Appendix II: Discussion Panel on Evaluation in Generation Research

Moderator: Inderjeet Mani

Evaluation is critical in offering feedback on progress to both developers and potential consumers of NLG technology. However, evaluation has thus far not been as well-established in NLG as it has become in NLU. This panel will discuss evaluation methods and resources. It is aimed at building a better understanding of NLG evaluation methods, and hopefully arriving at steps to facilitate future evaluations.

Applicable evaluation methods can be derived from work in NLG as well as Text Summarization and Machine Translation. The evaluation methods include intrinsic methods which test the generation system in itself, and extrinsic methods which test the generation system in relation to some other task.

Intrinsic methods can include assessing coverage of different varieties of generation input, the quality of the generated output, and comparison of generated output against reference output at some level (*e.g.*, by using subjective grading, comparison against templates, or comparing human correctness in answering questions based on each type of output, etc.) Of course, a fundamental problem in evaluating NLG is that there may be many acceptable outputs.

Extrinsic methods can include measuring efficiency in executing generated instructions (*e.g.*, how easy was it to install the component by following the generated manual?), assessing the relevance of generated output to some information need or goal (*e.g.*, are the generated business letters effective?), its impact on a system in which it is embedded (*e.g.*, how much does the generation help the question answering system?), measuring the amount of effort required to post-edit the output (*e.g.*, how much do the generated briefings need to be fixed up?), etc.

As the generation technology becomes more mature, it is useful to assess end-user acceptability of generated output, extensibility and portability, throughput, cost-benefit measures, etc. It is also interesting for evaluations address both features important to the overall task, as well as features unique to NL generation.

Participants will address the following issues:

1. What evaluation methods are applicable to NLG?

2. What are the pros and cons of NLG evaluations you have carried out?

3. Can we construct corpora to help evaluate NLG systems?

4. What steps can we collectively take to improve the role of evaluation in NLG?