THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

A Review of Techniques Used For Automated Scoring

Tarandeep Singh Walia[1], Gurpreet Singh Josan[2], Amarpal Singh[3]

1. Assistant Professor, School of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India, e-mail: taran_walia2k@yahoo.com,
2. Assistant Professor, Punjabi University, Patiala, Punjab
3. Associate Professor, BCET, Gurdaspur, Punjab, India

Abstract

This paper shows the importance of Automated Scoring (AS) that has a better degree of reproducibility in comparison to human evaluators. It basically presents a comparative study of some of the techniques used to achieve the automated scoring and the limitation of the respective techniques. The paper goes through the literature survey, valid findings have been concluded out on various issues in existing systems. It reviews the features used in the existing Automated Answer Scoring system and efforts to develop a new semantic features for Automated Scoring System.

1. Introduction

The automated scoring is an act of evaluating grades to responses automatically which is based on predefined algorithms. In automatic scoring, an answer is received form as an input which is in typed text and score is presented as output which is based on number of features of the text. When the score are generated, answer as an input is passed through different modules like pre-processing, extract features and classify (Fig 1).

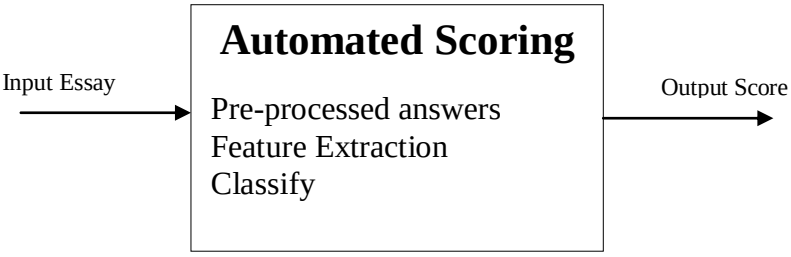

Fig. 1 Automated Scoring (AS) System

Although there are numerous AES systems existing, the focus of most studies is on the agreement between automated scores and human-assigned scores on a single essay. Furthermore, the agreement does not tell much about what is measured by automated scores. There is no sufficient evidence for validating

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

AES. Hence, it does not contribute in construction AES validation. The following Table 1 shows the strength of AES over manual scoring:

Table 1. Comparison between Manual And Automated Scoring

Sr. No.	Manual Scoring	Automated Scoring
1.	**Measurement Weaknesses**	**Measurement Strengths**
	Manual Scoring have measurement weaknesses: • Subjectivity • Lack of reproducibility • Inconsistency errors	Automated Scoring is able to achieve: • Consistency • Reproducibility • Traceability
2.	**Logistical Weaknesses**	**Logistical Strengths**
	• No quick rescoring • Takes more time to score • Not cost effective	• Quick rescoring • Time saving and possibility of immediate feedback. • Reduced cost
3.	**Other Weaknesses**	**Other Strengths**
	It requires: • Attention to basic human needs • Intensive direct labour and time • Calibration, training, recruiting and monitoring	It requires: • No basic human needs once the system is set • Only one trained operator is sufficient. Negligible labour and time • No more recruitment, training, calibration and monitoring

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

2. Background

The primary advantage of Automated scoring over the manual scoring are far enough with regards to its efficient feature, application of same evaluation criteria with grater consistency. Moreover, it ability to provide spontaneous feedback are its primary strength. Automated scoring achieves greater objectivity than manual scoring as computers are not being affected by external and emotional factors.

Table 2: Comparison of existing AES System

Existing System	Approach	Focus
PEG	Statistical	Style
IEA	LSA	Content
E-rater	NLP	Style and Content
IntelliMetric	NLP	Style and Content
BETSY	Bayesian text classification	Style and Content

Majority of automated scoring system generates nearly real time performance feedback on different aspects of writing. For examples: e-rater (ETS) model provides feedback on grammar, use of words, word mechanics, state and organization of a written typed text. Similar model Pearson's IEA covers the different aspects of writing for feedback. The aspects include ideas, organisations, conventions, fluency and choice of words. This advantage of AES is a limitation of human rating which is not able to provide such analytical feedback for huge quantities of essays. Also, human raters usually needs to train several score range linked with a specific rubric and certain tasks which further requires adequate training for shifting to a new grade. Such training is not at all required for AES which are able to evaluate the essays at different grading levels (for example: the e-rater, IEA and IntelliMetric). Comparison of the AES system as shown in above Table 2.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

3. Existing Techniques

It explains the review of the existing techniques on automated scoring system. The overall objective is to assess the shortcomings of earlier techniques. First, traditional automated systems have been discussed. Thereafter, other approaches have been discussed which are specifically related to the proposed research.

3.1 Short Answer Scoring Technique

The technique has been applied to short question answering because the domain of the system is fixed and moreover it is convenient to focus on meanings in short answers.

(C. Leacock 2003) defined an automated scoring engine as a C-rater which developed to grade answers to content-based short answer questions whereas C-rater utilizes morphological analysis, synonyms and predicate argument structure for assigning full or partial credit to a short answer questions, it cannot be referred merely as a sting machine program. C-rater agrees human raters to larger extent of 84% of the time.

(Song et al. 2010) explained the user interactive question answering by applying short text similarity assessment. The various applications of interactive question answering are: IR and text mining like text summarization, text categorization, content-based image retrieval and machine translation. It should be noted that the short text question-answers are used.

(Navjeet Kaur et al. 2012) explained short one-line free-text answers through automated assessment in the field of computer science. In their research, they have defined a segment of criteria for evaluation covering all the relevant areas of a short text evaluation system.

(Gomaa et al. 2012) used string similarity and corpus-based similarity in describing short answer grading. Scoring rules and predefined patterns are generated as these systems work in a supervised way. These two similarity measures when combined together proved to achieve a maximum correlation value of 0.504. These measures were tested separately before studying their combined effect.

(Gomaa et al. 2014) compared a different number of corpus-based and string-based similarities in order to explore text similarity approaches for automated short answer scoring in the Arabic language. The comparison between similarity measures reveals immediate feedback to the student. On analysis, resulted correlation and error rate findings proved that this system is useful for its application in a real scoring environment.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

(H. Rababah et al. 2017) forwarded a proposal of automated grading technique for Arabic essay questions in short answers. For this purpose of applying scoring process cosine similarity measure was used on the similarity between the student answer and standard answer. The results after experiments evaluated that the competitive scores were achieved when compared to other such approaches.

3.2. Vector Space Model (VSM) Approach

(Tsatsaronic et al. 2009) discussed a generalized VSM for Text Retrieval Based on Semantic Relatedness. The most difficult task is the modification of the standard interpretation of the VSM and other which deals with incorporating the semantic information in a theoretically sound and rigorous manner.

(Ekba et al. 2012) elaborated plagiarism detection in the text using Vector Space Model. In order to detect external plagiarism, they proposed technique based on textual similarity. Further it identifies the set of source documents from where the copying of suspicious document is carried out. This approach was based on the traditional VSM for selection.

(J.N. Singh et al. 2012) studied Vector Space Model information retrieval for analysis. It is one of the best traditional applied retrieval models for evaluating web page for its relevance. Various techniques of VSM to compute similarity score of the search engine hits were important.

(Jahan et al. 2014) discussed detection of plagiarism on electronic text-based answers using vector space model. On analysis, even though trigram utilizes enough time, it is more suitable for detecting plagiarism using cosine similarity measure in all text documents. VSM was used in retrieving information using query processing. Cosine similarity measure showing higher results was preferred over Jaccard similarity measure. The future work is to concentrate lesser time for dealing with a large amount of assignments with long length document and detect plagiarism optimally.

(Lilleberg J. et al. 2015) performed demonstration for classification of text with semantic features on the support vector machines and word2vec. Based on this, effectiveness of word2vec demonstrated by showing that tf-idf and word2vec combination can outperform tf-idf. Their approach was incomplete as it only scratches the surface; ideal results can still be expected. Recommendations for a future work depend on the ways to bring much improvement in consistency. This is achieved in many ways such as modification of stopword list or changing the weights.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

(Nguyen et al. 2017) combines word2vec with revised vector space model for better code retrieval. They perform a preliminary study that combining traditional IR with Word2Vec achieves better retrieval accuracy.

3.3. Other related concepts

(Brian E. Clauser et al. 2010) conducted a comparative study of the generalizability of scores produced by automated scoring systems and expert graders. In addition to the available information, their paper description is based on the performance of AES systems through various reports collected from expert raters and computer-produced scores. After analysis, performance was checked for physician's patient management skills through computer delivered assessment. Final results exhibit a relatively positive outcome regarding performance of the regression-based scoring algorithm.

(Safaa I. Hajeer 2012) conducted a study on various statistical similarity measures for their effectiveness. The project study on the several statistical measures in Information Retrieval (IR) is the highly effective on document retrieval taking a unified set of documents. Two issues were addressed viz. firstly, to study the different statistical measure for its effectiveness on a unified set of documents and secondly, to find the most appropriate one to classify documents through comparing them in an orderly manner. After analysis, it was concluded that the Cosine Similarity measure is the best for document retrieval technique. In future work, he hopes to extend this project to test other measures.

(Weigle S.C. 2013) presented the numerous considerations which are critical for English language learners and automated grading of essays. His study projected the various considerations to use automated scoring systems in evaluating second language writing. There were other aspects like challenges and opportunities which were listed in this presentation. His article analyses the extent to which system developers can assess the particular needs of learners in English language. It concludes that greater the evaluators and authorities posses knowledge regarding automated scoring system, the more will be chance of this technology to be used widely to meet the ever growing demands of huge population.

(B. Paskaleva et al. 2014) developed a vector space model for information retrieval with generalized similarity measures. They developed a new set of similarity functions for information retrieval. Records were considered as multisets of tokens which map records into real vectors. In their research, for bridging the gap between set-based models and Vector Space Model consistent extensions of set-based similarity functions were developed.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

(McNamaraa D.S. et al. 2015) explained in their study the significance of approach based on hierarchical classification in automated grading of essays. It significantly relies on machine learning approaches which are meant for computing essay scores involving a set of text variables. On analysis, 55% exact accuracy is revealed and along with 92% adjacent accuracy. Although, features which inform the overall assessment will differentiate depending on the specific problem, yet this technique is able to get performance models with high accuracy and information in comparison to simple one-shot regression.

(Md Arafat Sultan et al. 2016) discussed fast and easy short answer grading with high accuracy. In their research, student's short answer question is given with the correct answer; the principle of grading student response is derived from its semantic similarity with the correct answer. Key measure employed in their supervised model utilizes the recent approach of identifying the short-text similarity features. In Addition, the term weighting mechanisms are needed to identify important answer words in many cases. Accuracy for answer scoring can be achieved by evaluating a simple base model that can be easily extended with new features.

(Tianqi Wang et al. 2018) conducted a study on identifying current issues in short answer grading (SAG). In order to observe the issues involved in SAG, they analysed the results of a simple SAG approach. They used KNN to score query answers, where vector representations of answers are generated from weighted, pre-trained word embedding. By analyzing the errors in the given approach, it was shown how the diversity and short length of answers caused problems to SAG. Properties of short answer scoring such as diversity of answers were statistically analysed.

(Kevin Raczynski and Allan Cohen 2018) in their research article "Appraising the scoring performance of automated essay scoring systems—some additional considerations", provided useful validation framework for assessment of the automated scoring system. They determined the type of essays which can be used to calibrate and test Automated Essay Scoring (AES) systems. They also discussed what human grades should be used when there are scoring disagreements among multiple human raters.

(Yoav Cohen et al. 2018) discussed validation on manual and automated scoring of essays against "True" scores. Raters were divided into two groups (14 or 15 raters per group) rated 250 essays in two sets which were all written in response to the same prompt, thereby providing an approximate true score to the essay. Training on the datasets was provided to an automated essay scoring (AES) system in order to score the essays using a cross-validation scheme. The study

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

concluded that the correlation between human scores with automated scores is to the same extent as human graders correlate with one other.

4. Research Gap

The above review of preceding approaches brings to the surface their drawbacks at different levels. During literature survey research gaps found for the reliable automated scoring system. This paper presented the existing literature review to assess the shortcomings of earlier techniques in order to define the overall objective of the proposed study. First, traditional automated systems have been discussed. Thereafter, other approaches have been discussed which are specifically related to the proposed research. From the literature survey, it has been found that there are many aspects which have been overlooked in the past; especially feature based automated short answer scoring. The major research gaps in previously reported work are highlighted as refinement for an accurate scoring.

Conclusion

The automated scoring is an act of evaluating grades to responses automatically which is based on predefined algorithms. Although there are numerous AES systems existing, the focus of most studies is on the agreement between automated scores and human-assigned scores on a single essay. The overall objective of this paper is to assess the shortcomings of earlier techniques. The major research gaps in previously reported work are highlighted as refinement for an accurate scoring.

REFERENCES

[1] Al-Ashmoery, Yahya, and Messoussi, R., (2015), "Learning analytics system for assessing students' performance quality and text mining in online communication," *2015 Intelligent Systems and Computer Vision (ISCV)*.

[2] Alzahrani, A., Alzahrani, A., AlArfaj, F.K., Almohammadi, K., and Alrashid, M., (2015), "AutoScor: An Automated System for Essay Questions Scoring," *International Journal of Humanities Social Sciences and Education (IJHSSE)*, 2(5), pp. 182-187.

[3] Attali, Y., and Burstein, J., (2006), "Automated essay scoring with e-rater V.2.," *The Journal of Technology, Learning, and Assessment,* 4(3), pp. 3–30.

[4] Aziz, MJ.A., Ahmad, F.D., Ghani, A.A.A., and Mahmod, R., (2009), "Automated Marking System for Short Answer Examination (AMS-SAE)," *IEEE Symposium on Industrial Electronics and Applications (ISIEA 2009)*, 1, pp. 47-51.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

[5] Bejar, I. I., (2011), "A validity-based approach to quality control and assurance of automated scoring," *Assessment in Education: Principles, Policy & Practice,* 18(3), pp. 319–341. http://www.tandfonline.com/doi/abs/10.1080/0969594X.2011.555329

[6] Burrows, S., Gurevych, I., and Stein, B., (2015), "The Eras and Trends of Automatic Short Answer Grading," *Int J Artif Intell Educ, Springer,* 25, pp. 60–117.

[7] Chen, Yen-Yu, Liu C., Chang, T.H., and Lee, C.H., (2010) "An Unsupervised Automated Essay Scoring System," *IEEE Intelligent Systems,* 25(5), pp. 61-67.

[8] Dwivedia, S.K., and Singh, V., (2013), "Research and reviews in question answering system," *Procedia Technology, Elsevier,* 10, pp. 417 – 424

[9] Gomaa, W. H., and Fahmy, A. A., (2014), "Automatic scoring for answers to Arabic test questions," *International Journal of Computer Speech & Language,* 28(4), pp. 833-857.

[10] Gupta, P., and Gupta, V., (2012), "A Survey of Text Question Answering Techniques," *International Journal of 139 Computer Applications,* 53(4), pp. 0975 – 8887.

[11] Gupta, V. and Lehal, G.S., (2011), "Preprocessing Phase of Punjabi Language Text Summarization," *Information Systems for Indian Languages Volume 139 of the series Communications in Computer and Information Science,* pp. 250-253.

[12] Hajeer, S.I., (2012), "Comparison on the Effectiveness of Different Statistical Similarity Measures," *International Journal of Computer Applications,* 53(8), pp. 0975-8887.

[13] Heie, M.H., Whittaker, E.W.D., and Furui, S., (2012), "Question answering using statistical language modelling," *Elsevier, Computer Speech and Language,* 26, pp. 193–209.

[14] Jovita, Linda, Hartawan, A., and Suhartono, D., (2015), "Using Vector Space Model in Question Answering System," *ICCSCI 2015, Procedia Computer Science, Elsevier,* 59, pp. 305 – 311.

[15] Kerr, D., and Hamid, (2013), "Automatic Short Essay scoring using NLP to extract semantic information," *Journal of National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

[16] Kolomiyets, O., and Moens, M., (2011), "A survey on question answering technology from an information retrieval perspective," *Information Sciences, Elsevier,* 181, pp. 5412–5434.

[17] Kosseim, L., and Yousefi, J., (2008), "Improving the performance of question answering with semantically equivalent answer patterns," *Data & Knowledge Engineering,* 66, pp. 53–67.

[18] Landauer, T. K., Foltz, P. W., and Laham, D., (2004), "What is LSA?," web site from http://lsa.colorado.edu/whatis.html

[19] Leacock, C., and Chodorow, M., (2003), "C-rater: Automated Scoring of Short-Answer Questions," *Computers and the Humanities,* 37(4), pp. 389-405.

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

[20] Liu, H., Bao, H., and Xu, D., (2012), "Concept Vector for Semantic Similarity and Relatedness Based on WordNet Structure," *Journal of Systems and Software,* 85(2), pp.370-381.

[21] Li, H., Tian, Y., and Cai, Q., (2011), "Improvement of semantic similarity algorithm based on WordNet," *6th IEEE Conference on Industrial Electronics and Applications (ICIEA),* pp. 564-567.

[22] Li,Y., and Yan, Y., (2012), "An effective Automated essay scoring system using support Vector regression," *In proceeding of International Conference of Intelligent Computation Technology and Automation(ICICTA), IEEE,* pp. 65-68.

[23] Lilleberg, J., Zhu, Y., and Zhang, Y., (2015), "Support Vector Machines and Word2vec for Text Classification with Semantic Features," *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), IEEE, Beijing, China,* ISBN 978-1467372909.

[24] Ogheneovo, E. E., and Japheth, R. B., (2016), "Application of Vector Space Model to Query Ranking and Information Retrieval," *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE),* 6(5), pp. 42-47.

[25] Shaptala, R., Kyselova, A., and Kyselov, G., (2017), "Exploring the Vector Space Model for Online Courses," *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON),* Kiev, pp. 861-864.

[26] Sharma, D.V., and Aarti, (2011), "Punjabi Language Characterstics and Role of Thesaurus in NLP," *International Journal of Computer Science and Information Technology*, 2(4), pp. 1434-1437.

[27] Singh, J. N., and Dwivedi, S. K., (2012), "Analysis of Vector Space Model Information Retrieval," *Int'l Journal of computer Applications (IJCA), pp.14-18.*

[28] Tandalla, L., "Scoring Short Answer Essays(2012)," *International Journal of Comp. Sci. & Tech.(IJCST),* 2(1).

[29] Thomas, Ani, et al., (2011), "Extracting Noun Phrases in Subject and Object Roles for Exploring Text Semantics," *International Journal on Computer Science and Engineering,* 3(1).

[30] Toranj, S., and Nejad, D., (2012), "Automated versus Human Essay scoring: A Comparative Study," *IASCSIT Press, Singapore,* 33.

[31] Tsatsaronic, G., and Panagiotopoulous, V., (2009), "A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness," *In Proceedings of the EACI 2009 Student Research Workshop, Athens, Greece, Association of Computational Linguistics,* pp. 70-78.

[32] Xia, T., Chai, Y., and Lu, H., (2013), "E-learning support system aided by VSM based Question Answering System," *2013 8th International Conference on. IEEE Computer Science & Education (ICCSE),* pp. 1281-1285.

[33] XU, H., ZENG, W., GUI, J., QU, P., ZHU, Q., and WANG, L., (2015), "Exploring Similarity Between Academic Paper and Patent Based on Latent Semantic Analysis and Vector Space Model,"

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, pp. 801-805.

[34] Xu, L.H., and ShuTao, "Text Similarity Algorithm Based on Semantic Vector Space Model", *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Japan,* pp. 1-4.

[35] Ye, X., Shen, H., Ma, X., Bunescu, R., and Liu, C., (2016), "From Word Embeddings to Document Similarities For Improved Information Retrieval In Software Engineering," *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*, Austin, TX, pp. 404-415.

[36] Walia T.S., Josan G.S. & Singh A., "An Efficient Automated Answer Scoring System for Punjabi Language", Elsevier, Egyptian Informatics Journal (2019), Vol. 20, pp. 89-96

[37] Walia T.S., Josan G.S. & Singh A., "Quantum Based Automated Short Answer Scoring System", World Scientific, Modern Physics Letters B (2018), Vol. 32, No 33, pp. 1-27

[38] Walia T.S., Josan G.S. & Singh A., "Statistical Technique for Automated Answer Scoring: An Overview," International Journal of Management, Technology And Engineering (2018), ISSN No.: 2249-7455, Vol 8, pp. 1835-40.

[39] Weigle, and Cushing, S., (2013), "English language learners and automated scoring of essays: Critical considerations," *Assessing Writing,* 18(1), pp. 85-99.

[40] Weigle, S.C., (2013), "English as a second language writing and automated essay evaluation," *Handbook of Automated Essay Evaluation: Current Applications and New Directions, Routledge, New York,* pp. 36–54.

[41] Zhang, and Lee, W.S., (2003), "Question classification using support vector machines," *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and evelopment in Information Retrieval,* pp. 26–32.

[42] Zupanc, K., and Bosnic, Z., (2017), "Automated essay evaluation with semantic analysis," *Elsevier, Knowledge-Based Systems,* 120, pp. 118–132.