Automatic Procedures in MT Evaluation

MT Summit XI, Copenhagen, 11 September 2007

Workshop organised by the ELRA Evaluation Committee

Gregor Thurmair, Linguatec
Khalid Choukri, ELDA
Bente Maegaard, University of Copenhagen

The purpose of this workshop is to discuss automatic evaluation procedures in MT. Among the discussion points are:

- What do the scores really measure?
- What kind of implicit assumptions do they make?
- What kind initial effort do they require (e.g.: pre-translate test corpus)?
- What kind of resources do they need (e.g.: third party grammars)?
- Are they biased towards specific MT technologies?
- What kind of diagnostic support can they give? (where to improve the system)
- What kind of evaluation criteria (e.g. related to the FEMTI framework) do they support (adequacy, fluency, ...)

The objective of the workshop is to have a better understanding of the strengths and limitations of the respective approaches, and perhaps make steps towards defining a common methodology for MT output evaluation.

Programme

9.00	Welcome and introduction
9.20	*The place of automatic evaluation metrics in external quality models for machine translation* Andrei Popescu-Belis, University of Geneva
10.00	*Evaluating Evaluation --- Lessons from the WMT'07 Shared Task* Philipp Koehn, University of Edinburgh
10.30	Coffee break
11.00	*Investigating Why BLEU Penalizes Non-Statistical Systems* Eduard Hovy, University of Southern California
11.30	*Edit distance as an evaluation metric* Christopher Cieri, Linguistic Data Consortium (TBC)
12.00	*Experience and conclusions from the CESTA evaluation project* Olivier Hamon, ELDA

12.30	Lunch
13.30	*Automatic Evaluation in MT system production* Gregor Thurmair, Linguatec
14.00	*Sensitivity of performance-based and proximity-based models for MT evaluation* Bogdan Babych, Univ. Leeds
14.30	*Automatic & human Evaluations of MT in the framework of a speech to speech communication* Khalid. Choukri, ELDA
15.00	Coffee break
15.30	Discussion and conclusions
17.00	Close