

ADVANCES ON  
INFORMATION  
PROCESSING AND  
MANAGEMENT

## **CONFERENCE ORGANIZERS INSTITUTES**

The International Conference on Integrated Information is supported by the following Institutes:

Emerald Group Publishing Limited  
Technological educational Institute of Athens, Greece  
University of Peloponnese, Greece  
National And Kapodistrian University of Athens, Greece  
Mednet Hellas, The Greek Medical Network  
2<sup>nd</sup> AMICUS Workshop

To learn more about I-DAS, including the Book Series, please visit the webpage  
<http://www.i-das.org/>

# INTEGRATED INFORMATION

International Conference on Integrated Information

Kos, Greece      September, 29 – October, 3 2011

EDITORS

Georgios A. Giannakopoulos  
*Technological Educational Institute of Athens, Greece*

Damianos P. Sakas  
*University of Peloponnese, Greece*

**All papers have been peer-reviewed**



**Piraeus, Greece, 2011**

## Editors

Georgios A. Giannakopoulos

Technological Educational Institute of Athens  
Faculty of Management and Economics  
Department of Library Science and Information Systems  
Address: Aghiou Spyridonos Street, 12210, Egaleo  
E-mail: [gian@teiath.gr](mailto:gian@teiath.gr)

Damianos P. Sakas

University of Peloponnese  
Faculty of Science and Technology  
Department of Computer Science and Technology  
Address: End of Karaiskaki St., 22100, Tripolis, Greece  
E-mail: [D.Sakas@uop.gr](mailto:D.Sakas@uop.gr)

The copyrights will be owned by the authors under the Creative Commons Attribution-Non Commercial license (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted use, distribution, and reproduction in any non commercial medium, provided the original work is properly cited.

**ISSN:**

Printed in the Greece, EU

# CONTENTS

<b>PREFACE: Proceedings of the International Conference on Integrated Information (IC-INFO 2011)</b>	<b>1</b>
Georgios A. Giannakopoulos, Damianos P. Sakas	
<b>Conference Details</b>	<b>3</b>
<b>Keynote Speaker</b>	<b>5</b>
<b>SYMPOSIUM ON INFORMATION AND KNOWLEDGE MANAGEMENT</b>	<b>6</b>
Prof. Christos Skourlas	
<b>Towards the Preservation and Availability of Historical Books and Manuscripts: A Case Study</b>	<b>8</b>
Eleni Galiotou	
<b>An Extensive Experimental Study on the Cluster-based Reference set Reduction for Speeding-up the k-nn Classifier</b>	<b>12</b>
Stefanos Ougiaroglou, Georgios Evangelidis and Dimitris A. Dervos	
<b>Exploiting the Search Culture Modulated by the Documentation Retrieval Applications</b>	<b>16</b>
Nikitas N. Karanikolas and Christos Skourlas	
<b>Information and Knowledge Organization: The Case of the TEI of Athens</b>	<b>22</b>
Anastasios Tsolakidis, Manolis Chalaris and Ioannis Chalaris	
<b>Providing Access to Students with Disabilities and Learning Difficulties in Higher Education through a Secure Wireless framework</b>	<b>26</b>
Catherine Marinagi and Christos Skourlas	
<b>Improving Query Efficiency in High Dimensional Point Indexes</b>	<b>30</b>
Evangelos Outsios and Georgios Evangelidis	
<b>Text Segmentation Using Named Entity Recognition and co-Reference Resolution in Greek Texts</b>	<b>34</b>
Pavlina Fragkou	
<b>KINISIS, a Graphical XQuery Language</b>	<b>42</b>
Euclid Keramopoulos, Achilleas Pliakas, Konstantinos Tsekos and Ignatios Deligiannis	
<b>Dimensionality Curse, Concentration Phenomenon and the KDB-tree</b>	<b>46</b>
Nikolaos Kouiroukidis and Georgios Evangelidi	

<b>Applying Balanced Scorecard Strategic Management in Higher Education</b>	<b>50</b>
Manolis Chalaris, Anastasios Tsolakidis and Ioannis Chalaris	
<b>A Web Portal Model for NGOs' Knowledge Management</b>	<b>54</b>
Zuhal Tanrikulu	
<b>The Digital Archives System and Application Optimized for the Tradition Knowledge Archives</b>	<b>58</b>
Jeon Hong. Chan, In Deok. Hwang, Jae Hak. Park, Hyeok. Sim, U won. Gwon and Soon Cheol. Park	
<b>A Semi-automatic Emerging Technology Trend Classifier Using SCOPUS and PATSTAT</b>	<b>62</b>
Seonho Kim, Woondong Yeo, Byong-Youl Coh, Waqas Rasheed, Jaewoo Kang	
<b>Presenting a Framework for Knowledge Management within a Web Enabled Living Lab</b>	<b>66</b>
Lizette de Jager and Albertus AK Buitendag and Potjie (JS) van der Walt	
<b>4TH SYMPOSIUM ON BUSINESS AND MANAGEMENT AND DYNAMIC SIMULATION MODELS SUPPORTING MANAGEMENT STRATEGIES</b>	<b>71</b>
Dr. Damianos Sakas	
<b>New Political Communication Practices: No Budget Events Management. The New Challenge</b>	<b>73</b>
Evangelia N. Markaki, Damianios P. Sakas and Theodore Chadjipantelis	
<b>Free Software – Open Source Software. A Powerful Tool for Developing Creativity in the Hands of the Student</b>	<b>78</b>
Nasiopoulos K. Dimitrios, Damianos P. Sakas, Konstantinos Masselos	
<b>Open Source Web Applications. How it Spread Through the Internet and their Contribution to Education.</b>	<b>82</b>
Nasiopoulos K. Dimitrios, Damianos P. Sakas, Konstantinos Masselos	
<b>Culture in Modern Times in the Frame of Luhmann's System Theory</b>	<b>85</b>
Anastasia J. Chournazidis	
<b>Managing Scientific Journals: A Cultural Viewpoint</b>	<b>87</b>
Marina C. Terzi, Damianos P. Sakas, and Ioannis Seimenis	
<b>A Conceptual Framework for Analyzing Knowledge-based Entrepreneurship</b>	<b>92</b>
Nikos S. Kanellos	

<b>SESSION ON INFORMATION HISTORY: PERSPECTIVES, METHODS AND CURRENT TOPICS</b>	<b>96</b>
Prof. Laszlo Karvalics	
<b>Emerging Research Fields in Information History</b>	<b>98</b>
Laszlo Z. Karvalics	
<b>Information Management through Elementary Data Clusters: New Observations on Pridianum-Type Roman Statistical Documents</b>	<b>102</b>
Gergő Gellérfi	
<b>Information and Secrecy on the Silk Road. Methods of Encryption of Legal Documents in Inner Asia (3th-4th century)</b>	<b>106</b>
Szabolcs Felföldi	
<b>The Role of Information and Disinformation in the Establishment of the Mongolian Empire: A Re-examination of the 13th century Mongolian History from the Viewpoint of Information History</b>	<b>110</b>
Márton Gergő Vér	
<b>Early Warning Systems and the Hospitallers in the Eastern Mediterranean</b>	<b>114</b>
Zsolt Hunyadi	
<b>Information Management as Establishment Dutch Navigational Knowledge on Japan, 1608-1641</b>	<b>118</b>
Gabor Szommer	
<b>Files Everywhere - Register and Training of Men for Military and Civil Purpose in Prussia in the early 18th century</b>	<b>123</b>
Marton Holczer	
<b>SYMPOSIUM ON INTEGRATED INFORMATION: THEORY, POLICIES, TOOLS</b>	<b>126</b>
Prof. Georgios Giannakopoulos	
<b>Approaching Information as an Integrated Field: Educating Information Professionals</b>	<b>128</b>
Georgios Giannakopoulos, Daphne Kyriaki Manesi and Stryidon Zervos	
<b>Special Libraries as Knowledge Management Centers</b>	<b>132</b>
Eva Semertzaki	
<b>Digital Libraries' Developers and their Suitability: A Case Study</b>	<b>136</b>
Maria Monopoli	

<b>A Preliminary Study for the Creation of a Greek Citation index in the Humanities and the Social Sciences (GCI – H&amp;SS)</b>	<b>140</b>
Daphne Kyriaki-Manessi and Evi Sachini	
<b>Archiving as an Information Science. Evidence from a Survey Carried out on a Sample of Greek Students</b>	<b>144</b>
Georgios Giannakopoulos and Ioannis Koumantakis	
<b>Transition Process of E-records Management and Archiving System in Universities: Ankara University</b>	<b>147</b>
S. Özlem Bayram and Fahrettin Ozdemirci	
<b>Government Information: Access and Greece's Efforts for Access</b>	<b>150</b>
Aikaterini Yiannoukakou	
<b>School Archives and their Potentials in Teaching: Aspects of Greek Reality</b>	<b>156</b>
Sonia Geladaki and Panagiota Papadimitriou	
<b>Research on School Libraries in Greece and Suggestions on its Further Development</b>	<b>160</b>
Georgios D. Bikos	
<b>Building Digital Collections for Archeological Sites: Metadata Requirements and CIDOC CRM Extension</b>	<b>164</b>
Georgios S. Gkrous and Mara Nikolaidou	
<b>Museological Claims to Autonomous Knowledge: Rethinking the Conceptual Mode of Display and its Claims to Knowledge</b>	<b>169</b>
Assimina Kaniari and Georgios Giannakopoulos	
<b>Use of Library Loan Records for Book Recommendation</b>	<b>172</b>
Keita Tsuji, Erika Kuroo, Sho Sato, Ui Ikeuchi, Atsushi Ikeuchi, Fuyuki Yoshikane and Hiroshi Itsumura	
<b>Developing a National Database on Librarianship and Information Science. The Case of E-VIVA, the Hellenic Fulltext Database</b>	<b>176</b>
Filippos Ch. Tsimpoglou, Vasiliki V. Koukounidou and Eleni K. Sakka	
<b>Integrated Access to Cultural Heritage Information Pieces in Iran Astan-Quds Razavi's Organization of Libraries, Museums and Documents Center: A Theory of Unionization Disparate Information Assets over Imam Reza's Zarih</b>	<b>181</b>
Ms. Mitra Zarei and Ms. Maliheh Farrokhnia	
<b>Attitudes of University Librarians and Information Scientists towards the Draft Code of</b>	<b>185</b>



<b>Library Ethics to Present a Model for Final Library Ethical Codes</b>	
Mahsoomeh Latifi, Fatemeh Zandian and Hasan Siamian	
<b>SESSION ON OPEN ACCESS REPOSITORIES: SELF-ARCHIVING, METADATA, CONTENT POLICIES, USAGE</b>	<b>188</b>
Dr. Alexandros Koulouris	
<b>Geographical Collections in Greek Academic Libraries: Current Situation and Perspectives</b>	<b>189</b>
Ifigenia Vardakosta and Sarantos Kapidakis	
<b>Information Seeking Behavior: Factors that Affect the Behavior of Greek Astronomers</b>	<b>194</b>
Hara Brindesi and Sarantos Kapidakis	
<b>Aggregating Metadata for Europeana: The Greek Paradigm</b>	<b>198</b>
Alexandros Koulouris, Vangelis Banos and Emmanouel Garoufallou	
<b>Integrating a Repository with Research Output and Publications: The Case of the National Technical University of Athens</b>	<b>202</b>
Dionysis Kokkinos	
<b>Implementation of Workflows as Finite State Machines in a National Doctoral Dissertations Archive</b>	<b>205</b>
Nikos Houssos, Dimitris Zavaliadis, Kostas Stamatis and Panagiotis Stathopoulos	
<b>Practices of “Local” Repositories of Legally Protected Immovable Monuments. A Global Scheme for ‘Designation – Significance’ Information</b>	<b>209</b>
Michail Agathos and Sarantos Kapidakis	
<b>Integration of Metadata in BWMETA-2.0.0 Format</b>	<b>213</b>
Katarzyna Zamlynska, Jakub Jurkiewicz and Lukasz Bolikowski	
<b>SESSION ON EVIDENCE-BASED INFORMATION IN CLINICAL PRACTICE</b>	<b>216</b>
Dr. Evangelia Lappa	
<b>Applicability of Data Mining Algorithms on Clinical Datasets</b>	<b>218</b>
Wilfred, Bonney	
<b>Changing Roles of Health Librarians with Open Access Repositories</b>	<b>221</b>
Christine Urquhar and Assimina Vlachaki	
<b>From Medical Records to Health Knowledge Management Systems: The Coding to Health Sector</b>	<b>225</b>
Evangelia C. Lappa and Georgios A. Giannakopoulos	

<b>The Survey of Skill, Attitude and Use of Computer and Internet among Faculty Members</b>	<b>229</b>
Hasan Siamian, Azita Bala Ghafari, Kobra Aligolbandi, Mohammad Vahedi and Gholam Ali Golafshani Jooybari	
<b>Trends in Scholarly Communication among Biomedical Scientists in Greece</b>	<b>232</b>
Assimina Vlachaki and Christine Urquhart	
<b>SESSION ON ELECTRONIC PUBLISHING: A DEVELOPING LANDSCAPE</b>	<b>236</b>
Dr. Dimitris Kouis	
<b>E-Journal and Open Access Journal Publishing in the Humanities: Preliminary Results from a Survey among Byzantine Studies Scholars</b>	<b>238</b>
Victoria Tsoukala and Evi Sachini	
<b>Preliminary Results on a Printed VS Electronic Text Books Assessment Through Questionnaire</b>	<b>242</b>
Dimitrios A. Kouis and Kanella Pouli	
<b>An Interpretation of Aristotelian Logic According to George Boole</b>	<b>246</b>
Markos N. Dendrinis	
<b>SESSION ON INFORMATION CONTENT PRESERVATION AS OUTCOME OF CONSERVATION OF CULTURAL HERITAGE: ETHICS, METHODOLOGY AND TOOLS</b>	<b>251</b>
Prof. George Panagiaris and Dr. Spiros Zervos	
<b>Intrinsic Data Obfuscation as the Result of Book and Paper Conservation Interventions</b>	<b>254</b>
Spiros Zervos, Alexandros Koulouris and Georgios Giannakopoulos	
<b>Mass Deacidification: Preserving More than Written Information</b>	<b>258</b>
Michael Ramin, Evelyn Eisenhauer and Markus Reist	
<b>Information Literacy of Library Users: A Case Study of Mazandaran Public Library Users, Iran</b>	<b>260</b>
Hussein Mahdizadeh and Hasan Siamian	
<b>The Narratives of Paper in The Archives of the New Independent Greek State (Mid 19th c.)</b>	<b>264</b>
Ourania Kanakari and Maria Giannikou	
<b>From Macro to Micro and from Micro to Nano: The Evolution of the Information Content Preservation of Biological Wet Specimen Collections</b>	<b>268</b>
Nikolaos Maniatis and Georgios Panagiaris	

<b>Digital images: A valuable scholar's tool or misleading material?</b>	<b>272</b>
Patricia Engel	
<b>Attitudes of University Librarians and Information Scientists Towards the Draft Code of Library Ethics to Present a Model for Final Library Ethical Codes</b>	<b>277</b>
Mahsoomeh Latifi, Fatemeh Zandianand and Hasan Siamian	
<b>Investigation of the Degradation Mechanisms of Organic Materials: From Accelerated Ageing to Chemometric Studies</b>	<b>280</b>
Ekaterini Malea, Effie Papageorgiou and Georgios Panagiaris	
<b>SESSION ON DIVERGENCE AND CONVERGENCE: INFORMATION WORK IN DIGITAL CULTURAL MEMORY INSTITUTIONS</b>	<b>285</b>
Dr. Susan Myburgh	
<b>Extending Convergence and Divergence in Cultural Memory Institutions: The Old Slave Lodge in the New South Africa</b>	<b>287</b>
Archie L Dick	
<b>The Transfer of Knowledge from Large Organizations to Small: Experiences from a Research Project on Digitization in Wales</b>	<b>289</b>
Clare Wood-Fisher, Richard Gough, Sarah Higgins, Menna Morgan, Amy Staniforth and Lucy Tedd	
<b>The Usage of Reference Management Software (Rms) in an Academic Environment : A Survey at Tallinn University</b>	<b>293</b>
Enrico Francese	
<b>Varialog : How to Locate Words in a French Renaissance Virtual Library</b>	<b>297</b>
Marie-Hélène Lay	
<b>The Urge to Merge: A Theoretical Approach</b>	<b>301</b>
Susan Myburgh	
<b>SYMPOSIUM ON ADVANCES INFORMATION FOR STRATEGIC MANAGEMENT</b>	<b>304</b>
Professor Nikolaos Konstantopoulos	
<b>Empowerment in the Tax Office of Greece</b>	<b>306</b>
Antonios E. Giokas and Nikolaos P. Antonakas	
<b>Building Absorptive Capacity Through Internal Corporate Venturing</b>	<b>310</b>
Ioannis M. Sotiriou and Alexandros I. Alexandrakis	

<b>The Monitoring Information System (M.I.S.) - An information and Management System for Projects Co-financed Under the National Strategic Reference Framework (NSRF) and the Community support framework (CSF)</b>	<b>314</b>
Catherina G. Siampou, Eleni G. Fassou and Athanassios P. Panagiotopoulos	
<b>Corruption in Tax Administration: The Entrepreneurs View Point</b>	<b>318</b>
Nikolaos P. Antonakas, Antonios E. Giokas and Nikolaos Konstantopoulos	
<b>Conflicts between the IT Manager and the Software House after the Strategic Choice of Outsourcing of the Information Processes in Maritime Companies.</b>	<b>322</b>
Anthi Z. Vaxevanou, Nikolaos Konstantopoulos, Damianos P. Sakas	
<b>Contemporary Forms of Ordering Between the Supply Department and Ship Chandler Companies in the Shipping Industry</b>	<b>325</b>
Anthi Z. Vaxevanou, Nikolaos Konstantopoulos, Damianos P. Sakas	
<b>Strategies Implemented and Sources Used for the Acquisition of Information on Foreign Markets</b>	<b>329</b>
Myropi Garri, Nikolaos Konstantopoulos and Michail G. Bekiaris	
<b>The Effect of High Performance Working Systems on Informative Technology in Enterprises after Organisation Changes such as Mergers &amp; Acquisitions</b>	<b>333</b>
Nikolaos Konstantopoulos and Yiannis Triantafyllopoulos	
<b>Personnel's Absorptive Capacity as a Guiding Concept for Effective Performance in Informative Technology</b>	<b>337</b>
Nikolaos Konstantopoulos and Yiannis Triantafyllopoulos	
<b>SESSION ON CONTEMPORARY ISSUES IN MANAGEMENT: ORGANISATIONAL BEHAVIOUR, INFORMATION TECHNOLOG, EDUCATION &amp; HOSPITAL LEADERSHIP</b>	<b>341</b>
Dr. Panagiotis Trivellas	
<b>Investigating the Importance of Sustainable Development for Hotel SMES</b>	<b>343</b>
Panagiotis Reklitis and Anestis Fotiadis	
<b>Strategic Alignment of ERP, CRM and E-business: A Value Creation</b>	<b>347</b>
Catherine C. Marinagi and Christos K. Akrivos	
<b>The Impact of Occupational Stress on Performance in Health Care</b>	<b>351</b>
Panagiotis Trivellas Panagiotis Reklitis and Charalambos Platis	

<b>The Impact of Emotional Intelligence on Job Outcomes and Turnover Intention in Health Care</b>	<b>356</b>
Panagiotis Trivellas Vassilis Gerogiannis and Sofia Svarna	
<b>SYMPOSIUM ON BUSINESS MANAGEMENT AND COMMUNICATION STRATEGIES SUPPORTING DECISION MAKING PROCESS IN TOURISM SECTOR</b>	<b>360</b>
Dr. Panagiota Dionysopoulou	
<b>The Human Factor as a Mediator to the Total Quality in the Tourism Companies. The impact of Employees' Motivation to Quality Improvements</b>	<b>362</b>
Christos K. Akrivos and Panagiotis Reklitis	
<b>Tourist Destination Marketing and Management Using Advanced ICTS Technologies</b>	<b>365</b>
Anastasia Argyropoulou, Panagiota Dionyssopoulou, Georgios Miaoulis	
<b>G.N.T.O. (Greek National Tourism Organization) Communication Strategy in Advertising Campaigns 1991-2006</b>	<b>370</b>
George Stafylakis and Panagiota Dionyssopoulou	
<b>GENERAL PAPERS</b>	<b>375</b>
<b>The role of Environmental Education within the Framework of the Environmental Policy of a Regional Municipality</b>	<b>376</b>
Vassiliki Delitheou and Dimitra Thanasia	
<b>Issues of Social Cohesion: A case study from the Greek Urban Scenery</b>	<b>380</b>
Evgenia Tousi	
<b>Merging Activity and Employee Performance: The Greek Banking System</b>	<b>384</b>
Panagiotis Liargovas and Spyridon Repousis	
<b>Sustainable Development and Corporate Social Responsibility in Higher Education: Some Evidence from Greece</b>	<b>387</b>
Anastasios Sepetis and Fotios Rizos	
<b>Exploring the Effects of Organizational Culture on Collaborative vs. Competitive Knowledge Sharing Behaviors</b>	<b>395</b>
Hanan Abdulla Mohammed Al Mehairi and Norhayati Zakaria	

# **Preface: Proceedings of the International Conference on Integrated Information (IC-ININFO 2011)**

GEORGIOS A. GIANNAKOPOULOS

*Department of Library Science and Information Systems, Technological Educational Institute of Athens, Aghiou Spyridonos, Egaleo, 12210, Greece*

DAMIANOS P. SAKAS

*Department of Computer and Technology Science, University of Peloponnese, Praxitelous 89-91, Piraeus, 18532, Greece*

## **Aims and Scope of the Conference**

The International Conference on Integrated Information 2011 took place in Kos Island, Greece, between September, 29 and October, 3, 2011. IC-ININFO is an international interdisciplinary conference covering research and development in the field of information management and integration.

The conference aims at creating a forum for further discussion for an Integrated Information Field incorporating a series of issues and/or related organizations that manage information in their everyday operations. Therefore, the call for papers is addressed to scholars and/ or professionals of the fields of Library and Archives Science (including digital libraries and electronic archives), Museum and Gallery Studies, Information Science, Documentation, Information Management, Records Management, Knowledge Management, Data management and Copyright experts the latter with an emphasis on Electronic Publications. Furthermore, papers focusing on issues of Cultural Heritage Management and Conservation Management are also be welcomed along with papers regarding the Management of Nonprofit Organizations such as libraries, archives and museums.

One of the primary objectives of the IC-ININFO will be the investigation of information-based managerial change in organizations. Driven by the fast-paced advances in the Information field, this change is characterized in terms of its impact on organizations that manage information in their everyday operations.

Grouping emerging technologies in the Information field together in a close examination of practices, problems and trends, IC-ININFO and its emphases on integration and management will present the state of the art in the field. Addressed jointly to the academic and practitioner, it will provide a forum for a number of perspectives based on either theoretical analyses or empirical case studies that will foster dialogue and exchange of ideas.

## **Topics of general Interest**

Library Science, Archives Science, Museum and Gallery Studies, Information Science, Documentation, Digital Libraries, Electronic Archives, Information Management, Records / Document Management, Knowledge Management, Data Management, Copyright, Electronic Publications, Cultural Heritage Management, Conservation Management, Management of Nonprofit Organizations, History of Information, History of Collections, Health Information

## **Symposia**

The Conference offered a number of sessions under its patronage, providing a concise overview of the most current issues and hands-on experience in information-related fields.

- Symposium on Integrated information: Theory, Policies, Tools
- 4th Symposium on Business and Management and Dynamic Simulation Models supporting management strategies

- Session on Open Access Repositories: Self-archiving, Metadata, Content policies, Usage
- Session on Evidence-Based Information in Clinical Practice
- Session on Business Management and Communication Strategies supporting Decision Making Process in Tourism Sector
- Session on Electronic Publishing: A Developing Landscape
- Session on Information and Knowledge Management
- Session on Information Content Preservation as Outcome of Conservation of Cultural Heritage: Ethics, Methodology and Tools
- Session on Advances Information for Strategic Management
- Session on Information History: Perspectives, Methods and Current Topics
- Session on Divergence and Convergence: Information Work in Digital Cultural Memory Institutions
- Session on Contemporary issues in Management: Organisational Behaviour, Information Technology, Education & Hospital leadership.

The wide range of aspects that the sessions covered, highlighted future trends in the Information Science.

### **Paper Peer Review**

More than 300 papers had been submitted for consideration in IC-ININFO 2011. From them, 91 were selected for presentation, after peer review in a double blind review process. The accepted papers were presented at IC-ININFO 2011.

### **Thanks**

We would like to thank all members that participated in any way in the IC-ININFO 2011 Conference and especially:

- The famous publishing house Emerald for its communication sponsorship.
- The co-organizing Universities and Institutes for their support and development of a high-quality Conference scientific level and profile.
- The members of the Scientific Committee that honored the Conference with their presence and provided a significant contribution to the review of papers as well as for their indications for the improvement of the Conference.
- All members of the Organizing Committee for their help, support and spirit participation before, during and after the Conference.
- The Session Organizers for their willing to organize sessions of high importance and for their editorial work, contributing in the development of valued services to the Conference.
- PhDC Marina Terzi for her excellent editorial work, contributing in the production of the Conference proceedings.

## CONFERENCE DETAILS

### Chairs

Georgios A. Giannakopoulos, Technological Educational Institute of Athens, Greece  
Damianos P. Sakas, University of Peloponnese, Greece

### Co-Chairs

Daphne Kyriaki – Manesi, Technological Educational Institute of Athens, Greece  
Dimitrios Vlachos, University of Peloponnese, Greece

### Scientific Committee

Amanda Spink, Queensland University of Technology  
Andreas Bagias, European Court  
Andreas Rauber, Vienna University of Technology  
Astrid van Wesenbeeck, SPARC Europe  
Christine Urquhart, Aberystwyth University  
Christos Schizas, University of Cyprus  
Christos Skourlas, Technological Educational Institute of Athens  
Claire Farago, University of Colorado at Boulder  
Claus-Peter Klas, FernUniversität in Hagen  
Costas Vassilakis, University of Peloponnese,  
Dimitris Dervos, Technological Educational Institute of Thessaloniki  
Eelco Ferwerda, OAPEN  
Elena Garcia Barriocanal, University of Alcalá  
Emmanuel Garoufallou, Technological Educational Institute of Thessaloniki  
Filippos Tsimpoglou, University of Cyprus  
Fillia Makedon, University of Texas at Arlington  
George Korres, University of Newcastle  
Georgios Evangelidis, University of Macedonia  
Georgios Panagiaris, Technological Educational Institute of Athens  
Johan Oomen, Netherlands Institute for Sound and Vision  
José Aldana, University of Malaga  
Konstantinos Masselos, University of Peloponnese  
Luciana Duranti, The University of British Columbia  
Markos N. Dendrinis, Technological Institute of Athens  
Milena Dobрева, University of Strathclyde  
Prodromos Tsiavos, London School of Economics and Political Science  
Sándor Darányi, University of Borås  
Sarantos Kapidakis, Ionian University  
Sirje Virkus, Tallinn University  
Spiros Zervos, Technological Educational Institute of Athens  
Susan Myburgh, University of South Australia  
Theodoros Pitsios, University of Athens, Faculty of Medicine

### Organizing Committee

Alexandros Koulouris (Chair), Technological Educational Institute of Athens  
Christos Christopoulos, SCEV Scientific Events Ltd  
Marina Terzi, University of the Aegean, Greece  
Evangelia Markaki, Aristotle University of Thessaloniki



Assimina Kaniari, Athens School of Fine Arts  
Evangelia Lappa, General Hospital Attikis K.A.T.  
Dimitris Kouis, Greek Ministry of Education, Lifelong Learning and Religious Affairs  
Dionysis Kokkinos, National Technical University of Athens

## KEYNOTE SPEAKER



Professor Amanda Spink

Professor Amanda Spink has published over 340 scholarly journal articles, refereed conference papers and book chapters, and 6 books. Many of her journal articles are published in the Journal of the American Society for Information Science and Technology, Information Processing and Management, and the Journal of Documentation. She is Editor of the Emerald journal Aslib Proceedings. Amanda's research has been published at many conferences including ASIST, IEEE ITCC, CAIS, Internet Computing, ACM SIGIR, and ISIC Conferences. Her recent books include Information Behavior: An Evolutionary Instinct and Web Search: Multidisciplinary Perspectives, both published by Springer. Amanda's research focuses on theoretical and empirical studies of information behavior, including the evolutionary and developmental foundations. The National Science Foundation, the American Library Association, Andrew R. Mellon Foundation, Amazon.com, Vivisimo. Com, Infospace.com, NEC, IBM, Excite.com, AlltheWeb.com, AltaVista.com, FAST, and Lockheed Martin have sponsored her research. In 2008 Professor Spink had the second highest H-index citation score in her field from 1998 to 2008 [Norris, M. (2008)]. Ranking Fellow Scholars and their H-Index: Preliminary Survey Results. Loughborough University, Dept of Information Science Report].

# Text Segmentation Using Named Entity Recognition and Co-Reference Resolution in Greek Texts

Pavlina Fragkou<sup>†</sup>

<sup>†</sup> *Technological Educational Institution of Athens (TEI-A), Dept. of Library Science And Information Systems,  
Ag. Spyridonos, 12210, Egaleo, Athens, Greece.  
pfragkou(at)teiath.gr*

**Abstract:** *In this paper we examine the benefit of performing named entity recognition and co-reference resolution to a Greek corpus used for text segmentation. Segments consist of portions among one of the 300 documents published by ten different authors in the Greek newspaper "To Vima". The aim here is to examine whether the combination of text segmentation and information extraction (and most specifically the named entity recognition and co-reference resolution steps) can prove to be beneficial for the identification of the various topics that appear in a document. Named entity recognition was performed using an already existing tool which was trained on a similar corpus. The produced annotations were manually corrected and enriched in order to cover four types of named entities (i.e. person name, organization, location and time). Co-reference resolution and most specifically substitution of every reference of the same instance with the same named entity identifier was performed in a subsequent step. The evaluation using three well known text segmentation algorithms leads to the conclusion that, the benefit highly depends on the segment's topic, the number of named entity instances appearing in it, as well as the segment's length.*

**Keywords:** *Text segmentation, Named entity recognition, Co-reference resolution, Information extraction.*

## I. INTRODUCTION

The information explosion of the web aggravates the problem of effective information retrieval. To address this, various techniques such as text segmentation and information extraction provide partial solutions to the problem. More specifically, text segmentation methods are useful in identifying the different topics that appear in a document. On the other hand, information extraction methods try to identify portions of text that refer to a specific topic, by focusing on the appearance of instances of specific types of named entities (such as person, date, location, etc.) according to the thematic area of interest.

The question that arises is whether the combination of text segmentation and information extraction (and most specifically the named entity recognition and co-reference resolution steps) can prove to be beneficial for the identification of the various topics that appear in a document.

This paper examines the benefit of performing named entity recognition and co-reference resolution on a Greek corpus consisting of portions of documents taken from the Greek newspaper "To Vima". This corpus was previously used for examining the performance of text segmentation algorithms (Fragkou *et al.*, 2007). It must be stressed that, the focus is not on finding the algorithm that achieves the best segmentation performance on the corpus, but on the benefit of performing named entity recognition as well as co-reference resolution on a corpus used for text segmentation.

The structure of the paper is as follows. Section II provides an overview of related methods. Section III presents the steps performed for the creation of the "annotated" corpus. Section IV presents evaluation results obtained by using three well known text segmentation algorithms, while Section V provides conclusions and future steps.

## II. RELATED WORK

The text segmentation problem of concatenated text can be stated as follows: given a text which consists of several parts (each part dealing with a different subject), it is required to find the boundaries between the parts. A starting point to this is the calculation of the within-segment similarity based on the assumption that, parts of a text having similar vocabulary are likely to belong to a coherent topic segment. It must be stressed that, within-segment similarity is calculated on the basis of words but not on the basis of the application of other more sophisticated techniques such as named entity recognition or co-reference resolution. In the literature, several word co-occurrence statistics are proposed (Choi, 2000; Choi *et al.*, 2001; Hearst, 1997; Utiyama and Isahara, 2001). A significant difference between text segmentation methods is that, some authors evaluate the similarity between *all* parts of a text (Choi, 2000; Choi *et al.*, 2001; Ponte and Croft, 1997; Reynar, 1994; Xiang and Hongyuan, 2003), while other between adjacent parts (Hearst, 1997; Heinonen, 1998). To penalize deviations from the expected segment length, several methods use the notion of "length model" (Heinonen, 1998; Ponte and Croft, 1997). Dynamic programming is often used in order to calculate the globally minimal segmentation cost (Heinonen, 1998; Reynar, 1994; Xiang and Hongyuan, 2003; Kehagias *et al.*, 2004; Qi *et al.*, 2008). Current approaches involve the improvement of the dotplotting technique (Yen *et al.*, 2005), the improvement of Latent Semantic

Analysis (Bestgen, 2006) and the improvement of Hearst's TextTiling method (Hearst, 1997) presented by Kern and Granitzer (2009).

Information extraction, from a different point of view, aims to locate within a text passage domain-specific and pre-specified facts (e.g., in a passage about athletics, facts about the athlete participating in a 100m event, such as name, nationality, performance, as well as facts about the specific event, like the event name). More specifically, information extraction is about - among others - extracting from texts: (a) *Entities*: textual fragments of particular interest, such as persons, places, organizations, dates, etc. (b) *Mentions*: the identification of all lexicalisations of an entity in texts. For example, the name of a particular person can be mentioned in different ways inside a single document, such as "Lebedeva", "Tatiana Lebedeva", or "T. Lebedeva". The following pre-processing steps are applied in order to perform information extraction: (a) *Named Entity Recognition*, where entity mentions are recognized and classified into proper types for the thematic domain in question (b) *Co-reference*, where all the mentions that represent the same entity are identified and grouped together according to the entity they refer to.

Co-reference resolution complementary includes the step of anaphora resolution. The term anaphora denotes the phenomenon of referring to an entity already mentioned in a text -most often with the help of a pronoun or a different name. Co-reference basically involves the following steps: (a) pronominal co-reference (which is about finding the proper antecedent for personal pronouns), possessive adjectives, possessive pronouns, reflexive pronouns and pronouns this and that (b) identification of cases where both the anaphor and the antecedent refer to identical sets or types. This identification requires some world knowledge or specific domain knowledge. It also includes cases such as reference to synonyms or the case where the anaphor matches exactly or is a substring of the antecedent (c) ordinal anaphora for cardinal numbers and adjectives such as "former" and "latter".

The importance of text segmentation and information extraction is apparent in a number of applications, such as noun phrase chunking, tutorial dialogue segmentation, focused crawling, text summarization, semantic segmentation and web content mining. In Fragkou (2011) the use of information extraction techniques in the text segmentation process was examined on an English corpus. In this paper the same problem is examined in a Greek text corpus. Those techniques are applied on a corpus used for text segmentation, resulting in the creation of an "annotated" corpus. Evaluation was performed using three well-known segmentation algorithms (Choi *et al.*, 2001; Kehagias *et al.*, 2004; Utiyama and Isahara, 2001) applied both in the original as well as the "annotated" corpus.

To the best of our knowledge no similar work appears in the literature that combines named entity recognition and co-reference resolution to assist the text segmentation task performed in a Greek corpus.

### III. METHOD

Existing algorithms performing text segmentation exploit a variety of word co-occurrence statistic techniques in order to calculate the homogeneity between segments, where each segment refers to a single topic. However, they do not exploit the importance that several words may have in a specific context. Examples of such words are person names, locations, dates, group of names, scientific terms etc. The importance of those terms is further diminished by the application of word processing techniques, i.e., stop list removal and stemming on words such as pronouns or adjectives. We aim to exploit whether the identification of such words can be beneficial for the segmentation task. This identification requires the application of named entity recognition and co-reference resolution thus, their (manual or not) annotation effort is under examination.

#### A. The Corpus

While several papers regarding the segmentation of English texts have appeared in the literature, little work was performed for Greek texts. It should be stressed that, due to the fact that Greek is a highly inflected language the segmentation problem is harder for Greek. More specifically, to the best of our knowledge the only work that refers to segmentation of Greek texts is that presented in Fragkou *et al.* (2007). There, the authors used a text collection compiled from Stamatatos's corpus (Stamatatos *et al.*, 2001) comprising of text downloaded from the website <http://tovima.dolnet.gr>. Stamatatos *et al.* (2001) constructed a corpus collecting texts which includes essays on Biology, Linguistics, Archeology, Culture, History, Technology, Society, International Affairs and Philosophy from ten different authors, where 30 texts were selected from each author. Table 1 lists the authors contributing to Stamatatos *et al.* (2001) collection as well as the thematic area(s) dealt by each of them.

In the work presented in Fragkou *et al.* (2007) each of the 300 texts of the collection of articles compiled from this newspaper was pre-processed using the POS tagger developed Orphanos and Christodoulakis (1999). The tagger is based on a Lexicon capable of assigning full morphosyntactic attributes to 876.000 Greek word forms. In Fragkou *et al.* (2007) experiments, every noun, verb, adjective or adverb in the text was substituted by its lemma, determined by the tagger. For those words that their lemma was not determined by the tagger, no substitution was made. Fragkou *et al.* (2007) created two groups of experiments (which are described in details in Section IV) whose difference lies in the length of the created segments and the number of authors used for the creation of the texts to segment, where each text

being a concatenation of ten text segments. Each author was characterized by her/his vocabulary, hence Fragkou *et al.* (2007) goal was to segment the text into the parts written by the various authors.

<i>Author</i>	<i>Thematic Area</i>
Alachiotis	Biology
Babiniotis	Linguistics
Derilis	History, Society
Kiosse	Archeology
Liakos	History, Society
Maronitis	Culture, Society
Ploritis	Culture, History
Tassios	Technology, Society
Tsulakas	International affairs
Vokos	Philosophy

**Table 1.** List of authors and thematic areas dealt by each of them.

## B. Named Entity Annotation

The named entity recognition task for Greek texts has been examined in the literature by a number of researchers. Previously published work on Greek Named Entity recognition usually relies on hand-crafted rules or patterns, (Boutsis *et al.*, 2000; Farmakiotou *et al.*, 2000; Farmakiotou *et al.*, 2002) and/or decision tree induction with C4.5 (Karkaletsis *et al.*, 1999; Petasis *et al.*, 2001). The only exception is the work of Diamantaras *et al.*, (2005) and Michailidis *et al.* (2006), where SVMs, Maximum Entropy, Onetime and manually crafted post-editing rules were employed. Special attention must be given to two works. The first is the one presented in (Papageorgiou *et al.*, 2002), in which the problem of pronominal anaphora resolution was examined. The second is the one proposed by Lucarelli *et al.* (2006) where a freely available named-entity recognizer for Greek texts was constructed, which identifies temporal expressions, person and organization names. For temporal expressions, the named entity recognizer uses manually constructed token lists and automatically generalized regular expression patterns. For person and organization names, it uses an ensemble of SVMs that scan the input text in two passes. The second pass takes into account the decisions of the first, which allows it to learn how to correct mistakes of the first pass. It also considers whether or not the first pass has classified a token elsewhere in the same text as a person or organization name with high confidence. This allows it to identify re-occurrences of person and organization names in more difficult contexts. Apart from its two-pass architecture, another novelty of by Lucarelli *et al.* 's (2006) named entity recognizer is the use of active learning, which allows the system to select by itself candidate training instances to be annotated by a human during training. The aforementioned system is

one of the very few named entity recognizer, regardless of language, that have exploited active learning.

For our experiments, we used the corpus created by Fragkou *et al.* (2007) and applied the annotation tool implemented by Lucarelli *et al.* (2006). This annotation tool was chosen due to the fact that, it is publicly available and it was trained on documents taken from the newspaper "Ta Nea" having similar content with that of the newspaper "To Vima". The annotation tool was thus applied in our corpus without requiring training. For the annotation, he have chosen four types of named entities i.e. person name, group name, location and date. The application of the annotation tool produced annotations for some but not all instances of person names, group names and dates. In order to annotate all named entities appearing in each text a second pass was performed. During this pass, in each text manual named entity annotation of proper names belonging to one of the four categories was performed. We believe that the substitution of words with named entity instances does not have an effect in the performance of a segmentation algorithm. Based on this, during manual named entity annotation, we additionally: (a) annotated all instances of locations (b) substituted every reference of the same instance with the same named entity identifier. For example in the sentences "James P. Mitchell and Sen. Walter H. Jones R-Bergen, last night disagreed on the value of using as a campaign issue a remark by Richard J. Hughes,... . Mitchell was for using it, Jones against", we first identified three instances of person names. We further used the same entity identifier for James P. Mitchell and Mitchell and the same entity identifier for Sen. Walter H. Jones R-Bergen and Jones (c) we substituted every reference of the same instance, resulted from co-reference resolution, with the same named entity identifier (for example in the sentences "Mr. Hawksley, the state's general treasurer,... He is not interested in being named a full-time director", we substituted He with the named entity identifier given to Mr. Hawksley).

Group names involved expressions such as "House Committee on Revenue and Taxation" or "City Executive Committee". The annotation of location instances included possible derivations of them such as "Russian". The annotation of date instances included both simple date form (consisting only of the year or month) and more complex forms (containing both month, date and year). It must be stressed that, co-reference resolution was performed only on portions of text that refer to named entity instances and not on the text as a whole.

The annotation process led to the conclusion that, texts having social subject usually contain a small number of named entity instances. On the other hand, texts issuing politics, science, archeology, history and philosophy usually contain an important number of named entity instances. For example, texts belonging to the author Kiosse contain on average an important number of named entities because they describe

historical events issuing person names, dates and locations.

<i>Author</i>	<i>No. of docs per set</i>
Alachiotis	44
Babiniotis	70.23
Derilis	33.33
Kiosse	121.9
Liakos	77.7
Maronitis	40.4
Ploritis	94.2
Tassios	40
Tsulakas	37.12
Vokos	52.16

**Table 2.** Statistics regarding the average number of named entity instances appearing in the annotated documents per author.

This can be also justified by the fact that, the Stamataos *et al.* (2001) corpus deals with a number of areas i.e. Biology, Linguistics, Archeology, Culture, History Technology, Society, International Affairs, Philosophy. It must be stressed that, the annotation took place before the application of the Orphanos and Christodoulakis (1999) POS tagger and the selection of lemmas that are either noun or verb or adjective or adverb, determined by the tagger.

#### IV. EVALUATION

The "annotated" corpus that resulted from the previously described process was evaluated using three text segmentation algorithms. The first is Choi's C99b (Choi, 2001), which creates a similarity matrix for sentences appearing in a text using Latent Semantic Analysis. C99b then finds topic boundaries by recursively seeking the optimum density along the matrix diagonal. The second algorithm is the one proposed by Utiyama and Isahara (2001). This algorithm finds the optimal segmentation of a given text by defining a statistical model which calculates the probability of words to belong to a segment. To find the maximum probability segmentation, it calculates the minimum-cost segmentation obtained by the minimum cost path in a graph. Both algorithms benefit from the fact that, they do not require training and they are publicly available.

The third algorithm used is introduced by Kehagias *et al.* (2004) which, contrary to the previous ones, requires training. More specifically, this algorithm uses dynamic programming to find both the number and the location of segment boundaries. The algorithm decides the locations of boundaries by calculating the globally optimal splitting (i.e., global minimum of a segmentation cost) on the basis of a similarity matrix, a preferred fragment length, and a defined cost function.

#### A. Experiments – Results

We evaluate the performance of the algorithms in the original and "annotated" corpus using three widely known indices: Precision, Recall Beeferman's *Pk* metric (Beeferman at al., 1999) and WindowDiff (Pevzer and Hearst, 2002). Precision is defined as "the number of the estimated segment boundaries which are actual segment boundaries" divided by "the number of the estimated segment boundaries". Recall is defined as "the number of the estimated segment boundaries which are actual segment boundaries" divided by "the number of the true segment boundaries". Beeferman's metric *Pk* measures the proportion of "sentences which are wrongly predicted to belong to different segments (while they actually belong in the same segment)" or "sentences which are wrongly predicted to belong to the same segment (while they actually belong in different segments)". A variation of the *Pk* measure named WindowDiff index was proposed by Pevzer and Hearst (2002) and remedies several of *Pk*'s problems. The subsections that follow contain the results of the experiments that were performed in the two groups of experiments and compare the obtained results with those appearing in the literature for the same task.

#### B. Experiment group 1

The collection of texts used for the first group of experiments consists of 6 datasets: Set0,..., Set5. Each of those datasets differ in the number of authors used for the generation of the texts to segment and consequently in the number of texts used from the entire collection, as listed in Table 3.

<i>Dataset</i>	<i>Authors</i>	<i>No. of docs per set</i>
Set0	Kiosse, Alachiotis	60
Set1	Kiosse, Maronitis	60
Set2	Kiosse, Alachiotis, Maronitis	10
Set3	Kiosse, Alachiotis, Maronitis, Ploritis	120
Set4	Kiosse, Alachiotis, Maronitis, Ploritis, Vokos	150
Set5	All Authors	300

**Table 3.** List of the sets compiled in the 1rst group of experiments using Greek texts and the author's texts used for each of them.

For each of the above datasets, four subsets were constructed which differ in the number of the sentences appearing in each segment. Let  $L_{\min}$  and  $L_{\max}$  be the smallest and largest number of sentences which a segment may contain. Four different  $(L_{\min}, L_{\max})$  pairs were used: (3,11), (3,5), (6,8) and (9,11). Hence Set0 contains 4 subsets: Set01, Set02, Set03 and similarly for Set1, Set2, ..., Set5. The datasets Set\*1 are the ones with  $(L_{\min}, L_{\max}) = (3,11)$ , the datasets Set\*2 are the ones with  $(L_{\min}, L_{\max}) = (3,5)$ ,

and so on. Let also  $\{X_1, \dots, X_n\}$  be the authors contributing to the generation of the dataset. Texts belonging in each dataset are generated by the following procedure:

Each text is the concatenation of ten segments. For each segment we do the following.

1. An author from  $\{X_1, \dots, X_n\}$  is randomly selected. Let  $I$  be the selected author.
2. A text among the 30 available that belong to the  $I$  author is randomly selected. Let  $k$  be the selected text of author  $I$ .
3. A number  $l \in (L_{\min}, L_{\max})$  is randomly selected.
4.  $l$  consecutive lines from text  $k$  (starting at the first sentence of the text) are extracted. Those sentences constitute the generated segment.

Algorithm	Dataset	Precision	Recall	$P_K$	Window diff
<b>Kehagias et al. (2004)</b>	Set*1 (3-11)	64.90%	61.77%	15.69%	15.59%
	Set*2 (3-5)	85.13%	85.11%	6.45%	6.27%
	Set*3 (6-8)	90.51%	90.51%	2.54%	2.42%
	Set*4 (9-11)	91.92%	91.92%	1.29%	1.21%
<b>Choi's C99b</b>	Set*1 (3-11)	61.64%	61.66%	18.43%	19.37%
	Set*2 (3-5)	71.70%	71.70%	16.93%	17.93%
	Set*3 (6-8)	68.29%	68.29%	15.37%	15.89%
	Set*4 (9-11)	66.75%	66.75%	13.93%	14.07%
<b>Utiyama &amp; Isahara (2001)</b>	Set*1 (3-11)	64.00%	61.10%	17.37%	17.47%
	Set*2 (3-5)	70.00%	54.70%	20.79%	21%
	Set*3 (6-8)	75.42%	73.03%	10.84%	10.96%
	Set*4 (9-11)	73.13%	74.29%	8.83%	8.91%

**Table 4.** The Precision, Recall,  $P_K$  and WindowDiff values obtained by all algorithms for the 1rst group of experiments, without use of named entities.

Table 4 lists the values of Precision, Recall,  $P_K$  and WindowDiff reported in the literature after applying Kehagias *et al.* (2004), Choi's C99b and Utiyama and Isahara's (2001) algorithms on the same task averaged over all datasets which have segments of same length. Table 5 lists the values of Precision, Recall,  $P_K$  and WindowDiff obtained after applying the same

algorithms on the same datasets where annotation was previously performed.

We reach the following conclusions based on the obtained results. Regarding the algorithm of Utiyama and Isahara (2001), a significant improvement was obtained in all measures and for all datasets of the Experiment Group 1. This can be justified by the fact that, Utiyama and Isahara's algorithm (2001) performs global optimization of local information in contrast to Choi's C99 and Kehagias and al., (2004) algorithms which perform local optimization of global information and global optimization of global information respectively. The same observation holds for the results obtained after applying the two other algorithms where improvement was obtained in all datasets and all evaluation metrics. This improvement appears to be more important in datasets Set \*1(3-11) and Set \*2(3-5) in all algorithms. This is an indication that annotation succeed in identifying critical information which, in other ways, was lost. For datasets Set \*3(6-8) and Set \*4(9-11) the segmentation accuracy remains high. This is justified by the fact that, in those datasets the segment length is high leading to a high number of named entity instances.

Algorithm	Dataset	Precision	Recall	$P_K$	Window diff
<b>Kehagias et al. (2004)</b>	Set*1 (3-11)	70.12%	67.92%	13.12%	13.03%
	Set*2 (3-5)	87.58%	87.48%	5.15%	4.96%
	Set*3 (6-8)	92.29%	92.29%	2.04%	1.93%
	Set*4 (9-11)	93.11%	93.11%	1.10%	1.02%
<b>Choi's C99b</b>	Set*1 (3-11)	63.26%	63.26%	15.96%	17.40%
	Set*2 (3-5)	70.46%	70.46%	14.53%	15.91%
	Set*3 (6-8)	71.26%	71.26%	11.92%	12.45%
	Set*4 (9-11)	68.46%	68.46%	11.43%	11.89%
<b>Utiyama &amp; Isahara (2001)</b>	Set*1 (3-11)	70.74%	66.96%	13.72%	13.63%
	Set*2 (3-5)	76.65%	61.55%	16.83%	16.66%
	Set*3 (6-8)	80.31%	78.18%	8.43%	8.32%
	Set*4 (9-11)	76.75%	78.40%	7.15%	7.07%

**Table 5.** The Precision, Recall,  $P_K$  and WindowDiff values obtained by all algorithms for the 1rst group of experiments with use of named entities.

### C. Experiment group 2

The second group of experiments also uses Stamatatos *et al.* (2001) collection. There, a single dataset was constructed which contains 200 texts, with every author represented in each text. Each text is the concatenation of ten segments. More specifically, the construction of each segment is performed as follows:

1. An author among the 10, named  $I$  is randomly selected.
2. A text (named  $k$ ) among the 30 available that belong to the  $I$  author is randomly selected. Let  $Z$  be the number of paragraphs that  $k$ -th text contains.
3. A number  $l$  ( $1 < l < Z$ ) corresponding to the number of paragraphs that the generated segment will contain is randomly selected.
4. A number  $m$  ( $1 < m < Z-l$ ) corresponding to the "starting paragraph" was randomly selected. Thus, the segment contains all the paragraphs of text  $k$  starting from paragraph  $m$  and ending at the paragraph  $m+l$ .

The procedure described above produced segments and consequently concatenated texts which were longer than the ones used in Experiment Group 1. Hence the segmentation task in the current group is more difficult than the previous one. Table 6 lists the values of Precision, Recall,  $P_k$  and WindowDiff reported in the literature after applying Kehagias *et al.* (2004), Choi's C99b and Utiyama and Isahara's (2001) algorithms on the original i.e. non-annotated corpus. Table 7 lists the values of Precision, Recall,  $P_k$  and WindowDiff obtained after applying the same algorithms on this unique dataset where annotation was previously performed.

Algorithm	Precision	Recall	$P_k$	Window Diff
Kehagias et al. (2004)	60.60%	57.00%	11.07%	11.06%
Choi's C99b	44.62%	44.62%	19.44%	21.62%
Utiyama & Isahara (2001)	56.76%	67.22%	12.28%	12.26%

**Table 6.** The Precision, Recall,  $P_k$  and WindowDiff values obtained by all algorithms for the 2nd group of experiments, without use of named entities.

Algorithm	Precision	Recall	$P_k$	Window Diff
Kehagias et al. (2004)	63.46%	62%	9.06%	9.30%
Choi's C99b	49.4%	49.4%	18.12%	20.47%
Utiyama & Isahara (2001)	59.78%	69%	10.83%	13.57%

**Table 7.** The Precision, Recall,  $P_k$  and WindowDiff values obtained by all algorithms for the 2nd group of experiments, with use of named entities.

It can be seen that, the segmentation performance was improved in the annotated corpus for all accuracy metrics and in all algorithms. This is justified by the fact that, in these dataset the segment length is high leading to a high number of named entity instances. It must be stressed that, co-reference resolution contributed significantly to the increase of the number of entity instances per segment.

We also draw attention to the fact that, the type of named entity instance acts indirectly as a discriminative factor in the segmentation process. This is in contrast with information extraction, where the learning process takes into account the type of named entities occurring in a passage of text.

### V. CONCLUSIONS

In this paper we evaluated the benefit of incorporating information extraction techniques to enhance the performance of text segmentation algorithms. More specifically, we performed manual named entity recognition and co-reference resolution on a Greek corpus used by text segmentation algorithms. We then compared the performance of three well-known segmentation algorithms in both the original and the resulting "annotated" corpus. The results obtained show that, the benefit resulted from the use of annotation is apparent in all algorithms and for all metrics and datasets. The contribution of co-reference resolution in this improvement is high and deserves special attention. The potential benefit of the annotation is strongly related to the segment's topic as well as the number of named entity instances appearing in it. This approach may further prove beneficial for other problems, such as web mining and focused crawling.

We outlook several directions of future work. The first direction considers performing text segmentation on a different corpora with fewer topics than the one used, such as the corpus used by Lucarelli *et al.* (2006) as well as the one used by Papageorgiou *et al.* (2002) where co-reference resolution was also performed. The second direction is oriented towards the examination of other named entity recognition systems with special attention to those containing co-reference resolution tools. Regarding co-reference resolution, focus will be given to the types of co-reference examined as well as their scope (i.e. the examination of co-reference within the same sentence and/or with the previous one appearing in the text).

We further seek to examine the addition of other types of named entities that will be more oriented to the segment's topic. In the same direction lies the extraction and annotation of relations between named entities and the examination of their contribution to the segmentation task. The aim is to reinforce the role and identity of named entities in the segmentation process. Finally, it is interesting to examine the identification of events related to specific named entity types.



## REFERENCES

- Beeferman, D., Berger, A. and Lafferty, J. "Statistical models for text segmentation", *Machine Learning*, **34**, 177-210 (1999).
- Bestgen, Y. "Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings Deterministic and Moore", *Computational Linguistics*, **1**, 5-12 (2006).
- Boutsis S., Demiros I., Giouli V., Liakata M., Papageorgiou H. and Piperidis S. "A system for recognition of named entities in Greek", *In Proceedings of the 2nd International Conference on Natural Language Processing*, 424-435, Patra, Greece (2000).
- Choi, F.Y.Y. "Advances in domain independent linear text segmentation", *In Proc. of the 1st Meeting of the North American Chapter of the ACL*, 26-33 (2000).
- Choi, F.Y.Y., Wiemer-Hastings, P. and Moore, J. "Latent semantic analysis for text segmentation", *In Proceedings of the 6th Conf. on EMNLP*, 109 - 117 (2001).
- Diamantaras K., Michailidis I. and Vasileiadis S. "A very fast and efficient linear classification algorithm", *In Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, Mystic, CT* (2005).
- Farmakiotou D., Karkaletsis V., Koutsias J., Sigletos G., Spyropoulos C.D. and Stamatopoulos P. "Rule-based named entity recognition for Greek financial texts", *In Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries*, 75-78 (2000).
- Farmakiotou D., Karkaletsis V., Samaritakis G., Petasis G. and Spyropoulos C.D. "Named entity recognition in Greek Web pages", *In Proceedings of the 2nd Hellenic Conference on Artificial Intelligence, companion volume*, 91-102 (2002).
- Fragkou, P., Petridis, V. and Kehagias, A. "Segmentation of Greek Text by Dynamic Programming", *In Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, **2**, 370-373 (2007).
- Fragkou P. "Text Segmentation using Named Entity Recognition and Co-reference Resolution", *In ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence, Volume 1 - Artificial Intelligence, Rome*, 349-354 (2011).
- Hearst, M. A. "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages", *Computational Linguistics*, **23(1)**, 33-64 (1997).
- Heinonen, O. "Optimal Multi-Paragraph Text Segmentation by Dynamic Programming", *In Proc. of 17th COLING -ACL '98*, 1484-1486 (1998).
- Karkaletsis V., Paliouras G., Petasis G., Manousopoulou N., and Spyropoulos C.D. "Named-entity recognition from Greek and English texts", *Intelligent and Robotic Systems*, **26**, 123-135 (1999).
- Kehagias, Ath., Nicolaou A., Fragkou P. and Petridis V. "Text Segmentation by Product Partition Models and Dynamic Programming", *Mathematical & Computer Modeling*, **39**, 209-217 (2004).
- Kern, R. and Granitzer, M. "Efficient linear text segmentation based on information retrieval techniques", *In Proceedings of the International Conference on Management of Emergent Digital EcoSystems* (2009).
- Lucarelli G., Vasilakos X. and Androutsopoulos I. "Named Entity Recognition in Greek Texts with an Ensemble of SVMs and Active Learning", *International Journal on Artificial Intelligence Tools*, **16(6)**, 1015-1045 (2007).
- Michailidis I., Diamantaras K., Vasileiadis S. and Frere Y. "Greek named entity recognition using Support Vector Machines, Maximum Entropy and Onetime", *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, 45-72 (2006).
- Orphanos, G. and Christodoulakis, D. "Part-of-speech disambiguation and unknown word guessing with decision trees", *In Proc. of EACL '99*.
- Papageorgiou, H., Prokopoulos, P., Demiros, I., Giouli, V., Konstantinidis, A. and Piperidis, S. "Multi-level XML-based Corpus Annotation", *In Proceedings of the 3rd Language Resources and Evaluation Conference, Las Palmas* (2002).
- Petasis G., Vichot F., Wolinski F., Paliouras G., Karkaletsis V. and Spyropoulos C.D. "Using machine learning to maintain rule-based named-entity recognition and classification systems", *In Proceedings of the 39th Annual Meeting of ACL and 10th Conference of EACL*, 426-433 (2001).
- Pevzner, L. and Hearst, M. "A critique and improvement of an evaluation metric for text segmentation", *Computational Linguistics*, **28(1)**, 19-36 (2002).
- Ponte, J. M. and Croft, W. B. "Text segmentation by topic", *In Proc. of the 1st Europ. Conf. on Research and Advanced Technology for Digital Libraries*, 120 - 129 (1997).
- Qi S., Runxin L., Dingsheng L. and Xihong W. "Text segmentation with LDA-based Fisher kernel", *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, 269-272 (2008).
- Reynar, J.C. "An automatic method of finding topic boundaries", *In Proc. of the 32nd Annual Meeting of the ACL*, 331-333 (1994).
- Stamatatos, E., Fakotakis, N. and Kokkinakis, G. "Computer-based authorship attribution without lexical measures", *Computer and the Humanities, Kluwer Academic Publishers*, **35**, 193-214 (2001).

- Utiyama, M. and Isahara, H. "A statistical model for domain independent text segmentation", *In Proc. of the 9th EACL*, 491-498 (2001).
- Xiang J. and Hongyuan Z. "Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming", *In Proc. of the 26th ACM SIGIR Conf* (2003).
- Ye, N., Zhu, J., Luo H., Wang, H. and Zhang, B. "Improvement of the dotplotting method for linear text segmentation", *In Proc of Natural Language Processing and Knowledge Engineering*, 636- 641 (2005).