Jadt 2010

Statistical Analysis of Textual Data

Proceedings of 10th International Conference
Journées d'Analyse statistique des Données Textuelles
9-11 June 2010 - Sapienza University of Rome

Sergio Bolasco
Isabella Chiari
Luca Giuliano
editors

LED Edizioni Universitarie di Lettere Economia Diritto

JADT 2010
http://jadt2010.uniroma1.it http.//jadt.org

Organised by: Dip. di Studi Geoeconomici Linguistici Statistici Storici per l'Analisi Regionale, Sapienza Univ. di Roma

With the support of: SIS - Società Italiana di Statistica

Sponsored by: SAPIENZA UNIVERSITÀ DI ROMA
Dip. di Studi Sociali Economici Attuariali Demografici
Dip. di Sociologia e Comunicazione
Dip. di Matematica e Statistica, Università di Napoli 'Federico II'
Dip. di Sociologia, Università di Padova
CISU (Casa Editrice)

and by:
Istat - Istituto Nazionale di Statistica
Enel - Ente Nazionale di Energia Elettrica
Percorsi srl
SAS Institute

Graphic and Design: Augusto Frascatani, Ida Potenza
Artwork: Caterina Bolasco

Subjects:
Statistical Methods, Exploratory Textual Data Analysis, Text Mining, Information Extraction, Text Categorization and Classification, Natural Language Processing, Computational Linguistics, Semantic Analysis, Content Analysis, Discourse Analysis

ISBN 978-88-7916-450-9

LED Edizioni Universitarie di Lettere Economia Diritto

Via Cervignano 4 - 20137 Milano
www.lededizioni.com – www.ledonline.it – E-mail: led@lededizioni.com

Pagination and Layout: Linda Cazzaniga
Printed by: Digital Print Service

Contents

Volume 1

Text Mining

Stylometry and Textometry

Lexical and Semantic Analysis

PART II
DOMAIN SPECIFIC ANALYSIS

Newspaper Analysis

New Media and Social Networks

Volume 2

Legal, Scientific and Political Discourse

Society and Psycho-/Socio-Linguistics

PART III
APPLICATIONS AND TOOLS

Software Applications and Tools

Text Corpora Encoding

Monolingual and Multilingual Lexical Resources

Preface

The JADT (*Journées d'Analyse Statistique des Données Textuelles*) conference has become a reference point for all scholars interested in the methods and application of the statistical analysis of textual data. This year, it was held for the second time at the "La Sapienza" University of Rome (June 9-11, 2010). The meeting drew together a large group of statisticians, linguists, psychologists, sociologists and communication and IT experts, and was the tenth in the series – previous ones having been held in Barcelona (1990), Montpellier (1993), Rome (1995), Nice (1995), Lausanne (2000), Saint Malo (2002), Louvain-la-Neuve (2004), Besançon (2006) and Lyon (2008). Every two years, the JADT conference presents the state of the art for what concerns theories, problems, methods, algorithms and applications that share a 'metric' approach to the study of lexical, textual, pragmatic or discursive features of information expressed in natural language. The cover of the present book collects keywords extracted from the titles of the 800 papers presented so far at the JADT meetings.

The Proceedings of the 2010 conference, published in two volumes and, for the first time, on CD-ROM, collect 120 contributions by 250 scholars from 15 countries spread all over the world. Since the many applicative possibilities and problems of the statistical analysis of textual data correspond to a strong interdisciplinary character, these contributions are not easily classified. Nevertheless, an attempt has been made to help the reader by identifying three broad areas of interest: methods and exploratory textual data analysis; domain specific analysis; applications and tools.

The first part - METHODS AND EXPLORATORY TEXTUAL DATA ANALYSIS - deals with the use of original and innovative exploratory methodologies and analyses. Much attention has been given to the methods of automatic classification and indexing, ranging from the classification of autobiographical narratives or positive-negative evaluations to the comparison of different classification techniques, from SVM based strategies to Text Mining solutions. This section also includes papers on stylometry and textometry, two areas that share several measuring tools, interpretative criteria and areas of application. A focus on specific linguistic aspects (for instance, the automatic extraction of collocations, the measurement of textual coherence and the relationship between lexis, usage and semantics) characterizes the chapter on lexical and semantic analysis.

The second part - DOMAIN SPECIFIC ANALYSIS – covers the whole range of subjects that are explored by disciplines that make use of statistical analyses of textual data: for instance, the language of the press, the presentation of events by the media, and new communication technologies and applications in web 2.0 – such as social networks, blogs, forum and Wikipedia. The second volume begins with a chapter devoted to the analysis of legal, administrative, scientific and political discourse. Among other things, it includes studies of presidential discourses, analyses of semantic prosody in academic discourse, and an investigation of the lexis of the JADT network itself. This section ends with a series of contributions of a sociological and psycho-/socio-linguistic nature. These examine, for example, man-machine interactions, the spoken language of Alzheimer patients, and the lexis of trade union negotiations and job advertisements.

The third part presents research contributions on APPLICATIONS AND TOOLS. The first chapter of this section discusses software and tools for text mining and text analysis by means of web services, open-source platforms (TreeTagger, SATO, TXM) or automatic translation

tools for statistical applications. The second chapter is a collection of papers on the encoding and annotation of texts. It deals, in particular, with the difficulties involved in tagging spoken corpora, syntactic annotation, and annotation standards. Lastly, the third chapter collects works on the creation and use of multi- and mono-lingual lexical resources with a view to measuring similarities between bi-lingual dictionaries, creating dictionaries on idioms and building and using parallel corpora and frequency lists.

Acknowledgements

We express our gratitude to the 58 reviewers who offered their invaluable assistance in the selection and anonymous refereeing of the papers in this volume: Alessandra Areni, Enrica Aureli, Harald Baayen, Simona Balbi, Barbara Baldazzi, Roberto Basili, Francois Bavaud, Valérie Beaudouin, Monica Becue, Sergio Bolasco, Giuseppe Bove, Martine Cadot, Paola Cavalieri, Isabella Chiari, Mirella Conenna, Michele Cortelazzo, François Daoust, Jean-Claude Deroubaix, Anne Dister, Francesca Dolcetti, Jules Duchastel, Ramón Álvarez Esteban, Cedrick Fairon, Serge Fleury, Francesca Gambarotto, Kim Gerdes, Luca Giuliano, Benoit Habert, Serge Heiden, Stella Iezzi, Michel Jacobson, Margareta Kastberg, Mauro La Torre, Pierre Lafon, Ludovic Lebart, Jean-Marc Leblanc, Alain Lelu, Dominique Longrée, Pascal Marchand, William Martinez, Damon Mayaffre, Sylvie Mellet, Michelangelo Misuraca, Denis Monière, Hubert Naets, Rosamaria Paniccia, Bénédicte Pincemin, Gérald Purnelle, Max Reinert, Liliane Rodriguez, André Salem, Max Silberztein, Monique Slodzian, Maurice Tournier, Arjuna Tuzzi, Flavia Ursini, Jean-Marie Viprey, François Yvon.

JADT2010 was held under the patronage of the SIS, Società Italiana di Statistica (Italian Statistics Society), and was funded by the "Sapienza" University of Rome and in particular, by the Department of Social, Economic, Demographic and Actuarial Studies and the Department of Sociology and Communication Science; by the Department of Mathematics and Statistics at the University of Naples 'Federico II'; and by the Department of Sociology at the University of Padua. We are also very grateful to the following sponsors: Enel - Ente Nazionale di Energia Elettrica, Istat - Istituto Nazionale di Statistica, Percorsi S.r.l., and the SAS Institute. Not only have the above institutions contributed to the publication of the Proceedings, their in-house researchers have also made interesting and inspiring contributions to the conference as a whole.

As regards the organization of the conference, we would like to thank all the members of the local organizing committee and the staff of the Department of Geo-economic, Linguistic, Statistical and Historical Studies for Regional Analysis, Faculty of Economics. The following people deserve special mention: Rinaldo Coluccio, Paola D'Alonzo, Francesca Gargiulo, Alessandra Leotta, Patrizia Passacantilli, Raffaele Principe and Antonio Santini.

Special thanks go to Caterina Bolasco and Ida Potenza for providing the conference with a visual identity through their creative work; and to Francesco Baiocchi, Arianna Gattoni and Jonathan Anderlucci for their assistance with the web site.

For their precious help and painstaking revision and editing of the texts, we are grateful to Manuela Lo Prejato and Manuela Senza Peluso. We also wish to thank Giorgia Domanico and Marta Vincenzi for their editing assistance and secretarial work.

The editors