The Evaluation of Electronic Marking of Examinations

Pete Thomas
Open University
Milton Keynes
UK, MK7 6AA
+44 (0) 1908 652695

p.g.thomas@open.ac.uk

ABSTRACT

This paper discusses an approach to the electronic (automatic) marking of examination papers, in particular, the extent to which it is possible to mark a candidate's answers automatically and return, within a very short period of time, a result that would be comparable with a manually produced score. The investigation showed that there are good reasons for manual intervention in a predominantly automatic process. The paper discusses the results of tests of the automatic marking process that in two experiments yielded grades for examination scripts that are comparable with human markers (although the automatic grade tends to be the lower of the two). An analysis of the correlations between the human and automatic markers shows highly significant relationships between the human markers (between 0.91 and 0.95) and a significant relationship between the average human marker score and the electronic score (0.86).

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: *Distance Learning.*

K.3.2 [**Computer and Information Science Education**]: *Self assessment.*

General Terms

Experimentation

Keywords

Electronic examinations, Automatic marking.

1. INTRODUCTION

This paper discusses part of a project looking at all aspects of remote electronic examinations in which candidates construct their answers to conventional, time-limited examinations using a computer, submit their free-form answers via the Internet to a central facility where the answers are marked, and results are returned electronically (EAP, 2001). In the part of the project described here, we were interested in discovering the extent to which it would be possible to mark a candidate's script automatically and return, within a very short period of time, a result that would be comparable with a manually produced score.

Similar work has been carried out on the grading of free-form essays particularly in relation to style (Deerwester, Dumais et al., 1990; Burstein, Kukich et al., 2001; Christie, 1999; Whittington and Hunt, 1999; Leacock and Chodorow, 2000; Burstein, Leacock et al., 2001). Typically, in marking essays, the approach is to mark a number of scripts manually in order to 'prime' the automatic marker before using it to mark the remaining essays. Usually, the accuracy required when marking style is a grade taken from a narrow range. In examinations, the typical requirement is a finer-grained percentage mark.

In a typical time-limited examination the emphasis, in technical subjects, is on the assessment of content (often good style is not a feature of many answers!) and little work exists in this area. A discussion of a prototype for the automatic analysis of a short free-text answers which compares an examiner's 'expertly written model answer' with a student's answer can be found in (Callear, Jerrams-Smith et al., 2001). It parses sentences into concept dependency groups which is claimed to deal with grammatically incorrect and ambiguous answers. The work of ETS Technologies (Leacock and Chodorow, 2000; Burstein, Leacock et al., 2001) includes an investigation into the feasibility of automating the scoring of short answer content-based questions such as those that appear in a textbook's chapter review section. Their C-rater software identifies concepts and scores an answer as either correct or incorrect. It is based on analyzing the logical relations between the syntactic components for each sentence in the answer. C-rater uses the single correct answer that is found in an instructor's guide or answer key to grade student answers. The area of use for C-rater is 'relatively low stakes quizzes'. Thus, there are qualitative and quantitative differences between the areas covered by these efforts and our work on examinations.

In the first of two experiments reported on here, candidates attended an examination centre where they answered the examination questions in a strictly invigilated environment and wrote their answers manually on paper. In the second experiment, candidates used their own computers to construct their answers, which were then submitted via the Internet to a

ITiCSE'03, June 30–July 2, 2003, Thessaloniki, Greece.
Copyright 2003 ACM 1-58113-672-2/03/0006…$5.00.
.

central database. In both examinations candidates had up to three hours in which to answer their selected questions.

In the first experiment we chose a random sample of 20 hand-written scripts and converted one short-answer and one long-answer from each into to electronic format. We used this data to investigate algorithms for automatic marking and the process by which an accurate score could be obtained. We had access to the manually produced scores to validate the automatic marking system and the results are presented in this paper. This experiment revealed a number of issues related to manual marking that have to be taken into account when marking automatically. The second experiment candidates downloaded a copy of the examination paper onto their own computer and entered their answers into web-based forms. The electronic scripts were submitted to a central database where they were automatically marked. The scripts were also marked by three independent human markers. This paper compares the results from the automatic marker and the human markers.

2. MARKING EXAM SCRIPTS

There are four sets of objects in the examination process: questions that often consist of sub-questions that we refer to as *queries*, sample or *specimen solutions*, that are complete answers to the queries (not to be confused with indicative solutions that are simply guides for the graders who will use their own judgment in awarding individual grades), *mark schemes* that specify how marks are to be allocated, and student *answers*. The purpose of *marking* is to compare an answer with a specimen solution and associate a *mark* which represents a measure of the correctness of the answer.

Before any automatic grading algorithm can be contemplated, it is essential to be aware of the difficulties that normally arise in a conventional, manually graded, examination process. For example, incomplete, erroneous or inconsistent data may be specified in a question or a candidate may be asked to draw a conclusion that is either inherently incorrect or cannot be derived from the available data (either specified in the question or from the knowledge base of the associated course). A more insidious problem concerns ambiguity. That is, when a candidate's understanding of the meaning of the question differs from that of the question setter. Whatever kinds of error there may be in a question, the result is a discrepancy between the specimen solution and the 'true' solution. Of course, it is usual for examiners to go to great lengths to ensure that errors and ambiguity in questions are minimized prior to the examination being offered, but it can be the case that problems only come to light when candidates read the paper or when the examiners attempt to grade answers.

A specimen solution has several roles, but in the marking process it represents a guide for the marker. It may be an idealized solution that is not representative of the answers provided by candidates. That is, it is the examiner's answer and therefore relates to the examiner's interpretation of the question, which may be different from those of the candidates. In cases where a question is ambiguous, the differences between the specimen solution and candidates' answers could be quite dramatic.

There is a wide range of potential problems with candidates' answers. For example, a candidate may be unable to express ideas clearly either because of a lack of language skills or a lack of knowledge or their interpretation of a question can be markedly different from that of the question setter. A candidate's lack of knowledge can lead to answers that are difficult to comprehend, and excessively lengthy answers can mask this lack of knowledge. Spelling and typing errors occur frequently, making the recognition of phrases difficult. Candidates sometimes use abbreviations of their own devising, which often lead to incomprehensible phrases. In conventional manually marked examination processes, many measures are taken to minimize the problems inherent in the examination process. For example, using several examiners to agree questions, specimen solutions, and mark schemes. However, it is often left to the marker to identify alternative solutions and mark schemes. Automatic marking is no different: steps still have to be taken to minimize the impact that the problems identified above can have.

3. ELECTRONIC MARKING

A major aim of this work is to develop a marking system that would be capable of providing an accurate mark very quickly for possibly thousands of scripts (undergraduate computing courses at the Open University have several thousand students per presentation). In addition, we were interested in marking content but not style. Therefore, we decided to employ a relatively simple algorithm based on matching key-phrases between a student's answer to a question and the corresponding specimen solution. We employed a thesaurus to cope with language issues such as synonyms, plurals and verb tenses (there are other well-known techniques such as stemming that could have been used, but some form of thesaurus would still have been necessary (Salton, 1989; Jurafsky and Martin, 2000)).

For the reasons discussed earlier, a specimen solution can consist of possibly more than one solution to the question and each such solution could have a different allocated mark. In addition, a solution may be split into parts with each part having a mark. That is, the mark for a solution is the sum of the marks of its parts. Indeed, it is possible to envisage having alternatives for each part of a solution. Ultimately, therefore, a specimen solution may be composed of a set of alternatives, each alternative is composed of parts and each part can have alternatives. Hence, our view of a specimen solution is a set of phrases incorporated into a Boolean expression. For example, suppose that the solution to a (trivial) question consists of the following three phrases, all of which should be found in a correct answer:

bounds registers,

range memory locations,

address outside range exception

Further, suppose that the phrase 'storage protection keys' is a legitimate alternative to 'bounds registers'. This information can be represented by a tree as shown in Figure 1.

The notation AND(3, 6) means that there are three phrases, all of which are required and, if they all appear in an answer, would be awarded 6 marks. The notation OR(2, 1, 3) means that there are two choices, only one of which needs to appear in an answer and, if it does, is worth 3 marks. This choice of marks for the alternatives would force the remaining part of the conjunction to be worth 3 marks and, if no further information were given, would result in the two phrases 'memory locations range' and 'address outside range exception' to be allocated 1.5 each.

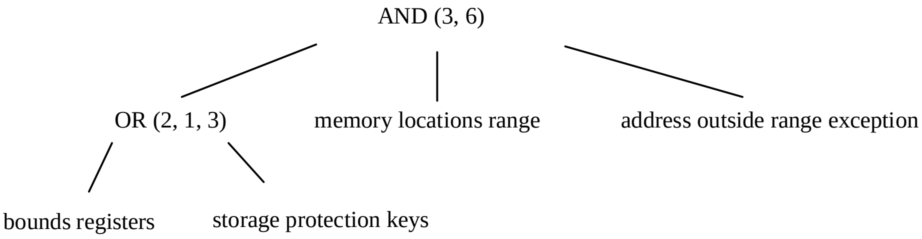

Figure 1 A solution tree with alternatives

4. RESULTS

A prototype automatic marker was first tested on a sample of 20 student scripts from a conventional hand written exam. The examination paper consisted of two parts so it was decided to transcribe one part 1 question (No. 7) and one part 2 question (No. 12). The specimen solutions were used to hand-build both the initial solution trees and an initial thesaurus. The results of this experiment are shown in Table 1. Question 12 consisted of 4 parts labelled 12(a) to 12(d).

Table 1 Results of the initial tests.

Question	7	12 (a)	12 (b)	12 (c)	12 (d)	12 Total	Overall
Marks allocated	4	8	2	8	2	20	24
Marker averages	1.7	6.1	1.1	4.8	1.13	13.0	14.6
Tool averages	1.75	5.5	0.98	4.8	0.8	12.4	14.2

The **marks allocated** row shows the maximum marks available for each (part of a) question. The **marker averages** are the average marks for the 20 scripts awarded by the three human markers. The **tool averages** are the results produced by the automatic marker. The average difference in the overall mark over the 20 scripts was 2.5 (10.5%) with a standard deviation of 1.54 marks. The difference in overall marks between the human graders and the tool for each candidate's script ranged between 0 and 6.

The second experiment was part of a broader experiment to assess the merits or otherwise of electronic examinations (Thomas, Price et al. 2001). In this experiment, students volunteered to take a mock examination: an examination with the same structure and kinds of question to be found on the final examination of the course. The mock exam was a synchronous electronic examination in which all candidates sat the exam at the same time, downloaded a copy of the examination paper over the Internet and were given three hours to complete it. The candidates typed their answers into a sequence of web-forms that were transferred to a central database at the end of the exam. In the event, 20 students volunteered to take part, but on the day only 11 students took the exam. The automatic marker was applied to the complete set of candidates' answers from the second experiment with the results shown in Table 2.

All scripts were marked by three independent markers, and Table 2 gives their average mark. An initial scan of Table 2 shows that the automatic marker underestimated every candidate's score when compared with the human markers' average score.

Table 2 Initial results from the second experiment.

Student	**Automatic score**	**Markers mean score**
1	40	56.7
2	41	60.0
3	51	60.8
4	43	60.8
5	48	76.0
6	49	48.5
7	32	43.5
8	58	62.0
9	53	61.5
10	37	47.0
11	34	42.0
Mean	**44.18**	**56.25**
St.Dev.	**8.28**	**10.09**

In some cases, the discrepancy is quite significant, and on average the difference was 12 marks. (Students had been warned of the experimental nature of the automatic marking and the likelihood that the quoted mark would be lower than their actual mark - based on our experiences with the first experiment.)

To understand the reasons for the discrepancies, a closer inspection of the solutions and answers was undertaken. The solution trees and thesaurus were examined in detail for deficiencies in dealing with one (arbitrarily chosen) candidate's answers. The problems are categorized in Table 3.

Table 3 Categories of deficiencies.

Description	**Number**
Number of significant spelling errors in answers	8
Number of deficient solution trees	15
Number of thesaurus deficiencies	12
Number of deficient specimen solutions	2
Number of lexical deficiencies	1
Number of language parsing errors	2
Number of deficient questions	2

In this table, the term 'deficiency' covers errors (e.g. spelling mistakes, incorrectly coded solution trees, and incorrect specimen solutions) and omissions (e.g. significant phrases omitted from the solution trees). For this candidate's choice of questions there were answers to 39 queries (and hence there were 39 solution trees).

A significant spelling error is any error in a word in a candidate's answer which, if corrected, would result in the grader increasing the mark for that answer. Eight significant spelling errors were found in this candidate's 39 answers.

Approximately one third of all solution trees (15 out of 39) were found to be deficient. The deficiencies included missing alternative phrases, inappropriate tree structures, incorrect mark allocations, and miscoding (the solution tree was not input correctly). The thesaurus deficiencies (12) consisted primarily of missing key words. Deficiencies were also found in queries and specimen solutions (2 each out of 39).

Lexical deficiencies and language parsing deficiencies indicate problems with the algorithms for recognizing phrases. For example, by not being able to distinguish the different uses of the oblique symbol (/), such as between i/o (a single phrase) and urgent/important (alternative phrases). In general, the prototype grader did not deal satisfactorily with abbreviations.

The next stage of the experiment was to amend the thesaurus and the solution trees in an attempt to rectify as many of the noted deficiencies relating to the one candidate's answers. The amendments included those required to deal with the two deficient specimen solutions. The scores after the first amendment show, in many cases, an increased score that is closer to the markers' average. The amendments gave an increase in the average examination script mark from 44.2 to 46.0.

The same process of solution and thesaurus improvement was repeated for three further scripts randomly chosen, each time providing an increase in the average script mark. The results are shown in Table 4.

The average of the human markers is shown in the column headed Marker Av. The difference between the average of the human markers and the final automatic mark in given in the final column headed Difference. As can be seen from Table 4, there was an improvement at each stage and the rate of improvement due to analyzing further scripts decreases as more scripts are analyzed.

After 4 rounds of improvement, the average difference between the human markers and the automatic marking over all scripts was 4.73. The difference between the averages of the human markers and the automatic marker was 2.65. These results indicate that although the automatic marker consistently underestimates the final mark and, at most, differs by 11 from the human markers' average, there is good agreement. However, it should be noted that the difference between individual human markers could be quite large (up to 8.5 in the case of student 5). The comparison between human markers and automatic marking is shown graphically in Figure 2.

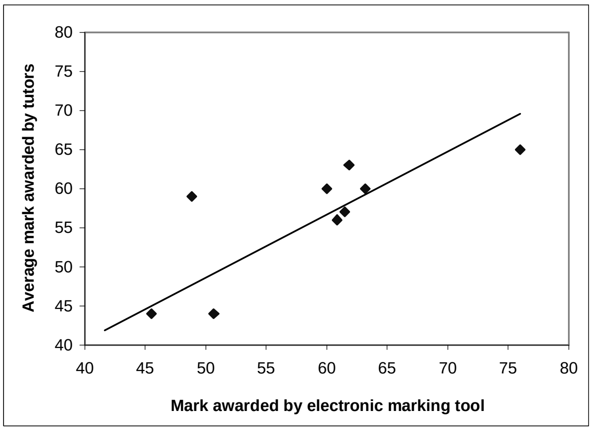

Figure 2 Comparison of human and automatic markers

It is useful to look at the correlations between the markers and between the markers and the electronic mark. Table 5 shows the Pearson pmcc values for the data in Table 4.

In all comparisons the correlations are significant. The correlation between the average of the three markers and the electronic score is also significant at 0.86 (the critical value for a two-tailed test with N=11 is 0.735 with 99% confidence).

Table 4 Comparative scores after four sets of improvements

Student	Markers			Marker Av.	Automatic Marker					Difference
	A	B	C		0	1	2	3	4	
1	50.5	54	47.5	50.67	40	43	43	44	44	6.67
2	60	59	61	60	41	41	48	60	60	0
3	64	62	56.5	60.83	51	52	53	53	56	4.83
4	60	61.5	63	61.5	43	43	56	57	57	3.75
5	81	72.5	74.5	76	48	48	64	65	65	11
6	50	47	49.5	48.83	49	53	57	59	59	10.17
7	44	43	46	44.3	32	39	39	39	39	5.3
8	60	64	61.5	61.83	58	56	60	63	63	1.17
9	61	62	66.5	63.17	53	54	55	58	59	4.17
10	47	47	42.5	45.5	37	42	44	44	44	1.5
11	44	40	41	41.67	34	35	37	39	39	2.67
Mean	56.5	55.6	57.44	55.85	44.2	46.0	50.5	52.8	53.2	4.73

1

Table 5 Pearson pmcc for correlations between markers.

	Part 1	Part 2	Total
Tutor 1 & Tutor 2	0.8260	0.9680	0.9456
Tutor 1 & Tutor 3	0.9531	0.9350	0.9281
Tutor 2 & Tutor 3	0.9279	0.9170	0.9362
All tutors & electronic	0.9026	0.8244	0.8604

5. FUTURE WORK

The marking algorithm seems to be effective. However, in the same way that variations in human markers are taken into account by multiple marking, we are investigating the idea that two (or more) sufficiently different marking algorithms could be incorporated into the system. If nothing else, discrepancies between the methods would highlight scripts that require additional scrutiny and might identify those scripts that would be most appropriate for the process of specimen solution improvement.

Clearly, larger scale experiments are required to prove the effectiveness of the algorithms and new experiments are under way. In addition, a more user-friendly way of representing specimen solutions and mark schemes has been developed, and software support for the process is being designed.

Ultimately, we wish to develop a fully integrated software system in which the needs of all users (candidates, examiners and administrators) are met and which effectively supports those areas requiring human intervention. The results of candidates' attempts at answering examination questions are an important resource in reviewing the effectiveness of both the teaching and the questions. If questions are to be reused (either directly or as templates for new questions) it is important to know how effective they have been. Therefore, an integrated system would enable the examiner to see the results of previous attempts.

6. CONCLUSIONS

We have identified the problems in trying to reach an accurate mark using manual marking which turns out to be a feature of examinations and not the method of marking. Therefore, electronic marking must be based on a process that takes these issues into account.

The experiments reported here have shown that acceptable scores that compare favourably with manual marking can be obtained. The average difference between the human markers and the automatic marking over all scripts was 4.73. The difference between the averages of the human markers and the automatic marker was 2.65. These results indicate that although the automatic marker consistently underestimates the final mark and, at most, differs by 11 from the human markers' average, there is good agreement. It should be noted that the difference between individual human markers could also be quite large: up to 8.5 in the case of one student.

We also catalogued the types of deficiencies found in the electronic marking process. Whilst it was not surprising that the largest numbers of errors occurred in the solution trees and thesaurus, it was clear that the iterative solution tree improvement process did rectify the specimen solutions and led to a more comprehensive thesaurus. However, the results also confirm the existence of deficiencies in questions and specimen solutions, although these were detected early on in the process. Candidate's spelling errors can have a significant effect on the marking process and we need to improve the system to deal with them effectively.

7. ACKNOWLEDGEMENTS

Thanks are due in no small measure to the EAP Team: Blaine Price, Carina Paine and Mike Richards, Open University, UK.

8. REFERENCES

[1] Burstein, J., K. Kukich, et al. (1998). Computer Analysis of Essays. NCME Symposium on Automated Scoring, USA.

[2] Burstein, J., C. Leacock, et al. (2001). Automated Evaluation of Essays and Short Answers. Fifth International Computer Assisted Assessment Conference, Loughborough University, UK, Learning & Teaching Development, Loughborough University, 41-45.

[3] Callear, D., J. Jerrams-Smith, et al. (2001). CAA of Short Non-MCQ Answers. Fifth International Computer Assisted Assessment Conference, Loughborough University, Learning & Teaching Development, Loughborough University, 55-69.

[4] Christie, J. (1999). Automated essay marking for both style and content. Third International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK, 39-48.

[5] Deerwester, S., S. Dumais, et al. (1990). Indexing by Latent Semantic Analysis. JASIS **41**(6): 391-407.

[6] EAP (2002). Electronic Assessment Project. http://mcs.open.ac.uk/eap

[7] Jurafsky, D. and J. H. Martin (2000). Speech and Language processing, Prentice Hall. ISBN 0-13-095069.

[8] Leacock, C. and M. Chodorow (2000). Automated Scoring of Short-Answer Responses, ETS Technologies. http://www.etstechnologies.com

[9] Salton, G. (1989). Automatic text processing: the transformation, analysis and retrieval of information by computer. Reading, Mass, USA, Addison-Wesley.

[10] Thomas, P. G., B. Price, et al. (2001). Experiments with Electronic Examinations over the Internet. Fifth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK, Loughborough University, 487-502.

[11] Thomas, P. G., B. Price, et al. (2002). Remote Electronic Examinations: student experiences. British Journal of Educational Technology **33**(5): 537-549.

[12] Whittington, D. and H. Hunt (1999). Approaches to the Computerised Assessment of Free Text Responses. 3rd International Conference on Computer Assisted Assessment, Loughborough University, Loughborough, UK, 207-219.