Deanship of Graduate Studies
Al-Quds University

Automatic Essays Scoring

Hamzeh Abdel Hamid Mujahed Mujahed

M.Sc. Thesis

Jerusalem-Palestine

1430 / 2009

Automatic Essays Scoring

Prepared By:
Hamzeh Abdel Hamid Mujahed Mujahed

B.Sc.: Computer Systems Engineering, 2002, Palestine Polytechnic University, Palestine

Supervisor: Dr. Labib Arafeh

A thesis submitted in Partial fulfillment of requirements for the degree of Master of Electronics and Computer Engineering/ Department of Electronics and Computer Engineering/ Faculty of Engineering/ Graduate Studies
Al-Quds University

1430 / 2009

Al-Quds University
Deanship of Graduate Studies
Electronics and Computer Engineering Department

Thesis Approval

Automatic Essays Scoring

Prepared By: Hamzeh Abdel hamid Mujahed
Registration No.: 20520140

Supervisor: Dr. Labib Arafeh

Master thesis submitted and accepted, Date:.....June, 20, 2009
The names and signatures of the examining committee members are as follows:

1. Head of Committee:.Dr. Labib Arafeh Signature:...................
2. Internal Examiner :.Dr. Amjad RATTROUT.... Signature:..............
3. External Examiner :Dr. Adi Anani. Signature:..............
4. External Examiner :Dr. Mahmoud SAHEB.. Signature:..............

Jerusalem - Palestine

1430 / 2009

Dedication

This thesis is dedicated to:

Words fail me to express my appreciation to my wife Rabab whose dedication, love and persistent confidence in me, has taken the load off my shoulder. I owe her for being unselfishly let her intelligence, passions, and ambitions collide with mine.

My dear three children, Taima', Basel and Saja have been very tolerant of three years worth of hours schooling to obtain this degree. They always encourage me through their love to improve myself and I greatly appreciate that.

My parent's who offered me unconditional love and support throughout the course of this thesis.

To Dr. Labib Arafeh, Dr. Hussien Jaddu and Dr. Ali Jamoos from Al-Quds University; you have all been extraordinary teachers, dedicating your valuable time and energy to the growth and development of your students.

Finally, this thesis is dedicated to all those who believe in the richness of learning.

January, 2009
Hamzeh Abdel Hamid Mujahed,

Declaration:

I certify that this thesis submitted for the degree of Master, is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed:........Hamzeh.........................

Hamzeh Abdel Hamid Mujahed Mujahed

Date:.........20, June, 2009...........................

Acknowledgement

There are a number of people without whom this thesis might not have been written, and to whom I am greatly indebted.

Dr. Labib Arafeh has been the ideal thesis supervisor. His sage advice, insightful criticisms, and patient encouragement aided the writing of this thesis in innumerable ways.

I gratefully acknowledge Professor Younis Amro (president of Al-Quds Open University) for his advice and support during my study.

I am grateful to Mr. Yousef Oriqat from Al-Quds Open University for his support and help.

I would like to show my special thanks to the staff of the Engineering College -Department of Electronic and Computer Engineering in Al-Quds University who taught me during my Master degree study.

I am grateful to Dr. Nadia Qawasmi, from Al-Quds Open University (QOU) and Mr. Mohammad Jabary from Palestine Polytechnic University(PPU) for their fruitful cooperation to support us with data to check the developed models. The valuable support and encouragement is so important. Your assistance went far beyond the call of duty, and I appreciate it immensely.

I would like to express my gratitude to all those who gave me the possibility to complete higher studies especially my colleagues at Al-Quds Open University.

Finally, I would like to thank everybody who was important to the successful realization of thesis, as well as expressing my apology that I could not mention personally one by one.

Abstract

In this study, an AES system has been developed. The idea behind our proposed AES is to grade the essays by identifying the main keywords in the essays and its synonyms that determined by the teacher, and processing these keywords using the modeling approach-based techniques including Fuzzy Logic, Clustering, and Nuero-Fuzzy. Three models have been developed; the first model is Multiple Input Single Output (MISO) Mamdani Model; the second model is MISO Sugeno with back propagation optimization technique and the last one is Sugeno subtractive clustering with hybrid optimization technique. These developed AES models are capable to identify up to 15 keywords, each of which has up to 4 synonyms. A 100-word history essay has been used to test the developed AES. A 1080-datasets have been generated using an automatic answers generator depending on 13 questions. Using the cross-validation method, the data has been splitted into 718 samples for training and 362 samples for testing. To check the adequacy of the models, we have used the correlation coefficient to measure the agreement between the theoretical (actual) and predicted marks. Two error measures have also been used to check the accuracy including; Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).

Different scoring dimensions (content, structure, syntax, etc.) commonly used in developing AES systems. In our developed models, we adopted content dimension in scoring the written essay. One of the notable gaps in AES based on content dimension is ignoring the negation's issue in the sentences and the order of the words in the sentence. In our models we have succeeded to address the problem of negation issue and to identify the necessary procedures to solve the problem of words order in the sentence by suggesting the development of language parser. The obtained results for the answers to some questions containing negation are promising and show high agreement between actual and predicted marks.

The obtained correlation coefficient between the theoretical (actual) and the predicted marks for the first model(Multiple Input Single Output (MISO) Mamdani Model) ranges between 0.887 and 0.9969 with an average value of 0.9863 for training and 0.9599 for the testing data. The MAPE values range between 0.017 and 0.1619 with an average value of 0.0854 for the training and 0.1189 for the testing data. The RMSE values range between 0.0378 and 0.2947 with an average value of 0.087 for training and 0.2073 for testing data.

The correlation coefficient obtained between the theoretical and predicted marks in the second model (MISO Sugeno with back propagation optimization technique) ranges between 0.9347 and 0.9966 with an average correlation of 0.9906 for training and 0.9675 for the testing data. The MAPE value ranges between 0.015 and 0.399 with an average value of 0.1039 for the training and 0.1071 for the testing data. The RMSE ranges between 0.043 and 0.3775 with an average value of 0.0809 for training and 0.2069 for testing data.

The third model is subtractive clustering Sugeno. The obtained correlation coefficient between the theoretical and predicted marks ranges between 0.9121 and 0.9977 with an average correlation of 0.9948 for training and 0.9712 for the testing data. The MAPE value ranges between 0.015 and 0.1516 with an average value of 0.0399 for the training and 0.1024 for the testing data. The RMSE ranges between 0.0328 and 0.2763 with an average value of 0.0526 for training and 0.1837 for testing. It is noted that the results obtained with subtractive clustering model are the best results since the correlation between predicted and

actual marks are the highest even though the number of Membership Function(MF) are fixed(seven MF for each input), the type of MF are triangular one and the number of rules is less than other two models. However, other researchers have obtained correlation between actual and predicted marks ranges between 0.87 (i.e. Project Essay Grader PEG) to 0.98 (i.e. Modeling Techniques Applied to Short Essay Auto-grading Problem), although the sample essays under testing and other factor (modeling techniques used, number of samples, scoring dimension used)are not similar. The results show that we can adopt these models for AES purposes with high correlation between actual and predicted marks when we tested it using the online system. Thus, we may conclude that the preliminary and promising results demonstrate the suitability, adequacy and competitive of using the modeling techniques to solve the automated essay scoring problem.

Further more, a powerful online-supported graphical user interface system(Fuzzy Automatic Essay Scoring System (FAESS)) have been developed to allow the user to feed the system with keywords and synonyms to score the essays in a more usable and flexible way. The interface is simple to use and has the ability to score the essays in normal mode and fuzzy based scoring mode. Stand alone application was built to make the system easy and simple to use for testing by different teachers. We have examined the stand alone application using the unseen data obtained from the local universities and the results are promising.

For further investigation, we have established a contact with other local universities (Al-Quds Open University (QOU) and Palestine Polytechnic University (PPU)) to support us with data. From QOU we obtained a sample of data (18 samples) related to one essay-type question in subjects other than history such as English literature. When this data was tested using FAESS and subject to the developed models (MISO Grid partition Sugeno), the calculated correlation coefficient value between theoretical and predicted marks in QOU data was 0.9954, the value for MAPE is 0.1292 and the value of RMSE equal to 0.348. We have been also in contact with PPU. From PPU we obtained answers for two questions related to information technology with 11 sample of answers for the first question and 23 samples for the second one. The obtained correlation value between actual and predicted marks for the first sample was 0.9755. whereas the a correlation value for the second sample equal to 0.9745 with error measures MAPE equal to 0.0696 and RMSE equal to 0.266. The obtained results from unseen data from local universities support us with promising results since we had a high correlation coefficient between actual and predicted marks and small values in the error measures (MAPE and RMSE). We may notice that the differences between the actual (human) score and the predicted one range between 0% and 8.3%. That is in marks, it ranges between 0 to 2.5 marks out of the total 30 marks. Although, the number of these samples are not so large (23, 11, and 18), we may conclude that this approach is adequate and suitable to address and solve the AES problem.

Further investigation is still required with more samples of data (essays) and of different types of field other than historical questions. Further investigation required also to identify and to implement the required procedures to solve the problem of words order in the sentence using language parser.

Table of contents

List of Tables

List of Figures

List of Appendices

Chapter One

Introduction

1.1 Introduction

Essays scoring fall into two categories: short answers and open ended essays answers. Short answer essays compromised of few words to few lines mainly related to specific topic and mainly evaluated for their content dimension. For example define questions. Open ended answers essays constitute of several lines or paragraphs and several writing dimensions (i.e. content, style, vocabulary, rhetorical structure and syntax /grammar) must be combined to produce an effective scoring. Educational institutions have strong interest in the development of scoring methods and constantly exploring new techniques to effectively score essays (Oriqat, 2007).

Automated Essay Scoring (AES) has been a topic of research for over four decades. A limitation of all past work is that the essays have to be in computer readable form (Srihari, et al, 2007). Automated scoring of exams consisting of written text would be doubtless of advantage to teachers. However, recent advances in computer techniques have opened up the possibility of being able to automate the scoring of free text responses typed into a computer without having to create systems that fully understand the answers (Pulman and Sukkarieh, 2005).

When large numbers of answers are submitted to be scored, teachers find themselves bogged down in their attempt to provide consistent evaluations and high quality feedback to students within as short a timeframe as is reasonable, usually a matter of days rather than weeks. Educational administrators are also concerned with quality and timely feedback, but in addition must manage the cost of doing this work. Clearly an automated system would be a highly desirable addition to the educational tool-kit, particularly if it can provide less costly and more effective outcome (Palmer, et al, 2002).

A number of studies have been conducted to assess the accuracy (measurement of the degree of agreement between actual marks and predicted marks) of the AES systems with respect to writing assessment. The results of several AES studies reported high agreement rates between AES systems and human raters. AES systems have been criticized for lacking human interaction, and their need for a large corpus of sample text to train the system. Despite its weaknesses, AES continues attracting the attention of public schools, universities, testing companies, researchers and educators (Dikli, 2006).