EACL-2006

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

Proceedings of the 2nd International Workshop on

Web as Corpus

Chairs:
Adam Kilgarriff
Marco Baroni

April 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/

WAC2: Programme

9.00-9.30	Marco Baroni and Adam Kilgarriff *Introduction*
9.30-10.00	András Kornai, Péter Halácsy, Viktor Nagy, Csaba Oravecz, Viktor Trón and Dániel Varga *Web-based frequency dictionaries for medium density languages*
10.00-10.30	Mike Cafarella and Oren Etzioni *BE: a search engine for NLP research*
	Break
11.00-11.30	Masatsugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro and Satoshi Sato *A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the web*
11.30-12.00	Gemma Boleda, Stefan Bott, Rodrigo Meza, Carlos Castillo, Toni Badia and Vicente López *CUCWeb: a Catalan corpus built from the web*
12.00-12.30	Paul Rayson, James Walkerdine, William H. Fletcher and Adam Kilgarriff *Annotated web as corpus*
	Lunch
2.30-3.00	Arno Scharl and Albert Weichselbraun *Web coverage of the 2004 US presidential election*
3.00-3.30	Cédrick Fairon *Corporator: A tool for creating RSS-based specialized corpora*
3.30-4.00	**Demos, part 1**
	Break
4.30-4.50	**Demos, part 2**
4.50-5.20	Davide Fossati, Gabriele Ghidoni, Barbara Di Eugenio, Isabel Cruz, Huiyong Xiao and Rajen Subba *The problem of ontology alignment on the web: a first report*
5.20-5.50	Kie Zuraw *Using the web as a phonological corpus: a case study from Tagalog*
5.50-6.00	*Organization, next meeting, closing*
Reserve paper	Rüdiger Gleim, Alexander Mehler and Matthias Dehmer *Web corpus mining by instance of Wikipedia*

Programme Committee

Toni Badia
Marco Baroni (co-chair)
Silvia Bernardini
Massimiliano Ciaramita
Barbara Di Eugenio
Roger Evans
Stefan Evert
William Fletcher
Rüdiger Gleim
Gregory Grefenstette
Péter Halácsy
Frank Keller
Adam Kilgarriff (co-chair)
Rob Koeling
Mirella Lapata
Anke Lüdeling
Alexander Mehler
Drago Radev
Philip Resnik
German Rigau
Serge Sharoff
David Weir

Preface

What is the role of a workshop series on web as corpus?

We argue, first, that attention to the web is critical to the health of non-corporate NLP, since the academic community runs the risk of being sidelined by corporate NLP if it does not address the issues involved in using very-large-scale web resources; second, that text type comes to the fore when we study the web, and the workshops provide a venue for nurturing this under-explored dimension of language; and thirdly that the WWW community is an important academic neighbour for CL, and the workshops will contribute to contact between CL and WWW.

High-performance NLP needs web-scale resources

The most talked-about presentation of the ACL 2005 was Franz-Josef Och's, in which he presented statistical MT results based on a 200 billion word English corpus. His results led the field. He was in a privileged position to have access to a corpus of that size. He works at Google.

With enormous data, you get better results. (See e.g. Banko and Brill 2001.) It seems to us there are two possible responses for the academic NLP community. The first is to accept defeat: "we will never have resources on the scale Google has, so we should accept that our systems will not really compete, that they will be proofs-of-concept or deal with niche problems, but will be out of the mainstream of high-performance HLT system development." The second is to say: we too need to make resources on this scale available, and they should be available to researchers in universities as well as behind corporate firewalls: and we can do it, because resources of the right scale are available, for free, on the web. We shall of course have to acquire new expertise along the way – at, *inter alia*, WAC workshops.

Text type

The most interesting question that the use of web corpora raises is text type. (We use 'text type' as a cover-all term to include domain, genre, style etc.) The first question about web corpora from an outsider is usually "how do you know that your web corpus is representative?" to which the fitting response is "how do you know whether any corpus is representative (of what?)". These questions will only receive satisfactory answers when we have a fuller account of how to identify and distinguish different kinds of text.

While text type is not centre-stage in this volume, we suspect it will be prominent in discussions at the workshop and will be the focus of papers in future workshops.

The WWW community: links, web-as-graph, and linguistics

One of CL's academic neighbours is the WWW community (as represented by, eg, the WWW conference series). Many of their key questions concern the nature of the web, viewing it as a large set of domains, or as a graph, or as a bag of bags of words. The web is substantially a linguistic object, and there is potential for these views of the web contributing to our linguistic understanding. For example, the graph structure of the web has been used to identify highly connected areas which are "web communities". How does that graph-theoretical connectedness relate to the linguistic properties one would associate with a discourse community? To date the links between the communities have been not been strong. (Few WWW papers are referenced in CL papers, and vice versa.) The workshops will provide a venue where WWW and CL interests intersect.

Recent work by co-chairs and colleagues

At risk of abusing chairs' privilege, we briefly mention two pieces of our own work. In the first we have created web corpora of over 1 billion words for German and Italian. The text has been de-duplicated, passed through a range of filters, part-of-speech tagged, lemmatized, and loaded into a web-accessible corpus query tool supporting a wide range of linguists' queries. It offers one model of how to use the web as a corpus. The corpora will be demonstrated in the main EACL conference (Baroni and Kilgarriff 2006).

In the second, WebBootCaT (work with Jan Pomikalek and Pavel Rychlý of Masaryk University, Brno), we have prepared a version of the BootCaT tools (Baroni and Bernardini 2004) as a web service. Users fill in a web form with the target language and some "seed terms" to specify the domain of the target corpus, and press the "Build Corpus" button. A corpus is built. Thus, people without any programming or software-installation skills can create corpora to their own specification. The system will be demonstrated in the "demos" session of the workshop.

The workshop series to date

This is the second international workshop, the first being held in July 2005 in Birmingham, UK (in association with Corpus Linguistics 2005). There was an earlier Italian event in Forlì, in January 2005. All three have attracted high levels of interest. The papers in this volume were selected following a highly competitive review process, and we would like to thank all those who submitted, all those on the programme committee who contributed to the review process, and the additional reviewers who helped us to get through the large number of submissions. Special thanks to Stefan Evert for help with assembling the proceedings. (Cafarella and Etzioni have an abstract rather than a full paper to avoid duplicate publication: we felt their presentation would make an important contribution to the workshop, which was a distinct issue to them not having a new text available.)

We are confident that there will be much of interest for anyone engaged with NLP and the web.

References

Banko, M. and E. Brill. 2001. "Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing." In Proc. Human Language Technology Conference (HLT 2001)

Baroni, M and S. Bernardini 2004. BootCaT: Bootstrapping corpora and terms from the web. Proc. LREC 2004, Lisbon: ELDA. 1313-1316.

Baroni, M. and A. Kilgarriff 2006. "Large linguistically-processed web corpora for multiple languages." Proc EACL, Trento, Italy.

Màrquez, L. and D. Klein 2006. Announcement and Call for Papers for the Tenth Conference on Computational Natural Language Learning. http://www.cnts.ua.ac.be/conll/cfp.html

Och, F-J. 2005. "Statistical Machine Translation: The Fabulous Present and Future" Invited talk at ACL Workshop on Building and Using Parallel Texts, Ann Arbor.

Adam Kilgarriff and Marco Baroni, February 2006

Table of Contents

Web-based frequency dictionaries for medium density languages
András Kornai, Péter Halácsy, Viktor Nagy, Csaba Oravecz, Viktor Trón and Dániel Varga 1

BE: A search engine for NLP research
Mike Cafarella and Oren Etzioni 9

A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the Web
Masatsugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro and S. Sato . . . 11

CUCWeb: A Catalan corpus built from the Web
Gemma Boleda, Stefan Bott, Rodrigo Meza, Carlos Castillo, Toni Badia and Vicente López 19

Annotated Web as corpus
Paul Rayson, James Walkerdine, William H. Fletcher and Adam Kilgarriff 27

Web coverage of the 2004 US Presidential election
Arno Scharl and Albert Weichselbraun 35

Corporator: A tool for creating RSS-based specialized corpora
Cédrick Fairon 43

The problem of ontology alignment on the Web: A first report
Davide Fossati, Gabriele Ghidoni, Barbara Di Eugenio, Isabel Cruz, Huiyong Xiao and Rajen Subba . . . 51

Using the Web as a phonological corpus: A case study from Tagalog
Kie Zuraw 59

Web corpus mining by instance of Wikipedia
Rüdiger Gleim, Alexander Mehler and Matthias Dehmer 67