Computing Attitude and Affect in Text: Theory and Applications

## THE INFORMATION RETRIEVAL SERIES

*Series Editor:*

**W. Bruce Croft**

*University of Massachusetts, Amherst*

**Also in the Series:**

**INFORMATION RETRIEVAL SYSTEMS:** ***Theory and Implementation,*** *by Gerald Kowalski;* ISBN: 0-7923-9926-9

**CROSS-LANGUAGE INFORMATION RETRIEVAL,** *edited by Gregory Grefenstette;* ISBN: 0-7923-8122-X

**TEXT RETRIEVAL AND FILTERING:** ***Analytic Models of Performance***, *by Robert M. Losee;* ISBN: 0-7923-8177-7

**INFORMATION RETRIEVAL: UNCERTAINTY AND LOGICS:** ***Advanced Models for the Representation and Retrieval of Information*****,** *by Fabio Crestani, Mounia Lalmas, and Cornelis Joost van Rijsbergen;* ISBN: 0-7923-8302-8

**DOCUMENT COMPUTING:** ***Technologies for Managing Electronic Document Collections***, *by Ross Wilkinson, Timothy Arnold-Moore, Michael Fuller, Ron Sacks-Davis, James Thom, and Justin Zobel;* ISBN: 0-7923-8357-5

**AUTOMATIC INDEXING AND ABSTRACTING OF DOCUMENT TEXTS,** *by Marie-Francine Moens*; ISBN 0-7923-7793-1

**ADVANCES IN INFORMATIONAL RETRIEVAL:** ***Recent Research from the Center for Intelligent Information Retrieval***, *by W. Bruce Croft*; ISBN 0-7923-7812-1

**INFORMATION RETRIEVAL SYSTEMS:** ***Theory and Implementation,*** **Second Edition**, *by Gerald J. Kowalski and Mark T. Maybury;* ISBN: 0-7923-7924-1

**PERSPECTIVES ON CONTENT-BASED MULTIMEDIA SYSTEMS,** *by Jian Kang Wu; Mohan S. Kankanhalli;Joo-Hwee Lim;Dezhong Hong;* ISBN: 0-7923-7944-6

**MINING THE WORLD WIDE WEB:** ***An Information Search Approach***, *by George Chang, Marcus J. Healey, James A. M. McHugh, Jason T. L. Wang;* ISBN: 0-7923-7349-9

**INTEGRATED REGION-BASED IMAGE RETRIEVAL**, *by James Z. Wang;* ISBN: 0-7923-7350-2

**TOPIC DETECTION AND TRACKING: Event-based Information Organization**, *edited by James Allan;* ISBN: 0-7923-7664-1

**LANGUAGE MODELING FOR INFORMATION RETRIEVAL,** *edited by W. Bruce Croft; John Lafferty;* ISBN: 1-4020-1216-0

**MACHINE LEARNING AND STATISTICAL MODELING APPROACHES TO IMAGE RETRIEVAL,** *by Yixin Chen, Jia Li and James Z. Wang;* ISBN: 1-4020-8034-4

**INFORMATION RETRIEVAL:** ***Algorithms and Heuristics,*** *by David A. Grossman and Ophir Frieder,* 2nd ed.; ISBN: 1-4020-3003-7; PB: ISBN: 1-4020-3004-5

# Computing Attitude and Affect in Text: Theory and Applications

Edited by

**James G. Shanahan**

*Clairvoyance Cooperation, Pittsburgh, PA, U.S.A.*

**Yan Qu**

*Clairvoyance Cooperation, Pittsburgh, PA, U.S.A.*

and

**Janyce Wiebe**

*University of Pittsburgh, PA, U.S.A.*

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-4026-1 (HB)
ISBN-13 978-1-4020-4026-9 (HB)
ISBN-10 1-4020-4102-0 (e-book)
ISBN-13 978-1-4020-4102-0 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

*www.springer.com*

*Printed on acid-free paper*



Printed in the Netherlands.

# TABLE OF CONTENTS

**PREFACE** ..... **xi**

**1. CONTEXTUAL VALENCE SHIFTERS**

**Livia Polanyi, Annie Zaenen** ..... **1**

1. INTRODUCTION ..... 1
2. FROM SIMPLE VALENCE TO CONTEXTUALLY DETERMINED VALENCE ..... 2
3. CONTEXTUAL VALENCE SHIFTERS ..... 3
4. CONCLUSION ..... 9

**2. CONVEYING ATTITUDE WITH REPORTED SPEECH**

**Sabine Bergler** ..... **11**

1. INTRODUCTION ..... 11
2. EVIDENTIAL ANALYSIS OF REPORTED SPEECH ..... 12
3. PROFILE STRUCTURE ..... 14
4. EXTENDED EXAMPLE ..... 16
5. SOURCE LIST ANNOTATION ..... 17
6. EXTENSION TO OTHER ATTRIBUTION ..... 20
7. CONCLUSION ..... 20

**3. WHERE ATTITUDINAL EXPRESSIONS GET THEIR ATTITUDE**

**Jussi Karlgren, Gunnar Eriksson, Kristofer Franzén** ..... **23**

1. RESEARCH QUESTIONS TO MOTIVATE THE STUDY OF ATTITUDINAL EXPRESSIONS ..... 23
2. STARTING POINTS – PROTOTYPICAL ATTITUDINAL EXPRESSIONS ..... 24
3. TEXT TOPICALITY: PLAYERS ..... 24
4. TEXT TOPICALITY: MOVES ..... 25
5. IDENTIFYING PLAYERS ..... 25
6. THE CASE FOR ANIMACY: ADJECTIVAL ATTRIBUTES AND GENITIVE ATTRIBUTES ..... 25
7. THE CASE FOR SYNTACTIC STRUCTURE: SITUATIONAL REFERENCE ..... 26
8. USING SYNTACTIC PATTERNS MORE SYSTEMATICALLY ..... 28
9. GENERALIZING FROM SYNTACTIC PATTERNS TO THE LEXICON ..... 29
10. CONCLUSIONS ..... 29

**4. ANALYSIS OF LINGUISTIC FEATURES ASSOCIATED WITH POINT OF VIEW FOR GENERATING STYLISTICALLY APPROPRIATE TEXT**

**Nancy L. Green** ..... **33**

1. INTRODUCTION ..... 33
2. PERSPECTIVES IN CORPUS ..... 34
3. ASSOCIATED FEATURES ..... 36

4. IMPLICATIONS FOR NATURAL LANGUAGE GENERATION AND AUTOMATIC RECOGNITION OF POINT OF VIEW ..... 38

**5. THE SUBJECTIVITY OF LEXICAL COHESION IN TEXT**

**Jane Morris, Graeme Hirst ..... 41**

1. INTRODUCTION ..... 41
2. THEORETICAL BACKGROUND ..... 42
3. EXPERIMENTAL STUDY ..... 43
4. DISCUSSION ..... 45

**6. A WEIGHTED REFERENTIAL ACTIVITY DICTIONARY**

**Wilma Bucci, Bernard Maskit ..... 49**

1. INTRODUCTION ..... 50
2. METHODS ..... 52
3. RESULTS ..... 58

**7. CERTAINTY IDENTIFICATION IN TEXTS: CATEGORIZATION MODEL AND MANUAL TAGGING RESULTS**

**Victoria L. Rubin, Elizabeth D. Liddy, Noriko Kando ..... 61**

1. ANALYTICAL FRAMEWORK ..... 62
2. PROPOSED CERTAINTY CATEGORIZATION MODEL ..... 65
3. EMPIRICAL STUDY ..... 68
4. APPLICATIONS ..... 74
5. CONCLUSIONS AND FUTURE WORK ..... 74

**8. EVALUATING AN OPINION ANNOTATION SCHEME USING A NEW MULTI-PERSPECTIVE QUESTION AND ANSWER CORPUS**

**Veselin Stoyanov, Claire Cardie, Diane Litman, Janyce Wiebe ..... 77**

1. INTRODUCTION ..... 77
2. LOW-LEVEL PERSPECTIVE INFORMATION ..... 78
3. THE MPQA NRRC CORPUS ..... 80
4. MULTI-PERSPECTIVE QUESTION AND ANSWER CORPUS CREATION ..... 80
5. EVALUATION OF PERSPECTIVE ANNOTATIONS FOR MPQA ..... 83
6. CONCLUSIONS AND FUTURE WORK ..... 89

**9. VALIDATING THE COVERAGE OF LEXICAL RESOURCES FOR AFFECT ANALYSIS AND AUTOMATICALLY CLASSIFYING NEW WORDS ALONG SEMANTIC AXES**

**Gregory Grefenstette, Yan Qu, David A. Evans, James G. Shanahan ..... 93**

1. INTRODUCTION ..... 94
2. THE CURRENT CLAIRVOYANCE AFFECT LEXICON ..... 95
3. EMOTIVE PATTERNS ..... 97
4. SCORING THE INTENSITY OF CANDIDATE AFFECT WORDS ..... 101
5. FUTURE WORK ..... 105
6. CONCLUSIONS ..... 106

**10. A COMPUTATIONAL SEMANTIC LEXICON OF FRENCH VERBS OF EMOTION**

**Yvette Yannick Mathieu ..... 109**

1. INTRODUCTION ..... 109
2. SEMANTIC LEXICON DESCRIPTION ..... 109
3. *FEELING* SYSTEM ..... 116
4. EVALUATION ..... 122
5. RELATED WORK ..... 123
6. CONCLUSION ..... 123

**11. EXTRACTING OPINION PROPOSITIONS AND OPINION HOLDERS USING SYNTACTIC AND LEXICAL CUES**

**Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, Dan Jurafsky 125**

1. INTRODUCTION ..... 125
2. DATA ..... 127
3. OPINION-ORIENTED WORDS ..... 130
4. IDENTIFYING OPINION PROPOSITIONS ..... 132
5. RESULTS ..... 136
6. ERROR ANALYSIS ..... 138
7. DISCUSSION ..... 139

**12. APPROACHES FOR AUTOMATICALLY TAGGING AFFECT**

**Nathanael Chambers, Joel Tetreault, James Allen ..... 143**

1. INTRODUCTION ..... 143
2. BACKGROUND ..... 144
3. ROCHESTER MARRIAGE-COUNSELING CORPUS ..... 145
4. APPROACHES TO TAGGING ..... 146
5. EVALUATIONS ..... 153
6. DISCUSSION ..... 154
7. CATS TOOL ..... 156
8. RELATED WORK ..... 157
9. CONCLUSION ..... 157

**13. ARGUMENTATIVE ZONING FOR IMPROVED CITATION INDEXING**

**Simone Teufel ..... 159**

1. CITATION INDEXING AND CITATION MAPS ..... 159
2. ARGUMENTATIVE ZONING AND AUTHOR AFFECT ..... 161
3. META-DISCOURSE ..... 163
4. HUMAN ANNOTATION OF AUTHOR AFFECT ..... 165
5. FEATURES FOR AUTHOR AFFECT ..... 167
6. EVALUATION ..... 167
7. CONCLUSION ..... 168

**14. POLITENESS AND BIAS IN DIALOGUE SUMMARIZATION: TWO EXPLORATORY STUDIES**

**Norton Trevisan Roman, Paul Piwek, Ariadne Maria Brito Rizzoni Carvalho ..... 171**

1. INTRODUCTION ..... 172
2. FIRST STUDY: POLITENESS AND BIAS IN UNCONSTRAINED DIALOGUE SUMMARIZATION ..... 174
3. SECOND STUDY: POLITENESS AND BIAS IN CONSTRAINED DIALOGUE SUMMARIZATION ..... 178
4. COMPARISON ..... 180
5. CONCLUSION AND OUTLOOK ..... 181

**15. GENERATING MORE-POSITIVE AND MORE-NEGATIVE TEXT**

**Diana Zaiu Inkpen, Ol'ga Feiguina, Graeme Hirst ..... 187**

1. NEAR-SYNONYMS AND ATTITUDINAL NUANCES ..... 187
2. RELATED WORK ..... 189
3. ESTIMATING THE RELATIVE SEMANTIC ORIENTATION OF TEXT ..... 189
4. WORD SENSE DISAMBIGUATION ..... 190
5. ANALYSIS ..... 190
6. GENERATION ..... 191
7. EXPERIMENTS ..... 192
8. EVALUATION ..... 195
9. CONCLUSION ..... 196

**16. IDENTIFYING INTERPERSONAL DISTANCE USING SYSTEMIC FEATURES**

**Casey Whitelaw, Jon Patrick, Maria Herke-Couchman ..... 199**

1. INTRODUCTION ..... 200
2. SYSTEMIC FUNCTIONAL LINGUISTICS ..... 200
3. REPRESENTING SYSTEM NETWORKS ..... 204
4. IDENTIFYING REGISTERS ..... 209
5. CONCLUSION ..... 212

**17. CORPUS-BASED STUDY OF SCIENTIFIC METHODOLOGY: COMPARING THE HISTORICAL AND EXPERIMENTAL SCIENCES**

**Shlomo Argamon, Jeff Dodick ..... 215**

1. INTRODUCTION ..... 216
2. BACKGROUND ..... 216
3. SYSTEMIC INDICATORS AS TEXTUAL FEATURES ..... 219
4. EXPERIMENTAL STUDY ..... 222
5. EXAMPLE TEXTS ..... 227
6. CONCLUSIONS ..... 228

**18. ARGUMENTATIVE ZONING APPLIED TO CRITIQUING NOVICES' SCIENTIFIC ABSTRACTS**

**Valéria D. Feltrim, Simone Teufel, Maria das Graças V. Nunes, Sandra M. Aluísio ..... 233**

1. INTRODUCTION ..... 234

2. THE SCIPO SYSTEM ..... 234
3. ARGUMENTATIVE ZONING FOR PORTUGUESE TEXTS ..... 237
4. EVALUATION OF SCIPO'S CRITIQUING TOOL ..... 242
5. CONCLUSIONS ..... 244

**19. USING HEDGES TO CLASSIFY CITATIONS IN SCIENTIFIC ARTICLES**

**Chrysanne Di Marco, Frederick W. Kroon, Robert E. Mercer ..... 247**

1. SCIENTIFIC WRITING, THE NEED FOR AFFECT, AND ITS ROLE IN CITATION ANALYSIS ..... 247
2. HEDGING IN SCIENTIFIC WRITING ..... 248
3. CLASSIFYING CITATIONS IN SCIENTIFIC WRITING ..... 250
4. DETERMINING THE IMPORTANCE OF HEDGES IN CITATION CONTEXTS ..... 252
5. A CITATION INDEXING TOOL FOR BIOMEDICAL LITERATURE ANALYSIS ..... 256
6. CONCLUSIONS AND FUTURE WORK ..... 261

**20. TOWARDS A ROBUST METRIC OF POLARITY**

**Kamal Nigam, Matthew Hurst ..... 265**

1. INTRODUCTION ..... 265
2. RELATED WORK ..... 266
3. CLASSES OF POLAR EXPRESSION ..... 268
4. RECOGNIZING POLAR LANGUAGE ..... 269
5. TOPIC DETECTION IN ONLINE MESSAGES ..... 270
6. THE INTERSECTION OF TOPIC AND POLARITY ..... 272
7. EMPIRICAL ANALYSIS ..... 273
8. METRICS FOR TOPIC AND POLARITY ..... 275
9. CONCLUSIONS AND FUTURE WORK ..... 277

**21. CHARACTERIZING BUZZ AND SENTIMENT IN INTERNET SOURCES: LINGUISTIC SUMMARIES AND PREDICTIVE BEHAVIORS**

**Richard M. Tong, Ronald R. Yager ..... 281**

1. INTRODUCTION AND MOTIVATION ..... 281
2. LINGUISTIC SUMMARIES ..... 282
3. EXAMPLE APPLICATIONS ..... 289
4. TRENDS-2™ INFRASTRUCTURE ..... 292
5. PREVIOUS AND RELATED WORK ..... 293
6. OPEN R&D AND APPLICATION ISSUES ..... 293

**22. GOOD NEWS OR BAD NEWS? LET THE MARKET DECIDE**

**Moshe Koppel, Itai Shtrimberg ..... 297**

1. INTRODUCTION ..... 297
2. EXPERIMENTS ..... 298
3. RESULTS ..... 299
4. CONCLUSIONS ..... 300

**23. OPINION POLARITY IDENTIFICATION OF MOVIE REVIEWS**

**Franco Salvetti, Christoph Reichenbach, Stephen Lewis ..... 303**

1. INTRODUCTION ..... 303
2. RELATED RESEARCH ..... 304
3. PROBABILISTIC APPROACHES TO POLARITY IDENTIFICATION ..... 305
4. FEATURES FOR ANALYSIS ..... 306
5. PART OF SPEECH FEATURE SELECTION ..... 307
6. EXPERIMENTS ..... 308
7. SYNONYMY AND HYPERNYMY FEATURE GENERALIZATION ..... 312
8. SELECTION BY RANKING ..... 314
9. DISCUSSION ..... 314
10. CONCLUSION ..... 315

**24. MULTI-DOCUMENT VIEWPOINT SUMMARIZATION FOCUSED ON FACTS, OPINION AND KNOWLEDGE**

**Yohei Seki, Koji Eguchi, Noriko Kando ..... 317**

1. INTRODUCTION ..... 318
2. EXPERIMENT OVERVIEW: MULTI-DOCUMENT VIEWPOINT SUMMARIZATION WITH SUMMARY TYPES ..... 319
3. SENTENCE-TYPE ANNOTATION ..... 323
4. GENRE CLASSIFICATION ..... 325
5. EXPERIMENT RESULTS ..... 328
6. CONCLUSION ..... 333

**INDEX ..... 337**

# Preface

Human Language Technology (HLT) and Natural Language Processing (NLP) systems have typically focused on the "factual" aspect of content analysis. Other aspects, including pragmatics, opinion, and style, have received much less attention. However, to achieve an adequate understanding of a text, these aspects cannot be ignored.

The chapters in this book address the aspect of subjective opinion, which includes identifying different points of view, identifying different emotive dimensions, and classifying text by opinion. Various conceptual models and computational methods are presented. The models explored in this book include the following: distinguishing attitudes from simple factual assertions; distinguishing between the author's reports from reports of other people's opinions; and distinguishing between explicitly and implicitly stated attitudes. In addition, many applications are described that promise to benefit from the ability to understand attitudes and affect, including indexing and retrieval of documents by opinion; automatic question answering about opinions; analysis of sentiment in the media and in discussion groups about consumer products, political issues, etc.; brand and reputation management; discovering and predicting consumer and voting trends; analyzing client discourse in therapy and counseling; determining relations between scientific texts by finding reasons for citations; generating more appropriate texts and making agents more believable; and creating writers' aids. The studies reported here are carried out on different languages such as English, French, Japanese, and Portuguese.

Difficult challenges remain, however. It can be argued that analyzing attitude and affect in text is an "NLP"-complete problem. The interpretation of attitude and affect depends on audience, context, and world knowledge. In addition, there is much yet to learn about the psychological and biological relationships between emotion and language.

To continue to progress in this area in NLP, more comprehensive theories of emotion, attitude and opinion are needed, as are lexicons of affective terms and knowledge of how such terms are used in context, and annotated corpora for training and evaluation.

This book is a first foray into this area; it grew out of a symposium on this topic that took place at Stanford University in March, 2004, under support from American Association for Artificial Intelligence (AAAI). Several of the presentations were extended into the chapters that appear here. The chapters in this collection reflect the majors themes of the workshop, corresponding to a balance among conceptual models, computational methods, and applications. The chapters in this book are organized along these themes into three broad, overlapping parts.

**Linguistic and Cognitive Models**

The chapters in the first part of this book explore linguistic and cognitive models which could support developing richer computational models of attitude and affect. This section begins with Polanyi and Zaenen's fascinating study of attitudinal valence (or polarity) as it is expressed in context. While individual words often suggest a negative or positive attitude, such as "horrible" and "great", respectively, the context of a word may change its base valence. Polanyi and Zaenen describe and illustrate a number of such contextual valence shifters, both intra-sentential (e.g., negatives and modals) and inter-sentential (e.g., discourse connectives and multi-entity evaluation).

The next chapter, by Bergler, also explores linguistic devices for conveying attitudes in text, namely reported speech expressions which convey attitudes. Bergler argues that reported speech serves to segment information into discourse segments called profiles. Each profile involves such things as degree of credibility of the source of the information, and the role the source has in the argumentative structure of the text. Bergler performs a detailed profile analysis of an extended story, and discusses extending this type of analysis to other attributes than reported speech.

Like Polaryi and Zaenen, and Bergler, Karlgren et al. focus on contextual aspects of linguistic expressions of attitudes. Their particular objects of study are attitude expressions which are internally structured. They argue that simple lists of attitudinal terms are not sufficient for recognizing attitudes in texts: it is often only in particular lexical and syntactic patterns that words convey attitudes. They present interesting results of a corpus study suggesting that certain syntactic contexts are more likely to be loci of attitudes, and that this is realized in stylistic differences between opinionated text types such as editorials and more objective text types such as reporting news articles.

The second set of chapters in this section address cognitive as well as linguistic issues in understanding attitude and affect in text. Green's chapter presents the results of a qualitative analysis of letters written by genetic counselors to their clients. The goal is to find stylistic features that would be salient for natural generation systems in this genre. Her study suggests that perspective must be taken into account to generate stylistically appropriate text. Green identifies a number of perspectives in this genre, including specific agents such as the author and client, as well as abstract perspectives such as education and research. As a generation system assumes different perspectives while generating such a letter, it should choose forms of reference, tenses, types of evidential language, and so forth to reflect that perspective.

Morris and Hirst's chapter addresses readers' perceptions of lexical semantic relations, in particular the perceived subjectivity of such relations. They perform a study to assess the degree of individual differences in readers' interpretations of lexical chains, which are groups of related words that create lexical cohesion. The results showed that subjects identified a common core of groups of related words in text, but also exhibited individual differences. Such knowledge could help NLP systems recognize which types of text meaning can be expected to be shared by most readers, and understand and generate text appropriately.

Bucci and Maskit's presentation of a "Weighted Referential Activity Dictionary" is the most psychologically oriented chapter in the collection. Bucci and Maskit use their dictionary in computer modeling of a psycholinguistic variable which they call Referential Activity (RA). RA ratings measure the degree to which language connects to nonverbal experiences such as bodily and emotional experience. In Bucci and Maskit's model, RA is mainly indicated by domain-independent stylistic attributes of language, aspects of which are included in their dictionary. The chapter presents compelling RA analyses of literary passages, and describes a method for assigning RA weights to dictionary entries. Their study reveals differential linguistic roles of particular lexical items in producing vivid versus abstract texts. They plan to investigate the psychological significance of these differences in future work.

The final two chapters in this part of the book present annotation schemes, i.e., schemes for manually labeling texts to create data for training and evaluating NLP systems. The chapter by Rubin et al. presents a framework for coding the writer's certainty in text. They categorize a set of linguistic certainty markers (such as "probably"and "allegedly") along four dimensions – level

(degree of certainty), perspective (whose certainty is being encoded), focus (abstract versus factual information), and time (past, present, or future). They perform an empirical study of their framework in which they applied their annotation scheme to 32 newspaper articles. Among their findings are that editorials contain more explicit certainty markers than news articles, and that a few specific combinations of dimension values dominate in editorials. The framework and empirical results will be informative for developing automatic certainty identification systems.

The chapter by Stoyanov et al. addresses annotations for Multi-Perspective Question Answering (MPQA), whereby an NLP system answers opinion-oriented questions. To be successful, an MPQA system will presumably need to recognize and organize the opinions expressed in one or more documents. An annotation scheme for encoding such opinions has been developed and evaluated in previous work. This chapter investigates the utility of that annotation scheme for MPQA processing. It first describes a new corpus of multi-perspective questions and answers. It then presents the results of a study investigating the usefulness of the earlier opinion annotations for multi-perspective versus fact-based question answering. Their findings are that opinion annotations can be useful for MPQA if used appropriately.

### Lexical Resources and Attitude/Affect Recognition and Generation

The first two chapters in this part of the book focus on lexical resources that could support recognition and generation of attitude and affect. Lexicons of words of emotion-conveying potential have been used in much work for identifying and generating affect. The chapter by Grefenstette et al. addresses the problems of automatically extracting affect words for expanding the coverage of existing affect lexicons, and of automatically assigning the affect words along multiple semantic axes. Emotive patterns are used as seeds to extract affect words from the Web. Through evaluation of the precision and recall of the extracted words, the authors show that it is possible to identify lexical patterns for finding emotion-bearing affect words with high precision. Once the affect words are extracted, the authors discuss ways to automatically assign the words along the different semantic axes using measures similar to point-wise mutual information. The measures show promise for finding degrees of belongingness to the semantic classes while at the same time assigning degrees of intensity to the affect words.

In the following chapter, Mathieu first presents a manually constructed lexicon of French verbs of emotion with positive, negative, or neutral affect. Thirty-three semantic classes are proposed and the classes are arranged in graphical structures through links of intensity and antonyms. French verbs of each class are described by simple attribute-pair type properties such as whether a verb accepts a non-agentive subject or not. The lexicon is evaluated for identifying positive, negative, or neutral affect of sentences from French Letters to the Editors texts. The evaluation shows that taking into account the intensity of verbs of emotion produces better classification results.

The next three chapters present computational methods for recognizing attitude and affect in text. The first chapter by Bethard et al. addresses the tasks of detecting propositional opinions and detecting holders of these opinions. Unlike a variety of previous work on separating facts from opinions at the document or sentence level, this paper focuses on determining the opinion status of a smaller piece of text. Propositional opinions are opinions that are generally found as the sentential complements of a predicate. The authors use supervised statistical classification methods for proposition detection and opinion-holder detection, incorporating semantic constituent labeling, opinion-oriented words, and syntactic features such as the presence of complex adjective phrases.

The next chapter by Chambers et al. presents approaches for automatically tagging the attitude of the speakers in transcribed dialogues. The authors explore several n-gram- and vector-based approaches and present results in a marriage-counseling domain and the Switchboard Corpus. In the marriage-counseling domain, each transcript is broken into thought units that are manually annotated with tags classifying the attribute and emotional commitment of the participants to a particular topic of discussion. The Switchboard corpus consists of conversations of random topics and has a richer tagging scheme. The performance results over both corpora are comparable, and the simple n-gram based approaches outperform or perform as well as the vector-based approaches. The authors also describe a Java tool for tagging attitude and affect which integrates the automatic classification capability.

The chapter by Teufel addresses the problem of automatically classifying academic citations in scientific articles according to author affect. The two rhetorical roles for citation analysis that are associated with affect in text include Contrast (comparison with, criticism of, or contract to other work) and Basis (agreement with or continuation of other work). Teufel examines discourse features such as section structure, history to classify author affect, in addition to other features such as semantic class of main verb, indicator phrases, etc. Such analysis aims at improving citation indexing through better detection of subjectivity in scientific text.

The last two chapters in this part of the book explore attitude and affect in text generation and summarization. Roman et al. explore the influence of affect and attitude on summarizing dialogues. In particular, they address the question of whether politeness should be reported in dialogue summaries and, if so, how politeness is reported. The chapter presents empirical studies designed to gather information about how people summarize dialogues. In these studies, a collection of four dialogues, involving a customer and vendor about buying a car, was used. Each dialogue was generated by an automated system with the politeness of the dialogue participants manipulated. Subjects were asked to summarize the dialogues from a particular dialogue participant's point of view. The studies showed that the percentage of summaries reporting some behavior information was higher when the dialogues were more impolite. This result is independent of the point of view and summary size. The studies also indicated that the point of view adopted by the summarizer biases the reporting of behaviors in their summaries. In particular, negative reporting of behavior information depends on the point of view of the summarizer rather than on the actual dialogue behavior. Tentative evidence showed that positive reporting is less subject to such bias.

While Roman et al. study how people's points of view influence human generation of text and summaries, Inkpen et al. explore a way of producing text with different attitudinal nuances by varying word choices. In particular, they examine nuances that differentiate near-synonyms relating to expressed attitude and text, and propose to transform the semantic orientation of a text automatically by choosing near-synonyms accordingly. The transformation of semantic orientation involves first representing text as an inter-lingual representation and a set of lexical nuances, and then replacing the words with attitudinal nuances in the original text by their near-synonyms according to the desired nuances.

## Applications

The third part of this book focuses on applications of attitude and affect. The first two chapters in this section explore the categorization of text based on the manner in which a document is written rather than its content. In particular, both chapters use a computational model based on different aspects of systemic functional linguistic (SFL) theory. Whitelaw et al. present a study that

demonstrates that the pronominal and determination systems of SFL are indeed powerful ways of characterizing interpersonal distance (between author and reader). They show empirically that this characterization of text is a robust means of recognizing financial scam email from regular email with a performance accuracy of 98% using a variety of machine learning algorithms. In contrast, Argamon and Dodick focus on conjunction, modality, and comment subsystems of SFL for genre-based text categorization of scientific articles in the historical and experimental sciences. Using a support vector machine trained on a systemic functional feature set (with no domain specific terms), they achieved over 83% accuracy for classifying articles according to field. The most highly-weighted features for each were consistent with hypothesized methodological differences between historical and experimental sciences.

The next two chapters in this section deal with applications that analyze the rhetorical structures of scientific papers. The first chapter by Feltrim et al. describes a system that uses argumentative zoning, a technique for identifying the rhetorical structure of text, as a thesis writing aid for graduate students working in Portuguese. Argumentative zoning techniques assign a label (drawn from possible rhetorical role labels such as background, purpose, results, and conclusion) to each sentence, indicating its argumentative role in a portion of text. The argumentative zoning algorithm (realized through a Naïve Bayes classier) is used to label each sentence as being a background, gap, purpose, methodology, result, conclusion, or outline. These rhetorical labels are then used by a rule-based system to identify problems in scientific text abstracts. A reported user study highlights the value of such a system for masters-level students. The work represents a successful adaptation of the argumentative zoning technique to the Portuguese language. In contrast, Di Marco et al., in their chapter, empirically validate a hypothesis that the use of hedges (words that make text more or less vague) is highly correlated with sentences that contain or surround citations. This study is based upon 985 peer-reviewed recent biology journal articles from the BioMed Central corpus. In addition, Di Marco et al. describe a system for classifying citations into 35 categories using a hand-built decision tree over cue-words, polarity switching words, and knowledge of the discourse structure of the article, among other features. Citation categories vary depending on the function of the citation, e.g., support or contrast.

The next two chapters in this section focus on aggregation of opinion from multiple sources. Nigam and Hurst describe an interesting polarity classifier which uses shallow NLP techniques and a topic-based classifier. They propose using a Bayesian statistical approach to aggregate the opinions expressed about a specific topic in Internet forums. Tong and Yager explore aggregating and characterizing opinion over time. They first create a time series of the subjects, opinions, and attitudes expressed in Internet sources. Subsequently, they generate linguistic summaries, using fuzzy set theory, which provide perspicuous overviews of the opinions expressed towards an event over a period of time.

The final three chapters of this section focus on empirical studies of deploying opinion-based systems. Koppel and Shtrimberg examine the use of sentiment analysis as a means of predicting future stock prices. Though their findings highlight that this is not a useful investment strategy, one potentially useful outcome of their work is a method for collecting labeled data for sentiment analysis, where data is labeled based upon the direction of relative large changes in stock price.

Salvetti, Reichenbach and Lewis describe an approach to opinion classification of movie reviews based upon feature selection (using part of speech tags), feature generalization (in terms of synonymy and hypernymy), and probabilistic classifiers (namely Naïve Bayes and Markov

Models). They note that using a simple thresholding of the log odds ratio of the positive and negative posterior probabilities can dramatically improve performance.

The final chapter, by Seki, Eguchi and Kando, focuses on multi-document summarization based on a topic/query and investigates the impact of using sentence and document-level genre information on building three types of summaries: summaries that concentrate on facts (events), opinions, or knowledge (definitions), respectively. The topic, characterized as a query, is used to retrieve/select documents from a collection of documents. The retrieved documents are then summarized using a clustering-based approach, where clusters and sentences within clusters are ranked and selected based upon similarity to the topic. The user is further allowed to select the type of summary required. The reported results on Japanese newswire articles show significant improvement in summary coverage and precision when combining sentence-level typing and genre classification information over baseline multi-document summarization techniques.

## Target Audience

The book is intended for advanced undergraduate and graduate students, as well as a broad audience of professionals and researchers in computer science, engineering, information science, and content analysis who have an interest in the subjective aspects of text. The subject matter in this book is far ranging, including conceptual models, computational models, and applications.

## Acknowledgements

We thank Helen Perilioux, Jesse Montogomery, and Debbie Moran for their help in editing this book.

James G. Shanahan
Yan Qu
Janyce Wiebe
Pittsburgh, June 2005