Al Quds University

Deanship of Graduate Studies

Electronics and Computer Engineering Dept.

Thesis Approval

Modeling Techniques Applied to Short Essay Auto-Grading Problem

By: Yousef Mohammad Mousa Oriqat

Reg. No. 20410864

Supervisor: Dr. Labib Arafeh

Master thesis submitted and accepted, Date:..............................

The names and signatures of the examining committee members as follows:

1-Head of Committee:..Dr. Labib. Arafeh. Signature:..........

2-Internal Examiner Dr. Rashid. Jayousi. Signature:..R. Jayousi

3-External Examiner :. Dr. Mohammad. SAHEB Signature:..........

Jerusalem – Palestine

1428 (Hijri) / 2007

Abstract

The various techniques of Automated Essay Grading (AEG) systems that are currently available have been reviewed. They include; Latent Semantic Analysis based on Singular Value Decomposition, Vector Space Models and Natural Language Processing techniques to extract features from the written prose. The features they extract are based on specific written dimensions including; contents, style, vocabulary, syntax and grammar, etc. Short answers essays are mainly evaluated for their content dimensions.

In this thesis, we have concentrated on developing a system that tries to mimic the way we, as human being think, reason, evaluate, grade, etc. Several approaches have been adopted to solve the short essay AEG problem by applying Fuzzy based techniques including integrated and adaptive Neuro-Fuzzy approaches.

The main idea behind our approach is to identify the number of main keywords (5 input), each of which has 4 synonyms. These inputs have been processed by developing four general models including; Multiple Input Single Output (MISO) Mamdani FIS using rapid prototype design, MISO Sugeno using Adaptive Neuro FIS based on Hybrid optimization techniques, ANN back-propagation optimization technique, and Subtractive Clustering model. A fifth general model have been developed to incorporate more main keywords by processing double (10 inputs) the number of previous models inputs, single output with specific constraints.

We have constructed 900 datasets based on ten questions and splitted them into training and testing sets using cross-validation methods. We have shared our datasets with other

systems including; (Institute for Information systems and Computer Media (IICM), Graz University of Technology, Austria) to compare our obtained results with other systems. Furthermore, we have generated 300 datasets for and additional question (Q11) with ten inputs.

To check the adequacy of the models, we have used the correlation coefficient to measure the agreements between the actual and predicted marks of our developed models, and used two error measures including; Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) to indicate the performance of the developed models.

The highest results obtained among the models that where developed for the five main keywords were from the Hybrid technique General model with an average correlation value of 0.9992 for the ten questions. While MAPE average values 0.0024, where as RMSE average values of 0.0109. To further check the suitability of our approaches we have used the 10 inputs datasets (Q11) to train and test our fifth general model and the promising results obtained shows an average correlation of 0.9998 with MAPE value of 0.001 and RMSE value of 0.003. In addition the datasets received from IICM shows a correlation of 0.9913 and MAPE 0.084, and RMSE 0.083.

The fifth general model results show a correlation coefficient of 0.9999, which means it can handle longer answers with high correlation between actual and predicted marks. The average correlation results obtained for all five general models are 0.98. This indicates the suitability and adequacy of our developed models to solve the short essay with different number of inputs AEG problem and shows that our preliminary results are promising.

ملخص

قام الباحث بمراجعة بعض الأنظمة الحالية المتعلقة بالتصحيح التلقائي للأسئلة الإنشائية. سيما وأن تحليل النصوص الانشائية يعتمد على التحليل النحوَي، والبلاغي، والتركيبي، و المحتوى الموضوعي إضافة إلى تحليل المرادفات اللغوية. حيث تعتمد الأنظمة التي تقوم بالتصحيح التلقائي للإجابات القصيرة على تحليل المرادفات اللغوية. ولقد تم تطوير العديد من الطرق لحل مشكلة التصحيح التلقائي منها: تحليل المعاني المخفي (Latent Semantic Analysis) بطريقة تحليل القيمة المفردة (Singular Value Decomposition)، تحليل المتجهات (Vector Space Model) و طرق معالجات اللغات الطبيعية (Natural Language Processing).

وهَدَفَ البحث الى إيجاد طرقاً ونماذج تستطيع محاكاة الإنسان من حيث التفكير، والمنطق، والاستنباط ، والتقييم، الخ. لذلك تم استخدام طرق المنطق الغامض، والاستنباط المنطقي الوصفي من خلال الأنظمة المتكاملة والمتكيفة والتي تعتمد على طرق التعلم في الشبكات العصبية الصناعية. وتتمحور الفكرة الرئيسة في هذه الطريقة على تبيان عدد الكلمات الرئيسية (5 مدخلات) و أربع مرادفات لكل مدخل. وتم معالجة المُدخلات باستحداث وتطوير أربعة أنظمة تستطيع التصحيح التلقائي للإجابات الانشائية القصيرة كما يلي: "مامداني" (Mamdani) متعدد المُدخلات مفرد المخرجات باستخدام تصاميم النماذج الأولية السريعة (Rapid prototype design)، "سوجينو" (Sugeno) متعدد المدخلات مفرد المخرجات باستخدام التقنيات المركبة او المهجنة (Hybrid techniques)، الشبكات العصبية الاصطناعية باستخدام تقنية التغذية الراجعة (ANN back-propagation)، تقدير الأنظمة باستخدام تقنية التكتل الطرحي (Subtractive Clustering)، ونموذج خامس عام يقوم على طريقتين لمعالجة عدد اكبر من المدخلات.

تمَ جمع أكثر من 900 إجابة قصيرة لعشرة اسئلة في مجال الحاسوب، وتم تقسيم البيانات إلى قسمين،

قسم لاستخدامها في تعليم الأنظمة (Learning) وقسم آخر لفحص النتائج (Testing) باستخدام طريقة (Cross-Validation). حيث تم مشاركة جامعة غراز النمساوية ببعض البيانات من الأسئلة والإجابات التي تم جمعها، وأيضا الحصول على بعض الاسئلة والإجابات من جامعة غراز النمساوية.

للتأكد من مدى ملائمة وصلاحية الأنظمة الخاصة بالباحث، تم انجاز نظام عام لكل نموذج استحدث وتم فحص النتائج باستخدام معامل درجة التوافق (Correlation coefficient) لقياس مدى تطابق العلامات الحقيقية مع المستنبطة أو المتنبئة (Predicted) ، كما تم استخدام طريقتين لحساب الخطأ هما متوسط نسبة القيمة المطلقة "ماب" (MAPE, Mean Absolute Percentage Error)، ومتوسط مربع الجذر "ارمس" (RMSE, Root Mean Square Error).

وأشارت النتائج الأولية إلى أن نظام التقنية المركبة (Hybrid) حصل على أعلى مستوى مقارنة مع الأنظمة الأخرى ذات الخمس مُدخلات التي تم تطويرها حيث كانت متوسطات قيم درجة التوافق 0.9992 لكافة الأسئلة العشرة. كما أن متوسط قيم "ماب" 0.0024، ومتوسط قيم "ارمس" 0.0109. وللتأكد من الإمكانيات المتوفرة للأنظمة المطورة تم فحص النظام الخامس العام بمضاعفة عدد المُدخلات، حيث كانت متوسطات قيم التوافق 0.9998 ومتوسط قيم "ماب " 0.001 ، و "ارمس" 0.003 . ثم تم استخدام البيانات المزودة من جامعة غراز النمساوية على نفس النظام الخامس وكانت نتيجة قيمة التوافق 0.9913، وقيمة "ماب" 0.084، وقيمة "ارمس" 0.083 .

للتأكد من مدى فاعلية هذه الطرق لحل مشكلة التصحيح التلقائي للأسئلة الإنشائية القصيرة تم استخدام النموذج الخامس حيث كانت قيم التوافق 0.9999. وبلغ معدل قيم التوافق لجميع الأنظمة التي تم تطويرها 0.98 . لذلك، فأن النتائج الأولية الواعدة تشير إلى مقدرة الأنظمة التي تم تطويرها باستخدام عدد متغير من المدخلات على حل مشكلة التصحيح التلقائي للأسئلة الإنشائية القصيرة.

Contents

CHAPTER 1

INTRODUCTION

1.1 Introduction

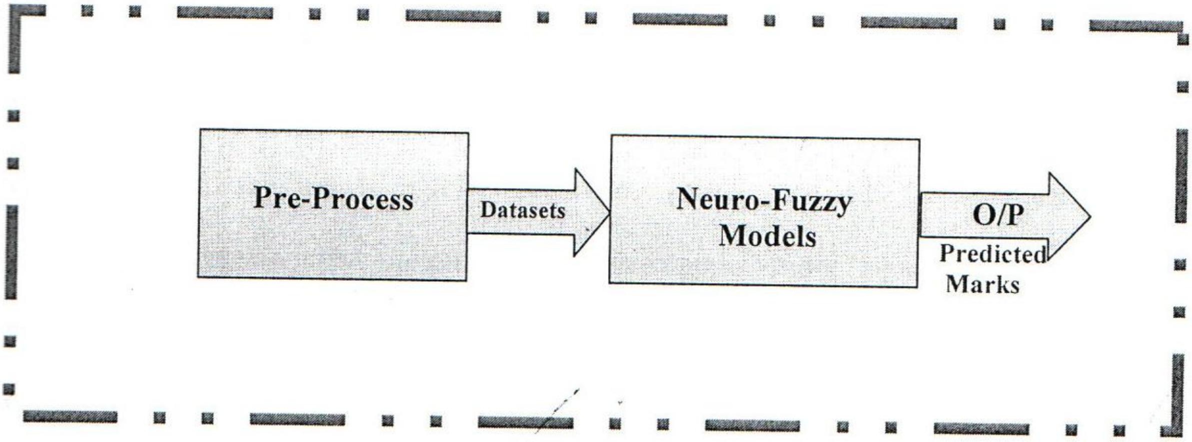

Fig. 1.1 Our Short Answers AEG system block diagram

E-learning systems now days are used by most universities, colleges and other institutions to support face to face training in the classroom or to implement distance learning programs. The growth of e-learning systems has increased greatly in recent years due to the demand by students for more flexible learning options and economic pressures on educational institutions, which see technology as a cost saving measure. One of the major aspects for developing e-learning systems is how to assess student knowledge based on essay type answers. For example, the Moodle (Modular Object-Oriented Dynamic Learning Environment) (Moodle, 2004) is a course management system (CMS), and sometimes also called Learning Management Systems (LMS) have an assessment module to Auto Generate Questions (AGQ) and also to Auto Grade (AG) student answers. The AGQ module based on predetermined set of questions that the assessment module chooses randomly and does not support essay type questions. It has various types of questions, which include:-

(1) Multiple choice questions

(2) Fill in the blank questions

(3) Match the right answers

(4) Specific abbreviation definitions questions

Essays fall into two categories: short answers and open ended essays. Short answers essays compromised of few words to few lines mainly related to a specific topic and mainly evaluated for their content dimensions. For example: Define ? questions. Open ended essays constitute of several lines or paragraphs and several writing dimensions must be combined to produce an effective grading. Educational institutions have strong interest in the developments of grading methods and constantly exploring new techniques to effectively grade essays.

Our Automated Essay Grading (AEG) (Fig. 1.1) systems have Pre-process stage to prepare the questions, reference answers and process the datasets. The second stage is the development of various integrated and adaptive Neuro-Fuzzy models. The output will be the predicted marks.

In referring to a previous thesis work under the supervision of Dr. Labib Arafeh done on AEG at Al-Quds university (Al-Amayreh, 2006) the work concentrated on using fuzzy logic based on short question and developed an algorithm called **FLASA** "**F**uzzy **L**ogic in **A**utomarking **S**hort text **A**nswers" .

In this thesis, we have explored the various modeling techniques approaches to automatically evaluate short answers essays; our work have focussed on building the models and examining these techniques aiming at producing a very high measure of accuracy between the predicted and actual grading. The modeling techniques then could be integrated within a stand alone auto-grading application or online through the web. More than 900 datasets have been constructed, some datasets are shared with IICM and also we received datasets from IICM to test it on our developed models.

AEG is a developing technology and are used to overcome time, cost, and standardization issues in writing assessment. The search for excellence in machine grading of essays is continuing and numerous studies are being conducted to improve the effectiveness of the AEG systems. AEG is defined as the computer technology that evaluates and scores the written prose (Page, 1994).

AEG systems can be very useful because they can provide the student with a score as well as feedback within seconds (Page, 2003). The reason for developing AEG systems is not only to provide students with opportunities to practice writing, but also to provide them with quick and accurate feedback regarding grammatical errors, style, content, and organization (Burstein et al. 2003) .

1.2 Motivation

Modeling techniques based on integrated and adaptive Neuro - Fuzzy approaches have produced outstanding results when used to solve real world problems in many fields like control, decision support, system identification, etc. (Jang ,1992; Jang and Roger, 1993; Karayiannis, 1996; Babuska, 1997; Guillaume, 2001). Researchers haven't yet fully explored the implementation of Neuro- Fuzzy to solve the short answers auto-grading problem. Furthermore neither published an extensive work on applying modeling techniques as part of the auto-grading problem. Nor mentioned the potential of Fuzzy based techniques on solving the content dimension of AEG problem. Mainly their work concentrated on using Natural Language Processing (NLP) techniques based on extracting features from he written prose. Therefore we have aimed at exploring the use of FIS models to solve the AEG problem. Thus we have adopted the Neuro-Fuzzy approaches to try to solve the short essay AEG problem.

1.3 Problem definition

AEG can be used in several ways in a given assessment process:

(1) As a sole measure of writing ability.

(2) As a joint measure along with human rating, namely weighted with human rating in order to produce an average score.

(3) As a monitoring or quality control procedure over the human rating process.

The use of Computer-generated scores contributes significantly to the standardization and accuracy of the test scores. This is always a highly desirable feature, particularly in high – stakes testing. Most AEG procedures were originally developed for the purpose of assessing writing skills by trying to mimic human readers as possible. Though few procedures were developed with a keen focus on specific writing characteristics, most AEG procedures attempt to produce score that at least correspond to common writing dimensions.

Table 1.1 (Cohen, 2003) presents five scoring dimensions and features commonly used in assessment of writing tasks. Each of the writing features associated with each dimension can be further broken down into more specific features that may later be translated into quantifiable measures. The tabulated dimensions present the highly complicated problem of making the computer (Auto Grade) assess based on human knowledge, background,

Chapter 7

Conclusions and further work

Automated Essay Grading is a complicated problem and an ongoing research and development. Many systems claimed that they have solved this problem yet they constantly developing their techniques and approaches. The importance of AEG systems has been recognized world wide by educational institutions of various levels and by e-learning system that need to further its assessment from multiple choice, etc., to assess student online on short or open ended questions. Three major components that constitute the AEG systems; the data corpus, the technique used to auto-grade and the software interface. We successfully applied the FIS, Neural learning and hybrid of both, techniques to solve the real world problem of Auto-Grading short essays. In the next section, we will conclude other system approaches and our findings from the developed modeling techniques to solve the Automated Essay Grading problem. Followed by suggested further work to further investigate this problem.

7.1 Conclusions

In general, AEG systems developed techniques to extract features from student answers to effectively Auto-grade the written prose. Whether the extracted features are surface or deep structure they develop techniques that could mimic the way as human thinking. Human thinking are based on knowledge, background, experience and intelligence that make us associate thinks (words) with each other; sense that make us reason things (word) with specific writing dimension.

Knowledge of specific subjects is based on learned information and the way we can communicate that information to best describe our answers for specific question. These facts about human learnability and accusation of information were widely noticed by most researchers of AEG. To this extent AEG systems and success stories were based on the availability of data corpus, like for example C-rater and their success stories with GRE/GMAT data available from the same producers of C-raters (ETS). Each AEG system have their own data corpus developed by their experts on specific field and live answers collected from students. The data corpuses need to be manipulated to fit the developed AEG systems so their approaches can incorporate the data by developed techniques to understand this data. Back to the learnability fact, part of this data will be used as training sets for the development and implementation of their techniques.

The development of such systems takes few years of development and commercial deployment. The importance of accurately mark student essays embarked the researches to constantly develop their techniques. Some AEG systems use LSA based on SVD techniques to extract features from the written prose, some use VSM, others use NLP, etc..

In our review of literature survey of AEG approaches we listed the main features for fifteen current AEG software systems as produced by (Ros'e et al, 2003), and focused our review on four main AEG systems that are well known for their achievements and establishments of research and developments in the field of AEG. Most systems are claimed to be very accurate and the agreement between each system and expert human raters are found to be high. Among those systems we can include that:

- PEG is considered the best choice for evaluating writing style.

- IEA and E-rater are better for grading contents.
- IEA doesn't perform any NLP processing.
- E-rater is the most complex system.
- Bayesian network combine the approaches of PEG, LSA and E-rater
- Main weakness of all these system is the lack of a very large corpus of essays that may become a reference for everyone interested in AEG, because most of these systems performance are strongly affected by the reference texts.

In fact, just combining the main advantages of each AEG systems, the resulting one could be considered the best choice.

Short answers essays mainly evaluated based on its content dimensions, yet several international institutions that are known for their advancement and achievements in the field of essay's auto grading system like ETS and IICM researching for new methods on improving their developed techniques of assessing short essays. IICM researching focused on developed platform like GATE answer assessment and Tomcat web application systems, both systems are Java platform, (Gutl, 2007).

AEG systems implemented their developed approaches within stand alone application or within e-learning systems recognizing the importance of assessing student based on short or open ended essays based on the learned subjects. Some examples are GATE, Tomcat web application and E-rater2, also some auto generate assessment questions.

In this thesis, our work concentrated on developing a system that tries to mimic the way we, as human kind think, reason, evaluate, grade, etc. Therefore we explored the integrated and adaptive Neuro-Fuzzy approaches. The use of FIS and Neural learning have been effective in solving real world problems and used within application such as control (Robotics, Automation, etc.), Information systems (DBMS, Information Retrieval) and decision support.

Our AEG system was based on input, process and output. The inputs is the assessed subject , that is related to information technology. The output will be the predicted marks. The process was divided into pro-process system and the developed modeling techniques. The pre-process system were used to produce the datasets, we have constructed 900 sets of five input main keywords with specific constraints and due to the successful and feasible results obtained we extended our datasets by doubling the no. of inputs to ten main keywords and further work was done on implementing IICM datasets within our system, we used IICM datasets for checking the adequacy of our developed models. We used FIS and Neural learning approaches to develop our five models. Each model use different integrated and adaptive fuzzy inference systems with hybrid and back-propagation optimization techniques. We started by developing MISO Mamdani FIS model and manually trained the system using rapid prototype design. For the second model, we develop an MISO Sugeno FIS model by using ANFIS based on Hybrid optimization technique. Third model was developed using back-propagation of feed forward neural learning network. Fourth model was developed using system identification techniques using subtractive clustering. The fifth model developed is the general model that we develop to incorporate various no. of input with specific constraints.

Comparison between our models using average results of correlation, RMSE and MAPE shows that using Hybrid technique produced the best results and similar results were drawn from the general model. Furthermore, we compared our findings with other systems and other systems results. The preliminary results obtained by the developed models are promising and shows the feasibility and adequacy of using our developed models to solve this kind of problem.

7.2 Further work

Further work suggesting includes the following:

1- Integrate our system (developed models) within e-learning systems.
2- Design the interface for stand alone application or on line using the web.
3- Designs a pre-process system to auto generate essay type questions.
4- Integrate our models within ready developed systems such as GATE.
5- Comparisons with other techniques and systems using same datasets.
6- Construct more expert and live data to examine our approach.
7- Modify the type of questions to handle longer questions and open ended questions
8- Release the constraints of the system.
9- Automate the pre-process system.
10- Further testing of these models with different types of questions from different fields.