

Intelligent and Proactive Approach for The Optimal Handling of Low Chatbot Quality of Services (CQoS)

by Ebtessam Hussain Almansor

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of Farookh Hussain

University of Technology Sydney
Faculty of Engineering and Information Technology

May 2022

Certificate of Original Authorship

I, Ebtesam Hussain Almansor declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature

Production Note:
Signature removed prior to publication.

May 2022

Acknowledgements

In the beginning, praise be to God, who helped me to complete the most important stage of my education, which is PhD degree. Secondly, thanks to my successful supervisor and ideal leader who helped me complete my PhD degree. Thank you for everything you have done, words are always unable to express thanks and expressions. Thank you to my dear parents who helped me to be in this place and get this degree. Thank you for your prayers and support.

I would also like to thank my husband who helped me in my absence and who supported me a lot. Thank you to my beloved children who were friends of my scientific journey. Thank you to all my friends for their support.

Finally, I would also like to express my thanks to Saudi Arabian Cultural Mission (SACM) in Australia and Najran University for supporting me during my PhD journey.

Abstract

Recently, the chatbot has evolved into a trending topic in the area of computer science. The rapid growth of intelligent chatbots as conversational agents with artificial intelligence has recently attracted much research attention. This significant increase in the use of chatbots across different domains, such as education, business, and health care, raises a problematic issue, this being the quality of the responses provided by the chatbot. Although most of the research studies attempted to build a chatbot that provides an intelligent response, in some cases, a chatbot might not understand the end-user's request, which leads to producing inappropriate utterances that cause a negative user experience and conversation breakdown.

While several studies focus on dialogue breakdown detection, they still face several challenges, such as the lack and bias of human annotation for the dataset. Also, when they detect a dialogue breakdown point, they do not provide a solution to handle the breakdown. In the current literature, there is no model to determine the quality of responses from a chatbot to make intelligent and proactive decisions to transfer the conversation from the chatbot to a live agent.

To tackle these challenges, in this thesis, we developed intelligent, automated, and data-driven approaches to address the aforementioned research issue of determining the chatbot quality of service (CQoS) and make proactive and intelligent decisions as to when to transfer the control of the conversation to a live agent. Various machine learning approaches are proposed to detect CQoS, including supervised and unsupervised

approaches. Also another key aspect is considered, which is the human thinking and reasoning using the fuzzy logic detection model. Importantly, the use of a sentiment score is introduced to trigger the breakdown without the need for annotated dataset. The proposed solutions are evaluated using real-time datasets. The key finding of our research was based on the evaluation process. We concluded that our proposed method for modeling CQoS outperforms other similar methods. Also, based on the evaluation process, the deep learning model was able to more accurately detect the need for handover mechanism compared with the other models.

Publications

1. Almansor, Ebtesam H., and Farookh Khadeer Hussain. "Survey on intelligent chatbots: State-of-the-art and future research directions." Conference on Complex, Intelligent, and Software Intensive Systems. Springer, Cham, 2019.
2. Almansor, E.H., Hussain, F.K. and Hussain, O.K., 2021. Supervised ensemble sentiment-based framework to measure chatbot quality of services. *Computing*, 103(3), pp.491-507.
3. Almansor, E.H. and Hussain, F.K., 2020, January. Modeling the Chatbot Quality of Services (CQoS) Using Word Embedding to Intelligently Detect Inappropriate Responses. In *International Conference on Advanced Information Networking and Applications*. Springer.
4. Almansor, E.H. and Hussain, F.K., 2021, July. Fuzzy Prediction Model to Measure Chatbot Quality of Service. In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-4). IEEE.
5. Almansor, E.H. and Hussain, F.K., 2021, May. Sentiment-Driven Breakdown Detection Model Using Contextual Embedding ElMo. In *International Conference on Advanced Information Networking and Applications* (pp. 163-171). Springer, Cham.

Abbreviations

CQoS	Chatbot Quality of Service
NLP	Natural language processing
DL	Deep learning
SA	Sentiment Analysis
ML	Machine Learning
AI	Artificial intelligent
CNN	Convolution Neural network
NN	Neural network
LSTM	Long Short-Term Memory
RNN	Recurrent Neural network
Char-CNN-Bi-LSTM	Character Convolutional Bi-directional- LSTM
AIML	Artificial Intelligence Markup Language
ALICE	Artificial Linguistic Internet Computer Entity
UIMA	Unstructured Information Management Architecture
LUSI	Language Understating Information Service
NLU	Natural Language Understanding
DST	Dialogue Stat Tracking
HMM/CFG	Hidden Markov model context-free grammars
CRFs	Conditional random fields
Seq2Seq	Sequence-to-sequence
NLG	Natural Language Generation
IRIS	Informal Response Interactive System

VSM	Vector Space Model
DBDC	Dialogue Breakdown Detection Challenge
MAP	Mean Average Precision
LSA	Latent Semantic Analysis
WWW	World Wide Web
MRR	Mean Reciprocal Rank
nDCG	Normalized Discounted Cumulative Gain
TAARNN	Topic-Aware Attentive Recurrent Neural Network
HRED	Hierarchical Recurrent Encoder-Decoder
RLM	Recurrent Neural Network Language Model
BLEU	Bilingual Evaluation Understudy
MMI	Maximum Mutual Information
ECM	Emotional Chatting Machine
MSE	Mean Square Error
BERT	Bidirectional Encoder Representations from Transformers
VADER	Valence Aware Dictionary and sEntiment Reasoner
TF-IDF	Term frequency-inverse document frequency
MELD	Multimodal Emotion Lines Dataset
CIC	Conversation Intelligent Challenge
DBDCs	Dialogue Breakdown Detection Challenges

Table of contents

Abstract	iii
Abbreviations	vi
List of figures	xv
List of tables	xvii
1 Introduction	1
1.1 Introduction	1
1.2 Research Challenges	3
1.3 Objectives of this Thesis	4
1.4 Significance of the Thesis	5
1.4.1 Scientific Contributions	5
1.4.2 Social Contributions	6
1.5 Structure of the Thesis	6
1.6 Conclusion	9
2 Literature review	10
2.1 Overview	10
2.2 Background of existing chatbots	11
2.2.1 Elizabot	11

2.2.2	Alicebot	12
2.2.3	Elizabeth bot	12
2.2.4	Mitsuku	13
2.2.5	IBM Watson	13
2.2.6	Microsoft LUIS	14
2.2.7	Google Dialogflow	14
2.2.8	Amazon Lex	15
2.3	Taxonomy of chatbot	15
2.3.1	Chatbot based on purpose	16
2.3.1.1	Task-oriented conversational agent	17
2.3.1.2	Non-task-oriented conversational agent (chatbot)	21
2.3.1.3	Retrieval response generation approach	21
2.3.1.4	Generation-based approach	28
2.3.2	Chatbot-dialogue based on response generation ability	33
2.3.2.1	Goal-based chatbot	33
2.3.2.2	Knowledge-based chatbot	33
2.3.2.3	Service-based chatbot	34
2.3.2.4	Response generated-based chatbot	34
2.4	Breakdown detection models in dialogue	34
2.4.1	Supervised models	35
2.4.2	Unsupervised approaches (Deep learning model)	36
2.4.3	Others related approach	37
2.5	Sentiment analysis in dialogue	47
2.6	Critical evaluation of the existing literature	48
2.7	Conclusion	49
3	Problem definition	50
3.1	Introduction	50
3.2	Key Terms and Concepts	50
3.2.1	The dialogue-based system	51

3.2.2	Natural language processing	51
3.2.3	Breakdown in dialogue	51
3.2.4	Chatbot Quality of Services (CQoS)	51
3.2.5	Utterance	51
3.2.6	Inappropriate or low utterance	51
3.2.7	Intelligent chatbot quality of service platform	52
3.2.8	Live agent	52
3.2.9	Administrator	52
3.2.10	Handover mechanism	52
3.2.11	Sentiment analysis	52
3.2.12	Machine learning	53
3.2.13	Supervised approach	53
3.2.14	Deep learning	53
3.2.15	Fuzzy logic	53
3.3	Problem Overview and Problem Definition	53
3.4	Research Questions	56
3.4.1	Research Question 1	56
3.4.2	Research Question 2	56
3.4.3	Research Question 3	56
3.4.4	Research Question 4	57
3.5	Research Objectives	57
3.5.1	Research Objective 1	57
3.5.2	Research Objective 2	57
3.5.3	Research Objective 3	57
3.5.4	Research Objective 4	58
3.6	The Research Approach to Problem-Solving	58
3.7	Existing Research Methods	58
3.7.1	Science and Engineering Approach	58
3.7.2	Social Science Approach	59

3.8	The Methodology of Science and Engineering-Based Research	60
3.8.1	Conceptual Level	62
3.8.1.1	Literature Review	62
3.8.1.2	Problem Formulation	62
3.8.1.3	Definition of Key Concepts	62
3.8.1.4	Conceptual Solution	62
3.8.2	Perceptual Level	63
3.8.2.1	Methodology Development	63
3.8.2.2	Development of the Prototype	63
3.8.3	Practical Level	63
3.9	Conclusion	63
4	Solution Overview	65
4.1	Introduction	65
4.2	Overview of the Solution for Measuring CQoS using an Ensemble Supervised Model	67
4.3	Overview of the Solution for Measuring CQoS Considering Human Thinking and Reasoning Using Fuzzy Logic	71
4.3.1	Preprocessing	72
4.3.2	Fuzzy logic model	72
4.3.2.1	Rule inference	73
4.3.3	Measuring CQoS	73
4.4	Overview of the Solution for Measuring CQoS Considering Automated Conversational Feature Extraction using Deep Learning	73
4.5	Overview of the Solution to Validate the Proposed Methodology	75
4.5.1	Evaluation metrics	75
4.5.1.1	Accuracy	75
4.5.1.2	Precision	76
4.5.1.3	Recall	76
4.5.1.4	F1-measure	77

4.6	Conclusion	77
5	Supervised Ensemble Sentiment-Based Framework to Measure Chatbot Quality of Services	78
5.1	Introduction	78
5.2	The proposed framework	79
5.2.1	Task formulation	79
5.2.2	Automated labeling	80
5.2.3	Pre-processing	81
5.2.4	Feature extraction	81
5.2.5	The supervised ensemble detection model	82
5.2.6	Decision-making system	83
5.3	Experimental Setting	84
5.3.1	Datasets	85
5.3.2	Baseline and State of the Art	86
5.3.3	Evaluation metrics	86
5.4	Results and discussion	86
5.4.1	Comparison with the state-of-the-art	87
5.4.2	Validation with other models and datasets	89
5.5	Conclusion and future work	91
6	Fuzzy-Based Monitoring System for Predicting the CQoS	92
6.1	Introduction	92
6.2	Proposed fuzzy-based framework for detecting and handling CQoS	93
6.2.1	Preprocessing	94
6.2.2	Fuzzy logic model	95
6.2.2.1	Fuzzification step	95
6.2.2.2	Rule inference	95
6.2.3	Measuring CQoS	97
6.3	Evaluation of the Prototype	97

6.3.1	Experiment setup and implementation	97
6.3.2	Datasets	97
6.3.3	Evaluation metrics	98
6.4	Results and discussion	98
6.5	Conclusion	100
7	Measuring Chatbot Quality of Service to Predict Human-Machine Handover using a Character Deep Learning Model	101
7.1	Introduction	101
7.2	The proposed hybrid Char-CNN-Bi-LSTM	102
7.2.1	Pre-processing	102
7.2.1.1	Pre-processing	102
7.2.1.2	Data labeling	102
7.2.2	Character Convolution Neural Network (char-CNN)	104
7.2.3	Bi-LSTM	105
7.3	Experiment setup and Evaluation criteria	106
7.3.1	Dataset	106
7.3.2	Evaluation metrics	107
7.3.3	Parameter setting	107
7.3.4	Baselines and state-of-the-art	108
7.4	Results and discussion	108
7.4.1	Comparison of the proposed method using different tasks	109
7.4.2	Comparison of the proposed method with state-of-the-art methods	110
7.5	Conclusion and future work	115
8	Conclusion and Future Work	116
8.1	Introduction	116
8.2	Problems addressed in this thesis	117
8.3	Contributions of the Thesis	118

8.3.1	Contribution 1: State-of-the-art comprehensive survey of the existing literature	118
8.3.2	Contribution 2: Develop a supervised ensemble sentiment based model to measure chatbot quality of services . .	119
8.3.3	Contribution 3: Develop a fuzzy logic model to measure chatbot quality of services	119
8.3.4	Contribution 4: Develop an ensemble deep learning model to measure chatbot quality of services	119
8.3.5	Evaluation of the proposed chatbot quality of service framework	120
8.4	Conclusion and future work	120
	Bibliography	123

List of figures

1.1	The structure of the thesis	8
2.1	Structure of the literature review	11
2.2	Taxonomy of chatbot	16
2.3	Illustration of the pipeline structure in the non-task-oriented approach	18
2.4	Classifying chatbots into two categories: retrieval-based chatbots and generation-based chatbots	21
2.5	The outline of the proposed classification of dialogue breakdown	35
3.1	Illustration of the proposed handover process in action	55
3.2	Overview of the science and engineering-based research methods applied in this thesis	61
4.1	Overview of the proposed detection solution and handling of low CQoS	66
4.2	Overview of the proposed intelligent framework to measure CQoS using the ensemble supervised model	67
4.3	Overview of the proposed ensemble supervised model	68
4.4	Overview of the solution for measuring CQoS considering human thinking and reasoning using fuzzy logic	71

4.5	Overview of the solution for measuring CQoS considering automated conversational feature extraction using deep learning	74
5.1	The proposed intelligent framework to measure CQoS	80
5.2	The proposed ensembles classifier	83
5.3	The Decision-making system	84
5.4	Evaluation results on DBDC3 with state-of the-art-model with EUS . .	88
5.5	Evaluation results on DBDC3 with state-of the-art-model with EUSS .	89
6.1	The structure of the proposed fuzzy rule-based framework	94
6.2	Comparison results of the proposed fuzzy model	99
7.1	Architecture of the proposed Char-CNN-Bi-LSTM approach for measuring Chatbot Quality of Services	103

List of tables

2.1	Overview of existing task-oriented approaches used for Chatbots	20
2.2	Overview of the existing retrieval approaches used for Chatbots	24
2.2	Overview of the existing retrieval approaches used for Chatbots	25
2.2	Overview of the existing retrieval approaches used for Chatbots	26
2.2	Overview of the existing retrieval approaches used for Chatbots	27
2.3	Overview of existing generating- approaches used in Chatbots	29
2.3	Overview of existing generating- approaches used in Chatbots	30
2.3	Overview of existing generating- approaches used in Chatbots	31
2.3	Overview of existing generating- approaches used in Chatbots	32
2.4	Comparative analysis of the existing studies of breakdown detection models.	38
2.4	Comparative analysis of the existing studies of breakdown detection models.	39
2.4	Comparative analysis of the existing studies of breakdown detection models.	40
2.4	Comparative analysis of the existing studies of breakdown detection models.	41
2.4	Comparative analysis of the existing studies of breakdown detection models.	42

2.4	Comparative analysis of the existing studies of breakdown detection models.	43
2.4	Comparative analysis of the existing studies of breakdown detection models.	44
2.4	Comparative analysis of the existing studies of breakdown detection models.	45
2.4	Comparative analysis of the existing studies of breakdown detection models.	46
5.1	Cornell dataset details	85
5.2	MELD dataset details	86
5.3	Evaluation results on the proposed model in DBDC3 Dataset	87
5.4	Evaluation results on the proposed model in DBDC3 Dataset	88
5.5	Comparison of the accuracy using our proposed framework on MELD and Cornell dataset	90
5.6	Comparison of F1-measure, recall, and precision on the Cornell	90
5.7	Comparison of F1-measure, recall, and precision score on MELD with the Automated sentiment	91
6.1	Input and output parameters	96
6.2	A fuzzy system using sentiment and word count for measuring CQoS	96
6.3	Comparison results of the proposed fuzzy model against the breakdown detection models on the DBDC3 dataset	99
6.4	Comparison results of the proposed fuzzy model against the supervised model	100
7.1	Comparison of various baseline models with Char CNNBi-LSTM on dialogue breakdown detection based on the original label	109
7.2	Comparison of various baseline models with Char CNNBi-LSTM on dialogue breakdown detection based on the sentiment score as a label	110

7.3	Comparison of accuracy of various baseline models with Char CNNBi-LSTM on measuring CQoS for DBDC3 and MELD	111
7.4	Comparison of various baseline models with Char CNNBi-LSTM using MELD, Joker and IRIS for Accuracy, F1 Score, Precision and Recall	112
7.5	Comparison of accuracy of various baseline models with Char CNNBi-LSTM on measuring CQoS using sentiment score	113
7.6	Comparison of various baseline models with Char CNNBi-LSTM using MELD, Joker, IRIS and Cornell with sentiment score using Accuracy, F1 Score, Precision and Recall	114