



Bernardo Lins de Albuquerque Compagnoni

**Development of Prediction Models for the
Quality of Spoken Dialogue Systems**

Dissertação de Mestrado

Dissertation presented to the Postgraduate Program in Electrical Engineering of the Departamento de Engenharia Elétrica, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Engenharia Elétrica.

Advisor: Prof. Cristiano Augusto Coelho Fernandes

Rio de Janeiro

April 2011



Bernardo Lins de Albuquerque Compagnoni

**Development of Prediction Models for the
Quality of Spoken Dialogue Systems**

Dissertation to the presented Postgraduate Program
in Electrical Engineering, of the Departamento de
Engenharia Elétrica do Centro Técnico Científico da
PUC-Rio, as partial fulfillment of the requirements for
the degree of Mestre

Prof. Cristiano Augusto Coelho Fernandes
Advisor

Departamento de Engenharia Elétrica – PUC-Rio

Prof. Fernando Gil Vianna Rsende
UFRJ

Prof. Abraham Alcaim
Centro de Estudos em Telecomunicações – PUC-Rio

Prof. José Eugenio Leal
Coordinator of the Centro Técnico Científico–
PUC-RIO

Rio de Janeiro, 04 de abril de 2011

All rights reserved.

Bernardo Lins de Albuquerque Compagnoni

Graduated in Electrical Engineering from PUC-Rio in 2008. Participated in a double degree program with the University of Braunschweig (Germany), between 2004 and 2006. Worked at Deutsche Telekom Laboratories in Berlin in 2006, where he did research on the topic of this dissertation. After graduating, he joined the Post-Graduate Program in Electrical Engineering from PUC-Rio to obtain a Masters degree

Bibliographic data

Compagnoni, Bernardo Lins de Albuquerque

Development of prediction models for the quality of spoken dialogue systems / Bernardo Lins de Albuquerque Compagnoni; advisor: Cristiano Augusto Coelho Fernandes. – 2011.

137 f. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2011.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Reconhecimento de voz. 3. Linguagem falada. 4. Spoken dialogues systems. 5. Regressão linear. 6. Árvores de regressão. 7. Árvores de classificação. 8. Redes neurais. 9. Avaliação de performance de sistemas. I. Fernandes, Cristiano Augusto Coelho. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

À minha filha Alice

Agradecimentos

À minha família. Ao meu pai Luiz, minha mãe Maria Regina, meu irmão Tiago e minha filha Alice, além de todos os meus amigos, pelo carinho, amor e apoio neste desafio. Sem eles seria impossível dar este passo.

Ao orientador Cristiano Fernandes, pela oportunidade concedida e confiança nas responsabilidades envolvidas. Agradeço pela excelente orientação em todas as etapas deste trabalho.

Aos orientadores Sebastian Möller e Tim Fingscheid, pela motivação e oportunidades concedidas durante a minha estadia na Alemanha.

À PUC-Rio, pelos auxílios concedidos e pelo ótimo ambiente de estudo.

Resumo

Compagnoni, Bernardo Lins de Albuquerque; Fernandes, Cristiano Augusto Coelho (Orientador). **Desenvolvimento de Modelos para Previsão de Qualidade de Sistemas de Reconhecimento de Voz**. Rio de Janeiro, 2011. 137p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Spoken Dialogue Systems (SDS's) são sistemas baseados em computadores desenvolvidos para fornecerem informações e realizar tarefas utilizando o diálogo como forma de interação. Eles são capazes de reconhecimento de voz, interpretação, gerenciamento de diálogo e são capazes de ter uma “voz” como saída de dados, tentando reproduzir uma interação natural falada entre um usuário humano e um sistema. SDS's provém diferentes serviços, todos através de linguagem falada com um sistema. Mesmo com todo o desenvolvimento nesta área, há escassez de informações sobre como avaliar a qualidade de tais sistemas com o propósito de otimização do mesmo. Com dois destes sistemas, BoRIS e INSPIRE, usados para reservas de restaurantes e gerenciamento de “casas inteligentes“, diversos experimentos foram conduzidos no passado, onde tais sistemas foram utilizados para resolver tarefas específicas. Os participantes avaliaram a qualidade do sistema em uma série de questões. Além disso, todas as interações foram gravadas e anotadas por um especialista. O desenvolvimento de métodos para avaliação de performance é um tópico aberto de pesquisa na área de SDS's. Seguindo a idéia do modelo PARADISE (PARAdigm for Dialogue System Evaluation – desenvolvido por Walker e colaboradores na AT&T em 1998), diversos experimentos foram conduzidos para desenvolver modelos de previsão de performance de sistemas de reconhecimento de voz e linguagem falada. O objetivo desta dissertação de mestrado é desenvolver modelos que permitam a previsão de dimensões de qualidade percebidas por um usuário humano, baseado em parâmetros instrumentalmente mensuráveis utilizando dados coletados nos experimentos realizados com os sistemas BoRIS e INSPIRE, dois sistemas de reconhecimento de voz (o primeiro para busca de restaurantes e o segundo para Smart Homes). Diferentes algoritmos serão utilizados para análise (Regressão linear, Árvores de Regressão, Árvores de Classificação e Redes Neurais) e para cada um dos algoritmos, uma ferramenta diferente será programada em MATLAB, para poder servir de base para análise de experimentos futuros, sendo facilmente modificado para sistemas e parâmetros novos em estudos subsequentes. A idéia principal é desenvolver ferramentas que possam ajudar na otimização de um SDS sem o envolvimento direto de um usuário humano ou servir de ferramenta para estudos futuros na área.

Palavras-chave

Engenharia Elétrica; Reconhecimento de Voz; Linguagem Falada; Spoken Dialogue Systems; Regressão linear; Árvores de Regressão; Árvores de Classificação; Redes Neurais; Avaliação de Performance de Sistemas.

Abstract

Compagnoni, Bernardo Lins de Albuquerque; Fernandes, Cristiano Augusto Coelho (Advisor). **Development of Prediction Models for the Quality of Spoken Dialogue Systems**. Rio de Janeiro, 2011. 137p. MSc Dissertation – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

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, 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, (used for Restaurant Booking Services and Smart Home Systems), 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 in 1998), several experiments were conducted to develop predictive models of spoken dialogue performance. The objective of this dissertation 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. 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 will be made 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.

Keywords

Electrical Engineering; Speech Recognition; Spoken Language; Spoken Dialogue Systems; Linear Regression; Regression Trees; Classification Trees; Neural Networks; System Performance Evaluation.

Summary

LIST OF DIAGRAMS AND FIGURES	10
LIST OF TABLES	11
CHAPTER 1: INTRODUCTION	15
1.1 OBJECTIVES AND OVERVIEW OF THIS DISSERTATION	15
CHAPTER 2: EVALUATION OF SPOKEN DIALOGUE SYSTEMS	18
2.1 SUBJECTIVE EVALUATION	18
2.2 PARAMETRIC DESCRIPTION OF INTERACTION	20
2.2.1 INSTRUMENTALLY-MEASURED AND EXPERT-ANNOTATED VARIABLES	21
CHAPTER 3: EXPERIMENT	33
3.1 BORIS AND INSPIRE SYSTEMS	33
3.2 EXPERIMENTAL SETUP	37
CHAPTER 4: PREDICTION MODELS	39
4.1 PERFORMANCE METRICS	39
4.1.1 R^2 AND $\overline{R^2}$	39
4.1.2 PEARSON'S CORRELATION	41
4.2 SELECTION OF INPUT AND TARGET VARIABLES	42
4.3 MODELLING TECHNIQUES	45
4.3.1 LINEAR REGRESSION	45
THEORY	45
PROGRESS SO FAR WITH LINEAR REGRESSION	47
4.3.2 NEURAL NETWORKS	48
THEORY	48
3-CLASS EVALUATION	49
THE CROSS-VALIDATION METHOD	51
GENERALIZATION	52
APPLICABILITY AND PROPERTIES ON QUALITY PREDICTION	54
4.3.3 CLASSIFICATION AND REGRESSION TREES	54
THEORY	54
PROGRESS SO FAR WITH REGRESSION TREES	55
PRUNING TECHNIQUE TOWARDS GENERALIZATION	56

KNOWN ISSUES WITH CLASSIFICATION AND REGRESSION TREES	57
4.3 MODEL EXPERIMENTAL SETTINGS	57
CHAPTER 5 – RESULTS AND STATISTIC ANALYSIS	59
5.1 ENHANCEMENTS AND OPTIMIZATION	59
5.1.1 CLASSIFICATION AND REGRESSION TREES - SELECTION OF THE ADEQUATE PRUNING TECHNIQUE	59
5.1.2 NEURAL NETWORKS - EXPERIMENT SETUP	63
5.2 OVERALL RESULTS	68
5.3 RESULTS PER APPROACH	83
5.3.1 LINEAR REGRESSION	83
5.3.2 REGRESSION TREES AND CLASSIFICATION TREES	85
5.3.3 NEURAL NETWORKS	92
5.4 INTERSYSTEM/INTERCONFIGURATION PREDICTION MODELS	107
CHAPTER 6 - CONCLUSIONS AND FUTURE WORK	114
6.1 OVERALL CONCLUSION	114
6.2 FUTURE PROSPECTS	115
BIBLIOGRAPHY	121
APPENDIXES	123
TABLES	123
CORRELATIONS – BORIS SYSTEM	123
CORRELATIONS – INSPIRE SYSTEM	125
QUESTIONNAIRES	128
BORIS (ENGLISH VERSION):	128
INSPIRE SYSTEM (GERMAN VERSION):	132

List of diagrams and figures

Diagram 1 – Instrumentally measured variables	23
Diagram 2 – Expert annotated variables	29
Figure 1: Bodden-Jekosch Scale with allocated concepts	19
Figure 2 - Structure of a neural network with 3 ‘logsig’ neurons as output.....	50
Figure 3: Example of pruned regression tree using ‘splitmin’	63
Figure 4: Example of classification tree	88
Figure 5: Classification Tree using no_turns, ca_no_ia, pa:co, wa_iso, uct., target Mean Questions B.	89
Figure 6: Classification Tree using Input3b ‘space’, wa_iso’, ‘verb’. / Target: ‘Use again’	91
Figure 7: Scatter plot from simulation using 4 input variables	96
Figure 8: Histogram from the target values (0 to 6 – Mean B)	97
Figure 9: Histogram from the prediction values (0 to 6) made by the neural network done with configuration 3.....	97
Figure 10: No. of neurons X average value of correlation	109
Figure 11: No. of neurons X difference between highest and lowest correlation.....	110
Figure 12: Classification tree (3 input variables #turns, #CA:IA, PA:CO to predict B0) that has the best accuracy(56,5%) on intersystem models.	113

List of Tables

Table 1: Interaction variables instrumentally measured during the experiments..	22
Table 2: Interaction variables collected by experts during the experiments.	24
Table 3: Interaction variables (errors) collected by experts during the experiments.....	31
Table 4: Example of Target Data, 3-class Question B0 (overall quality)	38
Table 5: Highest Pearson’s correlations between input and target variables from the BoRIS System.....	43
Table 6: Correlations from the variable “Weighted CA:IA” and the main target values.....	44
Table 7: Correlation between %UCT and UER	44
Table 8: Highest Pearson’s correlations between input and target variables from the INSPIRE System that will be used on the experiments.....	45
Table 9: Example of prediction accuracy per pruning percentage.	60
Table 10: Example of prediction accuracy with ‘best level’ pruning.....	60
Table 11: Example of prediction accuracy with pruning according to validation data.	61
Table 12: Example of prediction accuracy with pruning using a fixed number of levels.	61
Table 13: Comparison from prediction accuracy with the ‘catidx’ and ‘splitmin’ options.	62
Table 14: Example of how output values look like.....	64
Table 15: Example from output values from 7 cases using 3 ‘logsig’ neurons.....	64
Table 16: Example on test data.....	65
Table 17: Statistics from the dependency of the number of neurons for a neural network using #turns, ca:#ia, ts_ord and uct as input variables and Mean B as target value.	68
Table 18: Results from experiment 1	69
Table 19: Results from experiment 2 – linear regression.....	70
Table 20: Results from experiment 2 – regression trees	71
Table 21: Results from experiment 2 – classification trees.....	71
Tables 22a and 22b: Results from experiment 2 – neural networks.....	72

Tables 23a and 23b: Results from experiment 3 – all methods, input 9 – Question B0	73
Tables 24a and 24b: Results from experiment 3 – all methods, input 9 – Mean B Questions	74
Table 25: Correlations between the ‘error classification’ variables and the overall quality from INSPIRE config. 1	76
Tables 26a and 26b: Correlations between the input3b and ‘use again’ from INSPIRE config. 1 (see INSPIRE System questionnaire, chapter 8.2).	77
Tables 27a and 27b: Correlations between the input3 and ‘overall quality’ from INSPIRE config. 1	78
Tables 28a and 28b: Correlations between the input2 and ‘overall quality’ from INSPIRE config. 1	79
Tables 29a and 29b: Correlations between the input4 (‘Unprogressive state’, ‘verb’, ‘Space’, ‘repetition’) and ‘overall quality’ from INSPIRE config. 1	80
Tables 30a and 30b: Correlations between the input5 (‘%uct, ‘wa_iso’, ‘ts_ord’, ‘space’, ‘repetition’) and ‘overall quality’ from INSPIRE config. 1	81
Tables 31a and 31b: Correlations between the input8 (#turns, CA:#IA, PA:CO, WA_iso, WPST, UCT, IC, #Sys.Questions) and ‘overall quality’ from INSPIRE config. 1	82
Table 32: : Correlations between the input8 and ‘overall quality’ from INSPIRE config. 1, linear regression	84
Table 33: Correlation per user as he/she is left out as test data, Input 8, Inspire System 1	84
Table 34: Results from experiment 6 - classification trees	86
Table 35: Results from experiment 7 - classification trees	87
Table 36: Accuracy from classification tree on experiment 8, Leave-one-out technique, done on 23 users.	90
Table 37: Pearson’s correlation per user	92
Table 38: Pearson’s correlation for user 9	94
Table 39: Pearson’s correlation for user 27	94
Table 40: Pearson’s correlation for user 34	94
Table 41: Pearson’s correlation for user 20	95
Table 42: Pearson’s correlation for user 34	95

Table 43: Accuracy in relation to the amount of neurons in the hidden layer	98
Table 44: Statistics from #turns, ca:#ia, ts_ord and uct as input variables on a NN to predict Mean Questions B.....	99
Table 45: Statistics from #turns, ca:#ia, ts_ord, uct and age as input variables on a NN to predict Mean B.....	99
Table 46: Statistics from #turns, ca:#ia, ts_ord, uct and gender as input variables on a NN to predict ‘Mean B Questions’	100
Table 47: Statistics from #turns, ca:#ia, ts_ord, uct, age and gender as input variables on a NN to predict ‘Mean B Questions’	100
Table 48: Statistics from #turns, ca:#ia, ts_ord, uct and B1 as input variables on a NN to predict Mean B.....	100
Table 49: Statistics from experiment 14, weighted error analysis (PA:CO, WA_iso, IC to predict overall quality on the INSPIRE system config. 2.....	101
Table 50: Correlation between the 3 factors originated from input9	103
Table 51: Statistics from input 9 on a NN to predict ‘Mean B Questions’	103
Table 52: Statistics from #turns, ca:#ia, ts_ord and uct as input variables on a NN to predict ‘Mean B Questions’	104
Table 53: Statistics from 5 factors representing 52 input variables on a NN to predict overall quality	104
Table 54: Statistics from 5 factors representing 52 input variables on a linear regression to predict overall quality	105
Table 55: Correlations from 5 factors acquired from all “error coding” variables from the INSPIRE system.....	105
Table 56: Statistics from NN with input variables	106
Table 57: Statistics from Neural Networks with input variables: 3 factors from “error classifications”, target value: “easy learning”.....	106
Table 58: Statistics from NN with input variables: 2 factors from “error classifications”, target value: “easy learning”.....	107
Table 59: Statistics from NN with input variables: 1 factor from “error classifications”, target value: “easy learning”.....	107
Table 60: Statistics about intersystem/interconfiguration using linear regression (3 input variables: #turns, #CA:IA, PA:CO).....	108

Table 61: Statistics about intersystem/interconfiguration using neural networks	110
Table 62: Statistics about intersystem/interconfiguration using regression trees (3 input variables: #turns, #CA:IA, PA:CO)	111
Table 63: Statistics about intersystem/interconfiguration using classification trees.....	112
Table 64: Statistics with different validation techniques on NN's.....	116
Table 65: Statistics with different validation techniques on NNs with linear output. The function 'Trainbfg' was described on section 5.1.2.....	117
Table 66: Statistics with different validation techniques on NN's. The function 'Trainbr' was described on section 5.1.2.	117
Table 67: Statistics with different validation techniques and input variables on NN's.	118
Table 68: Statistics from Experiment to predict 'Mean Questions B', using 1 'purelin' output neuron (0 to 6).	118
Table 69: Statistics on linear regression with and without the new parameter (weighted consecutive CA:IA)..	119
Table 70: Accuracy comparison with and without the new parameter (weighted consecutive CA:IA) on classification trees.	120