Introduction

The aim of this workshop is to identify and to synthesize current needs for language-technology evaluation. The first part of the workshop will focus on one of the most challenging current issues in language engineering: the evaluation of dialogue systems and models. The second part will extend the discussion to address the problem of evaluation in language engineering more broadly and on more theoretical grounds.

The space of possible dialogues is enormous, even for limited domains like travel information servers. The generalization of evaluation methodologies across different application domains and languages is an open problem. Review of published evaluations of dialogue models and systems suggests that usability techniques are the standard method. Dialogue-based system are often evaluated in terms of standard, objective usability metrics, such as task-completion time and number of user actions. In the past, researchers have proposed and debated theory-based methods for modifying and testing the underlying dialogue model, but the most widely used method of evaluation is usability testing, although more precise and empirical methods for evaluating the effectiveness of dialogue models have been proposed. For task-based interaction, typical measures of effectiveness are time-to-completion and task outcome, but the evaluation should focus on user satisfaction rather than on arbitrary effectiveness measurements. Indeed, the problems faced in current approaches to measurement of effectiveness dialogue models and systems include:

1. Direct measures are unhelpful because efficient performance on the nominal task may not represent the most effective interaction
2. Indirect measures usually rely on judgment and are vulnerable to weak relationships between the inputs and outputs
3. Subjective measures are unreliable and domain-specific

Representative questions to be addressed include but are not limited to:

1. How do we deal with the combinatorial explosion of dialogue states?
2. How can satisfaction be measured with respect to underlying dialogue models?
3. Are there useful direct measures of dialogue properties that do not depend on task efficiency?
4. What is the role of agent-based simulation in evaluation of dialogue models?

Of course, the problems faced in evaluating dialogue and system models are found in other domains of language engineering, even for non-interactive processes such as part-of-speech tagging, parsing, semantic disambiguation, information extration, speech transcription, and audio document indexing. So the issue of evaluation can be viewed at a more generic level, raising fundamental, theoretical questions such as:

1. What are the interest and benefits of evaluation for language engineering?
2. Do we really need these specific methodologies, since a form of evaluation should always be present in any scientific investigation?
3. If evaluation is needed in language engineering, is it the case for all domains?
4. What form should it take? Technology evaluation (task-oriented in laboratory environment) or field/user Evaluation (complete systems in real-life conditions)?

5. We have seen before that the the evaluation of dialogue models is still unsolved, but for domains where metrics already exists, are they satisfactory and sufficient? How can we take into account or abstract from the subjective factor introduced by human operators in the process?

6. Do similarity measures and standards offer appropriate answers to this problem? Most of the efforts focus on evaluating process, but what about the issue of language resources evaluation?

In the second part of the workshop we wish to address the problem of evaluation both from a broader perspective, including novel applications domain for evaluation, new metrics for known tasks and resource evaluation, as well as look at the problem from a more theoretical point of view, including the isssue of formal theory of evaluation and infrastructural needs of language engineering. David

Novick & Patrick Paroubek.