EEOP2012: Exploring and Exploiting Official Publications

Sunday May 27, 2012

ABSTRACTS

Editor:

Steven Krauwer
Workshop Programme

Sunday May 27, 2012

09:00-09:05 Welcome and introduction by Steven Krauwer

09:05-09:50 **Invited talk:** Maarten Marx
*Open Official Documents: Requirements and Opportunities*

09:50-10:10 Michael Rosner and Andrew Attard
*Intelligent Exploitation of Local Government Resources*

10:10-10:30 Maria Palmerini, Ruben Cerolini, Giulio Santini and Francesco Cutugno
*From Recording to Retrieving: A Proposal of a Complete System for Semi-automatic Reporting for Local and National Governments*

10:30-11:00 Coffee break

11:00-11:20 Vidas Daudaravicius
*Automatic Multilingual Annotation of EU Legislation with Eurovoc Descriptors*

11:20-11:40 Francesca Frontini, Carlo Aliprandi, Clara Bacci, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno and Tiziana Soru
*GLOSS, an Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain*

11:40-12:00 Oliver Mason, Aleksander Trklja and Dominik Vajn
*Requirement Extraction from Transport Policy Documents*

12:00-12:50 Open discussion
*What are possible actions that could be undertaken to enhance the exploration and exploitation of official publications at the international, cross-national and national level?*

12:50-13:00 Winding up and closing by Steven Krauwer
# Workshop Organizers and Programme Committee

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steven Krauwer</td>
<td>Utrecht University / CLARIN ERIC</td>
</tr>
<tr>
<td>Ralf Steinberger</td>
<td>European Commission – Joint Research Centre (JRC)</td>
</tr>
<tr>
<td>Arjan van Hessen</td>
<td>University of Twente / CLARIN-NL</td>
</tr>
<tr>
<td>Nicoletta Calzolari</td>
<td>CNR-ILC / ELRA</td>
</tr>
<tr>
<td>Hans Uszkoreit</td>
<td>DFKI / META-NET</td>
</tr>
</tbody>
</table>
Introduction

The EEOP2012 workshop is dedicated to the exploitation and exploration of official publications in digital format, both at the international level (often multilingual) and at the national level (mostly monolingual, but in some cases multilingual as well). These publications can be in written, spoken or visual form or combinations thereof (e.g. written proceedings of parliaments, legislative documents, audio or video recordings of parliament sessions, simultaneous translations by interpreters or in sign language).

The workshop covers the whole lifecycle of these publications, ranging from acquisition, annotation, instrumentation, exploration of data and content, exploitation of data and content to support research and the development of tools and applications.

The main objectives of the workshop are:

• To create awareness of the importance of official publications by showing the research and development possibilities they offer;
• To share results, experiences and problems emerging from work on a variety of corpora, modalities and languages;
• To identify actions that could be undertaken to enhance the exploration and exploitation of official publications at the international, cross-national and national level.

Official publications can be of tremendous importance for the research communities interested in human language technology (in the broadest possible sense) and for the communities interested in linguistics, psychology, history, social sciences and political sciences because they have a number of specific characteristics that make them different from other language resources:

• If they exist in digital form they are normally public and free;
• They grow continuously;
• They are often multilingual and parallel;
• They lend themselves for exploitation (as training material for tools and sometimes possibly even for niche applications);
• They lend themselves for exploration to support linguistic studies, studies about human behaviour, about changes in society, attitudes, and many other possible research topics in the humanities and social sciences;
• Because of their comparability they lend themselves for porting technologies, methods and expertise between languages;
• They lend themselves for educational purposes for technologists, linguists and other scholars.

Primary audience of this workshop is:

• Language and speech technology researchers from academia and industry;
• Humanities and social sciences scholars with an interest in digital methods;
• Educators in these fields.

Additional beneficiaries, not necessarily present at LREC 2012:

• Professionals interested in analysing political behaviour or processes (e.g. journalists, policy makers, policy watchers);
• Parties interested in providing or exploiting such analysis tools on a commercial basis;
• Translation studies scholars;
• Comparative linguists.

Acknowledgement: This workshop was supported by the European Commission – Joint Research Centre (JRC)
Open Official Documents: Requirements and Opportunities (*Invited talk*)

Maarten Marx

In this talk we show results on a survey of the quality of the Parliamentary data in a number of European countries. We measure quality using the (Linked) Open Government Data requirements set out by the W3C eGov working group. We also show opportunities for information science researchers given by Open Government Data, focusing on Parliamentary data.

Intelligent Exploitation of Local Government Resources

*Michael Rosner and Andrew Attard*

Malta is divided into sixty-eight local councils each contributing to the most basic form of local government. Several meetings take place during which the councillors gather to discuss the maintenance and embellishment of the locality, each of which are noted down in Maltese. This paper concerns a corpus of local government documents. We suggest an approach to the problem of developing an intelligent browsing system that offers improved access to the information, for example to assist local councils in decision making, or to give members of the public more transparent way to browse local council documentation.

From Recording to Retrieving: A Proposal of a Complete System for Semi-automatic Reporting for Local and National Governments

*Maria Palmerini, Ruben Cerolini, Giulio Santini and Francesco Cutugno*

The system we present here gives the possibility of bringing multimediality into the process of information retrieval from audio, video and Italian texts derived by parliament reports. The aim is not only to improve and increase the different ways the official documents can be watched and listened to and retrieved, but also to let all this information be available for different categories of users. Cedat 85 has produced a web service thought to satisfy the requests of the Basilicata Region Council and Verona Town Council, but that, given its premises, aims to be applicable in a wider range of parliamentary environments.

Automatic Multilingual Annotation of EU Legislation with Eurovoc Descriptors

*Vidas Daudaravicius*

Automatic document annotation from a controlled conceptual thesaurus is useful for establishing precise links between similar documents. This study presents a language independent document annotation system based on features derived from a collocation segmentation method. Using the multilingual conceptual thesaurus EuroVoc, we evaluate the method, comparing it against other language independent methods based on single words and bigrams. Testing the method against the manually tagged multilingual corpus Acquis Communautaire 3.0 (AC) using all descriptors found there, we attain improvements in keyword assignment precision from 50.7 to 57.6 percent over
three diverse languages (English, Lithuanian and Finnish) tested. We found high correlation between automatic assignment precision against document length and language features such as inflectiveness and compounding.

**GLOSS, an Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain**

*Francesca Frontini, Carlo Aliprandi, Clara Bacciu, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno and Tiziana Soru*

Efficient access to information is crucial in the work of organizations that require decision taking in emergency situations. This paper gives an outline of GLOSS, an integrated system for the analysis and retrieval of data in the environmental and public security domain. We shall briefly present the GLOSS infrastructure and its use, and how semantic information of various kinds is integrated, annotated and made available to the final users.

**Requirement Extraction from Transport Policy Documents**

*Oliver Mason, Aleksander Trklja and Dominik Vajn*

Requirements are an important concept in systems engineering. We present an approach to the automatic extraction of requirement statements from transport policy documents using a local grammar. The grammar has been developed using standard corpus linguistic methods. With a fairly straightforward grammar we can identify instances of requirements, as they are expressed using a small number of distinct surface patterns. One additional complication involved in this research is the issue of clearly separating requirements from policies or strategies, which are generally at a higher level. We have identified a number of different patterns that can help in distinguishing between such statements.