EACL-2006

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

Proceedings of the workshop on

Learning Structured Information in Natural Language Applications

April, 3, 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

Language processing largely deals with multidimensional and highly structured forms of information. Indeed, from the morphological up to the deep syntactic and semantic levels, linguistic information is often described by structured data, making the learning of the associated linguistic tasks more complex.

Traditional methods for the design of language applications involve the extraction of features that map data representations to vectors of attributes/values. Unfortunately, there is no methodology that helps in this feature modeling problem. Consequently, in order to encode structured data, the designer has to rely on his/her deep knowledge, expertise and intuition about the linguistic phenomenon associated with the target structures.

Recently, approaches that attempt to alleviate such modeling complexity by directly encoding structured data have been developed. Among other, kernel methods and conditional random fields provide interesting properties. The former use kernel functions to implicitly define richer feature spaces (e.g. substructure spaces) whereas the latter allow the designer to directly encode the probabilistic model on the structures. The promising aspects of such approaches open new research directions:

(a) the study of their impact on the modeling of diverse natural language structures,

(b) their comparative assessment with traditional attribute-value models, and

(c) the investigation of techniques which aim to improve their efficiency.

Additionally, the complementary study of mapping the classification function in structured spaces is very interesting. Classification functions can be designed to output structured data instead of simple values. In other words, the output values may be interpreted as macro-labels which describe configurations and dependencies over simpler components, e.g. parse trees or semantic structures.

The workshop was held on April 3, 2006, just preceding the 11th Conference of the European Chapter of the Association for Computational Linguistics. Its primary objective was to favor the discussing on the above topics. For this purpose, researchers from different communities such as machine learning, computational linguistics, information retrieval and data mining were invited to partecipate at the workshop to promote the discussion and development of new ideas and methods for the effective exploitation of "structured data" for natural language learning and applications.

Regarding these latter, *Coreference Resolution*, *Information/Relation Extraction*, *Machine Translation*, *Multilingual Corpus Alignment*, *Named Entity Recognition*, *Question Classification*, *Semantic Role Labeling*, *Semantic Parsing*, *Syntactic Parsing and Parse Re-Ranking*, *Text Categorization* and *Word Sense Disambiguation* were considered particularly interesting for the workshop discussion. Moreover, machine learning approaches based on *Kernel Methods*, *Maximal Margin Classifiers*, *Conditional Random Fields* and *Support Vector Machines* (SVMs) were judged those most promising to deal with structured data.

This volume contains twelve papers accepted for presentation at the workshop. We received a rather large number of high quality papers. Consequently, due to the restriction imposed by one day workshop, we decided to divide the papers in two categories: those reporting almost conclusive results and/or theories supported by a sound experimentation (full papers) and those proposing preliminary results and/or theories that would have been received significant benefits from a more extensive experimentation (position papers).

The program committee accepted eight submissions as full papers (about 50% of acceptance rate) and others four as position papers. The workshop papers deal with several interesting aspects of structured data in natural language learning. From a machine learning perspective, the contributions on: kernel methods within SVMs, probabilistic approaches and unsupervised methods, e.g. latent semantic analysis, support an interesting comparative discussion. Regarding the NLP tasks, the papers touch almost all the targeted applications: *Named Entity recognition*, *Relation Extraction*, *Discourse Parsing*, *Semantic Role Labeling*, *Prepositional Phrase Attachment problem*, *Text Categorization*, *Machine Translation* and *Question Answering*.

We believe that the workshop outcome will be helpful to increase the knowledge about advanced machine learning techniques in the modeling of structured data for Natural Language Applications.

We gratefully acknowledge all the members of the Program Committee for the excellent work done in reviewing and commenting the individual submissions within the very short time.

Roberto Basili
Alessandro Moschitti

Rome, February 15th, 2006.

SPONSOR:

Department of Computer Science, University of Rome "Tor Vergata"

ORGANIZING COMMITTEE:

Roberto Basili, University of Rome "Tor Vergata", Co-chair
Alessandro Moschitti, University of Rome "Tor Vergata", Co-chair

PROGRAM COMMITTEE:

Nicola Cancedda (Xerox Research Centre Europe, France)
Nello Cristianini (University of California, Davis , USA)
Aron Culotta (University of Massachusetts Amherst, USA)
Walter Daelemans (University of Antwerp, Netherlands)
Marcello Federico (ITC-Irst, Italy)
Attilio Giordana (University of Turin, Italy)
Marko Grobelink (J. Stefan Institute, Ljubljana, Slovenia)
Fred Jelinek (CLSP John Hopkins University, USA)
Thorsten Joachims (Cornell University, USA)
Lluis Marquez (Universitat Politecnica de Catalunya, Spain)
Giuseppe Riccardi (University of Trento, Italy)
Dan Roth (University of Illinois at Urbana-Champaign, USA)
Alex Smola (National ICT Australia, ANU)
Carlo Strapparava (ITC-Irst, Italy)
John Shawe Taylor (University of Southampton, UK)
Ben Taskar (University of California at Berkeley , USA)
Dimitry Zelenko (SRA international inc., USA)

WORKSHOP WEBSITE:

`http://ai-nlp.info.uniroma2.it/eacl2006-ws10/`

WORKSHOP PROGRAM

Monday, April 3

9:00-9:15	WELCOME AND INTRODUCTORY NOTES
	FULL PAPER SESSION
9:15-9:40	*Maximum Entropy Tagging with Binary and Real-Valued Features* Vanessa Sandrini, Marcello Federico and Mauro Cettolo
9:40-10:05	*Constraint Satisfaction Inference:* *Non-probabilistic Global Inference for Sequence Labelling* Sander Canisius, Antal van den Bosch and Walter Daelemans
10:05-10:30	*Decomposition Kernels for Natural Language Processing* Fabrizio Costa, Sauro Menchetti, Alessio Ceroni, Andrea Passerini and Paolo Frasconi
10:30-11:00	COFFEE BREAK
11:05-11:50	*Invited Speaker*
	POSITION PAPER SESSION
11:50-12:10	*A Multiclassifier based Document Categorization System:* *profiting from the Singular Value Decomposition Dimensionality Reduction Technique* Ana Zelaia, Iñaki Alegria, Olatz Arregi and Basilio Sierra
12:10-12:30	*Discourse Parsing: Learning FOL Rules based on Rich Verb Semantic Representations to automatically label Rhetorical Relations* Rajen Subba, Barbara Di Eugenio and Su Nam Kim
12:30-14:00	LAUNCH BREAK
	FULL PAPER SESSION
14:00-14:25	*Reranking Translation Hypotheses Using Structural Properties* Saša Hasan, Oliver Bender, and Hermann Ney
14:25-14:50	*Tree Kernel Engineering in Semantic Role Labeling Systems* Alessandro Moschitti, Daniele Pighin and Roberto Basili
14:50-15:15	*Syntagmatic Kernels: a Word Sense Disambiguation Case Study* Claudio Giuliano, Alfio Gliozzo and Carlo Strapparava

	POSITION PAPER SESSION
15:15-15:35	*Learning to Identify Definitions using Syntactic Features* Ismail Fahmi and Gosse Bouma
15:35-15:55	*An Ontology-Based Approach to Disambiguation of Semantic Relations* Tine Lassen and Thomas Vestskov Terney
15:55-16:30	COFFEE BREAK
	FULL PAPER SESSION
16:30-16:55	*Towards Free-text Semantic Parsing:* *A Unified Framework Based on FrameNet, VerbNet and PropBank* Ana-Maria Giuglea and Alessandro Moschitti
16:55-17:20	*Constructing a Rule Based Naming System for Thai Names Using the Concept of Ontologies* Chakkrit Snae
17:20-18:00	CONCLUSIVE REMARKS AND DISCUSSION

Table of Contents