**Information Retrieval and Information Extraction**
**for Less Resourced Languages**
**IE-IR-LRL**

SEPLN 2009 pre-conference workshop
University of the Basque Country
Donostia-San Sebastián. Monday 7th September 2009
Organised by the SALTMIL Special Interest Group of ISCA

**DOWNLOAD THE PROCEEDINGS:** IE-IR-LRL.pdf

**PROGRAMME**

09:00 Registration
09:15 Opening
09:30 Invited Talk. Lars Borin (University of Gothenburg, Sweden)
"Linguistic diversity in the information society"
10:30 Papers (20+5) min.
1. Information retrieval and extraction in Maltese and Hebrew:
Issues in creating web-based corpora and lexical tools for

less-resourced languages.
Adam Ussishkin, Jerid Francom, Dainon Woudstra
2. TETEYEQ: A mharic question answering for factoid question.
Seid Muhie Yimam, Mulugeta Libsie

11:20 Coffee break

11:40 Papers (20+5) min.
3. Using Wikipedia for Named Entities Translation
Izaskun Fernandez, Iñaki Alegria, Nerea Ezeiza
4. Ihardetsi: A Question Answering system for Basque built on reused linguistic processors.
Iñaki Alegria, Olatz Ansa, Xabier Arregi , Arantza Otegi, Ander Soraluze

12:30 Projects (10 min. each)
1. Babelium Project. Promoting the Use and Learning of Minority Languages.
Juan A. Pereira Varela, Silvia Sanz-Santamaría, Julián Gutiérrez Serrano.
2. A web-based system for multilingual school reports
David Chan, Dewi Jones, Oggy East
3. The SALT Cymru Special Interest Group – European Funding Encouraging Collaboration Between Academia and Business in Wales within the field of Speech and Language Technology.
Gruffudd Prys
4. Automated English subtitling of Welsh TV Programmes
Llio Humphreys
5. A Dictionary Shell
Florie Moulin, Laura Laluque, Geróid Ó Néill

13:20 Panel
"Less resourced languages and Language technology.
Short- and medium-term objectives"
SALTMIL

13:45 Closing

## CALL FOR PAPERS

SALTMIL: http://ixa2.si.ehu.es/saltmil/
SEPLN 2009: http://ixa2.si.ehu.es/sepln2009
Call For Papers: http://ixa2.si.ehu.es/saltmil/en/activities/lrec2008/sepln-2009-workshop-cfp.html
Paper submission: http://sepln.org/myreview-saltmil2009
Deadline for submission: 8 June 2009

Papers are invited for the above half-day workshop, in the format outlined below. Most submitted papers will be presented in poster form, though some authors may be invited to present in lecture format.

### CONTEXT AND FOCUS

The phenomenal growth of the Internet has led to a situation where, by some estimates, more than one billion words of text is currently available. This is far more text than any given person can possibly process. Hence there is a need for automatic tools to access and process his mass of textual information. Emerging techniques of this kind include Information Retrieval (IR), Information Extraction (IE), and Question Answering (QA)

However, there is a growing concern among researchers about the situation of languages other than English. Although not all Internet text is in English, it is clear that non-English languages do not have the same degree of representation on the Internet. Simply counting the number of articles in Wikipedia, English is the only language with more than 20 percent of the available articles. There then follows a group of 17 languages with between one and ten percent of the articles. The remaining 245 languages each have less than one percent of the articles. Even these low-profile languages are relatively privileged, as the total number of languages in the world is estimated to be 6800.

Clearly there is a danger that the gap between high-profile and low-profile languages on the Internet will continue to increase, unless tools are developed for the low-profile languages to access textual
information. Hence there is a pressing need to develop basic language technology software for less-resourced languages as well. In particular, the priority is to adapt the scope of recently-developed IE, IR and QA systems so that they can be used also for these languages.

In doing so,
several questions will naturally arise, such as:

- What problems emerge when faced with languages having different linguistic features from the major languages?
- Which techniques should be promoted in order to get the maximum yield from sparse training data?
- What standards will enable researchers to share tools and techniques across several different languages?
- Which tools are easily re-useable across several unrelated languages?

It is hoped that presentations will focus on real-world examples, rather than purely theoretical discussions of the questions. Researchers are encouraged to share examples of best practice -- and also examples where tools have not worked as well as expected. Also of interest will be cases where the particular features of a less-resourced language raise a challenge to currently accepted linguistic models that were based on features of major languages.

**TOPICS**

Given the context of IR, IE and QA, topics for discussion may include, but are not limited to:

- Information retrieval;
- Text and web mining;
- Information extraction;
- Text summarization;
- Term recognition;
- Text categorization and clustering;
- Question answering;
- Re-use of existing IR, IE and QA data;
- Interoperability between tools and data.
- General speech and language resources for minority languages, with particular emphasis on resources for IR,IE and QA.

**IMPORTANT DATES**

- 8 June 2009: Deadline for submission
- 1 July 2009: Notification
- 15 July 2009: Final version
- 15 July 2009: Dead line for early registration.
- 7 September 2009: Workshop
- 8-10 September: SEPLN Conference

## ORGANISERS

- Kepa Sarasola, University of the Basque Country
- Mikel Forcada, Universitat d'Alacant, Spain
- Iñaki Alegria. University of the Basque Country
- Xabier Arregi, University of the Basque Country
- Arantza Casillas. University of the Basque Country
- Francis Tyers, Universitat d'Alacant, Spain
- Briony Williams, Language Technologies Unit, Bangor University, Wales, UK

## PROGRAMME COMMITTEE

- Iñaki Alegria. University of the Basque Country.
- Atelach Alemu Argaw: Stockholm University, Sweden
- Xabier Arregi, University of the Basque Country.
- Jordi Atserias, Barcelona Media (yahoo! research Barcelona)
- Shannon Bischoff, Universidad de Puerto Rico, Puerto Rico
- Arantza Casillas. University of the Basque Country.
- Mikel Forcada: Universitat d'Alacant, Spain
- Xavier Gomez Guinovart. University of Vigo.
- Lori Levin, Carnegie-Mellon University, USA
- Climent Nadeu, Universitat Politècnica de Catalunya
- Jon Patrick, University of Sydney, Australia
- Juan Antonio Pérez-Ortiz, Universitat d'Alacant, Spain
- Bojan Petek, University of Ljubljana, Slovenia
- Kepa Sarasola, University of the Basque Country
- Oliver Streiter, National University of Kaohsiung, Taiwan
- Vasudeva Varma, IIIT, Hyderabad, India
- Briony Williams: Bangor University, Wales, UK

## SUBMISSION INFORMATION

We expect short papers of max 3500 words (about 4-6 pages) describing research addressing one of the above topics, to be submitted as PDF documents by uploading to the following URL:

http://sepln.org/myreview-saltmil2009

The final papers should not have more than 6 pages, adhering to the stylesheet that will be adopted for the SEPLN'09 Proceedings. We recommend using the LaTeX and Word templates that can be downloaded from the conference webpage:
**Latex**
,
**Word.**