2006

COLING • ACL

COLING • ACL 2006

Information Extraction Beyond The Document

Proceedings of the Workshop

Chairs:
Mary Elaine Califf, Mark A. Greenwood,
Mark Stevenson and Roman Yangarber

22 July 2006
Sydney, Australia

Production and Manufacturing by
BPA Digital
11 Evans St
Burwood VIC 3125
AUSTRALIA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 1-932432-74-4

Table of Contents

Preface

Traditional approaches to the development and evaluation of Information Extraction (IE) systems have relied on relatively small collections of up to a few hundred documents tagged with detailed semantic annotations. While this paradigm has enabled rapid advances in IE technology, it remains constrained by a dependence on annotated documents and does not make use of the information available in large corpora. Alternative approaches, which make use of large text collections and inter-document information, are now beginning to emerge – as evidenced by a parallel emergence of interest in learning from unlabelled data in AI in general. For example, some systems learn extraction patterns by exploiting information about their distribution across corpora; others exploit the redundancy of the Internet by assuming that facts with multiple mentions are more reliable. These approaches require large amounts of unannotated text, which is generally easy to obtain, and employ unsupervised or minimally supervised learning algorithms, as well as related techniques such as co-training and active learning. These alternative approaches are complementary to the established IE paradigm based on supervised training, and are now forming a cohesive emergent trend in recent research. They constitute the focus of this workshop.

There are several advantages to employing large text collections for IE. They provide enormous amounts of training data, albeit mostly unannotated. Facts can be extracted from, or verified across, multiple documents. Large text collections often contain vast amounts of redundancy in the form of multiple references to or mentions of closely related facts. Redundancy can be exploited in the IE setting to identify trends and patterns within the text, e.g., by means of Data Mining techniques.

For this workshop, we solicited papers presenting new, original work on learning extraction rules or identifying facts across document boundaries while exploiting sizable amounts of unlabelled text in the training stage, in the extraction stage, or both.

Eight papers were selected for inclusion in the workshop following a peer reviewing process. These papers cover a wide range of topics in Information Extraction including traditional IE tasks such as name tagging and relation extraction as well as other topics which are relevant to IE such as terminology extraction, trend identification and lexical chains. The papers describe a number of techniques including using the web as a data source and semi-supervised machine learning. We hope these will form the basis of a productive workshop, and will stimulate further research into this area, which we believe is worth pursuing.

Mary Elaine Califf
Mark A. Greenwood
Mark Stevenson
Roman Yangarber

Organizers

Chairs:

Mary Elaine Califf, Illinois State University
Mark A. Greenwood, University of Sheffield
Mark Stevenson, University of Sheffield
Roman Yangarber, University of Helsinki

Program Committee:

Markus Ackermann, University of Leipzig
Amit Bagga, AskJeeves
Roberto Basili, University of Rome, Tor Vergata
Antal van den Bosch, Tilburg University
Neus Catala, Universitat Politècnica de Catalunya
Walter Daelemans, University of Antwerp
Jenny Rose Finkel, Stanford University
Robert Gaizauskas, University of Sheffield
Ralph Grishman, New York University
Takaaki Hasegawa, NTT
Heng Ji, New York University
Nick Kushmerick, University College Dublin
Alberto Lavelli, ITC-IRST
Gideon Mann, John Hopkin's University
Ion Muslea, Language Weaver Inc.
Chikashi Nobata, Sharp
Ellen Riloff, University of Utah
Stephen Soderland, University of Washington
Yorick Wilks, University of Sheffield

Workshop Program

Saturday, 22 July 2006

9:10–9:20 Welcome

9:20–9:55 *Development of an Automatic Trend Exploration System using the MuST Data Collection*
Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru, Sachiyo Tsukawaki and Hitoshi Isahara

9:55–10:30 *Comparing Information Extraction Pattern Models*
Mark Stevenson and Mark A. Greenwood

10:30–11:00 Coffee Break

11:00–11:35 *Automatic Extraction of Definitions from German Court Decisions*
Stephan Walter and Manfred Pinkal

11:35–12:10 *Improving Semi-supervised Acquisition of Relation Extraction Patterns*
Mark A. Greenwood and Mark Stevenson

12:10–13:45 Lunch

13:45–14:20 *Automatic Knowledge Representation using a Graph-based Algorithm for Language-Independent Lexical Chaining*
Gaël Dias, Cláudia Santos and Guillaume Cleuziou

14:20–14:55 *Data Selection in Semi-supervised Learning for Name Tagging*
Heng Ji and Ralph Grishman

14:55–15:30 *LoLo: A System based on Terminology for Multilingual Extraction*
Yousif Almas and Khurshid Ahmad

15:30–16:00 Coffee Break

16:00–16:35 *Learning Domain-Specific Information Extraction Patterns from the Web*
Siddharth Patwardhan and Ellen Riloff

16:35–17:00 Discussion