Coling 2010

23rd International Conference on Computational Linguistics

Proceedings of the

Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)

28 August 2010
Beijing International Convention Center
Beijing, China

Produced by
Chinese Information Processing Society of China
All rights reserved for Coling 2010 CD production.

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Introduction

A long-standing problem in Natural Language Processing has been a lack of large-scale knowledge for computers. The emergence of the Web and the rapid increase of information on the Web brought us to what could be called the ”information explosion era,” and drastically changed the environment of NLP. The Web is not only a marvelous target for NLP, but also a valuable resource from which knowledge could be extracted for computers. Motivated by the desire to have a very first opportunity to discuss early approaches to those issues and to share the state-of-the-art technologies at that time, the first International Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2008) was successfully held in conjunction with WWW 2008 in Beijing.

Since the discussion of the first workshop, research and development activities on large-scale text processing and large-scale knowledge acquisition become much more popular these days. The large-scale NLP naturally requires large-scale infrastructures, such as neatly-prepared huge corpora, robust morpho-syntactic tools, and high-performance computing environments. However, such infrastructures can not be prepared by individual researchers nor research groups alone in general, although of course we know some exceptions. Based on this motivation, towards much larger-scale NLP, activities aiming at constructing and sharing the infrastructures have continued. Although we have found many publications presented in recent conferences/workshops including the above mentioned workshop, we still do not have opportunities to compare latest approaches, share analysis on advantages/disadvantages, and discuss possible directions towards further improvement and innovation.

Furthermore, beyond the success of large-scale NLP and knowledge acquisition, we are starting to face a new problem: how to manage and use the automatically acquired knowledge (AAK in short). We are still not confident that those large-scale AAK can actually solve real world problems. How to incorporate the AAK into existing NLP frameworks and how to manage them are yet unsolved issues. One approach could be some bootstrapping of extracting knowledge and enhancing NLP based on the knowledge. The representation and standardization of AAK are also emerging important issues. One of the most highly demanded applications for AAK-based NLP is a semantic search to cope with the information explosion on the Web. Though our daily life heavily depends on the Web information, our diversified needs have not been sufficiently satisfied by the existing search engines. AAK-based NLP can be a key technology to realize a new-generation semantic search, which incorporates enhanced information access, analysis and organization.

The aim of the second workshop of the series of International Workshop on NLP Challenges in the Information Explosion Era (NLPIX) is to bring researchers and practitioners together in order to discuss large-scale and sharable NLP infrastructures, and furthermore to discuss emerging NEW issues beyond them. The program committee accepted 9 papers that cover wide variety of topics such as lexical acquisition, lexical semantics, coreference, and information access, many of which are based on very large scale Web text data.

The invited talks were given by Hang Li (Microsoft Research Asia) and Hoifung Poon (University of Washington).

We are grateful to Info-plosion (New IT Infrastructure for the Information-explosion Era, Grant-in-Aid for Scientific Research on Priority Areas, MEXT (Ministry of Education, Culture, Sports, Science and Technology, Japan)) for partially sponsoring the workshop. We would like to thank all the authors who submitted papers. We express our deepest gratitude to the committee members for their timely reviews. We also thank the COLING 2010 organizers for their help with administrative matters.

Sadao Kurohashi and Takehito Utsuro

Co-Organizers

Organizers:

Sadao Kurohashi, Kyoto University (Japan)
Takehito Utsuro, University of Tsukuba (Japan)

Program Committee:

Pushpak Bhattacharyya, IIT (India)
Thorsten Brants, Google (USA)
Eric Villemonte de la Clergerie, INRIA (France)
Atsushi Fujii, Tokyo Institute of Technology (Japan)
Julio Gonzalo, UNED (Spain)
Kentaro Inui, Tohoku University (Japan)
Noriko Kando, NII (Japan)
Daisuke Kawahara, NICT (Japan)
Jun'ichi Kazama, NICT (Japan)
Adam Kilgarriff, Lexical Computing Ltd. (UK)
Gary Geunbae Lee, POSTECH (Korea)
Hang Li, Microsoft (China)
Dekang Lin, Google (USA)
Tatsunori Mori, Yokohama National University (Japan)
Satoshi Sekine, New York University (USA)
Kenjiro Taura, University of Tokyo (Japan)
Kentaro Torisawa, NICT (Japan)
Marco Turchi, European Commission - Joint Research Centre (Italy)
Yunqing Xia, Tsinghua University (China)

Additional Reviewer:

Wei Wu, Microsoft (China)

Invited Speakers:

Hang Li, Microsoft (China)
Hoifung Poon, University of Washington (USA)

Table of Contents

Workshop Program

Saturday, August 28, 2010

9:30 *Opening*

9:40–10:30 **Invited Talk I**

Query Understanding in Web Search - by Large Scale Log Data Mining and Statistical Learning
Hang Li

10:30–11:00 *Tea Break*

11:00–12:15 **Session I: Information Access**

Exploiting Term Importance Categories and Dependency Relations for Natural Language Search
Keiji Shinzato and Sadao Kurohashi

Summarizing Search Results using PLSI
Jun Harashima and Sadao Kurohashi

Automatic Classification of Semantic Relations between Facts and Opinions
Koji Murakami, Eric Nichols, Junta Mizuno, Yotaro Watanabe, Hayato Goto, Megumi Ohki, Suguru Matsuyoshi, Kentaro Inui and Yuji Matsumoto

12:15–13:30 *Lunch*

13:30–14:20 **Invited Talk II**

Statistical Relational Learning for Knowledge Extraction from the Web
Hoifung Poon

14:20–15:35 **Session II: Lexical Acquisition**

Even Unassociated Features Can Improve Lexical Distributional Similarity
Kazuhide Yamamoto and Takeshi Asakura

A Look inside the Distributionally Similar Terms
Kow Kuroda, Jun'ichi Kazama and Kentaro Torisawa

Utilizing Citations of Foreign Words in Corpus-Based Dictionary Generation
Reinhard Rapp and Michael Zock

Saturday, August 28, 2010 (continued)

15:35–16:00 *Tea Break*

16:00–17:15 **Session III: Coreference and Semantics**

Large Corpus-based Semantic Feature Extraction for Pronoun Coreference
Shasha Liao and Ralph Grishman

Mining Coreference Relations between Formulas and Text using Wikipedia
Minh Nghiem Quoc, Keisuke Yokoi, Yuichiroh Matsubayashi and Akiko Aizawa

Adverse-Effect Relations Extraction from Massive Clinical Records
Yasuhide Miura, Eiji Aramaki, Tomoko Ohkuma, Masatsugu Tonoike, Daigo Sugihara, Hiroshi Masuichi and Kazuhiko Ohe