Proceedings of the

ACL-2000 Workshop

on

Recent Advances in Natural Language Processing and Information Retrieval

Held in conjunction with
The 38th Annual Meeting of the
Association for Computational Linguistics

Edited by

Judith Klavans
(Columbia University, New York, USA)
and

Julio Gonzalo
(Universidad Nacional de Educación a Distancia, Madrid, Spain)

8 October 2000
Hong Kong University of Science and Technology (HKUST)
Hong Kong

Proceedings of the

ACL-2000 Workshop

on

Recent Advances in Natural Language Processing and Information Retrieval

Held in conjunction with
The 38th Annual Meeting of the
Association for Computational Linguistics

Edited by

Judith Klavans
(Columbia University, New York, USA)
and

Julio Gonzalo
(Universidad Nacional de Educación a Distancia, Madrid, Spain)

8 October 2000
Hong Kong University of Science and Technology (HKUST)
Hong Kong

Order copies of this and other ACL workshop proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Preface

This volume contains the papers presented at the workshop on Recent Advances in Natural Language Processing and Information Retrieval, held on 8 October 2000 in conjunction with the 38^{th} Annual Meeting of the Association for Computational Linguistics (ACL).

The aim of the workshop was to foster the interaction between researchers in the areas of Natural Language Processing (NLP) and Information Retrieval (IR), and furthermore, to promote discussion on the current and potential benefits of common approaches to related research challenges. In putting the workshop together, our goal was to bring these two communities together to establish opportunity for communication. We hope there will be similar workshops at IR related conferences so that a gradual closing of the existing gaps will occur. This workshop is not the first on the topic, but one which reflects what we believe are increasing trends as each field reaches its limits.

The growing research and application possibilities provided by the increased amount of networked information have motivated new attempts to explore the relationship between NLP and IR. For researchers in IR, a compelling challenge is to move from (monolingual) *document* retrieval within controlled text collections, to actually retrieving *information*, rather than individual documents, from multilingual, heterogeneous and dynamic webs of interlinked documents and online services. The reciprocal challenge for NLP research is to scale up, adapt and possibly reshape techniques and resources to help bridge the gap between document and information retrieval in practical applications.

The central topic of the workshop was the application of language technologies to information retrieval, including

- the role of lexical-syntactic information in mono- and multilingual IR, including morphology, phrase detection and treatment, word sense disambiguation adapted to IR needs, acquisition and use of lexical resources, etc.
- empirical evidence regarding the use of NL techniques in different retrieval scenarios, typification of such scenarios, and the discussion of evaluation measures beyond precision/recall variants.
- interaction between NLP and IR techniques in topics related to both areas such as Cross-Language and Interactive Text Retrieval, Question Answering, Information Extraction, Text Summarization, Text Data Mining, etc.

The original call for papers resulted in 25 submissions, from which 10 papers were selected on the basis of a thorough reviewing process. A majority of the papers contained in this volume study how to improve retrieval processes using syntactic and semantic information, including lexical expansions for web querying, semantic indexing, phrasal indexing in different languages, acquisition and use of lexical databases, etc. The remaining papers combine NL techniques from related areas - summarization, information extraction - to extend search capabilities and presentation of results, including summarization of search engine hit lists and summarization for text categorization, and a search interface that accepts template-like general constraints and is able to return specific information items such as locations, people or companies that satisfy user's constraints. Although none of these papers

deal directly with cross-language issues (well covered in the *Cross-Language Evaluation Forum* held at Lisbon two weeks earlier), the monolingual systems presented here cover five different languages (English, Japanese, Korean, Italian and Czech).

We would like to first of all thank the authors, whose research contributions have made this workshop possible. Thanks are due also to the following: Nicoletta Calzolari (the ACL workshop chair), David Yarowsky, Priscilla Rasmussen and the local organizing committee at HKUST. Finally, we would like to express our deep gratitude to the members of the program committee, who did a careful reviewing job under tight time constraints.

Judith Klavans and Julio Gonzalo (Program Chairs)

PROGRAM COMMITTEE

WORKSHOP PROGRAM

8 October 2000

8:45-9:00 Welcome

Using NL semantics for IR

9:00-9:35 *Adapting a synonym database to specific domains*
Davide Turcato, Fred Popowich, Janine Toole, Dan Fass, D. Nicholson and G. Tisher (Gavagai Technology Inc., Vancouver)

9:35-10:10 *Exploiting Lexical Expansions and Boolean Compositions for Web Querying*
Bernardo Magnini and Roberto Prevete (ITC-IRST, Trento)

10:10-10:40 Coffee break

10:40-11:15 *Use of Dependency Tree Structures for the Microcontext Extraction*
Martin Holub and Alena Böhmová (MFF UK, Praha)

11:15-11:50 *Semantic Indexing using WordNet Senses*
Rada Mihalcea and Dan Moldovan (Southern Methodist Univ., Dallas)

11:50-12:00 Short break

12:00-13:00 **Invited session**

13:00-14:30 Lunch

Effects of NL Phrasal indexing on IR

14:30-15:05 *Discriminative Power and Retrieval Effectiveness of Phrasal Indexing Terms*
Sumio Fujita (Justsystem corp., Tokushima)

15:05-15:40 *Corpus-Based Learning of Compound Noun Indexing*
Byung-Kwan Kwak, Jee-Hyub Kim, Geunbae Lee and Jung Yun Seo (Pohang Univ./Sogang Univ.)

15:40-16:15 *REXTOR: A System for Generating Relations from Natural Language*
Boris Katz and Jimmy Lin (MIT, Cambridge)

16:15-16:45 Coffee break

Applications involving NL and IR

16:45-17:20 *A Text Categorization Based on a Summarization Technique*
Sue J. Ker and Jen-Nan Chen (Soochow Univ./Ming Chuan Univ., Taipei)

17:20-17:55 *From Information Retrieval to Information Extraction*
David Milward and James Thomas (SRI International, Cambridge)

17:55-18:30 *Automatic summarization of search engine hit lists*
Dragomir R. Radev and Weiguo Fan (Univ. of Michigan, Ann Arbor)

Table of Contents