NAACL HLT 2010

Sixth Web as Corpus Workshop (WAC-6)

Proceedings of the Workshop

June 5, 2010
Los Angeles, California

USB memory sticks produced by
Omnipress Inc.
2600 Anderson Street
Madison, WI 53707
USA

©2010 The Association for Computational Linguistics

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

Introduction

More and more people are using Web data for linguistic and NLP research. The workshop, the sixth in an annual series, provides a venue for exploring how we can use it effectively and what we will find if we do, with particular attention to

- Web corpus collection projects, or modules for one part of the process (crawling, filtering, de-duplication, language-id, tokenising, indexing, . . .)
- characteristics of Web data from a linguistics/NLP perspective including registers, domains, frequency distributions, comparisons between datasets
- using crawled Web data for NLP purposes (with emphasis on the data rather than the use)

Previous WAC workshops have been in Europe and Africa. The west coast of the US is the global centre for web development, hosting Google, Microsoft, Yahoo and a thousand others, so we are glad to be here!

Organizers:

Adam Kilgarriff, Lexical Computing Ltd. (Workshop Chair)
Dekang Lin, Google Inc.
Serge Sharoff, University of Leeds (SIGWAC Chair)

Program Committee:

Adam Kilgarriff, Lexical Computing Ltd. (UK)
Dekang Lin, Google Inc. (USA)
Serge Sharoff, University of Leeds (UK)
Silvia Bernardini, University of Bologna (Italy)
Stefan Evert, University of Osnabrück (Germany)
Cédrick Fairon, UCLouvain (Belgium)
William H. Fletcher, U.S. Naval Academy (USA)
Gregory Grefenstette, Exalead, (France)
Igor Leturia, Elhuyar Fundazioa (Spain)
Jan Pomikálek, Masaryk University (Czech Republic)
Preslav Nakov, National University of Singapore
Kevin Scannell, Saint Louis University (USA)
Gilles-Maurice de Schryver, Ghent University (Belgium)

Invited Speaker:

Patrick Pantel, ISI, University of Southern California

Proceedings:

Jan Pomikálek

Table of Contents

Workshop Program

Saturday, June 5, 2010

	Session 1:
8:30	Start, introduction
8:40	*NoWaC: a large web-based corpus for Norwegian* Emiliano Raul Guevara
9:05	*Building a Korean Web Corpus for Analyzing Learner Language* Markus Dickinson, Ross Israel and Sun-Hee Lee
9:30	Invited talk by Patrick Pantel
10:30	Coffee break
	Session 2:
11:00	*Sketching Techniques for Large Scale NLP* Amit Goyal, Jagadeesh Jagaralamudi, Hal Daumé III and Suresh Venkatasubramanian
11:25	*Building Webcorpora of Academic Prose with BootCaT* George Dillon
11:50	*Google Web 1T 5-Grams Made Easy (but not for the computer)* Stefan Evert
12:15	Closing session