LREC2008 – W19 – Workshop on

The 4th Web as Corpus Workshop: Can we do better than Google?

Sunday, 1 June 2008 – Karam 2

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 and 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