3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

Automated Evaluation of School Children Essays in Arabic

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

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

Abstract

The most effective technique for improving students writing skills is for them to get immediate instructor's feedback and as often as possible. This, however, significantly expands the workload of the instructors. There is a growing need for automated systems to help students draft essays. Automated essay evaluation is increasingly popular in the field of educational evaluation technology. In this work, we present an automatic evaluator of student essays in Arabic language, a system that is modeled on the scheme followed by the school-teachers in Riyadh, the capital of Saudi Arabia. The main criteria for assessing the essays is: language proficiency; the structure of the essay; and the content which should match the topic. With this in mind, we developed a scheme that relies on latent semantic analysis, and rhetorical structure theory. The system was tested on over 300 different essays, all handwritten by schoolchildren covering various subjects. The performance was measured by machine-human correlation in grading. Our system achieved an overall correlation of 0.79 with the teachers' evaluation.

Keywords: Automatic essay scoring; Essay evaluation; Latent Semantic Analysis; Rhetorical Structure Theory

1. Introduction

Automatic evaluation of essays (AEE) is computer program that evaluates essays that written by students. The automated essay feedback has appeared with the advent of online tuition systems. There is an increasing need for such systems to help students draft essays. A field of Natural Language Processing (NLP), AEE automatically provides feedback for essays that humans write in natural language. The AEE system is also known as automated essay scoring and, automated essay grading.

In 1966, Ellis Page published his (controversial then) paper, "The Imminence of … Grading Essays by Computer," which discusses the use of computers to evaluate essays and provide feedback [6]. Page published the article to explain his ideas for the development of Project Essay Grade (PEG) [7]. Some additional forces at work ultimately facilitated the future of AEE. These included the creation and widespread adoption of the AEE systems to evaluate writing in different languages.

We may simply define AEE as a way of evaluating written prose automatically by the computer [10]. Evaluation means that the computer system can engage in the task of scoring an essay or assigning a number to it. The availability of AEE systems is necessary for the more direct improvement of writing quality and because of the growing need for large-scale testing programs for English, such as TOEFL and GMAT. There are many approaches to implement efficient AEE systems, and there do exist many successful AEE applications. Efficient AEE systems are implemented through various NLP techniques, and includes aspects of information retrieval (IR), and machine learning (ML).

There is a dire need for developing AEE systems for the Arabic language. The Saudi National Center for Assessment (or Qiyas) conducts General Aptitude Test (GAT), a sort of online-standardized test that is required for any student seeking admission to a

* Corresponding author. Tel.: +966-11-467-6574; fax: +966-11-467-5423.
  E-mail address: aqil@ksu.edu.sa

national college/university in the country. Though GAT evaluates proficiency of the Arabic language, it lacks essay writings. Obviously, it is owed to the lack of an automated essay grading system. Countrywide, over a million students take this test each year. The sheer number of students taking GAT test makes it impractical to consider manual assessment of the essays. It is hoped our system will help the Qiyas Center to seriously consider automated assessment of the essays written in Arabic language.

In this paper we briefly describe our proposed system for evaluating school children essays in the Arabic language. The paper is organized as follows. In Section 2 we go over the related work. Design of the system is in Section 3. In Section 4 we evaluate and discuss its performance, and finally we conclude in Section 5.

2. Related work

There are many works for essays in English language, see for example [9][2][8], and for other languages, e.g. [3][4]. Latent Semantic Analysis (LSA) has produced promising results in content analysis of essays, e.g. [9][3]. With very few works, the work on automated evaluation of Arabic essays lags behind. We go over what some of the more recent works on evaluating Arabic essays.

Alghamdi et al. [1] presented a hybrid AEE system for evaluating Arabic essays that makes use of efficient reduced dimensionality for LSA (LSAD). Some of the features that were used for the assessment are: spelling mistakes, and the proportion of spelling mistakes for the given length of the essay. For the dataset, the authors collected around 600 essays written by university students in Saudi Arabia. The essays were part of a test in Arabic language course. The length of an essay ranged between 100-200 words. It is a two-phase system, the training phase and the runtime phase. The training phase involves some pre-processing which includes Buckwalter stemming. This phase is made up of three parts: vector of words, vector of spelling mistake, and the LSA concept space. In the runtime phase, the input essays pass through a number of processes. These processes make it possible to get the minimum cosine distance (cosine distance LSAD) between the input essays and the training essays. The size of the LSAD vector is the essay's score. In this particular system, there are six marks. During this phase, they use linear regression approach to obtain features that reflect the human senses. The authors reported an accuracy of 96.72% on the test data. The correlation result between this system's score and the evaluation by humans was 0.78.

Nahar and Alsmadi [5] presented a system for grading online exams in Arabic involving essay questions. Unlike multiple-choice questions, where grading is straightforward, this is more challenging. The idea is to score the student answer against the model answer by the instructor. The authors used different statistical distributions to give weights to the keywords in the model answer. The instructor determines the weights, which tells how important the keyword is. There is a provision to handle synonyms in the student's answer; this however requires synonym words to be manually added into the system. To score the student's answer, the system needs to measure the distance between both answers (student and the model). The paper does not go beyond the schemes; it does not evaluate the system on some real exam dataset, so to compare the automatic grading with manual grading.

3. Our Proposed System

Our objective is to develop a system to automate the evaluation of schoolchildren essays written in Arabic. All the children belonged to the Intermediate level school. That is covering grades 7 to 9 inclusive. The assessment criteria is based on an online survey of Intermediate level school-teachers in Saudi Arabia. According to the survey. the criterias are: spelling and grammar mistakes, the coherence and organization of the essay, the essay should be related to the topic, and sticking to Modern Standard Arabic (MSA) words. There was no general agreement on how much weight to assign to each of the criterias, however, the consensus was 4 marks (out of 10) for spelling mistakes, 3 marks for grammar mistakes, and 3 marks for the organization of the essay.

To solve the problem at hand we opted for a hybrid approach that combines latent semantic analysis (LSA), rhetorical structure theory (RST), and some other features that we will cover latter in the paper. One reason for this approach is the need to assess essays by focusing on elements such as cohesion. This hybrid approach applies LSA for the semantic analysis of the essay, and the RST to assess the cohesion and the writing style of the essay. In our design, we assign 50% of the total score on the cohesion of the essay, 40% for writing style and the remaining 10% for spelling mistakes.

We already noted that LSA has been successfully applied to automate giving grades and feedback on free-text responses in several systems. The basic assumption behind LSA is that there is a close relationship between the meaning of a text and the words in that text. The power of LSA lies in the fact that it is able to map the essays with similar wordings closer to each other in the vector space. The LSA method is able to strengthen the similarity between two texts even when they do not contain common words. The general architecture of our multiple processes AEE system is shown in Figure 1.