Brief Contents

What This Book Is About xi
Overview for Instructors xiii
Acknowledgments xvii

1 Prologue: Encoding Language on Computers 1
2 Writers’ Aids 33
3 Language Tutoring Systems 69
4 Searching 91
5 Classifying Documents: From Junk Mail Detection to Sentiment Classification 127
6 Dialog Systems 153
7 Machine Translation Systems 181
8 Epilogue: Impact of Language Technology 215

References 221
Concept Index 227
Contents

What This Book Is About xi
Overview for Instructors xiii
Acknowledgments xvii

1 Prologue: Encoding Language on Computers 1
  1.1 Where do we start? 1
    1.1.1 Encoding language 2
  1.2 Writing systems used for human languages 2
    1.2.1 Alphabetic systems 3
    1.2.2 Syllabic systems 6
    1.2.3 Logographic writing systems 8
    1.2.4 Systems with unusual realization 11
    1.2.5 Relation to language 11
  1.3 Encoding written language 12
    1.3.1 Storing information on a computer 12
    1.3.2 Using bytes to store characters 14
  1.4 Encoding spoken language 17
    1.4.1 The nature of speech 17
    1.4.2 Articulatory properties 18
    1.4.3 Acoustic properties 18
    1.4.4 Measuring speech 20
      Under the Hood 1: Reading a spectrogram 21
    1.4.5 Relating written and spoken language 24
      Under the Hood 2: Language modeling for automatic speech recognition 26
Contents

2 Writers’ Aids 33
  2.1 Introduction 33
  2.2 Kinds of spelling errors 34
    2.2.1 Nonword errors 35
    2.2.2 Real-word errors 37
  2.3 Spell checkers 38
    2.3.1 Nonword error detection 39
    2.3.2 Isolated-word spelling correction 41
    Under the Hood 3: Dynamic programming 44
  2.4 Word correction in context 49
    2.4.1 What is grammar? 50
    Under the Hood 4: Complexity of languages 56
    2.4.2 Techniques for correcting words in context 58
      Under the Hood 5: Spell checking for web queries 62
  2.5 Style checkers 64

3 Language Tutoring Systems 69
  3.1 Learning a language 69
  3.2 Computer-assisted language learning 71
  3.3 Why make CALL tools aware of language? 73
  3.4 What is involved in adding linguistic analysis? 76
    3.4.1 Tokenization 76
    3.4.2 Part-of-speech tagging 78
    3.4.3 Beyond words 80
  3.5 An example ICALL system: TAGARELA 81
  3.6 Modeling the learner 83

4 Searching 91
  4.1 Introduction 91
  4.2 Searching through structured data 93
  4.3 Searching through unstructured data 95
    4.3.1 Information need 95
    4.3.2 Evaluating search results 96
    4.3.3 Example: Searching the web 97
    4.3.4 How search engines work 100
      Under the Hood 6: A brief tour of HTML 103
  4.4 Searching semi-structured data with regular expressions 107
    4.4.1 Syntax of regular expressions 108
    4.4.2 Grep: An example of using regular expressions 110
      Under the Hood 7: Finite-state automata 112
  4.5 Searching text corpora 115
    4.5.1 Why corpora? 116
    4.5.2 Annotated language corpora 117
      Under the Hood 8: Searching for linguistic patterns on the web 118
5  Classifying Documents: From Junk Mail Detection to Sentiment Classification 127
   5.1 Automatic document classification 127
   5.2 How computers “learn” 129
      5.2.1 Supervised learning 130
      5.2.2 Unsupervised learning 131
   5.3 Features and evidence 131
   5.4 Application: Spam filtering 133
      5.4.1 Base rates 135
      5.4.2 Payoffs 139
      5.4.3 Back to documents 139
   5.5 Some types of document classifiers 140
      5.5.1 The Naive Bayes classifier 140
      Under the Hood 9: Naive Bayes 142
      5.5.2 The perceptron 145
      5.5.3 Which classifier to use 148
   5.6 From classification algorithms to context of use 149

6  Dialog Systems 153
   6.1 Computers that “converse”? 153
   6.2 Why dialogs happen 155
   6.3 Automating dialog 156
      6.3.1 Getting started 156
      6.3.2 Establishing a goal 157
      6.3.3 Accepting the user's goal 157
      6.3.4 The caller plays her role 158
      6.3.5 Giving the answer 158
      6.3.6 Negotiating the end of the conversation 159
   6.4 Conventions and framing expectations 159
      6.4.1 Some framing expectations for games and sports 160
      6.4.2 The framing expectations for dialogs 160
   6.5 Properties of dialog 161
      6.5.1 Dialog moves 161
      6.5.2 Speech acts 162
      6.5.3 Conversational maxims 164
   6.6 Dialog systems and their tasks 166
   6.7 Eliza
      Under the Hood 10: How Eliza works 172
   6.8 Spoken dialogs 174
   6.9 How to evaluate a dialog system 175
   6.10 Why is dialog important? 176

7  Machine Translation Systems 181
   7.1 Computers that “translate”? 181
   7.2 Applications of translation 183
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2.1</td>
<td>Translation needs</td>
<td>183</td>
</tr>
<tr>
<td>7.2.2</td>
<td>What is machine translation really for?</td>
<td>184</td>
</tr>
<tr>
<td>7.3</td>
<td>Translating Shakespeare</td>
<td>185</td>
</tr>
<tr>
<td>7.4</td>
<td>The translation triangle</td>
<td>188</td>
</tr>
<tr>
<td>7.5</td>
<td>Translation and meaning</td>
<td>191</td>
</tr>
<tr>
<td>7.6</td>
<td>Words and meanings</td>
<td>193</td>
</tr>
<tr>
<td>7.6.1</td>
<td>Words and other languages</td>
<td>193</td>
</tr>
<tr>
<td>7.6.2</td>
<td>Synonyms and translation equivalents</td>
<td>194</td>
</tr>
<tr>
<td>7.7</td>
<td>Word alignment</td>
<td>194</td>
</tr>
<tr>
<td>7.8</td>
<td>IBM Model 1</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>Under the Hood 11: The noisy channel model</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Under the Hood 12: Phrase-based statistical translation</td>
<td>204</td>
</tr>
<tr>
<td>7.9</td>
<td>Commercial automatic translation</td>
<td>205</td>
</tr>
<tr>
<td>7.9.1</td>
<td>Translating weather reports</td>
<td>205</td>
</tr>
<tr>
<td>7.9.2</td>
<td>Translation in the European Union</td>
<td>207</td>
</tr>
<tr>
<td>7.9.3</td>
<td>Prospects for translators</td>
<td>208</td>
</tr>
<tr>
<td>8</td>
<td>Epilogue: Impact of Language Technology</td>
<td>215</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>221</td>
</tr>
<tr>
<td></td>
<td>Concept Index</td>
<td>227</td>
</tr>
</tbody>
</table>