EACL-2006

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

Proceedings of the Workshop on

Adaptive Text Extraction and Mining (ATEM 2006)

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/

INTRODUCTION

With the explosive growth of the Web and intranets, the amount of information that is available in unstructured and semistructured documents keeps increasing at an unprecedented rate. These terabytes of text contain valuable information for virtually every domain of activity, from education to business to counter-terrorism. However, existing tools for accessing and exploiting this data are just not effective enough to satisfy user expectations.

Recent years have brought significant interest and progress in developing techniques for the automatic extraction and mining of information from text. In contrast to the previous generation of extraction systems which relied on (expensive and brittle) hand-written rules, most recent approaches use machine learning techniques to uncover the structure of text. To further reduce the users' burden, researchers are also investigating a variety of active, semi-supervised, and unsupervised learning algorithms that minimize the amount of labeled documents required for training.

The increasing interest of adaptive extraction and mining from texts is also demonstrated by several recent initiatives:

- The Automatic Content Extraction program was started a few years ago by NIST (www.itl.nist.gov/iad/894.01/tests/ace).
- The PASCAL challenge on Machine Learning for Information Extraction, whose aim was to assess the current state of the art, dentifying future challenges and to foster additional research in the field (nlp.shef.ac.uk/pascal).
- Various initiatives related to the life sciences, such as BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology; www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html), and the Shared Task proposed at the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications; research.nii.ac.jp/ collier/workshops/JNLPBA04st.htm).

Adaptive extraction and mining from texts is an extremely active area of research that lies at the intersection of diverse fields such as information extraction, text mining, machine learning, data mining, link analysis and relationship discovery, information retrieval, natural language processing, information integration, distributed databases, and knowledge capture. Developments in any of these fields have an immediate effect on the others, so it is important to foster the free exchange of ideas among researchers that work on the various aspects of adaptive text extraction and mining. The purpose of this workshop is to bring together researchers and practitioners from all these communities, so that they can discuss recent results and foster new directions of research in the field. The workshop builds on the success of previous workshops on the same topic at AAAI-1999, ECAI-2000, IJCAI-2001, ECML 2003, and AAAI-2004 (see www.isi.edu/info-agents/RISE/Resources.html for details). The workshop will also serve to follow up ideas discussed at the 2005 Dagstuhl Workshop on Machine Learning for the Semantic Web (www.smi.ucd.ie/Dagstuhl-MLSW), much of which focused on adaptive information extraction.

The program includes nine papers: eight papers come from from Europe (Germany, Italy, Ireland, Spain, Sweden, The Netherlands, United Kingdom), one from Japan/USA. Each paper was reviewed by at least two reviewers.

We thank the reviewers for their cooperation in the reviewing process, especially considering the very short interval between submission and notification.

Fabio Ciravegna
Claudio Giuliano
Nicholas Kushmerick
Alberto Lavelli
Ion Muslea
February 2006

ORGANISING COMMITTEE

Fabio Ciravegna (University of Sheffield, UK)
Claudio Giuliano (ITC-irst, Italy)
Nicholas Kushmerick (University College Dublin, Ireland)
Alberto Lavelli (ITC-irst, Italy)
Ion Muslea (Language Weaver, USA)

PROGRAMME COMMITTEE

Mary Elaine Califf (Illinois State University, USA)
Fabio Ciravegna (University of Sheffield, UK)
Mark Craven (University of Wisconsin, USA)
Valter Crescenzi (Universita' Rome Tre, Italy)
Walter Daelemans (University of Antwerp, Belgium)
Dayne Freitag (Fair Isaac Corporation, USA)
Claudio Giuliano (ITC-irst, Italy)
Nicholas Kushmerick (University College Dublin, Ireland)
Alberto Lavelli (ITC-irst, Italy)
Ion Muslea (Language Weaver, USA)
Un Yong Nahm (University of Texas at Austin, USA)
Ellen Riloff (University of Utah, USA)
Roman Yangarber (University of Helsinki, Finland)

Workshop Program

9:00-10:30 Section 1

Learning Effective Surface Text Patterns for Information Extraction
Gijs Geleijnse and Jan Korst

A Hybrid Approach for the Acquisition of Information Extraction Patterns
Mihai Surdeanu, Jordi Turmo and Alicia Ageno

10:30-11:00 Coffee Break

11:00-12:30 Section 2

An Experimental Study on Boundary Classification Algorithms for Information Extraction using SVM
Jose Iria, Neil Ireson and Fabio Ciravegna

Simple Information Extraction (SIE): A Portable and Effective IE System
Claudio Giuliano, Alberto Lavelli and Lorenza Romano

Transductive Pattern Learning for Information Extraction
Brian McLernon and Nicholas Kushmerick

12:30-14:30 Lunch

14:30-16:00 Section 3

Spotting the 'Odd-one-out': Data-Driven Error Detection and Correction in Textual Databases
Caroline Sporleder, Marieke van Erp, Tijn Porcelijn and Antal van den Bosch

Recognition of synonyms by a lexical graph
Peter Siniakov

Active Annotation
Andreas Vlachos

16:00-16:30 Coffee Break

16:30-18:00 Section 4

Expanding the Recall of Relation Extraction by Bootstrapping
Junji Tomita, Stephen Soderland and Oren Etzioni

PANEL

Table of Contents