CALL FOR PAPERS

ELRA Workshop on Evaluation
Looking into the Future of Evaluation: when automatic metrics meet task-based and performance-based approaches

To be held in conjunction with the 6th International Language Resources and Evaluation Conference (LREC 2008)

27 May 2008, Palais des Congrès Mansour Eddahbi, Marrakech

Background

Automatic methods to evaluate system performance play an important role in the development of a language technology system. They speed up research and development by allowing fast feedback, and the idea is also to make results comparable while aiming to match human evaluation in terms of output evaluation. However, after several years of study and exploitation of such metrics we still face problems like the following ones:

- they only evaluate part of what should be evaluated
- they produce measurements that are hard to understand/explain, and/or hard to relate to the concept of quality
- they fail to match human evaluation
- they require resources that are expensive to create

etc. Therefore, an effort to integrate knowledge from a multitude of evaluation activities and methodologies should help us solve some of these immediate problems and avoid creating new metrics that reproduce such problems.

Looking at MT as a sample case, problems to be immediately pointed out are twofold: reference translations and distance measurement. The former are difficult and expensive to produce, they do not cover the usually wide spectrum of translation possibilities and what is even more discouraging, worse results are obtained when reference translations are of higher quality (more spontaneous and natural, and thus, sometimes more lexically and syntactically distant from the source text). Regarding the latter, the measurement of the distance between the source text and the output text is carried out by means of automatic metrics that do not match human intuition as well as claimed. Furthermore, different metrics perform differently, which has already led researchers to study metric/approach combinations which integrate automatic methods into a deeper linguistically oriented evaluation. Hopefully, this should help soften the unfair treatment received by some rule-based systems, clearly punished by certain system-approach sensitive metrics.

On the other hand, there is the key issue of « what needs to be measured », so as to draw the conclusion that « something is of good quality », or probably rather « something is useful for a particular purpose ». In this regard, works like those done within the FEMTI framework have shown that aspects such as usability, reliability, efficiency, portability, etc. should also be considered. However, the measuring of such quality characteristics

cannot always be automated, and there may be many other aspects that could be usefully measured.

This workshop follows the evolution of a series of workshops where methodological problems, not only for MT but for evaluation in general, have been approached. Along the lines of these discussions and aiming to go one step further, the current workshop, while taking into account the advantages of automatic methods and the shortcomings of current methods, should focus on task-based and performance-based approaches for evaluation of natural language applications, with key questions such as:

- How can it be determined how **useful** a given system is for a given task?
- How can focusing on such issues and combining these approaches with our already acquired experience on automatic evaluation help us develop new metrics and methodologies which do not feature the shortcomings of current automatic metrics?
- Should we work on hybrid methodologies of automatic and human evaluation for certain technologies and not for others?
- Can we already envisage the integration of these approaches?
- Can we already plan for some immediate collaborations/experiments?
- What would it mean for the FEMTI framework to be extended to other HLT applications, such as summarization, IE, or QA? Which new aspects would it need to cover?

We solicit papers that address these questions and other related issues relevant to the workshop.

Workshop Programme and Audience Addressed

This full-day workshop is intended for researchers and developers on different evaluation technologies, with experience on the various issues concerned in the call, and interested in defining a methodology to move forward.

The workshop feature invited talks, submitted papers, and will conclude with a discussion on future developments and collaboration.

Workshop Chairing Team

Gregor Thurmair (Linguatec Sprachtechnologien GmbH, Germany) - chair

Khalid Choukri (ELDA - Evaluations and Language resources Distribution Agency, France) – co-chair
Bente Maegaard (CST, University of Copenhagen, Denmark) – co-chair

Organising Committee

Victoria Arranz (ELDA - Evaluations and Language resources Distribution Agency, France)
Khalid Choukri (ELDA - Evaluations and Language resources Distribution Agency, France)

Christopher Cieri (LDC - Linguistic Data Consortium, USA)
Eduard Hovy (Information Sciences Institute of the University of Southern California, USA)
Bente Maegaard (CST, University of Copenhagen, Denmark)
Keith J. Miller (The MITRE Corporation, USA)
Satoshi Nakamura (National Institute of Information and Communications Technology, Japan)
Andrei Popescu-Belis (IDIAP Research Institute, Switzerland)
Gregor Thurmair (Linguatec Sprachtechnologien GmbH, Germany)

Important dates

Deadline for abstracts: Monday 28 January 2008
Notification to Authors: Monday 3 March 2008
Submission of Final Version: Tuesday 25 March 2008
Workshop: Tuesday 27 May 2008

Submission Format

Abstracts should be no longer than 1500 words and should be submitted in PDF format to Gregor Thurmair at g.thurmair@linguatec.de.