DECLARATION

I, Md. Jaynal Abedin, Registration No.: Ph.D/1781/2011 Dated 22/09/2011 hereby declare that the subject matter of the thesis entitled "Automated Multiword Expressions Detection" 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: (Md. Jaynal Abedin)

Date: Research Scholar

Dedicated to My Dearest Parents

Moulana Morthuza Hussain (father)

and

Late Reshma Khanom (mother)

ACKNOWLEDGEMENT

All praises should go to ALLAH for keeping me strong enough and giving me the courage and patience to finish this work. I would like to express my earnest appreciation and heartfelt thanks to **Prof. Bipul Syam Purkayastha**, Supervisor, Head & Dean, Department of Computer Science, Albert Einstein School of Physical Sciences, Assam University, Silchar for his patience, constructive guidance, kindness and logistic support throughout my studies. Through him only I have been introduced to this wonderful field of Natural Language Processing and above all, he opened my eyes to the exacting but so gratifying world of research. I convey my sincerest gratitude to Prof. Pushpak Bhattacharyya, Director, NIT, Patna, Bihar for his constructive suggestions and encouragement by providing his valuable time for my research work.

I am thankful to all the faculty members in the Department of Computer Science,for their valuable support and suggestions. I am also thankful to all the non teaching staff and Ph.D scholars of Computer Science Department for their kind co-operation.

I also would like to extend special thanks to Late Ramendu Bhattacherjee, Ex.Pro-Vice Chancellor of Assam University Silchar, Prof. Debajyoti Biswas, Retd. Dean, School of Physical Sciences, Prof. K.Hemachandran. Professor, Department of Computer Science, Dr. Sahin Ara Begum, Associate Professor, Department of Computer science, Prof. C.R. Bhattacharjee, Department of Chemistry, Assam University Silchar, for their encouragement, moral support, kind co-operation and valuable suggestions.

I also extend my heartfelt thanks to Kh Raju Singha, Bablu Yumnam Singh, Abul Kasim Md.Fujail, Rejaul Karim Barbhuiya, Mehfuj Alom Mazumdar, Arif Uddin, Abijit Paul, Arindam Dey, Ksh. Krishna Bati Singha, Aadil Ahmad Lawaye and Saiful Islam for their kind and supportive attitudes that have been very conducive for the completion of my research work. I also express my heartiest thanks to my mother **Late Reshma Khanom,** to my brothers and to my sisters. Last but not the least my heartfelt thanks should go to my father, **Moulana Morthuza Hussain.** His understanding and love encouraged me to pursue Ph.D with enthusiasm and complete this work.

Date: MD. JAYNAL ABEDIN

Contents

Page No

List of Abbreviations…………………………………………………….. ……… ..xi

List of Figures……………………………………………………………...xii

List of Tables…………………………………………………………..........xiii

Abstract ……………………………………………………………………xiv

CHAPTER 1……………………………….……..…………………..…….1

Introduction……………………………………..….……………..……...1

1.1 Motivation ……………………………………………………….……....3

1.2 Objectives………………………………………………………..........…6

1.3 Methodology……………………………………………………..……...7

1.4 Main Contributions.…………………………………………...……..7

1.5 Thesis Outline ………………………….……………………….… .8

1.6 Chapter Summary ……………………………………………..……….9

CHAPTER 2 ……………………………………………………………10

Review of Literature…………………………………………………….........10

2.1 Different running projects on MWEs………………………………….11

2.2 Chapter Summary……………………………………………............15

CHAPTER 3…………………………………………………………… ……16

Multiword Expressions ……………………………………………….……....16

3.1 Introduction ………………………………………………………...16

3.2 Necessary and sufficient condition for Multiword Expressions ….………...18

3.3 Linguistic properties of Multiword Expressions ……………….….........19

3.3.1 Idiomaticity ……………………………………………………..20

3.3.1.1 Lexical Idiomaticity ……………………………….….........20

3.3.1.2 Syntactic Idiomaticity ……………………………..….........21

3.3.1.3 Semantic Idiomaticity ……………………………….........21

3.3.1.4 Statistical idiomaticity ………………………………….....23

3.4 Other Properties of Multiword Expressions ………………………...26

3.4.1 Cross Lingual Variation……………………………………......27

3.4.2 Single-word Paraphrasability…………………………………..27

3.4.3 Proverbiality………………………………………………..…..27

3.4.5 Prosody ………………………………………….……..28

3.5 Collocations and Multiword Expressions …………………….….......28

3.6 Types of Multiword Expressions ..………………………………….........28

3.6.1 Compound Noun ……………………………………….…......29

3.6.2 Conjunct Verb……………..………………………………............29

3.6.3 Compound Verb ……………………………………………......30

3.6.4 Reduplication …………………………………………………...30

3.7 Onomatopoeic expressions ………………………………….........30

3.8 Complete Reduplication ………………………………………….….........30

3.9 Partial Reduplication ………………………………………….………..31

3.10 Semantic Reduplication ……………………………………….……..31

3.11 Idiosyncrasies Observed In Multiword Expressions ……………..31

3.12 Characteristics of Multiword Expressions ………………………………...32

3.13 Types of Multiword Expressions ……………………………….........33

3.14 Classification Of Multiword Expressions ……………………….……...36

3.15 Chapter Summary ……………………………………………….…......38

CHAPTER 4 ……………………………………………………………..…..39

Part of Speech Tagging In Bengali ……………………………………..…...39

4.1 Introduction ………………………………………………………………....39

4.2 Related Works ……………………………………………………….…40

4.3 Annotation …………………………………………………………..46

4.3.1 Standford Log-Linear Part of Speech(POS) Tagger …………………47

4.3.2 Part of speech Tagger for English ……………………………......48

4.3.3 TnT Tagger ……………………………………………………...48

4.3.4 Brill tagger …………………………………………………………48

4.3.5 Tree-Tagger ……………………………………………………..49

4.4 Part of Speech Tagging in Bengali for Multiword Expressions ….………49

4.5 General Framework For Bengali Part Of Speech Tagging………….........50

4.6 Bengali Part of Speech tagset(LDC-IL based) ……………………...……51

4.7 Bengali Part of speech Category ……………….………………….…...54

4.8 Some examples of POS with corresponding Tag ……………….….......59

4.9 Various Approaches to POS Tagging ……………………….……...........60

4.10 Applications of Part of Speech Tagging …………………………….……...64

4.11 Chapter Summary…….……………………………………………….……..65

CHAPTER 5 ……………………………………………………………66

Methodological Framework for Multiword Expressions Acquisition ……......66

5.1 Step 1 Corpora Collection and Preprocessing……………………………...66

5.1.1 Text Preprocessing ………………………………………………67

5.1.2 Stop words ………..……………………………………….....67

5.1.3 Stemming …………..…………………………………………....68

5.2 Step 2. Candidate Selection ……………………………………….72

5.2.1 Linguistic Approaches……………………………………………72

5.2.2 Statistical Approaches ……………………………………………74

5.2.3 Hybrid Approaches ……………………………………….…………76

5.3 Step3. Statistical Co-Occurrence Tests …………..……………….........77

5.4 Step 4. Extracting Multiword Expressions Features ………………….....77

5.4.1 Extraction Methods ……………………………………….........77

5.4.1.1 Statistical Association Measures …………………………77

5.4.1.2 LocalMaxs Algorithm ………………………………........78

5.4.1.3 Smoothed Probabilities of n-Grams ……………………...79

5.4.1.4 Normalized Sequence Probabilities ………..………….........81

5.5 Data Preparation ……………………………………………………. 82

5.6 Multiword Expression Extraction System …………………………………83

5.6.1 Method Based on Sequence Frequencies (SF) ……………..83

5.6.2 Method Based on Smoothed Probabilities (SP) ………………. 84

5.6.3 Method Based on Normalized Sequence Frequencies (NS) 84

5.6.4 Method Based on Normalized Smoothed Probabilities (NSP)….84

5.7 Step 5. Detecting Multiword Expressions Detection ….………………84

5.8 Chapter Summary ……………………………………………………85

CHAPTER 6 ……………………………………………………………...86

Multiword Expressions Detection in Bengali ………………………………….86

6.1 Extracting Multiword Expressions Features……………………….........86

6.2 Algorithm for Multiword Expressions Detection………………….………87

6.3 Steps for proposed Flowchart ………………..…………..………..90

6.4 Evaluation methods ……………………………..………………...90

6.4.1 Measures for MWEs Extraction……………….…………...…..90

6.4.1.1 Pointwise Mutual Information……………….……...…...90

6.4.1.2 Chi-Square Test ………………………………………………92

6.5 System Performance …………………………………….......….. 93

6.6 Multiword Expressions Detection Method ……………………….…..96

6.7 Comparative Analysis ……………………………………………..102

6.8 Chapter Summary……………………………………………………103

CHAPTER 7 ………………………………………………………… ….104

Conclusions and Future Works …………………………..…………… 104

7.1 Summary of Contribution ……………………………….……..104

7.2 Future Works for Automated Multiword Expressions Detection…....105

References ………………………………………………….…………….107

Appendix A . ……………………………………………………….……...118

Appendix B ………..…………………………………………………..…...120

Appendix C ….…………………………………………………………....121

Appendix D ……………………..…………………………………….....122

Appendix E ..…………………………………………………………….123

List of Abbreviations

AM	Association Measure
BIS	Bureau of Indian Standards
BNC	British National Corpus
CL	Computational Linguistics
D	Demonstrative
GA	Grammatical Annotation
GUI	GUI Graphical User Interface
HMM	Hidden Markov Model
IR	Information Retrieval
LM	Language Model
LT	Language Technology
LVC	Light Verb Construction
MWE	Multiword Expression
MWT	Multiword Term
NLP	Natural Language Processing
P	Precision
PV	Phrasal Verb
SA	Syntactic Annotation
SMT	Statistical Machine Translation
TnT	Trigrams'n'Tags
SVM	Support Vector Machine
TBL	Transformation Based Learning
VPC	Verb Particle Construction
WSD	Word Sense Disambiguation

List of Figures

Page No.

Figure 3.1 Examples of Syntactic non-markedness vs. markedness 21

Figure 3.2 Examples of Semantics Idiomaticity 22

Figure 3.3 Classifications of MWEs 37

Figure 4.1 Various approaches to POS Tagging 61

Figure 5.1 System architecture of MWEs Extraction and Detection 69

Figure 5.2 Syntagmatic relationships between Words 73

Figure 6.1 Flow chart for the system model 89

Figure 6.2 Statistical performance of system model 96

Figure 6.3 Bengali MWEs Detection Tool “AMWET” GUI Interface 99

Figure 6.4 Bengali MWEs Detection Tool “AMWET” GUI Interface (input text) 100

Figure 6.5 Bengali MWEs Detection Tool “AMWET” GUI Interface (POS) 101

Figure 6.6 Bengali MWEs Detection Tool “AMWET” GUI Interface (After Detection) 102

List of Graph

Graph 6.1 Result Analysis 95

List of Tables Page No.

Table 3.1 Different examples of statistical idiomaticity 24

Table 3.2 Classification MWEs in term of their different Idiomaticity 26

Table 4.1 List of POS tagging in Bengali 51-54

Table 4.2 Types of Noun 55

Table 4.3 Types of Pronoun 55

Table 4.4 Types of Demonstrative 56

Table 4.5 Types of Nominal Modifier 56

Table 4.6 Types of Verb 56

Table 4.7 Types of Adverbs 56

Table 4.8 Types of Participle 57

Table 4.9 Types of Postposition 57

Table 4.10 Types of Particles 57

Table 4.11 Types of Numeral 58

Table 4.12 Types of Reduplication 58

Table 4.13 Types of Residual 58

Table 4.14 Types of Unknown 58

Table 4.15 Types of Punctuation 59

Table 6.1 Experimental Result 94

Table 6.2 Result Analysis 94

Table 6.3 Comparative Analysis 103