		<b>70</b>
6.1	Experimental Results . . . . .	71
6.2	Performance of different number of words.....	80
6.3	Input/output of the Tagger.....	81
6.4	System Performance.....	83

## CHAPTER 7

<b>CONCLUSIONS &amp; FUTURE WORK .....</b>		<b>87</b>
7.1	Conclusions.....	87
7.2	Future work.....	88
<b>REFERENCES.....</b>		<b>90</b>

<b>Appendix A .....</b>		<b>97</b>
	List of Publications.....	97
<b>Appendix B.....</b>		<b>98</b>
	Participation in Seminars, Conferences and workshops.....	98
	Paper presented in the Conferences.....	99

## LIST OF FIGURES

---

<i>FIGURE NAME</i>	<i>PAGE NO.</i>
Fig 3.1 Classification of POS tagging models .....	23
Fig. 5.1 Proposed System Design of Nepali POS Tagger.....	58
Fig 5.2 Tag transition probabilities.....	61
Fig 5.3 Nepali POS Tagger Domain.....	67
Fig 5.4 Nepali POS Tagging tool.....	68
Fig 5.5 Online Nepali POS Tagging Tool.....	69
Fig 6.1: Experiment Set 1 based on lexicon size: 5000 words.....	72
Fig 6.2: Experiment Set 2 based on lexicon size: 10000 words.....	75
Fig 6.3: Experiment Set 3 based on lexicon size: 15000 words .....	77
Fig 6.4: Experiment Set 4 based on lexicon size: 20000 words .....	80
Fig 6.5 Performance of different number of words.....	81
Fig 6.6 Output of the tagger .....	83



## LIST OF TABLES

---

TABLE NAME.	PAGE NO.
Table 2.1: POS tagging accuracy in the NLP AI machine learning contest.....	18
Table 4.1: The number and case suffixes of nouns.....	34-35
Table 4.2: The number and case inflection of noun Book/किताब.....	35
Table 4.3: Inflections of adjectives.....	36
Table 4.4: List of Part-of-Speech tags for Nepali Language.....	43-46
Table 4.5: Classification of Pronoun with example.....	47
Table 4.6: Classification of Verb with example.....	49
Table 4.7: Classification of Postpositions with example.....	52
Table 4.8: Classification of Punctuation with example.....	55
Table 6.1: Experiment-1 (Size of lexicon- 5000 .....	72
Table 6.2: Experiment Set 2 based on lexicon size: 10000 words .....	74
Table 6.3: Experiment Set 3 based on lexicon size: 15000 words .....	77
Table 6.4: Experiment Set 4 based on lexicon size: 20000 words.....	79
Table 6.5: Performance of different number of words.....	81
Table 6.6: Precision and Recall accuracy on different corpus.....	85

## **LIST OF ABBREVIATIONS**

---

NLP	Natural Language Processing
IR	Information Retrieval
IE	Information Extraction
MT	Machine Translation
BIS	Bureau of Indian Standards
CL	Computational Linguistics
CRF	Conditional Random Field
GUI	Graphical User Interface
HMM	Hidden Markov Model
HON	Honorificity
ILPOST	Indian Language Part of Speech Tagset
MSD	Morpho Syntactic Descriptions
MWE	Multi Word Expression
POST	Part of Speech Tagging
TBL	Transformation Based Learning
WSD	Word Sense Disambiguation