Chapter 1: Introduction

Technological advances have made possible the existence of real-time, interactive spoken-dialogue systems for a wide variety of applications. Spoken Dialogue Systems (SDS´s) are computer-based systems developed to provide information and carry out tasks using speech as the interaction mode. They are capable of speech recognition, interpretation, management of dialogue and have speech output capabilities, trying to reproduce a more or less natural spoken interaction between a human user and the system. SDS´s provide several different services like obtaining train schedule information, booking a hotel, telephone banking, control of domestic devices and even assisting astronauts in performing procedures on the International Space Station, all through spoken language.

Even with all this development, there is scarcity of information on ways to assess and evaluate the quality of such systems with the purpose of optimization.

With two of these SDS´s (BoRIS and INSPIRE, details on Chapter 3), extensive experiments were conducted in the past, where the systems were used to resolve specific tasks. The evaluators rated the quality of the system on a multitude of scales. In addition to that, the interactions were recorded and annotated by an expert.

The development of methods for performance evaluation is an open research issue in this area of SDS´s. Following the idea of the PARADISE model (PARAdigm for DIalogue System Evaluation model, the most well-known model for this purpose, developed by Walker and co-workers at AT&T (see [4] and [5])), several experiments were conducted to develop predictive models of spoken dialogue performance.

1.1
Objectives and overview of this Dissertation

Quality is neither an absolute nor an inherent property of a system (or a product in general), but depends on the specific users. Whenever users find it useful for themselves, it has quality under their viewpoint. The quality perceived

by the user is a combination between what he/she expects or desires, and the characteristics he/she perceives while using the service. Thus, any engineering approach to quality includes consideration of how the systems are received/perceived by the users and of how the needs and expectations of the users develop.

The objective of this study is to develop and assess models which allow the prediction of quality dimensions as perceived by the human user, based on instrumentally measurable variables using all the collected data from the BoRIS and INSPIRE systems. Different types of algorithms will be compared to their prediction performance and to how generic they are.

The spoken language used on all studied systems is German, so therefore all analyses and conclusions from this work should be generalized carefully to other languages.

Four different approaches will be used for these analyses: Linear Regression, Regression Trees, Classification Trees and Neural Networks.

For each of these methods, a different tool will be programmed using MATLAB, that can carry out all experiments from this work and be easily modified for new experiments with data from new systems or new variables on future studies. All the used MATLAB programs are available on the attached CD with an "operation manual" for future users as well as a guide to modify the existing programs to work on new data.

The main idea is to develop tools that would help on the optimization of a spoken dialogue system without a direct involvement of the human user or serve as tools for future studies in this area.

Until now, the most well-known model for this purpose is the PARADISE (PARAdigm for Dialogue System Evaluation) model, which has been developed by Walker and co-workers at AT&T (see [4] and [5]). It is used to predict performance evaluation, using multivariate linear regression, but it only gives results from training (in-sample) data, making the results restricted. The focus of these experiments that will be done with the developed tools is to build a generic model that could deal well with independent test (out-of-sample) data, since this is the main goal of a prediction model for performance evaluation.

After all experiments are made with models built to predict performance on its own systems, experiments where the models generalize across the two systems or system's configurations will be made as well.

Our work therefore can be divided in six chapters:

1- Choosing the adequate variables to be used on the predictions.

2- Programming the tools for each of the approaches.

3- Trying out and enhancing tools with the adequate techniques to generalize methods for test (out-of-sample) data.

4- Carrying out experiments.

5- Writing conclusions based on the experiments results, comparing methods on several different aspects.

6- Writing prospects and propositions for enhancements on the methods presented and future studies in this area.

The work done on each of these six phases can then be divided on the next chapters that compose this dissertation:

-Chapter 2: Evaluation of Spoken Dialogue Systems – the objective of the dissertation is explained in detail. All the variables, independent and dependent, are described.

-Chapter 3: Experiment – all experiments are described in detail.

-Chapter 4: Prediction Models – all models used for prediction are explained in detail.

-Chapter 5: Results and Statistic Analysis – the results from the experiments are presented and analysed.

-Chapter 6: Conclusion – conclusions extracted from the results are presented and future prospects for the area are proposed.