TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	iii
ABSTRACT (ENGLISH)	v
ABSTRACT (THAI)	vii
LIST OF TABLES	xiii
LIST OF FIGURES	xiv
ABBREVIATIONS AND SYMBOLS	xvi
CHAPTER 1: INTRODUCTION	1
1.1 Objectives of this thesis	4
1.2 Scope and Outline of this thesis	4
CHAPTER 2: INFORMATION EXTRACTION	BACKROUND 9
2.1 Introduction	9
2.2 Effectiveness of General Term in Biologi	cal Literature 12
2.2.1 Introduction	12
2.2.2 Methods and Dataset	12
2.2.3 Results and Discussion	
2.3 Biological Information Retrieval using La	atent Semantic Indexing 18
2.3.1 Introduction	
2.3.2 Methods and Dataset	19
2.3.3 Results and Discussion	eser 25

2.4 Biological Text Classification using Machine Learning Techniques 27

2.4.1 Introduction	27
2.4.2 Methods and Dataset	27
2.4.3 Results and Discussion	42
2.5 Conclusion	43
CHAPTER 3: BIOLOGICAL TERMS RECOGNITION	45
3.1 Introduction	46
3.2 Biological Named Entity Recognition with Graphical Models	48
3.2.1 Hidden Markov Model (HMM)	48
3.2.2 Maximum Entropy Markov Model (MEMM)	50
3.2.3 Conditional Random Field Model (CRFs)	51
3.2.4 The comparison of HMM, MEMM and CRFs	55
3.3 Biological Information Retrieval using Latent Semantic	57
Indexing and Biological Named Entity Recognition	
3.3.1 Introduction	57
3.3.2 Methods and Dataset	57
3.3.3 Results and Discussion	61
3.4 Biological Named Entity Recognition using the Conditional	62
Random Fields Technique	
3.4.1 Introduction	62
3.4.2 Methods and Dataset	63
3.4.3 Results and Discussion	69
3.5 Comparison of Biological Named Entity Recognition Models	75
Performance	

Х

3.5.1 Introduction	75
3.5.2 Methods and Dataset	75
3.5.3 Results and Discussion	80
3.6 Conclusion	81
CHAPTER 4: KNOWLEDGE EXTRACTION AND FUNCTIONAL	83
GENOMICS KNOWLEDGE BASE	
4.1 Introduction	83
4.2 Lexical-Relation Extraction	85
4.2.1 Biological Information Extraction using Poisson	85
Collocations	
4.2.2 Detecting Biological Terms Collocations from Literature	95
with Biomedical Named Entity Recognition Models	
4.2.3 Comparison to the log-Likelihood Measure	99
4.2.4 Conclusion	101
4.3 Semantic-Relation Extraction	102
4.3.1 Semantic-relations extraction by Part of speech: Parse Tree	102
4.3.2 Bio-molecular event extraction	106
4.3.3 Bio-molecular Event Extraction from Biological	107
4.3.4 Conclusion	112
4.4 Establishing Functional Genomics Knowledge Base	113
CHAPTER 5. CONCLUSION	
REFERENCES	121
	141

xii	
APPENDICES	132
APPENDIX A SAMPLE OF NAMED ENTIES	133
RECOGNITION DATA SET	
APPENDIX B SAMPLE OF BIO-EVENT DATA SET	136
APPENDIX C PUBLICATIONS BY AUTHOR	140
CURRICULUM VITAE	141

ลิขสิทธิมหาวิทยาลัยเชียงไหม Copyright[©] by Chiang Mai University All rights reserved

LIST OF TABLES

Table	Page
1.1 Scope of platform in thesis	5
2.1 Information retrieval model performance which generated by general term	ns 27
and LSI	
2.2 The performance of classifications models with features selections	42
3.1 Performance of Models for Information Retrieval	61
3.2 Performance of biological named entity recognition based on the 1 st CRF	s 70
3.3 Performance of biological named entity recognition based on the 2 nd CRH	Fs 71
3.4 Comparison of biological named entity recognition performance with other	ers 73
existing models	
3.5 Comparison of Biomedical Named Entity Recognition Models Performan	.ce 81
4.1 Biological Terms Collocations based on Poisson Collocations	90
4.2 The significant values collocations from biological NER Models	98
4.3 The significant Values Collocations form Poisson and log-Likelihood	101
4.4 The meaning of predicate argument to generate CFG	110
A.1:Training Data Set	133
A.2:Test Data Set	134
B.1: SampleData Set (File.txt)	136
B.2: SampleData Set (.a1*)	137
B.3: SampleData Set (.a2*)	138
B.4: Sample Data Set in Tree Structures	139

LIST OF FIGURES

Figure	Page
1.1 Framework of the thesis	6
2.1 Number of MEDLINE-indexed articles published per year	10
2.2 The framework of effectiveness of general term in biological literature	14
2.3 Term-document matrix scaling by TF-IDF	17
2.4 Abstracts classification with general term by TF-IDF Scaling	18
2.5 The framework of biological information retrieval using LSI	19
2.6 Cosine measure of document similarity	24
2.7 The framework of bioinformatics-text classification using machine	28
learning techniques	
2.8 A Bayesian network for the Naive Bayes classifier under the Bernoulli	31
document-based event model.	
2.9 Alternative linear decision boundaries for a binary classification problem	36
2.10 Illustration of the optimal separating hyperplane and margin. Circled	37
points are support vectors	
3.1 Sentence with <dna>, <rna>, <protein>, <cell line="">, and <cell< td=""><td>45</td></cell<></cell></protein></rna></dna>	45
Type> tags generated by Biomedical Named Entities Recognition	
3.2 State transition of Hidden Markov Model (HMM)	49
3.3 Example of Named Entity Recognition of HMM	50
3.4 MEMM states are conditioned on the previous state and the observation	51
3.5 Graphical structure of the linear chain CRFs for sequence labeling	52
3.6 Framework of Biological Information Retrieval using Latent Semantic	58

Indexing and Biological Named Entity Recognition	
3.7 Graphical structures of chain-structured CRFs for sequences.	59
3.8 The framework of Biological Named Entity Recognition using the	64
Conditional Random Fields Technique	
3.9 Example of Biological Named Entity Recognition	76
4.1The framework of biological information extraction using Poisson	86
Collocations	
4.2 Biological Terms Collocations Network	93
4.3 The framework of detecting biological terms collocations from literature	96
with biological named entity recognition models	
4.4 Cyanobacteria's abstract in NCBI corpus	97
4.5 The framework of Semantic-relations extraction by Part of speech: Parse	103
Tree	
4.6 Biological Relation Extractions with Parse Tree	104
4.7 Parse tree of biological terms collocations	105
4.8 Information of Oxygen Evolving Complex (OEC) in GO Project Method	106
4.9 The framework of Bio-molecular event extraction from biological	108
literature using Tree Kernels method	
4.10 Example of bio-molecular event dataset	109
4.11 Bio-molecular event in Tree structure based on Context-Free grammar	109
4.12 Specific functional genomics knowledge base	113
4.13 Sample of knowledge which establishing in the graph network form	114

XV

ABBREVIATIONS AND SYMBOLS

CRFs	Conditional Random Fields
NER	Named Entity Recognition
LSI	Latent Semantic Indexing
SVMs	Support Vector Machines
NLP	Natural Language Processing
POS	Part of Speech
IG	Information Gain
MI	Mutual Information
НММ	Hidden Markov Model
MEMM	Maximum Entropy Markov Mode

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่ Copyright[©] by Chiang Mai University All rights reserved