ABSTRACT

This paper discusses issues in evaluating spoken language dialogue systems in terms of technical performance and end-user acceptance. Recent efforts in this domain have been carried out in the framework of two major research initiatives: the European Esprit long-term project Spoken Language Dialogue Systems and Components - Best practice in development and evaluation (DISC) and the US American DARPA COMMUNICATOR project.

1. INTRODUCTION

Recently spoken Language Dialogue systems (SLDSs) researchers and system developers have begun to investigate systems evaluation in its full complexity. SLDS performance measures may address different aspects. Technical evaluations measure the system or component performance on the basis of pre-defined references with respect to the actions and reactions taken. The development of a spoken language system is an incremental process by which errors are analyzed and systems refined. This process can be performed by periodic technical performance assessments. Subjective evaluations, on the other hand, are used to investigate the user friendliness, reliability and ease of use of a SLDS.

Widely known experiments for technical SLDSs evaluation have been carried out for information retrieval tasks in the American ARPA-ATIS\(^1\) project [1]. ARPA, now DARPA, has been one of the first coordinators to establish periodic evaluation campaigns and to encourage objective system performance comparisons across the participating sites. The campaigns, accessible to various public and industrial laboratories, were aimed at sharing knowledge and scientific competencies. Speech recognition accuracies of spontaneous speech were provided, as well as an assessment of the ability of a spoken language system to provide the correct answer based on both speech recognition and natural language understanding.

The aim of the COMMUNICATOR project [2], also funded by DARPA, is to support rapid and cost-effective development of SLDSs. In order to achieve these goals, systems evaluation is considered as a central research issue. It is based on a methodology that learns general performance functions for SLDSs from experimental dialogue data, collected from the project partners in a common data collection effort. The evaluation framework is based on the assumption, that user satisfaction is the overall objective to be maximized and that task success and various interaction costs can be used to predict this user satisfaction.

Analogous efforts to better coordinate system evaluations have been pursued in several European projects. Having multilingual aspects in mind, evaluations needed to be organized on a larger number of languages. To date, European programs and activities have been concerned with the definition of common baselines for system evaluations. The ESPRIT project SAM (Multilingual Speech Input/Output Standardization, Assessment and Methodology) was one of the first projects to define standards for speech recognition systems, speaker verification and speech synthesis.

The DISC project was launched in 1997 in order to develop a first dialogue engineering best practice methodology [3]. The project addresses technical evaluation of SLDSs for different application domains as well as the components of these systems. Evaluation from the point of view of the end-user has not been fully addressed, mainly because of the scarcity of available dialogue data. Other European activities focused on establishing representative spoken language corpora. They include COCOSDA (Coordinating Committee on Speech Databases and Speech I/O Systems Assessment) for coordination of multilingual databases [4], RELATOR for defining a structure for data collection and distribution of linguistic resources, and ELRA (European Language Resources Association), which is the European equivalent of LDC (Linguistic Data Consortium).

\(^1\) ATIS was a designated common research task for data collection and evaluation support within the ARPA (Advanced Research Projects Agency) Speech and Natural Language program. An ATIS system allows the user to acquire information derived from the Official Airline Guide about fares and flight schedules between a restricted set of cities within the United States and Canada. Other information, such as the meals served on the flight or the type of aircraft, is also available.
(Speech Recognition Quality Assessment for Linguistic Engineering) project [5] aimed to adapt the ARPA-LVCSR (Large Vocabulary Continuous Speech Recognition) evaluation paradigm [1] to a multilingual context. The project EAGLES (Expert Advisory Group on Language Engineering Standards) has created a manual for the development of linguistic resources and the evaluation of spoken language systems [6]. In France, AUPELF-UREF (Association des universités partiellement ou entièrement de langue française - université des réseaux d’expression française), today AUF (Agence universitaire de la francophonie) has organized evaluation campaigns on speech dictation and speech synthesis for the French language [7, 8].

2. EVALUATION IN THE DISC PROJECT

The DISC project [3], terminated in 1999, mainly addresses technical evaluation of SLDSs for different application domains. It presented a methodology for complete and correct evaluation of SLDSs and its components, and elaborated an evaluation template for describing the criteria needed in performing this task.

The results of the project are based on a detailed analysis of SLDSs and their components for different application domains and human languages. The SLDSs that have been made available to the consortium are research prototypes and commercial systems.

In the DISC approach, a SLDS is a speech-only system that includes the following six aspects: speech recognition, speech generation, natural language understanding and generation, dialogue management, human factors and systems integration.

The best practice of the DISC methodology consists of choosing among a set of available options within each aspect (e.g., a decision on whether the dialogue manager uses a system-directed, mixed initiative, or user-directed dialogue strategy), and the best development and evaluation practices for these options.

The methodology is based on an underlying grid and life-cycle best analysis:

- **Grid.** It defines a space of aspect-specific issues which the system developer needs to account for (e.g., in terms of dialogue-management: who should have the initiative?).

- **Life-cycle.** It includes recommendations on how the development process (of the aspect and its options) should be carried out (e.g., a requirement analysis in the dialogue manager specification phase).

The DISC evaluation template [3] to be filled in by the system developer for each SLDS aspect is a model of what the developer needs to know if he is planning to evaluate a particular property of a SLDS and its components.

This knowledge is specified by the following ten entries:

1. **What is being evaluated?** The generic or specific property or properties of a SLDS or the component(s) that is (are) being evaluated.

2. **System part evaluated.** The component(s) of a SLDS that is (are) being evaluated.

3. **Type of evaluation.** It may be objective (quantitative and qualitative), subjective or comparative (internal and external).

4. **Methods of evaluation.** These are design analysis, WOZ (Wizard of Oz) data analysis, questionnaires and interviews, diagnostic evaluation (glass box and black box), test suites, user-system interaction and data analysis.

5. **Symptoms to look for.** The symptoms an expert evaluator should look for in the evaluation data.

6. **Life-cycle phases.** The life cycle phases in which evaluation of the property in question should be performed.

7. **Importance of evaluation.** Importance of evaluating a certain property of a SLDS (relevance with respect to all or only some current systems or components, possible penalties).

8. **Difficulty of evaluation.** Comments on the difficulties involved in performing the evaluation (various forms of complexities, unresolved research problems).

9. **Cost of evaluation.** Costs involved in performing the evaluation in terms of time and manpower.

10. **Evaluation tools.** Software tools or any other kinds of support which may help in performing the evaluation.

In summary, the DISC project is concerned with the task of specializing software engineering best development and evaluation practice applied to the SLDSs design. The project states that a SLDS can be modeled as having at most six aspects, each of which can be presented as a space of issues to be addressed, and each of which can be described by a development process.

The DISC best practice methodology aims to support the system developer in choosing a set of appropriate options for the application at hand and in properly developing and evaluating the SLDS that incorporates those options. Both aims are expressed by the filled evaluation template to be used in the development of the SLDS and its components.
3. EVALUATION IN THE COMMUNICATOR PROJECT

The aim of the COMMUNICATOR project [2, 9] is to support rapid and cost-effective development of SLDSs. This goal specifies that the project participating sites, mostly US American Research Laboratories, are required to use a common system architecture, the COMMUNICATOR HUB, by following a set of standards that promote interoperability and plug and play of similar components. Some of these standards have been drawn from the commercial domain, while others have been established by the project itself.

The COMMUNICATOR architecture enables the development of sharable human-computer interface components for speech recognition, speech synthesis, dialogue management, natural language understanding, contextual interpretation and natural language generation. A shared research environment using the common task of air travel information services, common data and a common evaluation framework has been created on a Web site. It allows developers to quickly assemble and test new, architecture-compliant interfaces. The software repository enables the project partners to contribute and to access architecture-compliant modules. The Web-accessible testbed allows developers to plug and play the various components in the repository and thereby coordinate their component with the others. In addition, the testbed provides data-gathering facilities.

The COMMUNICATOR evaluation paradigm is an extension and refinement of the PARADISE framework [10, 9]. This framework provides a methodology for learning general performance functions for SLDS systems from experimental data collected and analyzed from the project partners in a common data collection effort. The goal is to maximize user satisfaction, measured by some comparable diagnostic metrics, that are automatically logged and computed [10]:

- **Dialogue efficiency.** Total elapsed time, time to accomplish the task, number of system turns, turns on task, time per turn for each system module.

- **Dialogue quality.** Word error rate, re-prompts, error messages, help messages, response latency, mean word error rate, mean response latency, variance response latency.

- **Task success.** Perceived task completion, objective task completion.

In addition, user satisfaction questions on a survey probe a number of different aspects of the user’s perceptions of their interaction with the system in order to focus the user on the task of rating the system on the basis of a five point multiple choice scale. The values for all the responses are summed, resulting in a user satisfaction measure for each dialogue ranging from 5 to 25.

In summary, evaluation in the COMMUNICATOR project is based on the application of an empirical approach to developing general models of system usability. User satisfaction is used as a measure of system usability. The general models for predicting user satisfaction are based on experimental data collection and analyses.

4. DISCUSSION AND OUTLOOK

With research in the area of SLDSs having made rapid progress over the last decade, these systems are proliferating on the market for a large variety of applications in an increasing number of languages. Prototype systems have become subject to evaluation by real-users and under realistic conditions, which is well illustrated by the EU-project SENECA (Speech Control modules for Entertainment, Navigation and communication Equipment in Cars) [11]. This recent development shows that assessing the quality of such a SLDS becomes a central issue in research and development.

However, the twofold evaluation task reveals some considerable difficulties. The system developer has first to consider high-level technical aspects of the evaluation in order to guarantee the good performance of the SLDS. This type of evaluation is arbitrarily complex and depends on:

- on the number of the individual system components and their complexities,

- on the way of how these components interact,

- on the application at hand that strongly influences the overall SLDSs architecture and complexity,

- on the human language employed that at least governs systems portability, but in most cases alters the interaction type due to cultural differences.

One of the strong points of the best practice methodology elaborated in the framework of the European DISC project is its focus on the technical evaluation of systems and their components. This evaluation aspect could be established because of the availability of a wide range of operational SLDSs in different application domains and for different human languages. The SLDSs global community strongly benefits from having best practice standards in this field.

However, rather than in simply technical aspects, end-users tend to be more interested in global evaluations that take into account many different properties. Such evaluations are at best qualitative (technical) but also require subjective elements.
One of the weaknesses of the DISC project seems to be the lack in large-scale end-user evaluations. This represents a significant effort and requires a coordinated collection, costly analyses and annotation of large data amounts. Unlike US American research projects, such as those coordinated by DARPA, where working on a common task using common data and a common evaluation method has already become a culture, the focus of European programs has always been on the multi-task and multi-lingual aspects of SLDSs. One of the results of this research orientation is that to date, systems have been developed for a large variety of application domains and in an increasing number of languages.

In turn, evaluation programs and paradigms, such as ATIS and COMMUNICATOR motivate SLDSs researchers and developers to compete in building advanced systems and to share information in resolving problems. They help to provide rapid technical progress and yield a significant deal of technology transfer and communication.

The attempt of quantifying user satisfaction and to draw conclusions on system performance, as proposed in the COMMUNICATOR project, can be seen as complementary to the exclusively technical considerations such as those primarily addressed by the DISC evaluation template. However, it seems to be dangerous to conclude on systems performance exclusively from data analyses.

A weak point of the COMMUNICATOR project status seems to be that the research effort is limited to the specific task of information retrieval in the travel domain. This implies that the evaluation does not account for different types of SLDSs in completely different application domains (machine translation as an example for human-human dialogue support). Finally, a SLDSs developer should consider that users from different cultural groups may behave different and therefore provide different interaction results.

In a general perspective, coordinated evaluation programs are labor-intensive and compete in time and resources with other activities, specifically with the development of new approaches. Excessive focus on competitive performance evaluation may lead to risk-avoidance strategies, where getting a good score becomes crucial and takes over innovative research.

5. ACKNOWLEDGEMENTS

The author is grateful to J. Aberdeen, N.O. Bernsen, L. Dybkjaer, L. Hirschmann, and M.A. Walker for the project descriptions that provide the basis of this paper.

6. REFERENCES


