

# An Automatic Arabic Essay Grading System based on Text Similarity Algorithms

Abdulaziz Shehab, Mahmoud Faroun, Magdi Rashad

Information Systems Dept.

Faculty of Computers and Information, Mansoura University, Egypt

**Abstract**—Manual grading process has many problems such as time consuming, costly, enormous resources, lot of efforts and huge pressure on instructors. These problems place the educational community in a dire need to have auto-grading systems so as to address these problems associated with manual grading. Auto-grading systems are wide spread over the world because they play a critical role in educational technology. Additionally, the auto-grading system is introducing many advantages as it saves cost, effort and time in comparison to manual grading. This research compares the different algorithms used in automatic grading systems in Arabic languages using string and corpus algorithms separately. This process is a challenging task following the necessity of inclusive assessment to authenticate the answers precisely. Moreover, the challenge is heightened when working with the Arabic language characterized by complexities in morphology, semantics and syntax. The research applied multiple similarity measures and introduces Arabic data set that contains 210 students' answers. The results obtained prove that an automatic grading system could provide the teacher with an effective solution for article grading systems.

**Keywords**—Auto-grading systems; string-based similarity; corpus-based similarity; N-Gram

## I. INTRODUCTION

Nowadays, auto-grading systems are a very important scientific topic within the educational community. The huge amounts of tests and students have brought about the necessity for automatic grading systems. Therefore, researchers pay more attention to such systems which have become a vital component in the educational community because it is capable of reducing the load of the teaching staff due to graded students' exams. These systems have witnessed a continuous development in the last few years [1]-[5]. In fact, most of the teachers' time is wasted due to manual grading. Teaching staff all over the world suffer from wasted time spent on students' essay marking.

There is an increase in the number of student enrolment that makes the marking operation more difficult. Essay Questions are more difficult than other objective questions because it takes much time to mark these questions. But, the essay questions are preferable than objective questions because it makes the cheating process more difficult and raises the student's skills in writing.

In addition, the essay marking can differ from human grader and others in marks, which is unfair. Therefore, the auto-grading systems could automatically grade the student

answers and teachers could utilize the time wasted for marking the essays. In this way, teachers get an opportunity to concentrate on other critical assignments, for example, develop more viable instructional materials and other correspondence capacities.

Essay-typed questions are classified into two main types: short and long answers. For the first, student solutions are written in short sentences. For the latter, it gives the students the freedom to write as long as he or she wants. Hence, teachers look for special characteristics to be graded, for example, style of writing, mechanics, and content [6]. The short answers assessment depends on the content of answers and the style is not significant [7].

Auto-grading system has one of the most convoluted tasks as it depends on the overall semantic meaning, since short answers have many common words [8]. Moreover, the similarity on long essay is complicated to discover as every word could have other synonyms, meanings [9], [10]. The marks could be given to the students based on similar sentences that are prepared in the marking scheme. The majority of automated short answer grading systems are done in English ignoring Arabic because there are many challenges in Arabic those needs to be tackled in different ways. Arabic context has unique characteristics and features so it is challenging task.

In the last few years, many researchers have proposed a number of automated Arabic short answer grading systems [11]-[14]. They proposed many algorithms like Latent Semantic Analysis (LSA) [15], Damera-Levenshtein (DL) [16], [17], N-Gram [18], Extracting Distributionally similar words using Co-occurrences (DISCO) [19], [20]. The main idea of automatic grading systems is using n-grams in different applications [21]-[23]. More recently, there are approaches that used LSA model to evaluate written answers automatically [21], [22], [24]-[26]. However, their proposals lack a comparative study that deeply shows the advantages and disadvantages of such algorithms.

Dependently, in this paper, a comparative study between different approaches that are oriented for automatic grading is presented. The main goal of this paper is to find the best suitable technique suitable for Arabic language essay questions. So, we applied four algorithms on the same dataset of questions to compare them and find the optimal algorithm which gives the best correlation between manual grading and automatic grading. The remainder of this paper is as follows: Section II discusses related works; Section III presents the

used dataset. Section IV presents a comprehensive view on string-based and corpus-based while Section V presents the proposed architecture. The experimental results are shown in Sections VI and VII is our conclusion and future work.

## II. RELATED WORKS

There are several automatic grading systems conducted in English such as E-rater [27], Automark [28], C-rater [29], and LVQ [30]. On the other hand, a few studies have been done in Arabic [4], [11], [12].

Gomaa and Fahmy [11], one of the most recently published works, present a system that scores Arabic short answers. They prepared a dataset which contains 610 student answers written in the Arabic language. Their proposal evaluates the student answers after they have been translated into English. Their objective was to overcome the challenges that were faced in Arabic text processing. However, the proposed system has many problems namely the missing of utilization of good stemming techniques, the translation from Arabic into English causes the loss of context structure where many words in Arabic are not semantically translated, and the results obtained have to be passed to a machine learning that demands a high processing time.

Mezher and Omar [12] proposed a hybrid method based on a modified LSA and syntactic feature in a trial for automatic Arabic essay scoring. They relied on the dataset proposed in [11]. Their proposal focused on part of speech (POS) tagging in a way to identify the syntactic feature of words within the similarity matrix. Their study sought to resolve the drawbacks of standard LSA which laid emphasis on the limited syntactic analysis. However, utilizing only LSA technique in their study did not guarantee a high correlation ratio.

Emad [13] presents a system based on stemming techniques and Levenshtein edit operations for evaluating student's online exams. The proposed system was mainly based on the capabilities of light and heavy stemming. The dependence only on the string-based algorithm (Levenshtein) is counted as one of the main defects in his study as it ignores corpus-based algorithms that support semantic similarity.

Rashad et al. [31] proposed an Arabic online examination environment that provides an easy interaction between students and instructors. In fact, their proposal is only developed for grading objective (non-essay) questions. For essay grading, the proposed is nothing more than a storage system where the instructor has to manually assess the student's writings.

Alghamdi et al. [14] present a hybrid automatic system that combines LSA and three linguistic features: 1) word stemming, 2) word frequency, and 3) number of spelling mistakes. Their proposal determine the optimal reduced dimensionality used in LSA to evaluate the performance of

their proposed system. Khalid and Izzat [4] present a method depending on synonym and stemming of words. They assign weights the instructor's answer words to benefit the assessing process. Their study was impracticable and had neither a dataset nor an experimental result.

## III. DATASET

In this research, general methodology will be used to advance a grading system of Arabic short answers based on text similarity matching methods. To evaluate the methods for short answer grading, we used the dataset prepared in a general sociology course taken by secondary level 3 students where the total short answers in this dataset are 210 short answers (21 questions/assignment x 10 student answers/question). It was evaluated by judges where the scores ranged between 0 (completely wrong) and 5 (completely right). Each judge was unaware of the other's correction and grade. We considered the average score of two annotators as the gold standard to test the automatic grading task. Table I displays a sample question, model answer, student answers, and the average score.

TABLE I. EXPERIMENT'S DATASET SAMPLE

Question	عرف الصراع الاجتماعي؟	Score
Model Answer	هو عملية اجتماعية سلبية هدامة لأنه يعبر عن القوى الاجتماعية و مدى تصادمها و ينشأ نتيجة للظروف الاجتماعية و الاقتصادية و السياسية غير المستقرة و قد يصل الصراع الى حد التناحر من اجل البقاء	
1	هو عملية اجتماعية سلبية هدامة لأنه يعبر عن القوى الاجتماعية و مدى تصادمها و ينشأ نتيجة للظروف الاجتماعية و الاقتصادية و السياسية غير المستقرة و قد يصل الصراع الى حد التناحر من اجل البقاء.	5
2	عملية هدامة تعبر عن القوى الاجتماعية و الصدام بينهما.	2
3	عملية اجتماعية سلبية تنشأ عن الصراع بين القوى الاجتماعية و الاوضاع غير المستقرة.	3
4	عملية اجتماعية تحدث نتيجة الاوضاع السياسية و الاقتصادية و الاجتماعية غير المستقرة.	2.5
5	عملية هدامة تحدث نتيجة الاوضاع غير المستقرة في المجتمع.	1.5
6	عملية اجتماعية سلبية تؤدي لهدم المجتمع.	1
7	هو عملية اجتماعية سلبية هدامة لأنه يعبر عن القوى الاجتماعية و مدى تصادمها و ينشأ نتيجة للظروف الاجتماعية و الاقتصادية و السياسية غير المستقرة و قد يصل الصراع الى حد التناحر من اجل البقاء.	5
8	هو عملية اجتماعية سلبية هدامة لأنه يعبر عن القوى الاجتماعية و مدى تصادمها و ينشأ نتيجة للظروف الاجتماعية و الاقتصادية و السياسية غير المستقرة و قد يصل الصراع الى حد التناحر من اجل البقاء.	5
9	عملية اجتماعية سلبية هدامة.	0.5
10	صراع القوى الاجتماعية التي تؤدي الى التصادم و التناحر.	2

## IV. STRING-BASED AND CORPUS-BASED APPROACH

### A. String-Based Text Similarity

Damerau-Levenshtein (DL) algorithm works on counting the quorum of processes that are required to map one string into another string. These operations could insert, or a character obliterate from the string. Moreover, it could be a replacement of a single character, or a reversal of two adjacent



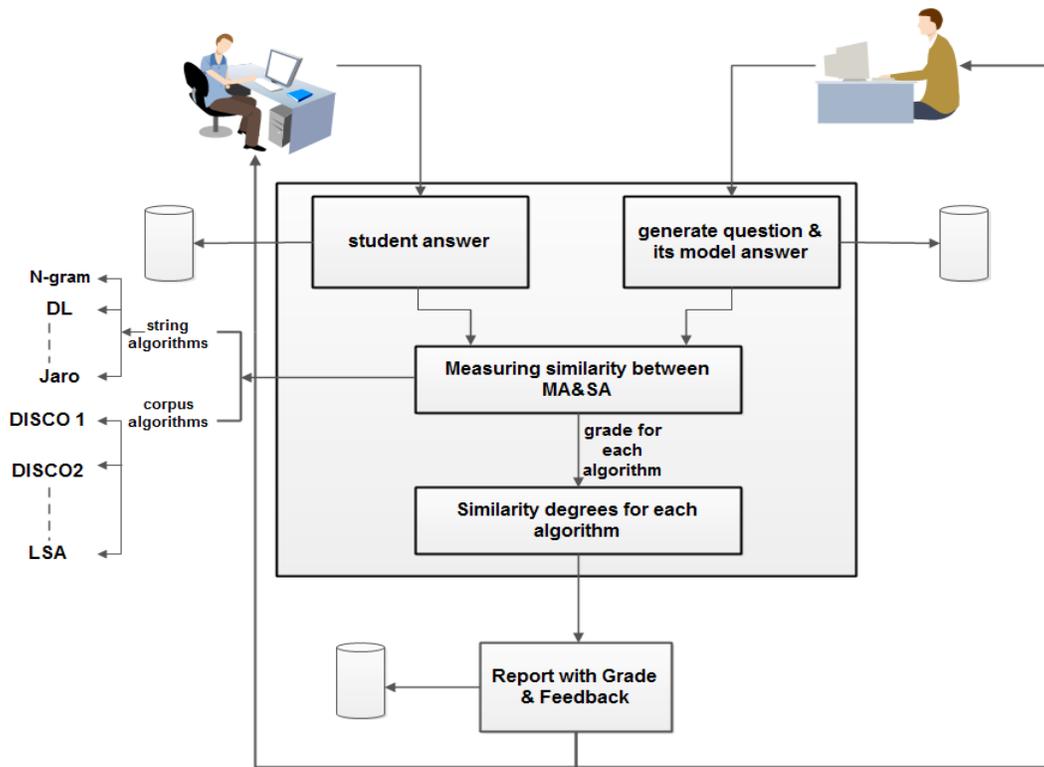


Fig. 1. Proposed system.

The objective of stemming is to find the representative indexing form of a word by truncating these affixes.

Antefixes (وبال, وال, بال, فال, كال, ول, ال, وب, ول, لل, فس, فب, فل, وس, ك, ف, و, ب, ل)

Prepositions meaning respectively: and with the, and the, with the, then the, as the, and to (for) the, the, and with, and to (for), then will, then with, then to (for), and will, as, then, and, with, to (for).

Prefixes (ن, ي, ت)

Letters inflecting morphologically for person.

Suffixes (تما, يون, تين, تان, ات, ان, ون, ين, وا, تا, تم, تن, نات, ن, ا, ي, و)

Word endings inflecting for verb ending, and dual/plural/female markers

Postfixes (كما, هما, كن, هن, تي, ها, نا, هم, كم, ك, ه, ي)

Pronouns meaning respectively: your, their, your, their, my, her, our, their, your, your, his, me.

E. Stopstem

A combination of stop-words and stemming tasks are applied.

VI. EXPERIMENTAL RESULTS

Our experiments were performed using two string-based algorithms and two corpus-based algorithms. Four different methods Stop, Raw, Stop-Stem and Stem are used in testing for string-based algorithms. However, only the Stop method is

used in corpus-based algorithms where Raw, Stem, and Stop-Stem methods cannot be utilized because there is no need to measure the semantic similarity between the Stop words.

The correlation constant was calculated between automatic system and the human grading. The correlation coefficient is used to determine to what extent the system and the human are correlated in assigning the grades.

Equation (1) is correlation constant where X, Y are two sets and  $\bar{x}, \bar{y}$  are the averages of each set in series.

$$Cor(X, Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \tag{1}$$

As shown in Fig. 2 and 3 the correlation between the applied algorithms and the manual grading are presented for the same question. The main target of any algorithm is to be very near to the manual grading to prove the efficiency of this technique. The degrees are from 0 to 5 and the number of students is ten. The figures show that the N-gram can be used as the nearest algorithm to the manual grading.

As noticed from Table II, in string -based distance measures; DL resemblance got the best association value 0.803 practical to the Stop-Stem text. The reason behind this could be Stop stem works on the origin of the words comparing the characters against the model answer and neglects the stop-words. This enables the algorithm provide very high results.

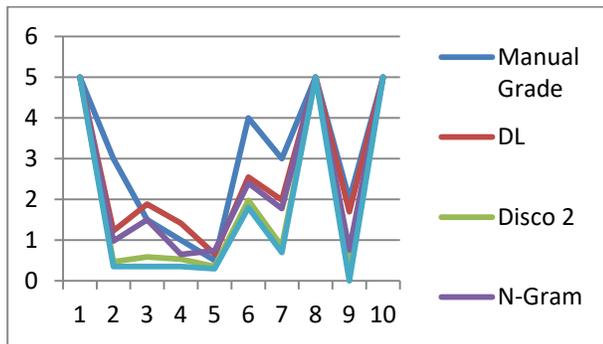


Fig. 2. The grades assigned by teacher, DL, for 10 students' response to the question 'عرف الظاهرة الاجتماعية؟'.

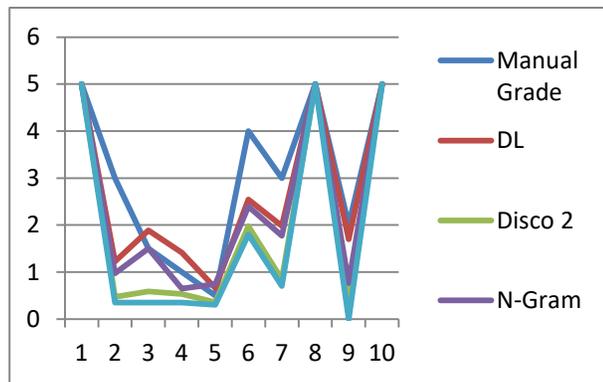


Fig. 3. The grades assigned by teacher, DL, for 10 students' response to the question 'عرف علم الاجتماع؟'.

Moreover, for N-gram algorithm, the best correlation achieved while using Stop method is 0.820. The character-based N-gram algorithm achieved better results than the other three types. In general, the character-based N-gram approach has many advantages, such as: simplicity; it is more reliable for noisy data such as misspellings and grammatical errors; it outputs more N-grams in given strings than N-grams resulting from a word-based approach, which leads to collecting a sufficient number of N-grams that are significant for measuring similarity, and easy to conceptually understand, fast to calculate, language-neutral (i.e., it allows neglected characteristics of language features), error-tolerant.

For corpus-based similarity, LSA achieved 0.781 correlation value while DISCO2 achieved 0.796. Dependently, DISCO2 achieved a higher correlation compared to LSA because DISCO2 depends on groups of distributionally similar words.

TABLE II. THE CORRELATION RESULTS BETWEEN MODEL ANSWER AND STUDENT ANSWERS USING DL, N-GRAM, LSA, AND DISCO2.

Algorithm	Correlation			
	Raw	Stop	Stem	Stop-Stem
DL	0.799	0.800	0.798	0.803
N-Gram	0.800	0.820	0.796	0.795
LSA	N/A	0.781	N/A	N/A
Disco2	N/A	0.796	N/A	N/A

## VII. CONCLUSION AND FUTURE WORK

The automatic grading system is an efficient way for grading even if it is used for article questions. The automatic grading system provides many advantages, it is very quick in implementation of results, provides an easier and flexible platform for subjective questions, workload of invigilators is reduced, manual performance is reduced by performing everything online and it is Low cost. The character-based N-gram algorithm achieved better results than the other three types. The N-gram approach has many advantages, such as: simplicity; it is more reliable for noisy data such as misspellings and grammatical errors; and it outputs more N-grams in given strings than N-grams resulting from a Word-based approach, which leads to collecting a sufficient number of N-grams that are significant for measuring the similarity. The paper proved that using string algorithms gives the teacher an effective solution to help them to undertake student grading with high precision. The paper presented the comparison between four effective algorithms to prove the possibility of automatic grading over manual grading systems.

In future work, we are aiming to combine the string algorithm and corpus algorithm together to achieve the highest possible results depending on the synonymous method and content method to decrease automation grading errors. We are aiming to use different datasets in different subjects to find out the reliability of those algorithms in applications.

## REFERENCES

- [1] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," *Journal of Information Technology Education*, vol. 2, pp. 319-330, 2003.
- [2] T. Miller, "Essay assessment with latent semantic analysis," Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada, 2002.
- [3] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem," *The Journal of Technology, Learning, and Assessment*, vol. 1, no. 2, 2002.
- [4] K. M. Nahar and I. M. Alsmadi, "The automatic grading for online exams in Arabic with essay questions using statistical and computational linguistics techniques," *MASAUM Journal of Computing*, vol. 1, no. 2, 2009.
- [5] S. Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) system in Indian context," in *Proceedings of TENCON 2008-2008 IEEE Region 10 Conference*, 2008, pp. 1-6. "doi:10.1109/TENCON.2008.4766677"
- [6] Callear D., Jennifer Jerrams-Smith and Victor Soh. CAA of Short Non-MCQ Answers. *Proceedings of the 5th CAA Conference*, Loughborough, UK, 2001.
- [7] Pulman S. G. and Sukkarieh J. Z. Automatic Short Answer Marking. *Proceedings of Second Workshop on Building Education Application Using NLP*, Ann Arbor, Michigan, 2005.
- [8] Aziz, M. and F. Ahmad, 2009. Automated Marking System for Short Answer Examination (AMS-SAE). *Proceedings of the IEEE Symposium on Industrial Electronics and Applications*, Oct. 4-6, IEEE Xplore Press, Kuala Lumpur, pp: 47- 51. DOI: 10.1109/ISIEA.2009.5356500.
- [9] F. Mozgovoy, T. Kakkonen and E. Sutinen, "Using natural language parsers in plagiarism detection," in *Proceedings of the SLATE Workshop on Speech and Language Technology in Education*, Farmington, Pennsylvania, USA, 2007.
- [10] S. Alzahrani, N. Salim and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods," *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 42, no. 2, pp. 133-149, 2012.

- [11] Goma, Wael Hassan, and Aly Aly Fahmy. "Automatic scoring for answers to Arabic test questions." *Computer Speech & Language* 28.4 (2014): 833-857.
- [12] R. Mezher and N. Omar." A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring." *Journal of Applied Sciences* ,16(5): 209-215(2016).
- [13] Al-Shalabi, Emad Fawzi." An Automated System for Essay Scoring of Online Exams in Arabic based on Stemming Techniques and Levenshtein Edit Operations." *International Journal of Computer Science Issues(IJCSI)* 13.5(2016).
- [14] Alghamdi, M.,M. Alkanhal,M. Al-Badrashiny, A.Al-Qabbany,A. Areshey,A. Alharbi. A hybrid automatic scoring system for Arabic essays.*J.AI Commun.*,27:103-111.
- [15] Landauer, T.K., Dumais, S.T., 1997. A solution to Platos problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104 (2), 211–240.
- [16] Hall, P.A., Dowling, G.R., 1980. Approximate string matching. *ACM Computing Surveys (CSUR)* 12 (4), 381–402.
- [17] Peterson, J.L., 1980. Computer programs for detecting and correcting spelling errors. *Communications of the ACM* 23 (12), 676–687.
- [18] Barrón-Cedeno, A., Rosso, P., Agirre, E., Labaka, G., 2010. Plagiarism detection across distant language pairs. In: *Proceedings of the 23rd International Conference on Computational Linguistics*, August. Association for Computational Linguistics, pp. 37–45.
- [19] Kolb, P., 2008. Disco: A multilingual database of distributionally similar words. In: *Proceedings of KONVENS-2008*, Berlin.
- [20] Kolb, P., 2009. Experiments on the difference between semantic similarity and relatedness. In: *Proceedings of the 17th Nordic Conference on Computational Linguistics-NODALIDA'09*, May.
- [21] Noorbehbahani F, Kardan AA (2011) The automatic assessment of free text answers using a modified bleu algorithm. *Comput Educ* 56:337–45.
- [22] He Y, Hui SC, Quan TT (2009) Automatic summary assessment for intelligent tutoring systems. *Comput Educ* 53:890–9.
- [23] Chu-Carroll J, Carpenteru B (1999) Vector-based natural language call routing. *J Comput Linguist* 25:361–88.
- [24] Landauer TK, Foltz PW, Laham D (1999) An introduction to latent semantic analysis. *Discourse Process* 25:259–84.
- [25] Bestgen Y (2006) Improving text segmentation using latent semantic analysis: a reanalysis of Choi, Wiemer-hastings, and Moore. *Comput Ling* 32:5–12.
- [26] Magliano JP, Graesser AC (2012) Computer-based assessment of student-constructed responses. *Behav Res Methods* 44:608–21.
- [27] Burstein, J., "The e-rater scoring engine: Automated Essay Scoring with natural language processing", In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring: A cross disciplinary approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates, 2003.
- [28] Mitchell, T., Russel, T., Broomhead, P., & Aldridge N., "Towards robust computerized marking of free-text responses", In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughboroug University, Loughborouh, UK., 2002.
- [29] Leacock C., Chodorow M. C-rater: Automated Scoring of Short-Answer Questions. *Computers and Humanities* 37, 2003.
- [30] A. Shehab, M. Elhoseny, and A. E. Hassanien, "A hybrid scheme for Automated Essay Grading based on LVQ and NLP techniques," in *2016 12th International Computer Engineering Conference (ICENCO)*, 2016, pp. 65-70.
- [31] Rashad, Magdi Z., et al. "An Arabic web-based exam management system." *International Journal of Electrical & Computer Sciences IJECS-IJENS* 10.01 (2010): 48-55.
- [32] Dunning, T. (1994). *Statistical identification of language* (pp. 10-03). Computing Research Laboratory, New Mexico State University.
- [33] Karen R and and Ramsayb J. Now what was that password again? A more flexible way of identifying and authenticating our seniors. *Behaviour and Information Technology* 2007; 26(4): 309–322.