MARKING SHORT FREE TEXT RESPONSES IN E-ASSESSMENT

Tendai Dube and Minhua Ma
University of Derby, UK
{T.Dube, M.Ma}@derby.ac.uk

ABSTRACT

While many e-assessment systems support a variety of question types that are used in traditional methods of assessment, like multiple choices, true or false, hotspot, drag and drop, and gap filling questions, it has however been a challenge for developers to support automatic marking of essay type questions due to the diversity of responses that can be submitted. Marking short free text response questions has been considered instead as expected answers are specific making it reasonably easier to distinguish between those which are correct and incorrect. In this paper we present a critical review of some of the techniques and algorithms that have been adopted in systems which have attained some success in supporting these free text response question types. We give descriptions of these methods and then we discuss them considering usefulness as the review is aimed at informing the selection of one such method, for adoption in an e-assessment system which is currently in development.

Keywords

E-assessment, free text response, automatic marking

1. INTRODUCTION

Technology has found a place in education, not only in the delivery of course materials but also in assessment. Virtual Learning Environments (VLEs), like Blackboard and Moodle, which have been chosen by different educational institutions for the delivery of learning resources, also come embedded with assessment tools. The question types supported by the VLEs are however limited to multiple choices, true or false, hotspot and gap filling questions. E-assessment systems like Questionmark Perception have been developed with improved functionality to support these closed question types which include drag and drop. Developers of such systems have also attempted to support free text response questions but because of the open nature of these question types, they are difficult to mark automatically. Free text responses are not predictable enough to be marked by the computer as they can be very diverse. For that reason, automatic marking of free-text response questions remains an area open for research. It is worth pursuing investigations in free text marking due to the pedagogical value of the free text question type in promoting deep learning rather than surface learning. Objective question types have been criticized for only being able to test lower order cognitive skills [12]. Other benefits of automatic free-text marking include saving time and costs and also improved efficiency [7] and consistency in the allocation of marks as human markers tend to be subjective and to make errors resulting in inconsistencies and unfairness. Human marking is also a very slow process which results in delayed feedback to the students.

Those who have made some success [12] in this area of automatic free-text marking shed some light on how this challenge can be addressed as described in the following sections. Free-text marking engines have been developed and though they may use different techniques, a general pattern is followed in their use. Firstly, marking guidelines are generated by the assessment owner. Templates of model answers are then produced using sample responses. Students' answers are then analyzed and marked in relation to the sample answers. Manual moderation is then conducted to check the marking.

2. AUTOMATIC MARKING METHODS

There are considerable techniques and algorithms that have been devised to mark free text response questions. These involve computational linguistics, Natural Language Processing (NLP), statistics, information extraction (IE) and pattern matching. Some of these techniques and algorithms are combined in existing marking engines to achieve better accuracy. Descriptions of how some of these techniques and algorithms are used are given below.

2.1 Natural Language Processing in e-marking

Natural Language Processing (NLP) is defined as theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications [4]. With its origins from the field of Linguistics, NLP focuses on language processing and language generation. This respectively means that it is concerned with the analysis of language to produce meaningful representations and also with the production of language from a representation. Besides Linguistics, other areas which have influenced NLP techniques are artificial intelligence, machine learning, computational statistics and cognitive science.

NLP can be approached in different ways which include the symbolic approach, statistical, connectionist and hybrid. The symbolic approach is based on explicit representation of facts about language through representation of schemes and associated algorithms. Deep analyses of linguistic phenomena are performed using this approach and rule-based systems are good examples. Mathematical techniques are used in the statistical approach which also uses large text corpora (bodies of text containing large sentences) to develop approximate generalized models of linguistic phenomena. The statistical approach has been used in parsing and also in speech recognition. The connectionist approach combines statistical learning with various theories of representation allowing transformation, inference, and manipulation of logic formulae [4].

Research has identified applications which use NLP and they include Information Retrieval (IR), Information Extraction (IE), question-answering, summarization, and machine translation. Following a user's query, Information Retrieval provides a list of potentially relevant documents. Information Extraction extracts key elements of information into structured representations. In a question-answering application, in response to a query, the user is given the particular answer text or the passage where the answer lies.

The evolutionary development of Natural Language Processing technique has won some popularity among developers as evidenced by its adoption in various systems which boast some success. Among these systems is the Automark system which was developed by Intelligent Assessment Technologies (IAT). Using NLP to mark free-text and open-ended questions, Automark looks for specific content within free-text answers based on the marking templates. Each template specifies one particular form of acceptable or unacceptable answer. The OpenMark e-assessment system of the Open University in the United Kingdom uses IAT's free-text assessment software which uses NLP's IE techniques. Also, both the Electronic Essay Rater (E-Rater) and the Conceptual Rater (C-Rater) systems employ NLP. E-Rater uses NLP techniques combined with statistical techniques to extract linguistic features from essays to be graded [4].

The popularity of NLP may be linked to its goal to accomplish human-like language processing. Like a human marker, e-assessment engines which use NLP techniques analyse sentences, check spellings and process the semantics of the students' text responses.

2.2 Information Extraction

Information Extraction (IE) has already been introduced in the section above on NLP but due to its significance in text manipulation, it is worth describing it further. IE is described as any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts [1]. [10] identifies the objective of IE as that of extracting from text certain pieces of information that are related to a prescribed set of concepts which they call "extraction scenario". A popular example is that of the domain of *Management Succession* which was used in the Sixth Message Understanding Conference in 1995 [8, 10] as shown in Figure 1. The IE technique is applied on the input text to get the new officer (PersonIn), the officer leaving (PersonOut) and the position in question (Post).

Input text:

C. Vincent Protho, chairman and chief executive officer of this maker of semiconductors, was named to the additional post of president, succeeding John W. Smith, who resigned to pursue other interests.

Succession event

PersonIn:	C. Vincent Protho
PersonOut:	John W. Smith
Post:	president

Figure 1. Management Succession domain: An input sentence and the output case frame [10]

In the automatic marking of free-text responses, the crucial prerequisite is therefore a set of text extraction rules that help identify the relevant information to be extracted for acceptable responses. [8] claims that the extraction rules for free text are typically based on patterns involving syntactic relations between words or semantic classes of words. So besides extraction rules, IE technology also involves syntactic analysis and semantic tagging plus recognition of domain objects like names, and also inferences.

2.3 Pattern Matching

Pattern matching is the act of checking for the presence of the constituents of a given pattern. The pattern which is specifically defined focuses on sequences or tree structures. Text is compared to the defined pattern looking for the structures. Matching parts are retrieved and, in the case of a marking engine, are then marked as appropriate. The Automark system mentioned in section 2.1 above is made up of modules which focus on different functionalities. One of the modules is a pattern-matching module which searches for matches between the mark scheme templates and the syntactic constituents of the student text. The sentence analysis module above the pattern matching module is responsible for identifying the main syntactic constituents of the student text, and their relations. The result of the pattern match is then processed by a feedback module [5] to get an appropriate mark for the student's response. Figure 2 shows the place of the pattern matching technique in the architecture of the Automark system. A similar structure is given in [7].

The mark scheme is the key element in pattern matching as it provides the input against which text to be marked should be compared and ideally all possible answers should be included. As an example, for a Biology question like: **What is the function of the red blood cells?** the mark scheme will contain such answers as: Red blood cells **transport oxygen** from the lungs to the body cells, or Red blood cells **carry oxygen** from the lungs to the body cells or Red blood cells **take oxygen** from the lungs to the body cells. Unfortunately, answers which may be acceptable to a human marker will be marked wrong if they are not represented in the templates. Assessment owners need to create exhaustive mark schemes to avoid disadvantaging some students and they also need to ensure manual moderation is done accordingly.

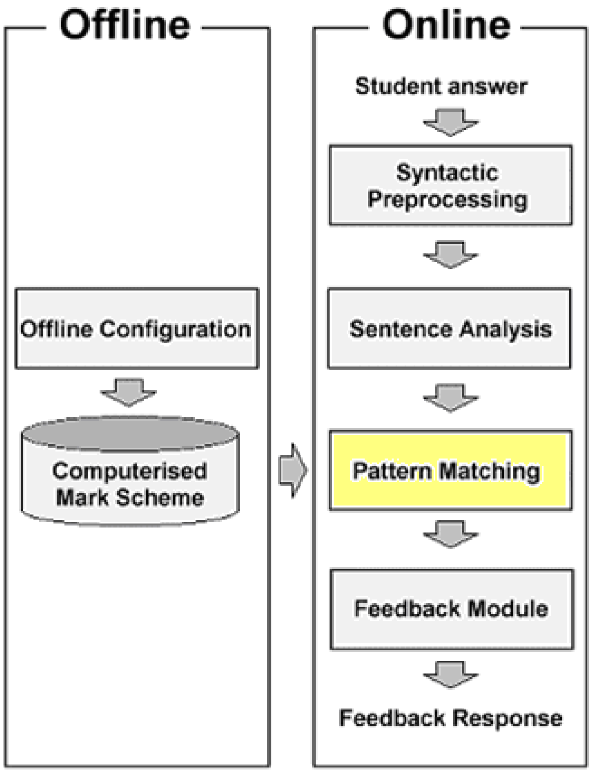

Figure 2. Pattern Matching in Automark System Architecture

2.4 Latent Semantic Analysis

Latent Semantic Analysis (LSA), also known as Latent Semantic Indexing (LSI), is a technique for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text [3] (corpus refers to a body of text containing large sentences). In other words, LSA is a way of analyzing relationships between a set of documents. The documents and their words are represented in a large two-dimensional matrix semantic space which is then decomposed using a matrix algebra technique called Singular Value Decomposition (SVD) [11] to derive a particular latent semantic structure model. A matrix represents the words and their contexts and each word being analyzed represents a row in the matrix whereas each column represents the context in which the word occurs, like sentences or paragraphs (in the case of marking students' texts) or documents. A technique called dimension reduction is then applied to the matrices to reveal relationships between words and context.

Figure 3 is an example of an initial matrix derived from the sentences: A) "The lady prefers travelling abroad on a vacation." B) "Japan was her most recent vacation choice." Only significant words are extracted from the texts leaving stopwords like *the, on, etc*.

$$
\begin{array}{ll}
 & \begin{array}{cc} \text{Context A} & \text{Context B} \end{array} \\
\begin{array}{l} \text{lady} \\ \text{vacation} \\ \text{Japan} \end{array} & \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}
\end{array}
$$

Figure 3. Initial Matrix

This first matrix **I** is then decomposed through SVD which involves some complex mathematical computations to produce three other matrices **M1, M2,** and **M3**. Thus **I = M1, M2, M3**. Dimension reduction is then applied to these special matrices to show a breakdown of the original relationships of the words into linearly independent components or vectors [2]. The occurrence of each word is weighted as an estimate of its importance in the passage.

The first use of LAS was in indexing documents and in information retrieval (IR). LAS been used for scoring students' answers in a system known as The Intelligent Essay Assessor (IEA) which was developed by Knowledge Analysis Technologies (KAT) [9]. Students' answers are compared to a model answer by calculating the distance between their vector projections.

LSA, also described as a bag-of-words, is limited in that it does not take into account syntactic relations of words. Sentences like "The cat ate the rat" and "The rat ate the cat" cannot be distinguished using this technique making it less accurate. LAS also requires large amounts of data to enable construction of a suitable matrix representation of the use or occurrence of words [12].

3. DISCUSSION OF THE TECHNIQUES

The techniques and algorithms described above have their strengths and weaknesses making some of them better than the others from different perspectives. The fact that most of these techniques are combined in systems where they have been applied is evidence of that imbalance. The Latent Semantic Analysis technique has been criticised for its failure to consider syntactic relations of words in a learner's response. It therefore provides a macro level of marking. However, in that right, LSA is a suitable technique for marking work for international students whose first language may not be English. In such students' responses, only the content is considered and not their language skills making it a fair process. On the other hand, Natural Language Processing's rule-based methods consider syntax in the responses. NLP therefore renders itself to inclusion in automatic free-text marking engines for the teaching of languages like English. Learners of English for Speakers of Other Languages (ESOL) can benefit from NLP, for example if it is included to mark formative assessments which enable students to practice on their own. They will be able to master grammatical concepts of the language from the automatic feedback. NLP features in many automatic marking engines and researchers' prototypes due to its consideration of syntax in the marking process.

While the formalisms used in LSA may aid accuracy in the marking process, the complexities of the mathematical computations involved make it deterrent for adoption in e-assessment systems or marking engines. Strong mathematical background is required in order to comprehend how the technique actually works and more so how to adopt it.

Information Retrieval, Information Extraction and Pattern Matching techniques are interlaced with LSA and NLP. It is difficult to separate these techniques as they overlap. What is common in all of their use is that there should be input text to be manipulated by retrieving pieces of information from it and matching that information to predefined rules or patterns and a decision is made on what to do with the result, i.e. whether the input text gets marks or not. It should be noted that pattern matching is key in the use of these techniques in automatic marking of free text. In all cases, there should be a reference to a document containing the expected answers for comparison to be made with what is contained in a student's answer.

4. CONCLUSION

While free-text automatic marking using the computer has always been a challenge, continued research in this area has seen some success stories on which improvements can be made. Progress was made as a result of research in various fields which include Linguistics, artificial intelligence, machine learning, computational statistics and cognitive science. The outcomes of that research are the various techniques and algorithms that were created in order to manipulate text thereby providing an essential ingredient to the development of automatic marking engines for free-text responses. The techniques involve LSA, IR and IE, NLP, pattern-matching, computational linguistics and statistics. Many systems use a combination of these techniques for better accuracy thereby complementing their strengths and weaknesses.

Further research and development in the area of free-text automatic marking is inevitable because of the growth of computer-based technologies for teaching and learning especially in further and higher education, where assessment questions with free text responses are more common than questions with fixed-answers. E-Assessment systems are already a common entity of the academic provision package. Including free-text response question types in e-assessment systems will not only help presentation of different question types but ensure that higher cognitive skills are also covered, thus making such systems pedagogically valuable. In education, technology uptake should not be about the freshly released "cool" system, but it should be informed by how that system fits into the pedagogic strategies of a course or an institution at large.

Free-text automatic marking engines will benefit the learners by enabling timely feedback as marking takes a short time unlike human marking. It will benefit the teachers by reducing marking load freeing up some time for preparation of teaching materials and for research, and help towards achievement of the government's goal of provision of Higher Education for degrees attainment to many people. Educational institutions will benefit from saving costs from teachers' marking, e.g. no scripts will be printed for marking, and also possibly from student satisfaction as a result of timely feedback of their work.

5. REFERENCES

[1] Cowie, J. and Wilks, Y., *Information Extraction,* [Online] http://www.dcs.shef.ac.uk/~yorick/papers/infoext.pdf, (Retrieved on 23/02/2010).

[2] Dewester, S., Dumais, S. and Harshman, R., *Indexing by Latent Semantic Analysis,* [Online] lsi.research.telcordia.com/lsi/papers/JASIS90.pdf, (Retrieved on 11/02/10).

[3] Landauer, T. K., Foltz, P. W., & Laham, D., *Introduction to Latent Semantic Analysis.* Discourse Processes, **25**, 259-284 (1998).

[4] Liddy, E.D. *Natural Language Processing.* Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc. [Online] www.cnlp.org/publications/03NLP.LIS.Encyclopedia.pdf.4, (Retrieved on 13/02/10).

[5] Mitchell, T., Russell, T., Broomhead, P., and Aldridge, N., *Towards Robust Computerised Marking of Free-Text Responses,* In Proceedings of the 6th CAA Conference, Loughborough: Loughborough University (2002).

[6] Mohler, M. and Mihaclea, R., *Text-to-text semantic similarity for automatic short answer grading*, Proceedings of the 12th Conference of the European Chapter of the Computational Linguistics, Association for Computational Linguistics, (2009).

[7] Siddiqi, R. and Harrison, C., *On the Automated Assessment of Short Free-Text Responses*, [Online] www.iaea2008.cambridgeassessment.org.uk/.../180392_Siddiqi___Harrison.pdf (Retrieved on 11/02/2010).

[8] Soderland, S., *Learning Information Extraction Rules for Semi-Structured an Free Text,* Machine Learning Vol. 34 233-272 (1999).

[9] Sukkarieh, J. Z., Pulman, S. G. and Raikes, N., *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. (2003).

[10] Turmo, J., Ageno, A. and Catala, N., *Adaptive Information Extraction,* ACM Computing Surveys, Vol. 38, No. 2, Article 4, (July 2006).

[11] Valenti, S., Neri, F., and Cucchiarelli, A., *An Overview of Current Research on Automatic Essay Grading.* Journal of Information Technology Education. Volume 2 (2003).

[12] Whittington, D. and Hunt, H., *Approaches to the Computerized Assessment of Free Text Responses*, In Proceedings of the Sixth International Computer Assisted Assessment Conference (1999)