AUTOMATED FREE TEXT MARKING WITH PAPERLESS SCHOOL

Oliver Mason and Ian Grove-Stephenson

Automated free text marking with Paperless School

Oliver Mason
Department Of English
Birmingham University

Ian Grove-Stephensen
Paperless School
The Chalkface Project

O.Mason@bham.ac.uk

Abstract

The Paperless School automarking system utilises a number of novel approaches to address the challenge of providing both summative and formative assessments with little or no human intervention.

The Paperless School system is designed primarily for day-to-day, low stakes testing of essay and short-text student inputs. It intentionally sacrifices some degree of accuracy to achieve ease of set up, but nevertheless provides an accurate view of the abilities of each student by averaging marks over a number of essays.

The system is designed to function as a back-end service to an Learning Management System (LMS), thus facilitating the marking of large numbers of texts. This should enable considerable teacher resources to be freed up for other teaching tasks.

In this paper we will discuss some of the issues involved in bringing computational linguistics to bear in the educational context. We will cover

- how Blooms Taxonomy (the pedagogical model underlying most formal grading schemes) can be represented in software.
- an overview of the steps required to derive a grade that will sufficiently closely predict the grade a human marker would give.
- extending the system to include formative assessment, via intelligent comment banks.

Introduction

The Paperless School free-text marking engine is designed as an integrated component of our schools' Learning Management System (LMS). In principle, it can also be used as a standalone application. It is still under development; this is our report on progress so far.

Why Free Text Marking is Important

The most highly valued activity for a teacher is to teach - yet in Britain, school teachers spend only 40% of their time in the classroom. Another 30% of their time is spent on marking. This activity does have value, but less so. British schools are traditionally wary of objective marking, so if we are to free up that 30% (worth 3 Billion UK Pounds / year to the taxpayer by the way) then we must find an effective way, that teachers will trust, to mark essays and short-text responses.

How Does a Human Being Mark an Essay?

Most school teachers express confidence that they could mark an essay on a subject they did not themselves know, provided

they can read an appropriate text on the subject first.
they are given a formal mark scheme specifying how many points they should give to which features.

Typical Things a Teacher Looks for

Coverage of relevant concepts
Evidence of understanding; typically through the relationships between concepts
Evidence of evaluation; sometimes the teacher is looking for a specific view, sometimes not.
Evidence of misconceptions: teachers usually know what these would be e.g. "seeing goes from the eye to the object" rather than "seeing involves light travelling from the object to the eye".

Natural Language Processing (NLP)

Natural Language Processing (NLP) is concerned with processing 'natural' language (as used by humans, in contrast to artificial languages such as computer programming languages) using the computer. Due to the sophisticated nature of human language(s), this has been an active research topic since the 1950s, and though a lot of progress has been made since, there are still far more unsolved problems than there are solutions.

The first major area where research in NLP was applied to was machine translation (MT). This was at the beginning of the Cold War, when it became important for the US military to be able to read Russian newspapers and other documents without being able to master the language. After being disbanded as unworkable, MT became important again with the introduction of the EU, where a large number of documents has to be translated into a number of different languages in the minimum amount of time. Abandoning fully

automatic translation, it is now computer-aided translation which is used sucessfully.

A second area where NLP plays a pivotal role is information retrieval (IR), where the limitations of simple key word searches become clear as soon as the size of a full-text database reaches a non-trivial dimension. Internet search engines partly suffer from problems connected with the multitude of meanings that apparently simple words can carry in different contexts. The use of NLP can increase the efficiency of such systems by providing a more in-depth analysis of the context in which a key word is used, giving useful clues as to whether it is used in a meaning relevant to the original query.

Initially most NLP systems followed a rule-based paradigm, based on the predominant linguistic formalism of Phrase Structure Grammar. After it became increasingly clear from the analysis of actually occurring language (as opposed to invented examples, on which much of mainstream linguistics is still based) that language is far too complex to handle efficiently with rule-based mechanisms, most up-to-date NLP systems nowadays operate probabilistically, or perhaps following a hybrid approach with both rule-based components and stochastic elements. This means that there will inevitably be mistakes in the processing, but the coverage is far broader than it would be possible with purely rule-based systems.

In the context of automatic assessment NLP is the key for advancing from simple multiple-choice-style questions (which can easily be processed by computer) to so-called 'free text' questions, where the student writes the answer without any computer-imposed constraints on the format used for formulating the answer. Instead the student simply writes the answer to a question in essay form, to be analysed by the computer through an NLP module.

This module faces a number of challenges:

- analysis the grammatical structure of the text
- extracting the 'meaning' of the text
- deciding whether the extracted information is relevant to the question, and to what degree

The key to solving the first problem is to realise that the level of analysis can be fairly shallow. The fewer details are identified, the fewer mistakes can be made by the computer. This also means that the system is quite robust, which means it can cope with essays which have some amount of grammatical mistakes, as long as they are not relevant to the analysis (see below for treatment of mistakes in the student essays).
The next problem is the meaning of 'meaning': this is a very elusive issue, and philosophers and linguists have tried to get a grip on it since the times of Aristotle. In order to reduce the problem down to a manageable level, Bloom's taxonomy was taken as the starting point. As this is used to define a student's performance it seemed like the obvious choice as a model for computer assessment.

Bloom's Taxonomy

Bloom (1956) identified a six-element taxonomy of educational objectives, or competencies, which is widely used in the assessment of student work.

Knowledge: remembering of previously learned material; recall (facts or whole theories); bringing to mind.
Comprehension: grasping the meaning of material; interpreting (explaining or summarizing); predicting outcome and effects (estimating future trends).
Application: ability to use learned material in a new situation; apply rules, laws, methods, theories.
Analysis: breaking down into parts; understanding organization, clarifying, concluding.
Synthesis: ability to put parts together to form a new whole; unique communication; set of abstract relations.
Evaluation: ability to judge value for purpose; base on criteria; support judgment with reason. (No guessing).

The English National Curriculum (NC), followed by children from ages 5-14, defines 8 Level Descriptors for each subject, using a subset of Bloom's Taxonomy:

Knowledge
Understanding
Evaluation

Note that these were selected by the creators of the NC specifically because they correlate with the way teachers already evaluate students' work.

Lingustic Predicates of Bloom's Categories

Our approach has been to apply NLP to test a piece of student text for the presence of each of these competencies.
The three key stages (K, U, and E) had to be matched on linguistic categories that can be recognised by computer. This section describes the different approaches used for that.

a) Knowledge

The subject area covered by a text is generally reflected in the terms used, and the objects and entities described. In a text about animals in Africa one would expect to find mentions of lions, zebras, antelopes, elephants, and so on. But the phrase 'word processing skills' would be rather odd in this context. If the student answering a question on this subject mentions a lot of unusual objects, entities, or concepts, one could conclude that the relevance of the answer was not very high, whereas a large number of relevant items clearly indicates that the student possesses sufficient knowledge of the area, and thus would score well on the K component.

Assessing the K-score then boils down to identifying concepts in the student's essay, and evaluating their relevance to the subject area. This can be achieved to a high degree of accurracy using current NLP techniques. All that is required for the knowledge-based part is a list of relevant concepts against which the concepts from the student's essay are compared.

In our system, this list is created by abstraction from an authoritative (and extensive) master text. This would normally consist of textbooks used for teaching the subject, using existing resources. It is not necessary to manually build a knowledge base.

b) Understanding

In Bloom's Taxonomy, Understanding is actually four separate categories:

Comprehension, Application, Analysis and Synthesis. Each of these then breaks down into a number of skills such as separating or identifying components and rearranging elements. The exact way in which we have modelled these processes in software is commercially sensitive and currently not something we can yet discuss. However, we hope it will form the substance of a future paper.

c) Evaluation

A rough measure of the evaluative content of an essay is relatively easy; one simply counts adjectives and adverbs. One can in principle refine this in three ways;

analyse syntactic patterns expressing evaluation. This would typically involve an embedded sentence, where the main clause has a verb related to mental processes, such as "I think that X", where 'X' is the embedded sentence that is evaluated. Other cases are instances of predicative adjectives with embedding, as in "It is obvious that X".
consider the context of each evaluation, by measuring the closeness of its link with a relevant concept. We have achieved this through a similar approach to our measurement of understanding.
consider the correctness of each evaluation. As well as the difficult technical problems this presents, there are pedagogical ones as well. A student's own opinion is often valued more highly by an assessor than is the opinion expressed in even an authoritative text. We believe that where correctness of evaluation needs to be measured, this is better done through objective testing.

The Core Process of the Paperless School Automarker

The automarker is implemented as a component of a web-based managed learning environment. A 'marking engine' runs on the LMS server; due to its processing requirements, it does not mark essays in realtime.

On submission, the student essay is sent to the server, together with information about the task, in order to identify the correct master texts for comparison. Each task is defined via a number of master texts which are relevant to the question to be answered, and there can even be 'negative' master texts which effectively contain a set of false statements, which have been established as typical student mistakes.

The student essay is then compared against each relevant master text to derive a number of parameters which reflect knowledge and understanding as exhibited by the student. The evaluation parameter is calculated through a linguistic analysis as described above.

When multiple master texts are involved in the comparison, each result from an individual comparison gets a weighting, which could be negative in the case of a master text containing common misconceptions. The weights are derived during the initial training phase.

The individual parameters computed during the analysis phase are then combined in a numerical expression to yield the assignment's grade - typically a National Curriculum grade or a GCSE level. Apart from that they are also used to select specific comments from a comment bank relevant to the task. With a fine grained set-up it is possible to give formative feedback to the student regarding his or her performance in different areas within the subject.

The output from the marking process is then passed back to the client for presentation to the teacher. This includes details on sections of the essay which are particularly good (or bad) in relation to the K, U, and E factors.

Setting up the Automarker for Summative Assessment

The process of setting up the automarker for a particular task is very straightforward:

Select master texts
Have a sample hand-marked (can be as few as 30)
Run the same sample through the marker and perform regression analysis.
Upload the resulting data to the server

Master texts can be drawn from a number of sources such as textbooks, encyclopedias or relevant websites. The system is highly tolerant of duplication of content between master texts, but can lose accuracy if the master texts use extremely complex grammar.

A small sample of student's essays is then hand-marked. This needs to be done once only per task, in order to derive the right weightings for the parameter values computed by the marking system. Once the weightings have been calculated, they can be re-used whenever that particular task is set again.

Computing the parameters is currently achieved through a regression analysis, which tries to get a best fit between the grades given by the marker and those resulting from the combination of the parameters. In principle there are other methods which might be used, for example genetic algorithms or techniques from machine learning.

Finally the task-related data (master texts and parameter weights) needs to be uploaded on the MLE server, and can then be accessed via an identifier attached to a student's essay.

Setting up the Automarker for Formative Assessment

This is more guesswork than science, and as a result the comments are best couched as suggestions. Surprisingly, teachers often see this as a benefit. Knowing that the system is fallible makes students evaluate the comments and creates the conditions for fruitful debate.

Integrating the Marking Engine into a Learning Management System (LMS)

Integration presents problems of scale and management. The processing required to mark even a short essay can be several seconds; in some instances we have recorded times as long as 30 seconds, even on a standalone system. We decided early on that it would be impractical to offer realtime marking, and instead implemented a queuing system.

We have chosen a web-based interface to allow users (typically commercial publishers) to set up and test automarking on their own content. This produces an XML-based Task Definition File (TDF) for each marking task. The TDF includes the IDs of the master texts, the parameter weightings, the comment bank and threshold values for each comment.

Problems Encountered

Human markers tend not to agree with each other, so we have no gold standard to calibrate against. The solution is better-written mark schemes, and we are working with one exam board on this issue at the moment.

There is no clear set of rules for selecting master texts. This puts a new skill requirement into the system.

Graduate and postgraduate-level material has so far proved difficult to automark with any accuracy. We believe that the problem is in the range of content that higher level students draw on.

Misspellings and bad grammar can throw the system out, but autocorrection would cause inaccuracies of its own. Thus far, we have put the onus onto the student to check his or her own spelling and grammar.

Conclusion

The Paperless School automarker is still a work in progress, and we are promoting it more for its ability to provide formative assessment (comments) than summative assessment (grades). We expect by this time next year to be able to report with confidence that the system is grading to an acceptable degree of accuracy for low-stakes coursework in a wide variety of contexts.

References

Bloom, B.S. (Ed.) (1956) *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain.* New York; Toronto: Longmans, Green.