Position Paper: Towards Standardized Metrics and Tools for Spoken and Multimodal Dialog System Evaluation

Sebastian Möller, Klaus-Peter Engelbrecht,
Florian Kretzschmar, Stefan Schmidt, Benjamin Weiss
Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin
Ernst-Reuter-Platz 7
10587 Berlin, Germany
sebastian.moeller@telekom.de

Abstract

We argue that standardized metrics and automatic evaluation tools are necessary for speeding up knowledge generation and development processes for dialog systems.

1 Introduction

The Spoken Dialogue Challenge launched by CMU (Black et al., 2011) provides a common platform for dialog researchers in order to test the performance of their systems and components against the state-of-the-art. Still, evaluations are individual undertakings in most areas, as common metrics and procedures which would be applicable for a range of systems are sparse. In the following, it is argued that significant progress can be made if three prerequisites are available:

- Common metrics for quantifying user and system interaction behavior and perceived quality
- Reliable models for predicting user judgments on the basis of automatically-extracted or annotated interaction metrics
- Methods for realistically simulating user behavior in response to dialog systems

The state-of-the-art and necessary research in these three areas is outlined in the following paragraphs. The Spoken Dialogue Challenge can contribute to validating such metrics and models.

2 Common Metrics

Whereas early assessment and evaluation cycles were based on ad-hoc selected metrics, approaches have been made to come up with a standard set of metrics for quantifying interactions between users and systems which would make evaluation exercises comparable. The International Telecommunication Union (ITU-T) has standardized two sets of metrics: ITU-T Suppl. 24 to P-Series (2005) for spoken dialog systems, and ITU-T Suppl. 25 to P-Series Rec. (2011) for multimodal dialog systems. These metrics describe system performance (e.g. in terms of error rates) and user/system interaction behavior (e.g. in terms of meta-communication acts, durations) in a quantitative way, and can thus serve as an input to the models discussed below. Input is welcome to stabilize these metrics, so that they are of more use to researchers and system developers. The proper conjunction between such metrics and standardized annotation schemes (e.g., Bunt et al., 2010) will strengthen the establishment and spreading of a specific set of metrics.

When it comes to user-perceived quality, Hone and Graham (2000) have made a first attempt to come up with a validated questionnaire (SASSI), which, however, lacks a scale to assess speech output quality. The approach has been put forward in ITU-T Rec. P.851 (2003) by including speech output and dialog managing capabilities. A framework structure was preferred over a fixed (and validated) questionnaire, in order to more flexibly address the needs of researchers and developers. This approach still needs to be extended towards multimodal systems, where modality appropriateness, preference and perceived performance have to be considered. ITU-T welcomes contributions on this topic.

NAACL-HLT 2012 Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data, pages 5–6,
Montréal, Canada, June 7, 2012.

For practical usage, it is desirable to have evaluation methods which provide diagnostic value to the system developer, so that the sources of misbehavior can be identified. The diagnosis can be based on perceptual dimensions (effectiveness, efficiency, mental effort, etc.) or on technical characteristics (error rates, vocabulary coverage, etc.) or both. Approaches in this direction are welcome and would significantly increase the usefulness of evaluation exercises for the system developers.

3 User-perceived Quality Prediction

The first approach to predict user judgments on the basis of interaction metrics is the well-known PARADISE model (Walker et al., 1997). The main challenge to date is the low generalizability of such models. The reason is that many of the underlying input parameters are interdependent, and that a simple linear combination does not account for more complex relationships (e.g. there might be an optimum length for a dialog, which cannot be easily described by a purely linear model).

However, other algorithms such as non-linear regression, classification trees or Markov models, have not shown a significantly improved performance (Möller et al., 2008; Engelbrecht, 2011). The latter are however adequate to describe the evolution of user opinion during the dialog, and thus might have principled advantages over models which use aggregated interaction performance metrics as an input.

4 User Behavior Simulation

During system development, it would be useful to anticipate how users would interact with a dialog system. Reflected to the system developer, such anticipations help to identify usability problems already before real users interact with the system.

Whereas user behavior simulation has frequently been used for training statistical dialog managers, only few approaches are documented which apply them to system evaluation. Early approaches mainly selected possible utterances from a set of collected data. The MeMo workbench (Engelbrecht, 2011) tried to combine statistical selection of probable interaction paths with the knowledge of usability experts about what typically influences user behavior. Such knowledge can also be generated by a conversational analysis and categorization (Schmidt et al., 2010).

A different approach has been followed in the SpeechEval project (Möller et al., 2009) where statistical dialog managers have been trained on a large diverse dataset to generate utterances on a conceptual level. The system is then amended with ASR and TTS to allow for a speech-based black-box interaction with telephone-based dialog systems. Combined with diagnostic quality prediction models, such tools can support system developers to evaluate different dialog strategies early in the design cycle and at low costs, and thus avoid dissatisfied users. The approach still has to be extended towards multimodal dialog systems.

References

Alan W Black et al, *Spoken Dialog Challenge 2010: Comparison of Live and Control Test Results*, Proc. SIGDIAL2011, Portland, OR.

H. Bunt, et al.: *Towards an ISO standard for dialogue act annotation*. Proc. LREC 2010, 19-21.

K.-P. Engelbrecht. 2011. *Estimating Spoken Dialog System Quality with User Models*, Doctoral Dissertation, TU Berlin, to appear with Springer, Berlin.

K.S. Hone, R. Graham. 2000. Towards a Tool for Subjective Assessment of Speech System Interfaces (SASSI), *Natural Language Eng.*, 6(3-4):287-303.

ITU-T Rec. P.851. 2003. *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems*, Int. Telecomm. Union, Gerneva.

ITU-T Suppl. 24 to P-Series Rec. 2005. *Parameters Describing the Interaction with Spoken Dialogue Systems*, Int. Telecomm. Union, Geneva.

ITU-T Suppl. 25 to P-Series Rec. 2011. *Parameters Describing the Interaction with Multimodal Dialogue Systems*, Int. Telecomm. Union, Geneva.

S. Möller, K.-P. Engelbrecht, R. Schleicher. 2008. Predicting the Quality and Usability of Spoken Dialogue Services, *Speech Communication* 50:730-744.

S. Möller, R. Schleicher, D. Butenkov, K.-P. Engelbrecht, F. Gödde, T. Scheffler, R. Roller, N. Reithinger. 2009. Usability Engineering for Spoken Dialogue Systems Via Statistical User Models, in: *First IWSDS 2009*, Kloster Irsee.

M.A. Walker, D.J. Litman, C.A. Kamm, A. Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents, *Proc. ACL/EACL 35th Meeting*, Madrid, 271-280.

S. Schmidt, J. Stubbe, M. Töppel, S. Möller. 2010. Automatic Usability Evaluation for Spoken Dialog Systems Based on Rules Identified by a Sociotechnical Approach, in: *Proc. PQS 2010*, Bautzen.