Educ Inf Technol (2018) 23:2855–2866
https://doi.org/10.1007/s10639-018-9745-z

Short answer scoring system using automatic reference answer generation and geometric average normalized-longest common subsequence (GAN-LCS)

Feddy Setio Pribadi[1,2] · **Adhistya Erna Permanasari**[1] · **Teguh Bharata Adji**[1]

Received: 28 February 2018 / Accepted: 18 May 2018 / Published online: 19 June 2018

Abstract The Automatic Short Answer Scoring (ASAS) system is one of the tools that can be used to conduct assessment process on e-learning system. One of the methods applied in the ASAS system is a method for measuring similarities between the reference and student answers. There are two issues to be considered in the assessment process using this method. First, this method should be able to provide a variety of reference answers that can handle the diversity of student answers. Secondly, this method should be able to provide an accurate sentence similarity between the reference answers and student answers. Therefore, two methods are proposed to solve both problems. The first method is to generate a variety of reference answers automatically using Maximum Marginal Relevance (*MMR*) method, which obtains an accuracy of 91.95%. The second method is to measure accurately sentence similarity between student answers and reference answers that have significantly different length using GAN-LCS. The performance of the proposed method shows an improvement of the Root Mean Square Error (RMSE) value of 0.884 and a correlation value of 0.468.

Keywords Automatic reference answer generation · Short answer scoring · Maximum marginal relevance · Sentence similarity

✉ Teguh Bharata Adji
adji@ugm.ac.id

Feddy Setio Pribadi
feddy.setio.p@mail.ugm.ac.id

Adhistya Erna Permanasari
adhistya@ugm.ac.id

[1] Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Jl. Grafika No. 2, Sleman, Yogyakarta 55281, Indonesia

[2] Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Semarang, 2nd Floor E11 Building, Sekaran Gunungpati, Semarang 50229, Indonesia

1 Introduction

Automatic short answer scoring (ASAS) system is one of the assessment tools in e-learning process. The system does not only act as a compliment to e-learning system but also reduces the scoring time of a manual system (Pérez and Alfonseca 2005). Automatic scoring can also provide more objective scoring compared to manual corrections (Xi and Liang 2011). According to Burrows (Burrows et al. 2015), as many as 35 ASAS systems were developed until 2015. Howerver, automatic short answer scoring algorithms are unable to deliver equal performance as human does until recently (Shermis 2015; Jayashankar and Sridaran 2017). Thus, a new approach in ASAS system is demanded to improve its accuracy.

On one hand, the accuracy can be obtained using varied reference answers to handle heterogeneity in student answers (Noorbehbahani and Kardan 2011; Rodrigues and Araújo 2012). The heterogeneity of student answers emerge naturally since students construct sentences on their own versions by using synonyms, paraphrasing, and different sentence structures. The varied reference answer generation can be carried out both manually and automatically. An example of manual reference answer generation is using C-rater system, developed by (Leacock and Chodorow 2003; Sukkarieh and Blackmore 2009). C-rater generates several variation of reference answers that lead to the same abstraction. Similar techniques were also demonstrated by (Bachman et al. 2002; Noorbehbahani and Kardan 2011; Senthil Kumaran and Sankar 2015). However, in those research, the generations of reference answer variation were manually performed by evaluators. Unlike the research as mentioned earlier, Mohler (Mohler and Mihalcea 2009) tried to generate reference answer variation automatically from available student answers. The work used Rocchio method to find student answers that have sentence's similarity level closest to the gold standard of reference answer. The one of the proposed methods in this paper is to revise the utilization of Rocchio method to obtain alternative reference answers. The problem is that Rocchio method requires training process that is time-consuming. To solve this problem, this work utilizes Maximum Marginal Relevance (*MMR*) method, proposed in (Carbonell and Goldstein 1998) to generate alternative reference answers automatically.

On the other hand, the accuracy of ASAS system can be increased using improved methods which measure the sentence similarity between student answers and reference answers. Some researchers applied methods that requires the training process as done by (Wolska et al. 2014; Ziai et al. 2012; Mohler et al. 2011), i.e., the system needs to be trained to recognize the patterns of answers before carrying out the actual assessment. There was also a need to develop a knowledge base, such as corpus, that was used as a learning media system as done by (Gomaa and Fahmy 2012; Mohler and Mihalcea 2009; Senthil Kumaran and Sankar 2015). Meanwhile, some methods required sentences with a minimum length of one paragraph (Klein et al. 2011; Adhitia et al. 2009) because the applied method used terms weigthing to measure the similarity of short sentences. A simple method measured directly between two short sentences to calculate the similarity between student answer and reference answer as done in (Noorbehbahani and

Kardan 2011). The research applied Modified Bleu (M-Bleu), in which words having significant roles have higher weights. The problem is that the method is weak in measuring two sentences with different length significantly. Hence, this work proposes a method so-called GAN-LCS (Geometric Average Normalized - Longest Common Subsequence) that removes non contributive words in the reference answer so that it will increase the similarity coefficient between the student answer and the reference answer of different lengths.

2 Related works

The public dataset for ASAS was first published in (Mohler and Mihalcea 2009), but in that paper, Mohler uses only 630 data from a total dataset of 2442 data. In the second publication, Mohler used the entire dataset (Mohler et al. 2011). Some researchers used the dataset published by Mohler including works in (Ziai et al. 2012), (Gomaa and Fahmy 2012), (Sultan et al. 2016). The studies will be described in the following lines.

A research in (Mohler and Mihalcea 2009) tried to develop the ASAS system in attempting to test eight knowledge-based methods and two corpus-based methods. All tested methods were categorized into unsupervised category. The aim was to find the best way that can be applied to ASAS system. The result showed that corpus-based method has better performance compared with the knowledge-based method. The research used correlation value to show the system performance and the resulted correlation value was 0.463. The existing dataset provided only one reference answer. This caused the applied method was not able to handle the diversity of student answers. To improve the performance, Mohler applied the Rocchio method to get alternative reference answers taken from the student answers. Through this process, the reference answers became various so that the sentence similarity result has a better correlation value.

In 2011 (Mohler et al. 2011) an ASAS using a learning engine was also published. The research applied graph method combined with semantic lexical measurement. In this study, Mohler used two parameters to show the performance of the system that is the correlation value (a value obtained from the comparison between human score and machine score) and RMSE value (the square root of the mean square of the different value between human score and machine score). The result shows the increase correlation value from the previous research (Mohler and Mihalcea 2009) that is equal to 0.464 with the RMSE value is equal to 0,978. In the study, Mohler stated that the performance test using the RMSE value is able to measure the proximity value between the score generated by the machine and the score produced by humans (Mohler et al. 2011).

Another research in (Ziai et al. 2012) used natural language processing (NLP) to develop the ASAS system. Application of this method required adequate language tools such as part of speech tagger, stemmer, and Treebank. This research obtained correlation and RMSE values of 0.405 and 1.016, respectively. A recent research involving the same dataset was done in (Sultan et al. 2016) that implements a supervised learning-based method. The system

performance was measured using correlation and RMSE values, which were 0.59 and 0.887, respectively.

3 Dataset

In this paper, we use Texas Corpus as proposed in (Mohler and Mihalcea 2009; Mohler et al. 2011), which is known as Texas Corpus. The dataset consists of 12 assignments, in which each consists of seven to 10 questions, with 87 questions in total. Each question is answered by approximately 30 students and hence the dataset has 2442 student answers. There are two evaluators of which each gives an appropriate score to each student answer. The two scores are then averaged and normalized to a scale from 0 to 5. The highest score is 5 that indicates the answer is 100% correct while 0 indicates otherwise.

On one hand, the average length of the student answers contained in the Texas Corpus dataset is 19 words (the length of a sentence is the number of words that compose the sentence). The shortest sentence has a length of one word and the longest reaches 170 words. This range of length is considered as a type of short answer question as supported by (Siddiqi et al. 2010), which mentioned that student answers generally consist of one phrase up to three or four sentences. This condition is also supported by (Burrows et al. 2015), which explained that student answers comprise one phrase to one paragraph. On the other hand, the student answers having the average length of 19 words in the dataset are also considered as short answers according to which stated in (Sukkarieh and Blackmore 2009) that each student answer consists of maximally 100 words. Thus, this dataset reflects the characteristics of short answer questions. In addition, this dataset can be accessed through web.eecs.umich.edu/~mihalcea/download.html. Some researchers who have used this dataset are (Mohler and Mihalcea 2009; Mohler et al. 2011; Gomaa and Fahmy 2012; Ziai et al. 2012; Senthil Kumaran and Sankar 2015; Sultan et al. 2016).

4 System architecture

Figure 1 illustrates the proposed system architecture in this study. The figure shows that the system is divided into three stages, namely preprocessing, reference answer generation, and sentence similarity measurement using GAN-LCS. The preprocessing stage consists of three processes, namely lowercase, tokenization, and removing punctuation. The preprocessing stage is intended to get the canonical form so that the data is in accordance with the algorithm used in the next process. The second and third stages will be explained in detail in Sections 4.1 and 4.2.

4.1 How *MMR* works to generate reference answers

MMR is developed as an algorithm to perform automatic text summarization (Carbonell and Goldstein 1998) which involves comparison process between each sentence and the document title and between sentences within the document. Sentences with highest similarity scores are put in the top-ranked and marked as selected sentences of the corresponding document. The process of determining variant sentences is the base of

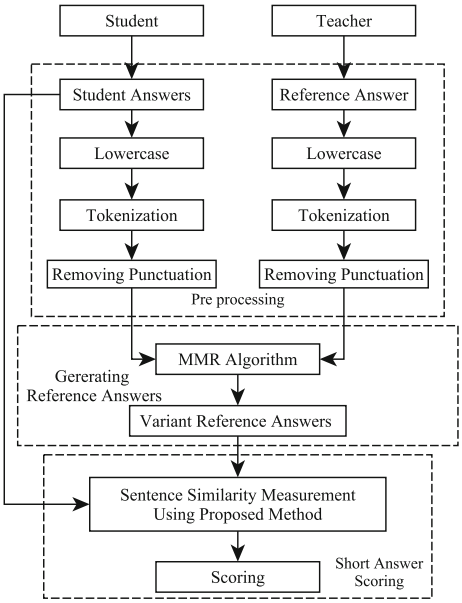

Fig. 1 System architecture

MMR method in choosing appropriate reference answer candidates from the student answers. Eq. (1) is the formula of *MMR*.

$$MMR = \underset{D_i \in R \setminus S}{\operatorname{argmax}} \left[\lambda \left(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right) \right] \tag{1}$$

Each component of *MMR* formula is stated as follows. D_i is the student answers. Q is the reference answer for each question. D_j is the selected student answer with the highest *MMR* value from previous iteration. λ is a constant to adjust the relevance or diversity ranking among the sentences. In this paper, λ is equal to 0.85, which is based on (Carbonell and Goldstein 1998) stating that the closer λ to 1, the more similar the sentence with Q, while the closer λ to 0, the less similar the sentence with Q.

Equation (1) can be used for varying the student answers based on the highest *MMR* value in each iteration process. We only use 3 iterations in this paper and thus will yield in three variant reference answers. Hence, there will be four reference answers including the sentence constructed by the human. *Sim* is a method to measure the similarity between sentences which uses Cosine Coefficient (*CC*) as stated in Eq. (2). R is the reference answers and S is the student answers.

$$\mathrm{CC} = \frac{R \cap S}{\sqrt{R}.\sqrt{S}} \tag{2}$$

The following lines ilustrate the selection of variant reference answers by implementing *MMR* method. Table 1 shows examples of a reference answer (Q) and five student answers (D_i). The candidate variation selection that is based on the

Table 1 Reference answer and student answers

Q	to simulate the behavior of portions of the desired software product
D_1	it simulates the behavior of portions of the desired software product
D_2	program that simulates the behavior of portions of the desired software product
D_3	a prototype program simulates the behaviors of portions of the desired software product to allow for error checking
D_4	a prototype program is used in problem solving to collect data for the problem
D_5	to show that a certain part of the program works as it is supposed to

reference answers starts with a calculation of sentence similarity between D_i and Q in function Sim_1 (a function on the left of negative sign). In the beginning, Sim_2 is 0 as there is no selected sentence candidate (D_j). The first iteration result is the sentence chosen from the student answers with the highest *MMR* value.

Table 2 shows that D_1 has the highest *MMR* value, which is used as D_j. Afterward, the next iteration excludes student answer with the highest *MMR* value which is D_1. The remaining D_i are D_2, D_3, D_4, and D_5.

The second iteration calculates the values of Sim_1 and Sim_2. Sim_2 will not be zero value since $D_j \neq 0$ ($D_j = D_1$). Table 3 shows that D_2 has the highest *MMR* value and thus is assigned as the second D_j. Later, the process continues to the third iteration.

In the third iteration, the process begins with determining D_j between two candidates, D_1 and D_2. The decision is determined by taking the maximum value between $\mathrm{Sim}(D_i,D_2)$ and $\mathrm{Sim}(D_i,D_1)$, as seen in Table 4. $Sim(D_i,D_2)$ is higher than $Sim(D_i,D_1)$, then D_2 becomes D_j. Next, the *MMR* value will be calculated in the same manner as in the second iteration to obtain the third reference answer variation candidate.

Sim_1 measures the similarity level between D_i (D_3, D_4, D_5) with Q, while Sim_2 measures the similarity level between D_i and D_j (D_2). Table 5 shows that D_3 has the highest *MMR* value in the third iteration. As this system is only aimed to generate three reference answer variation candidates, the resulting reference answer variation are D_1, D_2, and D_3. Table 6 shows the final results of the reference answer variation using *MMR* method.

The next process is manual inspection to determine whether the resulting sententes are appropriate as reference answers. Table 7 shows the result and the scores given from the dataset. In this study, we set the threshold for the candidates with score equals 4. The scores of equal or greater than 4 will be accepted, while the scores of less than 4 will be rejected. Note that five (5) is the maximum score of a correct student answer.

Table 2 Result of the first iteration

	Sim_1 (D_i,Q), Sim_2(0)				
	D_1	D_2	D_3	D_4	D_5
Iteration 1	0.5667	0.5376	0.4958	0.1572	0.2271
Iteration 2					
Iteration 3					

Table 3 Result of Second iteration

	$Sim_1(D_i,Q)$, $Sim_2(D_i,D_1)$			
	D_2	D_3	D_4	D_5
Iteration 1				
Iteration 2	0.4111	0.4083	0.1433	0.1870
Iteration 3				

4.2 Short answer scoring using GAN-LCS

Variation of reference answers obtained in the previous process are then used as references to perform the assessment process of student answers. The evaluation process is not only focused on sentence similarity level but is also focused on resulting in the coefficient value, which is $Sim(R,S)$. This coefficient value can be directly converted to final reward for the student as the effort to respond a question. The sentence similarity calculation in the assessment process implements a new method so-called GAN-LCS as formulated in Eq. (3):

$$Sim(R,S) = \frac{2\sqrt{|R||S|}}{|R|+|S|}.\exp\left(\log\left(\frac{lcs_{R\cap S}}{\min(|R|,|S|)}\right)\right) \tag{3}$$

R is the reference answer and S is the student answer. lcs_{RnS} is the longest common subsequence characters between two sentences that are the intersection between the reference answer and the student answer. Equation (3) is developed to deal with the problem of measuring sentences having significantly different lengths that is somewhat different from other available methods. In ASAS system, it is possible that two sentences of different lengths can be considered to be similar. Thus, the *log* part of Eq. (3) is used to omit non contributive words by using the denominator $min(|R|,|S|)$. This will increase the coefficient value of similarity between the reference and student answers of different length.

The coefficient value is then converted into the student grade. For each available reference answer, we calculate the similarity between student and reference answers using Eq. (3) to get the maximum similarity coefficient. The final student's grade is

Table 4 Determining D_j between two candidates, D_1 and D_2

	$Max(Sim(D_i,D_1), Sim(D_i,D_2)$	
	D_1	D_2
Iteration 1	0.5833	
Iteration 2		0.6324

Note that all the iterations as seen in Table 4 is local for *Max(Sim(Di,D1), Sim(Di,D2))*

Table 5 Result of third iteration

	$Sim_1(D_i,Q)$, $Sim_2(D_i,D_2)$		
	D_3	D_4	D_5
Iteration 1			
Iteration 2			
Iteration 3	0.4009	0.1308	0.1764

then obtained from the multiplication between the maximum similarity coefficient and the maximum score, as shown in Eq. (4).

$$Score = \max\left(Sim\left(S_i, R_j\right)*ms\right) \tag{4}$$

$Sim(S_i,R_j)$ is the similarity coefficient between the i-th student answer and j-th reference answer and ms is the maximum score. The maximum score is given as a reward when the student answer is very similar with one of the reference answers. In this study, ms is set to 5 since this value is the maximum score.

5 Result and discussion

The result of the proposed study is aimed to yield three kinds of outputs. The first output is the accuracy level of *MMR* method to obtain reference answer variation. The second output is the correlation value of the manual and automatic evaluation. Meanwhile, the third output is the RMSE value of the manual and automatic evaluation. The correlation and RMSE values are then compared with prior research. Each of the outputs is explained in the following paragraphs.

The first output is the accuracy level of the *MMR* method. The *MMR* method select 261 reference answer candidates from the total 2442 student answers in the dataset. Table 8 shows the accepted or rejected reference answer candidates and the last column is the total amount of the selected candidates per question. The percentage values of the accepted and rejected candidates are also provided.

The reference answer candidate is accepted if equal or greater than 4, while it is rejected if lower. Table 8 shows that the *MMR* method is able to generate reference answer variation with an accuracy of 91.95%. The second and third outputs of the

Table 6 Reference answer variation candidates

	Reference answer variation candidates
D_1	it simulates the behavior of portions of the desired software product
D_2	program that simulates the behavior of portions of the desired software product
D_3	a prototype program simulates the behaviors of portions of the desired software product to allow for error checking

Table 7 Justification of reference answer variation

	Reference answer variation candidates	Human score	Justify
D_1	it simulates the behavior of portions of the desired software product	5	Accept
D_2	program that simulates the behavior of portions of the desired software product	5	Accept
D_3	a prototype program simulates the behaviors of portions of the desired software product to allow for error checking	4	Accept

proposed method are the correlation and RMSE values between human score and machine score. Table 9 shows the correlation and RMSE values.

Table 10 shows the correlation and RMSE results of the proposed method and the previous methods as conducted in (Mohler et al. 2011; Gomaa and Fahmy 2012; (Ziai et al. 2012; Sultan et al. 2016). There were two works using indentical dataset carried out in (Mohler and Mihalcea 2009) and (Senthil Kumaran and Sankar 2015), which resulted in correlation of 0.509 and 0.79, respectively. However, those works only incorporated 630 data of Texas Corpus so that the research cannot be compared with the proposed method.

Based on Table 10, the RMSE value of the proposed method is the best (the lower the better) and the correlation of the proposed method is the third best. In fact, the correlation values of all methods are still in the same range namely moderate category (Evans 1996). The advantages of the proposed method is that the scoring process is done without training process and without using corpus. Meanwhile, the methods in (Mohler et al. 2011; Sultan et al. 2016; Ziai et al. 2012) applied SVM method to conduct the training process, while the methods in (Mohler and Mihalcea 2009) and (Gomaa and Fahmy 2012) required corpus as

Table 8 *MMR* accuracy in obtaining reference answer variation

Question No.	Accepted	Rejected	Total
1	18	3	21
2	19	2	21
3	21	0	21
4	21	0	21
5	11	1	12
6	18	3	21
7	18	3	21
8	19	2	21
9	20	1	21
10	20	1	21
11	28	2	30
12	27	3	30
Total	240	21	261
Percentage (%)	91.95	8.04	

Table 9 The correlation and RMSE values

Assignment	Response	Correlation	RMSE
1	203	0.481	0.889
2	210	0.369	1.051
3	217	0.580	0.781
4	210	0.634	0.972
5	112	0.516	0.825
6	182	0.485	0.669
7	182	0.472	0.801
8	189	0.365	0.952
9	189	0.380	0.842
10	168	0.441	0.880
11	300	0.527	0.950
12	280	0.364	0.999
Average		0.468	0.884

learning medium to extract semantically similar meanings. The proposed method also does not require manually generated reference answer variation, which was mandatory in the methods in (Siddiqi et al. 2010; Senthil Kumaran and Sankar 2015; Noorbehbahani and Kardan 2011).

Two main contributions are proposed in this study. The first is the use of *MMR* method (as formulated in Eq. 1) for automatically generate reference answer variation. This variation will enrich the scoring process compared to other available methods which are merely based on the original reference answer.

The second is a novel sentence similarity so-called GAN-LCS calculation as formulated in Eq. (3). This equation can measure similarity between the student answer and the reference answer of different lengths. However, this equation is yet to cope with the shape of student knowledge diversity as a result of a systemic network explained in (Spiliotopoulou-Papantoniou 2007). This student knowledge diversity emerges from adult's description and the history of human development. This will be a challenge in combining the proposed ASAS method with a systematic network. Firstly, student knowledge categorization is developed to grab the depth of knowledge in the student answers. These student answers of each

Table 10 Performance comparison with other methods

System	Correlation	RMSE
Mohler, 2011	0.464	0,978
Gomaa and Fahmy 2012	0.470	–
Ziai, 2012	0.405	1.016
Sultan, 2016	0.592	0.887
Proposed Method	0.468	0.884

knowledge category are then assessed using our proposed method. Finding the correlation between each category assessment will also be challenging.

However, when the proposed ASAS method will be used for assessing combined categories, then the problem will be how to map each student answer into the appropriate category with the help of available methods such as adding the corpus of each category and applying keyword extraction algorithm.

6 Teaching implication

The use of *MMR* method in this work imitates how teachers assess student answers. Most teachers use a fraction of the student answers that are perceived to obtain good scores as reference answer variation. This will be beneficial to cover the diversity of the student answers. The other benefit is that the *MMR* method is done automatically so that the reference answer variation is generated consistently. This unlikely happens during the manual process of selecting reference answers that is inconsistent due to the subjectivity of each individual teacher. The next benefit is that this automatic reference answers selection for each question only consumes 5.28 milliseconds averagely based on our experiment. This result is very significant since the manual selection of reference answers is time consumed. The other time reduction is coming from the scoring process where the proposed ASAS using GAN-LCS only consumes 52.63 milliseconds which is also very significant. The last benefit is that the proposed ASAS using GAN-LCS method can be implemented without training process and does not require corpus. Hence, the method is easily aplicable in any domain and any language so that it will give high impacts on educational institutions.

7 Conclusion

This study presents that the use of *MMR* method in ASAS system is able to decrease the time consumed for generating reference answer variation. The generation process is done automatically with an average accuracy of 91.95%. The study also present the novel sentence similarity between two sentences of different length using GAN-LCS, which is able to remove non contributive words. The correlation and RMSE values are 0.468 and 0.884, respectively. The RMSE value is the best while the correlation value is the third best among available methods. The most impact to educational institutions is that the proposed method is done without training and without the use of corpus. It will thus ease the implentation on any domain and any language.

In the future, the accuracy of the ASAS can be increased using automatic reference answer generation by implementating other sentence similarity techniques. The techniques may incorporate natural language processing methods, such as semantic network and Part of Speech Tagger so that the techniques are able to identify the sentences with similar meaning and have different structures. Moreover, the future research can be how to capture the diversity of student knowledge using systemic network.

References

Adhitia, R., Purwarianti, A., & Bandung, I. T. (2009). Automated essay grading system using SVM and LSA for essay answers in Indonesian. *Journal of Information Systems, 5*(1), 33–41.

Bachman, L. F., et al. (2002). A reliable approach to automatic assessment of short answer free responses. *Proceedings of the 19th International Conference on Computational Linguistics, 2*, 1–4.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60–117.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98* (pp. 335–336).

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove: Brooks/Cole Pub. Co.

Gomaa, W. H., & Fahmy, A. A. (2012). Short answer grading using string similarity and corpus-based similarity. *International Journal of Advanced Computer Science and Applications, 3*(11), 115–121.

Jayashankar, S., & Sridaran, R. (2017). Superlative model using word cloud for short answers evaluation in eLearning. *Education and Information Technologies, 22*(5), 2383–2402.

Klein, R., Kyrilov, A., & Tokman, M. (2011). Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education* (pp. 158–162).

Leacock, C., & Chodorow, M. (2003). C-rater: automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405.

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09* (pp. 567–575).

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 752–762).

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education, 56*(2), 337–345.

Pérez, D., & Alfonseca, E. (2005). Application of the Bleu algorithm for recognising textual entailments. In *Workshop Recognising Textual Entailment* (pp. 1–4).

Rodrigues, F., & Araújo, L. (2012). Automatic assessment of short free text answers. In *CSEDU 2012 - Proceedings of the 4th International Conference on Computer Supported Education* (pp. 50–57).

Senthil Kumaran, V., & Sankar, A. (2015). Towards an automated system for short-answer assessment using ontology mapping. *International Arab Journal of e-Technology, 4*, 17–25.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment, 20*(1), 46–65.

Siddiqi, R., Harrison, C. J., & Siddiqi, R. (2010). Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies, 3*(3), 237–249.

Spiliotopoulou-Papantoniou, V. (2007). Models of the universe: children's experiences and evidence from the history of science. *Science Education, 16*, 801–833.

Sukkarieh, J., & Blackmore, J. (2009). c-rater: Automatic content scoring for short constructed responses. In *Proceedings of the Twenty-Second International FLAIRS Conference* (pp. 290–295).

Sultan, M. A., Salazar, C., & Sumner, T. (2016). Fast and easy short answer grading with high accuracy. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1070–1075).

Wolska, M., Horbach, A., & Palmer, A. (2014). Computer-assisted scoring of short responses: the efficiency of a clustering-based approach in a real-life task. *Advances in Natural Language Processing*, 298–310.

Xi, Y., & Liang, W. (2011). Automated computer-based CET4 essay scoring system. In *Proceedings - PACCS 2011: 2011 3rd Pacific-Asia Conference on Circuits, Communications and System*.

Ziai, R., Ott, N., & Meurers, D. (2012). Short answer assessment: Establishing links between research strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 190–200).