PREFACE

The Workshop on Corpus Profiling for Information Retrieval and Natural Language Processing took place in London, in October 2008, in conjunction with IliiX2008. Our aim was to bring together people from different research communities interested in exploring how specific properties of a corpus or collection affect the behaviour of techniques in Information Retrieval (IR) and Natural Language Processing (NLP), and to start mapping out a shared research agenda. These eWiCs Proceedings capture the final versions of papers presented at the workshop.

It is well known in NLP and IR that the properties of a collection or corpus have a significant impact on the effectiveness with which a technique can perform some task. In 1973, Spärck-Jones attributed differing degrees of success at automatic classification to differences in dataset characteristics. Since Croft and Harper (1979), IR performance has repeatedly been related to collection size and other features, though no upper bound has been found. Sensitivity to data collection characteristics has been highlighted, for instance, in several sub-areas of IR, anaphora resolution, automatic summarization and in word sense disambiguation. Recently, web enterprise and retrieval systems have emerged, whose performance is sensitive to URL properties, link graph properties, click streams, and other non-linguistic markers in a collection.

Systematic exploration of features that are relate collection characteristics to technique performance, has been missing from IR/NLP research. This gap creates deep-seated problems, for instance with replicability of experimental results (because the results of an experiment cannot necessarily be replicated on a different collection) and the development of applications (because it is unclear which techniques would work best for a new collection).

Over the past 15 years, several subject areas have converged with IR, and adopted research methodologies centred on large corpora and test collections. The time is right to address topics tracking the effects of dataset profiles. A BCS-IRSG workshop in London is an ideal opportunity to start with articulating a cross-disciplinary research agenda, building on the high concentration in Western Europe of IR, NLP and Semantic Web groups.

REFERENCES


ACKNOWLEDGEMENTS

We thank our invited speakers, the authors, the reviewers and all participants for their support and contribution. Our thanks are also due several people at the Open University: Damien Dadswell and Harriet Cornish who offered great help in design and maintenance of the workshop website Catherine McNulty, who took care of much of the formatting of the final papers published in the eWiC proceedings, and Debbie Briggs, who assembled the workshop hand-outs. We gratefully acknowledge the generous support from the British Computer Society and the BC-IRSG.

Anne De Roeck (The Open University)
Dawei Song (The Robert Gordon University)
Udo Kruschwitz (University of Essex)
Leif Azzopardi (University of Glasgow)