*****SECOND CALL FOR PAPERS*****

NLP4UGC 2012: "@NLP can u tag #user_generated_content?! via lrec-conf.org"

Workshop in conjunction with LREC 2012
Istanbul, May 21 - 27 2012

Workshop date: Saturday May 26 2012 (afternoon)
Submission deadline: February 15 2012

URL: http://nlp4ugc.barcelonamedia.org

Background and Motivation

The Web 2.0 has transferred the authorship of contents from institutions to the people; the web has become a channel where users exchange, explain or write about their lives and interests, give opinions and rate others' opinions. The so-called User Generated Content (UGC) in text form is a valuable resource that can be exploited for many purposes, such as cross-lingual information retrieval, opinion mining, enhanced web search, social science analysis, intelligent advertising, and so on.

In order to mine the data from the Web 2.0 we first need to understand its contents. Analysis of UG content is challenging because of its casual language, with plenty of abbreviations, slang, domain specific terms and, compared to published edited text, with a higher rate of spelling and grammar errors. Standard NLP techniques, which are used to analyze text and provide formal representations of surface data, have been typically developed to deal with standard language and may not yield the expected results on UGC. For example, shortened or misspelled words, which are very frequent in the Web 2.0 informal style, increase the variability in the forms for expressing a single concept.

This workshop aims at providing a meeting point for researchers working in the processing of UGC in textual form in one way or another, as well as developers of UGC-based applications and technologies, both from industry and academia.

Topics of Interest

We are mainly interested in, but not restricted to, the following research questions:

- What characterises UGC? Linguistic and textual phenomena that distinguish UGC from standard written text, and may pose a challenge for NLP.
- Definition of norm, concept of error, deviation and variation in UGC.
- Criteria and standards for the annotation of evaluation corpora in UGC at various levels of linguistic analysis (form, part of speech, constituents, dependencies, speech acts, deviation types, etc.).
- How quality of text affects processing tasks (tokenization, POS tagging, chunking, parsing, named-entity detection, etc.)

- Architecture and software design for flexible adaptation of NLP processing pipelines to new domains (topic domains and text-genre domains)
- Text normalisation vs adaptation of processing tools:
    - Pros and cons
    - Task dependent?
    - Costs and benefits
    - Hybrid solutions
- Approaches to normalisation (text checking, ASR, MT techniques, etc.)
- Evaluation issues related to processing and normalising UGC

Intended Audience

The workshop aims at bringing together researchers and developers from academia and industry. In particular, perspectives from the following user groups are welcome:

- UGC-based application developers, from both research and industry
- Researchers from the NLP, IR and IE communities
- Ph.D students interested or working in the processing of UGC

Submissions

Oral papers and posters should follow the main conference formatting requirements (http://www.lrec-conf.org/lrec2012/).
To submit contributions, please follow the instructions at
https://www.softconf.com/lrec2012/UGC2012/
The contributions will undergo a double review by members of the programme committee. When submitting a paper from the START page, authors will be asked to provide essential information about resources (in a broad sense, i.e. also technologies, standards, evaluation kits, etc.) that have been used for the work described in the paper or are a new result of your research. For further information on this new initiative, please refer to:
http://www.lrec-conf.org/lrec2012/?LRE-Map-2012

Important Dates

February 15: Paper submission deadline
March 15: Acceptance notifications
March 30: Camera-ready papers
May 26: Afternoon workshop at LREC

Organising Committee

Laura Alonso i Alemany, *Universidad Nacional de Córdoba (Argentina)*
Jordi Atserias, *Yahoo! Research (Spain)*
Toni Badia, *Universitat Pompeu Fabra (Spain)*
Maite Melero, *Barcelona Media Innovation Center (Spain)*
Martí Quixal, *Barcelona Media Innovation Center (Spain)*

Programme Committee

Rafael Banchs, *Institute for Infocomm Research - A*Star (Singapore)*
Steven Bedrick, *Oregon Health & Science University*
Richard Beaufort, *Cental, Catholic University of Louvain (Belgium)*
Joan Codina, *Universitat Pompeu Fabra (Spain)*

Louise-Amélie Cougnon, *Cental, Université Catholique de Louvain (Belgium)*
Jennifer Foster, *Dublin City University (Ireland)*
Michael Gamon, *Microsoft Research (USA)*
Dídac Hita, *Infojobs (Spain)*
Dan Lopresti, *CS&E, Lehigh University (USA)*
Fei Liu, *Bosch Research (USA)*
Ulrike Pado, *VICO Research&Consulting GmbH (Germany)*
Lluís Padró, *Universitat Politècnica de Catalunya (Spain)*
Alan Ritter, *CSE, University of Washington (USA)*
Roser Saurí*, Barcelona Media Innovation Center (Spain)*
Paul Schmidt, *Institut für Angewandten Informationsforschung (Germany)*
L Venkata Subramaniam, *IBM Research (India)*