Information Retrieval and Information Extraction for less resourced languages

SEPLN 2009 pre-conference workshop
Donostia-San sebastián. Monday 7th September 2009 Call for papers

Paper submission: 8 June 2009
Organised by the SALTMIL Special Interest Group of ISCA

CALL FOR PAPERS
Information Retrieval and Information Extraction
for less resourced languages.

SEPLN 2009 pre-conference workshop
Donostia-San sebastián. Monday 7th September 2008

Organised by the SALTMIL Special Interest Group of ISCA
SALTMIL: http://ixa2.si.ehu.es/saltmil
SEPLN 2009: http://ixa2.si.ehu.es/sepln2009
Call For Papers: 23 March 2009
Paper 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 this 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
7 September 2009 Workshop

ORGANISERS

Kepa Sarasola, Department of Computer Languages and Systems, University of the Basque Country.
Mikel Forcada, Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain
Xabier Arregi, Department of Computer Languages and Systems, University of the Basque Country.
Iñaki Alegria. Department of Computer Architectures and Technologies, University of the Basque Country.
Arantza Casillas. Department of Electricity and Electronics, University of the Basque Country.
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
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 Proceedings (to be announced later on the Conference web site).