Computers & Education 76 (2014) 30–41

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

A system for formative assessment and monitoring of students' progress

Fátima Rodrigues*, Paulo Oliveira

GECAD – Knowledge Engineering and Decision-Support Research Center, Institute of Engineering – Polytechnic of Porto (ISEP/IPP), Portugal

ARTICLE INFO

Article history:
Received 29 December 2013
Received in revised form
28 February 2014
Accepted 4 March 2014

Keywords:
Computer-assisted assessment
Free-text answers
Natural language processing
Semantic meaning
Automatic feedback

ABSTRACT

Assessment plays a central role in any educational process as a way of evaluating the students' knowledge on the concepts associated with learning objectives. The assessment of free-text answers is a process that, besides being very costly in terms of time spent by teachers, may lead to inequities due to the difficulty in applying the same evaluation criteria to all answers. This paper describes a system composed by several modules whose main goal is to work as a formative assessment tool for students and to help teachers creating and assessing exams as well monitoring students' progress. The system automatically creates training exams for students to practice based on questions from previous exams and assists teachers in the creation of evaluation exams with various kinds of information about students' performance. The system automatically assesses training exams to give automatic feedback to students. The correction of free-text answers is based on the syntactic and semantic similarity between the student answers and various reference answers, thus going beyond the simple lexical matching. For this, several pre-processing tasks are performed in order to reduce each answer to its more manageable canonical form. Besides the syntactic and semantic similarity between answers, the way the teacher evaluates the answers is also acquired. To accomplish that, the assessment is done using sub scores defined by the teacher concerning parts of the answer or its subgoals. The system has been trained and tested on exams manually graded by History teachers. There is a good correlation between the evaluation of the instructors and the evaluation performed by our system.

1. Introduction

Periodical information about a student's state or level of his knowledge is crucial for an efficient and personalized learning. Several strategies have been developed to make teaching more effective and accessible to a larger number of students. E-Learning is a true realization of these goals because it is always accessible and allows students to practice their levels of knowledge anytime. Another advantage of an E-learning system is its ability to provide with immediate feedback, which can be used by the students to identify their main misconceptions. E-learning systems are also useful for teachers to know how well the concepts have been understood and to allow them to keep track of each student and the whole class's learning progress. Most of E-learning tools are based on the use of the so-called objective-type questions i.e. multiple choice, selection/association and visual identification. However, some aspects of complex achievement corresponding to the higher levels of the Bloom's taxonomy (Bloom, 1974) are difficult to measure using objective-type questions. Free-text answers require students to recall, organize and integrate ideas and are considered the most adequate form to evaluate educational skills and competencies. For these reasons the (open-question based) assessment through open questions is still a widely used method for evaluating knowledge, being an essential part of any teaching-learning process.

The assessment of questions in an evaluation process involving a large number of free-text answers presents teachers with three major problems. Firstly, this procedure is very time expensive; according to Mason and Grove-Stephensen (2002) evaluation is a task to which teachers dedicate approximately 30% of their time. In addition to this, it is also very difficult to ensure an equitable application of evaluation criteria. This is due not only to the subjective nature of the assessment of free-text answers but also to the lengthy evaluation process.

* Corresponding author.
E-mail addresses: mfc@isep.ipp.pt (F. Rodrigues), pjo@isep.ipp.pt (P. Oliveira).

http://dx.doi.org/10.1016/j.compedu.2014.03.001
0360-1315/

Teachers must be highly concentrated for long periods of time, and therefore assessment is subject to variations of level of concentration and of mood. This can lead to different evaluation grades for answers with similar quality, thus creating inequities in the assessment process that could be even more pronounced if the evaluation is conducted by different evaluators. Moreover, such task involves human labor that cannot be reused.

In the latest Computer-Assisted Assessment (CAA) systems, assessment strategies are based on the existence of a correct answer for each question to be used as a Reference Answer (RA) in the evaluation of Students Answers (SAs). The similarity between the RA and the SAs is determined using traditional Information Retrieval (IR) techniques based on co-occurrence of terms. Such methods are usually effective when dealing with long texts, because similar long texts usually contain a high degree of co-occurring words. However, in short free-text answers, word co-occurrence may be rare or even null, and yet have similar meanings. Also, the use of only one RA can be very penalizing because of the multiplicity of possible correct different answers to a question (Noorbehbahani & Kardan, 2011). Another negative aspect of these systems is not considering the teacher criteria evaluation, but only the similarity between the RA and SAs.

We have dealt with these aspects by developing Assisted Study (AssiStudy) system, which for students works as a formative assessment tool whilst helping teachers creating and assessing exams as well monitoring students' progress. AssiStudy system automatically creates training exams for students to practice, built with questions from past exams, picked from a question repository. These training exams are student-personalized, they are defined based on the questions already answered by the student in order not to repeat the same question correctly answered and, whenever possible, progressively increasing the level of difficulty of the questions. The system also assists teachers in the creation of evaluation exams with various kind of information about students' performance. The evaluation exams are constructed by the teachers using three main types of questions: enumeration, specific knowledge and essay. The training exams are automatically assessed in order to provide instantaneous feedback to students. The same assessment procedure is applied to evaluation exams, but these are also checked by the teacher in order to have a term of comparison between AssiStudy system assessment and teacher's assessment. The automatic assessment is performed using various Natural Language Processing (NLP) techniques to put the RA and SAs into a canonical form easier to process. Several paraphrases of the RA are derived using a semantic network in order to turn the answers comparison more effective. The similarity between answers is calculated combining corpus-based and knowledge-based measures, which are more adequate to evaluate the similarity of short answers with few common terms.

The remainder of this paper is organized as follows: Section 2 provides with an overview about some approaches of computer-assisted assessment. In Section 3 a description of our system is made in terms of its main modules and in the following subsections each module is described with more detail. In the following section a comparison between the assessment results delivered by AssiStudy and the teachers' assessment results, as well as the impact of the use of AssiStudy in the students' marks, is presented. In the last section, conclusions and future work are disclosed.

2. An overview of approaches on computer-assisted assessment

CAA of students' essays is an interesting development domain that has been ongoing since the 1960s up to today (Larkey, 1998; Malmberg, Jarvenoja, & Jarvela, 2013; Pérez-Marín, Pascual-Nieto, & Rodriguez, 2009). CAA systems can be distinguished by the way they primarily evaluate essays either for style or for content, or both. Another distinction criterion is the approach adopted for assessing style and/or content. The most important approaches found in existing CAA systems are Statistical, Latent Semantic Analysis (LSA) and Natural Language Processing (NLP). The first CAA systems, which were focused on statistical approaches, captured only the structural similarity of texts. The following systems, which were based on LSA, did more than a simple analysis of co-occurring terms. In fact, they introduced two new approaches to the problem: a comparison based on a corpus and an algebraic technique that allowed identifying similarities between two texts with different words (Thomas, Haley, DeRoeck, & Petre, 2004). The latest systems are based on NLP techniques and can do intelligent analyzes that capture the semantic meaning of free-text documents. As mentioned above, the distinction is made between grading essays based on content and those based on style. While there are systems that evaluate primarily based on style, e.g. Project Essay Grade (PEG) (Page, 1994); or on content, e.g. Intelligent Essay Assessor (IEA) (Callear, Jerrams-Smith, & Victor, 2001), Educational Testing Service (ETS I) (Whittington & Hunt, 1999), Conceptual Rater (C-Rater) (Burstein, Leacock, & Swartz, 2002), most of the latest systems aim to grade across both dimensions, e.g. Bayesian Essay Test Scoring sYstem (BETSY) (Rudner & Liang, 2002), Automark (Mitchell, Russell, Broomhead, & Aldridge, 2002), Paperless School free text Marking Engine (PS-ME) (Mason & Grove-Stephensen, 2002). Another feature of these systems is that they widely differ in the methods used. For example, IEA is based on LSA, whereas E-rater is based on NLP, BETSY uses Bayesian networks and PEG works on linguistic features via proxies.

More recently, Educational Data Mining (EDM) has emerged as a relevant research field (Baker & Yacef, 2009). EDM aims to provide insight into instructional practices and student learning, often using data from assessments and learning experiences, both formal and informal (Romero, Ventura, Pechenizkiy, & Ryan, 2011). The works concerning EDM has been organized into four functionalities: student modeling, tutoring, content and assessment (Peña-Ayala, 2014). Regarding assessment, which is more closely related to our work, several recent works may be mentioned. TeamAnalytics (Kim, Shaw, Xu, & Adarsh, 2012) is an instructional tool that facilitates the analysis of the student collaboration process by creating dynamic summaries of team member contributions over time. Also, López, Luna, Romero, and Ventura (2012) have developed a specific Moodle mining tool that permits predicting final marks on the basis of forum data in a Moodle course. The authors compare the performance and usefulness of different data mining techniques for classifying university students through the use of different datasets from Moodle courses. Another work studied the learning patterns that emerge in learning situations that are favorable and challenging (Malmberg et al., 2013). The authors identify differences between high and low-achieving students' strategic actions in varying learning situations. Their outcomes unveil that both kinds of students adopted similar strategies in favorable learning situations. However, when the learning situation was considered challenging, high-achieving students used deep strategies for learning, whereas low-achieving students did not. An interactive environment, named Science Assistments, (Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012), is being developed to assess students' inquiry skills as they engage in an inquiry using science Microworlds. In order to support both assessment and the provision of scaffolding, the environment makes inferences about student inquiry skills using models developed through a combination of text replay tagging, a method for rapid manual coding of student log files and educational data mining.