AAEE - Automated evaluation of students' essays in Arabic language

Article *in* Information Processing & Management · July 2019

DOI: 10.1016/j.ipm.2019.05.008

Aqil Azmi
King Saud University

Maram F. Al-Jouie
King Saud University

Muhammad Hussain
King Saud University

AAEE — Automated evaluation of students' essays in Arabic language[☆]

Aqil M. Azmi[*], Maram F. Al-Jouie, Muhammad Hussain

Department of Computer Science, College of Computer & Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Abstract

Assessing student's essay writing and providing thoughtful feedback is a truly labor-intensive and time-consuming task. With human instructors already overwhelmed, the alternate is to consider a computer-based grading. Recent advances have generated renewed interest in automatic evaluation of essays (AEE). The AEEs instantaneous feedback and more consistent grading helps students draft better essays. This work presents a system to automatically grade the school children essays in Arabic, calling it AAEE for "automatic Arabic essays evaluator". The system is modeled upon the scoring scheme followed by the school instructors in Saudi Arabia. The instructors had specific criteria upon which an essay is assessed. Putting these criteria together we developed a system that relies on Latent Semantic Analysis, and Rhetorical Structure Theory. With this design we are able to assess individual components of the essay such as language proficiency, structure of the essay etc. To test the system, we collected essays by local school children covering grades 7–12. A total of 350 different handwritten essays—spanning eight different topics—each transcribed into computer readable format. The AAEE shows that 90% of the test essays were correctly scored, and a correlation of 0.756 between automatic and teachers' scoring. This exceeds the human-human correlation of 0.709 for the Arabic essays.

Keywords: Arabic, Automatic essay scoring, Latent Semantic Analysis, Rhetorical Structure Theory, Improving classroom teaching, Interactive learning environments

1. Introduction

The preferable way to improve students' writing skills is to get feedback from the instructor and, subsequently, engage in this process, recursively, as often as possible. This is a grueling task for an already overburdened instructor, who has to read the essays and provide feedback. As computers' intelligence is rapidly developing, there are plenty of tools that could help teachers become more efficient. One of the interesting tools is automatic computer grading of written essays. Formally abbreviated AEE for "automatic evaluation of essays", a software that evaluates essays that are written by students. Historically, several different names have been used for it interchangeably. The names automated essay scoring (AES), and automated essay grading (AEG) are slowly being replaced with the term AEE. The term evaluation within the name (AEE) came in to use as a consequence of modern systems are expected to provide some kind of feedback (Zupanc & Bosnic, 2015). In specific terms, AEE is defined as the way of evaluating written prose automatically by the computer (Shermis & Burstein, 2003). Evaluation means that the computer system can engage in the task of scoring an essay, or assigning a number to it while providing feedback. AEE is a

[☆] This paper is a substantially enhanced and extended version of the paper presented at the Third International Conference on Arabic Computational Linguistics (ACLing 2017) (Al-Jouie & Azmi, 2017).

[*] Corresponding author. Tel.: +966 (0)11 467-6574; Fax: +966 (0)11 467-5423

Email addresses: aqil@ksu.edu.sa (Aqil M. Azmi), maramfj@gmail.com (Maram F. Al-Jouie), mhussain@ksu.edu.sa (Muhammad Hussain) multi-disciplinary field which incorporates research from diverse areas such as computer science, educational measurement, linguistics, writing research, and cognitive psychology (Shermis et al., 2013).

The history of auto-grading essays in English goes back at least half a century. In 1966, Ellis B. Page took the first steps towards automatic grading of essays. As a Professor of Educational Psychology and Director of Bureau of Educational Research, Page began the development of PEG (for Project Essay Grading) software, inspired by the convergence of computational linguistics, artificial intelligence, and his own experience as a high school English teacher. Page published his (then controversial) paper, "The Imminence of ... Grading Essays by Computer," which discusses the use of computers to evaluate essays and provide feedback (Page, 1966). This was at a time when computers were reserved for the most advanced tasks, and access to them was highly restricted. Using computers to grade essays wasn't realistic, from either a practical or economical standpoint.

The fifty year progress in automatic grading of English essays has yet to reflect in students' performance. The main reason is that AEE has not been fully embraced by the educators. A recent report claims that most college graduates have difficulty writing a decent essay (Lynch, 2018). Students had limited experience in writing essays, largely because the teachers themselves have difficulty grading them. To save time, the teachers stopped giving informative feedback in favor of holistic scoring. There is a high cost associated with every essay having to be graded manually by humans, and the standardized state tests with a written part of the examination have become increasingly expensive. This cost has led many states to ditch this important part of assessment test. On the other hand, the writing sections of TOEFL iBT and GRE tests make use of a combination of human raters and e-rater, a commercial automated writing evaluation engine (ETS Research, 2017). Recently, NPR reported that several states including Utah and Ohio already use automated grading on their standardized tests (Smith, 2018). Initially, their respective State Board of Education were skeptical, and every machine-graded essay was double-checked by a real human. But, as the computer scoring has proven "spot-on" the state now allows machines be the sole judge of the vast majority of essays. In case the computer detects something unusual, the essay is flagged for human review.

According to (www.youthpolicy.org/mappings/regionalyouthscenes/mena/), about 60% of the Arab population—of approximately 300 million—is under 25 years old. This makes the Arab countries one of the most youthful regions in the world, with a median age of 22 years compared to a global average of 28. Though the educational enrollment rates are high, with nearly universal access to primary level and approximately 70% enrollment at the secondary level, unfortunately, the quality of education remains low (same site). The Arabian gulf countries are among the wealthiest in the Arab world. In 2017, the total number of students attending school (all levels) in Saudi Arabia is almost 6.1 million, the largest among the gulf countries (www.statista.com/statistics/628665/saudi-arabia-total-number-of-students/). However, a recent report in a local newspaper stated, "overcrowded Saudi classrooms 'hampering learning process'." (Al-Sughair, 2014). The report states the government education regulations stipulate a maximum of 30 students in a classroom, but in reality there are over 40 students in many of these classes. This has created a scenario where it becomes difficult for instructors to teach, and students to learn, since there is not enough time for in-depth discussion on schoolwork in a typical 50 minute class.

Thus, developing AEE system for the Arabic language is important to make up for the poor state of education. Beside that, there is a need for a standardized mean to assess essay writing. Following the lead by ETS TOEFL Practice Online (TPO), where the examinees take online test to get feedback on among others, their essay writings—everything scored purely by machine including the written essays—other countries such as Saudi Arabia is following the track. Locally, the National Center for Assessment (Qiyas) conducts General Aptitude Test (GAT), an online-standardized test. Like SAT in the US, the admission into any national college/university in Saudi Arabia requires achieving a certain minimum in Qiyas' GAT test, whose grade is incorporated into a composite score which includes high schools' GPA. The current Qiyas system evaluates proficiency of Mathematics, and Arabic language. The SAT offers the essay writing as an option, but no such option is available in Qiyas' GAT. This should be obvious. Qiyas employs an automated grading system, and the center lacks the technique to assess essays in the same fashion as ETS TPO. We believe this—to the best of our knowledge—to be the first study showing the feasibility of AEE system for the school children in the Arab countries. The history of Arabic natural language processing (NLP) is relatively recent compared to that of English, and thus it is lacking in many research areas including grading Arabic essays automatically.

One of the reasons is the complexity of the Arabic language which prevents the development of a strong evaluation system. For instance, the Arabic language is characterized by high ambiguity, rich morphology, complex morpho-syntactic agreement rules, etc (Habash, 2010). Yet another factor is the current trend in Arabic NLP where most of the research is geared towards tackling tweets.

Though we failed to find a definitive study on the nature of written text by the Arabs, we found some information hidden in works related to teaching English as Foreign Language (EFL). The Arabic learners of English face a more arduous task than other EFL learners, specifically when it comes to writing (Meehan, 2013). According to the author, the difficulty stems from the vast differences between the structural systems of Arabic and English writing. Rass (2015) noted that Arab students usually transfer the stylistic features and patterns of thinking of their mother tongue when writing in English. For instance, they tend to write long sentences with coordinating conjunctions (Al-Khatib, 2001), repeat themselves and argue through presentation and elaboration (Johnstone, 1991), and often talk around the topic and repeat phrases before stating the main points (Dweik, 1986). Another stylistic difference is the degree of explicitness and implicitness of the narrative text (Mohamed, 1993). For example, Arab writers usually avoid conveying their messages explicitly, assuming that readers are responsible for understanding the message. The above argument makes it clear that the writing rules and style in Arabic is different from that of English. This goes to say, we can't just take an existing English AEE system and use it to assess some essays in Arabic language.

In this work, we explore the Arabic written essays in general, how they are assessed and graded by instructors and then develop a system that closely mimics human evaluators. Our objective is to save instructors' time and resources spent on improving essays written in Arabic, and writing skills in general. We named our system AAEE for automatic Arabic essays evaluator. The AAEE we devise makes use of latent semantic analysis (LSA), and rhetorical structure theory (RST). The RST is one of the leading theories in computational linguistics, and we use it to assess the coherence of the text. We believe this is the first time RST has been used in assessing the essays. The system is intended to grade typical classroom essays written by school children while providing feedback. Though AAEE is intended for automated grading of essays written in Arabic, it can be adapted for other languages as well. Farsi (Persian) and Urdu speaking countries may benefit from the system as they share a lot culturally with Arabic speaking countries, and this will be reflected in their general writing style.

To evaluate our system we collected a total of 350 different handwritten essays by local Saudi school children. All the essays were transcribed—exactly as is—into computer readable format. We tried to be faithful to the original essay, so they were transcribed line-by-line including typos. However, we ignored the crossed-out text. The essays were graded out of maximum score of 10 marks. The performance was measured by machine-human correlation in grading. AAEE achieved an overall correlation of 0.756 between automatic and the teachers' evaluation. This is close to the inter-human correlation of 0.85 in English (Hearst, 2000), and slightly better than 0.709 for Arabic essays (as we will see later). The inter-human correlation coefficients are computed by taking the human scores for essays for which they were given at least two human judgments and compute the Pearson's correlation coefficient for them.

This paper is organized as follows. Section 2 is an overview of the related works. A brief look at some of the background material in Section 3. Section 4 is a look at our proposed AAEE system. The results of evaluating the system is in Section 5. The last Section concludes the paper with some directions for future work.

2. Literature review

The automatic grading of essays (AEE) is tasked with constructing a computer-based system to grade essays, with the aim of eliminating the need of human raters or minimize their involvement. Until recently, the field has been dominated by commercial organizations keen on protecting their investments by restricting access to their technological details (Zupanc & Bosnic, 2015). As recent as 2015, Zupanc & Bosnic (2015) compared 21 state-of-the-art AEE systems, 19 of which were commercial ones. It goes to show the extent commercial entities controlled this research area. Only in the last few years there were several attempts to make the field more mainstream. This included a recently published Handbook on the subject (Shermis &

Burstein, 2013). Traditional approaches treated AEE as a classification problem (Rudner & Liang, 2002), regression (Phandi et al., 2015), and ranking classification (Chen & He, 2013). We start by looking into some of the different approaches used to solve AEE problem, and we include systems for languages other than Arabic. Then we cover AEE systems that were devised for Arabic essays.

2.1. Different approaches used in AEE system

Burstein & Marcu (2000) argued that vocabulary in 40% and 60% discourse-based summaries contain considerable information from the original essay. They hypothesize that essays are written under time constraint, leaving little time for the student to edit and proof read the text. As a result, essays could have a lot of noise, such as repetitive statements and extraneous comments irrelevant to the main arguments in the essay. Summaries, the authors state, are able to clean up the essay by removing some of the noise leaving the more important part out. This technique could be used to provide feedback for subject-based essays such as in College Board's AP exams in US History.

In (Bin et al., 2008) proposed applying k-Nearest Neighbor (k-NN) used in text categorization model to essay scoring. Implementing the k-NN algorithm for AEE systems occurs in steps. First step, transforming the problem into text categorization involves transforming the required scoring to categories. Next, transforming the essays into vectors. This is followed by feature selection and reduction to minimize the vector space dimensions. The authors used two methods for feature selection: term frequency (*TF*) and information gain (*IG*). Search for the k-NN of the training essays and computing the nearest objects from the new essay to all essays in the training set. After that, the results are sorted in descending order, and the first k essays are selected. The final step is text categorization, which involves classifying the essay into the same category as that of nearby essays. Experiments on CET4 essays in the Chinese Learner English Corpus show achieving accuracy above 76%. In this study, the best result is achieved when k = 3.

Chang et al. (2008) argues that semantic characteristics of an essay must be taken into consideration when scoring an essay. For this, the authors propose using the literary concepts in the essay to score the essay. The difficulty entails from extracting the literary concepts from the small training set. To solve this problem they use the semantic network tool, where all the essays' concepts, including the literary ones, are transformed into sememes. Sememe is a semantic language unit of meaning, analogous to a morpheme, and is a major element in the semantic network. Their work involves the use of one or more semes to represent the semantics of any concept. The authors use three step method to extract the literary sememes: (a) calculate the numbers of sememes in a training set that contains scored essays; (b) employ the number of sememes in the test essay; and (c) calculate the accuracy of scoring essays based on the predictive scores. The data set comprises 689 written essays in Chinese language. They reported an accuracy of 92%.

In (Kakkonen et al., 2008) looked into three different dimension reduction schemes, namely: latent semantic analysis (LSA), probabilistic LSA (PLSA), and latent Dirichlet allocation (LDA); on automated essay grading. All three schemes assume that we can model documents as mere collection of words without any regard to their order. Following experiments, the authors concluded that LSA yield slightly more accurate essay grades than the other two schemes. However, PLSA and LDA tend to provide better means in giving students feedback on the essay.

Latif & Wood (2009) proposed using text summarization as a preprocessing step for essay assessment. Then use the summaries instead of full essays to group like-quality essays. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), is a widely used metric to evaluate the summaries. ROUGE uses n-gram co-occurrences between auto-generated summaries and human generated ones. According to the authors, their experimental results show that there is a positive high correlation between ROUGE scores (of computer generated summaries of student's essays) and human scoring of the essays. This correlation can be used as a basis to classify students' essays into wide-ranging of quality. The data set are sample essays from TOEFL test.

In (Islam & Hoque, 2010) proposed an approach for automated essay evaluation through extending LSA. The paper developed an AEE system using generalized latent semantic analysis (GLSA) which takes an n-gram (sequence of n consecutive words) by document to build a matrix, rather than taking a word by document as in the LSA model. The traditional word by document matrix creation of LSA preclude word sequence in a document (Devi & Mittal, 2013). Thus, the word pair “care free” makes the same result as “free care”. Here, LSA fails to capture the semantic effect of collocations in the document. GLSA resolves this issue by treating *n*-gram as atomic unit of the document instead of individual word. Thus now, “care free” is not the same as “free care”. The training set comprises 960 essays written by undergraduate, and the test set comprises 120 essays. Experimental results show that GLSA achieved an accuracy better than 89%, indicating that this system is close to the human rater. In a later work, the same authors applied GLSA for scoring essays written in Bangla language (Islam & Hoque, 2013).

Razon et al. (2010) proposed to evaluate essays in aspect of semantic analysis using the dimensionality reduction algorithm called concept indexing (CI). CI is similar LSA, which takes semantic similarities between parts of textual input. From the information retrieval (IR) point of view, CI is better than LSA when it comes to execution speed, accuracy, and storage costs. The authors aim was to compare the performance of CI implemented with k-means, and CI implemented with fuzzy C-means, against LSA. They concluded that CI could replace LSA in the evaluation of essays. Table 1 summarizes the result of this paper, which compares both CI’s and LSA. The EAA stands for exact agreement accuracy, which indicates the number of essays that receive the same percentage score from the AEE system and human rater. The Pearson product-moment correlation coefficient (PMCC) measures the relationship between the number of test essays and the instructor’s score against the automated score of each essay. The data sets for these experiments were essays from two sections in high school English classes from the University of the Philippines Integrated School.

Table 1: Summary of experimental results of the University of Philippines Integrated School English classes (Razon et al., 2010). EAA refers to exact agreement accuracy, and PMCC is the Pearson correlation coefficient.

	University of Philippines Integrated School			
	First Year		Second Year	
Algorithm	EAA	PMCC	EAA	PMCC
LSA	78.95%	0.486	62.50%	0.626
CI with k-means	84.21%	0.669	81.25%	0.680
CI with fuzzy C-means	89.47%	0.819	81.25%	0.832

In (Chen et al., 2012) considered AEE as a ranking problem rather than treating it as classification or regression problem. The authors proposed rank-based learning approach, a supervised machine learning (ML) technique which automatically construct a ranking model of the retrieved essay. The paper showed that learning to rank is much better than other traditional ML techniques in AEE. However, machine-learned ranking is typically used to rank documents in IR. The accuracy of this system is not as good as that associated with the practical use of e-rater and IntelliMetric, both of which are commercial AEE systems. These commercial systems has the advantage of having huge data set from the examinations held every year. The authors proposed two algorithms: SVMRank, and LambdaMart. Respectively, these are pair-wise, and list-wist ranking algorithms. The data set comprises essays from a private company in US. The best result was achieved by SVMRank with a reported accuracy of 73%.

The next system is geared towards grading open-ended questions in a specific domain. Ade-Ibijola et al. (2012) devised ES4ES (Expert System for Essay Scoring), an expert system for scoring free text answers. The system is composed of three modules: knowledge base (KB), inference engine, and working memory. The designed KB uses semantic network approach for knowledge representation, and the knowledge base for the system is based on “Software Engineering-I”, a course offered at the author’s institute. The inference engine uses shallow NLP technique for information extraction purposes. It uses inference rules pattern to match the data contained in the knowledge base. The synonyms handler module further improves the performance. The authors propose using fuzzy-scoring approach to evaluate the students’ answers. The system was tested in a classroom exam of ten questions to five students. The exam was graded independently by ES4ES and by a human expert. The result showed a 0.71 correlation between the system score and the instructor score. This system has some limitations. For instance, it is not suitable to assess free-text answers where the word order matters. It is also more geared towards short answers rather than large texts.

Doğan et al. (2014) looked into the reliability of AEE scoring system, in particular the commercial e-rater (ETS Research, 2017), by comparing it with human scoring. The methodology used is to compare the scoring of 50 essays between e-rater and three humans. All three humans have similar education and experience. The essays were scored in the range from 1 to 6 marks. The authors calculate the correlation as a matrix between the grades that the humans and the electronic rater award. Table 2 summarizes the results of this comparison between humans and the electronic rater. It is apparent that the AEE system gives a grade close to the human raters. However, the electronic evaluation gives the essay a slightly higher score than the humans.

Table 2: Correlation of grading 50 essays by human raters and e-rater (a commercial scoring engine) (Doğan et al., 2014). The Table is symmetric.

	Human raters		
	First	Second	Third
Second human rater	0.583		
Third human rater	0.652	0.841	
E-rater (electronic rater)	0.716	0.712	0.890

Zupanc & Bosnic (2015) compared 21 state-of-the-art AEE systems, which includes 19 commercial and two publicly available systems, looking into their highlight and weaknesses. The authors conclude that the AEE systems has matured to the point where the system can provide useful feedback on students' writing; and that these systems can be a useful complement to human scoring, i.e. should not replace human rater. According to the authors, NLP has influenced the growing development of the AEE systems in the last two decade. The main challenges presented in the paper are reliability, validity, validation, and the semantic evaluation of essays. Validation is defined as the accumulation of evidence to support interpretation of the test score, while validity is a teacher's perception that the validation evidence is sufficient. Reliability is concerned with the consistency of test scores. For evaluation the authors mention a variety of common metrics, such as Pearson's and Spearman's correlation, F-measure etc. Two of the best performing AEE systems were: Semantic Automated Grader for Essays (SAGE) (Zupanc & Bosnic, 2014), and PEG (Page, 1966, 1994) with a reported accuracy of 0.83, and 0.79 (respectively).

2.2. Arabic-based AEE systems

Though the research into Arabic NLP has picked up steam in recent years, it does lag behind in similar research in other languages, particularly English. In the area of AEE systems we came across two papers only, one of them is our paper reporting preliminary results of this work (Al-Jouie & Azmi, 2017). It is not a surprise, there are many more work on scoring short essays to questions. These systems are closed-domain, and as thus, less challenging since the systems are more or less after specific keywords, and some kind of similarity measures.

Alghamdi et al. (2014) presented Abbir, a hybrid AEE for essays in Arabic language. The system combines LSA, and some linguistic features which included number of words, and the number of spelling mistakes. The system was tested on two group of undergraduates, 329 and 311 students respectively. The first group was asked to write an essay on the impact of Internet on the society, while the second group was asked to write on smoking in public places. The size of the essay was limited to 100–200 words. Each essay was scored by two human raters with a grade from 1 to a maximum of 6 (best score). A third rater gets involved when the absolute difference between the original two human raters is over 1 for a particular essay. Each essay is assigned the average score of the involved raters, and rounded to the nearest integer. According to the authors, the correlation between human raters was 0.7. The system has two phases: training, and runtime. The training phase is made up of three components: the word vector, the vector of spelling mistake, and the LSA concept space. In addition, this phase involves some pre-processing, including Buckwalter stemming. In the runtime phase, the input essays pass through a number of processes. These processes make it possible to get the minimum cosine distance between the input and the training essays. During this phase, the authors use a linear regression approach to obtain features that reflect the human senses. They used 579 documents for training, and 61 documents for the testing. The best performance, correlation of 0.78 between machine-human, was reported when LSA with stemming and including the linguistic features were used.

The rest of the systems reviewed in this Section scores short free text answers (short essays) to questions. Nahar & Alsmadi (2009) presents a theoretical research that did not use a data set and has no experimental results. The essay questions require specific answers, so the proposed approach requires the instructors to present a model answer for each question. For each model answer, the instructor is expected to mark selected keywords and assign them weights. The system uses stemming process and word synonyms that were manually predefined. In the end, the distance between the student answer and the model answer determines its score.

In (Gomaa & Fahmy, 2014) is yet another system for auto-grading short answers in Arabic language. The paper presents an evaluation of the system that combines string, corpus, and knowledge-based similarity measures. The system also provides a limited feedback in terms of comments that describes answer's level of correctness. However, the system is not fully automated and does require human intervention. The system was tested using a data set of 61 questions with 10 answers for each, pertaining to an official Egyptian curriculum for a course on environmental science. The answers were graded using a value between 0 and 5, and the average length of student's answer is 20 words. The authors reported inter-human correlation of 0.86, and the best result reported for their system was a machine-human correlation of 0.83.

Ewees et al. (2014) presented a comparison between cosine similarity and k-NN algorithm in LSA method to automatically score Arabic essays. Cosine similarity with LSA led to better performance than using k-NN with LSA. The former yielded a correlation of 0.88 between machine-human, versus 0.50 for the latter. The study is based on 29 answering papers (essays) by sophomore students attending "System Designing" course, answering a single essay question in Arabic "What are the objectives of the design phase?". Only five essays were used for training, and the rest of 24 essays were used for testing. The data set used in this study is too small to warrant any serious consideration.

In (Abbas & Al-Qaza, 2014) presented AAES (for "automated Arabic essay scoring") a web-based system that is based on Vector Space Model (VSM). The system is intended to score short essay type answers to questions. The two phase system initially applies IR to extract important information from the essays. And in the second phase, the VSM is applied to find the similarity degree between the written essay and a model essays written by the instructor. The similarity is measured using term frequency and inverse document frequency (*TF-IDF*) and cosine similarity. The authors tested the system using 30 written essays answering the single question "What is computer network?". These essays where scored against four model answers prepared by human instructor. The authors computed the similarity of a written essay with the model answers. The essay is assigned the score corresponding to the highest similarity. The authors did not provide an overall assessment of their system, but rather showed a figure with two curves, one showing AAES score and another with instructor score for individual written essay. In another paper (Abbas & Al-Qaza, 2015), the same authors improved their system by further employing Latent Semantic Indexing (LSI). Using the same dataset of 30 written essays graded against four model answers, the authors reported a correlation of 0.978.

Shehab et al. (2018) is another system to score Arabic short essays to questions. For similarity measure the authors used string-based and corpus-based text similarities. To evaluate the system the author used a dataset of 210 short answers (10 students answering 21 questions) in general sociology course. The students' answer were scored against a single model answer per question. The best correlation of 0.82 was reported when using n-gram with stop-word removal.

3. Background

3.1. Arabic language—a brief look

Arabic is a Semitic language and one of the official United Nations languages. It is the fifth most spoken language in the world, in addition to the many Muslims worldwide who are keen on learning it. Arabic is written from right-to-left, and can be classified as Classical, Modern, and vernacular. In this work we will be dealing with Modern Standard Arabic (MSA), an expanding form of the language to embrace the modern challenges (Farghaly & Shaalan, 2009). MSA is the literary standard across all the Arab countries. Most of the printed material—books, newspapers, magazines, reading primers for children—is written in MSA.

The Arabic language alphabet consists of 25 consonants, three long vowels, and thirteen diacritical markings to represent the different combinations of three short vowels, the syllabification mark, and the nunation. The optional markings are used to indicate the phonetic information associated with each letter, and to clarify the sense and meaning of the word. Modern Arabic is printed and written without diacritical markings. Occasionally, diacritics appear either only on homographic words when not disambiguated by surrounding text or in full text such as in religious or elementary level educational texts (Hermena et al., 2015). This does help the humans in resolving ambiguity, however, it is useless for computers as they do not understand the text. Every single unvowalized word in Arabic has multiple meaning, which is disambiguated by the set of diacritical markings. For example, the following words all have the same three letters (ع ل م: *E l m*) but the different diacritical markings results in different meaning,[1] (عَلَم: *Ealam*), (عِلْم: *Eilom*), and (عَلَّم: *Eal~am*) meaning *flag*, *science*, and *taught* respectively. The lack of diacritical markings is a source of ambiguity in computational NLP. Another problem, there are no capital letters in Arabic alphabet. This complicates certain NLP tasks, e.g. named-entity recognition.

Some of the challenges associated with using Arabic language are, the ambiguity in the writing system, the rich morphology, and the complex word formation process. Arabic morphology is highly complex, though systematic. One of the prominent phenomenon of Arabic is the presence of two types of morphemes (the smallest meaningful unit in a language, e.g., “girls” has two meaningful units: “girl”, and the plural marker “-s”), templatic and concatenative (Habash, 2010). The templatic morphology is used to generate templatic stem which is composed of root, pattern and diacritical markings. While the concatenative morphology helps in forming words through a sequential concatenation process on top of stems by adding affixes and clitics. The stems are mostly based on a triliteral, and scarcely on larger roots. In the triliteral roots, the three consonants which are known as radicals are designated by C_1, C_2, and C_3. The pattern is an abstract template where diacritics and the generic letters of the root are replaced by the given letters of the root. Consider for example the root word (ك ت ب: *k t b*) and the pattern $C_1aC_2aC_3a$ results in the word (كَتَبَ: *kataba*) meaning *write*, and for the pattern $C_1iC_2aAC_3$ results in (كِتَاب: *kitaAb*) *book*. The complex morphology allows a single Arabic word, e.g. (فسيكفيكهم), the power to convey a whole sentence (see Table 3).

Table 3: Example of a complex morphology of an Arabic word. It means “Then He will be sufficient for you against them”.

ـهم *hm*	كـ *k*	كفيـ *kfy*	يـ *y*	سـ *s*	فـ *f*
them	you	be sufficient	he	will	then
suffix		root	prefix		

3.2. Overlook at techniques employed in AAEE

There are many approaches used in implementing AEE systems in general. In this Section we briefly look at the techniques used in implementing AAEE. Broadly these are: information retrieval, machine learning, and NLP.

The goal of information retrieval (IR) system is to rank documents optimally given a query so that relevant documents would be ranked above non-relevant ones (Zhai, 2008). Over the years, many different types of retrieval models have been proposed and developed, mainly: the Boolean model, the Statistical model, which includes the vector space and the probabilistic retrieval model, and the Linguistic and Knowledge-based models. The standard Boolean approach has several shortcomings. For instance, users find it difficult

[1] We use Buckwalter transliteration scheme (`http://www.qamus.org/transliteration.htm`) for rendering Arabic script into Roman.

to construct effective Boolean queries. When writing a query the users resort to their knowledge of English, where the natural language terms AND, OR or NOT have a different meaning when used in a query. Also, the traditional Boolean approach does not provide a relevance ranking of the retrieved documents.

In vector space and probabilistic models, both use statistical information in the form of term frequencies to determine the relevance of documents with regard to the query. Both schemes produce—as their output—a list of documents ranked by their estimated relevance. The latent semantic indexing (LSI, sometimes referred to as latent semantic analysis, or LSA for short) is based on statistical retrieval approach. LSI measures distributional semantics between documents and terms. It assumes there is some latent structure in the pattern of word usage across documents, and that statistical techniques can be used to estimate this structure. The beauty of this approach is that we can retrieve documents even if they have no words in common with the query. The model involves a deeper process within the structure. LSI algorithm produces a matrix that includes the words in the rows and the documents in columns. Since the term-document matrix is huge, we often lower its rank using SVD (singular value decomposition).

Machine learning (ML) systems automatically learn programs from data (Domingos, 2012). The advantage of ML in AEE system, in general, is the ability to integrate various kinds of document features into the process of ranking (Shermis & Hamner, 2013). There are different ML algorithms, e.g. k-nearest-neighbor (k-NN), and linear regression function. The k-NN is a simple classification algorithm that takes the data points that are separated into several classes to predict the classification of a new data point. The new object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. In linear regression we attempt to model the relationship between two features by fitting a linear equation to observed data. One feature is considered to be an explanatory feature, and the other is considered to be a dependent feature. In AEE systems linear regressions is used to solve the ranking problem, e.g. e-rater (ETS Research, 2017) a commercial AEE system that uses multiple linear regression techniques to predict a score for the essay.

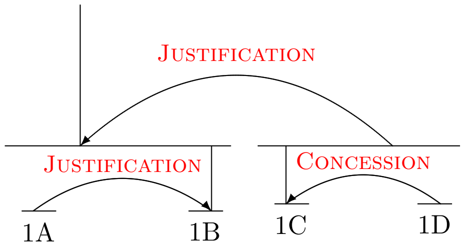
One of the things we seek to address, does part(s) of the essay fit together in a natural or reasonable way. In other words, is the essay coherent. Rhetorical Structure Theory (RST) (Mann & Thompson, 1987), Segmented Discourse Representation Theory (SDRT) (Asher & Lascarides, 2003) etc, are different theories that tackle discourse analysis, a study of how texts are organized and attempts to grasp their underlying structure. Both discourse theories, RST and SDRT, have been adapted for the Arabic language. The reason we decided to use RST as opposed to, say SDRT, is that RST has been around for a while, and has been applied in different Arabic NLP applications, see e.g. (Salem et al., 2010; Azmi & Alshenaifi, 2016; Azmi & Altmami, 2018). RST assumes texts to consist of at least two spans, that are linked by a particular (discourse) relation. The spans themselves may consist of smaller spans linked by relations etc, down to the level of independent clauses (Mann & Thompson, 1992). Text structures in RST are hierarchic, built on patterns called schemas, which describe the functions of the parts rather than their form characteristics (Mann & Thompson, 1987). The rhetorical structure tree (RS-tree) is a representation of elementary discourse units and the rhetorical relations among them. Table 4 illustrates the RS-tree of a sample text. According to (Taboada & Stede, 2009) there are two underlying principles to RST: (a) coherent texts consist of minimal units (text spans) that are recursively linked to each other through rhetorical relations; and (b) there should be no gaps in coherent texts, i.e. there must be some relation attributable to different parts of the text. The set of RST rhetorical relations are not universal, and they vary from one language to another. Mann & Thompson (1988) defined a set of twenty-three rhetorical relations for English. For Arabic, however, it is only eleven (Al-Sanie, 2005).

4. Our proposed system

In this chapter we present the general design and implementation of AAEE, an automated essays evaluation system for the Arabic language. The aim of AAEE is to score students essays written in Arabic while providing some feedback. To be more precise, essays written by Intermediate (Middle School) and Secondary (High School) students in Saudi Arabia. In the Saudi education system, the twelve year schooling is divided into six, three and three years for Primary, Intermediate, and Secondary education (respectively).

Table 4: A sample text and its corresponding RS-tree. The height of the tree is two.

[No matter how much one wants to stay a non-smoker,]^{1A} [the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.]^{1B} [We know that 3,000 teens start smoking each day,]^{1C} [although it is a fact that 90% of them once thought that smoking was something that they'd never do.]^{1D} (text$_{0786e}$)



4.1. Scoring methodology

We used questionnaire to gather information from the instructors grading the essays. The questionnaire focused on the criterion used to grade the essays, and the weights assigned to each criteria. The consensus among the instructors was to look at three different criteria when assessing an essay. These are: semantic analysis, writing style, and spelling mistakes. This is known as analytic scoring of the essays (Foltz et al., 2000; Diederich et al., 1961), as opposed to holistic scoring (White, 1993). What they disagreed on was, how much weight to assign to each criteria. The majority favored assigning 50% of the grade on the semantic analysis, 40% for the writing style, and just 10% for the spelling mistakes. All the essays were graded using integers ranging from 1 (weak) to 10 (excellent).

4.2. System detail

Our system comprises latent semantic analysis (LSA), rhetorical structure theory (RST), and other features that will be explained later in this Section. This design gives us the flexibility to do analytic scoring of the essay as well as providing informative feedback on different components of the essays (e.g., coherence, spelling). Figure 1 shows the general architecture of AAEE. The system is divided into two phases: training and testing. The collected essays (data set) is divided into two disjoint sets, training and testing. In the training phase we use a set of pre-scored essays, which we will designate $E_{Training}$. These essays were scored by human instructors out of 10. To avoid bias, each essay was graded by two human raters. In case the absolute difference between the first two raters of a particular essay was beyond a given threshold, then a third human rater was involved. This is similar to the approach used by (Alghamdi et al., 2014). In the testing phase, the score of a new essay E' is calculated through a number of steps.

Due to the unstructured representation of the essays, direct scoring leads to low-quality results, we thus preprocess the essays. The preprocessing stage includes: (a) breaking an essay into sentences; (b) segmenting the sentences into disconnected words; (c) normalizing the text, where variants of a letter is represented by a single form, e.g. unify the different forms of the character Alif (ا،أ،إ،آ) → (ا); and (d) stemming. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base, or root form. For example, we have two forms for *teachers*, (المعلمون: *AlmElmwn*) (masc.) and (المعلمات: *AlmElmAt*) (fem.), both are derived from the stem (معلم: *mElm*). For stemming we use Buckwalter Morphological Analyzer version 2.0, an open-source morphological analyzer.[2] Buckwalter stemming breaks down the Arabic word into its root/stem and affixes (prefix, suffix, etc.). We this step, the same word in its various forms can be identified; we can find the same word by removing the prefixes and suffixes of the input word.

In the training phase we use a set of pre-scored essays, which we will designate $E_{Training}$. These essays were scored by human instructors out of 10. In this phase, we generate composite word vector, and calculate the average number of words per essay in the training set (α). The training stage includes LSA steps for extracting the essays' features, creating a general LSA model. It calculates the LSA concept matrix of the training set, which is a composite word vector. This LSA model will help in predicting the score of a new essay in the testing phase. In the testing phase, the score of a new essay E' is calculated through a number

[2] `https://catalog.ldc.upenn.edu/products/LDC2004L02`.

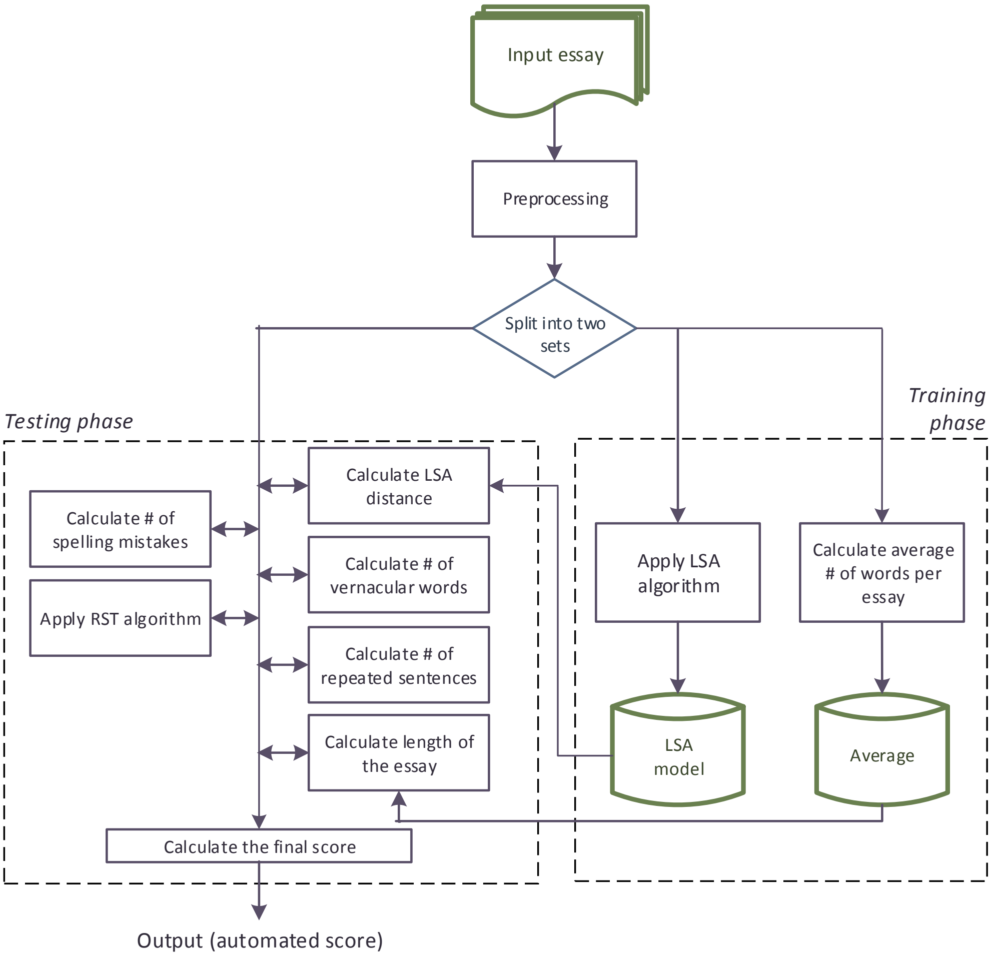

Input essay
Preprocessing
Split into two sets
Testing phase
Training phase
Calculate # of spelling mistakes
Apply RST algorithm
Calculate LSA distance
Calculate # of vernacular words
Calculate # of repeated sentences
Calculate length of the essay
Calculate the final score
Apply LSA algorithm
Calculate average # of words per essay
LSA model
Average
Output (automated score)

Figure 1: The proposed architecture for AAEE, the automatic Arabic essay evaluation system.

of steps. The phase starts with the preprocessing of the input essay. Next, the cohesion of the written essay in relation to the topic is checked. Then, we count the number of spelling mistakes, the length of the essay E' compared to α, and the number of repeated statements. We apply RST to check for the cohesion of the essay, and its writing style, and we also calculate the LSA distance that facilitates the semantic analysis of the essay. Finally, we calculate the overall score of the essay E'.

The main algorithm is listed as Algorithm 1. This algorithm accepts an essay E' and auto-grades it out of 10. As we said earlier, there are three criteria AAEE looks at. The score of each criteria is assigned to a variable, with a value ranging between zero and 10. These variables are: score_{Sem}, which carries the score assigned to the semantic analysis; score_{W}, holds the score set to the writing style; and for spelling mistakes, we assign the score to score_{Sp}. The overall score of the essay is the weighted sum of the three variables (line 7 in Algorithm 1), which reflects the consensus among instructors when manually grading the essays (Section 4.1). This breakdown of scores gives us the ability to offer the student a limited feedback.

4.2.1. Semantic analysis

In this step, the LSA is first trained on relevant materials that have been pre-scored ($E_{Training}$). The training results in a semantic representation of the knowledge of the topic. By using this representation,

Algorithm 1: Main algorithm - assesses the overall score of an essay out of 10.

```
Input: the essay to score ($E'$).
Output: total score (Overall_score) of the essay $E'$.
begin
    Overall_score ← 0
    $\text{score}_{Sem}$ ← LSAdistance($E_{Training}$, $E'$)
    $\alpha$ ← average # words in $E_{Training}$
    $\text{score}_W$ ← WritingStyle($E'$, $\alpha$)
    $\text{score}_{Sp}$ ← 10 − (# of spelling mistakes)/3
    // Ensure $\text{score}_{Sem}$, $\text{score}_W$, and $\text{score}_{Sp}$ are non-negative
    Overall_score ← $^1/_2 \cdot \text{score}_{Sem} + {}^2/_5 \cdot \text{score}_W + {}^1/_{10} \cdot \text{score}_{Sp}$
end
```

bits of textual information (e.g., words, sentences, or essays) can be compared against each other. This comparison provides a measure of semantic relatedness. In a way indicates the extent to which the chunks of text are discussing the topic in the same way. We apply LSA to evaluate the semantic aspect of the new essay E', and accordingly score score_{Sem}. The advantage of using LSA is that some of the retrieved documents may not have words in common with the essay E'.

4.2.2. Writing style

The system automatically evaluates the writing style using several different criteria. These include: (a) the cohesion of an essay, (b) checking for duplicate sentences, (c) do we have vernacular words in the essay, and (d) length of the essay. Algorithm 2 appraises the writing style of the essay.

Algorithm 2: Computes the writing style of an essay, out of a max score of 10. Four factors are considered: cohesion, usage of vernaculars, avoid duplicate text, and the overall length of the essay.

```
Input: a single essay ($E'$), and average number of words in the training essays ($\alpha$)
Output: score of the writing style.
Procedure WritingStyle($E'$, $\alpha$)
    Build RS-tree($E'$)                          // Rhetoric Structure Theory (RST) tree
    $S_{\text{RST}} \leftarrow \min\left\{3, \text{maximum level of RS-tree}(E')\right\}$
    Normalize $S_{\text{RST}}$ score to be out of 10
    size_of_essay ← # words in essay $E'$
    if size_of_essay < $\alpha/4$ then $S_{\text{length}} \leftarrow 2$
    else $S_{\text{length}} \leftarrow (4, 6, 8, 10)$ provided respectively that size_of_essay $\geq (\alpha/4, \alpha/2, 3\alpha/4, \alpha)$
    $S \leftarrow 0.8 \cdot S_{\text{RST}} + 0.2 \cdot S_{\text{length}}$
    // Penalize duplicate text and use of vernaculars
    if # duplicate text > threshold then Subtract one from $S$ for each duplicate text
    if exist vernaculars in $E'$ then Subtract one from $S$ for each vernacular word
    return $S$
```

Cohesion of an essay is detected by applying RST (see Section 3.2). We use RST to analyze the structure of the written text, and to detect the coherence of the essay. The text structures in RST are hierarchic, and are represented using RS-tree (rhetorical structure tree). In the AAEE system, we determine the coherence relations in a given essay E' by constructing an RS-tree. If the entire essay can be converted into an RS-tree, then we assume it coherent, otherwise it is not. We determine the quality of the cohesion of the essay by the number of levels in the generated RS-tree. We assume the more levels the RS-tree has, the better the essay is. In this work, we treat any essay yielding an RS-tree of height three or more as worthy of full mark for the cohesiveness (line 3 in Algorithm 2).

Repeated sentences beyond a certain threshold are not acceptable. It is possible that a smart student, knowing that the essay will be auto-graded, may attempt to game the system. One of the simplest way is by repeating the same sentence verbatim, or close to it, multiple times in the essay. Another possibility is the student is incapable of expressing his or her ideas, and therefore repeat the same sentence just to lengthen the essay. In either case, this is an unacceptable practice. We count the number of words that occur sequentially in the essay, that is after we remove the stop-words. The algorithm counts the number of sentences that have the same three consecutive words (we remove the prefix (ال: *Al*) which means "the"). If the number of such sentences is greater than some threshold, then we subtract the number of repeated sentences from the score as a penalty (line 9 in Algorithm 2).

Vernacular words in the essay are penalized. We manually collected and compiled a list of about three hundred vernacular words. In rare cases, certain MSA words have a different meaning when we consider them vernacular. For instance, (نفسي: *nfsy*) which means *myself* in MSA, while in vernacular it means *I wish*. When compiling the list of vernaculars we avoided such words, by giving the student the benefit of doubt and assuming the word is in MSA. It is not infeasible trying to discriminate whether a given word is MSA or vernacular, but it requires going into semantics, and this will certainly complicate our system. As a penalty, we take off one mark for each vernacular word in the essay (line 10 in Algorithm 2).

Length of the essay must be reasonable. We use the average length of the essays in the training set as a guide. Let αbe this average. If the size of the essay E' is greater than α, then the student is entitled to full mark on this criteria. However, if it is shorter then the student gets less mark proportionally. For example, if the length of the essay $\in [\alpha/2, 3\alpha/4)$ then the student gets 60% of the mark designated for this criteria. The length of the essay is based on word count.

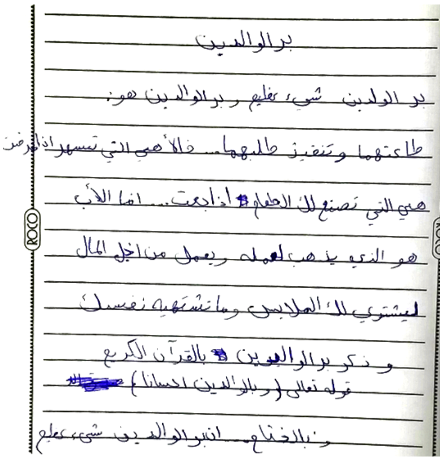
The score for writing style (score_W) is split into 80% of the mark on cohesion, and 20% on the length of the essay (line 8 in Algorithm 2). Then, we penalize the usage of vernaculars and duplicate texts by taking points out of score_W.

4.2.3. Spelling

The words in the essay E' are compared against a set of words in a huge dictionary. We deduct one mark out of 10 for every three typos (line 6 in Algorithm 1), while ensuring the score for the spelling (score_{Sp}) does not go into negative. The dictionary consists of over nine million MSA words Shaalan et al. (2012). These were compiled using AraComLex (Attia et al., 2011), an open-source large-scale finite state morphological transducer to detect MSA spelling error.

5. Evaluation and results

The standard way to evaluate the accuracy of a set of essay grades is to measure how well two independent scorings agree with each other, which in our case is an automatic grader and a human judge. So, for the evaluation of AAEE we manually compiled a corpus of 350 different essays in Arabic, covering eight different topics. These essays were written by students from four different schools (two Intermediate, and two Secondary) in Riyadh, Saudi Arabia. The essays were part of regular assignment, and were graded out of 10 by resident teachers in their respected schools. To eliminate any chance of bias, we had two more human raters. The second human rater graded all the essays. The third human rater was involved when the absolute difference between the first two raters is big. In defining big, we followed (Alghamdi et al., 2014) who set it over 1. Since our essays are scored out of 10, this translates to difference ≥ 2.5 (see Eq. 1 for the details). The final score of each essay is thus the rounded average of 2 (or 3 raters). We call this, their actual score, and it is always integer. It is worth noting, all the human raters scored the essays using whole numbers. The correlation of the two human raters was 0.709, with less than 10% of the essays were marked by the third rater. This is close to the reported correlation of 0.7 in (Alghamdi et al., 2014). As the essays were handwritten, we transcribed them verbatim, and that included the spelling, line-by-line into computer readable format. Figure 2 is a sample handwritten essay in Arabic on the topic "Being dutiful to one's parents", and is followed by line-by-line transcribing and English translation. For those who can read the sample essay (Figure 2) there is a typo in the third line, which we did not edit and it appears in the transcribed text. In the translated text the missing letter appears inside the square bracket.



بر الوالدين

بر الوالدين شيئ عظيم وبر الوالدين هو:

طاعتهما وتنفيذ طلبهما .. فالأهي التي تسهر اذا مرضن

هي التي تصنع لك الطعام اذا جعت .. اما الأب

هو الذي يذهب لعمله ويعمل من اجل المال

ليشتري لك الملابس وما تشتهيه نفسك

وذكر بر الوالدين بالقرآن الكريم

قوله تعالى (و بالوالدين احسانا)

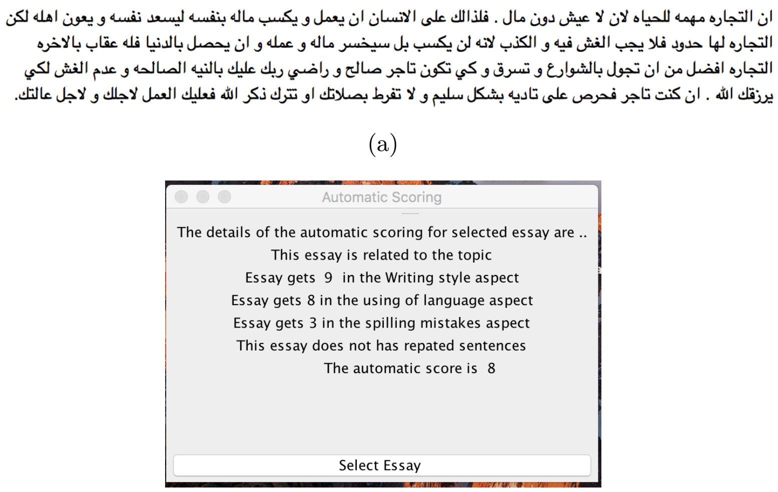
وبالختام .. ان بر الوالدين شيء عظيم

Figure 2: Sample handwritten essay in Arabic.

Being Dutiful to One's Parents	بر الوالدين
Being dutiful to one's parents is a great thing and it is:	بر الوالدين شيئ عظيم وبر الوالدين هو:
Obeying and implementing their request ... the [m]other is the one that watches you if you get sick	طاعتهما وتنفيذ طلبهما ... فالأ هي التي تسهر اذا مرضت
She is the one that make you food if you are hungry ... the father	هي التي تصنع لك الطعام إذا جعت ... اما الأب
is the one who goes to work and works for the money	هو الذي يذهب لعمله ويعمل من اجل المال
to buy you clothes and whatever you desire	ليشتري لك الملابس و ما تشتهيه نفسك
And being dutiful to one's parent is mentioned in the Holy Qur'an	وذكر بر الوالدين بالقرآن الكريم
Allah Says, "and that ye be kind to parents".	قوله تعالى (و بالوالدين احسانا)
In closing ... being dutiful to parents is a great thing	وبالختام ... ان بر الوالدين شيء عظيم

The corpus was split into two disjoint sets: training, and testing. A training set of 300 essays, and the remaining 50 essays as unseen data to be used for testing. For the training phase we use 10-fold cross validation (CV). Generally speaking, in *k*-fold CV, the data set is split into *k* mutually exclusive subsets of approximately equal size. The classification model is trained and tested *k* times. Each time it is trained on all but one fold and tested on the remaining single fold. Each CV yields a model along with an accuracy associated with the model. The practice is to pick the model yielding the best accuracy and test it on the unseen data. The empirical study in (Kohavi, 1995) showed that ten seem to be an optimal number of folds, so we use $k = 10$. Figure 4 shows the actual score distribution for the training, and the testing data sets. As seen, the score distribution in the test set differs from that of the training one. This is because we decided to test the system against a different distribution. This, we believe, is a more natural scenario. Figure 3 is a screen shot of the input essay and the AAEE GUI showing the result of automatic scoring of an essay.

Table 5 shows the list of topics covered by the 350 essays, and the number of essays per topic. For each topic, we tried to maintain the same ratio (≈ 17%) between the number of essays in the training and those in the testing data sets (Figure 5a). We reserved 300 essays for the training phase using 10-fold CV. This translates to 30 essays per fold. To maintain a balance, it is normal that every fold must have essays from each of the eight topics in our corpus. Figure 5b shows the breakdown of each fold. For instance, the fifth fold contains six essays from topic 1, two essays from topic 2, five from topic 3, etc.



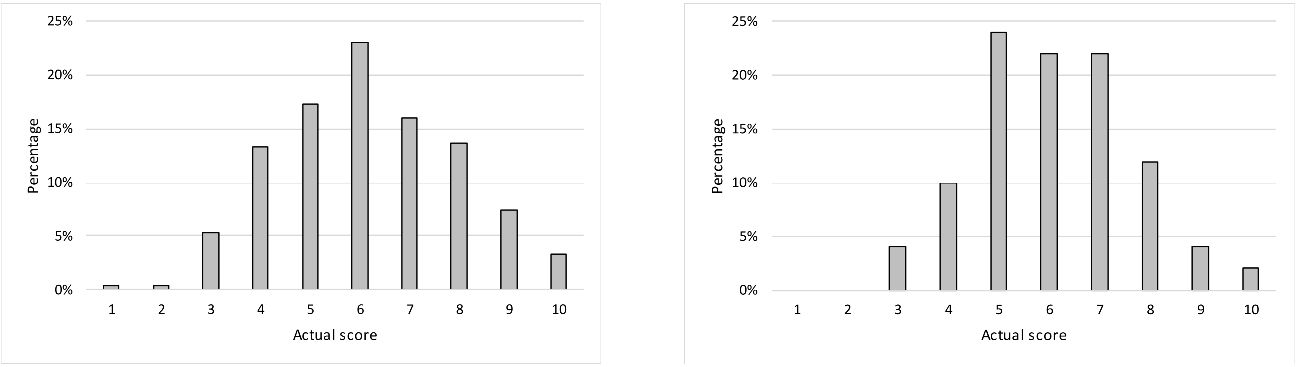
ان التجاره مهمه للحياه لان لا عيش دون مال . فلذالك على الانسان ان يعمل و يكسب ماله بنفسه ليسعد نفسه و يعون اهله لكن
التجاره لها حدود فلا يجب الغش فيه و الكذب لانه لن يكسب بل سيخسر ماله و عمله و ان يحصل بالدنيا فله عقاب بالاخره
التجاره افضل من ان تجول بالشوارع و تسرق و كي تكون تاجر صالح و راضي ربك عليك بالنيه الصالحه و عدم الغش لكي
يرزقك الله . ان كنت تاجر فحرص على تاديه بشكل سليم و لا تفرط بصلاتك او تترك ذكر الله فعليك العمل لاجلك و لاجل عالتك.

(a)

The details of the automatic scoring for selected essay are ..
This essay is related to the topic
Essay gets 9 in the Writing style aspect
Essay gets 8 in the using of language aspect
Essay gets 3 in the spilling mistakes aspect
This essay does not has repated sentences
The automatic score is 8

(b)

Figure 3: Screen shot of (a) input essay, and (b) AAEE GUI showing the evaluation of the input essay.



(a)

(b)

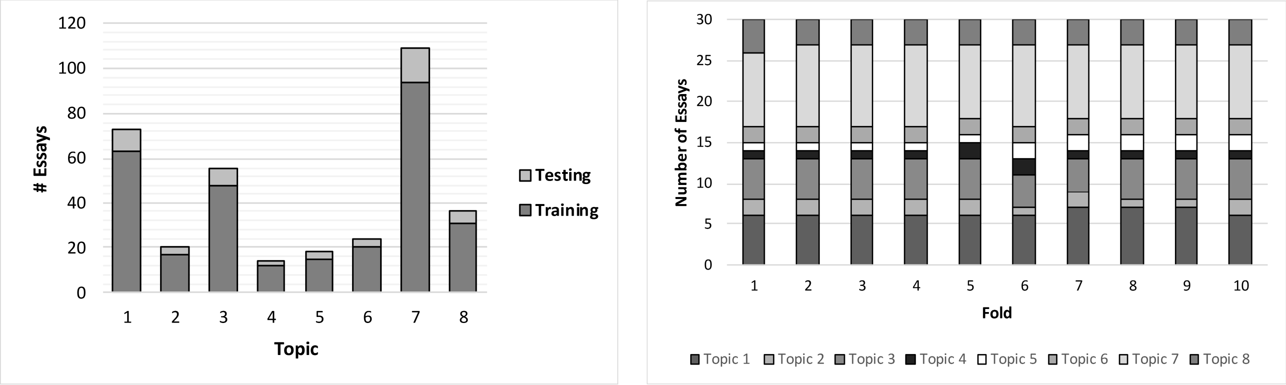
Figure 4: The distribution of the actual score for (a) the training data set, and (b) the testing data set. The size of the training data set is 300 documents (essays), and 50 documents for the testing. The actual score is the average of the human raters (teachers), rounded to the nearest integer.

Following the 10-fold CV and picking the best performing model we are now ready to apply it on the 50 essays which we left aside (unseen data). Each of the test essays will be auto-scored using Algorithm 1. The calculated automatic score is a real number that does not exceed 10. We will discuss the results based on two experiments. In the first experiment the automatic score is reported as is, and in the second experiment it is rounded to the nearest integer. Mathematically speaking, the rounding of positive real number $x \in \mathbb{R}$ is $\lfloor x + 0.5 \rfloor$.

The authors in (Alghamdi et al., 2014) classified the automatic score of the test essays according to the absolute difference between its automatic and the actual score. Let Δ designate this difference. The auto-score was classified as: “exact”, “within range”, or “out of range” depending on the absolute value of Δ. The conditions set forth for this classification reflects (Alghamdi et al., 2014) approach to auto-grading, which is whole numbers out of 6. However, we need to revise the conditions since our essays are auto-graded out of 10 using fractional numbers. Next, we state the original condition used in (Alghamdi et al., 2014),

Table 5: List of different topics covered by the corpus (the full data set), the number of essays per topic, and the number of essays (topic wise) used for training and for testing phase.

Topic	Topic title	Total # essays	Number of essays	
			Training	Testing
1	Unemployment as a terminal disease	73	63	10
2	Positive aspect of business trading	20	17	3
3	An issue within the society	56	48	8
4	Write a story	14	12	2
5	Prayer	18	15	3
6	Tire squeal and traffic problems	24	20	4
7	On social media	109	94	15
8	Being dutiful to one's parents	36	31	5
Total		350	300	50



(a)

(b)

Figure 5: The input corpus of 350 essays contains eight topics. (a) The number of essays in the training and testing data set per topic. (b) Number of essays per topic in each of the ten folds.

and our proposed revision. Later we will justify this redefinition. For the test essay we say that,

$$\text{auto-scores are} \underbrace{\begin{cases} \text{exact (E)}, & \text{if } |\Delta| = 0 \\ \text{within range (WR)}, & \text{if } |\Delta| = 1 \\ \text{out of range (OOR)}, & \text{if } |\Delta| \geq 2. \end{cases}}_{\text{Original condition in (Alghamdi et al., 2014)}} \implies \underbrace{\begin{cases} \text{exact}, & \text{if } |\Delta| < 0.5 \\ \text{within range}, & \text{if } 0.5 \leq |\Delta| < 2.5 \\ \text{out of range}, & \text{if } |\Delta| \geq 3.33. \end{cases}}_{\text{Our revised condition}} \quad (1)$$

To explain the revised condition in Eq. 1, consider the condition for within range (WR). The original condition for WR states that the absolute difference between the auto-score and the actual score is 1. It thus, excludes the condition for exact. Suppose (Alghamdi et al., 2014) auto-graded the essays out of 6 using real numbers. A possible revised condition for WR would be $0.5 \leq |\Delta| < 1.5$. Rounding this value yields $|\Delta| = 1$, same as defined originally in (Alghamdi et al., 2014). For our case, we need to rescale the conditions since our essays are auto-graded out of maximum score of 10 in continuous range. The upper bound for WR becomes $|\Delta| < (^{1.5}/_{6}) * 10 = 2.5$.

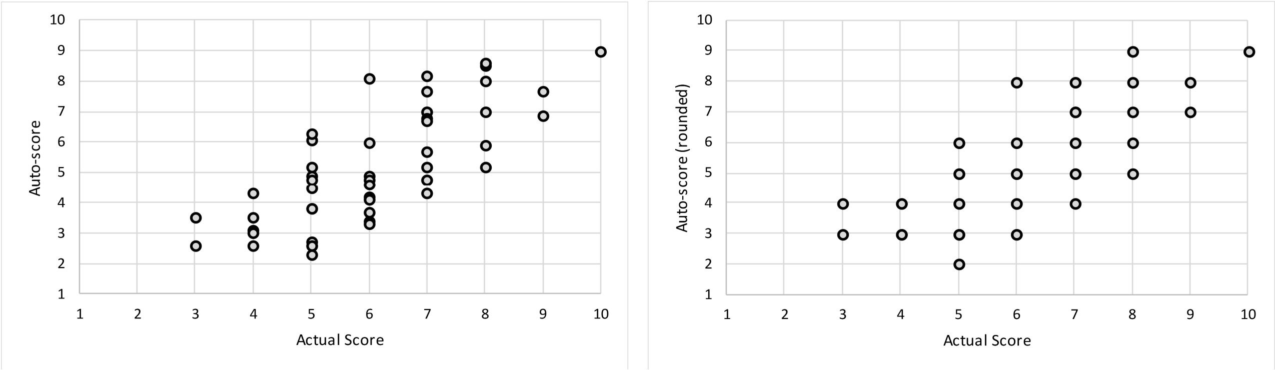
The performance will be measured using two different metrics, namely: accuracy (Acc), and Pearson's r

correlation. Alghamdi et al. (2014) defined the accuracy as,

$$\text{Acc} = \frac{\text{\# of predictions that can be considered correct}}{\text{Total \# of predictions}}, \tag{2}$$

where all the test essays whose auto-score are either exact or within range are considered correct. In other words, the accuracy is the sum of exact and within range percentages (E % + WR %). The Pearson's r correlation is defined as follows. Let n be the number of essays, and let $\{\,(x_i, y_i) \mid i = 1, 2, \ldots, n\,\}$ be the set of score pairs of essay i, where x_i is the actual score and y_i is the auto-score of the same essay. The correlation is given by,

$$r = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{n \sum_i x_i^2 - (\sum_i x_i)^2}\,\sqrt{n \sum_i y_i^2 - (\sum_i y_i)^2}}. \tag{3}$$



(a) Experiment 1

(b) Experiment 2

Figure 6: Detail results for experiments 1 and 2 on unseen data (50 essays): (a) scatter plot of AAEE auto-score vs actual score of all the 50 test essays with auto-score in continuous range, and (b) do but with auto-score rounded to the nearest integer.

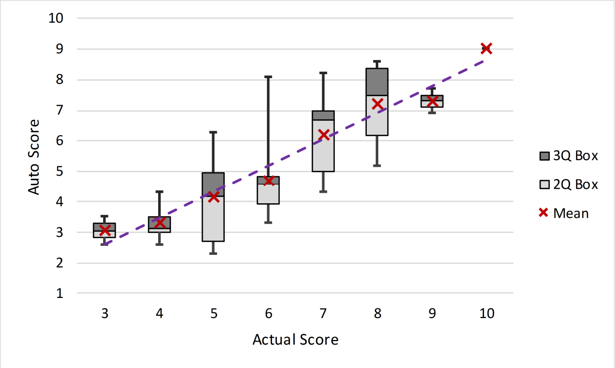


Figure 7: Detail results for experiment 1 on unseen data. Box and whisker plot of actual score vs auto-score as is. Note, the actual score is always integer. The whiskers span the first and the fourth quartile. The lowest and the highest auto-score are marked by the small horizontal bar at the end of the whiskers. The dashed line represents the trend line.

Figures 6-7 shows the detailed performance of evaluating 50 test essays for both experiments. We show various plots to assess AAEE automatic scoring of test essays versus independent scoring by human judge (actual score). The scatter graph in Figure 6a plots the auto-score value exactly as output by AAEE, while in Figure 6b the auto-score is rounded to the nearest integer. Figure 7 is a box and whisker plot showing the auto-score (real numbered) and the actual score. There is very little difference between real valued

Table 6: Summarization of the results on 50 unseen test essays. Experiment 1 reports the result for auto-score expressed as real number, and in Experiment 2 it is rounded to the nearest integer.

	Experiment 1	Experiment 2
E (%)	22%	26%
WR (%)	68%	64%
OOR (%)	0%	0%
Acc	90%	90%
Correlation r	0.756	0.740

Table 7: Comparison of different aspects between our system (AAEE) and Abbir (Alghamdi et al., 2014). Both systems automatically grades essays written in Arabic. The values for Abbir are as reported by its authors.

	AAEE	Abbir
Domain	Middle and high school children	College undergraduates
Human score range	1–10	1–6
Size of training/testing data set	300/50	579/61
# topics covered by the essays	8	2
Acc	90%	96.7%
Correlation r	0.756	0.78

auto-score and the rounded to the nearest integer. This is evident in Table 6 where we summarize the results of both experiments.

From Figure 7 we see that AAEE mostly under-grades the students' essays. In other words, it tends to give lower scores. For instance, 12 test essays (24% of all the test essays) were scored by human raters assigning them 5 marks (see Figure 4b). However, AAEE auto-scored these essays giving them (all but three cases) less than the human judgment. From Table 6, we note that rounding the auto-score did not impact the accuracy (Acc) which remained unchanged at 90%, also the percent of documents designated out of range (OOR) did not change which stayed at 0%. On the other side, it did impact the number of essays which were scored exact (E) from 22% to 26%, while the correlation dipped somewhat. The results we achieved for automatically grading Arabic essays using AAEE in terms of machine-human correlation is 0.756, which is better than the inter-human correlation of 0.709. The accuracy of 90% means that 45 test essays (out of 50) where scored within acceptable range of the human judgment.

Based on the Pearson's correlation we can say it is better to keep the automatic score as is, i.e. not to round them. But, if it was decided that the auto-scores must be whole numbers, then there should be no qualm as the impact on the correlation is not that significant. We believe the closeness of results between experiments 1 and 2 in Table 6 owes to our formulation of the revised condition in Eq. 1.

We wanted to compare the performance of AAEE and Abbir (Alghamdi et al., 2014), the only other system that automatically grades the Arabic essays. We contacted the authors of Abbir to get an executable copy of their developed tool. This would have been a chance to evaluate two different algorithms on a single dataset. Unfortunately, Abbir was no longer maintained, and the lead project investigator was not interested in salvaging it. Table 7 summarized both systems, keeping in mind each system was evaluated on different dataset. The performance (Acc and correlation) reported for Abbir is better than AAEE. There are two factors that contributed to Abbir's better performance. First, the size of the training data set. Abbir's training data set is almost twice as big as ours (579 essays versus 300). Second, Abbir had only two topics covered by the entire data set, while our system had to deal with eight different topics. Simply put, compared to AAEE, Abbir had a larger pool to train from per topic. No doubt, the higher the number of training samples per topic will be reflected in a better overall performance. Nonetheless, we feel our system, AAEE, did an excellent job for the first of its kind.

6. Conclusion and future work

Automated evaluation of essays (AEE) evaluates and marks written essays. AEE uses computers in the assessment of student learning, which is an important part of the teaching process; AAE is one of the newest techniques for assessment of student learning. Automated evaluation of student essays is being used to improve students writing skills, given the increasing need to use technology to improve education. It is easier and faster for students to get feedback about their essays from a system than from an instructor.

The lack of AEE systems for the Arabic language necessitates developing such a system for one of the widely spoken languages. In this work we devised AAEE (Automatic Arabic Essays Evaluator), an AEE system for the Arabic language. The system comprises the latent semantic analysis, rhetorical structure theory, and some additional features. The AAEE is intended to automatically grade Arabic essays written by school children. The system is modeled on the essay grading by school teachers in Saudi Arabia. To assess our system we gathered a data set of 350 handwritten pre-graded essays from four different schools spanning two different school levels. The essays were part of standard classroom assignment, and covered eight different topics. Each essay was graded by two (or sometimes three) human raters. Three hundred essays were used to train the system, and a different set of 50 essays to test it. We did two experiments. In the first experiment the automatic scores were reported as is (real number out of 10), and in the second experiment the automatic scores were rounded to the nearest integer. Both experiments show that 90% of the test essays were correctly scored. The machine-human correlation varied slightly between experiments, 0.756 and 0.740 respectively for both experiments. It is marginally better than the inter-human correlation of 0.709 for the Arabic essays. This study justifies considering AEE for Arabic essays as part of a standardized test setup, e.g. Qiyas GAT test in Saudi Arabia, a SAT like test but with no essay option.

For future work, we intend to improve the evaluation results by adding some additional features into the system. In particular we intend to use word2vec, one of the popular approaches for the semantic similarity. We also intend to improve the evaluation results by collecting a larger number of essays and having each essay graded by more than one teacher to obtain more accurate results.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no. RG-1439-067.

References

Abbas, A. R., & Al-Qaza, A. S. (2014). Automated Arabic essay scoring (AAES) using vector space model (VSM). www.iasj.net/iasj?func=fulltext&aId=103659 [Accessed Sep 10, 2018].

Abbas, A. R., & Al-Qaza, A. S. (2015). Automated Arabic essay scoring (AAES) using vector space model (VSM) and latent semantics indexing (LSI). *Engineering and Technology Journal (University of Technology - Iraq)*, *33B*, 410–426.

Ade-Ibijola, A. O., Wakama, I., & Amadi, J. C. (2012). An expert system for automated essay scoring (AES) in computing using shallow NLP techniques for inferencing. *International Journal of Computer Applications*, *51*, 37–45.

Al-Jouie, M. F., & Azmi, A. M. (2017). Automated evaluation of school children essays in Arabic. *Procedia Computer Science*, *117*, 19–22.

Al-Khatib, M. A. (2001). The pragmatics of letter-writing. *World Englishes*, *20*, 179–200.

Al-Sanie, W. (2005). *Towards an infrastructure for Arabic text summarization using rhetorical structure theory*. Master's thesis Riyadh, Saudi Arabia.

Al-Sughair, S. (2014). Overcrowded saudi classrooms 'hampering learning process'. Arab News. http://www.arabnews.com/saudi-arabia/news/644571 [Accessed Jun 27, 2018].

Alghamdi, M., Alkanhal, M., Al-Badrashiny, M., Al-Qabbany, A., Areshey, A., & Alharbi, A. (2014). A hybrid automatic scoring system for Arabic essays. *AI Communications*, *27*, 103–111.

Asher, N., & Lascarides, A. (2003). Cambridge University Press.

Attia, M., Pecina, P., Toral, A., Tounsi, L., & van Genabith, J. (2011). An open-source finite state morphological transducer for modern standard Arabic. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing* (pp. 125–133). Association for Computational Linguistics.

Azmi, A. M., & Alshenaifi, N. A. (2016). Answering Arabic why-questions: Baseline vs. RST-based approach. *ACM Transactions on Information Systems (TOIS)*, *35*, 6:1–6:19.

Azmi, A. M., & Altmami, N. I. (2018). An abstractive Arabic text summarizer with user controlled granularity. *Information Processing & Management*, *54*, 903–921.

Bin, L., Jun, L., Jian-min, Y., & Qiao-ming, Z. (2008). Automated essay scoring using the KNN algorithm. In *International Conference on Computer Science and Software Engineering* (pp. 735–738).

Burstein, J., & Marcu, D. (2000). Toward using text summarization for essay-based feedback. In *Traitement Automatique de la Langue Naturelle (TALN 2000)* (pp. 245–256).

Chang, T., Lee, C., Tsai, P., & Tam, H. (2008). Automated essay scoring using set of literary sememes. In *Proceeding of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE '08)* (pp. 1–5).

Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)* (pp. 1741–1752).

Chen, H., He, B., Luo, T., & Li, B. (2012). A ranked-based learning approach to automated essay scoring. In *Second International Conference on Cloud and Green Computing* (pp. 448–455).

Devi, M. S., & Mittal, H. (2013). Subjective evaluation using LSA technique. *International Journal of Computers and Distributed Systems*, *3*, 15–19.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgements of writing ability.* Technical Report Research Bulletin RB 61-–15 Princeton, NJ: Educational Testing Services.

Doğan, A., Akbarova, A. A., Aydoğan, H., Gönen, K., & Tuncdemir, E. (2014). Automated essay scoring versus human scoring: A reliability check. *International Journal of Linguistics, Literature and Translation*, *3*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communication of the ACM*, *55*, 78–87.

Dweik, B. (1986). *Research Papers in Applied Linguistics.* Hebron University Press: Hebron.

ETS Research (2017). Automated scoring of writing quality. `www.ets.org/research/topics/as_nlp/writing_quality/` [Accessed Jun 28, 2018].

Ewees, A., Eisa, M., & Refaat, M. M. (2014). Comparison of cosine similarity and k-NN for automated essays scoring. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*, 8669—8673.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: challenges and solutions. *ACM Transaction on Asian and Low-Resource Language Information Processing (TALLIP)*, *8*, 1–22.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with lsa. *Interactive Learning Environments*, *8*, 111–127.

Gomaa, W. H., & Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. *Computer Speech & Language*, *28*, 833–857.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, *15*, 22–37.

Hermena, E. W., Drieghe, D., Hellmuth, S., & Liversedge, S. P. (2015). Processing of Arabic diacritical marks: phonological-syntactic disambiguation of homographic verbs and visual crowding effects. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 494–507.

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

Islam, M. M., & Hoque, A. S. M. L. (2013). Automated Bangla essay scoring system: ABESS. In *International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 1–5).

Johnstone, B. (1991). *Repetition in Arabic Discourse: Paradigms, Syntagms, and the Ecology of Language.* John Benjamins: Amsterdam.

Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society*, *11*, 275—288.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of 14th International Joint Conference on Artificial Intelligence - Volume 2* IJCAI'95 (pp. 1137–1143). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Latif, S., & Wood, M. M. (2009). A novel technique for automated linguistic quality assessment of students' essays using automatic summarizers. (pp. 144–148). volume 5.

Lin, C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries. In S. S. Marie-Francine Moens (Ed.), *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out (WAS '04)* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

Lynch, M. (2018). Why can't most college graduates write a decent essay? [The Edvocate] `https://www.theedadvocate.org/cant-college-graduates-write-decent-essay/` [Accessed Jul 8, 2018].

Mann, W. C., & Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of text structures. In G. Kempen (Ed.), *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics* (pp. 85–95). Dordrecht: Springer Netherlands.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text*, *8*, 243–281.

Mann, W. C., & Thompson, S. A. (1992). Relational discourse structure: A comparison of approaches to structuring text by 'contrast'. In S. J. Hwang, & W. R. Merrifield (Eds.), *Language in Context: Essays for Robert E. Longacre* (pp. 19–45). SIL Dallas.

Meehan, S. (2013). *An Investigation into the Structural Errors of Arabic Learners' Written Persuasive Discourse in English.* Master's thesis University of Lancashire UK.

Mohamed, A. (1993). *A Contrastive Study of Syntactic Relations, Cohesion and Punctuation as Markers of Rhetorical Organisation in Arabic and English Narrative Texts.* PhD dissertation University of Exeter.

Nahar, K. M. O., & Alsmadi, I. M. (2009). The automatic grading for online exams in Arabic with essay questions using statistical and computational linguistics techniques. *MASAUM Journal of Computing*, *1*, 215–220.

Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *47*, 238–243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, *62*, 127–142.

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP '15)* (pp. 431–439).

Rass, R. A. (2015). Challenges face arab students in writing well-developed paragraphs in english. *English Language Teaching*, *8*, 49–59.

Razon, A. R., Vargas, M. L. J., Guevara, R. C. L., & Naval, P. C. (2010). Automated essay content analysis based on concept indexing with fuzzy C-means clustering. In *2010 IEEE Asia Pacific Conference on Circuits and Systems* (pp. 1167–1170).

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

Salem, Z., Sadek, J., Chakkour, F., & Haskkour, N. (2010). Automatically finding answers to "Why" and "How to" questions for Arabic language. In R. Setchi, I. Jordanov, R. Howlett, & J. Lakhmi (Eds.), *Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 586–593). volume 6279.

Shaalan, K. F., Samih, Y., Attia, M., Pecina, P., & van Genabith, J. (2012). Arabic word generation and modelling for spell checking. In *The Eighth International Conference on Language Resources and Evaluation (LREC)* (pp. 719–725).

Shehab, A., Faroun, M., & Rashad, M. (2018). An automatic Arabic essay grading system based on text similarity algorithms. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *9*, 263–268.

Shermis, M., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* chapter 19. (pp. 313–346). New York: Routledge.

Shermis, M. D., & Burstein, J. (2003). Introduction. In M. D. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. xiii–xvi). Manwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay assessment: Current applications and new directions*. New York, NY: Routledge.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* chapter 1. (pp. 1–15). New York: Routledge.

Smith, T. (2018). More states opting to 'robo-grade' student essays by computer. NPR. `www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer` [Accessed Jul 2, 2018].

Taboada, M., & Stede, M. (2009). Introduction to RST (Rhetorical Structure Theory). `edu.cs.uni-magdeburg.de/EC/lehre/wintersemester-2011-2012/dokumentverarbeitung/folien-und-materialien/RST_Introduction.pdf` [Accessed Jul 14, 2018].

White, E. M. (1993). Holistic scoring: Past triumphs, future challenges. In M. M. Williamson, & B. A. Huo (Eds.), *Validation of holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79–108). Cresskill, NJ: Hampton Press.

Zhai, C. X. (2008). Statistical language models for information retrieval a critical review. *Foundations and Trends in Information Retrieval*, *2*, 137–213.

Zupanc, K., & Bosnic, Z. (2014). Automated essay evaluation augmented with semantic coherence measures. In *Proceedings of 14th IEEE International Conference on Data Mining* (pp. 1133–1138).

Zupanc, K., & Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica*, *39*, 383–395.