SPONSORS:

European Network in Human Language Technologies (ELSNET)
(http://www.elsnet.org)

Institute for Language, Speech and Hearing (ILASH), University of Sheffield, UK
(http://www.dcs.shef.ac.uk/research/ilash/)

WORKSHOP ORGANIZATION COMMITTEE:

Katerina Pastra
Department of Computer Science
University of Sheffield
Sheffield, S1-4DP, UK
Phone: +44 114-222-1945
katerina@dcs.shef.ac.uk
http://www.dcs.shef.ac.uk/~katerina

Yorick Wilks
Department of Computer Science
University of Sheffield
Sheffield, S1-4DP, UK
Phone: +44 114-222-1804
yorick@dcs.shef.ac.uk
http://www.dcs.shef.ac.uk/~yorick

SCIENTIFIC COMMITTEE:

Kalina Bontcheva	(University of Sheffield, UK)
Hamish Cunningham	(University of Sheffield, UK)
Rob Gaizauskas	(University of Sheffield, UK)
Donna Harman	(NIST, USA)
Lynette Hirschman	(MITRE, USA)
Maghi King	(ISSCO, Switzerland)
Steven Krauwer	(Utrecht University, The Netherlands)
Inderjeet Mani	(MITRE, USA)
Joseph Mariani	(LIMSI, France)
Patrick Paroubek	(LIMSI, France)
Katerina Pastra	(University of Sheffield, UK)
Martin Rajman	(EPFL - Switzerland)
Karen Spärck-Jones	(University of Cambridge, UK)
Horacio Saggion	(University of Sheffield, UK)
Beth Sundheim	(SPAWAR Systems Center, USA)
Simone Teufel	(University of Cambridge, UK)
Yorick Wilks	(University of Sheffield, UK)

WORKSHOP WEBSITE:

http://www.dcs.shef.ac.uk/~katerina/EACL03-eval/index.html

INTRODUCTION

Systems that accomplish different Natural Language Processing (NLP) tasks have different characteristics and therefore, it would seem, different requirements for evaluation. However, are there common features in evaluation methods used in various language technologies? Could the evaluation methods established for one type of systems be ported/adapted to another NLP research area? Could automatic evaluation metrics be ported? For instance, could Papineni's MT evaluation metric be used for the evaluation of generated summaries? Could the extrinsic evaluation method used within SUMMAC be applied to the evaluation of Natural Language Generation systems? What are the reusability obstacles encountered and how could they be overcome? What are the evaluation needs of system types such as dialogue systems, which have been less strenuously evaluated till now, and how could they benefit from current practices in evaluating Language Engineering technologies? What are the evaluation challenges that emerge from systems that integrate a number of different language processing functions (e.g. multimodal dialogue systems such as Smartkom)? Could resources (e.g. corpora) used for a specific NLP task, be reused for the evaluation of an NLP system and if so, what adaptations would this require?

Cross-fertilization of evaluation resources has taken place to some extent: in MUC, the extraction-specific adaptation of the standard Information Retrieval precision metric has been accepted as a standard for the evaluation of Information Extraction systems. In SUMMAC, parts of the TREC collection (documents, relevance assessments and even assessment software) have been reused. Both MTEval and SUMMAC have used conceptually similar approaches to evaluation (i.e. subject-based evaluation by testing reading comprehension). Many U.S. and European funding initiatives have been devoted to the evaluation of specific NLP systems, such as: MUC, SUMMAC, TREC and its follow-up initiative CLEF, MTEval and DUC. ISLE, the European initiative for establishing standards in Language Engineering has a working group on the evaluation of Machine Translation systems and its predecessor, EAGLES, has addressed evaluation issues for Language Engineering in general.

The ELSE project (1998-2000) was concerned with the evaluation infrastructure that could be deployed within the scope of the IST Key Actions of the 5th Framework Program of the European Community and indeed, the funding of evaluation activities has been addressed within the 5th Framework (as reported by Mariani and Paroubek 1999). Transatlantic co-operation for the evaluation of Human Language Technologies has also been stressed, among other issues, within an extensive report that was submitted to both the U.S. National Science Foundation and the European Commission's Language Engineering Office in 1999. This report mentions that evaluation techniques in the different Language Engineering areas grow more similar, a fact that emphasizes the need for co-ordinated and reusable evaluation resources.

The time has come to bring together all the above attempts to address the evaluation of NLP systems as a whole and explore ways for reusing established evaluation methods, metrics and other resources, thus, contributing to a more co-ordinated approach to the evaluation of language technology. This is exactly what this workshop has achieved: to bring together leading researchers from various NLP areas (such as Machine Translation, Information Extraction, Information Retrieval, Automatic Summarization, Question-Answering, Dialogue Systems and Natural Language Generation) in order to discuss this topic.

The papers included in this volume address issues of reuse of evaluation resources within and across NLP research areas. We cordially thank the authors and the members of the Programme Committee whose significant contributions made this workshop possible. We are especially grateful to our invited speakers: Donna Harman and Kevin McTait and ELSNET for its support and endorsement.

Katerina Pastra,
April 2003