

# Contents

## Part I Natural Language Processing Core-Technologies

<b>1</b>	<b>Linguistic Introduction: The Orthography, Morphology and Syntax of Semitic Languages .....</b>	<b>3</b>
	Ray Fabri, Michael Gasser, Nizar Habash, George Kiraz, and Shuly Wintner	
1.1	Introduction .....	3
1.2	Amharic .....	5
1.2.1	Orthography .....	6
1.2.2	Derivational Morphology .....	7
1.2.3	Inflectional Morphology .....	9
1.2.4	Basic Syntactic Structure .....	11
1.3	Arabic .....	13
1.3.1	Orthography .....	14
1.3.2	Morphology .....	15
1.3.3	Basic Syntactic Structure .....	18
1.4	Hebrew .....	19
1.4.1	Orthography .....	20
1.4.2	Derivational Morphology .....	22
1.4.3	Inflectional Morphology .....	23
1.4.4	Morphological Ambiguity .....	25
1.4.5	Basic Syntactic Structure .....	25
1.5	Maltese .....	26
1.5.1	Orthography .....	26
1.5.2	Derivational Morphology .....	27
1.5.3	Inflectional Morphology .....	29
1.5.4	Basic Syntactic Structure .....	30
1.6	Syriac .....	32
1.6.1	Orthography .....	32
1.6.2	Derivational Morphology .....	33

1.6.3	Inflectional Morphology.....	33
1.6.4	Syntax .....	34
1.7	Contrastive Analysis .....	34
1.7.1	Orthography.....	34
1.7.2	Phonology.....	35
1.7.3	Morphology.....	36
1.7.4	Syntax .....	37
1.7.5	Lexicon.....	37
1.8	Conclusion .....	38
	References.....	38
<b>2</b>	<b>Morphological Processing of Semitic Languages .....</b>	<b>43</b>
	Shuly Wintner	
2.1	Introduction.....	43
2.2	Basic Notions.....	44
2.3	The Challenges of Morphological Processing .....	45
2.4	Computational Approaches to Morphology.....	47
2.4.1	Two-Level Morphology .....	48
2.4.2	Multi-tape Automata .....	48
2.4.3	The Xerox Approach .....	49
2.4.4	Registered Automata .....	50
2.4.5	Analysis by Generation.....	50
2.4.6	Functional Morphology .....	51
2.5	Morphological Analysis and Generation of Semitic Languages ...	51
2.5.1	Amharic .....	52
2.5.2	Arabic .....	52
2.5.3	Hebrew .....	54
2.5.4	Other Languages.....	55
2.5.5	Related Applications .....	55
2.6	Morphological Disambiguation of Semitic Languages.....	56
2.7	Future Directions .....	58
	References.....	58
<b>3</b>	<b>Syntax and Parsing of Semitic Languages .....</b>	<b>67</b>
	Reut Tsarfaty	
3.1	Introduction.....	67
3.1.1	Parsing Systems .....	69
3.1.2	Semitic Languages.....	74
3.1.3	The Main Challenges .....	80
3.1.4	Summary and Conclusion .....	84
3.2	Case Study: Generative Probabilistic Parsing.....	84
3.2.1	Formal Preliminaries .....	85
3.2.2	An Architecture for Parsing Semitic Languages .....	91
3.2.3	The Syntactic Model .....	99
3.2.4	The Lexical Model .....	113

3.3	Empirical Results .....	117
3.3.1	Parsing Modern Standard Arabic .....	117
3.3.2	Parsing Modern Hebrew .....	120
3.4	Conclusion and Future Work .....	123
	References .....	124
<b>4</b>	<b>Semantic Processing of Semitic Languages .....</b>	<b>129</b>
	Mona Diab and Yuval Marton	
4.1	Introduction .....	129
4.2	Fundamentals of Semitic Language Meaning Units .....	130
4.2.1	Morpho-Semantics: A Primer .....	130
4.3	Meaning, Semantic Distance, Paraphrasing and Lexicon Generation .....	135
4.3.1	Semantic Distance .....	136
4.3.2	Textual Entailment .....	138
4.3.3	Lexicon Creation .....	138
4.4	Word Sense Disambiguation and Meaning Induction .....	139
4.4.1	WSD Approaches in Semitic Languages .....	140
4.4.2	WSI in Semitic Languages .....	141
4.5	Multiword Expression Detection and Classification .....	142
4.5.1	Approaches to Semitic MWE Processing and Resources .....	143
4.6	Predicate–Argument Analysis .....	145
4.6.1	Arabic Annotated Resources .....	146
4.6.2	Systems for Semantic Role Labeling .....	148
4.7	Conclusion .....	152
	References .....	152
<b>5</b>	<b>Language Modeling .....</b>	<b>161</b>
	Ilana Heintz	
5.1	Introduction .....	161
5.2	Evaluating Language Models with Perplexity .....	162
5.3	N-Gram Language Modeling .....	164
5.4	Smoothing: Discounting, Backoff, and Interpolation .....	166
5.4.1	Discounting .....	166
5.4.2	Combining Discounting with Backoff .....	168
5.4.3	Interpolation .....	168
5.5	Extensions to N-Gram Language Modeling .....	170
5.5.1	Skip N-Grams and FlexGrams .....	170
5.5.2	Variable-Length Language Models .....	171
5.5.3	Class-Based Language Models .....	173
5.5.4	Factored Language Models .....	174
5.5.5	Neural Network Language Models .....	175
5.5.6	Syntactic or Structured Language Models .....	177
5.5.7	Tree-Based Language Models .....	178
5.5.8	Maximum-Entropy Language Models .....	178

5.5.9	Discriminative Language Models .....	180
5.5.10	LSA Language Models .....	183
5.5.11	Bayesian Language Models .....	184
5.6	Modeling Semitic Languages .....	187
5.6.1	Arabic .....	188
5.6.2	Amharic .....	189
5.6.3	Hebrew .....	191
5.6.4	Maltese .....	191
5.6.5	Syriac .....	192
5.6.6	Other Morphologically Rich Languages .....	192
5.7	Summary .....	193
	References .....	193

## Part II Natural Language Processing Applications

<b>6</b>	<b>Statistical Machine Translation .....</b>	<b>199</b>
	Hany Hassan and Kareem Darwish	
6.1	Introduction .....	199
6.2	Machine Translation Approaches .....	200
6.2.1	Machine Translation Paradigms .....	200
6.2.2	Rule-Based Machine Translation .....	202
6.2.3	Example-Based Machine Translation .....	202
6.2.4	Statistical Machine Translation .....	203
6.2.5	Machine Translation for Semitic Languages .....	203
6.3	Overview of Statistical Machine Translation .....	204
6.3.1	Word-Based Translation Models .....	204
6.3.2	Phrase-Based SMT .....	205
6.3.3	Phrase Extraction Techniques .....	206
6.3.4	SMT Reordering .....	207
6.3.5	Language Modeling .....	207
6.3.6	SMT Decoding .....	208
6.4	Machine Translation Evaluation Metrics .....	209
6.5	Machine Translation for Semitic Languages .....	210
6.5.1	Word Segmentation .....	210
6.5.2	Word Alignment and Reordering .....	211
6.5.3	Gender-Number Agreement .....	212
6.6	Building Phrase-Based SMT Systems .....	213
6.6.1	Data .....	213
6.6.2	Parallel Data .....	213
6.6.3	Monolingual Data .....	214
6.7	SMT Software Resources .....	214
6.7.1	SMT Moses Framework .....	214
6.7.2	Language Modeling Toolkits .....	214
6.7.3	Morphological Analysis .....	215

6.8	Building a Phrase-Based SMT System: Step-by-Step Guide .....	215
6.8.1	Machine Preparation.....	215
6.8.2	Data .....	216
6.8.3	Data Preprocessing .....	216
6.8.4	Words Segmentation.....	216
6.8.5	Language Model.....	217
6.8.6	Translation Model .....	217
6.8.7	Parameter Tuning .....	217
6.8.8	System Decoding .....	218
6.9	Summary.....	218
	References.....	218
<b>7</b>	<b>Named Entity Recognition.....</b>	<b>221</b>
	Behrang Mohit	
7.1	Introduction.....	221
7.2	The Named Entity Recognition Task .....	222
7.2.1	Definition .....	222
7.2.2	Challenges in Named Entity Recognition .....	223
7.2.3	Rule-Based Named Entity Recognition .....	224
7.2.4	Statistical Named Entity Recognition .....	225
7.2.5	Hybrid Systems .....	228
7.2.6	Evaluation and Shared Tasks.....	228
7.2.7	Evaluation Campaigns.....	229
7.2.8	Beyond Traditional Named Entity Recognition .....	230
7.3	Named Entity Recognition for Semitic Languages .....	230
7.3.1	Challenges in Semitic Named Entity Recognition .....	231
7.3.2	Approaches to Semitic Named Entity Recognition .....	232
7.4	Case Studies .....	233
7.4.1	Learning Algorithms .....	234
7.4.2	Features .....	234
7.4.3	Experiments.....	235
7.5	Relevant Problems .....	236
7.5.1	Named Entity Translation and Transliteration .....	236
7.5.2	Entity Detection and Tracking .....	238
7.5.3	Projection .....	238
7.6	Labeled Named Entity Recognition Corpora .....	239
7.7	Future Challenges and Opportunities.....	240
7.8	Summary.....	241
	References.....	241
<b>8</b>	<b>Anaphora Resolution.....</b>	<b>247</b>
	Khadiga Mahmoud Seddik and Ali Farghaly	
8.1	Introduction: Anaphora and Anaphora Resolution .....	247
8.2	Types of Anaphora .....	248
8.2.1	Pronominal Anaphora .....	248
8.2.2	Lexical Anaphora.....	249
8.2.3	Comparative Anaphora .....	249

8.3	Determinants in Anaphora Resolution .....	249
8.3.1	Eliminating Factors .....	250
8.3.2	Preferential Factors .....	251
8.3.3	Implementing Features in AR (Anaphora Resolution) Systems .....	252
8.4	The Process of Anaphora Resolution .....	256
8.5	Different Approaches to Anaphora Resolution .....	257
8.5.1	Knowledge-Intensive Versus Knowledge-Poor Approaches .....	257
8.5.2	Traditional Approach .....	259
8.5.3	Statistical Approach .....	259
8.5.4	Linguistic Approach to Anaphora Resolution .....	260
8.6	Recent Work in Anaphora and Coreference Resolution .....	262
8.6.1	Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree [24] .....	262
8.6.2	A Twin-Candidate Model for Learning-Based Anaphora Resolution [47, 48] .....	263
8.6.3	Improving Machine Learning Approaches to Coreference Resolution [36] .....	264
8.7	Evaluation of Anaphora Resolution Systems .....	265
8.7.1	MUC [45] .....	265
8.7.2	B-Cube [2] .....	267
8.7.3	ACE (NIST 2003) .....	267
8.7.4	CEAF [23] .....	268
8.7.5	BLANC [40] .....	269
8.8	Anaphora in Semitic Languages .....	269
8.8.1	Anaphora Resolution in Arabic .....	270
8.9	Difficulties with AR in Semitic Languages .....	272
8.9.1	The Morphology of the Language .....	272
8.9.2	Complex Sentence Structure .....	273
8.9.3	Hidden Antecedents .....	273
8.9.4	The Lack of Corpora Annotated with Anaphoric Links .....	273
8.10	Summary .....	274
	References .....	274
<b>9</b>	<b>Relation Extraction</b> .....	<b>279</b>
	Vittorio Castelli and Imed Zitouni	
9.1	Introduction .....	279
9.2	Relations .....	280
9.3	Approaches to Relation Extraction .....	281
9.3.1	Feature-Based Classifiers .....	281
9.3.2	Kernel-Based Methods .....	285
9.3.3	Semi-supervised and Adaptive Learning .....	288

9.4	Language-Specific Issues .....	291
9.5	Data .....	292
9.6	Results .....	294
9.7	Summary.....	295
	References.....	295
<b>10</b>	<b>Information Retrieval .....</b>	<b>299</b>
	Kareem Darwish	
10.1	Introduction.....	299
10.2	The Information Retrieval Task.....	299
10.2.1	Task Definition.....	301
10.2.2	The General Architecture of an IR System .....	302
10.2.3	Retrieval Models.....	303
10.2.4	IR Evaluation .....	305
10.3	Semitic Language Retrieval.....	309
10.3.1	The Major Known Challenges .....	309
10.3.2	Survey of Existing Literature .....	313
10.3.3	Best Arabic Index Terms .....	316
10.3.4	Best Hebrew Index Terms.....	318
10.3.5	Best Amharic Index Terms .....	318
10.4	Available IR Test Collections .....	318
10.4.1	Arabic .....	318
10.4.2	Hebrew .....	319
10.4.3	Amharic .....	319
10.5	Domain-Specific IR .....	319
10.5.1	Arabic–English CLIR .....	320
10.5.2	Arabic OCR Text Retrieval .....	322
10.5.3	Arabic Social Search .....	326
10.5.4	Arabic Web Search .....	328
10.6	Summary.....	329
	References.....	329
<b>11</b>	<b>Question Answering .....</b>	<b>335</b>
	Yassine Benajiba, Paolo Rosso, Lahsen Abouenour, Omar	
	Trigui, Karim Bouzoubaa, and Lamia Belguith	
11.1	Introduction.....	335
11.2	The Question Answering Task.....	336
11.2.1	Task Definition.....	336
11.2.2	The Major Known Challenges .....	338
11.2.3	The General Architecture of a QA System .....	339
11.2.4	Answering Definition Questions and Query Expansion Techniques.....	341
11.2.5	How to Benchmark QA System Performance: Evaluation Measure for QA .....	343

11.3	The Case of Semitic Languages .....	344
11.3.1	NLP for Semitic Languages .....	344
11.3.2	QA for Semitic Languages .....	345
11.4	Building Arabic QA Specific Modules .....	347
11.4.1	Answering Definition Questions in Arabic .....	347
11.4.2	Query Expansion for Arabic QA .....	353
11.5	Summary .....	366
	References .....	367
<b>12</b>	<b>Automatic Summarization .....</b>	<b>371</b>
	Lamia Hadrich Belguith, Mariem Ellouze, Mohamed Hedi Maaloul, Maher Jaoua, Fatma Kallel Jaoua, and Philippe Blache	
12.1	Introduction .....	371
12.2	Text Summarization Aspects .....	372
12.2.1	Types of Summaries .....	374
12.2.2	Extraction vs. Abstraction .....	375
12.2.3	The Major Known Challenges .....	376
12.3	How to Evaluate Summarization Systems .....	376
12.3.1	Insights from the Evaluation Campaigns .....	377
12.3.2	Evaluation Measures for Summarization .....	377
12.4	Single Document Summarization Approaches .....	378
12.4.1	Numerical Approach .....	379
12.4.2	Symbolic Approach .....	379
12.4.3	Hybrid Approach .....	380
12.5	Multiple Document Summarization Approaches .....	380
12.5.1	Numerical Approach .....	381
12.5.2	Symbolic Approach .....	382
12.5.3	Hybrid Approach .....	383
12.6	Case of Semitic Languages .....	385
12.6.1	Language-Independent Systems .....	385
12.6.2	Arabic Systems .....	386
12.6.3	Hebrew Systems .....	388
12.6.4	Maltese Systems .....	388
12.6.5	Amharic Systems .....	389
12.7	Case Study: Building an Arabic Summarization System (L.A.E) .....	389
12.7.1	L.A.E System Architecture .....	390
12.7.2	Source Text Segmentation .....	390
12.7.3	Interface .....	400
12.7.4	Evaluation and Discussion .....	401
12.8	Summary .....	402
	References .....	403



<b>13 Automatic Speech Recognition</b>	<b>409</b>
Hagen Soltau, George Saon, Lidia Mangu, Hong-Kwang Kuo, Brian Kingsbury, Stephen Chu, and Fadi Biadisy	
13.1 Introduction	409
13.1.1 Automatic Speech Recognition	410
13.1.2 Introduction to Arabic: A Speech Recognition Perspective	411
13.1.3 Overview	412
13.2 Acoustic Modeling	413
13.2.1 Language-Independent Techniques	413
13.2.2 Vowelization	418
13.2.3 Modeling of Arabic Dialects in Decision Trees	423
13.3 Language Modeling	428
13.3.1 Language-Independent Techniques for Language Modeling	428
13.3.2 Language-Specific Techniques for Language Modeling	432
13.4 IBM GALE 2011 System Description	434
13.4.1 Acoustic Models	434
13.4.2 Language Models	439
13.4.3 System Combination	441
13.4.4 System Architecture	441
13.5 From MSA to Dialects	443
13.5.1 Dialect Identification	443
13.5.2 ASR and Dialect ID Data Selection	446
13.5.3 Dialect Identification on GALE Data	447
13.5.4 Acoustic Modeling Experiments	448
13.5.5 Dialect ID Based on Text Only	452
13.6 Resources	453
13.6.1 Acoustic Training Data	453
13.6.2 Training Data for Language Modeling	454
13.6.3 Vowelization Resources	454
13.7 Comparing Arabic and Hebrew ASR	455
13.8 Summary	456
References	457

Natural Language Processing of Semitic Languages

Zitouni, I. (Ed.)

2014, XXIV, 459 p. 61 illus., 23 illus. in color., Hardcover

ISBN: 978-3-642-45357-1