AUTOMATIC EVALUATION OF CONTINUOUS ASSESSMENT TESTS

Mireia Farrús
Universitat Oberta de Catalunya
mfarrusc@uoc.edu

Marta R. Costa-jussà
Barcelona Media Innovation Centre
marta.ruiz@barcelonamedia.org

Germán Cobo
Universitat Oberta de Catalunya
gcobo@uoc.edu

David García
Universitat Oberta de Catalunya
dgarciaso@uoc.edu

Luis Villarejo
Universitat Oberta de Catalunya
lvillarejo@uoc.edu

Rafael E. Banchs
Barcelona Media Innovation Centre
rafael.banchs@barcelonamedia.org

ABSTRACT

The UOC is an online university where students follow an online continuous assessment system in order to successfully complete their studies. Due to the growing number of students, the task of marking their assessment tests has become tedious and time-consuming. In order to speed up and standardise the task, the UOC started to develop a pre-evaluation system to provide a preliminary, automatic evaluation of the tests by using the latent semantic analysis technique. Given a set of ideal reference answers for the evaluation tests, the students' answers are evaluated with respect to the reference, providing an estimated score. This paper shows some preliminar results, from which we can conclude that the correlation between automatic and human evaluations is higher and statistically significant when both the math formulas codification and the language used by the students are the same than in the reference answers.

1. INTRODUCTION

Assessment in education is the process of obtaining, organising and presenting information about what and how the student is learning, by using several techniques during the teaching-learning process. Assessment is especially useful when evaluating open answers questions, since they allow teachers to know more in depth the assimilation of the student in the subject. As Tyner (1999) stated, in some cases, students with high punctuation in closed answer tests report subjacent conceptual errors when being interviewed by a teacher.

The Universitat Oberta de Catalunya (Open University of Catalonia, UOC) is an online university based in Barcelona with more than 54,000 students. Over 2,000 tutors and faculty work alongside an administrative staff of around 500 to provide services to all these students. Students follow a continuous assessment system in order to complete their studies successfully. Assessment is carried out online throughout the semester. At the end of the year, a final test is held on at one of the university's centers.

Due to the growing number of students each year, the task of marking their continuous assessment tests has become tedious and time-consuming. Likewise, more external tutors are needed to carry out this task, which makes it difficult to come to agreement on criteria.

In order to speed up and standardise the task, the UOC has developed a pre-evaluation system which aims to provide a preliminary, automatic assessment of continuous tests assignments. To this end and following (Miller, 2003), a technique based on latent semantic analysis (LSA) is used. LSA is a natural language processing algorithm that analyses semantic relationships between a set of document and the terms they contain. The use of a computer with assessment purposes can be described under the following terms:

- Computer-Aided Assessment (CAA)
- Computer-Mediated Assessment (CMA)
- Computer-Based Assessment (CBA)
- Online Assessment

The goal of the present task involves the development of a text-free assessment tool through internet (e-assessment), so that it would lie in the CBA group. The main objectives of the task are based on achieving the most innovative features of the assessment strategy:

1. To use several techniques to monitories the progress in the conceptual understanding of the student during the learning process.
2. To optimise and improve the interpretation of the information obtained by means of assessment techniques.
3. To reduce, and even to eliminate the use of objectives exams and other traditional evaluation techniques.

On the other hand, the aims of the task include also achieving and consolidating the advantages of a system with such characteristics (Brown et al., 1999):

1. To reduce the professors work charge by automating part of the student evaluation task.
2. To provide the students with the detailed information about their learning period in a more efficient way than in the traditional evaluation.
3. To integrate the assessment culture to the students daily work in an e-learning environment.

Although there are many works in the literature oriented to automated essay scoring (AES) research (Miller, 2003; Shermis, 2003), to the best of our knowledge, there is no AES research oriented to math and/or technical subjects. This paper tries to apply success techniques for AES to a university engineering subject written mostly in math language.

The structure of this paper is as follows. Next section introduces the application framework of the assessment task at the UOC. Section 3 presents a brief description of the Latent Semantic Analysis technique used to evaluate the tests. The assessment experiments and results are presented in section 4 and, finally, the conclusions of the study can be found in section 5.

2. WORKING FRAMEWORK

As previously stated, one of our main aims is to develop a tool in order to help teachers in their continuous assessment tasks of a large number of students. However, the experiments presented in this paper are preliminary, since we are taking the first steps in the development process of the tool. Therefore, these first experiments involve a controlled and relatively small amount of students, in order to establish the groundwork for further and more complete experiments. Thus, the application framework of the preliminary experiments showed in this paper covers the students in two consecutive semesters (with 54 and 70 registered students, respectively) of a single UOC's subject called *Fonaments Tecnològics II* (Circuit Theory, CT).

CT is a core subject belonging to the first year of UOC's Telecommunications Engineering Grade, whose matter is the analysis of analog electronic circuits (a general vision of analog electronics and an introduction to the frequency domain are given to the students). Its objectives and contents are structured in the six following modules:

- **Module 1: Electrical circuits**. The fundamentals of analog circuits, their basic elements and the laws they're governed by are shown. The Kirchhoff's laws and the differences between direct current (DC) and alternating current (AC) are introduced.

- **Module 2: RLC circuits**. The behavior and modeling of capacitors, inductors and diodes are explained, as well as the systematic analysis of RLC circuits in static regime and the nodal and mesh analysis techniques.
- **Module 3: Dynamic circuits**. The Laplace Transform and its usefulness are presented, and concepts like network function, stability, and natural and forced response are established.
- **Module 4: AC circuits**. The fundamentals of AC circuits analysis by means of phasors are shown. Concepts like impedance, admittance resonance and power are introduced, as well the circuits with transformers.
- **Module 5: Analog filters**. Frequency-domain-based analysis of circuits is studied in depth. Both Bode magnitude and Bode phase plots, and the basic filters are explained.
- **Module 6: Applications**. The transistor and his behaviour are introduced, as well its main applications: the operational amplifier, regulators, oscillators… The negative feedback concept and some examples of real circuits are explained too.

The assessment model of the subject contemplates, apart from the single Final Test that takes places at the end of the semester, four different single Continuous Assessment Assignments (CAAs) distributed in the course of the semester and a single Practical Work that includes computer simulation exercises. These four CAAs are structured as follows. The first three CAAs are made up of two different sections: a short questions section and an exercises section. The fourth and last CAA contains an exercises section only. More specifically, the short questions sections consist of a set of 5-6 questions about very concrete issues. Each of these questions is attached with four possible answers, where one of them is only correct, in such a way that the students have to specify the correct answer and reason their choices out with a brief justification. Due to the technical nature of the subject matter, mathematical equations usually appear in both questions and answers wordings, as well as in the students' corresponding justifications. Finally, in order to complete the assessment model description, the relationship between the four CAAs and the six modules that constitute the whole subject is detailed below:

- **CAA 1 – Modules 1 & 2**
- **CAA 2 – Module 3**
- **CAA 3 – Modules 4 & 5**
- **CAA 4 – Module 6**

In this context, the short questions sections of the first three CAAs of the CT subject have been chosen as specific application framework to take place our automatic evaluation experiments, due to the suitability of the structure and length of both the questions and answers, as well as to the nature (short text + few mathematical equations) of the justifications the students have to provide.

3. LATENT SEMANTIC ANALYSIS

The task of evaluating a document in our education context implies judging the semantic content of such document. To this end and following Miller (2003), the latent semantic analysis (LSA) technique has been chosen in order to compare the documents in the concept space.

For the document comparison and/or document retrieval, documents are typically transformed into a suitable representation, usually a vector-space model. A document is represented as a vector, in which each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is *tf-idf* (term frequency inverse document frequency weighting. The tf-idf weight defines statistically how important a word is to a document in a collection. Such a representation is known to be noisy and sparse. That is why in order to obtain more efficient vector-space representations;

space reduction techniques using LSA (Deerwester, 1990) and probabilistic latent semantic analysis (Hofmann, 1999) are applied.

The space reduction technique using LSA consists in performing a singular value decomposition (svd) over the tf-idf matrix. Then, out of this decomposition the *k* largest singular values and their corresponding singular vectors are chosen, which gives the rank *k* approximation to the original *tf-idf* matrix. The more important thing of this approximation is that the new reduced space is supposed to capture semantic relations among the documents in the collection. Figure 1 shows a schematic representation of the use of latent semantic analysis for automatic essay scoring.

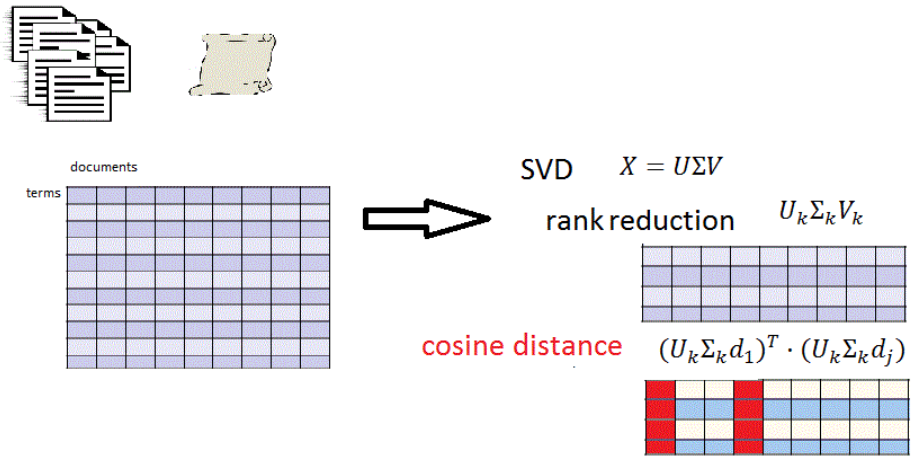

Figure 1. Schematic representation of the use of latent semantic analysis for automatic essay scoring

Finally, the cosine distance similarity measure among each exam and its solution in the reduced space is calculated. As a result, a score is obtained, which shows how a particular set of exams is similar in semantics with their corresponding solution.

4. ASSESSEMENT EXPERIMENTS

This section describes the assessment experiments performed in this study. Section 4.1 presents the experimental framework of the continuous assessment tests, and section 4.2 shows the results obtained in the preliminar experiments.

4.1 Experimental framework

As stated above, students at the UOC follow a continuous assessment through a series of tests called *Proves d'Avaluació Continuada* (Continuous Assessment Tests, PAC). The preliminar experiments carried out in the current paper used the PACs from two consecutive semesters: Spring Semester in 2008 (code 20082) and Fall Semester in 2009 (code 20091), with 54 and 70 registered students, respectively. Each semester included a set of 3 different PACs. (PACs1, PACs2 and PACs3). Especifically, semester 20081 included 20 PACs1, 19 PACs2 and 15 PACs3, while semester 20092 included 28PACs1, 25 PACs2 and 20 PACs3. Table 1 shows the statistics of the material used in the experiments.

The material used in the current paper presented three main problems. On the one hand, the students delivered their PACs in different formats, mainly PDF, Word and Open Office Writer. Some of them were even scanned documents pasted as image files in Word or Writer documents. Therefore, not all the PACs could be transformed into TXT format; PDF PACs and those PACs containing image files were removed, and the final number of PACs was reduced, as shown in Table 1. The table also shows the size of the vocabulary for each PAC and semester collection. It can be seen that the vocabulary size is not correlated with the number of PACs.

Table 1. Registered students, number of PACs and vocabulary size used for each semester

Semester	Students	PAC1		PAC2		PAC3	
		number	vocabulary	number	vocabulary	number	vocabulary
20082	54	14	857	13	730	10	712
20091	70	20	1027	9	699	16	1291

On the other hand, given that we are using an approach that uses bag-of-words, the other two problems in the remaining material are the following. First, the formulation extracted from Open Office documents was coded with MathML, while the formulation extracted from Word documents was not, which makes a big difference between PACs regarding the final vocabulary. Second, the students submitted the PACs in both Catalan and Spanish languages. In this case, we assume that the method presented in the current paper is able to take advantage of the vocabulary that is language independent such as the math variables and others.

4.2 Results

In order to carry out the preliminary assessment experiments, PAC1 and PAC2 from the spring semester of 2008 (20082) were used as development material. This development material allowed concluding that the best rank reduction was 5.

The results obtained are reported in Table 2, which shows the correlation between the qualifications obtained automatically (using LSA and the cosine distance as explained in section 3) and the human qualifications, together with the statistical significance of the results.

Table 2. Correlations between automatic and human evaluations obtained in the preliminar experiments

Semester	PAC1	PAC2	PAC3
2008	16% (p=0.60)	12% (p=0.68)	15%(p=0.68)
2009	52% (p=0.04)	69% (p=0.04)	29% (p=0.27)

As it can be seen from the table, in significant results ($p<0.05$), correlation varies from 52% to 69%. Although they are a little bit behind the ones presented in Miller (2003), they are promising given that we are not facing a complete textual subject. The rest of the results are not significant.

In order to understand the results obtained, the material used in the experiments has been carefully analysed. On the one hand, the reference answers were written in Catalan, while the students could choose whether to answer the tests in Catalan or Spanish, so that the language of the tests was not the same in all the students' PACs. On the other hand, unlike the students' PACs, all the reference solutions were available in Writer format. Since only the mathematical formulas of the Writer documents were transformed into MathML, there was also disparity in the formulas in each set of PACs.

Therefore, and in order to see how these disparities could have affected the results, we computed the percentage of PACs in each set that satisfied the following two requirements at the same time, i.e. the same two requirements satisfied by the reference solutions:

1. The formulas were coded in MahtML
2. The students answered in Catalan language

The percentage of PACs satisfying both requirements are shown in Table 3. It can be seen that the two significant results with a correlation over the 50% correspond to those in which the codification and the language used as the same as the reference solutions in more than 25% of the cases. Therefore, it could be stated from the results that the correlation results depend on the coherence of both the math codification and the language used in the tests.

Table 3. Percentatge of PACs satisfying the same requirements than the reference solutions

Semester	PAC1	PAC2	PAC3
2008	14%	15%	10%
2009	30%	28%	25%

5. CONCLUSION

This paper has presented some preliminar experiments in order to analyse the possibility of automatically evaluating the continuous assessment tests carried out in an online university. One of the future works derived from the present study is to embed this free-text assessment tool as part of virtual classrooms in UOC's web-based teaching-learning environment, in order to help students' self-assessment by providing them with an instant feedback. Thereby, adult e-learners, who usually have lack of time, do not have to await teachers assess to be graded.

The study carried out in this paper has had to overcome some problems regarding the available material. First, the existence of a lot of mathematical formulas in the subjects treated. Although many research works have dealt with automated essay scoring, as far as we are concerned, they have not dealt with math language. Moreover, the students' tests are available in different languages and file formats, which makes even more difficult to treat the mathematical formulas by converting them into a homogeneous code.

Nevertheless, despite the difficulties in the material used, the preliminar experiments have shown some interesting results. After computing the correlation between the automatic and the human assessment tests it was shown that only two from the six evaluation tests performed provided correlation above 50% with statistically significant results. These two sets correspond to those set of PACs that have more similarity with the reference solution PACs: the math formulas are coded in MathML and the students answers were mostly written in the same language.

Although for the time being the correlation results are not satisfactory enough, they have set a starting point that allow us to deal and work with this kind of material in engineering subjects. These are only preliminar experiments and we intend to progress in the future. Thus, future work will focus on improving the format of the materials and give coherence to them (i.e. by using the same formulation and dealing with the language issue) and, additionally, we plan to experiment with non-linear space reduction like multidimensional scalability in order to find further semantic similarities.

ACKNOWLEDGEMENT

This work has been partially funded by the Universitat Oberta de Catalunya under the Teaching Innovation Project number IN-PID1043 and by the Spanish Department of Education and Science through the Juan de la Cierva fellowship program and the BUCEADOR project number TEC2009-14094-C04-01. The authors also want to thank the Barcelona Media Innovation Centre for its support and permission to publish this research.

REFERENCES

Brown, S.; Race, R., & Bull, J. (1999). *Computer-assisted assessment in higher education.* Kogan Page.

Deerwester, S., Dumanis, S., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Hofmann (1999). Probabilistic Latent Semantic Analysis. *Proceedings of Uncertainty in Artificial Intelligence, UAI99I*, 289-296.

Miller, T. (2003). Essay assessment with Latent Semantic Analysis. *Journal of Educational Computing Research, 29*(4), 495-512.

Shermis, M.D., & Burstein, J.C. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Tyner, K. (1999). *Development of mental representation: Theories and applications.* Lawrence Erlbaum Associates.

References and Citations should follow the APA Style Manuel 5th edition.