Research Journal of Applied Sciences, Engineering and Technology 12(3): 258-263, 2016
DOI: 10.19026/rjaset.12.2332
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 2, 2015 Accepted: August 2, 2015 Published: February 05, 2016

Research Article

An Automated Assessment System for Evaluation of Students' Answers Using Novel Similarity Measures

Madhumitha Ramamurthy and Ilango Krishnamurthi
Department of CSE, Sri Krishna College of Engineering and Technology, Coimbatore-641008, TamilNadu, India

Abstract: Artificial Intelligence has many applications in which automating a human behavior by machines is one of very important research activities currently in progress. This paper proposes an automated assessment system which uses two novel similarity measures which evaluate students' short and long answers and compares it with cosine similarity measure and n-gram similarity measure. The proposed system evaluates the information recall and comprehension type answers in Bloom's taxonomy. The comparison shows that the proposed system which uses two novel similarity measures outperforms the n-gram similarity measure and cosine similarity measure for information recall questions and comprehension questions. The system generated scores are also compared with human scores and the system scores correlates with human scores using Pearson and Spearman's correlation.

Keywords: Artificial intelligence, assessment, education, sentence similarity, similarity, WordNet

INTRODUCTION

Artificial intelligence creates machines with intelligence. Many e-learning applications are examples for machines with intelligence. Automation is also an important research area where automation of assessment of students' answers is an important research in the educational sector. Computer Assisted Assessment (CAA) helps to automate the assessment of answers by using computers.

Students' answers can be divided into objective and subjective answers where objective answer assessment is the most common one when compared to subjective answers assessment which includes short answers and long answers. In subjective assessment, more focus is given on short answers and have many approaches for assessment when compared to assessment of long answers.

Evaluation of answers is based on six types of questions according to bloom's taxonomy. Those six categories of questions are information recall, comprehension, application, analysis, synthesis and evaluation questions.

Information recall questions (Questions Skills, year) makes the students to recall the studied information. The students remember the studied information to answer an information recall question.

Comprehension questions (Questions Skills, year) make the students to use the studied information and express the information in their own words.

Application questions (Questions Skills, year) make the students use the studied information and to apply what they have learned to solve the problem.

Analysis questions (Questions Skills, year) make the students to analyze the questions by the studied information and answer those questions and they also reason out their findings.

Synthesis questions (Questions Skills, year) make the students to answer the questions by thinking innovatively by finding their own ways for solving the problems.

Evaluation questions (Questions Skills, year) make the students to answer the questions by evaluating and judging their idea and coming to a conclusion why an idea is better than the another idea and they should also give based on what criteria they have given this evaluation.

The humans can evaluate all these types of questions. But there is a challenge for computers to do this task. The proposed assessment system automates the evaluation of information recall type questions and comprehension questions.

LITERATURE REVIEW

PEG (Project Essay Grade), (Whittington and Hunt, 1999). This was one of the earliest implementations for automatic assessment. It did not consider NLP and lexical content to grade the essays and focused on only simple style analysis. The strength

Corresponding Author: Madhumitha Ramamurthy, Department of CSE, Sri Krishna College of Engineering and Technology, Coimbatore-641008, TamilNadu, India

of this method is that the correlation between human and computer for grading is 0.83. The weakness with this method is that only writing style of the essay is checked and a number of training essays which are manually graded are used to score a new essay.

E-rater (Electronic Essay Grade), (Burstein *et al*., 2001) checks the writing style and the structure of the essays rather than the specific content. The strength is the agreement between the E-rater and human is above 97%. The weakness of E-rater is that it requires a number of manually scored training essays to score the answers.

C-rater (Concept rater) (Burstein *et al*., 2001; Yigal *et al*., 2008) was also developed by ETS (Educational Testing Service) and it is also called as content rater. The scoring is based on the content and concept. It uses natural language processing techniques i.e., it uses lexical semantic techniques are used to build the scoring system. This system uses domain related, concept based data in evaluation. When compared to E-rater, in C-rater the number of training essays is reduced and it mainly focuses on the information that must be present in the correct answer. The semantic domain is limited and cannot recognize wrong concept. The strength is the agreement between the C-rater and the human judge is 84%.

IEA (Intelligent Essay Assessor) (Saxena and Gupta, 2009) uses LSA (Latent Semantic Analysis) technique. Semantic and syntactic information is considered to evaluate the essay. The word-document co-occurrence matrix is constructed and singular value decomposition is performed to find the match by calculating the cosine similarity measure between the words in the matrix. The correlation with the human judge is 86%.

SELSA (Syntactically Enhanced LSA) (Kanejiya *et al*., 2003) still improves the performance of the automatic evaluation of students' answers by considering syntactic and semantic information along with the word of the previous word. It considers POS tag of the word as well as the preceding word. Therefore SELSA has better performance in evaluating students' answers than LSA. The disadvantage is that this method also uses many training essays.

WebLAS (Bachman *et al*., 2002) uses Natural Language Processing and Pattern Matching Techniques. The student answer is compared with model answer and uses WordNet for checking the semantic information. This method does not need any manually graded training answers. The disadvantage is that it works for only short answers.

RARE (Pérez *et al*., 2005) is a free text answers assessor. It compares the answers of students' with teacher's reference answers. If there is no match between the students' and teachers' answers then the system cannot evaluate and award marks. Two solutions to solve this problem are paraphrased reduction and creation of new reference answers. For the first case AR identifies referential expressions for same referents and gathers them in co-referential chains. For the second case manually asks the teachers to write many reference answers and use it for evaluation. Atenea processes the students' and teachers' answers using the NLP techniques like stemming, removal of closed class words and word sense disambiguation. Then the processed answers are sent into the comparison module which in-turn calculates the students' score. The comparison module is based on BLEU algorithm. The disadvantage is that it considers a number of training answers to evaluate the student answer.

TANGOW (Alfonseca *et al*., 2005) is a system which supports adaptive web-based courses. Atenea is CAA system which is based on BLEU i.e., n-gram co-occurrence metrics and it performs vocabulary analysis and compares students' and teachers answer to score them automatically. They use NLP and statistical based techniques. There should be at least three reference answers written by various teachers and these answers are to be stored in the database. The reference answers can also be taken from the best student answer to have more alternative answers. The internal architecture of Atenea has a statistical module called ERB (Papineni *et al*., 2001) (Evaluating Responses with BLEU) and NLP modules to score automatically. The main disadvantage is that this approach considers a number of training answers to evaluate the student answers.

Pattern Matching techniques are used (Siddiqi and Harrison, 2008) for evaluating the answers. The students' answers are tagged for POS and extracts noun and verb phrases. These are used by the pattern matcher to match the patterns which conforms to the rules set by grammar. Then marking process gives the evaluated score. The main disadvantage is that it also considers a number of training answers to evaluate the student answers.

Automatic segmentation techniques (Hu and Xia 2010) and subject ontology is also for evaluating students' answers in this study. The reference answers and students' answers are converted to term document matrix and projected to k-dimensional LSI space by singular value decomposition. The answers are evaluated based on the similarity between projected vectors. This approach also makes use many reference answers to evaluate the student answers.

Kerr *et al*. (2013) uses NLP techniques to evaluate the answers. This approach also uses pre-graded essays to evaluate the answers.

Pattern based Information Extraction for Automatic Assessment (Saxena and Gupta, 2009) uses Information Extraction, NLP and Pattern Matching Techniques to evaluate the students' answers. It considers both syntax and semantic information. POS is generated for each answer and stored as patterns. Metonymy is found for

each word. This method classifies patterns and update patterns as and when patterns are generated for every answer for matching. If the pattern is matched with the answer then marks are awarded to the answer. It matches with only existing patterns in the knowledge base. This method does not need any pre-graded training essays. The disadvantage is that it matches with the existing patterns in the knowledge base and suits for only short answers.

Kumaran and Sankar (2013) uses concept map and ontology for assessing the students' knowledge. Concept map is created for the student answer and ontology is constructed for the concept map. This ontology is matched with the reference answer ontology by using ontology mapping to evaluate the knowledge of students.

The limitations of automated assessment of answers is that most of the automated systems considers many training essays and answers to evaluate student answer, Human graded essays are considered for automatic evaluation. But the methodology used in this paper uses only one key answer to evaluate the answer.

SYSTEM ARCHITECTURE FOR AUTOMATED ASSESSMENT SYSTEM

The system architecture in the Fig. 1 consists of student answer, key answer, sentence segmentation, POS tagger, sentence similarity computation module and scoring module. The student answers in the text form are stored in the text file. Sentence segmentation module segments the text given in paragraph into individual sentences both in student and key answer. Now the key answer is in the form of key points/sentences. Each key point is associated with a score. Similarly the student answer is also in the form of sentences. POS tagger module extracts the parts of speech of each sentence in the student and key answer and stores them separately. This POS tagging is done with the Stanford Parser.

The sentence similarity computation module uses a sentence similarity algorithm (Madhumitha, Ilango, 2015) compares the key points written by the student with the key point written by the teacher. Each key point in the key answer is compared with all the key points in the student answer, the scoring module gives the score. The sentence similarity algorithm takes the POS extracted by the POS Stanford Parser.

Once the POS are tagged, among all the POS only nouns, verbs, adjectives and adverbs are extracted in all the sentences and stored separately and these are passed into WordNet to extract the synsets of those four POS. WordNet is like a dictionary or thesaurus which is a lexical database of English language. It groups English words into synsets. Synsets are the sets of synonyms for a particular word. This is done separately for each sentence in the student and key answer sentences. After

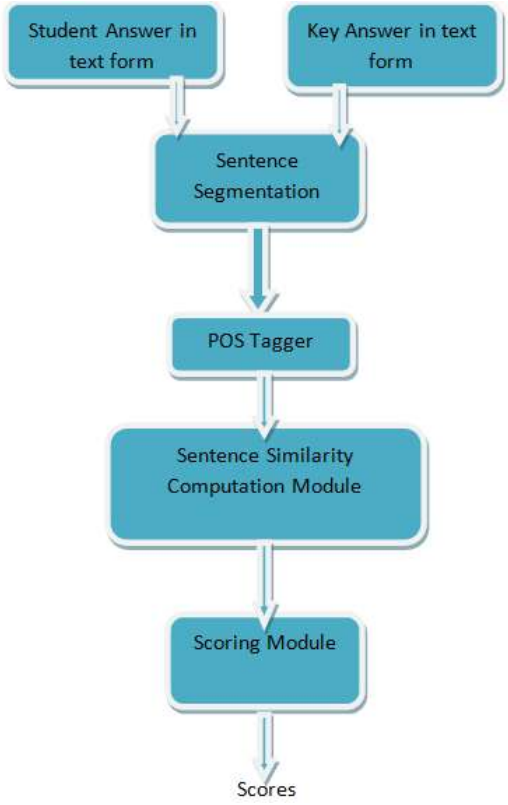

Fig. 1: System architecture for automated assessment system

getting the synsets from the WordNet, the sentence formation of all the above mentioned POS i.e., nouns, verbs, adjectives and adverbs are done and the corresponding sets are formed. The Eq. (1) and (2) explains how the POS sets i.e., noun set, verb set, adjective set and adverb set are formed as separate sets. In Eq. (1) and (2) POS is the Parts of Speech and Syns is the synonyms of those POS. Both the equations form the appropriate set by the union of POS of a sentence (e.g., noun words) and the synonyms of those POS words:

$$POSSet(Sentence1) = \{POS(Sentence1) + Syns(POS(Sentence1)) \quad (1)$$

$$POSSet(Sentence2) = \{POS(Sentence2) + Syns(POS(Sentence2)) \quad (2)$$

These four different sets are considered as separate sets because only if we categorize the sets are noun, verb, adjective and adverb correct similarity of the sentence can be found. The reason is that nouns contribute more to evaluate the sentences, priority next comes to verb, then adverb and last the adjective. Therefore these four POS are separated and weightage are given according to the priority to evaluate the sentences. Each of the POS sets computed using the Eq. (1) and (2) are passed separately into the cosine

similarity to get the corresponding POS Similarity. For example the noun sets of both the sentences are sent to the cosine similarity and the maximum cosine value among the nouns sets of the sentences are taken as the noun similarity. Similarly, the verb, adjective and adverb similarity is computed. The Eq. (3) explains how the POS Similarity is formed:

$$POSSimilarity(POSSim) = MAX \{\frac{POSSet(Sentence1) \bullet POSSet(Sentence2)}{POSSet(Sentence1) \times POSSet(Sentence2)}\} \quad (3)$$

From the above equation *NounSim*, *VerbSim*, *AdjSim and AdvSim* are calculated. Finally Overall Similarity between sentences is computed using the Eq. (4) and (5).

Equation (4) computes the similarity between sentences using nouns and verbs POS i.e., *NounSim*, *VerbSim*. This formula takes two parameter values for nouns and verbs i.e., $\alpha = 0.7$, $\beta = 0.3$,. The parameter values are taken by considering the noun similarity to have more weightage of 70% i.e., $\alpha = 0.7$, verb with 30% i.e., $\beta = 0.3$:

$$OverallSentenceSimilarity = \alpha NounSim + \beta VerbSim \quad (4)$$

Equation (5) computes the similarity between sentences using nouns, verbs, adjectives and adverbs POS i.e., *NounSim*, *VerbSim*, *AdjSim and AdvSim*. This formula takes four parameter values for nouns, verbs, adjectives and adverbs i.e., $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.1$, $\delta = 0.2$. The parameter values are taken by considering the noun similarity to have more weightage of 40% i.e., $\alpha = 0.4$, verb with 30% i.e., $\beta = 0.3$, adjective with 10% i.e., $\gamma = 0.1$ and adverb with 20% i.e., $\delta = 0.2$:

$$OverallSentenceSimilarity = \alpha NounSim + \beta VerbSim + \gamma AdjSim + \delta AdvSim \quad (5)$$

The sentence similarity algorithm computes the similarity between the sentences by Eq. (4) and (5). This similarity is generated for each sentence in key answer with each sentence in student answer and the maximum matched sentence in the student answer with corresponding key point will return the highest score. This matching is computed with the sentence similarity algorithm. If the marks allotted for each key point is 1 and after evaluation using the sentence similarity algorithm, if it gets a similarity score more than 0.5, then the allotted mark is given for that key point. This procedure is repeated for all the key points and the total score is given for the answer.

This architecture is used to evaluate the text based answers of the student for the questions like information recall questions and comprehension questions.

Working of automated assessment system: The model works in the form of the matrix. Consider the key points in the key answer to be $k = \{k_i, k_j, k_{l,....} k_m\}$ and the key points in the student answer be $s = \{ s_i, s_j, s_l ....., s_n \}$. The evaluation of the answers using the cosine similarity measure will be of the form:

	s_i	s_j	s_l	..	..	s_n
k_i	v_{ii}	v_{ij}	v_{il}	..	..	v_{in}
k_j	v_{ji}	v_{jj}	v_{jl}	..	..	v_{jn}
k_l	v_{li}	v_{lj}	v_{ll}	..	..	v_{ln}
..	..	..	..	..	..	..
..	..	..	..	..	..	..
k_m	v_{mi}	v_{mj}	v_{ml}	..	..	v_{mn}

where, $v_{ii,}$ $v_{ij,}$ etc are the values of similarity obtained. Among these values the key point k_i with the maximum similarity value >0.5 with the other student answer key points is considered and then the allotted score for the key point k_i is given. Similarly it is done for all the key points.

DATA SETS

The dataset used for automated assessment of students' answers in this paper is taken from the students' answer scripts written in an educational institution in Coimbatore, TamilNadu. The key answer is the answer written by the teacher for the questions which are used for manual evaluation. The answers with different mark category which belong to information recall and comprehension questions were taken and compared with the two similarity measures. The questions considered for information recall questions evaluation are 'Define rational agent' and 'What is a device controller?', 'List the types of scanners', etc. The questions considered for comprehension questions evaluation are 'Explain in detail about computer organization', etc. The sample key and the students' answers for one question is shown in the Table 1.

Table 1: Sample key and student answers for question 1-what is device controller?

Key answer	Any I/O device connected to the CPU through a controller is called device controller.
Data set 1	It controls the transfer of data from the computer to peripheral device and vice-versa. When an I/O is connected with a controller then it is called as device controller. The entire activity of that device is controlled by the device controller.
Data set 2	Device controller is a device which acts as an interface. It controls the transfer of data from the CPU and any peripheral I/O device.

Table 2: Comparison of key points in key Vs student answers

Sample data set	Key answer key points Vs student answer key points	N-gram similarity value	Cosine similarity value	NV Similarity value	NVAA Similarity value
Data set 1	k1-(a) Vs s1-(a)	0.51	0.59	0.91	0.65
	k1-(a) Vs s1-(b)	0.3	0.32	0.35	0.25
	k1-(b) Vs s1-(a)	0.08	0.17	0.16	0.1
	k1-(b) Vs s1-(b)	0.11	0.25	0.1	0.1

Table 3: Comparison of the total score for the answers

Data sets	Marks allotted for the question	N-gram Score	Cosine score	NV Score	NVAA Score
Data set 1	2	1	1	1	1
Data set 2	2	0	1	2	1
Data set 3	2	0	0	2	2
Data set 4	2	0	0	1	1
Data set 5	2	0	0	1	0
Data set 6	2	0	0	1	1
Data set 7	2	0	1	2	1
Data set 8	2	0	1	2	1

EXPERIMENTAL RESULTS AND DISCUSSION

Comparison of key points in key vs student answer: The sample key answer given Table 1 is compared with the two students answer in i.e., dataset 1 and dataset 2. The key answer in Table 1 consists of two key points namely k1-(a) and k1-(b). The student answer in dataset 1 also consists of two key points namely s1-(a) and s1-(b). Table 2 shows the matrix form of evaluation of answers. Key point k1-(a) is compared with student answer s1-(a) and s1-(b) and the second key point k1-(b) is compared to s1-(a) and s1-(b). The similarity values using n-gram, cosine, proposed NV and NVAA similarity measures is given in the Table 2.

All the answers are evaluated in the above matrix form and similarity values between the key points in the key answer vs. key points in the student answer for the information recall question and comprehension questions are calculated.

Comparison of the total score for the answers: The similarity value of the separate key points for sample dataset using n-gram, cosine, proposed NV, NVAA similarity measure was shown in Table 2. In Table 2 when k1-(a) compared with s1-(a) and s1-(b), s1-(a) gives the value more than 0.5, therefore the allotted mark for the key point i.e., 1 is given. But the second key point k1-(b), when compared to s1-(a) and s1-(b) gives n-gram similarity values less than 0.5, therefore allotted mark is not assigned to k1-(b) key point. Totally the student answer i.e., Data Set 1 gets 1 mark out of 2 using n-gram similarity. Similarly the same Data set 1 is evaluated using cosine, NV and NVAA similarity measures and similarity values are given in Table 2. The total score for the Data Set 1 using cosine similarity measure is 1. The total score for NV and NVAA similarity measure for Data Set 1 is 1 and 1 respectively. The total score for the answers of various data sets taken is given in the Table 3 and Fig. 2 shows the performance comparison of the answers using n-gram, cosine, NV and NVAA similarity measures. The

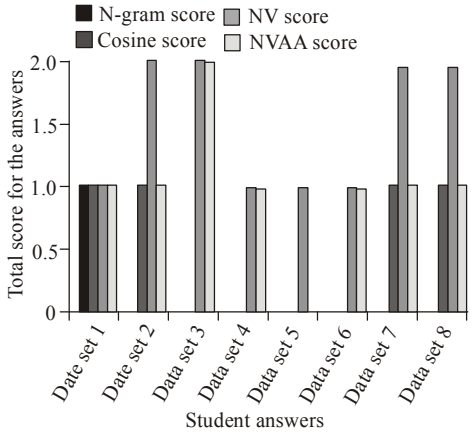

Fig. 2: Performance comparison of the total score for the answers

Table 4: Comparison of the system and human score

Data sets	NV score	NVAA score	Human score
Data set 1	1	1	1
Data set 2	2	1	1.5
Data set 3	2	2	2
Data set 4	1	1	1
Data set 5	1	0	1
Data set 6	1	1	1.5
Data set 7	2	1	1.5
Data set 8	2	1	1.5

performance shows that both the proposed measures outperform the n-gram and cosine similarity measure.

Comparison of the system and human scores with NV and NVAA measures: The comparison between the proposed system score i.e., NV Score, NVAA Score and the human score is shown in Table 4 and the performance comparison is shown Fig. 3.

Correlation between NV and NVAA score and human score: The Pearson correlation and Spearman's correlation between proposed measures and human scores are given in Table 5. The Pearson correlation between the NV score and human score is 0.75, NVAA score and human score is also 0.75. The Spearman's correlation between the NV score and human score is

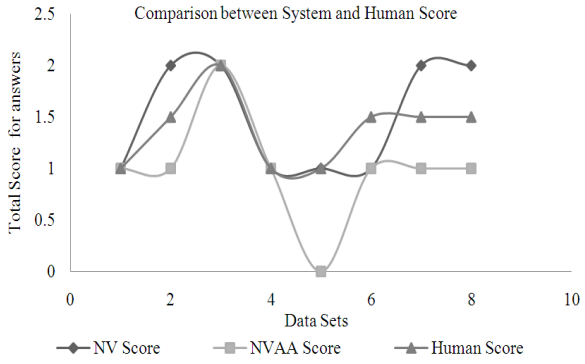

Fig. 3: Performance comparison of the system and human score

Table 5: Correlation between NV, NVAA approaches with human score

Approach/ Measures	Pearson correlation with human scores	Spearman's correlation with human scores
NV	0.75	0.77
NVAA	0.75	0.71

0.77, NVAA score and human score is also 0.715. The Table 5 shows that the proposed system which uses the two novel similarity measures has correlation with the human evaluators.

CONCLUSION

This paper describes the importance of automatic assessment of students' answers. This type of automatic assessment is very important in educational sector. The proposed system which uses NV and NVAA similarity measures outperforms n-gram and cosine similarity measure for evaluating information recall and comprehension questions. The proposed system scores also correlates with human scores using Pearson and Spearman's correlation.

REFERENCES

Alfonseca, E., R.M. Carro, M. Freire, A. Ortigosa, D. Pérez and P. Rodríguez, 2005. Authoring of adaptive computer assisted assessment of free-text answer. Educ. Technol. Soc., 8(3): 53-65.

Bachman, L.F., N. Carr, G. Kamei, M. Kim, M.J. Pan and C. Salvador, 2002. A reliable approach to automatic assessment of short answer free responses. Proceeding of the 19th International Conference on Computational Linguistics. Taipei, Taiwan, 2: 1-4.

Burstein, J., C. Leacock and R. Swartz, 2001. Automatic evaluation of essays and short answers. In: Danson, M. (Ed.), Proceeding of the 6th International Computer Assisted Conference. Loughborough, UK.

Hu, X. and H. Xia, 2010. Automated assessment system for subjective questions based on LSI. Proceeding of the 3rd International Symposium Intelligent Information Technology and Security Informatics (IITSI, 2010), pp: 250-254.

Kanejiya, D., A. Kumar and S. Prasad, 2003. Automatic evaluation of students' answers using syntactically enhanced LSA. Proceeding of the Workshop on Building Educational Applications using Natural Language Processing (HLT-NAACL-EDUC' 03), 2: 53-60.

Kerr, D., H. Mousavi and M. Iseli, 2013. Automatic short essay scoring using natural language processing to extract semantic information in the form of propositions. CRESST Report 831.

Kumaran, V.S. and A. Sankar, 2013. An automated assessment of students' learning in e-learning using concept map and ontology mapping. In: Wang, J.F. and R. Lau (Eds.), ICWL, 2013. LNCS 8167, Springer-Verlag, Berlin, Heidelberg, pp: 274-283.

Madhumitha, R. and K. Ilango, 2015. Parts of speech based sentence similarity computation measures. Int. J. Appl. Eng. Res., 10(21): 20176-20184.

Papineni, K., S. Roukos, T. Ward and W.J. Zhu, 2001. BLEU: A method for automatic evaluation of machine translation. IBM Research Report RC22176 (W0109-022).

Pérez, D., O. Postolache, E. Alfonseca, D. Cristea and P. Rodríguez, 2005. About the effects of using Anaphora Resolution in assessing free text student answers. Proceeding of the International Conference Recent Advances in Natural Language Processing (RANLP, 2005), pp: 380-386.

Questions Skills, year. 6 Categories of Questions. Karen Teacher Working Group. Retrieved from: http://ktwg.org/ktwg_ texts.html.

Saxena, S. and P.R. Gupta, 2009. Automatic assessment of short text answers from computer science domain through pattern based information extraction. Proceeding of the ASCNT, 2009. CDAC, Noida, India, pp: 109-118.

Siddiqi, R. and C.J. Harrison, 2008. On the Automated Assessment of Short Free-text Responses. Retrieved from: http://www.iaea.info/ documents/paper_ 2b711df83.pdf.

Whittington, D. and H. Hunt, 1999. Approaches to the computerized assessment of free text responses. Proceeding of the 3rd CAA Conference. Loughborough University, Loughborough.

Yigal, A., P. Don, F. Marshall, H. Marissa and O. Susan, 2008. Automated scoring of short-answer open-ended GRE subject test items. GRE Board Research Report No. 04-02, ETS RR -08-20, ETS, Princeton, NJ.