Arabic Short Answer Scoring with Effective Feedback for Students

Wael Hassan Gomaa
Modern Academy for Computer Science and Management Technology, Cairo, Egypt

Aly Aly Fahmy
Faculty of Computers and Information, Cairo University, Cairo, Egypt

ABSTRACT
In this paper, we explore text similarity techniques for the task of automatic short answer scoring in Arabic language. We compare a number of string-based and corpus-based similarity measures, evaluate the effect of combining these measures, handle student's answers holistically and partially, provide immediate useful feedback to student and also introduce a new benchmark Arabic data set that contains 50 questions and 600 student answers. Overall, the obtained correlation and error rate results prove that the presented system performs well enough for deployment in a real scoring environment.

General Terms
Natural Language Processing, Text Mining

Keywords
Short Answer Scoring, Text Similarity, Semantic Similarity, Arabic Corpus

1. INTRODUCTION
For many years, the learning process has been perceived as a closed circle between teachers and students in terms of assignments, quizzes and exams. The increasing number of students and computer technology entering the educational field; raised the need of an automatic scoring system that can replace teacher in the scoring process , guarantees fairness and saves time. The main concept behind automatic scoring is comparing students answer to model answer. Automatic scoring systems have many forms to adapt with all types of curriculums and forms of students answers as writing, speaking and mathematics. Writing assessment is either handled by Automatic Essay Scoring (AES) or Short Answer Grading Systems. Speaking assessment includes low and high entropy spoken responses while mathematical assessments include textual, numeric or graphical responses. The simplest kind of Automatic Scoring Systems in terms of implementation and design are the ones designed for questions as Multiple Choice, True-False, Matching and Fill in the blank. Implementation of AS systems that are meant to scoring essay questions a difficult and complicated task as student's answers require text understanding and analysis.

This research presents a system for short answer scoring in Arabic language. Arabic is a widespread language that is spoken by approximately 300 million people around the world. From a natural language point of view, the Arabic language is characterized by high ambiguity, rich morphology, complex morpho-syntactic agreement rules and a large number of irregular form. A new benchmark Arabic data set that contains 600 students' short answers is presented in this article. The system considers two types of text similarity techniques -String Similarity and Corpus Similarity- that were used separately and combined. Holistic and Partitioning models were examined. Holistic model measures the similarity between the complete form of student answer and model answer without partitioning for students answers. Partitioning model partitions student answer to a set of sentences automatically by using sentences boundary detection templates; then it maps each sentence to the highest similarity element of model answers. A key point of this research is a feedback module that gives the students automatic and immediate useful comments. Experiments showed high accuracy of the feedback module by calculating the human-system agreement rate. This paper is organized as follows: Section 2 presents related work of the main automatic short answer grading systems. Section 3 presents the Arabic data set used for benchmarking the short answer scoring systems. Section 4 introduces the two main categories of Similarity Algorithms used in this research. Section 5 shows experiments' results. Section 6 introduces the feedback module and finally section 7 presents conclusion of the research.

2. RELATED WORK
The system introduced in [1] is the most closely related research to our work as it was the only research that handled the Arabic language ; Environmental Science data set was created as part of the research, it contained 61 questions, 10 answers for each, with a total number of 610 answers. Many aspects were introduced that depend on translation to overcome the lack of text processing resources in Arabic, such as extracting model answers automatically from an already built database and applying K-means clustering to scale the obtained similarity values. The system scored each student's answer with 536 different automatic runs: 256 of the runs used String-Based Similarity, 64 used Corpus-Based Similarity, and the other 216 used Knowledge-Based Similarity measures. For each run, the Pearson Correlation Coefficient (r) and the Root Mean Square Error (RMSE) were computed. Combining the measures from different categories achieved r = 0.83 and RMSE = 0.75. These resulting values were very close to the values that were scored manually by two annotators. Concerning the English language, many scoring systems were applied as a text-to text similarity task in which the score is assigned according to a measure of the lexical and semantic similarity between a student answer and a model answer when using several measures, including string-based, knowledge-based and corpus-based [2,3,4]. Powergrading system [5] was based on training a similarity metric between student responses and then using this metric to group responses into clusters and subclusters. An excellent and more detailed overview of related work can be found in [6,7], such as CarmelTC [8], C-Rater [9], Intelligent Assessment Technologies (IAT) [10] and Oxford-UCLES [11]. A substantial amount of work has recently been performed in short-answer grading at the SemEval-2013 task #7:The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge [12].This task offered three problems: a 5-way task, with 5 different answer judgments, and 3-way and 2-way tasks, which conflate more judgment

categories each time. Two different corpora, Beetle and SciEntsBank, were labeled with the5 following labels: Correct, Partially correct incomplete, Contradictory, Irrelevant and Non Domain, as described in [13].

3. DATA SET

In this research, a new Arabic data set that can be used as benchmark for short answer grading is developed. Questions presented in the data set cover the official curriculum for Philosophy course. The data set is available in XML format and contains 50 questions with 12 answers per each with total number of 600 answers. Average length of a student's answer is 2.5 sentences, 24 words or 130 characters. It contains a collection of students' answers and grades which were scored by two specialists who gave marks within the range of 0 and 10 and obtained Pearson correlation coefficient and Root Mean Square Error of 0.87 and 0.73, respectively. Figure 1 shows the student's marks distribution. Model answer for each question is divided to set of elements; each element may contain Section(s) and Sub Section(s) with certain mark for each as shown in table 1. Assigning a certain mark for each section and subsection helps in: achieving justice in the manual scoring process, providing two ways of Automatic Scoring either by comparing Student Answer to Model Answer as a whole or partially and finally providing useful feedback to students depending on the description of each Section and Sub Section.

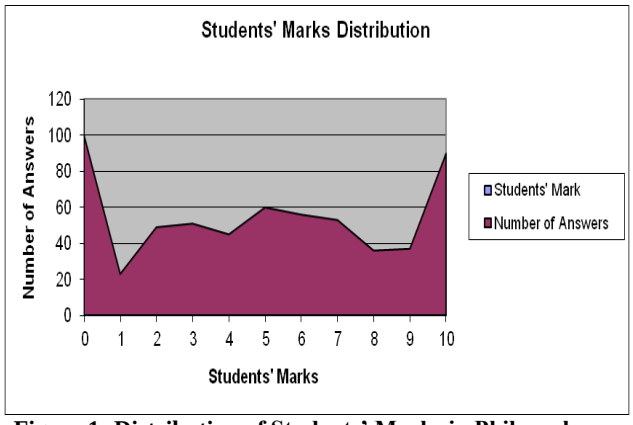

Figure 1: Distribution of Students' Marks in Philosophy data set

4. TEXT SIMILARITY MEASURES

The experiments are centered around the use of measures of text similarity for automatic short answer scoring. Fourteen String-Based and two Corpus-Based similarity algorithms were experimented through two models. The first model (Holistic Model) measures the similarity between the complete form of student answer and model answer without dividing the student answer and ignoring the partition scheme of model answer. The second model (Partitioning Model) automatically divides student answer into set of sentences using sentences boundary detection templates based on regular expression, then it maps each sentence to the highest similarity element of model answers.

String similarity measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison. Fourteen algorithms are used in our experiments ; Seven of them are character based –Longest Common SubString, Damerau, Jaro, Jaro Winkler, Needleman Wunch, Simth Waterman and N-gram - while the other are term-based distance measures – Block Distance, Cosine Similarity, Dice's Coefficient, Eclidean Distance, Jaccard Similarity, Matching Coefficient and Overlap Coefficient -. An excellent and more detailed overview of string-based similarity measures can be found in [14].

Table 1: A Sample Question with its Model Answer in Philosophy data set

```
<Questions>
        <Question_Text>وضح أهمية الفلسفة بالنسبة للفرد؟</ Question_Text>
        <Full_Mark>10</Full_Mark>
        <Section>
                <Section_Description>عنصر تعميق الوعى</ Section_Description>
                <Section_Text>تعميق الوعى لدى الفرد</Section_Text>
                <Section_Mark> 3 </Section_Mark>
                <SubSection>
                        <SubSection_Description>شرح عنصر تعميق الوعى</ SubSection_Description >
<SubSection_Text>
تجعل الإنسان يفهم حياته و يدرك مكانته فى الوجود و فى المجتمع و تحدد أهدافه و توقظه من نومه العميق
</ SubSection_Text>
                        <SubSection_Mark> 2 </SubSection_Mark>
        </SubSection>
        </Section>
        <Section>
                <Section_Description>عنصر الإرتقاء بالمستوى العقلى</ Section_Description>
                <Section_Text> الإرتفاع بالمستوى العقلى و حل المشكلات</Section_Text>
                <Section_Mark> 3 </Section_Mark>
                <SubSection>
                        <SubSection_Description>شرح عنصر الإرتفاع بالمستوى العقلى</ SubSection_Description>
                        <SubSection_Text>
            لأنها تعتمد على التفكير العقلى و دراسة وجهات نظر الفلاسفة فى مختلف المشاكل و تساعد الفرد فى حل مشاكله الخاصة
</ SubSection_Text>
                        <SubSection_Mark> 2 </SubSection_Mark>
                </SubSection>
        </Section>
</Question>
```

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. Extracting DIStributionally similar words using CO-occurrences (DISCO) package is used in our experiments[15]. DISCO represents the distributional similarity between words assumes that words with similar meaning occur in similar context. Large text collections are statistically analyzed to get the distributional similarity. DISCO is a method that computes distributional similarity between words by using a simple context window of size ±3 words for counting co-occurrences. When two words are subjected for exact similarity DISCO simply retrieves their word vectors from the indexed data, and computes the similarity according to Lin measure. If the most distributionally similar word is required; DISCO returns the second order word vector for the given word. DISCO has two main similarity measures DISCO1 and DISCO2; DISCO1 computes the first order similarity between two input words based on their collocation sets. DISCO2 computes the second order similarity between two input words based on their sets of distributionally similar words.

The automatic score is then calculated in two steps. First, the similarity between the student and model answers is measured using the text similarity measures described above. Second, the obtained similarity values (0-1) are scaled onto the original scale (0-10) for ease of comparison. The scaling step is essential and is highly related to the system accuracy evaluation. IsotonicScale method [1] was applied to all the experiments. This method simply trains an isotonic regression model on each type of system output. 10-fold cross validation was used for all of the experiments; 9/10th of the average marks given by the two annotators were used to acquire the

training data, and 1/10th was used for evaluation. The average performance over the 10 experiments is reported.

The system accuracy is evaluated by comparing the manual and automatic scores while considering two factors, which are the association and the error size. The association is measured by the Pearson correlation coefficient (r) to indicate the correctness of the score ranking. The error size is measured by the Root Mean Square Error (RMSE) to characterize the precision of automatic score prediction. To evaluate the system output, especially using RMSE, it is necessary for the obtained similarity values and the annotators' marks to be on the same (0-10) scale.

5. EXPERIMENTS RESULTS AND ANALYSIS

5.1 String-Based Similarity Measures

Fourteen String-Based similarity algorithms mentioned above were experimented through the Holistic and Partitioning Models. Four methods were used to deal with strings in model and students answer; Raw, Stop, Stem, StopStem. The similarity in Raw method is computed without applying any NLP task. In the Stop method; Stop Words Removing is applied using stop list that contains 387 words. Information Science Research Institute (ISRI) Arabic Stemmer [16] is used to replace each non-stop word with its stem without removing the stop words. Both Stop Words Removing and Stemming tasks were applied in StopStem method. Tables 3 and 4 represent experimental results of String-based measures applied to Arabic text through the holistic and partitioning models respectively.

The results showed that applying stop word removing task separately or merged with the stemming task is better than applying the stemming task separately. Also partitioning model achieved better results than holistic model in all cases although simple sentence boundary detection templates were used. Additionally, the results showed that several measures appear to be better when evaluating with the r correlation measure while others appear to be better when measuring with the RMSE. The best value of the r correlation was 0.826, which resulted from the N-gram character-based approach using the average results from both Bi-gram and Tri-gram methods. Although this correlated value is very promising, by scanning the error rate, we found that the best RMSE 1.16 resulted also from the N-gram character-based method applied to partitioning model, and it is not promising to rely on string-based similarity in a real scoring system.

5.2 Corpus-Based Similarity Measures

The two Corpus-Based similarity Disco1 and Disco2 were applied using Arabic Disco corpus explained in table 2.

Table 2. Disco Data Packet

Language	Arabic
Corpus-Size	518 Megabytes
Number of Tokens	188 million tokens
Number of queriable words	134,000
Source	Arabic Wikipedia Ajder Corpora

The similarity is calculated either by Max similarity (MaxSim) or Average similarity (AvgSim). MaxSim is the highest similarity value between a given word w and all of the words in the student answer. AvgSim is calculated by dividing the sum of the similarity values of a given word w by the number of words in the student answer. The experiment results emphasize that Corpus-Based algorithms give less error rate values. This finding was clear in table 5, using the MaxSim method clearly enhanced the correlation and error rate results in all of the cases in Corpus-Based measures. DISCO2 achieved the best correlation value, 0.852, using the partitioning model, and DISCO2 resulted in the best and most promising RMSE value, which is 0.84. Although correlation values that result from string-based measures are in the range of values that result from corpus-based measures, the RMSE values for the corpus-based measures were much less than the string-based measure values. In addition, the partitioning model achieved better results than holistic model in all cases. These findings confirm the important role of corpus-based measures as a type of semantic similarity approach.

Table 3. Experimental results of String-based measures applied to Arabic text through the holistic model

	Raw		Stop		Stem		StopStem	
	r	**RMSE**	**r**	**RMSE**	**r**	**RMSE**	**r**	**RMSE**
Character-Based Distance Measures								
LCS	0.405	1.31	0.454	1.28	0.405	1.29	0.474	1.27
DL	0.671	1.25	0.691	1.22	0.681	1.24	0.695	1.23
Jaro	0.413	1.31	0.426	1.26	0.393	1.27	0.333	1.27
Jaro-Winkler	0.412	1.30	0.427	1.26	0.394	1.27	0.322	1.28
Needleman-Wunsch	0.555	1.26	0.577	1.23	0.551	1.23	0.58	1.23
Smith-Waterman	0.312	1.33	0.322	1.28	0.286	1.29	0.32	1.29
N-gram (Bi+Tri)	0.764	**1.18**	**0.781**	**1.18**	0.747	1.21	0.764	1.21
Term-Based Distance Measures								
Block Distance	0.635	1.26	0.668	1.23	0.715	1.23	0.751	1.22
Cosine similarity	0.608	1.26	0.62	1.22	0.692	1.21	0.719	1.21
Dice's coefficient	0.627	1.25	0.648	1.22	0.709	1.21	0.74	1.20
Euclidean distance	0.613	1.25	0.633	1.22	0.697	1.20	0.715	1.20
Jaccard similarity	0.591	1.26	0.607	1.23	0.679	1.21	0.70	1.22
Matching Coefficient	0.63	1.26	0.667	1.22	0.674	1.22	0.749	1.21
Overlap coefficient	0.465	1.31	0.411	1.29	0.502	1.29	0.497	1.30

Table 4. Experimental results of String-based measures applied to Arabic text through the partitioning model

	Raw		Stop		Stem		StopStem	
	r	**RMSE**	**r**	**RMSE**	**r**	**RMSE**	**r**	**RMSE**
Character-Based Distance Measures								
LCS	0.469	1.28	0.509	1.24	0.461	1.24	0.530	1.22
DL	0.751	1.20	0.764	1.18	0.753	1.20	0.767	1.18
Jaro	0.483	1.27	0.488	1.23	0.455	1.24	0.395	1.23
Jaro-Winkler	0.482	1.26	0.489	1.22	0.456	1.23	0.384	1.26
Needleman-Wunsch	0.619	1.22	0.633	1.20	0.627	1.25	0.636	1.21
Smith-Waterman	0.376	1.29	0.378	1.26	0.342	1.24	0.376	1.26
N-gram (Bi+Tri)	0.818	**1.16**	**0.826**	1.17	0.804	1.18	0.821	1.18
Term-Based Distance Measures								
Block Distance	0.695	1.23	0.73	1.20	0.767	1.21	0.803	1.21
Cosine similarity	0.668	1.22	0.672	1.19	0.744	1.18	0.771	1.19
Dice's coefficient	0.691	1.20	0.704	1.20	0.765	1.21	0.796	1.18
Euclidean distance	0.677	1.20	0.688	1.20	0.753	1.21	0.771	1.18
Jaccard similarity	0.655	1.22	0.663	1.20	0.735	1.23	0.756	1.19
Matching Coefficient	0.694	1.23	0.743	1.21	0.73	1.24	0.804	1.19
Overlap coefficient	0.529	1.27	0.467	1.27	0.559	1.26	0.553	1.26

Table 5.. Experimental results of Corpus-based measures

Algorithm	Holistic Model				Partitioning Model			
	MaxSim		AvgSim		MaxSim		AvgSim	
	r	RMSE	r	RMSE	r	RMSE	r	RMSE
Disco1	0.841	0.90	0.433	1.02	0.854	0.86	0.532	0.99
Disco2	0.848	0.87	0.422	1.02	**0.852**	**0.84**	0.560	0.98

5.3 Hybrid Approach

Combining the similarity values of String-Based and Corpus-Based was applied in a supervised way by learning the obtained student marks through three models using Weka [17]. These models are Simple Linear Regression, Linear Regression and SMOreg. SMOreg is a sequential minimal optimization algorithm for training a support vector regression. To perform the training and testing tasks, 10-fold cross-validation is used for all of the experiments. We have submitted two methods, called CombineALL and CombineBest, for a hybrid task. The CombineALL method is performed by training on all of the obtained marks from all of the runs that were tested separately. In the CombineBest method, we choose the best measures that outputted best correlation and error rate results. Four algorithms were selected for CombineBest method; we choose the best measures that outputted best correlation and error rate results. For the String-based method, the character-based N-gram and the term-based Block Distance measures are selected. In the Corpus-based similarity measures, all of the runs of Partitioning model are selected. Table 6 represents the experiment results of the proposed hybrid models. These results indicate the benefit of combining multiple similarity measures by enhancing both the correlation and the error rate values. The best r value, 0.862, resulted from both Linear Regression and SMOreg models applied to the CombineBest method. The best RMSE value, 0.76, resulted from the SMOreg model applied to the CombineBest method.

Table 6: Experimental results of combining String-Based and Corpus-Based for Philosophy data set.

Combining Models	CombineALL		CombineBest	
	r	RMSE	r	RMSE
Simple Linear Regression	0.852	0.81	0.854	0.80
Linear Regression	0.846	0.80	0.848	0.78
SMOreg	0.861	0.77	**0.862**	**0.76**

An interesting research point is to benefit from the combination of similarity algorithms in reducing the total required time of measuring the automatic mark. Corpus-based similarity algorithms suffer from the large time needed to search for the relationship between given two words in a large corpus and then constructing the similarity matrix as explained previously. In the proposed system combination task was implemented so that; N-gram string-based algorithm was used to map each sentence in student's answer to each sentence in model's answer and then Corpus-based similarity algorithm was applied between the corresponding two sentences in the partitioning model. This task will reduce the required time by canceling the module of measuring the similarity between each sentence in student's answer and each sentence in model answer using Corpus-based similarity. Another advantage is the ability of using multithreading approach by measuring couples of sentences in the same time.

Table 7: Elapsed Time Comparison

Similarity Algorithm	Elapsed time of 600 Answers in minutes	Elapsed time of average one student answer
Corpus-based	720	1.2
Corpus-based and String-based	210	0.35
Corpus-based and String-based with Multithreading	125	0.208

Experiments were performed using a laptop with processor Intel Core i7, 1.6 Ghz and 8 GB Ram; table 7 presents a comparison between the elapsed time to obtain the automatic mark of 600 students' answers using Corpus-based algorithms separately and combined with N-gram string-based similarity. Combining the two types of similarity reduced the elapsed time to the sixth which is considered a real achievement. Also this combination paved the way to multithreading approach which accordingly decreased the elapsed time.

6. FEEDBACK

A very important point for students is to know how far are they going with their answers. Standing on the good points of students answers gives students a sense of achievement, which motivates them to learn more. Also knowledge of weak points help students take corrective actions for better education. Hence, it is essential for teachers to monitor students' learning and give them feedback. Another dimension of feedback is its immediacy. The longer the time gap between the completion of the work and its feedback, the less effective the feedback becomes. Ideally, feedback should be provided immediately after the completion of a task. The importance of generating automatic and immediate feedback was behind designing a module for feed back in the proposed system.

As previously mentioned, the model answer for each question is divided to set of elements; each may contain Section(s) and Sub Section(s) with certain mark for each. This scheme helped in providing useful feedback to students depending on the description of each Section and Sub Section. This research presents automatic feedback by generating a comment and optionally displaying the model answer within the automatic mark for each element. Four statements were chosen to comment the student's answers; الإجابة خاطئة ، الإجابة ناقصة ، الإجابة شبه صحيحة ، الإجابة صحيحة literally Wrong Answer, Incomplete Answer, Semi-Correct Answer and Correct Answer. The key point in the automatic feedback task is how to select the threshold similarity value of each of the four statements. The simple and accurate one-dimensional K-means clustering algorithm was applied to cluster the automatic marks into four categories or clusters. The maximum similarity value of each cluster was considered as a threshold of each category [18,19].

To calculate the accuracy of feedback module; a specialist manually mentioned a feedback statement to each element in student's answer considering that a blank answer is a wrong answer. Then the human-system agreement rate was calculated to represent the overall accuracy; human-system agreement rate for each feedback type is calculated by dividing the number of judgments where the human and system agree by the total number of judgments. The obtained results are encouraged specially for extreme "Wrong Answer" and "Correct Answer" types. Table 8 presents the range of similarity values for each feedback statement and the human-system agreement rate.

Table 8: Feedback Module

Feedback Statement	Similarity Values	Human-System Agreement Rate
Wrong Answer	0 to 0.27	95 %
Incomplete Answer	0.28 to 0.60	88 %
Semi-Correct Answer	0.61 to 0.82	89 %
Correct Answer	0.83 to 1	**97 %**

7. CONCLUSION

A new short answer benchmarking data set called Philosophy was presented in this research; it contains 50 questions with 12 answers per each with total number of 600 answers. Model answer for each question is divided to set of elements, each element may contain Section(s) and Sub Section(s) with certain mark for each. Assigning a certain mark for each section and subsection helped in scoring either by comparing Student Answer to Model Answer as a whole or partially and finally providing useful feedback to students depending on the description of each Section and Sub Section. Fourteen String-Based and two Corpus-Based similarity algorithms were experimented through two models. The first model (Holistic Model) measures the similarity between the complete form of student answer and model answer without dividing the student answer and ignoring the partition scheme of model answer. The second model (Partitioning Model) automatically divides student answer into set of sentences using sentences boundary detection templates based on regular expression, then it maps each sentence to the highest similarity element of model answers. partitioning model achieved better results than holistic model in all cases although simple sentence boundary detection templates were used. Combining multiple similarity measures enhanced both the correlation and the error rate values. An interesting research point was to benefit from the combination of similarity algorithms in reducing the total required time of measuring the automatic mark. Applying String-based measures to map each sentence in student answer to each element in model answer obtained similarity reduced the elapsed time to the sixth which is considered real achievement. Also this combination paved the way to multithreading approach which accordingly decreased the elapsed time. Providing students with useful feedback was introduced; this module was performed by selecting four thresholds according to K-means clustering. These thresholds defined the range of each type of feedback comments. The accuracy of feedback module was promising specially for the two extremes "Wrong Answer" and "Correct Answer" types. In conclusion, we presented two case studies which results indicated that the presented system walks along the same path as manual evaluation.

8. REFERENCES

[1] Gomaa, W. H., & Fahmy, A. A. (2013). Automatic scoring for answers to Arabic test questions. Computer Speech & Language. http://dx.doi.org/10.1016/j.csl.2013.10.005

[2] Gomaa, W.H. & Fahmy, A. A. (2012). Short Answer Grading Using String Similarity and Corpus-Based Similarity. In International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No.11.

[3] 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 (pp. 567-575). Association for Computational Linguistics.

[4] Mohler, M., Bunescu, R. C., & Mihalcea, R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In ACL (pp. 752-762).

[5] Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading, In Transactions of the ACL (TACL).

[6] Gomaa, W.H. & Fahmy, A.A. (2011). Tapping Into The Power of Automatic Scoring. the eleventh International Conference on Language Engineering, Egyptian Society of Language Engineering (ESOLEC '2011).

[7] Ziai, R., Ott, N. & Meurers, D. (2012). Short answer assessment: establishing links between research strands. In: Proceedings of the Seventh Workshopon Building Educational Applications Using NLP, June. Association for Computational Linguistics, pp. 190–200.

[8] Rosé, C. P., Roque, A., Bhembe, D., & VanLehn, K. (2003). A hybrid approach to content analysis for automatic essay grading. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2 (pp. 88-90). Association for Computational Linguistics.

[9] Leacock, C., Chodorow, M. (2003). C-Rater: automated scoring of short-answer questions. Computers and the Humanities 37 (4), 389–405.

[10] Mitchell, T., Russell, T., Broomhead, P.& Aldridge, N. (2002). Towards robust computerized marking of free-text responses. In: Proceedings of the Sixth International Computer Assisted Assessment Conference. Loughborough University, Loughborough, UK.

[11] Pulman, S.G.& Sukkarieh, J.Z. (2005). Automatic short answer marking. In: Proceedings of the Second Workshop on Building Educational Applications Using NLP, June. Association for Computational Linguistics, pp. 9–16.

[12] Dzikovska, M.O., Nielsen, R.D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., et al. (2013). SemEval 2013 task 7: the joint student response analysis and 8th recognizing textual entailment challenge. In: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval2013), in Conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013), Atlanta, Georgia, USA, June. Association for Computational Linguistics.

[13] Dzikovska, M.O., Nielsen, R.D. & Brew, C. (2012). Towards effective tutorial feedback for explanation questions: a dataset and baselines. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June. Association for Computational Linguistics, pp. 200–210.

[14] Gomaa, W.H., Fahmy, A.A. (2013). A Survey of text similarity approaches. International Journal of Computer Applications 68 (13), 13–18.

[15] Kolb, P. (2009). Experiments on the difference between semantic similarity and relatedness. In Proceedings of the 17th Nordic Conference on Computational Linguistics NODALIDA'09.

[16] Taghva, K., Elkhoury, R., & Coombs, J. (2005, April). Arabic stemming without a root dictionary. In Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on (Vol. 1, pp. 152-157). IEEE.

[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

[18] Mostafa, M.S., Haggag M. H. & Gomaa, W.H. (2008). Document Clustering using Word Sense Disambiguation, In proceeding of: 17th International Conference on Software Engineering and Data Engineering (SEDE 2008), June 30 - July 2, 2008, Omni Los Angeles Hotel at California Plaza, Los Angeles, California, USA.

[19] Wang, H. & Song, M. (2011). Ckmeans : optimal k means clustering in one dimension by dynamic programming. The R Journal 3, 29–33.