EACL-2006

11th Conference of the European Chapter of the Association for Computational Linguistics

Proceedings of the workshop on

NEW TEXT Wikis and blogs and other dynamic text sources

April, 4, 2006
Trento, Italy

The conference, the workshop and the tutorials are sponsored by:

Center for the Evaluation of Language and Communication Technologies

Celct
c/o BIC, Via dei Solteri, 38
38100 Trento, Italy
http://www.celct.it

Xerox Research Centre Europe
6 Chemin de Maupertuis
38240 Meylan, France
http://www.xrce.xerox.com

CELI s.r.l.
Corso Moncalieri, 21
10131 Torino, Italy
http://www.celi.it

Thales
45 rue de Villiers
92526 Neuilly-sur-Seine Cedex, France
http://www.thalesgroup.com

EACL-2006 is supported by

Trentino S.p.a.  and Metalsistem Group 

Order copies of ACL proceedings from:
Priscilla Rasmussen,
Association for Computational Linguistics (ACL),
3 Landmark Center,
East Stroudsburg, PA 18301 USA

Phone +1-570-476-8006
Fax +1-570-476-0860
E-mail: acl@aclweb.org
On-line order form: http://www.aclweb.org/

Preface

New types of text sources, multi-lingual, with numerous cooperating or even adversarial authors and little or no editorial control are one effect of the recently dramatically lowered publication threshold. Many contain linguistic items or features classically associated with spoken language — combining the high interactivity of dialogue with the low bandwidth of written text and with the multicasting capabilities of digital communication.

New material published today most noticeably includes *blogs* — a genre that has evolved from diaries, logbooks, commentaries, columns, and editorials into a multi-faceted and networked churn of text with widely ranging viewpoints and perspectives and varying application and ambition on the part of the creator. One of the most noticeable charateristics of the blog genre is its opinionated nature and its timeliness. Blog texts are often ill-edited and hastily cobbled together in a language reminiscent of brief notes, spoken asides, or short letters, rather than of essays or newsprint. This, at any rate, is the public perception.

Another emergent genre is that of the *wiki*, a shared workspace for many more or less equal participants: wiki texts are written and edited by open teams of authors. The best known application of the wiki is the *wikipedia* — closely patterned on a classic text genre, that of the encyclopedia; other applications include e.g. project management or creative text authoring. In contrast to blogs, wikis (especially the wikipedia applications) tend to have high ambitions as regards factual correctness, persistence, editorial quality, and trustworthiness.

Bridging the two are genres such as discussion boards, web fora, and mailing lists.

Let us call these various new types of text (or indeed other modes of linguistic communication) collectively New Text. This workshop is intended to discuss the analysis and application of new text, formulate research measures that are crying out to be taken, discuss which methodological steps are obsoleted, and which babies can be saved from the bath water.

Challenge questions

NEW TEXT provides a number of research issues, immediately obvious questions, and tentative applications for our research fields:

1. New possibilities for the philologically inclined: How does new text cast new light on human communicative behaviour? This includes question on style and genre: the characteristics of new text and relations to traditional media. Do blogs in fact resemble spoken language in any important way? Do wikis hold up their promise of qualitative information dissemination? How has research in textuality, discourse and linguistic behaviour been hindered by reliance on well-edited and well-groomed data sets? Or, in more positive words: what advances can we expect, either in terms of application or in terms of understanding human behaviour, by the new data sources available to us now?

2. New challenges for building text analysis tools – how are the today's algorithms portable to new text? This includes questions on multilinguality,

code-switching, register variation, and formality melange apparent in new text.

3. New challenges for evaluation methodologies for information access systems:
   - Can new text, with dynamic information sources and streams of variable quality and impact be plugged into relevance-oriented evaluation frameworks without revising the target notion of text relevance?
   - Some new texts have high social impact; some sink without a trace; some have high import in tightly knit circles and communities. Traditional media have sales figures, citation indices, and distribution analyses. How can the impact of new texts be analyzed?
   - New texts have variable perceived intellectual status and quality – how can it be measured and predicted?
4. New opportunities for new services – e.g. linking different types of text in dynamic and interactive sessions of information refinement and elaboration.
5. How new is "new"? Didn't we use to have new text before? What is the difference between "new" and "old", really?

Welcome!

Welcome to the workshop! Please join in the discussion!

Organizers

- Jussi Karlgren, SICS, Stockholm

Reviewers

- Shlomo Argamon, Illinois Institute of Technology, Chicago, IL
- Paul Clough, University of Sheffield
- Bjrn Gambck, SICS, Stockholm
- Michael Gamon, Microsoft, Redmond
- Julio Gonzalo, UNED, Madrid
- Gilad Mishne, University of Amsterdam
- Fredrik Olsson, SICS, Stockholm
- Martin Svensson, SICS, Stockholm
- zlem Uzuner, MIT, Cambridge

Workshop Program

Tuesday, April 4

9:00 - 10:30	**Session: Usage and the character of the net**
9:00 - 9:10	*Welcome* Jussi Karlgren
9:10 - 9:40	*Text Linkage in the Wiki Medium - A Comparative Study* Alexander Mehler
9:40 - 9:50	*Errors in wikis* Ann Copestake
9:50 - 10:30	*Discussion on quality, trust and authority*
10:30 - 11:00	**Italian Coffee**
11:00 - 12:30	**Session: Data**
11:00 - 11:20	*Linguistic features of Italian blogs: literary language* Mirko Tavosanis
11:20 - 11:40	*An analysis of Wikipedia digital writing* Antonella Elia
11:40 - 12:00	*Learning to Recognize Blogs: A Preliminary Exploration* Erik Elgersma, Maarten de Rijke
12:00 - 12:30	*Discussion on style*
12:30 - 14:30	**Italian lunch**
14:30 - 16:00	**Session: Experiments**
14:30 - 14:45	*Interpreting Genre Evolution on the Web* Marina Santini
14:45 - 15:00	*Novelle, a collaborative open source writing tool software* Federico Gobbo, Michele Chinosi, Massimiliano Pepe
15:00 - 15:15	*Anomaly Detecting within Dynamic Chinese Chat Text* Yunqing Xia, Kam-Fai Wong
15:15 - 15:30	*A proposal to automatically build and maintain gazetteers for Named Entity Recognition* Antonio Toral, Rafael Muoz
15:30 - 16:00	*Finding Similar Sentences across Multiple Languages in Wikipedia* Sisay Fissaha Adafre, Maarten de Rijke
16:00 - 16:30	**Italian Coffee**
16:30 - 18:00	**Winding Up**
16:30 - 16:40	*Multilingual interactive experiments with Flickr* Paul D Clough, Julio Gonzales, Jussi Karlgren
16:40 - 17:00	*Discussion on Common task*
17:00 - 17:30	*Discussion on Resources*
17:30 - 18:00	*Planning ahead: Continuing the discussion: How, Where, In what form?*

Contents