TOPIC DETECTION AND TRACKING

Event-based Information Organization

edited by

James Allan
University of Massachusetts at Amherst

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

Contents

Preface ix

1
Introduction to Topic Detection and Tracking 1
James Allan
1 Introduction 1
2 TDT tasks 3
3 History of TDT 7
4 TDT 1999 and TDT 2000 10
5 The Future of TDT 13

2
Topic Detection and Tracking Evaluation Overview 17
Jonathan G. Fiscus and George R. Doddington
1 Introduction 17
2 TDT Definitions: Stories, Events, and Topics 18
3 TDT Corpora 19
4 Evaluation Methodology 20
5 Task Definitions 25
6 Summary 30

3
Corpora for Topic Detection and Tracking 33
Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman
1 Introduction 33
2 Overview of TDT Corpus Development 35
3 Collection of Raw Data 36
4 Transcription 38
5 Story Segmentation 39
6 Topic Definition 42
7 Topic Annotation 45
8 Corpus Formats 54
9 Some Properties of the Corpus 61
10 Conclusion 64

4
Probabilistic Approaches to Topic Detection and Tracking 67
Tim Leek, Richard Schwartz, and Srinivasa Sista
1 Introduction 67
2 Core TDT Technologies 68
3 Corpus Processing 75
4 Tracking 75
5 Detection 77
6 Crosslingual TDT 80
7 Conclusions and Future Work 81
8 Acknowledgments 82

5
Multi-strategy Learning for TDT 85
Yiming Yang, Jaime Carbonell, Ralf Brown, John Lafferty, Thomas Pierce, and Thomas Ault
1 Introduction 85
2 Segmentation 87
3 Topic and Event Tracking 88
4 Topic Detection 96
5 First Story Detection 99
6 Story Link Detection 101
7 Multilingual TDT 107
8 Concluding Remarks 111

6
Statistical Models of Topical Content 115
J.P. Yamron, L. Gillick, P. van Mulbregt, and S. Knecht
1 Introduction 115
2 Models of Story Generation 117
3 Tracking Systems 120
4 Detection System 128
5 Summary 132

7
Segmentation and Detection at IBM 135
S. Dharanipragada, M. Franz, J.S. McCarley, T. Ward, and W.-J. Zhu
1 Story Segmentation 135
2 Topic Detection 142
3 Acknowledgements 147

8
A Cluster-Based Approach to Broadcast News 149
David Eichmann and Padmini Srinivasan
1 Introduction 149
2 Segmentation 152
3 Detection 154
4 Tracking 163
5 Acknowledgements 173

9
Signal Boosting for Translingual Topic Tracking 175
Gina-Anne Levow and Douglas W. Oard
1 Introduction 176
2 The Signal-to-Noise Perspective 177
3 Topic Tracking System Architecture 178
4 Contrastive Conditions 184
5 Conclusions and Future Work 191
6 Acknowledgments 194

10
Explorations Within Topic Tracking and Detection 197
James Allan, Victor Lavrenko, and Russell Swan
1 Introduction 197
2 Basic System 198
3 Tracking 203
4 Cluster Detection 205
5 First Story Detection 208
6 Link Detection 208
7 Bounds on Effectiveness 216
8 Automatic Timeline Generation 219
9 Conclusions 222

11
Towards a “Universal Dictionary” for Multi-Language IR Applications 225
J. Michael Schultz and Mark Y. Liberman
1 Introduction 225
2 Our TDT tracking algorithm 229
3 The “Universal Dictionary” experiment 236
4 Conclusions and Directions for Future Work 239

12
An NLP & IR Approach to Topic Detection 243
Hsin-Hsi Chen and Lun-Wei Ku
1 Introduction 243
2 General System Framework 245
3 Representation of News Stories and Topics 246
4 Similarity and Interpretation of a Two-threshold Method 248
5 Multilingual Topic Detection 250
6 Development Experiments 256
7 Evaluation 259
8 Discussion 261
9 Concluding Remarks and Future Works 262

Index 265