Proceedings of the

4^{th} Web as Corpus Workshop (WAC-4)

Can we beat Google?

Edited by Stefan Evert, Adam Kilgarriff and Serge Sharoff

Marrakech, Morocco
1 June 2008

Workshop Programme

9.15 – 9.30	Welcome & Introduction
Session 1: Can we do better than Google?	
9.30 – 10.00	**Reranking Google with GReG**
	Rodolfo Delmonte, Marco Aldo Piccolino Boniforti
10.00 – 10.30	**Google for the Linguist on a Budget**
	András Kornai, Péter Halácsy
10.30 – 11.00	Coffee break
Session 2: Cleaning up the Web	
11.00 – 11.30	**Victor: the Web-Page Cleaning Tool**
	Miroslav Spousta, Michal Marek, Pavel Pecina
11.30 – 12.00	**Segmenting HTML pages using visual and semantic information**
	Georgios Petasis, Pavlina Fragkou, Aris Theodorakos, Vangelis Karkaletsis, Constantine D. Spyropoulos
12.00 – 12.45	Star Talk: **Identification of Duplicate News Stories in Web Pages**
	John Gibson, Ben Wellner, Susan Lubar
12.45 – 13.30	Group discussion on **The Next CLEANEVAL**
13.30 – 15.00	Lunch break
Session 3: Compilation of Web corpora	
15.00 – 15.30	**GlossaNet 2: a linguistic search engine for RSS-based corpora**
	Cédrick Fairon, Kévin Macé, Hubert Naets
15.30 – 16.00	**Collecting Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision**
	Igor Leturia Azkarate, Iñaki San Vicente, Xabier Saralegi, Maddalen Lopez de Lacalle
16.00 – 16.30	Coffee break
Session 3 (cont'd)	
16.30 – 17.15	Star Talk: **Introducing and evaluating ukWaC, a very large Web-derived corpus of English**
	Adriano Ferraresi, Eros Zanchetta, Silvia Bernardini, Marco Baroni
Session 4: Technical applications of Web data	
17.15 – 17.45	**RoDEO: Reasoning over Dependencies Extracted Online**
	Reda Siblini, Leila Kosseim
17.45 – 18.15	General discussion
18.15	Wrap-up & Conclusion

Workshop Organisers

Stefan Evert, *University of Osnabrück*

Adam Kilgarriff, *Lexical Computing*

Serge Sharoff, *University of Leeds*

Programme Committee

Silvia Bernardini, *U of Bologna, Italy*

Massimiliano Ciaramita, *Yahoo! Research Barcelona, Spain*

Jesse de Does, *INL, Netherlands*

Katrien Depuydt, *INL, Netherlands*

Stefan Evert, *U of Osnabrück, Germany*

Cédrick Fairon, *UCLouvain, Belgium*

William Fletcher, *U.S. Naval Academy, USA*

Gregory Grefenstette, *Commissariat à l'Énergie Atomique, France*

Péter Halácsy, *Budapest U of Technology and Economics, Hungary*

Katja Hofmann, *U of Amsterdam, Netherlands*

Adam Kilgarriff, *Lexical Computing Ltd, UK*

Igor Leturia, *Elhuyar Fundazioa, Basque Country, Spain*

Phil Resnik, *U of Maryland, College Park, USA*

Kevin Scannell, *Saint Louis U, USA*

Gilles-Maurice de Schryver, *U Gent, Belgium*

Klaus Schulz, *LMU München, Germany*

Serge Sharoff, *U of Leeds, UK*

Eros Zanchetta, *U of Bologna, Italy*

Contents

Preface

> We want the Demon, you see, to extract from the dance of atoms only information that is genuine, like mathematical theorems, fashion magazines, blueprints, historical chronicles, or a recipe for ion crumpets, or how to clean and iron a suit of asbestos, and poetry too, and scientific advice, and almanacs, and calendars, and secret documents, and everything that ever appeared in any newspaper in the Universe, and telephone books of the future ...
>
> Stanisław Lem (1985). *The Cyberiad*, translated by Michael Kandel.

Can we beat Google? It is a big question.

First, it is as well to remember that Google is the *non plus ultra* of Internet startups. It is amazing. It is an outrageous fantasy come true, in terms of both speed and accuracy and the fabulous wealth accruing to its founders. If the Internet has fairy tales, this is it. We don't even think of it as an Internet startup any more: it transcended that long ago,[1] as it entered the lexicons of the world,[2] changed the way we live our lives, and diverted a substantial share of the world's advertising spend through its coffers.

Their success is no accident. What they do, they do very well. It would be a bad idea to compete head-on. The core of their business is to index as much of the Web as possible, and make it available very very quickly to people who want to find out about things or – better, from Google's point of view – buy things. And, of course, to carry advertisements and thereby to make oodles of money. In order to do that, they address a large number of associated tasks, including finding text-rich Web pages, finding the interesting text in a Web page, partitioning and identifying duplicates, near-duplicates and clusters.

Much of what they do overlaps with much of what we do, as Web corpus collectors with language technology and linguistic research in mind. But the goals are different, which opens up a space to identify tasks that they perform well from their point of view but that is different to ours, and others that they do, but are not central to their concerns and we can do better.

An example of the first kind is de-duplication. In an impressive study of different methods, Monika Henzinger, formerly Director of Research at Google, discusses pages from a UK business directory that list in the centre the phone number for a type of business for a locality. Two such pages differ in five or less tokens while agreeing in about 1000. From Google's point of view they should not be classified as near-duplicates. From ours, they should. The paper by Gibson *et al.* in this volume addresses duplication from a WAC point of view.

The two biggest languages in the world, one of which is Google's home language, don't have much, or any, inflectional morphology, which may be why Google doesn't consider it so important for search. Speakers of most of the world's languages might give it a higher priority. In general, Google's spectacular performance relates to the languages where they have applied most effort, notably English. For Basque (for which the Web is not so large, and which has ample inflectional morphology) Leturia and colleagues clearly do beat Google on a number of counts.

We know that Google must do lots of text cleaning, as they succeed in finding terms for indexing and also are able to provide, for example, HTML versions of PDF or Word pages. But they do not publish details, so how might we find out what they do, and how it compares to what we do?

[1] As far as anything is long ago in its ten-year life. It was not yet a company when Tony Blair became UK Prime Minister, and was only a two-year-old when George W. Bush arrived in the White House.

[2] Most of Google's 6570 hits for *googlant* are for the present participle of the French verb; most of the 57,900 for *googlest* are for the second person singular of the German verb; most of the 66,400 hits of гуглить are for the infinitive of the Russian verb.

One way to explore the question is by looking at Web1T, a remarkable resource that Google generously provided for academic research in 2006 which lists all 1-, 2-, 3-, 4- and 5-grams occurring more than 40 times on the Google-indexed English Web. According to the brief description of the resource that is all that is provided, it is based on a trillion words. It seems likely that the counts are from de-duplicated pages. The text in the pages has clearly been identified as text (in contrast to images, formatting, etc), tokenised, and has had its language identified.

This resource can be compared to results used in the WaC community[3] and to traditional corpora, such as the BNC. Preliminary results show that our corpora are not worse than the results of Google. Web1T unigrams and bigrams contain more boilerplate (*unsubscribe, rss, forums*), business junk (*poker, viagra, collectibles*) as well as porn (*porn, lingerie*). There are reasons why this information is kept by Google: it is necessary to keep them as relevant keywords if someone is searching for a forum, poker or pornography.

However, we are different: we are searching constructions, not products. So we need different tools and resources, which cannot be provided by Google. Submissions to this volume show that the tools and resources can be provided by us.

Adam Kilgarriff, Serge Sharoff, Stefan Evert

[3] Sharoff in `http://wackybook.sslmit.unibo.it/` or Ferraresi *et al.* in this volume