most likely to succeed. In addition, any preparation for the test should have a beneficial effect on the candidate, equipping them with skills that they will need as they progress through life.

References

Boyle, C. (1998). 'Organisations selecting people: how the process could be made fairer by the appropriate use of lotteries (with discussion)', *Journal of the Royal Statistical Society: Series D: The Statistician*, **47**, 2, 291–322.

Everson, H.T., and Millsap, R.E. (2004). 'Beyond individual differences: exploring school effects on SAT scores', *Educational Psychologist*, **39**, 3, 157–172. With correction **39**, 4, 261–261.

McDonald, A.S., Newton, P.E., Whetton, C. and Benefield, P. (2001). *Aptitude Testing for University Entrance: A literature review*. Slough: NFER.

McDonald, A.S., Newton, P.E. and Whetton, C. (2001). *A pilot of aptitude testing for University Entrance*. Slough: NFER.

Powers, D.E., and Rock, D.A. (1999). 'Effects of coaching on SAT I: Reasoning Test Scores'. *Journal of Educational Measurement*, **36**, 2, 93–118.

AUTOMATIC MARKING

Automatic marking of short, free text responses

Jana Z. Sukkarieh[1], **Stephen G. Pulman**[1] and **Nicholas Raikes**[2]

Introduction

Many of UCLES' academic examinations make extensive use of questions that require candidates to write one or two sentences. With increasing penetration of computers into schools and homes, a system that could partially or wholly automate valid marking of short, free text answers typed into a computer would be valuable, but would seem to pre-suppose a currently unattainable level of performance in automated natural language understanding. However, recent developments in the use of so-called 'shallow processing' techniques in computational linguistics have opened up the possibility of being able to automate the marking of free text without having to create systems that fully understand the answers. With this in mind, UCLES funded a three year study at Oxford University. Work began in summer 2002, and in this paper we introduce the project and the information extraction techniques used. A further paper in a forthcoming issue of *Research Matters* will contain the results of our evaluation of the automatic marks produced by the final system.

Uses for automatic marking

UCLES' traditional strength is in high stakes assessments that lead to qualifications. As more of our customers move to computer based assessments, an initial application of automatic free text marking in a high stakes context is as a quality control check on human marking, increasing the speed and efficiency of our quality control process. Every short, free text answer[3] could be marked both by computer and human markers, with any differences being resolved by a second human marker. Over time, as the capabilities and limitations of automatic marking became better understood, the proportion of answers marked by both computer and human could be reduced, with human marking targeted on the hardest to mark questions and on reviewing automatic marks that appear anomalous.

In the short term, however, the real opportunity for automatic free text marking is in low stakes tests. Many teachers and students use questions from our past papers, and we would like to be able to offer them an automatic marking service covering the free text questions as well as the 'objective' ones.

The challenge

Raikes and Harding (2003, p.270) state that an item's suitability for automatic marking depends on how near it can be placed to the objective end of what they call the objective-subjective continuum. The continuum is defined by the 'resolution' – the specificity and comprehensiveness – of an explicit marking guide that specifies how answers should be processed and marked. Traditionally, high resolution guides have been generated by greatly constraining the answers that students may give, as in multiple choice tests. More recently, attention has focussed on techniques for generating what are in effect high resolution marking guides for more open-ended item types, shifting them towards the objective end of the continuum where they may be automatically marked without affecting their validity.

In our automatic marking project we were concerned with marking short, factual answers varying in length from a few words up to around five lines, taken from GCSE biology examinations, where answers were marked for their correct content. The challenge was in coping with the myriad and sometimes unconventional ways in which credit-worthy answers were expressed, and the many mistakes in grammar and spelling found in some answers that nevertheless contained more or less the right content. Standard syntactic and semantic analysis methods would have been difficult to use, and even if we had fully accurate syntactic and semantic processing, many answers contained features that require a degree of inference that is beyond the state of the art. For example, in a question concerning asexual reproduction, a human marker inferred that

1. Computational Linguistics Group, Centre for Linguistics and Philology, Walton Street, Oxford OX1 2HG, United Kingdom. Email: first.lastname@clg.ox.ac.uk
2. Senior Research Officer, Evaluation & Validation Unit
3. Of course, completely objective items – multiple choice items and the like where every possible answer can be predicted – will be marked solely by computer.

a student who wrote *you do not have to wait until spring* meant to say *asexual reproduction can be done at any time*, a statement worth a mark according to the mark scheme. We have also found that students sometimes use a negation of a negation for a positive, as in *won't be done only at a specific time*, written for the same question. Contradictory or inconsistent information must also be detected, such as the inconsistent scientific information contained in the student statement *identical twins have the same chromosomes but different DNA*. These circumstances conspire to make the task too challenging for deep processing at present and so we decided to trade accuracy for robustness and investigate shallower 'information extraction' techniques, since they do not require complete and accurate parsing and are relatively robust in the face of ungrammatical and incomplete sentences.

Information extraction in a nutshell

Information extraction (IE) techniques pull out pertinent information from a partially syntactically analysed text by extracting those bits that match a set of domain-specific patterns typically built from training data. In our case, the training data are a sample of human marked answers – some human marking is necessary for setting up automatic marking – and the mark scheme, and a pattern is essentially all the paraphrases discovered for a particular entry in the mark scheme. The patterns include linguistic features as well as keywords.

Patterns

We wrote our initial patterns by hand, but have worked on a tool to take most of the tedious work out of this task. We base the patterns on recurring head words or phrases found in the training data, with syntactic annotation where necessary.

Consider the following six example training answers, which were written in response to the part question

Explain what has caused these two twins to be identical:

- the egg after fertilisation splits in two
- the fertilised egg has divided into two
- the egg was fertilised it split in two
- one fertilised egg splits into two
- one egg fertilised which split into two
- one sperm has fertilised an egg that split into two

These are all acceptable paraphrases of an answer given in the mark scheme as *They are formed from the same fertilised egg/same embryo*, and they and similar variants are captured by a pattern like:

singular_det + <fertilised egg> +{<split>; <divide>; <break>} + {in, into} + <two_halves>, where

<fertilised egg>	=	NP with the content of 'fertilised egg'
singular_det	=	{the, one, 1, a, an}
<split>	=	{split, splits, splitting, has split, etc.}
<divide>	=	{divides, which divide, has gone, being broken...}
<two_halves>	=	{two, 2, half, halves}

It is sometimes essential that the patterns incorporate the linguistic knowledge our syntactic analyser can generate at the moment, namely part-of-speech tags, noun phrases and verb groups. In the above example, the requirement that <fertilised egg> is a noun phrase (NP) will exclude something like *one sperm split in two and fertilised more than one egg* but accept something like *an egg which is fertilised*....

System architecture

The Student's View

Figure 1 is an annotated screenshot from a trial test that uses our automatic marking engine. Students enter their answers in the box at the bottom and may optionally click the 'check spelling and typing' button, which identifies any unrecognised words and suggests alternatives. We currently permit students to edit their answers as much as they wish without penalty. When they are happy with their answers to all the questions, students click the 'Finish test' button which submits their answers for marking. Students receive their item level marks together with an indication of what the marks were awarded for. The system can easily be reconfigured to, for example, provide marks and feedback after each question attempt, rather than at the end, depending on the context in which it is being used.

Figure 2 is a schematic diagram of how the marking system works. In this case the answer *When the caterpillars are feeding on the tomato plants, a chemical is released from the plants* is fed into the syntax analyser, which tags the different parts of speech (POS) and identifies chunks of text that represent noun phrases (NPs) and verb groups (VGs). The analyser makes use of a general lexicon, derived from the *Wall Street Journal* and the British National Corpus, and a specialised lexicon, derived in this case from a GCSE biology textbook and other specialised vocabulary encountered in the training data. The tagged and chunked answer then goes into the pattern-matching and marking system, where it is matched, if possible, with the patterns constructed from the training data. Marks and justifications are issued according to the scores and justifications pre-determined for the patterns matched and the rubric of the mark scheme.

Preliminary indications of marking accuracy

We will give the results of a larger and more detailed trial in a future article in *Research Matters*, but results from a preliminary trial involving nine part-questions are given in Table 1. The marking patterns were manually written using a training set of around 200 answers marked once by a human examiner, and the table gives the results when these patterns were tested on a further 60 answers that had not been seen by the developers until after the patterns were written. Note that the full mark for each question ranges between 1 and 4 – that is, the number of correct points required ranged from 1 to 4; there is a one-to-one correspondence between a correct point and a mark.

Column 3 records the percentage agreement between our system and the marks assigned by a human examiner. Sometimes humans make mistakes, however, and column 4 reflects the degree of agreement between the marks awarded by our system and those that would have been awarded by following the marking scheme consistently. Notice that agreement is correlated with the mark scale: the system appears less accurate on multi-point answers. We adopted an extremely strict measure of agreement, requiring an exact match.

Figure 1: Annotated screenshot of a trial automatically marked test

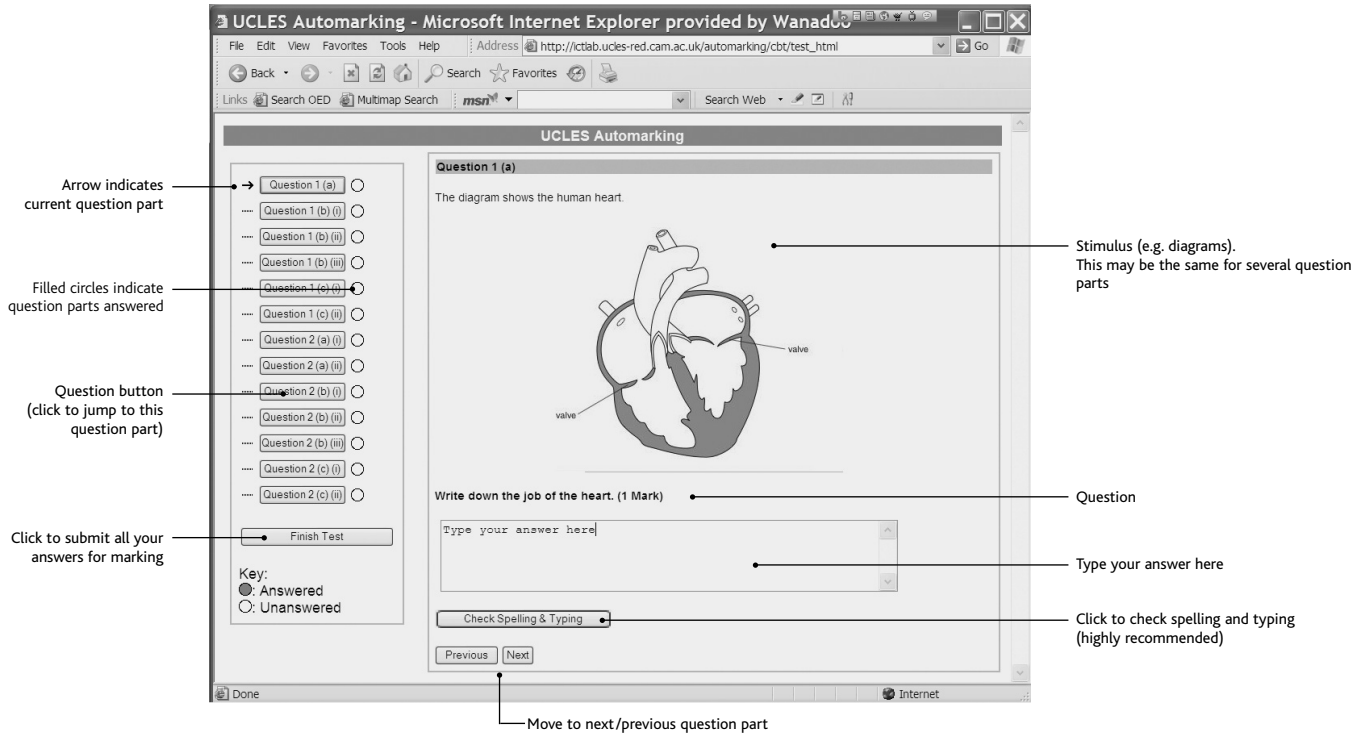

Figure 2: Schematic diagram of the marking system

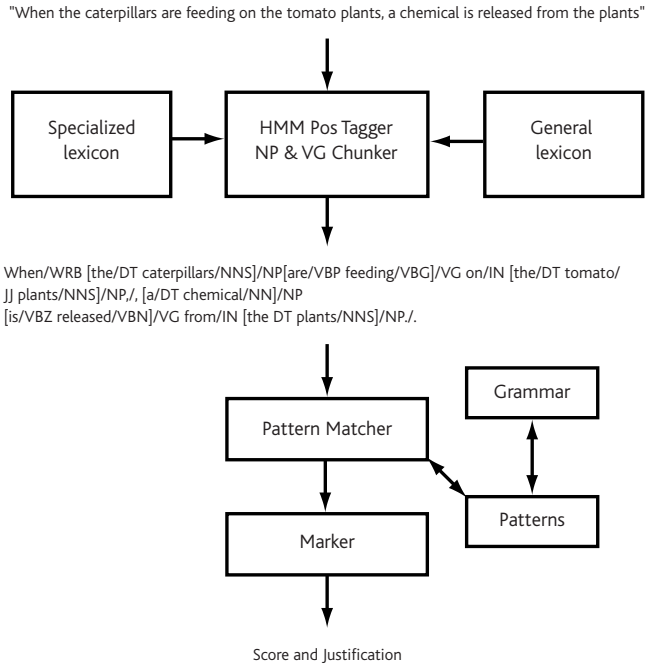

Table 1: Results for a preliminary evaluation of automatic marks compared with human marks, using manually written patterns and an Information Extraction approach

Question	*Full Mark*	*% Examiner Agreement*	*% Mark Scheme Agreement*
1	2	89.4	93.8
2	2	91.8	96.5
3	2	84.0	94.2
4	1	91.3	94.2
5	2	76.4	93.4
6	3	75.0	87.8
7	1	95.6	97.5
8	4	75.3	86.1
9	2	86.6	92.0
Average	—	85.0	92.8

Others' work

Several other groups are working in this area. The most prominent systems are C-Rater, developed by Leacock et al. (2003) at the Educational Testing Service (ETS), the IE-based system of Mitchell et al. (2003) at Intelligent Assessment Technologies, and that at Carnegie Mellon University described by Rosé et al. (2003). The four systems (these three and ours) are being developed independently, yet it seems they share similar characteristics. Commercial and resource pressures currently make it impossible to try these different systems on the same data, and so performance comparisons are meaningless.

Practical limitations and extensions

It takes around a day and half for a developer to discover and write the patterns manually for a new question, and we require around 200 definitively marked answers for pattern writing. We are currently evaluating whether a non-specialist programmer without experience in computational linguistics can do this task as effectively. We have also done some quite promising work on semi-automating the pattern writing process to make it quicker and less labour intensive – see Sukkarieh et al (2004) and Sukkarieh and Pulman (2005) for more information and results. Alternative, non-IE machine learning approaches have also been trialled with varying degrees of success –Sukkarieh and Pulman (2005) give details and results.

Conclusion

We have introduced our automatic marking project and described the information extraction techniques used and how we have applied them. Initial results are encouraging, with automatic marks correct 93% of the time on average. We will present the results of a more wide-ranging evaluation in a future edition of *Research Matters*.

References and further reading

Leacock, C. and Chodorow, M. (2003). 'C-rater: Automated Scoring of Short-Answer Questions'. *Computers and Humanities,* **37**, 4.

Mitchell, T., Russell, T., Broomhead, P. and Aldridge, N. (2003). 'Computerized marking of short-answer free-text responses': paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Raikes, N. & Harding, R. (2003). 'The horseless carriage stage: replacing conventional measures', *Assessment in Education*, **10**, 3, 267–77.

Rosé, C. P., Roque, A., Bhembe, D. and VanLehn, K. (2003). 'A hybrid text classification approach for analysis of student essays'. In *Building Educational Applications Using Natural Language Processing*, 68–75.

Sukkarieh, J. Z. and Pulman, S. G. (2005). 'Automatic Short Free-Text Answer Marking', *Natural Language Engineering*, (under review).

Sukkarieh, J. Z., Pulman, S. G. and Raikes N. (2003). 'Auto-marking: using computational linguistics to score short, free text responses', paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Sukkarieh, J. Z., Pulman, S. G. and Raikes N. (2004). 'Auto-marking 2: An update on the UCLES-Oxford University research into using computational linguistics to score short, free text responses': paper presented at the 30th annual conference of the International Association for Educational Assessment (IAEA), Philadelphia, USA.