

CONTENTS

<i>Illustrations</i>	<i>xvii</i>
<i>Preface</i>	<i>xix</i>
<i>Acknowledgments</i>	<i>xxi</i>
Chapter 1 Introduction	1
<i>Digital Information and Information Users</i>	1
Information	1
Digital Information	2
Information Users	2
<i>Digital Libraries and Digital Collections</i>	3
Digital Libraries Defined	3
Digital Library Architecture	5
Digital Objects	5
Metadata and Metadata Records	6
Digital Collections	7
<i>Services for Digital Collections</i>	8
Searching	8
Browsing	9
Digital Reference Services or Virtual Reference Services	9
Social Media and Interaction	10
<i>World Languages and Linguistics</i>	10
Language	10
World Languages	11
Linguistics—the Discipline of Languages	12

	<i>Related Concepts</i>	12
	<i>Culture</i>	12
	<i>Society</i>	13
	<i>Communication</i>	13
	<i>Multilingual Information Access</i>	13
	<i>Information Access and Its Language Barriers</i>	14
	<i>Defining MLLA</i>	15
	<i>MLLA for Digital Collections</i>	16
	<i>Related Fields</i>	18
	<i>Data Modeling and Database Design</i>	18
	<i>Full-Text Information Retrieval</i>	19
	<i>Natural Language Processing</i>	19
	<i>Machine Translation</i>	21
	<i>User Studies and User Interfaces</i>	22
	<i>Summary and Suggested Readings</i>	23
	<i>Questions</i>	24
	<i>References</i>	24
Chapter 2	Cross-Language Information Retrieval	27
	<i>Information Retrieval and Its Challenges</i>	27
	<i>IR Models</i>	29
	<i>Boolean Model</i>	30
	<i>Vector Space Model</i>	30
	<i>Probability Models</i>	31
	<i>Google's PageRank Model</i>	33
	<i>Cross-Language Information Retrieval</i>	34
	<i>CLIR Processes</i>	35
	<i>Transformation Strategies</i>	35
	<i>Translation Resources and Approaches</i>	37
	<i>Dictionaries, Ontologies, and Thesauri</i>	37
	<i>Parallel, Comparable, and Combined Monolingual Corpora</i>	38
	<i>MT Systems</i>	39
	<i>CLIR Translation Problems and Solutions</i>	39
	<i>Lexical Ambiguity</i>	40
	<i>Untranslatable Search Keys</i>	40
	<i>Phrase Identification and Translation</i>	41
	<i>Noise in Translation Resources</i>	41
	<i>Evaluation</i>	42
	<i>Evaluation Methods</i>	42
	<i>Evaluation Forums</i>	43
	<i>CLIR and MLIR at NTCIR</i>	45
	<i>CLIR</i>	45
	<i>MLIR</i>	46

<i>Summary and Suggested Readings</i>	47
<i>Questions</i>	48
<i>References</i>	48
Chapter 3 <i>Machine Translation Research and Practice</i>	53
<i>MT and Its History</i>	53
<i>Translation and MT</i>	53
<i>History of MT</i>	54
<i>MT Approaches</i>	56
<i>Rule-Based MT</i>	56
<i>Direct Translation</i>	56
<i>Interlingual Translation</i>	57
<i>Transfer-Based Translation</i>	58
<i>Example-Based MT</i>	59
<i>Statistical MT</i>	61
<i>Hybrid MT Approaches</i>	62
<i>Machine Translation between English and East Asian Languages</i>	63
<i>Characteristics of East Asian Languages</i>	63
<i>Chinese</i>	63
<i>Japanese</i>	65
<i>Korean</i>	66
<i>Chinese and Japanese Text Segmentation</i>	66
<i>n-Gram-Based Text Segmentation</i>	66
<i>Chinese Word Segmentation</i>	67
<i>Dictionary-Based Approaches</i>	68
<i>Statistical and Learning-Based Approaches</i>	69
<i>Hybrid Approaches</i>	70
<i>Use of Chinese Linguistic Knowledge in</i>	
<i>Word Segmentation</i>	70
<i>Japanese Word Segmentation</i>	71
<i>Word-Based Approaches</i>	72
<i>Character-Based Tagging Approach</i>	72
<i>Evaluation of Word Segmentation</i>	72
<i>MT Systems and Approaches</i>	73
<i>MT Evaluation</i>	74
<i>Overview</i>	75
<i>Human Evaluation Measures</i>	75
<i>Automatic Evaluation Measures</i>	76
<i>MT Evaluation Procedures</i>	76
<i>Data Preparation</i>	76
<i>Determination of Evaluation Measures</i>	77
<i>Reference Translation Creation</i>	77
<i>Evaluator Recruitment</i>	77

	<i>Evaluation Platform Development</i>	78
	<i>Evaluation Instruction and Progress</i>	78
	<i>Result Analysis and Decision Making</i>	78
	<i>MT for Information Service</i>	78
	<i>MT Applications</i>	79
	<i>Use Scenarios in Library and Information Science</i>	80
	<i>Summary and Suggested Readings</i>	81
	<i>Questions</i>	82
	<i>References</i>	83
Chapter 4	Machine Translation for Digital Collections	89
	<i>MRT Project: Metadata Records Machine Translation and Evaluation</i>	89
	<i>Objectives</i>	90
	<i>Research Design</i>	90
	<i>Metadata Records Extraction and MT</i>	90
	<i>Metadata Records Extraction</i>	90
	<i>MT of Selected Records</i>	93
	<i>HeMT: The Platform for Human Translation and MT</i>	93
	<i>Evaluation</i>	93
	<i>HeMT Design Principle</i>	93
	<i>Major Functions</i>	94
	<i>User Management</i>	94
	<i>Multilingual Lexicon Management</i>	94
	<i>Manual Translation</i>	95
	<i>User Training</i>	95
	<i>Evaluation</i>	95
	<i>Evaluation Result Visualization</i>	95
	<i>Database Design</i>	96
	<i>Website Architecture</i>	98
	<i>Implementation</i>	98
	<i>Usability Testing</i>	99
	<i>Workflow</i>	100
	<i>Testing Model</i>	101
	<i>Pretest and Posttest Questionnaires</i>	102
	<i>Test Results and Discussion</i>	103
	<i>Evaluation Tasks and Measures</i>	103
	<i>Evaluation Tasks</i>	103
	<i>Evaluation Measures</i>	104
	<i>Evaluator Recruitment and Training</i>	106
	<i>The Evaluation Process</i>	108
	<i>Results and Analysis</i>	108
	<i>Adequacy, Fluency, and the Best System</i>	109
	<i>Chinese Translations</i>	109

<i>Spanish Translations</i>	110
<i>Comments from the Evaluators</i>	110
<i>Translation Errors</i>	111
<i>Summary of the Results</i>	111
<i>Building Your Own MT System</i>	112
<i>Procedures for Building an MT System</i>	112
<i>Determining an MT Strategy</i>	112
<i>Selecting an Open-Source MT Platform</i>	113
<i>Analyzing Available Linguistic Resources</i>	113
<i>Installing and Configuring the MT Platform</i>	113
<i>Developing or Purchasing Desired Linguistic Resources</i>	114
<i>Creating an MT System</i>	114
<i>Testing and Adjusting the MT System</i>	114
<i>Documenting the MT System</i>	114
<i>MRT Project: Constructing a Multi-Engine MT System</i>	115
<i>Multi-Engine MT Strategy</i>	115
<i>Moses: The Statistical MT Platform</i>	116
<i>Available Linguistic Resources</i>	117
<i>Moses Installation and Configuration</i>	117
<i>Parallel Corpus Preparation and Monolingual Corpus Acquisition</i>	118
<i>MEMT Strategies and Experimentation</i>	118
<i>Results Analysis</i>	119
<i>Thoughts on Future Improvement</i>	120
<i>Appendix: Data Dictionary for HEMT Database</i>	121
<i>MDR Table</i>	121
<i>MDR-DET Table</i>	122
<i>USER Table</i>	122
<i>REF Table</i>	123
<i>REF-DET Table</i>	123
<i>MT Table</i>	124
<i>MT-DET Table</i>	124
<i>EVAL Table</i>	125
<i>EVAL-DET Table</i>	125
<i>EVAL-COM Table</i>	126
<i>TERM Table</i>	127
<i>Summary and Suggested Readings</i>	127
<i>References</i>	127
Chapter 5 <i>Multilingual Systems and Interfaces</i>	131
<i>Multilingual Digital Libraries and Collections</i>	131
<i>Multilingual Digital Libraries Defined</i>	131
<i>Multilingual Systems on the Internet</i>	131
<i>Multilingual Digital Collections</i>	132

<i>Multilingual Digital Libraries</i>	132
<i>A Framework for Analyzing Multilingual Digital Libraries</i>	133
<i>The Perseus Digital Library</i>	134
<i>Missions & Goals</i>	134
<i>Funding</i>	135
<i>People</i>	135
<i>Organization</i>	135
<i>Services</i>	136
<i>Architecture</i>	136
<i>Technologies</i>	137
<i>The World Digital Library</i>	137
<i>Missions & Goals</i>	137
<i>Funding</i>	138
<i>People</i>	138
<i>Organization</i>	138
<i>Services</i>	139
<i>Architecture</i>	140
<i>Technologies</i>	140
<i>ICDL</i>	141
<i>Missions & Goals</i>	141
<i>Funding</i>	141
<i>People</i>	141
<i>Organization</i>	142
<i>Services</i>	143
<i>Architecture</i>	143
<i>Technologies</i>	144
<i>Characteristics of Multilingual Digital Libraries</i>	144
<i>Challenges and Opportunities to Multilingual Digital Libraries</i>	145
<i>Developing Multilingual Services for Digital Collections</i>	146
<i>General Procedures</i>	146
<i>Planning</i>	146
<i>Analysis</i>	147
<i>Design</i>	147
<i>Implementation</i>	147
<i>User Testing and Modification</i>	148
<i>MLIA Strategies</i>	148
<i>Query Translation–Based</i>	148
<i>Metadata Records Translation–Based</i>	149
<i>Multilingual Interfaces</i>	150
<i>Search Interface</i>	150
<i>Result Display Interface</i>	151
<i>Browsing Interface</i>	152
<i>Related Projects</i>	152
<i>MultiMatch Project</i>	153
<i>MLIA4DC Project</i>	153

<i>Summary and Suggested Readings</i>	154
<i>Questions</i>	155
<i>References</i>	155
Chapter 6 Beyond Retrieval: Knowledge Discovery and Future Directions	157
<i>Knowledge Discovery from Multilingual Texts</i>	157
<i>Multilingual Information Extraction</i>	157
<i>Information Extraction Tasks</i>	158
<i>Entity Identification and Extraction</i>	158
<i>Scenario or Relationship Detection</i>	158
<i>Event Identification and Extraction</i>	158
<i>General Information Extraction Processes</i>	159
<i>Automatic Metadata Generation</i>	159
<i>Multilingual Information Extraction and Its Challenges</i>	160
<i>Evaluating Multilingual Information Extraction</i>	162
<i>Multilingual Question Answering</i>	162
<i>Introduction</i>	162
<i>Question Analysis</i>	163
<i>Question Answering Approaches</i>	165
<i>Multilingual Question Answering Research</i>	165
<i>NTCIR Multilingual QA Tracks</i>	166
<i>CLEF Multilingual QA Tracks</i>	168
<i>Evaluating Multilingual Question Answering</i>	168
<i>Challenges and Future Development</i>	169
<i>Other Knowledge Discovery Tasks</i>	169
<i>Multilingual Text Summarization</i>	169
<i>Rapid Prototyping of Cross-Language Tools for Low-Resource Languages</i>	170
<i>Cross-lingual Link Discovery</i>	170
<i>Multilingual Social Network Analysis</i>	171
<i>Multilingual Recommender Systems</i>	171
<i>Multilingual News Systems</i>	171
<i>MLIA Challenges and Future Directions</i>	171
<i>Developing and Sharing Knowledge Resources and Tools</i>	172
<i>Integrating Different Approaches</i>	173
<i>Involving Users into the MLIA Procedures</i>	173
<i>Developing and Evaluating Empirical MLIA Systems</i>	173
<i>Preparing For Multilingual Information Access</i>	173
<i>Studying Related Knowledge and Skills</i>	174
<i>Keep an Open Mind to New Technologies</i>	174
<i>Seek Collaboration with MLIA Researchers</i>	174
<i>Summary and Suggested Readings</i>	175

<i>Questions</i>	175
<i>References</i>	176
Chapter 7 Related Internet Language Resources and Tools	181
<i>Dictionaries and Parallel Corpora</i>	181
<i>Monolingual and Multilingual Dictionaries</i>	182
<i>Acronym Finder</i>	182
<i>Dictionary.com</i>	182
<i>ADL Gazetteer</i>	182
<i>On-line Chinese Tools</i>	183
<i>Japanese Language Learning Tools on the Web</i>	183
<i>Merriam-Webster Online</i>	183
<i>WordNet</i>	183
<i>Parallel Corpora and Comparative Corpora</i>	184
<i>Canadian Hansards</i>	184
<i>Europarl and JRC-Acquis</i>	184
<i>Chinese-English Parallel Corpora</i>	184
<i>Wikipedia</i>	184
<i>Test Collections</i>	185
<i>TREC Test Collections</i>	185
<i>NTGIR Test Collections</i>	185
<i>Open-Source Systems and Tools</i>	185
<i>Digital Library Management Systems</i>	185
<i>DSpace</i>	186
<i>Fedora</i>	186
<i>Greenstone</i>	186
<i>CONTENTdm</i>	186
<i>Text Processing Systems</i>	187
<i>Apache OpenNLP</i>	187
<i>NLTK</i>	187
<i>Stanford CoreNLP</i>	187
<i>Stemmers</i>	188
<i>Part-of-Speech Taggers</i>	188
<i>Syntactic Parsers</i>	188
<i>Word Segmenters</i>	188
<i>Information Retrieval Systems</i>	188
<i>Lemur & Indri</i>	188
<i>Apache Lucene</i>	189
<i>Terrier</i>	189
<i>Machine Translation Services and Platforms</i>	189
<i>Google Translate</i>	189
<i>Bing Translator</i>	190
<i>SYSTRAN</i>	190
<i>WorldLingo</i>	190
<i>Apertium</i>	190

<i>Moses</i>	190
<i>Joshua</i>	191
<i>NiuTrans</i>	191
<i>Associations and Groups</i>	191
<i>Professional Associations</i>	191
<i>ACM Special Interest Group on Information Retrieval (SIGIR)</i>	191
<i>The Linguistic Society of America</i>	191
<i>The Association for Computational Linguistics</i>	192
<i>Association for Computational Linguistics and Chinese Language Processing</i>	192
<i>Asian Federation of Natural Language Processing (AFNLP)</i>	192
<i>Research Groups</i>	192
<i>Research Groups in Information Retrieval</i>	192
<i>Industrial Research Groups</i>	193
<i>Other Organizations</i>	193
<i>Linguistic Data Consortium (LDC)</i>	193
<i>European Language Resources Association Corpora-List</i>	194
<i>Research and Publication Sites</i>	194
<i>ACL Anthology</i>	194
<i>ACM Digital Library</i>	194
<i>Other Repositories and Archives</i>	195
<i>Dr. Oard's Website</i>	195
<i>MT Archive</i>	195
<i>Statistical MT</i>	195
<i>Publications from Large-Scale Evaluations</i>	195
<i>References</i>	195
<i>Acronym List</i>	197
<i>Glossary</i>	201
<i>Index</i>	205