Applied Artificial Intelligence

An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Subjective Evaluation: A Comparison of Several Statistical Techniques

Himani Mittal & M Syamala Devi

To cite this article: Himani Mittal & M Syamala Devi (2018) Subjective Evaluation: A Comparison of Several Statistical Techniques, Applied Artificial Intelligence, 32:1, 85-95, DOI: 10.1080/08839514.2018.1451095

To link to this article: https://doi.org/10.1080/08839514.2018.1451095

Published online: 21 Mar 2018.

APPLIED ARTIFICIAL INTELLIGENCE
2018, VOL. 32, NO. 1, 85–95
https://doi.org/10.1080/08839514.2018.1451095

Taylor & Francis
Taylor & Francis Group

Subjective Evaluation: A Comparison of Several Statistical Techniques

Himani Mittal and M Syamala Devi

Department of Computer Science and Applications, Panjab University, Chandigarh, India

ABSTRACT

Evaluation of subjective examinations using computerized tools has been a topic of research for more than four decades. Several statistical and mathematical techniques have been proposed by various researchers. In this research work, the several methods proposed earlier like Latent Semantic Analysis (LSA), Generalized Latent Semantic Analysis (GLSA), Bilingual Evaluation Understudy (BLEU), and Maximum Entropy (MaxEnt) are compared on common input data. The techniques are implemented using Java programming language, MatLab, and other open source tools. Experiments have been conducted and developed prototypes are tested using a database of 4500 answers with approximately 50 questions of computer science. Comparison of these techniques on a common database is not available in the literature as far as the authors' review is concerned. The database used for testing is collected by conducting tests of students of graduate level in the field of computer science. The pros and cons of each technique on the basis of experiments are discussed in the paper.

Introduction

Evaluation is a systematic determination of a subject's merit, worth, and significance, using criteria governed by a set of standards. Student evaluation in subjective examination assess whether the student has gained knowledge as a result of teaching. The primary purpose of evaluation, in addition to gaining insights into prior or the existing initiatives, is to enable reflection and assist in the identification of future change. Efforts are being made to computerize subjective evaluation for last four decades.

In manual examination system, the student submits the answers written on answer-book. These answer-books are given to evaluator for grading. Then, the results are compiled. The use of computerized tools reduces the limitations of the manual process. The online examination system can transfer the answers submitted by students, electronically to the centralized database and thus avoiding the physical movement of answer books. Manual grading is time consuming and depends on the availability of the evaluator. The use of software tools

CONTACT Himani Mittal research.himani@gmail.com Assistant Professor, GGDSD College, Sector 32, Chandigarh 160030, India

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/UAAI.

provides fast and human error-free results. They ensure uniformity of the marking scheme because they have the same inference mechanism for checking all the answers. There is a need to remove jeopardize in examination like bias, lack of transparency; and to establish an effective and efficient evaluation system. Such a system should be different from any regular information system which requires constant monitoring. It should be capable of assisting the human examiner.

Several statistical and mathematical techniques have been proposed for subjective evaluation. In this research work, the statistical methods like Latent Semantic Analysis (LSA), E-rater, Generalized LSA, Bilingual Evaluation Understudy (BLEU) and Maximum Entropy (MaxEnt) are discussed. The LSA, GLSA, BLEU, MaxEnt techniques are implemented and tested.

The paper is organized as follows: Section 2 contains the review of related work. Section 3 includes the implementation details. In Section 4, testing and analysis of results is included. Section 5 includes the conclusions.

Review of related work

General review

In 1982, Macdonald, Frase, Ginrich, and Keenan (Macdonald et al. 1982) developed Writer's Workbench programs, which analyze English prose and suggest improvements. Project Essay Grader (PEG) (Page 1994), performs the evaluation based on style analysis. Its agreement with human graders is 87%. It measures features like essay length, word length and vocabulary used etc. PEG was brought to web in (Shermis et al. 2001). Electronic Essay Rater (E-Rater) was developed by Burstein, Kukich, Wolff, Chi, & Chodorow, 1998 (Burstein et al. 1998) and later enhanced in 2006 (Attali and Burstein 2006). Conceptual Rater (C-rater) (Burstein et al. 2000 (Burstein and Marcu 2000) and Sukkarieh et al. 2009 (Sukkarieh and Blackmore 2009) is a Natural language based prototype aimed at the evaluation of short answers related to content-based questions and has an accuracy up to 93%. In 2011, Sukkarieh (Sukkarieh 2011) discussed that the max-entropy technique is used in C-rater. C-Rater achieved over 80% agreement with the score assigned by a human-grader. In 1999, Foltz et al. [9,10,and 11] applied mathematical technique called Latent Semantic Analysis (LSA) to computerized evaluation in a tool called Intelligent Essay Assessor (IEA). The correlation between human and IEA scores is 80%. This technique is not capable of evaluating short-answer questions and technical answers. In 1999, Hofmann (Hofmann 1999) developed Probabilistic latent semantic analysis (PLSA). This method is an improvement over LSA, as it has strong statistical foundation in aspect modeling. It calculates document to word joint probability using estimation maximization algorithm. In 2001, Callear, Jerrams-Smith and

Soh (Callear, Jerrams-Smith, and Soh 2001) presented a survey of major systems for the automated evaluation and proposed a new technique called the Automated Text Marker (ATM) prototype. The two main components of ATM are the syntax and semantics analyzers. ATM is written in Prolog. The Bayesian Essay Test Scoring System (BETSY) (2002) (Rudner and Liang 2002) is a program that classifies text based on trained material. The models used for text classification are Multivariate Bernoulli Model (MBM) and the Multinomial Model. An accuracy of over 80% was achieved. In 2003, Blei et al. (Blei, Ng, and Jordan 2003) suggested a generalization of PLSA, by using mixture model and dirichlet allocation distribution for calculating co-occurrence probability of words. This technique is known as Latent Dirichlet Allocation (LDA). In 2003, Rose, Roque, Bhembe, and VanLehn (Rose et al. 2003), proposed a tool called CarmelTC. The hybrid CarmelTC approach uses decision trees and Naive Bayes text classifier statistical technique. The success rate of this CamelTC was 90% which was comparable to that of LSA. In 2005, P´erez, Gliozzo, Strapparava, Alfonseca, Rodr´ıguez and Magnini (Perez et al. 2005) have tested the hypothesis that combining different knowledge sources and algorithms is a viable strategy for an automatic evaluation of students' free-text answers. The combination schema for the techniques Bleu (machine translation effectiveness algorithm) and LSA (essay clustering technique) was found effective. The mean correlation to the human's scores has reached 50%. In 2008, Kakkonen, Myller, Sutinen and Timonen (Kakkonen et al. 2008), Automatic Essay Assessor (AEA) is a system that utilizes information retrieval techniques such as LSA, PLSA and LDA for automatic essay grading. They performed a series of experiments using LSA, PLSA, and LDA for document comparisons in AEA comparing the applicability of LSA, PLSA, and LDA to essay grading with empirical data. It was found that using LSA yielded slightly more accurate grading than PLSA and LDA. In 2008, Li Bin, Lu Jun, Yao Jian-Min, Zhu Qiao-Ming (Bin et al. 2008), the K-Nearest Neighbor (KNN) algorithm for text classification is applied. The experiments show accuracy above 76%. In 2010, Islam and Hoque (Islam and Hoque 2010) proposed a system that makes use of the generalized latent semantic analysis (GLSA) technique for evaluation. It has 89% of accuracy which show that the system is very closer to human grader. In 2010, Cutrone and Chang (Cutrone and Chang 2010) in their research paper proposed a short answer evaluation method using Natural language processing (NLP) techniques. It reduces the standard answer and student answer into its canonical form and compares them. It uses NLP techniques like parsing, stemming, stop-word removal, morpho-syntactic-variation handling, etc. The system in its current format is capable of processing answers containing a single sentence that is free of grammar and spelling mistakes. In 2016, Mittal and Devi (Mittal and Syamala Devi 2016) have extended the work given in (Perez et al. 2005) and combined LSA and BLEU using Fuzzy Logic. The hybrid developed by them is tested on database of 50 questions and accuracy rate is 72–99%.

Specific techniques chosen for this work

Latent semantic analysis

The LSA technique was first proposed by (Deerwester et al. 1990) Deerwester, Dumais, Furnas, Landauer and Harshman. This technique is used to establish similarity between two contents. Before LSA exact word match were found between the documents. This was not a good method as individual words may appear in more than one context. LSA tries to overcome this problem of term-matching retrieval by treating the unreliability of observed term matching document association as a statistical problem. It establishes the hidden semantic structure in data using statistical method which is indeterminate by the random word selection in retrieval. For this words and documents are represented as vectors by performing Singular Value Decomposition (SVD). These vectors if related to each other will have positive cosine similarity. It is a dimension reduction technique. LSA is used in search engine and plagiarism detection. In both these applications, not only the exact word matches but words similar in meaning also need to be identified. Latent Semantic Analysis is able to establish relation between synonyms and words of opposite meaning. The accuracy of such relation depends on concept model given as input. The LSA was tested for evaluating its performance in identifying the semantic relationship between texts (Foltz and Landauer 1998). It was established that LSA captures coherence of text in continuity of lexical semantics. It is also found that LSA captures coherence even when synonym words and phrases of related meaning .but containing different lexical items. LSA has the ability to segment discourse i.e., topic change. As the semantic distance between documents decreases the cosine falls. It is established that LSA can be used as a writing critic. However, the level of performance will depend on the word matrix provided to construct the word by a document matrix. LSA does not perform syntactic processing or parsing. The word order is not considered. If the same words are repeated again and again then high correlation is generated. Intelligent Essay Assesor (IEA) by Foltz, Lahman, Landauer (Foltz, Laham, and Landauer 1999) is a set of software tools for scoring the quality of essay content. They used IEA in TOEFL exam (ETS) organization and results were satisfying. Highest correlation between human and IEA was .87. Around 1205 essays with 12 topics were tested human-human correlation was .707 and human-LSA correlation was .701.

Electronic essay rater

It was developed in 2006 (Attali and Burstein 2006). It used MSNLP for parsing the text and extracting text features. Then weight-age is assigned to these features. Whenever a new essay is to be evaluated, its features are compared to already grade essays. It is successfully used for AWA test in GMAT with agreement rates between human expert and system consistently at 84%. It uses hybrid feature identification method, including syntactic structure analysis,

rhetorical structure analysis and topic analysis. In Syntactic Analysis, all sentences in the essay were parsed and types of clauses and phrases are identified. In Rhetorical analysis, discourse of the student answer is evaluated. In topical analysis, word weight and word frequency is calculated to find the content similarity. The emphasis of this software tool is to evaluate the effectiveness of English essays and evaluation is done with AWA test in mind. Testing of the tool is done for 57 features and outputs of feature values for each answer are combined using linear regression equation. The tool uses linear regression model fitting for combining feature values. It can be replaced with fuzzy logic or other similar techniques. The set of features used are specific to AWA and Toefl test, a generic model needs to be developed. The focus is on language grammar based feature identification. Its applicability to technical answers is not proven as yet. Discrepant essays are scored like regular essays, such cases are not explicitly pointed out. It was tested with 18000 essays with 64 questions and a maximum of 0.93 correlations with human examiner is calculated.

Generalized latent semantic analysis

In 2010, Islam and Hoque (Islam and Hoque 2010) proposed a system that makes use of Generalized latent semantic analysis (GLSA) technique for evaluation. In GLSA n-gram by document matrix is created instead of a word by document matrix of LSA. It has 89% of accuracy which show that the system is very closer to human grader. In GLSA n-gram by document matrix is created instead of word by document matrix of LSA. N-grams i.e., unigrams, bigrams, trigrams,. . ...*n*-grams are constructed using neighboring terms of the selected important words. For training 960 essays on 3 prompts were used. Testing was done using 120 essays. Accuracy of 86–97% is reported. The use of n-grams makes the software to consider the word order also to some extent which is a major drawback of LSA.

Conceptual rater

In 2011, Sukkarieh (Sukkarieh 2011) discussed that the Maximum Entropy technique is used in the ***Conceptual rater*** (C-rater). The C-rater achieved over 80% agreement with the score assigned by a human-grader. It uses a perceptron neural network to evaluate answers. It read the training data and studies the features. The feature extracted is what word follows and precedes the word under consideration. The probability is calculated for the current word to appear in a given context. Then, it reads one student answer at a time and finds if student answer entails the standard answer concepts. It finds lexicon similarity between phrases, provides for morphological analysis of the answers, matches the subject and predicates, matches the negative role with positive role.

Bleu-inspired algorithm

In 2005, Diana Perez et al. (Perez et al. 2005), developed a system using Latent semantic Analysis (LSA) and BLEU (bilingual Evaluation understudy) algorithm to essay evaluation. LSA performs semantic analysis and modified BLEU as used by the authors performs syntactic Analysis. The results of the two are combined by a linear equation. However, the amount of weightage that should be given to BLEU generated score and LSA generated score is not fixed. Author has shown multiple combinations and average success rate is 50% of times.

Hybrid technique

In 2016, Mittal and Devi (Mittal and Syamala Devi 2016) have extended the work in (Perez et al. 2005). They claim that LSA generated score is like an upper bound on the maximum marks the student answer can get. The score generated using BLEU is like a lower bound. The two have been combined using Fuzzy Logic. The accuracy of results is 72–99%.

Implementation of subjective evaluation techniques

The selected techniques as mentioned above, were implemented and tested. The implementation was primarily done using Java Programming language. The LSA technique is implemented in Java programming language and MatLab. MatLab is used for performing Singular Value Decomposition (SVD) for LSA. Open source libraries are used for invoking Matlab (Matlab Control- 2015) from java code. The GLSA technique is implemented by extending the LSA package by incorporating n-grams. All the programming and extension tools used are the same as in LSA. BLEU is implemented in java programming language. The maximum Entropy package is freely available at http://maxent.sourceforge.net .It is a Java-based maximum entropy package. The features and working of this package are understood and it is used for evaluation. The Hybrid Technique is implemented in java programming language and MatLab for Fuzzy logic.Various tools and techniques used for implementation are shown in Table 1.

Testing and analysis of results

Testing and results

There is no standard database in subjective answer evaluation which can be used for comparing all the techniques discussed in this paper. Therefore, the database was created over a period of time by conducting class tests. All the techniques (as implemented above) for answer evaluation have been tested using this common database. The database consists of 4500 answers with approximately 50 prompts

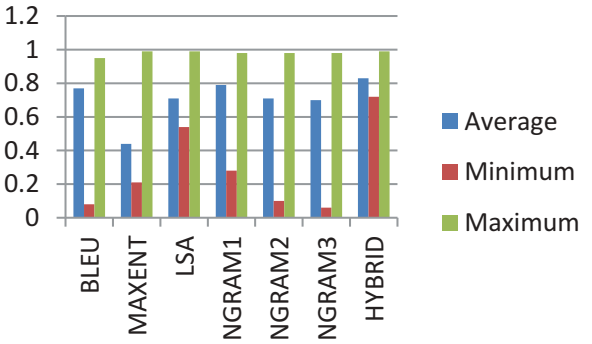

Figure 1. Accuracy of scores generated using different techniques.

from field of computers (technical answers). All the techniques were used to evaluate the answers and the comparison of accuracy of various techniques is given in Figure 1. The correlation for BLEU varies between 0.08 and 0.95, LSA varies between 0.54 and 0.99, maximum entropy varies between 0.21 and 0.99, GLSA Ng2 varies between 0.28 and 0.98, GLSA Ng3 varies between 0.10 and 0.98, GLSA Ng4 varies between 0.06 and 0.98, and Hybrid Technique varies between 0.72 and 0.99. The Blue bar shows maximum correlation achieved using the labeled techniques and red bar shows minimum correlation. The correlation is calculated between human assigned scores and computer generated scores. The maximum correlation for all the techniques is above 90%. The Hybrid (LSA, BLEU, and FUZZY) is giving more consistent results with 72–99% accuracy. The reasons for low minimum accuracy for each technique have been identified. LSA overrates repetition of keywords and word choice is not open to any words. In BLEU, word choice is limited to Model answer keywords. Maximum Entropy is unable to rate discrepant essays. GLSA (n-gram size 2,3,4) is only theoretic improvement. As the word affinity is different and varies according to sentence construction the ideal n-gram size cannot be fixed to one value. Also the execution time increases many fold and the technique is very slow (15–20 minutes for input data with 500 word model and 60 student answers with 500 words each). The Hybrid technique identifies discrepant essays and allows freedom of word choice.

Analysis of results

The BLEU algorithm matches exact word presence. It acts as a lower bound on the maximum number that should be given to a student answer. It behaves more like PEG and some features measured in E-rater like word average. The Maximum Entropy technique is a classification algorithm. The student answers are of varying nature as they use different words, different examples and some extreme cases where invalid content is present. It is not possible to model all the

classes beforehand. Also, the level of accuracy changes as the number of model answers is increased. It is not feasible to generate such a large database of model answers. The LSA algorithm assigns score for the presence of keyword with respect to semantic similarity of keywords. It does not consider the syntactic structure of the answers but measures the semantic aspect thoroughly. However, the system can be used as an upper bound as to maximum how many marks can be assigned on the basis of semantic similarity. It assigns high score if the keywords are repeated several times which is not a good performance measure as there can be invalid content also. The GLSA algorithm performs somewhat like LSA except that it looks for *n*-grams. Due to this modification it assigns less score to answers that do not have words appearing in certain order. This is a improvement in theoretic terms, but when we look at the results of comparison with human assigned scores, LSA outperforms GLSA.

Hybrid technique makes use of combination of LSA and BLEU using Fuzzy Logic. The standard LSA technique is modified by pre-processing the input to find synonyms and performing stemming. This makes the output of LSA more precise. By finding synonyms and stemming, all the forms of the words are provided for. It gives the students freedom to use any similar word or form of same word depending of tense and sentence structure being used. When word comparisons are made to calculate the term frequency, all the synonyms of the keyword are taken into consideration. A modified version of BiLingual Evaluation Understudy (BLEU) algorithm is used. The original BLEU algorithm makes use of n-grams (phrases of words). We are using individual words as we have already removed stop words and phrases cannot be constructed. Secondly, the original method calculates brevity factor. We are not calculating the brevity factor. We are just using BLEU to clip the max usage of keywords so that unnecessary repetition of keywords does not fetch more marks to students. LSA has two inherent problems: overrating presence of repetition of keywords and syntactic significance of words in sentence. BLEU takes care of the first problem by clipping the number of instances of keyword by maximum occurrence in model answer. So repetition of keywords is dealt with the help of BLEU. The syntactic significance of keywords is checked using WordNet and also WordNet is used to provide the student freedom to choose words different from keywords but having same meaning. The correlations generated by LSA and BLEU are combined using Fuzzy Logic with rules specified in Table 2. The interaction between the two fuzzy variables – LSA and BLEU is calculated with the help of empirical data study.

The comparison chart for comparing reported results of several techniques along with proposed hybrid technique is given in Table 3. IEA tool has accuracy of 59–89%. However, it is unable to point out discrepant essays which have unnecessary repetition of keywords. E-rater has an accuracy of 87–93% with its main focus on style of writing as it is used to evaluate the English essays. It treats the discrepant essays like regular essays. Its feature set is more specific to English

Table 1. Tools and techniques used for implementation.

Tool	Library	Technology/technique	Purpose
Java development Kit 1.7			It is used for the development of Evaluation Application
	Stemmer		Porter's Algorithm implemented by originator of the algorithm
MatLab 2013	Matlab Control Library	Fuzzy Logic Matrix Operations	Used for LSA matrix calculations and Fuzzy Logic implementation for hybrid technique.
WordNet 2.1	JWI 2.2.3	Semantic Networks	Used for finding word Synonyms
	Guava library	Multi Hash Lists	Used for counting frequency of words

Table 2. Rules for inference engine of fuzzy logic.

INPUT	BLEU	Bad	Average	Excellent
LSA	OUTPUT			
Bad		Bad	Ok	Ok
Average		Ok	Average	Average
Excellent		Ok	Average	Excellent

Table 3. Comparison of hybrid technique with the existing techniques (according to reported results).

Tool Criteria	IEA	E-rater	Atenea	C-rater	Hybrid technique
Accuracy maximum	89%	93%	79%	98%	99%
Accuracy minimum	59%	87%	23%	48%	72%
Technique(s) used	Latent Semantic Analysis	Latent Semantic Analysis, word average and grammar based feature extraction. Feature scores combined using linear regression.	Latent Semantic Analysis and Bilingual Evaluation Understudy combined using linear equation	Maximum Entropy based Technique	Latent Semantic Analysis, Bilingual evaluation Understudy and Fuzzy Logic

language and designed specifically for AWA test. Atenea tool has combined LSA and BLEU using linear equations. The combination coefficients of LSA and BLEU are left undefined. C-rater tool uses MaxEnt technique which has an inherent problem of over fitting. So, it cannot handle discrepant essays. Hybrid technique has an accuracy between 72 and 99% can identify discrepant essays easily.

Conclusion

The use of software tools can never replace a human examiner, because human evaluation is more holistic in nature. However, these techniques can be used to identify the discrepant answers and other exceptional cases and can be used for preliminary screening of student answers. The hybrid technique based software can help to a large extent the human examiner in evaluating subjective answers.

The techniques used for evaluation, LSA and BLEU are complementary combination. The fuzzy function gives balanced weight to LSA and BLEU depending on different combinations of outputs. The use of WordNet helps in reduction of number of keywords to be given, as it finds synonyms of given keywords. This ensures student can make use of words of his choice. The performance of technique can be improved by introducing domain specific ontology. It can be concluded from the results that Hybrid of LSA, BLEU and Fuzzy Logic can be a better choice as it performs consistently as compared to all the other techniques.

References

Attali, Y., and J. Burstein. 2006. Automated essay scoring with E-rater? V.2.0. *The Journal of Technology, Learning and Assessment* 4 (3):e–journal.

Bin, L., L. Jun, Y. Jian-Min, and Z. Qiao-Ming, (2008), Automated essay scoring using the knn algorithm, International Conference on Computer Science and Software Engineering, IEEE, Vol. 1, pp 735–738. DOI 10.1109/CSSE.2008.623

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (5):993–1022.

Burstein, J., K. Kukich, S. Wolff, L. Chi, M. Chodorow, L. Braden-Harder, and M. D. Harris, (1998), Automated scoring using a hybrid feature identification technique, Proceedings of the Annual Meeting of the Association of Computational Linguistics and International Conference on Computational Linguistics, 1, 206–10. Doi: 10.315/980451.980879.

Burstein, J., and D. Marcu, (2000), Towards using text summarization for essay-based feedback, Le 7e Conference Traitement Automatique des Langues Naturelles TALN'2000.

Callear, D., J. Jerrams-Smith, and V. Soh, (2001), Bridging gaps in computerized assessment of texts, Proceedings of the IEEE International Conference on Advanced Learning Techniques (ICALT'01), IEEE, 139–49.

Cutrone, L., and M. Chang, (2010), Automarking: automatic assessment of open questions, 10th IEEE International Conference on Advanced Learning Technologies, IEEE, 143–47.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41:391–407.

Foltz, P. W., D. Laham, and T. K. Landauer. 1999. Automated essay scoring: applications to educational technology. In *Proceedings of EDMedia'99*, Eds. B. Collis, and R. Oliver, pp. 939–944. Charlottesville, VA: Association of Computing in Education.

Foltz, P. W., and T. K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes* 25 (2&3):285–307.

Hofmann, T., (1999), Probabilistic Latent Semantic Indexing Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99).

Islam, M., and A. S. M. L. Hoque, (2010), Automated essay scoring using generalized latent semantic analysis, Proceedings of 13th International Conference on Computer and Information Technology, pp. 358–363.

Kakkonen, T., N. Myller, E. Sutinen, and J. Timonen. 2008. Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society* 11 (3):275–88.

Macdonald, N. H., L. T. Frase, P. S. Gingrich, and S. A. Keenan. 1982. The writer's workbench: computer aids for text analysis. *IEEE Transactions on Communications* 30 (1):105–10.

Matlab Control- https://code.google.com/p/matlabcontrol/(accessed on 31-Dec-2015

Mittal, H., and M. Syamala Devi, "Subjective evaluation using hybrid technique", Proceedings of International Conference on Innovations in Computer Science and Engineering, Springer-Verlag, 2016.

Page, E. B. 1994. Computer grading of student prose. *Using Modern Concepts and Software, the Journal of Experimental Education* 62 (2):127–42.

Perez, D., A. Gliozzo, C. Strapparava, E. Alfonseca, P. Rodrýguez, and B. Magnini, (2005), Automatic assessment of students' free-text answers underpinned by the combination of a bleu-inspired algorithm and latent semantic analysis, Proceedings of 18th International Florida Artificial Intelligence Research Society Conference, Flairs, AAAI Press.

Rose, C. P., A. Roque, D. Bhembe, and K. VanLehn. 2003. A hybrid approach to content analysis for automatic essay grading. *Proceedings of the 2003 Conference of the North American Chapter of Association of Computational Linguistics on Human Language Technology* 2:88–90. doi:10.3115/1073483.1073513.

Rudner, L. M., and T. Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning, and Assessment* 1 (2):e–journal.

Shermis, M. D., H. R. Mzumara, J. Olson, and S. Harrington. 2001. On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education* 26 (3):247–59.

Sukkarieh, J. Z. 2011. Using a maxent classifier for the automatic content scoring of free-text responses. *American Institute of Physics Conference Proceedings* 1305 (1):41.

Sukkarieh, J. Z., and J. Blackmore, (2009), c-rater: automatic content scoring for short constructed responses, Proceedings of the Twenty-Second International FLAIRS Conference, AAAI Press, pp. 290–295.