The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational & Social Sciences (EPESS), 2014

Volume 1, Pages 232-236

ICEMST 2014: International Conference on Education in Mathematics, Science & Technology

AN AUTOMATED SCORING APPROACH FOR ESSAY QUESTIONS

Ahmed Alzahrani
University of Essex

Abdulkareem Alzahrani
University of Essex

Fawaz Alarfaj
University of Essex

Khalid Almohammadi
University of Essex

Malek Alrashidi
University of Essex

ABSTRACT: The automated scoring or evaluation for written student responses have been, and are still a highly interesting topic for both education and natural language processing, NLP, researchers alike. With the obvious motivation of the difficulties teachers face when marking or correcting open essay questions; the development of automatic scoring methods have recently received much attention. In this paper, we developed and compared number of NLP techniques that accomplish this task. The baseline for this study is based on a vector space model, VSM. Where after normalisation, the baseline-system represents each essay by a vector, and subsequently calculates its score using the cosine similarity between it and the vector of the model answer. This baseline is then compared with the improved model, which takes the document structure into account. To evaluate our system, we used real essays that submitted for computer science course. Each essay was independently scored by two teachers, which we used as our gold standard. The systems' scoring was then compared to both teachers. A high emphasis was added to the evaluation when the two human assessors are in agreement. The systems' results show a high and promising performance.

Keywords: Automated essay scoring, project essay grade, e-pedagogy and e-assessment.

INTRODUCTION

The essays examinations are basically considered as an indispensable key in the educational processes. It helps instructors to know students achievements and their situations during the learning journey. Even more, they are considered as a measurement of the learner's ability to memorise, organise, analyse, and write thoughts focusing on specific desirable goals. In perspective aspects, the essay examination advantageously suits the small number of candidates as this gradually decreases when the number becomes larger. Furthermore, it eliminates the candidate's guessing, since it relies on his free answer rather than selecting the good answer such as in multiple choices tests.

However, when there is a vast number of examinations that need to be assessed at once, the instructor finds himself overwhelmed to provide high quality feedback to educators within as short a period of time as is reasonable. Furthermore, different instructors of such a module can have various feedback scores for one candidate, which is one of the essays examinations drawbacks commonly known as subjectivity (Nitko, 1996). As a result, the advancement in technological systems, especially natural language processing (NLP), is increasingly flourishing into to reduce effort, time, and cost of institution resources, for example the Intelligent Essay Assessor (IEA) (Landauer, 2003), which uses Latent Semantic Analysis (LSA) to extract semantic similarity of words and passages from text. However, this kind of system, based on (LSA) tends more towards the frequency of terms rather than understanding the meaning of human knowledge (language).

To address this issue, an automated scoring system based on vector space models (VSMs) is applied in this

- Selection and peer-review under responsibility of the Organizing Committee of the conference

*Corresponding author: Ahmed Alzahrani- e-mail: araalz@essex.ac.uk

paper. Using this model we address the issues of the aforementioned approaches and attempt to exploit the different NLP techniques to come up with an optimal solution.
In the remainder of this paper we demonstrate our system development, in five parts. The first part describes related work. In the second section, a methodology is described. The results and findings are presented in the third section. Section four has been dedicated to drawing conclusion following some recommendations for the future work.

LITERATURE REVIEW

Automated Essay Scoring (AES) is defined as the computer technology that enables us to evaluate and score the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). The aim of using AES is to tackle the issues related in writing assessments, such as: time, cost, generalisability and reliability (Burstein, 2003; Chung & O'Neil, 1997; Hamp-Lyons, 2001; Ellis Batten Page, 2003; Rudner & Gagne, 2001; Rudner & Liang, 2002; Sireci & Rizavi, 2000). The advantages of using AES have been attracting public schools, universities and researchers (Burstein et al., 1998; Shermis & Burstein, 2003; Sireci & Rizavi, 2000). Some of these advantages are: relieve the grading burden from the educators and adding a consistent level that unachievable sometimes by educators (Shermis & Barrera, 2002).
Nowadays, there are more than 12 programs in the Project Essay Grade (PEG) and Automated Essay Scoring field. These projects have been influenced by Page work(1966) (Williamson, 2009). In addition, they focus as much on assessing the essays' semantic relevance to a given prompt as on assessing the quality of the essay itself. Some of the popular programs are; the Educational Testing Services (ETS) e-Rater (Attali & Burstein, 2006), PearsonKTs KAT Engine, Intelligent Essay Assessor (IEA) (Landauer, Laham, & Foltz, 2003) and Vantage Learnings Intellimetric (Elliot, 2003).
The e-Rater engine marks writing essays by extracting a set of features representing important aspects of writing quality from each essay. It is based on a regression-based methodology, which is a number of properties derived from natural language processing (NLP). When the regression weights are determined for those properties, they can be employed to more essays to turn a predicted score out based on the calibrated feature weights. The on-going version of e-Rater system uses 10 such regression properties, with eight representing factors of writing quality and two representing content. A set of such sub-features computed from NLP techniques composes of primary scoring features (Attali & Burstein, 2006; Burstein, 2003).

PearsonKT's KAT Engine is a similar application which uses a Latent Semantic Analysis technique (Landauer, Foltz, & Laham, 1998). This technique uses a dimensionality-reduction method based on singular-value decomposition. This method represents the content of each essay as a vector. They use direction and length as two aspects of the computed vector which define the location of the essay in this multi-dimensional space. Then, the content score is calculated as a weighted sum, after normalisation and regression. The essay victor is compared to pre-scored essays by the angle separating. They select the closest pre-scored essays in angle, then take their average human scores, and then the cosine distance from the candidate essay is calculated to produce the direction value. The content score is the combination of the weighted sum with a vector length value of the essay. Finally, they combined this content score with other linguistic measures such as style and mechanics (e.g. spelling) features to produce the resultant essay score (Landauer et al., 2003).
Vantage Learning's IntelliMetric is used to evaluate students' essays as part of intelligent tutoring system (Kukich, 2000). As a consequence of a wide range of applications that are used for automated assessment to essays, these systems concentrate on assessing the semantic relevance or topicality of essays. The IntelliMetric is a scoring engine that assesses the skills of student's essays. The aim of this system is to mark the students writing based on the state achievement examinations. This approach, as other systems, uses a substantial number of computer-produced features, which are designed to represent different aspects of writing quality. Moreover, they use five main classes to aggregate these features. The classes are as follows: Focus/Coherence, Organisation, Elaboration/Development, Sentence Structure, and Mechanics/Conventions. Finally, the multiple concurrent statistical methods are used for aggregating the features in order to give the final score (Elliot, 2003; Learning, 2003; Rudner, Garcia, & Welch, 2006).
Similarity within the text is a fundamental and essential research topic within the processing of the natural language as well as the similarity measure of the variation of physical units. Many researchers have conducted the evaluation of the students' essays upon the matching between the model's and optimal answer and the candidate's answer. It is essential to emphasise the simplicity of the vector space model (VSM), yet effective techniques are needed to determine the similarities between documents or utterances. Such method has been widely used within the educational testing field .VSM technique has been applied by Attali & Burstein (2006)

for the purpose of measuring English writers that are non-native in terms of the choice of vocabularies. This technique, which is used during the students' essay, are scored through determining the relationship between the words that exist within a student's answer with the words that are contained within sample essays originating from individual scoring categories. A theory of this method is that outstanding essays would have higher similarity in the choice of words being used. This is especially true when two VSM-derived characteristics were utilised that includes the maximum cosine similarity as well as the cosine similarity related to the top scoring category. Furthermore, this technique has been used by Higgins et al. (2006) in order to discover the off-topic essays by students, though comparing the word based IV originating from an essay to an RV built from a series of essays that are on-topic. The essay is considered to be off-topic if the difference is more significant compared with a pre-defined threshold. Zechner and Xi (2008) also used VSM to assess whether content was relevant or not when marking work by non-native English speakers, while Xie et al. (2012) examined the viability of VSM methods for automated speech grading. They used a more advanced ASR than Zechner and Xi, and determined that the VSM results correlated quite highly with human scores. This study aims to expand on the scope of existing studies by employing the cosine similarity

METHODS

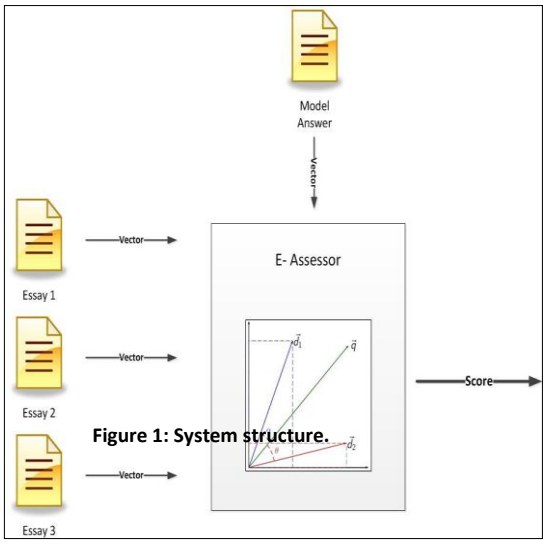

Figure 1: System structure.

Firstly, we defined the language based on the model answer and created a vector for each subject area. Subsequently these vectors are compared with each student submission to get the automated result (Figure 1).

Before creating the model answer vector we run the standard NLP normalisation techniques including removing stop-word (Figure 2).

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Figure 2: An example of 25 semantically non-selective stop words. Manning et al. (2009)

Removing such words before processing reduced the noise from the collection and helped to increase the accuracy of the system. Moreover, it helps in the computational process, as we will be dealing with much smaller vectors.

We expand the result text by electing the most relevant synonyms for each word to be including in the composed vector. This step is crucial to capture a higher level of semantic. Therefore, if the same concept is formulated using different words, it still is considered toward the correct answer. The WordNet[1] lexical database is used within the normalisation engine to extract the synonyms.

Finally, we normalise result texts by converting each word to its base form. Thus words like *organise*, *organizes*, and *organising* would be mapped to the same element in our vector. This will help in comparing the students' work to the model answer as they may use different forms of the same word to express the same idea. To accomplish this task we apply Porter's algorithm. We found that the aforementioned steps would greatly increase the accuracy of the results.

We emphasise on this where we examine the similarity of each part of the document separately, before calculating the full document similarity. This approach helps the problem faced when marking the documents organisation, not only the document content. At this stage we consider only three parts of the document, the abstract, document body, and the conclusion. We use these parts as they are clearly mentioned in the question description. We used a linear function for combining these similarities with weighting factors for each part.

[1] http://wordnet.princeton.edu

EXPERIMENT and RESULTS

To test our approach, we selected eight essays from two different knowledgeable areas. Each essay was assigned to two independent human assessors to mark it. The human marks are then compared with the marks generated by our systems.

	Knowledge area 1				Knowledge area 2			
Essay number	1	2	3	4	5	6	7	8
hVec	56	57	38	62	63	57	66	54
mVec	73	79	69	72	80	77	88	82
humanAvg	86	88	62	65	82	80.5	73.5	84.5

Table 1: Results summary

Two automated runs are computed: 1) *hVec*, where we use one vector to represent the model document and the students essays. 2) *mVec*, where we represent each document with multi vectors, one vector for each part of the document.
We compared both automated runs with the average of the two human marks, *humanAvg* (Table 1). As can be seen in (Figure 3) the multi vector model, *mVec*, is much closer to human judgment, which gives a more reliable and accurate indication.

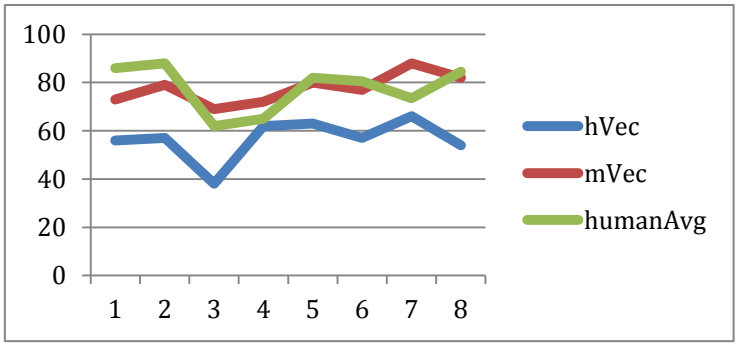

Figure 3: System results

CONCLUSION

This paper has given an account of, and the view for the widespread use of automated scores techniques. In this investigation, the aim was to assess the methods used in the automated scoring systems. Our system showed that normalisation and taking document structure into account gave a noticeable improvement in the results. The normalisation process was an important factor, which reduced noise in our data. Using document structure to compare with model answers instead of evaluating the whole document at once resulted in an increasing accuracy. For future work, we plan to improve our approach by investigating more ways to represent the document structure. We also plan to apply our methodology in different languages (e.g. Arabic).

REFERENCES

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Burstein, J. (2003). The E-rater® scoring engine: Automated essay scoring with natural language processing.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 206–210). Association for Computational Linguistics.

Chung, G. K., & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays*. Citeseer.

Elliot, S. (2003). IntelliMetric: From here to validity. *Automated Essay Scoring: A Cross-Disciplinary Perspective*, 71–86.

Hamp-Lyons, L. (2001). Fourth Generation Writing. *On Second Language Writing*, 117.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145–159.

Kukich, K. (2000). Beyond automated essay scoring. *IEEE Intelligent Systems*, *15*(5), 22–27.

Landauer, T. K. (2003). Automatic essay assessment, *10*(3), 295–308.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Automated Essay Scoring: A Cross-Disciplinary Perspective*, 87–112.

Learning, V. (2003). How does IntelliMetric score essay responses. *RB-929). Newtown, PA: Author*.

Nitko, A. J. (1996). *Educational assessment of students*. ERIC.

Page, E. B. (1966). The imminence of... grading essays by computer. *Phi Delta Kappan*, 238–243.

Page, E. B. (2003). Project essay grade: PEG. *Automated Essay Scoring: A Cross-Disciplinary Perspective*, 43–54.

Rudner, L. M., & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer*. ERIC Clearinghouse on Assessment and Evaluation.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4).

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, *1*(2).

Shermis, M. D., & Barrera, F. D. (2002). Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shermis, M. D., Raymat, M. V., & Barrera, F. (2003). Assessing Writing through the Curriculum with Automated Essay Scoring.

Sireci, S. G., & Rizavi, S. (2000). Comparing Computerized and Human Scoring of Students' Essays.

Williamson, D. M. (2009). A framework for implementing automated scoring. In *Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, San Diego, CA*.

Xie, S., Evanini, K., & Zechner, K. (2012). Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 103–111). Association for Computational Linguistics.

Zechner, K., & Xi, X. (2008). Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 98–106). Association for Computational Linguistics.