
Contents

FOREWORD BY A. ZAMORA AND R. SALVADOR	xi
FOREWORD BY H. SAGGION	xv
NOTATION	xvii
INTRODUCTION	xix
PART 1. FOUNDATIONS	1
CHAPTER 1. WHY SUMMARIZE TEXTS?	3
1.1. The need for automatic summarization	3
1.2. Definitions of text summarization	5
1.3. Categorizing automatic summaries	10
1.4. Applications of automatic text summarization	13
1.5. About automatic text summarization	15
1.6. Conclusion	21
CHAPTER 2. AUTOMATIC TEXT SUMMARIZATION: SOME IMPORTANT CONCEPTS	23
2.1. Processes before the process	23
2.1.1. Sentence-term matrix: the vector space model (VSM) model	26
2.2. Extraction, abstraction or compression?	28

2.3. Extraction-based summarization	30
2.3.1. Surface-level algorithms	31
2.3.2. Intermediate-level algorithms	33
2.3.3. Deep parsing algorithms	34
2.4. Abstract summarization	35
2.4.1. FRUMP	35
2.4.2. Information extraction and abstract generation	38
2.5. Sentence compression and Fusion	38
2.5.1. Sentence compression	38
2.5.2. Multisentence fusion	39
2.6. The limits of extraction	39
2.6.1. Cohesion and coherence	40
2.6.2. The HexTAC experiment	42
2.7. The evolution of text summarization tasks	43
2.7.1. Traditional tasks	43
2.7.2. Current and future problems	45
2.8. Evaluating summaries	50
2.9. Conclusion	51
CHAPTER 3. SINGLE-DOCUMENT SUMMARIZATION	53
3.1. Historical approaches	53
3.1.1. Luhn’s Automatic Creation of Literature Abstracts	57
3.1.2. The Luhn algorithm	59
3.1.3. Edmundson’s linear combination	61
3.1.4. Extracts by elimination	64
3.2. Machine learning approaches	66
3.2.1. Machine learning parameters	66
3.3. State-of-the-art approaches	69
3.4. Latent semantic analysis	73
3.4.1. Singular value decomposition (SVD)	73
3.4.2. Sentence weighting by SVD	74
3.5. Graph-based approaches	76
3.5.1. PAGERANK and SNA algorithms	77
3.5.2. Graphs and automatic text summarization	78
3.5.3. Constructing the graph	79
3.5.4. Sentence weighting	80

3.6. DIVTEX: a summarizer based on the divergence of probability distribution	83
3.7. CORTEX	85
3.7.1. Frequential measures	86
3.7.2. Hamming measures	87
3.7.3. Mixed measures	88
3.7.4. Decision algorithm	89
3.8. ARTEX	90
3.9. ENERTEX	93
3.9.1. Spins and neural networks	93
3.9.2. The textual energy similarity measure	95
3.9.3. Summarization by extraction and textual energy	97
3.10. Approaches using rhetorical analysis	102
3.11. Lexical chains	107
3.12. Conclusion	107

CHAPTER 4. GUIDED MULTI-DOCUMENT

SUMMARIZATION	109
4.1. Introduction	109
4.2. The problems of multidocument summarization	110
4.3. DUC/TAC & INEX Tweet Contextualization	112
4.4. The taxonomy of MDS methods	115
4.4.1. Structure based	115
4.4.2. Vector space model based	116
4.4.3. Graph based	117
4.5. Some multi-document summarization systems and algorithms	117
4.5.1. SUMMONS	118
4.5.2. Maximal marginal relevance	119
4.5.3. A multidocument biography summarization system	120
4.5.4. Multi-document ENERTEX	121
4.5.5. MEAD	123
4.5.6. CATS	126
4.5.7. SUMUM and SUMMA	128
4.5.8. NEO-CORTEX	131
4.6. Update summarization	134
4.6.1. Update summarization pilot task at DUC 2007	134

4.6.2. Update summarization task at TAC 2008 and 2009	135
4.6.3. A minimization-maximization approach	138
4.6.4. The ICSI system at TAC 2008 and 2009	142
4.6.5. The CBSEAS system at TAC	145
4.7. Multidocument summarization by polytopes	146
4.8. Redundancy	148
4.9. Conclusion	149
PART 2. EMERGING SYSTEMS	151
CHAPTER 5. MULTI AND CROSS-LINGUAL SUMMARIZATION	153
5.1. Multilingualism, the web and automatic summarization	153
5.2. Automatic multilingual summarization	156
5.3. MEAD	159
5.4. SUMMARIST	159
5.5. COLUMBIA NEWSBLASTER	161
5.6. NEWSEXPLORER	163
5.7. GOOGLE NEWS	166
5.8. CAPS	166
5.9. Automatic cross-lingual summarization	168
5.9.1. The quality of machine translation	169
5.9.2. A graph-based cross-lingual summarizer	172
5.10. Conclusion	177
CHAPTER 6. SOURCE AND DOMAIN-SPECIFIC SUMMARIZATION	179
6.1. Genre, specialized documents and automatic summarization	179
6.2. Automatic summarization and organic chemistry	183
6.2.1. YACHS2	183
6.3. Automatic summarization and biomedicine	189
6.3.1. SUMMTERM	189
6.3.2. A linguistic-statistical approach	196
6.4. Summarizing court decisions	201
6.5. Opinion summarization	204
6.5.1. CBSEAS at TAC 2008 opinion task	204

6.6. Web summarization	206
6.6.1. Web page summarization	206
6.6.2. OCELOT and the statistical gist	207
6.6.3. Multitweet summarization	211
6.6.4. Email summarization	215
6.7. Conclusion	216
CHAPTER 7. TEXT ABSTRACTING	219
7.1. Abstraction-based automatic summarization	219
7.2. Systems using natural language generation	220
7.3. An abstract generator using information extraction	222
7.4. Guided summarization and a fully abstractive approach	223
7.5. Abstraction-based summarization via conceptual graphs	226
7.6. Multisentence fusion	227
7.6.1. Multisentence fusion via graphs	228
7.6.2. Graphs and keyphrase extraction: the TAKAHÉ system	231
7.7. Sentence compression	232
7.7.1. Symbolic approaches	235
7.7.2. Statistical approaches	236
7.7.3. A statistical-linguistic approach	238
7.8. Conclusion	241
CHAPTER 8. EVALUATING DOCUMENT SUMMARIES	243
8.1. How can summaries be evaluated?	243
8.2. Extrinsic evaluations	245
8.3. Intrinsic evaluations	246
8.3.1. The baseline summary	247
8.4. TIPSTER SUMMAC evaluation campaigns	248
8.4.1. <i>Ad hoc</i> task	249
8.4.2. Categorization task	249
8.4.3. Question-answering task	250
8.5. NTCIR evaluation campaigns	250
8.6. DUC/TAC evaluation campaigns	251
8.6.1. Manual evaluations	252
8.7. CLEF-INEX evaluation campaigns	254
8.8. Semi-automatic methods for evaluating summaries	256

8.8.1. Level of granularity: the sentence	256
8.8.2. Level of granularity: words	257
8.9. Automatic evaluation via information theory	263
8.9.1. Divergence of probability distribution	265
8.9.2. FRESA	266
8.10. Conclusion	271
CONCLUSION	275
APPENDIX 1. INFORMATION RETRIEVAL, NLP AND ATS	281
APPENDIX 2. AUTOMATIC TEXT SUMMARIZATION RESOURCES	305
BIBLIOGRAPHY	309
INDEX	343