ENHANCED TERM EXTRACTION BASED ON PROBABILISTIC ESTIMATION FROM SYNTACTIC PARSE TREES

ZHANG XING

DOCTOR OF PHILOSOPHY

CITY UNIVERSITY OF HONG KONG

MAY 2011

CITY UNIVERSITY OF HONG KONG
香港城市大學

Enhanced Term Extraction Based on Probabilistic Estimation from Syntactic Parse Trees
基於句法樹概率測算的術語優化提取

Submitted to
Department of Chinese, Translation and Linguistics
中文、翻譯及語言學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Zhang Xing
張杏

May 2011
二零一一年五月

Abstract

This research is an effort to explore how the syntactic information of term candidates can be exploited for the task of term extraction. It proposes an approach that represents a novel, linguistically motivated perspective in the area of terminological processing. The hypothesis of this work is that terms tend to perform certain types of syntactic functions more prominently than others. This syntactic behaviour of terms can be captured as termhood by estimating term probabilities from their occurrences in different syntactic paths. Based on a large corpus of parse trees, this feature allows for highly reliable statistics on properties of term occurrences. In essence, this method is a weighting scheme that measures probabilistic relations between term occurring patterns and syntactic paths, which is discussed in this thesis as Syntactic Function Value (SF-Value) and implemented in a term extraction system.

Experiments conducted in this study begin by building up an automatic term extraction system that integrates such a weighting scheme. The purpose of these experiments is not to design a term extraction system with the best performance but to investigate the contributions of syntactic information to term extraction, including single-word terms, multi-word terms, and new terms. Specifically, these experiments are aimed at answering several research questions, including the following: whether linguistic knowledge as term rates in syntactic paths is useful for recognising candidate terms in medical texts; to what extent single-word terms can be extracted by this linguistic indicator; and subsequently how this linguistically based metric can be used to improve the ranking of multi-word terms, and whether term rates in syntactic paths can be used effectively for new term extraction. Finally, with the aim of investigating whether this linguistic metric can be used as an effective feature within a machine learning framework, a series of experiments are conducted on general term extraction and new term extraction using the method of Conditional Random Fields (CRF).

Empirical results strongly argue that the term extraction approach proposed in this study demonstrates superior performance when compared with two existing term extractors. The key technique of this term extraction system, SF-Value, proves to be especially useful in selecting single-word terms and is also effective in enhancing the ranking of multi-word term candidates after their initial ranking by a statistical measure, C-Value. With regard to new term extraction, results show that SF-Value does not perform as well, which suggests that more features are needed to distinguish new terms from known terms. CRF framework is subsequently applied with the uses of SF-Value and term rate as added features for the extraction of new terms. Results show that this machine learning framework performs quite well in general term extraction. However, for the task of generating a list of new term candidates, this framework does not show good performance as expected. This result indicates that, for the task of new term extraction, more features related to new term candidates should be taken into consideration, in addition to syntactic function information.

In conclusion, this study devises an innovative, linguistically motivated measure for term extraction and implements it in a software system. Comprehensive experiments are conducted to evaluate its performance, and empirical results demonstrate its superior performance in comparison with existing term extraction systems.

Table of Contents

CHAPTER 1 INTRODUCTION 1

1.1 THEORETICAL FOUNDATION OF TERMINOLOGY 1

1.1.1 Definitions of Terms 1

1.1.2 Basic Concepts in Term Extraction 3

1.1.3 Term Formation 4

1.2 GOALS OF AUTOMATIC TERM EXTRACTION AND SCOPE OF THIS THESIS 5

1.3 CHAPTER OVERVIEW 7

CHAPTER 2 LITERATURE REVIEW 9

2.1 OVERVIEW 9

2.2 LINGUISTIC METHODS 9

2.2.1 Early ATE Systems 10

2.2.2 Term Variants 11

2.2.3 Etymology in Medical Vocabularies 13

2.2.4 Verb and Verb Syntax Information 15

2.3 STATISTICAL METHODS 16

2.3.1 TF-IDF 16

2.3.2 Mutual Information 17

2.3.3 Log-Likelihood Ratio 17

2.3.4 Dice Factor 18

2.3.5 A Combination of Perplexity and Term Frequency (FPP) 18

2.4 HYBRID TECHNIQUES 19

2.4.1 C-Value 19

2.4.2 NC-Value 20

2.4.3 SNC-Value 21

2.4.4 Combining Technique 22

2.5 SUMMARY 24

CHAPTER 3 RATIONALE, METHODOLOGY AND IMPLEMENTATION 25

3.1 OVERVIEW 25

3.2 SYNTACTIC FUNCTION VALUE 26

3.2.1 Introduction 26

3.2.2 The Algorithm of SF-Value 27

3.3 THE ICE-GB CORPUS 31

3.4 THE AUTASYS TAGGER 33
3.5 THE SURVEY PARSER 34
3.6 THE PUBMED CORPUS 37
3.6.1 Studies on MEDLINE and MeSH *38*
3.6.2 MeSH Term Files *40*
3.7 THE CREATION OF TERM LISTS FROM MESH 42
3.8 THE ANNOTATION OF MESH TERMS 43
3.9 SYNTACTIC PATHS 44
3.9.1 Rules for Recording Syntactic Paths *44*
3.9.2 Term Rate and Weight of Syntactic Path *47*
3.9.3 The Conflation of Sparse Syntactic Paths *47*
3.10 THE GENERATION OF AN NP LIST AND DEFINITION OF LINGUISTIC FILTERS 50
3.10.1 The Generation of an NP List from a Parse Tree *50*
3.10.2 Stop Word List *58*
3.10.3 Selective Deletion Strategy *59*
3.11 SFATE SYSTEM DESIGN 60
3.12 EVALUATION METHOD 61
3.13 SUMMARY 62

CHAPTER 4 EXPERIMENTS AND EVALUATION 64

4.1 OVERVIEW 64
4.2 EXPERIMENTAL SETUP 64
4.2.1 Corpus Acquisition and Partition *64*
4.2.2 Corpus Tagging, Parsing, and Term Annotation *66*
4.2.3 Basic Statistics for Training and Testing corpora *70*
4.3 GOLD STANDARD 71
4.4 SYNTACTIC PATHS OBTAINED FROM TRAINING 72
4.5 OVERALL SYSTEM PERFORMANCE 75
4.5.1 Introduction to Baseline Systems *75*
4.5.2 Performance Evaluation *77*
4.5.2.1 Comparison between SFATE performance before and after conflation 77
4.5.2.2 Performance of TerMine and TermExtractor 81
4.5.2.3 Performance Comparison within TerMine, TermExtractor and SFATE 83
4.5.3 Conclusion *87*
4.6 SINGLE-WORD TERM EXTRACTION 88
4.6.1 Introduction *88*
4.6.2 Related Studies on Single-word Term Extraction *89*
4.6.3 Performance Evaluation *90*

4.6.4 Discussion and Conclusion *95*
4.7 MULTI-WORD TERM EXTRACTION 100
4.7.1 Related Studies on Multi-word Term Extraction *100*
4.7.2 Performance Evaluation *104*
4.7.2.1 Performances of C-Value 104
4.7.2.2 Performances of SF-Value 105
4.7.2.3 Performances of Combining C-Value with SF-Value 107
4.7.2.4 Performance Comparison within C-Value, SF-Value, and Combine Measure 110
4.7.2.4.1 Comparison of F-scores between C-Value and SF-Value Before Conflation (A with B) 112
4.7.2.4.2 Comparison of F-scores between C-Value and SF-Value After Conflation (A with C) 112
4.7.2.4.3 Comparison of F-scores between SF-Value Before Conflation and SF-Value After Conflation (B with C) 113
4.7.2.4.4 Comparison of F-scores between Combined Measure Before Conflation and Combined Measure After Conflation (D with E) 113
4.7.2.4.5 Comparison of F-scores between C-Value and Combined Measure Before Conflation (A with D) 114
4.7.2.4.6 Comparison of F-scores between C-Value and Combined Measure After Conflation (A with E) 115
4.7.2.4.7 Comparison of F-scores between Combined Measure Before Conflation and SF-Value Before Conflation (B with D) 115
4.7.2.4.8 Comparison of F-scores between SF-Value After Conflation and Combined Measure After Conflation (C with E) 116
4.7.3 Discussion and Conclusion *117*
4.8 SUMMARY 119

CHAPTER 5 THE EXTRACTION OF NEW TERMS 120

5.1 OVERVIEW 120
5.1.1 Introduction *120*
5.1.2 The Application of the TF-IDF Metric *121*
5.2 RELATED STUDIES ON THE EXTRACTION OF NEW TERMS 122
5.3 RESOURCE CONSTRUCTION 126
5.4 EXPERIMENTS 127
5.4.1 SF-Value Comparison *129*
5.4.2 TF Value Comparison *130*
5.4.3 IDF Value Comparison *131*
5.4.4 TF-IDF Value Comparison *132*
5.5 EVALUATION 134
5.6 CONCLUSION 136

5.7 SUMMARY 138

CHAPTER 6 APPLICATION OF CONDITIONAL RANDOM FIELDS TO TERM EXTRACTION 139

6.1 INTRODUCTION 139
6.2 RELATED STUDIES ON THE APPLICATION OF CRF TO ATE 140
6.3 CRF MODEL 143
6.4 EXPERIMENTS AND EVALUATION 144
6.4.1 Resource Construction 144
6.4.2 Conversion from Parse Trees to Matrices 146
6.4.3 Evaluation Method 149
6.4.4 Experiments on Training Corpus Size 149
6.4.5 Experiments of 10-fold Cross Validation 153
6.4.6 Comparisons with Other Related Studies 156
6.5 CONCLUSION 158
6.6 SUMMARY 159

CHAPTER 7 CONCLUSION AND FUTURE WORK 161

7.1 RESEARCH CONTRIBUTIONS 161
7.2 FUTURE RESEARCH 163

BIBLIOGRAPHY 165

APPENDIX 1 176

APPENDIX 2 178

APPENDIX 3 180

APPENDIX 4 181

APPENDIX 5 191

APPENDIX 6 197

APPENDIX 7 200

List of Tables

TABLE 1.WEIGHTS OF DIFFERENT CATEGORIES....21
TABLE 2.SYNTACTIC PATHS AND THEIR *FSPS* AND *WSPS*....27
TABLE 3.SUMMARY OF THREE TERM LISTS....43
TABLE 4.TWO TERMS AND THEIR SYNTACTIC PATHS....45
TABLE 5.CONFLATION OF SYNTACTIC PATHS....48
TABLE 6.EFFECTS OF CONFLATION ON SF-VALUES....49
TABLE 7.SYNTACTIC FUNCTIONS AND TERM OCCURRENCE FREQUENCIES IN BASE CORPUS....68
TABLE 8.BASIC STATISTICS FOR TRAINING CORPORA....71
TABLE 9.BASIC STATISTICS FOR TESTING CORPORA....71
TABLE 10.THE NUMBER OF SYNTACTIC PATHS BEFORE CONFLATION AND AFTER CONFLATION....72
TABLE 11.TOP TEN SYNTACTIC PATHS RANKED BY TERM RATES BEFORE CONFLATION....73
TABLE 12.TOP TEN SYNTACTIC PATHS RANKED BY TERM RATES AFTER CONFLATION....74
TABLE 13.THE NUMBERS OF TERMS AND MESH TERMS EXTRACTED BY SFATE BEFORE CONFLATION FROM THE TEN TESTING SUB-CORPORA....78
TABLE 14.PERFORMANCES OF SFATE BEFORE CONFLATION ON TEN TESTING SUB-CORPORA....78
TABLE 15.THE NUMBERS OF TERMS AND MESH TERMS EXTRACTED BY SFATE AFTER CONFLATION FROM THE TEN TESTING SUB-CORPORA....78
TABLE 16.PERFORMANCES OF SFATE AFTER CONFLATION ON THE TEN TESTING SUB-CORPORA....79
TABLE 17.PAIRED SAMPLES STATISTICS FOR F-SCORES BEFORE CONFLATION AND AFTER CONFLATION....80
TABLE 18.PAIRED SAMPLES CORRELATIONS OF F-SCORES BEFORE CONFLATION AND AFTER CONFLATION....80
TABLE 19.PAIRED SAMPLES TEST OF F-SCORES BEFORE CONFLATION AND AFTER CONFLATION....80
TABLE 20.THE NUMBERS OF TERMS AND MESH TERMS EXTRACTED BY TERMINE FROM THE TEN TESTING SUB-CORPORA....81
TABLE 21.PERFORMANCES OF TERMINE ON THE TEN TESTING SUB-CORPORA....81
TABLE 22.THE NUMBERS OF TERMS AND MESH TERMS EXTRACTED BY TERMEXTRACTOR FROM THE TEN TESTING SUB-CORPORA....82
TABLE 23.PERFORMANCES OF TERMEXTRACTOR ON THE TEN TESTING SUB-CORPORA....82
TABLE 24.F-SCORE COMPARISON WITHIN TERMINE, TERMEXTRACTOR, AND SFATE....83
TABLE 25.ANOVA....84
TABLE 26.POST HOC TESTS OF ONE WAY ANOVA....84
TABLE 27.HOMOGENEOUS SUB-CORPORA ON F-SCORES....85
TABLE 28.THE NUMBERS OF TERMS EXTRACTED BY SFATE, TERMINE, AND TERMEXTRACTOR FROM THE TEN TESTING SUB-CORPORA....86

TABLE 29.ONE-WAY ANOVA TEST BETWEEN THE NUMBERS OF TERMS EXTRACTED BY SFATE, TERMINE, AND TERMEXTRACTOR 86
TABLE 30.DISCRIPTIVE DATA ABOUT MEANS OF THE NUMBERS OF TERMS EXTRACTED BY SFATE, TERMINE, AND TERMEXTRACTOR 87
TABLE 31.THE NUMBERS OF SINGLE-WORD TERMS AND SINGLE-WORD MESH TERMS EXTRACTED BY SFATE BEFORE CONFLATION FROM THE TEN TESTING SUB-CORPORA 91
TABLE 32.THE NUMBERS OF SINGLE-WORD TERMS AND SINGLE-WORD MESH TERMS EXTRACTED BY SFATE AFTER CONFLATION FROM THE TEN TESTING SUB-CORPORA 91
TABLE 33.PAIRED SAMPLES STATISTICS FOR RECALL BC AND RECALL AC 93
TABLE 34.PAIRED SAMPLES CORRELATIONS OF RECALL BC AND RECALL AC 93
TABLE 35.PAIRED SAMPLES TEST OF RECALL BC AND RECALL AC 93
TABLE 36.SINGLE-WORD MESH TERMS EXTRACTED BY TERMEXTRACTOR FROM THE TEN TESTING SUB-CORPORA 94
TABLE 37.SINGLE-WORD MESH TERMS EXTRACTED BY SFATE AND TERMEXTRACTOR FROM THE TEN TESTING SUB-CORPORA 95
TABLE 38.RECALL RATES OF SINGLE-WORD MESH TERMS OF TERMEXTRACTOR AND SFATE....95
TABLE 39.MESH TERMS SELECTED BY SF-VALUE AT THRESHOLD OF 0.02 96
TABLE 40.THE NUMBERS OF MULTI-WORD TERMS AND MULTI-WORD MESH TERMS EXTRACTED BY C-VALUE FROM THE TEN TESTING SUB-CORPORA 104
TABLE 41.C-VALUE PERFORMANCES ON THE TEN TESTING SUB-CORPORA 105
TABLE 42.DESCRIPTIVE STATISTICS FOR C-VALUE PERFORMANCES....105
TABLE 43.THE NUMBERS OF MULTI-WORD TERMS AND MULTI-WORD MESH TERMS EXTRACTED BY SF-VALUE BEFORE CONFLATION FROM THE TEN TESTING SUB-CORPORA 105
TABLE 44.SF-VALUE PERFORMANCES BEFORE CONFLATION ON THE TEN TESTING SUB-CORPORA 106
TABLE 45.DESCRIPTIVE STATISTICS FOR SF-VALUE PERFORMANCES BEFORE CONFLATION....106
TABLE 46.THE NUMBERS OF MULTI-WORD TERMS AND MULTI-WORD MESH TERMS EXTRACTED BY SF-VALUE AFTER CONFLATION FROM THE TEN TESTING SUB-CORPORA 106
TABLE 47.SF-VALUE PERFORMANCES AFTER CONFLATION ON THE TEN TESTING SUB-CORPORA107
TABLE 48.DESCRIPTIVE STATISTICS FOR SF-VALUE PERFORMANCES AFTER CONFLATION 107
TABLE 49.THE NUMBERS OF MULTI-WORD TERMS AND MULTI-WORD MESH TERMS EXTRACTED BY COMBINED BC FROM THE TEN TESTING SUB-CORPORA....108
TABLE 50.PERFORMANCES OF COMBINED BC ON THE TEN TESTING SUB-CORPORA....108
TABLE 51.DESCRIPTIVE STATISTICS FOR PERFORMANCES OF COMBINED BC....109
TABLE 52.THE NUMBERS OF MULTI-WORD TERMS AND MULTI-WORD MESH TERMS EXTRACTED BY COMBINED AC FROM THE TEN TESTING SUB-CORPORA....109
TABLE 53.PERFORMANCES OF COMBINED AC ON THE TEN TESTING SUB-CORPORA 109

TABLE 54.DESCRIPTIVE STATISTICS FOR PERFORMANCES OF COMBINED AC 110
TABLE 55.F-SCORES OF C-VALUE, SF-VALUE BC, SF-VALUE AC, COMBINED MEASURE BC, COMBINED AC.....110
TABLE 56.MEAN F-SCORES OF THESE FIVE MEASURES 117
TABLE 57.NEW MESH TERMS IN 20 TESTING CORPORA.....128
TABLE 58.PAIRED SAMPLES STATISTICS FOR THE SF-VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....129
TABLE 59.PAIRED SAMPLES CORRELATIONS OF THE SF-VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....129
TABLE 60.PAIRED SAMPLES TEST OF THE SF-VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....130
TABLE 61.PAIRED SAMPLES STATISTICS FOR TF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....131
TABLE 62.PAIRED SAMPLES CORRELATIONS OF TF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....131
TABLE 63.PAIRED SAMPLES TEST OF TF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....131
TABLE 64.PAIRED SAMPLES STATISTICS FOR IDF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....132
TABLE 65.PAIRED SAMPLES CORRELATIONS OF IDF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....132
TABLE 66.PAIRED SAMPLES TEST OF IDF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....132
TABLE 67.PAIRED SAMPLES STATISTICS OF TF-IDF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....133
TABLE 68.PAIRED SAMPLES CORRELATIONS OF TF-IDF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....133
TABLE 69.PAIRED SAMPLES TEST OF TF-IDF VALUES OF THE 30 NEW MESH TERMS IN THE TESTING CORPORA OF 2008 AND 2009.....133
TABLE 70. SUMMARY TABLE OF THE TEN TESTING CORPORA.....135
TABLE 71. FEATURE TEMPLATES USED IN CRF BASED TERM TAGGING.....144
TABLE 72.BASIC STATISTICS FOR EIGHT TRAINING CORPORA OF DIFFERENT SIZES.....145
TABLE 73.BASIC STATISTICS FOR TEN TRAINING CORPORA FOR 10-FOLD CROSS VALIDATION 146
TABLE 74.MATRIX PRESENTATION OF THE SAMPLE SENTENCE 150
TABLE 76.PERFORMANCES ON GENERAL TERMS 151
TABLE 77.PERFORMANCES ON NEW TERMS.....151
TABLE 78.RESULTS OF 10-FOLD CROSS VALIDATION ON GENERAL TERMS.....153

TABLE 79.DESCRIPTIVE STATISTICS FOR RESULTS OF 10-FOLD CROSS VALIDATION ON GENERAL TERMS 154
TABLE 80. RESULTS OF 10-FOLD CROSS VALIDATION ON NEW TERMS 154
TABLE 81.DESCRIPTIVE STATISTICS FOR RESULTS OF 10-FOLD CROSS VALIDATION ON NEW TERMS 155
TABLE 82. PERFORMANCE COMPARSION WITHIN RELATED STUDIES 157
TABLE 83.PHRASAL FUNCTIONS DEFINED BY THE SURVEY PARSER 176
TABLE 84.CLAUSAL FUNCTIONS DEFINED BY THE SURVEY PARSER 177
TABLE 85.WORDS IN STOP WORD LIST 178
TABLE 86.WORDS USED FOR SELECTIVE SELECTION 180
TABLE 87.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 1 181
TABLE 88.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 2 182
TABLE 89.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 3 183
TABLE 90.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 4 184
TABLE 91.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 5 185
TABLE 92.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 6 186
TABLE 93.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 7 187
TABLE 94.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 8 188
TABLE 95.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 9 189
TABLE 96.SYNTACTIC PATHS IN TRAINING SUB-CORPUS 10 190
TABLE 97. PAIRED SAMPLES STATISTICS FOR C-VALUE AND SF-VALUE BEFORE CONFLATION 191
TABLE 98. PAIRED SAMPLES CORRELATIONS OF C-VALUE AND SF-VALUE BEFORE CONFLATION 191
TABLE 99. PAIRED SAMPLES TEST OF C-VALUE AND SF-VALUE BEFORE CONFLATION 191
TABLE 100. PAIRED SAMPLES STATISTICS FOR C-VALUE AND SF-VALUE AFTER CONFLATION 191
TABLE 101. PAIRED SAMPLES CORRELATIONS OF C-VALUE AND SF-VALUE AFTER CONFLATION 192
TABLE 102. PAIRED SAMPLES TEST OF C-VALUE AND SF-VALUE AFTER CONFLATION 192
TABLE 103. PAIRED SAMPLES STATISTICS FOR SF-VALUE BEFORE CONFLATION AND AFTER CONFLATION 192
TABLE 104. PAIRED SAMPLES CORRELATIONS OF SF-VALUE BEFORE CONFLATION AND AFTER CONFLATION 192
TABLE 105. PAIRED SAMPLES TEST OF SF-VALUE BEFORE CONFLATION AND AFTER CONFLATION 193
TABLE 106. PAIRED SAMPLES STATISTICS FOR COMBINED MEASURE BEFORE AND AFTER CONFLATION 193

TABLE 107. PAIRED SAMPLES CORRELATIONS OF COMBINED MEASURE BEFORE AND AFTER CONFLATION 193
TABLE 108. PAIRED SAMPLES TEST OF COMBINED MEASURE BEFORE AND AFTER CONFLATION .193
TABLE 109. PAIRED SAMPLES STATISTICS FOR C-VALUE AND COMBINED MEASURE BEFORE CONFLATION 194
TABLE 110. PAIRED SAMPLES CORRELATIONS OF C-VALUE AND COMBINED MEASURE BEFORE CONFLATION 194
TABLE 111. PAIRED SAMPLES TEST OF C-VALUE AND COMBINED MEASURE BEFORE CONFLATION 194
TABLE 112. PAIRED SAMPLES STATISTICS FOR C-VALUE AND COMBINED MEASURE AFTER CONFLATION 194
TABLE 113. PAIRED SAMPLES CORRELATIONS OF C-VALUE AND COMBINED MEASURE AFTER CONFLATION 194
TABLE 114. PAIRED SAMPLES TEST OF C-VALUE AND COMBINED MEASURE AFTER CONFLATION 195
TABLE 115. PAIRED SAMPLES STATISTICS FOR SF-VALUE BEFORE CONFLATION AND COMBINED MEASURE BEFORE CONFLATION 195
TABLE 116. PAIRED SAMPLES CORRELATIONS OF SF-VALUE BEFORE CONFLATION AND COMBINED MEASURE BEFORE CONFLATION 195
TABLE 117. PAIRED SAMPLES TEST OF SF-VALUE BEFORE CONFLATION AND COMBINED MEASURE BEFORE CONFLATION 195
TABLE 118. PAIRED SAMPLES STATISTICS FOR SF-VALUE AFTER CONFLATION AND COMBINED MEASURE AFTER CONFLATION 196
TABLE 119. PAIRED SAMPLES CORRELATIONS OF SF-VALUE AFTER CONFLATION AND COMBINED MEASURE AFTER CONFLATION 196
TABLE 120. PAIRED SAMPLES TEST OF SF-VALUE AFTER CONFLATION AND COMBINED MEASURE AFTER CONFLATION 196
TABLE 121.36 MESH TERMS SELECTED BY C-VALUE AT THE THRESHOLD OF 1.34 (RANKED BY C-VALUE).....197
TABLE 122.52 MESH TERMS SELECTED BY COMBINED MEASURE AFTER CONFLATION AT THE THRESHOLDS OF C=0.5, SF-VALUE=0.01(RANKED BY SF-VALUE) 198
TABLE 123.TOP TEN MESH TERMS IN THE NEW TERM LIST OF 2008 CORPORA AND THAT OF 2009 CORPORA.....200

List of Figures

FIGURE 1.RELATIONS BETWEEN OBJECT, CONCEPT AND TERM....2
FIGURE 2.AN EXAMPLE OF AN OBJECT AND ITS TERM REPRESENTATION....3
FIGURE 3.SYNTACTIC PATH 1 OF *PATHOGEN*....29
FIGURE 4.SYNTACTIC PATH 2 OF *PATHOGEN*....29
FIGURE 5.SYNTACTIC PATH 3 OF *PATHOGEN*....30
FIGURE 6.TREE STRUCTURE FOR EXAMPLE 6 (FANG, 2000)....36
FIGURE 7.A MULTI-WORD TERM CONSISTING OF COMPONENTS IN THE SAME NODE....43
FIGURE 8.A MULTI-WORD TERM CONSISTING OF COMPONENTS ACROSS TWO NODES....44
FIGURE 9.THE PARSE TREE OF EXAMPLE 7....46
FIGURE 10.THE PARSE TREE OF EXAMPLE 9....53
FIGURE 11.AN N NODE CONSISTING OF THREE COMPONENTS....56
FIGURE 12.AN NP NODE CONSISTING OF TWO NPPR COMPONENTS....57
FIGURE 13.AN NP NODE CONSISTING OF TWO CONJOINED COMPONENTS....58
FIGURE 14.PROCESS FLOW CHART OF SFATE....60
FIGURE 15. THE PARSE TREE OF EXAMPLE 12 (ANNOTATED WITH MESH TERMS)....67
FIGURE 16.TERM OCCURING IN PP FUNTIONING AS A....68
FIGURE 17.TERM OCCURING IN VP FUNCTIONING AS VB (CASE 1)....69
FIGURE 18.TERM OCCURING IN VP FUNCTIONING AS VB (CASE 2)....69
FIGURE 19.TERMS OCCURING IN AJPS FUNCTIONING AS CS....70
FIGURE 20.PERFORMANCE COMPARISON BETWEEN TERMINE, TERMEXTRACTOR, AND SFATE....85
FIGURE 21.RECALL OF SINGLE-WORD TERMS BEFORE CONFLATION AND AFTER CONFLATION FROM THE TEN TESTING SUB-CORPORA....92
FIGURE 22.SYNTACTIC PATH 1 OF NP *PATIENTS*....97
FIGURE 23.SYNTACTIC PATH 2 OF NP *PATIENTS*....98
FIGURE 24.SYNTACTIC PATH 32 OF NP *PATIENTS*....99
FIGURE 25.F-SCORES OF C-VALUE, SF-VALUE BC, SF-VALUE AC, COMBINED MEASURE BC, COMBINED AC....111
FIGURE 26. THE PARSE TREE OF THE SAMPLE SENTENCE....147
FIGURE 27.PERFORMANCES ON GENERAL TERMS....152
FIGURE 28.PERFORMANCES ON NEW TERMS....152
FIGURE 29.RESULTS OF 10-FOLD CROSS VALIDATION ON GENERAL TERMS....155
FIGURE 30.RESULTS OF 10-FOLD CROSS VALIDATION ON NEW TERMS....156