The Knowledge Engineering Review, Vol. 24:4, 353–374. © Cambridge University Press, 2009
doi:10.1017/S026988890999018X

Computer-assisted assessment of free-text answers

DIANA PÉREZ-MARÍN[1], ISMAEL PASCUAL-NIETO[2] and PILAR RODRÍGUEZ[2]

[1]*Language and Computer Systems I Department, Computer Science Faculty, Office 2025, Ampliación del Rectorado Building, Tulipán Street, 28933 Móstoles, Universidad Rey Juan Carlos, Madrid, Spain;*
e-mail: diana.perez@urjc.es
[2]*Computer Science Department of the Universidad Autónoma of Madrid, Calle Francisco Tomás y Valiente, 11, Cantoblanco 28049, Madrid, Spain;*
e-mail: ismael.pascual@uam.es, pilar.rodriguez@uam.es

Abstract

The automatic assessment of students' free-text answers has recently received much attention, due to the necessity of exploring and taking advantage of new and more complex computer-based assessment methods. In this paper, a review of the state-of-art of the field is presented, focusing on the techniques that underpin these systems and their evaluation metrics. Although there is still a long way to go so as to reach the ideal system, the fact that the existing systems are already being used commercially and as a second opinion in exams such as GMAT proves the uptake of this field.

1 Introduction

Teachers all over the world spend a great deal of time just marking students' works. Hence, they have to cut down the time they can devote to other duties. This, among other reasons such as reducing costs or serving double-checkers of human's scores to increase their reliability, has motivated the research in Computer-Assisted Assessment (CAA). Nowadays, CAA has many possibilities of application, such as scoring the students' assignments (summative assessment), producing feedback to discover if the students have learned what the teacher intended (formative assessment) and evaluating assessment effectiveness (Blayney & Freeman, 2003).

According to most authors, the main goal of CAA is not to substitute teachers, but to support them in their tutoring task (Mason & Grove-Stephenson, 2002). Therefore, CAA is typically formative although it can also be used with summative purposes. Furthermore, plagiarism detection systems and protections against cheating are included in most of these systems (Denton, 2003; Sealey *et al*., 2003).

Most of the initial work in CAA was devoted to designing closed questions, such as fill-in-the-blank or multi-choice questions (MCQ). However, many authors agree that MCQs do not really measure the higher cognitive skills (Birenbaum *et al*., 1992; Foltz *et al*., 1999; Mcgrath, 2003; Mitchell *et al*., 2003; Palmer & Richardson, 2003; Parsons *et al*., 2003).

Although there has always been hard critics about the idea of a computer grading human essays, the advances in Natural Language Processing (NLP) and Machine Learning techniques, the popularization of online e-learning environments, the lack of time to give students appropriate feedback (despite the general assumption of its importance) and the conviction that MCQs cannot be the only computer-based assessment method are favoring a change in this situation.

Automatic assessment of students' free-text answers can be seen as including two different subtypes: automatic assessment of *short answers* and automatic assessment of *essays*. Sometimes the same tool can evaluate both kinds but, in general, the boundaries between the two tasks are clear and most CAA tools only evaluate either essays or short answers. Some systems that will be considered out of the scope of this review are semi-automated computer-based essay marking systems (Marshall &

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Barron, 1987), systems that assess the student ability to summarize (Kintsch *et al.*, 2000) and systems to improve the student writing skills (Wiemer-Hastings & Graesser, 2000). Besides, the paper will center on automatic assessment of electronic texts, given that optical character recognition of hand-written text is a wholly different problem addressed by a separate research community.

The key question is how a computer can effectively measure either the conceptual accuracy of a text, or its technical writing quality. In fact, according to some authors, the key question is how a computer can effectively measure both the conceptual accuracy of the text and its technical writing quality (Christie, 2003). Concerning essay content evaluation, there are several approaches, most of which compare the student's answer against some reference (ideal answer) or template. In order to grade the technical writing quality, one traditional approach is to look for direct features in the text, such as word number or word lengths, and to use them to infer more abstract measures such as variety, fluency or quality (Page, 1966; Christie, 2003).

This paper is structured in the following way: Section 2 describes the techniques that are being currently used; Section 3 reports the evaluation metrics; Section 4 reviews the systems; and finally, Section 5 compares the systems and concludes with final remarks.

2 Techniques

There have been several classifications of techniques reported in the literature. Chung and O'Neill (1997) distinguished between *text classification* and *text understanding* approaches. Page classified system depending on whether they evaluate *content* or *style*, to which Valenti *et al.* (2003) added a third category for systems that evaluate *both*. Whittingdon and Hunt (1999) classified systems according to the knowledge representation formalism used to represent the contents of the texts (e.g. semantic networks, or Lexical Conceptual Structure).

Table 1 lists the currently available software tools for free-text CAA alphabetically ordered, and the main technique each of them uses. In this section, it is proposed to classify these techniques according to the level of NLP required. In particular, two categories have been distinguished: shallow NLP only involving statistical techniques for the lexical level, and full NLP involving more complex techniques not only for the lexical level but up to the syntax, semantics and/or discourse processing levels.

2.1 Shallow natural language processing

In general, all systems that rely on a statistical analysis of one or several features of the texts should be considered in this category. They usually need an initial training phase to calculate the parameters of the system. They do not use complex NLP techniques and, in most cases, the texts are only processed with a tokenizer, a sentence splitter and a part-of-speech (POS) tagger. As a consequence, they should be easy to port across languages and domains.

The following subcategories can be defined:

- **Keyword analysis.** It is the simplest technique and it consists in looking for coincident keywords between the student's answer and the references. A common model, also used in Text Categorization and Information Retrieval, is the Vector Space Model (VSM) (Salton *et al.*, 1975). In VSM, texts are represented as vectors in a hyperspace, where dimensions correspond to words. The magnitude of a document in a given dimension i is the frequency with which a word w_i appears in the document. Frequencies may be replaced with weights, such as $\mathrm{tf} \cdot \mathrm{idf}$ (Salton, 1989). Next, documents can be compared by calculating the cosine of the angle of their associated vectors. A lower cosine angle means a higher similarity with the reference answer and, therefore, a higher score. In general, if a student essay is very similar to a reference written by a teacher, it will receive a high score. VSM is used as an additional module in E-rater (Burstein *et al.*, 1998).

  Single keyword analysis may be substituted by N-gram analysis, where sequences of N consecutive words are compared between the reference answer and the student's answer. This is the approach of the statistical module in Willow (Pérez-Marín *et al.*, 2006).

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Table 1 Technical approaches taken in the current existing computer-assisted assessment of free-text answers systems

System	Reference	Technique
Automatic Essay Assessor	Kakkonen *et al.* (2005)	Variations of latent semantic analysis
Apex Assessor	Dessus *et al.* (2000)	Latent semantic analysis
Automated Text Marker	Callear *et al.* (2001)	Information extraction
Automark	Mitchell *et al.* (2002)	Information extraction
Auto-marking	Sukkarieh *et al.* (2003)	NLP and pattern matching
Bayesian Essay Test Scoring System	Rudner and Liang (2002)	Statistical
CarmelTC	Rosé *et al.* (2003)	Machine learning
C-rater	Burstein *et al.* (2001)	NLP
Essay Grading and Analysis Logic	Datar *et al.* (2004)	NLP
E-rater	Burstein *et al.* (1998)	NLP
Intelligent Essay Assessor	Foltz *et al.* (1999)	Latent semantic analysis
Intelligent Essay Marking System	Ming *et al.* (2000)	Pattern matching with clustering
IntelliMetric	Vantage Learning Technology (2000)	Artificial intelligence approach
Japanese Essay Scoring System	Ishioka and Kameda (2004)	Pattern matching
Larkey's system	Larkey (1998)	Text categorization technique
MarkIt	Williams and Dreher (2004)	NLP and pattern matching
MultiNet Working Bench	Lutticke (2005)	Comparison of semantic networks
Project Essay Grader	Page (1966)	Measurement of surface linguistic features
Paperless School Marking Engine	Mason and Grove-Stephenson (2002)	NLP
Research Methods Tutor	Wiemer-Hastings *et al.* (2004)	Latent semantic analysis
Schema Extract Analyse and Report	Christie (1999)	Pattern matching
Willow	Pérez-Marín *et al.* (2006)	Shallow NLP

NLP = Natural Language Processing.

This method is limited by the fact that it cannot extract a representation of the meaning of the student answer, and it is difficult for the method to deal with synonyms and polysemous terms.

- **Latent semantic analysis (LSA).** It can be considered as a further extension of the VSM. LSA is a complex statistical technique that was initially developed for indexing documents and Information Retrieval (Deerwester *et al.*, 1990). Like VSM, it can be applied to automated essay grading (Landauer & Dumais, 1997; Landauer *et al.*, 1997), using it to measure the similarity between the students' answers and the references written by the teacher. According to Dessus *et al.* (2000) this approach is quite robust and proves its name by finding the hidden relationships between words that could be in different documents or between documents that do not share words.

LSA works in the following way:

1. *The vectorial representation*. After, optionally, removing stopwords and stemming the text, each context (a whole document or portion of a document: sentences, paragraphs…) is represented by a vector in a space whose dimensionality is the size of the vocabulary. Each cell represents the frequency of a word in the context.
2. *The matrix representation*. All the vectors can be put together in a matrix, where columns represent dimensions, and rows represent the contexts under study.
3. *Transforming frequencies into weights*. The relevance of each word (each dimension) in the passage is now measured. The purpose of this step is to give low weights to words that are

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

equally common in every context, and high weights to words that are very representative of a particular context. Possible weight functions are tf · idf (van Rijsbergen, 1979; Salton, 1989) or the χ^2 function (Manning & Schutze, 2001).

4. *Singular value decomposition (SVD)*. The original matrix is decomposed into the product of three orthogonal matrices ULA^T. L is a diagonal matrix and contains the singular values (the eigenvalues).
5. *Dimensionality reduction transformation*. Only the k dimensions corresponding to the highest eigenvalues are kept in the three matrices. The reduction in dimensionality allows for faster calculations, and it ensures that most information is kept. Furthermore, the product of the reduced matrices corresponds to the matrix of rank k that better approximates the original matrix.
6. *Similarity computation*. The similarity between contexts is measured with the same procedures as in VSM, for example, as the cosine of the angle between the corresponding reduced row vectors.

The Intelligent Essay Assessor (IEA) and Apex Assessor are underpinned by LSA. Besides, this technique has also been employed in computer-based tutors with dialog interfaces to evaluate the quality of the students' answers (Wiemer-Hastings *et al.*, 1998). Some modifications have also been tried, for example, Kakkonen *et al.* (2005) use Probabilistic Latent Semantic Analysis (PLSA), which adds a sounder probabilistic model to LSA based on a mixture decomposition derived from the latent class model. This results in a more principled approach that has a solid foundation in statistics. However, it also has overfitting problems. In fact, the results achieved with PLSA are quite similar to those achieved with LSA.

- **Analysis of surface linguistic features.** This subcategory includes systems that require (a) a list of features that are going to be measured; (b) a training phase to discover the relative importance for each one of them in assessing a text; and (c) a calibration phase to adjust the weights to the optimal values. When evaluating a student's answer, it is necessary to extract the features from the text, and apply them to the learned model. For example, the values of the features may be used as independent variables in a linear regression function, learned in the previous step, whose result is the final score.

  Project Essay Grader (PEG) (Page, 1994) is based partly on this analysis, although it also uses some additional NLP software, such as a grammar checker, an electronic dictionary, a POS tagger and a parser.
- **Text categorization techniques.** This category comprises systems that face the free-text automatic scoring as a classification problem. In contrast with the previous techniques, in which it is possible to obtain a numerical score, now the classification is done using a discrete set of classes. The common practice is to have a set of predefined categories, such as *good* and *bad*, or a scale with N points indicating the degree of correctness. The purpose is to classify each student's answer in one of those categories. It is possible to use many different techniques: Bayesian networks (Larkey's system and Bayesian Essay Test Scoring sYstem (BETSY)), k-Nearest Neighbors classifiers, decision trees, and so on. A more advanced Bayesian model is the three-level hierarchical model called Latent Dirichlet Allocation (LDA) that has been tested in Automatic Essay Assessor (AEA) (Kakkonen *et al.*, 2005), although with worst results than using LSA.
- **Information extraction (IE).** IE consists in acquiring structured information from free text, for example, identifying entities of interest (Named Entities) in the text, relations between them and outputting the results either by filling a template (MUC7, 1998) or as a semantic network. In general, it does not require a deep parsing of the texts, so it can be considered as a shallow NLP technique. *Pattern-matching* is a technique commonly used for IE. It can be applied for evaluating students' answers in the following way: first, the text is broken into concepts and their relationships; next, the relations obtained are compared against a model template written by a human expert to produce the student's score. For instance, Automark, Automated Text Marker (ATM) and Schema Extract Analyse and Report (SEAR) are based on this approach.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

- **Clustering.** Answers that have similar word patterns are grouped to form a cluster with the same score. For example, the Intelligent Essay Marking System (IEMS) uses the clustering algorithm called Indextron (Mikhailov, 1998) to automatically assess the free-text students' answers.

2.2 Full natural language processing

Some more complex NLP tools that have been applied to free-text CAA are syntactic analyzers (Manning & Schutze, 2001), to identify constituents and syntactic dependencies between them, rhetorical parsers (Marcu, 2000), to find the discourse structure of a text (Burstein *et al.*, 2001), and semantic analyzers, to identify the role that constituents perform in the actions or states reported in the text (e.g. patient, agent, location and so on).

The combination of these techniques is expected to improve the earlier purely statistical approaches, as it is possible to obtain a discourse and semantic analysis, from which to effectively assess the student's answer. On the other hand, it is hard to accomplish and very difficult to port across languages, and the system may be very dependent on the quality of the NLP tools. Among the current systems, C-rater and Paperless School Marking Engine (PS-ME) are underpinned by these techniques.

A different approach, which can also be classified under this category, consists in comparing semantic networks expressing the answer of the student with the model semantic network given by the instructors. The comparison is currently done by focusing on which nodes appear in each net and which edges relate them. It is implemented in the MultiNet Working Bench (MRW) system (Lutticke, 2005).

Finally, as it is usually the case, it is possible to combine several techniques in order to take advantage of all of them. For instance, although E-rater is mainly based on discourse and linguistic analysis, it also makes use of VSM for capturing the use of vocabulary and performing the topical analysis. Auto-marking relies both on NLP and pattern-matching, Intellimetric on NLP, Machine Learning and statistical analyses, and CarmelTC on decision trees and Bayesian Network classifiers.

3 Evaluation procedures and metrics

This section describes some expected requisites from the teachers and students, and the evaluation metrics and procedures most commonly used.

3.1 Requisites

Darus and Stapa (2001) collected the requisites that both teachers and students would expect from a free-text CAA system. Therefore, they surveyed 88 lecturers (22 from Economics and Business, 26 from Arts and Social Science and 40 from Language and Education) about the requirements they expected for a system that they would be willing to use. The results obtained are shown in Table 2. In the second survey, they interviewed 190 students from the Faculty of Language Studies in order to find out what is the most important factor for them in a CAA system. The answer was rather unanimous: feedback. On the other hand, there was some dispersion about which kind of feedback was better, as shown in Table 3.

Although there were some differences between the two populations, it is relevant to note that there was ample agreement in that feedback about the errors in the answers is the most important output that a system should provide (formative assessment). Many students also consider important that the system scores their work (summative assessment). Curiously, most of the evaluation procedures used up to now, described in the next subsection, only refer to the summative side of CAA, so it seems that there is a lack of evaluation metrics for formative assessment.

3.2 Procedures

It is possible to make a distinction between moderated experiments, and blind or unmoderated experiments (Vantage Learning Technology, 2000; Mitchell *et al.*, 2002).

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Table 2 Lecturers' expectation about what a free-text computer-assisted assessment system should provide them in order to be useful (Darus *et al.*, 2001)

Wished function	Percentage of respondents (%)
Indicate errors	57.5
Mark syntax	47.5
Provide error statistics	47.5
Mark non-native speakers writing	42.5
Produce letter grade	42.5
Mark organization of ideas	40.0
Mark surface features	37.5
Mark rhetorical structure	37.5
Mark topic content e.g. look at vocabulary	35.0
Give individual feedback	35.0
Mark holistically	35.0
Mark knowledge content e.g. look at semantics	32.5
Mark analytically	32.5
Mark according to disciplines	30.0

Table 3 Students' expectation about what a free-text computer-assisted assessment system should provide them in order to be useful (Darus *et al.*, 2001)

Areas of feedback	Percentage of respondents (%)
Errors in the essay	84.2
Organization of ideas	65.3
Coherence of text	63.2
Rhetorical structure	60.0
Their English dominion	53.7
Knowledge content	52.6
Topic content	51.6
Creativity	51.6
Style of writing	50.5
Syntax	40.0

When a blind experiment is performed, the system does not know anything about the grades assigned by human experts. This procedure is useful whenever the goal of the evaluation is to find out the goodness of the techniques used by the system. For instance, when Mitchell *et al.* wanted to evaluate AutoMark, they created a set of computerized mark schemas to serve as template for the assessment, and asked a group of students to submit their answers using a blind procedure. Automark gave a score to each answer without having any knowledge about which score would have been given by a human teacher. Afterwards, the teachers of the subject were asked to mark the same set of students' answers, without knowing the score provided by Automark. Finally, the performance of Automark was measured by comparing the automatic scores with the human scores for the same set of students' answers. This comparison can be made using one of the metrics explained in the next subsection.

On the other hand, when a moderated experiment is performed, the grades assigned by human experts are known. This procedure is useful whenever the goal of the evaluation is to identify system errors. For instance, when Mitchell *et al.* wanted to identify errors in Automark, they used computerized mark schemas with test data, and asked a group of students to submit their answers. Automark gave a score to each answer that was checked against the human score. In case, the answer had been unexpected but correct, the automatic score would have been low, whereas the human teacher would have been high, and this discrepancy would have been taken into account to

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

improve the system, so that the next time a similar unexpected but correct answer is presented, the system can correctly automatically evaluate it.

3.3 Metrics

Some common evaluation metrics are the following:

- **Pearson correlation or inter-rater reliability.** It measures the standard correlation, that is, how much the teacher scores (X) are related to the system scores (Y). Sometimes, the true scores are the result of the average consensus of several teachers. It is calculated in the following way:

$$\text{Correlation}(X, Y) = \frac{\text{covariance}(X, Y)}{\text{stdDev}(X) \times \text{stdDev}(Y)} \tag{1}$$

- **Spearman or non-parametric correlation.** After ranking the teacher scores and the system scores, the Spearman correlation measures the Pearson correlation between the ranks. If there are n paired ranks, and d_i is the difference in the ranks of the two variables for the ith student's answer, the Spearman rank correlation value is calculated as:

$$r_s = 1 - \frac{6 \times \sum d^2}{n \times (n^2 - 1)} \tag{2}$$

- **Kappa measure.** Kappa is a measure of agreement obtained by comparing the observed levels of agreement with the levels of agreement expected by chance. If O_a is the observed count of agreement, E_a is the expected count of agreement and N is the total number of respondent pairs, Kappa is calculated as:

$$k = \frac{O_a - E_a}{N - E_a} \tag{3}$$

- **Exact agreement.** It measures the percentage of times that the system and the human rater have scored exactly the same value. If the scores are real values, it may be necessary to partition the possible set of scores. In fact, this metric is usually applied when the set of possible scores is discrete, such as when a Text Classification procedure is applied for classifying the students' answers in a few categories (e.g. *bad* or *good*) (Larkey, 1998).
- **Adjacent agreement.** It measures the percentage of times that the system and the teacher only differ within one point (Larkey, 1998). It can be generalized to any threshold for agreement:

$$\text{AdjAgr} = \frac{|\{t_i : |\text{true_score}(t_i) - \text{system}(t_i)| < \Theta\}|}{N}$$

- **Mean and standard deviation.** If the system's scores are reliable, their distributions (and, therefore, their mean and standard deviation) should be similar to those obtained for the teacher's scores (Vantage Learning Technology, 2000).
- **False alarm rate.** It measures the percentage of times that a system cannot score a student's answer because it is too different to the models or the training materials (Rosé *et al.*, 2003).

4 Existing systems

In this section, the current free-text CAA systems are presented in alphabetical order to study their main features. For each of them, a brief introduction is presented, followed by a short system description and evaluation.

4.1 Automatic essay assessor

The AEA (Kakkonen *et al.*, 2005) was created in the Computer Science Department of the University of Joensuu in Finland. It is able to assess essays written in Finnish by comparing the student's essay with a set of assignment-specific texts corpus such as textbook passages, lecture notes and so on.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

First of all, as Finnish is a morphologically complex language, they have to process the text with a morphological analyzer, constraint grammar parser for lemmatization and a syntactic parser. Next, they originally applied the LSA technique to the reference corpus. However, they have recently also tried PLSA and LDA in this phase. In any case, what they obtain is the representation to which the human-graded essays are compared and the threshold similarity values for each grade category are determined. Finally, the LSA, PLSA or LDA representation of the student's essay is compared to the LSA, PLSA or LDA representation obtained in the previous phase and the similarity value of the essay is matched to the grade categories according to their limits to determine the correct grade.

To evaluate the performance of the system, they carried out an experiment using three essay sets collected from courses on education, marketing and software engineering summing a total of 100–150 essays. They tested all the possible dimensions for LSA (i.e. from two to the number of passages in the comparison materials) and the same number of dimensions for PLSA (just to make fair the comparison since PLSA has no restrictions in the number of latent variables). The results indicated that the best technique between LSA, PLSA and LDA is: LSA achieving 75% correlation between the automatic grades and the human grades for the same set of questions.

4.2 Apex assessor

Apex Assessor (Dessus *et al*., 2000) is integrated inside the Apex Web-based learning environment. When students want to study a topic in Apex, they only have to select it and start reading. A final review is done to assess the students' progress in which open-ended questions are asked. The Apex Assessor is responsible for selecting these questions and evaluating them.

Apex Assessor was created in the year 2000 by Dessus, Lemaire and Vernier in the Laboratorie des Sciences de L'Éducation in the Université Pierre-Mendés in France. According to its authors, the aim of the assessment process guided by Apex Assessor is not just summative but formative. Dessus *et al*. want to engage the students in an iterative improvement process in which they write their texts, and then they receive feedback about the outline and the coherence of their essays in order to give the students the possibility of rewriting their essays and send them again.

This system is underpinned by LSA. Thus, it needs a set of unmarked texts for training. This set includes non-technical French texts too to allow the system to deal with non-domain terms that might appear in the student answer.

Apex Assessor has three main modules:

- **Content-based assessment module.** It compares the student LSA representation answer with the LSA model.
- **Outline assessment module.** For each paragraph in the student's text, the most similar portion of the course is showed so that the student can be given an outline view of the essay.
- **Coherence assessment module.** It measures the semantic distance between sentences with LSA. Hence, if the proximity between two consecutive sentences is below a threshold, a coherence break is detected and the student is warned.

To analyze the system performance, Dessus *et al*. took 31 essays of a graduate course on sociology of education and typed them into the computer in order to compare the teachers' grade with the Apex one. The result was 59% correlation with $p < 0.001$.

Finally, it is important to mention that one problem Dessus *et al*. found out was that very short answers of students could achieve high scores. To solve it they have set up the system in order to be stricter with texts that do not have at least 300 words.

4.3 Automated text marker

The ATM (Callear *et al*., 2001) was created in the year 2001 by Callear, Jerrams-Smith and Soh in the Portsmouth University in the UK. They were so convinced that both content and style should be taken into account that they designed their system in order to give two independent scores, one for each aspect and to leave the teacher the task of combining them to give the final grade.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

ATM relies on IE techniques to assess students' essays. It is important to highlight its syntax and semantics analyzer:

- **The syntax analyzer.** It checks the grammar of each input sentence. According to Callear *et al.* this can be done successfully. A codification in Prolog is given by Callear *et al.* (2001).
- **The semantics analyzer.** The system looks for concepts in the text and their dependencies, and then a pattern-matching Prolog procedure is performed between the dependency groups from the student's answer and the reference model.

According to its authors, ATM works better assessing short answers to factual questions (e.g. in Prolog programming, psychology and biology-related fields). To our knowledge, Callear *et al.* have not yet published information about their system's performance.

4.4 Automark

Automark (Mitchell *et al.*, 2002) was created in the year 1999 by Mitchell, Russell, Broomhead and Aldridge from the University of Liverpool and Brunel University in UK. At the beginning, it was an academic work, but in the year 2002 they founded their company, the so-called Intelligent Assessment Technologies, and they started using it commercially. Incidentally, in the year 2002 it was made available in ExamOnline just for registered users.

The aim of the system is mostly summative, that is, to grade the style and the content of a student essay in order to say whether it is acceptable or not, according to the criteria specified by the teacher to the system.

Automark uses IE techniques and some NLP techniques to ignore some mistakes in spelling, typing, syntax or semantics that should not be taken into account. The Automark-assessing process is the following: in the first step, human experts develop some computerized reference mark schemes for the acceptable and unacceptable answers; in the second step, the system standardizes the punctuation and spelling of the student's text; the third step identifies the main syntactic constituents in the student's text and their relationships; the fourth step applies the pattern-matching module that will look for the scheme templates features in the syntactic constituents of the student's answer, trying to cover multiple paraphrasing; and finally, the fifth step processes the output of the pattern-matching module and generates the feedback for the student. It usually consists only of the score, but it might be possible to produce additional information.

The system has been used in the Brunel University to test Java knowledge of first year engineering students, and it has also been applied to assess answers from the 1999 statutory national curriculum assessment of science. In this case, students were 11-year-old pupils, and there were four types of questions: single word generation, single value generation, generation of a short explanatory sentence and description of a pattern in data. The correlation achieved ranged between 93% and 96%.

Finally, four problems can be identified: to correctly identify misspelled words, to correctly analyze the sentence structure, to identify an incorrect answer and to assess information that is not represented in the mark scheme template.

4.5 Auto-marking

Auto-marking (Sukkarieh *et al.*, 2003) was developed by Pulman, Sukkarieh and Raikes in Oxford and in the Interactive Technologies in Assessment and Learning (ITAL) Unit of the University of Cambridge Local Examinations Syndicate (UCLES). Its aim is not to automatically score high-stakes exams, but to help in low-stakes ones. Each exercise is given a value between 0 and 2, where 0 means incorrect, 1 partially correct or incomplete and 2 correct and complete.

This system relies on a combination of NLP and pattern-matching techniques. It consists of three modules:

- **Customization and shallow processing module.** First, it uses a Hidden Markov Model POS tagger, and two finite-state machine chunkers to chunk the noun and verb phrases. Sometimes, an additional manual tuning is necessary.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

- **The pattern-matcher module.** It is very similar to the one used in Automark, that is, human experts have to design the information extraction patterns and then the students' answers are compared against them. However, given the difficulty of writing good rules, they have devised a language to express the rules for finding the Information Extraction patterns automatically from the human orders.
- **The marking algorithm module.** These rules are organized in classes and the algorithm described in Sukkarieh *et al*. (2003) matches them with the student's processed answer to score it.

The system has been applied with answers from the GCSE exam of Biology with 88% of exact agreement between the teacher and the system. On the other hand, the authors claimed that this system is not suitable for subjective general opinions and therefore it should not be used in those areas.

The main problem they encountered was the inaccuracy of taggers, which do not have enough knowledge about Biology. Besides, they stated that their system cannot deal with students' inferences and with contradictory or inconsistent information.

4.6 Bayesian essay test scoring system

The BETSY system (Rudner & Liang, 2002) was developed between 2001 and 2003 by Rudner and Liang at the College Park of the University of Maryland with funds from the US Department of Education. According to the authors, its aim is to classify essays using a four-point nominal scale (e.g. extensive, essential, partial, unsatisfactory) taking into account both the content and the style.

BETSY is underpinned by naive bayesian networks. The user is given the possibility of choosing one of two models: Multivariate Bernouilli Model (MBM) and Bernouilli Model (BM). Rudner and Liang claim that BM is quicker as it only looks if certain features are present, while MBM takes into account the uses in which these features have been employed.

BETSY has the possibility of stemming the text and removing the stopwords, which might improve the text classification task. Besides, as it is very CPU-demanding, the authors thought of adding the possibility of purging infrequent words and phrases that appear less than five times per thousand.

The system has been used to assess Biology items for the Maryland High School and the results were that BM achieved 80% accuracy and MBM 74%. Furthermore, Rudner and Liang say that their system could be applied to any text classification task.

4.7 C-rater

C-rater (Burstein *et al*., 2001) is an automated scoring engine created by Educational Testing Service (ETS). It is focused on measuring the students' understanding of content material, without taking into account the students' writing skills.

The building of the reference model in C-rater is manual. In particular, it is done with the Alchemist user-interface by the experts in the topic. They store what they consider that the students should know to pass the open-ended questions.

C-rater uses NLP techniques to find out whether a student response contains the linguistic information required by the experts to prove that the student can pass the question.

Moreover, C-rater uses several syntactic constraints to recognize correct responses so that not only word matching is applied. It also deals with syntactic variations, words in different inflections, misspelled words, synonyms, similar terms, and used of pronouns to recognize paraphrases of the correct linguist information required.

When C-rater was used in a small-scale study with a university virtual learning program, it achieved over 80% agreement with the instructor and, according to Leacock (2004), when it was used in a large-scale assessment to score 170,000 short-answer responses to 19 reading comprehension and five algebra questions, the result was 85% accuracy. Furthermore, the Kappa value for C-rater/human agreement was 0.77, which represents excellent agreement beyond chance.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

4.8 CarmelTC

Carmel is a Virtual Learning Environment system that has recently incorporated a new free text assessment module called CarmelTC (Rosé *et al.*, 2003). This module has been developed at the University of Pittsburgh by Rosé, Roque, Bhembe and Vanlehn. Apart from giving the student a score, it can be used to find out which set of correct features are present in student essays.

CarmelTC relies on the combination of machine learning classification methods using the features extracted from the Carmel's linguistic analysis of the text and the Rainbow Naive Bayes classification.

The procedure for assessing a student's answer is the following: the first step is to break the text in sentences, the second step is to use the Bayesian network to look for the possible correct feature that represents each sentence, in order to generate a vector indicating the presence or absence of each correct feature and finally, the third step induces the rules for identifying sentence classes based on these feature vectors with the ID3 tree learning algorithm (Quinlan, 1993).

It can be applied to many several domains, including causal ones such as physics, which are out of the limits of traditional bags of words approaches. In CarmelTC, thanks to the functional relations found by Carmel, they can be successfully processed.

The system was tested with 126 physics essays, and the results were 90% precision, 80% recall and 8% false alarm rate.

4.9 Essay grading and analysis logic

The Essay Grading and Analysis Logic (EGAL) (Datar *et al.*, 2004) is a system developed by a group of American students. It is a source open system based on four criteria: gibberish detection, relevance to the question, identification of facts and their accuracy. They can be used as independent modules or together. In fact, according to its authors, it is more efficient together as whenever a sentence is marked as gibberish, there is no point in continuing studying if it is relevant, as well as if the sentence is marked as irrelevant, as it is not necessary to continue checking whether it is a statement of fact.

Gibberish can be semantic or syntactic. In order to determine whether a certain sentence is semantically gibberish or not, the stopwords (without lexical meaning) are removed and the rest of the words are stemmed. Thus, it can be calculated by its semantic similarity using WordNet. The mean of the semantic similarities values is the semantic coherence (s) and the percentage gibberishness value is calculated as $100 \times (1 - s)$. Whenever it is above a certain threshold, the sentence is flagged as semantic gibberish. Provided that the sentence is not semantic gibberish, it is checked whether it is syntactic gibberish by parsing the sentence with the module Lingua:: LinkParser and looking at the percentage of unknown words and unused links in the sentence. Whenever it is above a certain threshold, the sentence is flagged as syntactic gibberish. The relevance is calculated by calculating the similarity using WordNet between the words of the essay and words in texts about the topic under assessment. Finally, the statement of facts are identified by looking at some rules such as that the sentence is not in future tense and that it contains fact words instead of opinion words.

The system has not been evaluated in depth yet. The authors only report results over seven essays. Thus, the sample is too small to infer any general conclusion. Nevertheless, they state that so far the system's performance has been satisfactory and, that the detection of gibberishness, relevancies and statement of facts is being as expected.

4.10 E-rater

E-rater is a writing analysis tool that automatically evaluates, and scores essays written in English (Burstein *et al.*, 1998). It has been created by ETS, where it was being used since 1999 to score exams such as GMAT or TOEFL.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

E-rater is also the scoring engine integrated in the ETS Web-based application called Criterion. Criterion has also another component called Critique, which uses E-rater to:

- Generate immediate feedback about errors in grammar, usage and mechanics.
- Identify the essay's discourse structure.
- Recognize undesirable stylistic features in the essay under evaluation.

E-rater is based on an analysis of writing features using NLP techniques to replicate the scoring performance of human teachers. Some of these features are: organization and development of ideas in the text, lexical complexity, vocabulary usage or essay length. Not all the features are used to produce the automatic holistic score. For instance, in E-rater v1.3 a variable subset from 8 to 12 predictive features were selected by a stepwise linear regression from a larger set of approximately 50 features. While in E-rater v2.0, the number of features selected is not only low but fixed, and it is not necessary to build a different model per topic.

Whenever E-rater is not able to score the text because it is too short, or too different from the rest, it generates an advisory message (Burstein *et al*., 2001).

The correlation between E-rater and human scorers for the same data set of students' answers is 93%. This value is higher than the correlation between two different human scorers for the same data set, which was evaluated by ETS as 50%. In fact, the test-retest reliability of E-rater v2.0 scores (for a single essay) in a 6th- to 12th-grade population (0.60) is higher than the test-retest reliability of a single human rater (0.50), and is comparable to the average of two human raters (0.58).

4.11 Intelligent essay assessor

The IEA (Foltz *et al*., 1999) was created in the 1997 year by Landauer, Foltz and Laham. It was originally conceived as an academic product but, some years later, they founded their own company called Knowledge Analysis Technology. They are now in the process of patenting their system. Moreover, IEA cannot be executed in an ordinary PC but on secure Web servers placed in their company in USA. The authors claimed that as IEA is a Web-based application it only takes 20 s for students to receive their feedback.

The main goal of IEA is to assess the knowledge conveyed in the essay, rather than its style, syntax or argument structure.

IEA is underpinned by LSA. This statistical technique has been briefly exposed in Section 2. More information about LSA can be found in Deerwester *et al*. (1990) and Landauer *et al*. (1997). One of its main advantages is its language independence, with the restriction that it is not able to process too complex morphological structure of the language.

According to Chung and O'Neill (1997) three main modules can be distinguished:

- **The content module.** It is the most important module. It uses the LSA vector to extract the quality score as the weighted average of the scores for the k most similar calibration essays, and the domain relevance score as the length of the essay's vector.
- **The mechanics module.** Punctuation and spelling are analyzed in order to grade the essay's mechanics.
- **The style module.** It takes into account the essay's coherence, which is measured with the LSA value of relatedness among contexts, and the essay's grammar, which is measured with the LSA value of resemblance between the grammatical structure of the essay's sentences and the sentences of the model.

It is also possible to perform synonym recognition in order to treat several synonyms with similar meanings as the same word. The system is also able to identify if students have based their essays more in one reference text or in other. Therefore, it can give the score and feedback to the students with the subtopics that are not enough covered in their essays and links to the reference texts. The system will allow them to resend their essay with the suggested modifications to improve it.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Another technique employed is the anomalous essay checking, that is, the use of a flag to warn the teacher that the essay is too different from the others to reliably assess it and that s/he should review it because maybe the student is having difficulties or maybe s/he is trying to cheat.

According to its authors, IEA can be used in many different applications within education, from the simple consistency checker, to help teachers to discover cheating and plagiarism, to the formative and summative assessment of the essays. It requires an initial training but it is not human supervised. The only input is a set of texts about the topic to evaluate.

IEA has been tested in the military environment with $\sim$2000-word essays achieving 0.35 inter-reliability between the teacher and the system (between teachers it was lower, 0.31) (Streeter *et al*., 2003). IEA has also been used for psychology, medicine and history texts, achieving 80%–90% exact agreement when a 0–100 scoring scale was being used.

Landauer *et al*. stated that one problem their system has is that it does not take into account the word order. Thus, it cannot interpret sentences in which word order is the discriminant factor. Besides, it is easily tricked because it does not perform any syntactical or grammatical analysis.

4.12 Intelligent essay marking system

The IEMS (Ming *et al*., 2000) was presented by Ming, Mikhailov and Kuan from the NGEE ANN Polytechnic in Singapore, in 2000. Its aim is both summative and formative.

IEMS is based on the Pattern Indexing Neural Network, the Indextron that performs pattern recognition and in this case the patterns are the words of the texts. Further information can be found in Mikhailov (1998).

This system has been mostly applied to qualitative questions (e.g. biology, psychology, history or anatomy) rather than numerical ones. In the instance of taking an 800-word passage entitled '*Crime in Cyberspace*' and asking 85 students of third-year Mechanical Engineering to write a summary of not more than 180 words about the text, IEMS achieved 80% correlation with the teacher's scores.

4.13 IntelliMetric

IntelliMetric (Vantage Learning Technology, 2000) was created by the company Vantage Learning, after having spent more than 10 millions dollars in its development. It is a commercial system whose focus is emulating the human scorer by grading the content, the style, the organization and the conventions of each response using a 1–4 scale.

IntelliMetric requires an initial training phase with a set of manually scored answers in order to infer the rubric and the human graders' judgments to be applied by the automatic system. From the initial 100 features that IntelliMetric could take into account, it chooses the most appropriate for the topic under study. For instance, some features could be the focus and unity that indicate the purpose and main idea of the text: the organization and structure, which indicate the logic of discourse; or the conventions, that indicate conformance to English language rules.

Because it is not an academic product, there is little published information about the techniques that it employs. However, Vantage Learning Technologies has stated that IntelliMetric relies on other of their proprietary systems, the so-called CogniSearch and the Quantum Reasoning Technologies. Moreover, they have claimed that they used an Intelligent Artificial approach, because IntelliMetric uses its intelligence to score the students' texts.

This system has extensively been used in schools, high schools and companies. For instance, it has been used with 594 texts written by students aged 11. Using 100 texts for training, they achieved 98% adjacent agreement. Furthermore, it has assessed essays that are not written in English language, such as Hebrew attaining 84% correlation (Vantage Learning Technology, 2001).

4.14 Japanese essay scoring system

The Japanese Essay Scoring System (JESS) (Ishioka & Kameda, 2004) is the first automated Japanese essay scorer. It has been created in the National Research Center for the University

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Entrance Exam in Japan. It examines three features in the essay: rhetoric (i.e. syntactic variety), organization (i.e. how ideas are presented and related in the essay) and content (i.e. how relevant is the information provided and how precise and related to the topic is the vocabulary employed).

For rhetoric assessment, Jess measures a set of items such as the ease of reading, diversity of vocabulary, percentage of big words and passive sentences. For organization, it attempts to determine the logical structure of the document by detecting the occurrence of certain conjunctive expressions. For content, it uses LSA (the training is done using editorials and columns taken from the Mainichi Daily News newspaper as learning models).

Jess has been evaluated with 480 applicants who wrote an essay about the meaning of work in their life. Three experts scored each essay independently. The correlation achieved between the system scores and the mean human scores was 57% higher than 48% found between the expert raters' scores. This result was improved in another experiment in which 143 university students were asked to write about '*festivals in Japan*' with 84% correlation between the automatic and by hand scores. Again higher than the inter-rater correlation that was 73%.

4.15 Larkey's system

Larkey had been working on text categorization techniques to assess students' essays in the University of Massachusetts in USA and she produced her system in 1998 for classifying the students' essays as 'good' or 'bad'. It considers both their content and their style (Larkey, 1998)

The assessing procedure could be one of the following, or a combination of them:

1. **Bayesian classifiers.** Each document is assigned a probability of belonging to one previously specified category of documents. To achieve this goal, two steps are performed: the first one is the feature selection that removes stopwords, stems the text and looks for the most representative features using Bayesian networks. It next trains using the Lewis binary model, so that 0 means that the feature is not in the text and 1 is just the opposite.
2. **Finding the *k* most similar reference essays.** The Inquery retrieval system is used to find the k essays closest to the student essay.
3. **Using eleven text complexity features.** Eleven features are automatically calculated from the text. Some of them are the number of characters in the document, the number of different words in the document, the average sentence length, the average word length and the number of words longer than seven characters.

The score is the result of the linear regression performed with the results of the values for the features, the results of the Bayesian classifiers or a combination of the three methods.

Larkey's system has been applied to essays on social studies, physic questions and legal arguments, achieving 60% exact agreement and 100% one-point-of-difference agreement when all the criteria for assessment were used. Therefore, she has even tried to assess general opinion questions, with the results of 55% exact agreement and 97% of one-point-of-difference. The correlation attained was always above 80% and, in particular, for the general opinion essays, it was 88%.

4.16 MarkIT

MarkIT (Williams & Dreher, 2004) is a free-text scoring system that gives feedback to the students about how they have used the concepts in the essays. It has been developed by a research team at the School of Information Systems in the Curtin University of Technology in Australia. It uses propriety technology based on NLP techniques, LSA and an electronic thesaurus to process and compare the student's answer with the model answer that has been extracted from a set of e-learning contents.

First of all, it is necessary to train the system by feeding it with 50–200 human-graded essays (the assessment is better as more human raters score the test and the score is averaged). In this way, it can be tuned to use multiple linear regression. Next, both the student's answer and the model answer are processed by a specially designed chunking algorithm called Context Free Phrase Structure Grammar parser to identify the noun phrases and verb clauses in them. During

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

this phase, a transformational grammar is used to represent the semantics of the content and the thesaurus to extract lexical information and, the internal knowledge representation of the answers is built. Finally, pattern-matching techniques are employed to ascertain the proportion of model answer knowledge present in the student answer that is scored accordingly.

The system has been tested with 390 essays written by year 10 high-school students on the topic of '*The school leaving age*'. They were asked to write their essays in Microsoft Word document format. Next, these essays were submitted to three different human graders and MarkIT. Two hundred of them were used as training and the other 190 essays as test. The results for the test were 75% correlation between human scorers and the authors claimed that MarkIT performed as well as human graders (although no numerical information was provided).

4.17 MultiNet working bench

The MRW (Lutticke, 2005) is a graphical tool to assess student knowledge. It has been created in the Computer Science Department of the FernUniversitat in Germany. It is based on the MultiNet paradigm whose core idea is to represent natural language as semantic networks, in which the nodes refer to discourse entities and the edges to semantic relations between them. Inner nodes are for more complex concepts and a fixed set of 110 relations has been defined.

MRW is able to represent, edit and assess semantic networks in MultiNet form. The analysis can be done from the net as drawn by hand by a student or from its natural language reformulation. In any case, the internal representation of the student answer as semantic network and a reference solution is compared using logic inference and the result can be that the text is wrong, with missing fragments, unverified or verified. This result is given to the student as textual and graphical information. For instance, wrong or unverified parts of the student's network are marked in red and verified parts in green. The feedback can also be enriched with support hints such as links to literature or examples.

The system is currently subject to further development and extension. Nevertheless, a preliminary version has been used in a practical NLP course imparted in the FernUniversitat with promising results. However, no numerical results are published yet.

4.18 Project essay grader

The PEG (Page, 1966) was first presented in 1966 by Page in the University of Duke in USA. It focused on the style of the essay.

At the beginning, no NLP technique was used and the system was based only on a statistical approach that consisted in looking for several features (proxes) that represent abstracter ones (trins). According to Chung and O'Neill (1997), PEG considered 28 different proxes such as the title, the average sentence length, the number of paragraphs, the punctuation and the number of prepositions in 1966. Incidentally to score the students' essays, PEG is introduced in a number of previously manually marked essays for proxes to calculate the coefficients for the regression equation that finally will give the students' scores. In 1990, the system was improved with a grammar parser and a POS tagger to improve the proxes discovery. Moreover, in conformity with Shermis *et al.* (2002), PEG currently includes content, organization, style, mechanics and creativity assessment.

PEG is suitable for most type of essays, achieving 87% correlation between its scores and human ones.

4.19 Paperless school marking engine

The PS-ME (Mason & Grove-Stephenson, 2002) is the system presented by Mason and Grove-Stephenson in the Birmingham University in UK in 2002, and it has also become commercially available. Its assessment objective is both summative and formative, with little or no human intervention. Besides, it can be integrated in a learning management system, or be used as a stand-alone application.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

PS-ME relies on NLP techniques to cover Bloom's taxonomy (Bloom, 1956):

- **Knowledge level.** It exactly corresponds to the Bloom knowledge competence. According to Mason and Grove-Stephenson, it is only necessary to create, from the reference texts, a list of the most relevant concepts that should be present in the student's essay in order to evaluate this level.
- **Understanding level.** It comprises the competencies between knowledge and evaluation in the Bloom's taxonomy: comprehension, application, analysis and synthesis. Mason and Grove-Stephenson refuse to give more details about this level arguing that it is commercially sensitive.
- **Evaluation level.** It matches the Bloom evaluation competence. PS-ME's authors are not too convinced that this level can be effectively measured by a computer, since the teacher usually scores higher a creative opinion of a student than one that is based on a reference text. However, they have included this option that could be based only on the frequency of adjectives and adverbs in the text; or even better, by looking for some syntactic patterns such as "*I think that X…*" or "*It is obvious that X*".

PS-ME requires an initial training phase with at least 30 hand-marked sample texts that could include not only reference texts, but also 'negative' texts with a very low score. Besides, Mason and Grove-Stephenson thought that due to processing requirements, their system should not be used for real-time essay grading. Instead, they implemented it as a Web-based system that sends the information in XML to a queuing system in the server. Finally, PS-ME does not only give the score but some formative feedback to the student in different areas within the subject.

This system has been applied to low-stakes coursework, National Curriculum Grade and GCSE exam in the academic field. In the commercial field, it has usually been employed by publishers. The main problems found were the difficulty for selecting master texts and the misspellings and bad grammar mistakes that, in words of Mason and Grove-Stephenson, could 'throw the system out'. To our knowledge, Mason and Grove-Stephenson have not yet published PS-ME results.

4.20 Research methods tutor

The Research Methods Tutor (RMT) (Wiemer-Hastings *et al.*, 2004) is a Web dialog-based tutoring system result of the joint effort of the Computer Science and Psychology departments of the DePaul University in USA. It has been designed to be flexible enough to integrate different tools and techniques for improving tutoring.

RMT is based on LSA. This means that, first of all, it needs to be trained with a set of reference texts. Next, it evaluates the student response by transforming it to its LSA representation and comparing it to the LSA representation of the expected answers. In this way, the intelligent tutor can continue asking the student according to the good or bad answer provided by the student. Currently, they are exploring the possibility of improving the technique by segmenting input sentences into subject, verb and object parts and comparing each separately.

The system was integrated as a regular component of the Research Methods in Psychology course in order to find out how students use such systems during a whole term. However, due to technical difficulties with the agent software and some compliance issues with the students, they did not get significant results. Further experiments are planned in the short future.

4.21 Schema extract analyse and report

The SEAR system was presented in the Robert Gordon University in UK to assess both the style and content of the students' essays (Christie, 1999, 2003). In general, the system is underpinned by Information Extraction techniques. However, the algorithms for assessing style and content are different:

- **For style**, four steps are needed: to pre-determine the candidate metrics, to have some manually marked system, to calibrate the system in order to find an acceptable agreement between the human expert and SEAR, and to process the student's essay just by looking for these features and applying the weight of each metric to compute the score as the result of a weighted linear function.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

- **For content**, neither training nor calibrating is necessary. The teacher needs to create some reference schemes. It uses Information Extraction techniques to fill in the students' schemes with the students' data and to compare them against the references.

It has been applied to assess essays about the potted history of Robert Gordon (the founder of the Robert Gordon University). The results attained are from 30% to 59.4% correlation between the system and the human scores.

4.22 Willow

Willow is an automatic system for evaluating students' answers (Pérez-Marín *et al.*, 2006) developed in the Universidad Autónoma de Madrid in Spain. It is the adaptive version of Atenea (Alfonseca *et al.*, 2004) and it is currently integrated in the Will Tools to automatically generate students' conceptual models (Pérez-Marín *et al.*, 2007).

It is based on the use of *N*-gram precision and recall metrics between the student's answer and a set of references written by the teacher. The texts are processed by several shallow NLP modules: stemming, closed-class words removal, naive Word Sense Disambiguation and LSA (Pérez *et al.*, 2005). It has also been integrated with the TANGOW educational system (Alfonseca *et al.*, 2004).

The average correlation is 54%. Although, it has attained correlations with the teachers' scores as high as 90% in short answers about definitions on Computer Science concepts (Operating Systems and Object-Oriented Programming).

5 Comparison and conclusions

Despite the core idea of all these systems is quite similar: to compare the student answer with one or more reference texts or model content, a complete objective comparison cannot be done

Table 4 Domains to which the current existing computer-assisted assessment of free text answers systems have been applied and their availability

System	Domain	Availability
Automatic Essay Assessor	Marketing and software engineering	Academic
Apex Assessor	Sociology of education	Academic
Automated Text Marker	Factual disciplines	Academic
Automark	Science	Academic
Auto-marking	Biology	Academic
Bayesian Essay Test Scoring System	Any text classification task	Free
CarmelTC	Physic	Academic
C-rater	Comprehension and algebra	Academic
Essay Grading and Analysis Logic	Opinion and factual texts	Free
E-rater	GMAT exam	Academic
Intelligent Essay Assessor	Psychology and military	Commercial
Intelligent Essay Marking System	Non-mathematical texts	Academic
IntelliMetric	K-12 and creative writing	Commercial
Japanese Essay Scoring System	General topic essays	Academic
Larkey's system	Social and opinion	Academic
MarkIt	General topic essays	Academic
MultiNet Working Bench	NLP course on semantic networks	Academic
Project Essay Grader	Non-factual disciplines	Academic
Paperless School Marking Engine	NCA or GCSE exam	Commercial
Research Methods Tutor	Research on Psychology	Academic
Schema Extract Analyse and Report	History	Academic
Willow	Conceptual domains	Academic

NLP = Natural Language Processing.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Table 5 Overview of the techniques, evaluation and language of the reviewed free-text computer-assisted assessment systems

System	Technique	Evaluation	Language
Automatic Essay Assessor	LSA, PLSA or LDA	Corr: 0.75	Finnish
Apex Assessor	LSA	Corr: 0.59	French
Automated Text Marker	Pattern matching	—	English
Automark	Information Extraction	Corr: 0.95	English
Auto-marking	NLP and pattern matching	EAgr: 0.88	English
Bayesian Essay Test Scoring System	Bayesian networks	CAcc: 0.77	English
CarmelTC	ML and Bayesian networks	f-S: 0.85	English
C-rater	NLP	Agr: 0.83	English
Essay Grading and Analysis Logic	NLP and statistics	—	English
E-rater	NLP and VSM	Agr: 0.97	English + others
Intelligent Essay Assessor	LSA	Agr: 0.85	English
Intelligent Essay Marking System	Pattern matching	Corr: 0.80	English
IntelliMetric	CogniSearch and Quantum	Agr: 0.98	English + others
Japanese Essay Scoring System	Pattern matching	Corr: 0.71	Japanese
Larkey's system	TCT	EAgr: 0.55	English
MarkIt	NLP, pattern matching and statistics	Corr: 0.75	English
MultiNet Working Bench	Logical inference	—	German
Project Essay Grader	Linguistic features	Corr: 0.87	English
Paperless School Marking Engine	NLP	—	English
Research Methods Tutor	LSA	—	English
Schema Extract Analyse and Report	Pattern matching	Corr: 0.45	English
Willow	Hybrid: statistics and NLP	Corr: 0.54	Spanish + English

LSA = Latent Semantic Analysis; PLSA = Probabilistic Latent Semantic Analysis; LDA = Latent Dirichlet Allocation; NLP = Natural Language Processing; ML = Machine Learning; VSM = Vector Space Model; TCT = Text Categorization Techniques.
Possible metrics are: Corr = correlation; Agr = agreement; EAgr = exact agreement; CAcc = classification accuracy; f-S = f-Score; and, — for not available. When the authors have presented several values for the results, the mean value has been taken.

because they use different corpora and evaluation metrics. Nevertheless, in an attempt to put together the different techniques and results provided by their authors, Tables 4 and 5 are presented with the systems alphabetically ordered.

It can be seen that each system has been applied to a somewhat different assessment area, depending on the technique that they are employing. For example, LSA, that has its focus on the content, is mostly used for the assessment of humanities' essays (Wiemer-Hastings *et al.*, 1998), while IE techniques, that rely on filling schemes, can be used both for factual and non-factual disciplines. In fact, LSA has performed poorly in causal domains such as research methods (Malatesta *et al.*, 2002). Moreover, according to Callear *et al.* (2001) neither Apex Assessor not IEA are suitable to assess short answers where the word order is important.

Therefore, it can be concluded that statistical techniques are mostly useful to deal with technical writing quality features, such as length of the essay, lexical variability or vocabulary usage. On the other hand, whenever the evaluation of the content is more relevant than the writing skills of the student, more NLP is required. At least, it is necessary to include some syntactic constraints to distinguish the subject and complement of the sentences, and to do some parsing to go beyond the simple word matching without taking into account the order of the words by identifying the basic structure of the answer.

The most useful metrics to evaluate free-text CAA systems are similarity metrics. This is because these systems are usually evaluated against human scores. In fact, it is considered that the best scoring engine is the one that produces automatic scores that are more similar to the scores given by one or more human teachers. Therefore, any measure of agreement will be a good metric,

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

and depending on the accuracy required on the automatic score it could be distinguished between exact agreement or adjacent agreement. In our opinion, adjacent agreement is more realistic as it is not usually necessary that the human and the automatic score are exactly the same, although it is usually necessary that the scores do not differ more than a certain value depending on the accuracy required (i.e. low-stakes exams will need a lower accuracy than high-stakes exams).

There is also much discussion and disagreement about which system can be considered as the best. For example, Rudner and Gagne (2001) stated that, among IEA, E-rater and PEG, the best choice for evaluating writing style is PEG. This is because it relies on writing quality features to determine the grades. Besides, it is simpler and it consumes less CPU. On the other hand, IEA and E-rater are better for grading content. However, since IEA can be tricked, as it does not perform any NLP processing, E-rater can be thought as the best one at the cost of being the most complex. This opinion is also shared by Williams (2001) who said that in terms of comparison with human markers, E-rater is best, followed by IEA, Appex Assessor, Larkey's system and finally PEG.

All the same, Rudner and Liang (2002) claimed that Bayesian networks are the best approach because they are easy to implement, they combine the advantages of PEG, LSA and E-rater and, they are perfectly suitable to assess short essays. Despite that, in conformity with Cucchiarelli *et al.* (2000) the main weakness of all these systems is the lack of a very large corpus of essays that may become a reference for everyone interested in automated essay grading.

Finally, it is interesting to mention how free-text CAA has prospered and has not been limited just to English texts and academic purposes. In fact, the level of performance achieved by some of these systems has made possible their use as commercial applications, and although there is still no current system that could be highlighted as the best one by all comparisons, the technology exists, and provided that the goals are kept realistic, the future of this field is very promising.

Acknowledgement

This work has been sponsored by Spanish Ministry of Science and Technology, project number TIN2007-64718.

References

Alfonseca, E., Carro, R., Freire, M., Ortigosa, A. & Pérez, D. 2004. Educational adaptive hypermedia meets computer assisted assessment. In *Proceedings of the International Workshop of Educational Adaptive Hypermedia, collocated with the Adaptive Hypermedia (AH) Conference*, Eindhoven, The Netherlands.

Birenbaum, M., Tatsuoka, K. & Gutvirtz, Y. 1992. Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement* **14**(4), 353–363.

Blayney, P. & Freeman, M. 2003. Automated marking of individualised spreadsheet assignments: the impact of different formative self-assessment options. In *Proceedings of the 7th Computer Assisted Assessment Conference*, Loughborough, UK.

Bloom, B. 1956. Taxonomy of educational objectives: the classification of educational goals. *Handbook I, Cognitive Domain*. Longman, Whiteplains (New York); Toronto.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Bradenharder, L. & Harris, M. D. 1998. Automated scoring using a hybrid feature identification technique. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, The Association of Computational Linguistics, Montreal, Quebec, Canada.

Burstein, J., Leacock, C. & Swartz, R. 2001. Automated evaluation of essays and short answers. In *Proceedings of the 5th International Computer Asssited Assessment Conference*, Loughborough, UK.

Callear, D., Jerrams-Smith, J. & Soh, V. 2001. CAA of short non-MCQ answers. In *Proccedings of the 5th International Computer Assissted Assessment conference*, Loughborough, UK.

Christie, J. 1999. Automated essay marking—for both style and content. In *Proceedings of the 3rd Computer Assisted Assessment International Conference*, Loughborough, UK.

Christie, J. 2003. Automated essay marking for content—does it work? In *Proceedings of the 7th International Computer Assisted Assessment Conference*, Loughborough, UK.

Chung, G. & O'Neill, H. 1997. *Methodological Approaches to Online Scoring of Essays*. Technical Report 461, UCLA, National Center for Research on Evaluation, Student Standards, and Testing, USA.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Cucchiarelli, A., Faggioli, E. & Velardi, P. 2000. Will very large corpora play for semantic disambiguation the role that massive computing power is playing for other AI-hard problems? In *Proceedings of the 2nd Conference on Language Resources and Evaluation*, Greece.

Datar, A., Doddapaneni, N., Khanna, S., Kodali, V. & Yadav, A. 2004. *EGAL—Essay Grading and Analysis Logic*, SourceForge project. http://egal.sourceforge.net

Darus, S., Hussin, S. & Stapa, S. 2001. Students' expectations of a computer-based essay marking system. In *Reflections, Visions and Dreams of Practice: Selected papers from the IEC 2001 International Education Conference*, Malaysia, 197–204.

Darus, S. & Stapa, S. 2001. Lecturers' expectations of a computer-based essay marking systems. *Journal of the Malaysian English Language Teachers' Association (MELTA)* **30**, 47–56.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* **41**(6), 391–407.

Denton, P. 2003. Evaluation of the 'electronic feedback' marking assistant and analysis of a novel collusion detection facility. In *Proceedings of the 7th Computer Assisted Assessment Conference*, Loughborough, UK.

Dessus, P., Lemaire, B. & Vernier, A. 2000. Free text assessment in a virtual campus. In *Proceedings of the 3rd International Conference on Human System Learning*, Paris, France, 61–75.

Foltz, P., Laham, D. & Landauer, T. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* **1**(2). Available online at http://imej.wfu.edu/articles/1999/2/04/index.asp

Ishioka, T. & Kameda, M. 2004. Automated Japanese Essay Scoring System: JESS. In *Proceedings of the 15th International Workshop on Database and Expert Systems Applications*, 4–8.

Kakkonen, T., Myller, N., Timonen, J. & Sutinen, E. 2005. Automatic Essay Grading with Probabilistic Latent Semantic Analysis. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, Association for Computational Linguistics, 29–36.

Kintsch, E., Steinhart, D., Stahl, G. & the LSA Research Group, 2000. Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments* **8**, 87–109.

Landauer, T. & Dumais, S. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* **104**, 211–240.

Landauer, T., Laham, D., Rehder, B. & Schreiner, M. 1997. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, Erlbaum, Mawhwah, New Jersey, 412–417.

Larkey, L. S. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, 90–95.

Leacock, C. 2004. Scoring free-responses automatically: A case study of a large-scale assessment. English version of Leacock, C. 2004. Automatisch beoordelen van antwoorden op open vragen; een taalkundige benadering. *Examens Journal* **1**(3).

Lutticke, R. 2005. Graphic and NLP Based Assessment of Knowledge about Semantic Networks. In *Proceedings of the Artificial Intelligence in Education conference*. IOS Press.

Malatesta, K., Wiemer-Hastings, P. & Robertson, J. 2002. Beyond the short answer question with research methods tutor. In *Proceedings of the Intelligent Tutoring Systems Conference*, Lecture Notes in Computer Science **2363**. Springer; San Sebastian.

Manning, C. & Schutze, H. 2001. *Foundations of Statistical Natural Language Processing*. MIT Press.

Marcu, D. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

Marshall, S. & Barron, C. 1987. Marc-methodical assessment of reports by computer. *System* **15**(2), 161–167.

Mason, O. & Grove-Stephenson, I. 2002. Automated free text marking with paperless school. In *Proceedings of the 6th International Computer Assisted Assessment Conference*, Loughborough, UK.

Mcgrath, P. 2003. Assessing students: Computer simulation vs MCQs. In *Proceedings of the 7th Computer Assisted Assessment Conference*, Loughborough, UK.

Mikhailov, A. 1998. Indextron. *Intelligent Engineering Systems Through Artificial Neural Networks* **8**, 57–67.

Ming, Y., Mikhailov, A. & Kuan, T. 2000. Intelligent essay marking system. In *Learners Together*, Cheers, C. (ed.). NGEE ANN Polytechnic.

Mitchell, T., Aldridge, N., Williamson, W. & Broomhead, P. 2003. Computer based testing of medial knowledge. In *Proceedings of the 7th Computer Assisted Assessment Conference*, Loughborough, UK.

Mitchell, T., Russell, T., Broomhead, P. & Aldridge, N. 2002. Towards robust computerised marking of free-text responses. In *Proceedings of the 6th Computer Assisted Assessment Conference*, Loughborough, UK.

MUC7. 1998. *Proceedings of the 7th Message Understanding Conference (MUC-7)*. Morgan Kaufmann, California, USA.

Page, E. 1966. The imminence of grading essays by computer. *Phi Delta Kappan* **47**, 238–243.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Page, E. 1994. Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education* **2**(62), 127–142.

Palmer, K. & Richardson, P. 2003. On-line assessment and free-response input—a pedagogic and technical model for squaring the circle. In *Proceedings of the 7th Computer Assisted Assessment Conference*, Loughborough, UK.

Parsons, H., Schofield, D. & Woodget, S. 2003. Piloting summative Web assessment in secondary education. In *Proceedings of the 7th Computer Assisted Assessment Conference*, Loughborough, UK.

Pérez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodríguez, P. & Magnini, B. 2005. Automatic assessment of students' free-text answers underpinned by the combination of a Bleu-inspired algorithm and latent semantic analysis. In *Proceedings of the 18th International Conference of the Florida Artificial Intelligence Research Society*, American Association for Artificial Intelligence (AAAI), Menlo Park, California.

Pérez-Marín, D., Alfonseca, E., Rodríguez, P. & Pascual-Nieto, I. 2006. Willow: Automatic and adaptive assessment of students free-text answers. In *Proceedings of the 22nd International Conference of the Spanish Society for the Natural Language Processing (SEPLN)*, Zaragoza, Spain.

Pérez-Marín, D., Alfonseca, E., Rodríguez, P. & Pascual-Nieto, I. 2007. Automatic generation of students' conceptual models from answers in plain text. In *Proceedings of the User Modeling International Conference*, Conati, C., McCoy, K. & Paliouras, G. (eds). Lecture Notes in Artificial Intelligence **4511**, 329–333. Springer-Verlag.

Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Rosé, C., Roque, A., Bhembe, D. & VanLehn, K. 2003. A hybrid text classification approach for analysis of student essays. In *Proceedings of the HLT-NAACL Workshop on Educational Applications of NLP*, Edmonton, Canada.

Rudner, L. & Gagne, P. 2001. An overview of three approaches to scoring written essays by computer. *Educational Resources Information Center (ERIC) digest*, ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

Rudner, L. & Liang, T. 2002. Automated essay scoring using bayes' theorem. In *Proceedings of the Annual Meeting of the National Council on Measurement in Education*, New Orleans, LA.

Salton, G. 1989. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

Salton, G., Wong, A. & Yang, C. 1975. A vector space model for automatic indexing. *Communications of the ACM* **11**(18), 613–620.

Sealey, C., Humphries, P. & Reppert, D. 2003. At the coal face. Experiences of computer-based exams. In *Proceedings of the 7th Computer Assisted Assessment Conference*, Loughborough, UK.

Shermis, M., Koch, C., Page, E., Keith, T. & Harrington, S. 2002. Trait rating for automated essay scoring. *Educational and Psychological Measures* **62**, 5–18.

Streeter, L., Pstoka, J., Laham, D. & MacCuish, D. 2003. The credible grading machine: Automated essay scoring in the DOD. In *Proceedings of Interservice/Industry, Simulation and Education Conference (I/ITSEC)*, Orlando, Florida, USA.

Sukkarieh, J., Pulman, S. & Raikes, N. 2003. Auto-marking: using computational linguistics to score short, free text responses. In *Proceedings of the 29th IAEA Conference, Theme: Societies' Goals and Assessment*, Philadelphia, USA.

Valenti, S., Neri, F. & Cucchiarelli, A. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education* **2**, 319–330.

van Rijsbergen, C. J. 1979. *Information Retrieval*. Butterworths.

Vantage Learning Technology 2000. *A Study of Expert Scoring and Intellimetric Scoring Accuracy for Dimensional Scoring of Grade 11 Student Writing Responses*. Technical Report RB-397, Vantage, USA.

Vantage Learning Technology 2001. *A Preliminary Study of the Efficacy of Intellimetric for Use in Scoring Hebrew Assessments*. Technical Report RB-561, Vantage, USA.

Whittingdon, D. & Hunt, H. 1999. Approaches to the computerised assessment of free-text responses. In *Proceedings of the 3rd International Computer Assisted Assessment Conference*, Loughborough, UK.

Wiemer-Hastings, P. & Graesser, A. 2000. Select-a-kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments* **8**(2), 149–169.

Wiemer-Hastings, P., Allbritton, D. & Arnott, E. 2004. RMT: A dialog-based research methods tutor with or without a head. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, Springer-Verlag, Berlin.

Wiemer-Hastings, P., Graesser, A., Harter, D. & the Tutoring Research Group, 1998. The foundations and architecture of Autotutor. In *Proceedings of the 4th International Conference on Intelligent Tutoring Systems*, Springer-Verlag, New York, 334–343.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press

Williams, R. 2001. Automated essay grading: an evaluation of four conceptual models. In *Proceedings of the 10th Annual Teaching and Learning Forum: Expanding Horizons in Teaching and Learning*, Curtin University of Technology, Perth, Australia.

Williams, R. & Dreher, H. 2004. Automatically Grading Essays with Markit. In *Proceedings of Informing Science Conference*, Rockhampton, Queensland, Australia.

https://doi.org/10.1017/S026988890999018X Published online by Cambridge University Press