An Overview of Current Research on Automated Essay Grading

Salvatore Valenti, Francesca Neri and Alessandro Cucchiarelli
DIIGA - Universita' Politecnica delle Marche, Ancona, Italy

valenti@inform.unian.it neri@inform.unian.it alex@inform.unian.it

Executive Summary

Essays are considered by many researchers as the most useful tool to assess learning outcomes, implying the ability to recall, organize and integrate ideas, the ability to express oneself in writing and the ability to supply merely than identify interpretation and application of data. It is in the measurement of such outcomes, corresponding to the evaluation and synthesis levels of the Bloom's (1956) taxonomy that the essay questions serve their most useful purpose.

One of the difficulties of grading essays is represented by the perceived subjectivity of the grading process. Many researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness. This issue may be faced through the adoption of automated assessment tools for essays. A system for automated assessment would at least be consistent in the way it scores essays, and enormous cost and time savings could be achieved if the system can be shown to grade essays within the range of those awarded by human assessors.

This paper presents an overview of current approaches to the automated assessment of free text answers. Ten systems, currently available either as commercial systems or as the result of research in this field, are discussed: Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing service I, Electronic Essay Rater (E-Rater), C-Rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free text Marking Engine and Automark. For each system, the general structure and the performance claimed by the authors are described.

In the last section of the paper an attempt is made to compare the performances of the described systems. The most common problems encountered in the research on automated essay grading is the absence both of a good standard to calibrate human marks and of a clear set of rules for selecting master texts. A first conclusion obtained is that in order to really compare the performance of the systems some sort of unified measure should be defined. Furthermore, the lack of standard data collection is identified. Both these problems represent interesting issues for further research in this field.

Keywords: Automated Essay Grading, NLP, Computer-based Assessment Systems

Introduction

Assessment is considered to play a central role in the educational process. The interest in the development and in use of Computer-based Assessment Systems (CbAS) has grown exponentially in the last few years, due both to the increase of the number of students attending universities and to the possibilities provided by e-learning approaches to asynchronous and ubiquitous education. According to our findings (Valenti, Cuc-

chiarelli, & Panti., 2002) more than forty commercial CbAS are currently available on the market. Most of those tools are based on the use of the so-called objective-type questions: i.e. multiple choice, multiple answer, short answer, selection/association, hot spot and visual identification (Valenti et al., 2000). Most researchers in this field agree on the thesis that some aspects of complex achievement are difficult to measure using objective-type questions. Learning outcomes implying the ability to recall, organize and integrate ideas, the ability to express oneself in writing and the ability to supply merely than identify interpretation and application of data, require less structuring of response than that imposed by objective test items (Gronlund, 1985). It is in the measurement of such outcomes, corresponding to the higher levels of the Bloom's (1956) taxonomy (namely evaluation and synthesis) that the essay question serves its most useful purpose.

One of the difficulties of grading essays is the subjectivity, or at least the perceived subjectivity, of the grading process. Many researchers claim that the subjective nature of essay assessment leads to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness. Furthermore essay grading is a time consuming activity. According to Mason (2002), about 30% of teachers' time in Great Britain is devoted to marking. "So, if we want to free up that 30% (worth 3 billion UK Pounds/year to the taxpayer by the way) then we must find an effective way, that teacher will trust, to mark essays and short text responses."

This issue may be faced through the adoption of automated assessment tools for essays. A system for automated assessment would at least be consistent in the way it scores essays, and enormous cost and time savings could be achieved if the system can be shown to grade essays within the range of those awarded by human assessor. Furthermore, according to Hearst (2000) using computers to increase our understanding of the textual features and cognitive skills involved in the creation and in the comprehension of written texts, will provide a number of benefits to the educational community. In fact "it will help us develop more effective instructional materials for improving reading, writing and other communication abilities. It will also help us develop more effective technologies such as search engines and question answering systems for providing universal access to electronic information."

Purpose of this paper is to present a survey of current approaches to the automated assessment of free text answers. Thus, in the next section, the following systems will be discussed: Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing service I, Electronic Essay Rater (E-Rater), C-Rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free text Marking Engine and Automark. All these systems are currently available either as commercial systems or as the result of research in this field. For each system, the general structure and the performance claimed by the authors are presented.

In the last section, we will try to compare these systems and to identify issues that may foster the research in the field.

Current Tools for Automated Essay Grading

Project Essay Grade (PEG)

PEG is one of the earliest and longest-lived implementations of automated essay grading. It was developed by Page and others (Hearst, 2000; Page, 1994, 1996) and primarily relies on style analysis of surface linguistic features of a block of text. Thus, an essay is predominantly graded on the basis of writing quality, taking no account of content. The design approach for PEG is based on the concept of "*proxes*", i.e. computer approximations or measures of *trins*, in*trins*ic variables of interest within the essay (what a human grader would look for but the computer can't directly measure) to simulate human rater grading. *Proxes* include: essay length (as the amount of words) to represent the trin of fluency; counts of prepositions, relative pronouns and other parts of speech, as an indicator of complexity of sentence structure;

variation in word length to indicate diction (because less common words are often longer). Proxes are calculated from a set of training essays and are then transformed and used in a standard multiple regression along with the given human grades for the training essay to calculate the regression coefficients. These regression coefficients represent the best approximation to human grades when obtained according to proxes. Then, they are used with the proxes obtained from unmarked essays to produce expected grades. PEG purely relies on a statistical approach based on the assumption that the quality of essays is reflected by the measurable proxes. No Natural Language Processing (NLP) technique is used and lexical content is not taken in account. PEG also requires training, in the form of assessing a number of previously manually marked essays for proxes, in order to evaluate the regression coefficients, which in turn enables the marking of new essays.

PERFORMANCE: Page's latest experiments achieved results reaching a multiple regression correlation as high as 0.87 with human graders.

Intelligent Essay Assessor (IEA)

IEA was developed in the late nineties (Hearst, 2000; Jerrams-Smith, Soh, & Callear, 2001) and is based on the Latent Semantic Analysis (LSA) technique that was originally designed for indexing documents and text retrieval (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). LSA represents documents and their word content in a large two-dimensional matrix semantic space (Whittington & Hunt, 1999; Williams, 2001). Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationship are modified to more accurately represent their true significance. A matrix represents the words and their contexts. Each word being analysed represents a row of in the matrix, while each column represents the sentences, paragraphs, and other subdivisions of the context in which the word occurs. The cells of the matrix contain the frequencies of the words in each context. This initial matrix is then transformed according to an inverse document frequency weighting approach, a very well known technique of the indexing and information retrieval domain. The SVD is then applied to the matrix to subdivide it into three component matrices that reproduce the original one if multiplied together. Using a reduced dimension of these three matrices in which the word-context associations can be represented, new relationships between words and context are induced when reconstructing close approximations to the original matrix from the reduced dimension component SVD matrices. These new relationships are made manifest, whereas prior to the SVD they were hidden or latent. To grade an essay, a matrix for the essay document is built, and then transformed by the SVD technique to approximately reproduce the matrix using the reduced dimensional matrices built for the essay topic domain semantic space. (The semantic space typically consists of human graded essays). Cosine correlation is used to measure the similarity of the reduced dimensional space constructed from a "model answer", such as an instructional text taken from a course text or an essay prepared by the class tutor, against a student essay.

LSA makes no use of word order since the authors claim that this is not the most important factor for grabbing the sense of a passage. It also requires large amounts of data to construct a suitable matrix representation of word use/occurrence, and due to the size of the matrices involved computations are very cumbersome.

Key features of IEA include relatively low unit cost, quick customized feedback, and plagiarism detection. Furthermore, the authors claim that the system is very well suited to analyse and score expository essays on topics such as science, social studies, history, medicine or business, but not suitable to assess factual knowledge. IEA automatically assesses and critiques electronically submitted text essay, and represents a useful domain-independent tool. It supplies instantaneous feedback on the content and the quality of the student's writing.

PERFORMANCE: A test conducted on GMAT essays using the IEA system resulted in percentages for adjacent agreement with human graders between 85%-91%.

Educational Testing Service (ETS I)

This system, developed by Burstein and Kaplan of the Educational Testing Service in early nineties, works only on a sentence fragment of between 15 and 20 words (Whittington & Hunt, 1999). The technique uses lexical-semantic techniques to build a scoring system, based on small data sets. It uses a domain-specific, concept-based lexicon and a concept grammar, both built from training data.

The training data essays are parsed by Microsoft Natural Language Processing (MsNLP) tool, any suffixes are removed by hand, and a list of stop words is also removed. This produces a lexicon. The list of words and terms in the lexicon remain constant whilst the features associated with each entry are modular, so can be replaced as necessary. Some manual classification is necessary, such as specification of some words as metonyms of each other and so on.

Grammar rules are then constructed, again manually, for each category of answer (each category should contain all the paraphrases for that possible answer) using syntactic parses of sentences from the training data along with the lexicon.

New essays are then parsed by the phrasal node extraction program that outputs the sentences' noun, verb, adjective, adverb and preposition phrases etc. The system does not make use of specific parts of speech at this stage so they are collapsed into a generic XP phrase type and the sentence, or what is left of it now, is checked for matches against the grammar and the lexicon. The XP phrase type is taken from X-bar syntax, which attempts to model common properties between the different syntactic components of noun, verb, adjective and preposition phrase etc. Instead of building a different grammar rule for each different type of phrase, X-bar syntax generalizes out to a single rule that applies universally to all.

The system involves lots of pre-processing and much of it is manual, although the authors argue that the cost, in time, is still worth the saving.

PERFORMANCE: The authors of the ETS I system claim an accuracy of 80% when marking the essay test set and 90% when marking both the training and test set, i.e. using the training the system on a set of essays and then including it in the test set for the marking part as well. The most common source for errors is due to gaps in the lexicon, i.e. to words not manually identified as metonyms.

In a second reported experiment, the authors adopted an augmented lexicon. Constructing this new lexicon involved examining the test set as well as training set to manually place metonyms. This achieved accuracy of 93% when marking on the test set and of 96% when marking both the training and test sets.

Electronic Essay Rater (E-Rater)

E-Rater was developed by Burstein and others (Burstein, Kukich, Wolff, Chi, & Chodorow, 1998; Burstein, Leacock, & Swartz, 2001). E-Rater uses the MsNLP tool for parsing all sentences in the essay.

E-Rater uses a combination of statistical and NLP techniques to extract linguistic features from the essays to be graded. Essays are evaluated against a benchmark set of human graded essays. With E-Rater, an essay that stays on the topic of the question, has a strong, coherent and well-organized argument structure, and displays a variety of word use and syntactic structure will receive a score at the higher end of a six-point scale. E-Rater features include the analysis of the discourse structure, of the syntactic structure and of the vocabulary usage (domain analysis). E-Rater adopts a corpus-based approach to model building by using actual essay data to analyse the features of a sample of essay responses. The application is designed to identify features in the text that reflect writing qualities specified in human reader scoring criteria and is currently composed by five main independent modules. Three of the modules identify features that may be used as scoring guide criteria for the syntactic variety, the organization

of ideas and the vocabulary usage of an essay. A fourth independent module is used to select and weigh predictive features for essay scoring. Finally, the last module is used to compute the final score. E-Rater is currently embedded in Criterion a web-based real-time version of the system developed by ETS Technologies. A further feedback component with advisory features has been added to the system. The advisories are based on statistical measures and are completely independent from the score generated by E-Rater, thus providing additional feedback about qualities of writing related to topic and fluency only. E-Rater is trained on a collection of 270 essays that have been manually scored by trained human raters. E-Rater is far more complex and requires more training than many other available systems. Furthermore, no on-line demonstration and no downloadable trial version of E-Rater have been made available to the scientific community.

PERFORMANCE: Over 750000 GMAT essays have been scored, with agreement rates between human expert and system consistently above 97%. By comparing human and E-Rater grades across 15 test questions, the empirical results range from 87% to 94%.

Conceptual Rater (C-Rater)

C-rater is a NLP based prototype aimed at the assessment of short answers related to content-based questions such as those that may appear in a textbook's chapter review section (Burstein et al., 2001). C-rater adopts many of the some natural language processing tools and techniques developed for E-Rater, even if the two systems differ in many important ways.

E-Rater assigns a score for writing skills rather than for specific-content while C-rater is aimed to score a response as being either correct or incorrect. This goal is achieved by evaluating whether a response contains information related to specific domain concepts; if the response expresses these concepts it is rated as correct, otherwise it is rated as incorrect without any regard to writing skills.

Moreover, E-Rater provides grades that are partly based on the rhetorical structure of an essay while C-rater needs to identify specific content. It generates a fine-grained analysis of the logical relations between the syntactic components of each sentence appearing in the response. C-rater does not require a large collection of graded answers for training. Instead, it uses the single correct answer that is found in an instructor's guide or answer key, because it is believed unrealistic to require extensive data collection for the purpose of grading relatively low stakes quizzes, especially given that a set of short questions is often provided at the end of chapters in a textbook.

PERFORMANCE: C-Rater achieved over 80% agreement with the score assigned by an instructor.

Bayesian Essay Test Scoring sYstem (BETSY)

BETSY is a program that classifies text based on trained material and is being developed by Lawrence M. Rudner at the College Park of the University of Maryland with funds from the U.S. Department of Education (http://ericae.net/betsy).

According to Rudner and Liang (2002) the goal of the system is to determine the most likely classification of an essay into a four point nominal scale (e.g. extensive, essential, partial, unsatisfactory) using a large set of features including both content and style specific issues. The underlying models for text classification adopted are the Multivariate Bernoulli Model (MBM) and the Bernoulli Model (BM). With the MBM each essay is viewed as a special case of all the calibrated features, and the probability of each score for a given essay is computed as the product of the probabilities of the features contained in the essay. With the BM the conditional probability of presence of each feature is estimated by the proportion of essays within each category that contain the feature. This model can require a long time to compute since every term in the vocabulary needs to be examined. Both the Bernoulli Model and the Multivariate Bernoulli Model are considered naive Bayes models because they assume conditional independence.

According to its authors BETSY relies on an approach that may incorporate the best features of PEG, LSA (Landauer, Foltz, & Laham, 1998) and E-rater "plus it has several crucial advantages of its own. It can be employed on short essays, is simple to implement, can be applied to a wide range of content areas, can be used to yield diagnostic results, can be adapted to yield classification on multiple skills, and is easy to explain to non statisticians" (Rudner & Liang, 2002).

BETSY is a Windows-based program written in Power Basic, and is computationally intensive. Moreover, BETSY is the only software in the field that may be freely downloadable and useable.

PERFORMANCE: Rudner and Liang (2002) report about two text classification models that were calibrated using 462 essays with two score points. The calibrated systems were then applied to 80 new prescored essays, with 40 essays in each score group. An accuracy of over 80% was achieved with the described dataset.

Intelligent Essay Marking Systems (IEMS)

IEMS is based on the Pattern Indexing Neural Network (the Indextron) developed at NGEE ANN Polytechnic (Ming, Mikhailov, & Kuan, 2000). The system can be used both as an assessment tools and for diagnostic and tutoring purposes in many content-based subjects.

Students can be given immediate feedback and can learn where and why they had done well or not made the grade. Thus it can be embedded in an intelligent tutoring system that will help students to write better by grading papers fast and providing the feedback quickly. The essay grading is based on qualitative type of questions rather than numerical type. Indextron is defined as a specific clusterisation algorithm. In itself, such an algorithm is not a neural network. However, the clusterisation algorithm can be implemented as a neural network. The Indextron-based neural network attempts to overcome a slow, non-incremental training, which is typical of traditional Artificial Neural Networks.

PERFORMANCE: According to Ming et al. (2000) an experiment involving the evaluation of the essays produced by 85 students doing a module on Project Report Writing and coming from six classes of third-year Mechanical Engineering, obtained a correlation of 0.8.

Automark

Automark is a software system developed in pursuit of robust computerized marking of free-text answer to open-ended questions (Mitchell, Russel, Broomhead, & Aldridge 2002). Automark employs NLP techniques to mark open-ended responses. This software has been under development for almost three years, and has been employed in a commercial e-Learning product for the last months (ExamOnline as cited in Mitchell et al., 2002). Computer Assisted Assessment procedures based on Automark are currently being developed at a number of higher education establishments, including Brunel University where an online Java test for first year engineering student is under developed. The system incorporates a number of processing modules specifically aimed at providing robust marking in the face of errors in spelling, typing, syntax and semantics. Automark looks for specific content within free-text answers, the content being specified in the form of a number of marking scheme templates. Each template represents one form of a valid or a specifically invalid answer. Development of the templates in the computerized marking scheme is an offline process, achieved through a custom-written system configuration interface. The representation of the templates is robust enough to cope with multiple variations of the input text.

The marking process progresses through a number of stages. First, the incoming text is pre-processed to standardise the input in terms of punctuation and spelling. Then, a sentence analyser identifies the main syntactic constituents of the text and how they are related. The pattern-matching module searches for matches between the marking scheme templates and the syntactic constituents of the student text. Finally, the feedback module processes the result of the pattern match. Feedback is typically provided as a mark, but more specific feedback is claimed to be possible.

PERFORMANCE: Automark has been tested on National Curriculum Assessment of Science for eleven years old pupils. The form of response was: single word generation, single value generation, generation of a short explanatory sentence, description of a pattern in data. The correlation achieved ranged between 93% and 96%.

Schema Extract Analyse and Report (SEAR)

SEAR is a software system developed by Christie (1999) as a result of his PhD research work. According to Christie, the automated grading of essays requires the assessment both of style and content (where appropriate). Thus, the system provides an expandable, flexible method for the automated marking of the essay content and provides a method for the automated marking of the essay style too.

The methodology adopted to assess style is based on a set of common metrics, and requires some initial calibration. In essence the computer-based marking of the style is based on pre-determining what would be candidate metrics, use a subset of essays as training set and mark them manually. Then a process of calibration is started by adjusting the weight of each metric until an acceptable agreement between human and computer marking is achieved, processing the whole essay set.

For content assessment only those essay that are technical are candidates for this type of marking (they lead to a bounded spectrum of content) are taken into consideration.

For SEAR the content schema is prepared once and is revised fairly quickly and easily. Further, the SEAR content schema requires neither 'training' nor 'calibration', although the usual practice of taking a sample to verify the method would be recommended. The schema is held as a simple data structure. Two measures have been devised to aid the automated marking process: usage and coverage. The former aimed to measure how much of each essay is used, while the latter measures of how much of the essay schema is used by the essay under examination. Both measures were devised to envision the relationship between each essay and the schema.

PERFORMANCE: SEAR is still a work in progress. At this point, no performance indicators are available yet.

Paperless School free-text Marking Engine (PS-ME)

PS-ME is designed as an integrated component of a Web-based Learning Management System (Mason & Grove-Stephenson, 2002) and is still under development. Due to its processing requirements, the PS-ME does not grade essays in real-time.

This system applies NLP techniques to assess student essays in order to reveal their level of competencies as for knowledge, understanding and evaluation. The student essay is submitted to the server, together with information about the task in order to identify the correct master texts for comparison. Each task is defined via a number of master texts that are relevant to the question to be answered. An interesting issue is raised by the existence of 'negative' master texts containing a set of false statement composed using typical student mistakes and misconception. The essay to be rated is compared against each relevant master text to derive a number of parameters reflecting the knowledge and understanding exhibited by the student. The ability to evaluate parameter is calculated through a linguistic analysis as described above. When multiple master texts are involved in the comparison, each result from an individual comparison gets a weighting, which could be negative in the case of a master text containing common misconceptions. The weights are derived during the initial training phase.

The individual parameters computed during the analysis phase are then combined in a numerical expression to yield the assignment's grade (typically a National Curriculum grade or a GCSE level). The parameters are also used to select specific comments from a comment bank relevant to the task. With a fine-grained set-up it is possible to provide the student with formative feedback regarding his/hers per-

formance in different areas within a given subject. The output from the marking process is then passed back to the client for presentation to the teacher. This includes details on sections of essay that are particularly good or bad in relation to the knowledge, understanding and evaluation factors.

The process of setting up the auto marker for a particular task is very straightforward: select master text, from a number of sources such as textbooks, encyclopedias or relevant websites (the system is highly tolerant of duplication of content between master texts, but can lose accuracy if the master texts use extremely complex grammar); have a sample hand-marked (can be as few as 30, this needs to be done once only per task, in order to derive the right weightings for the parameter values computed by the marking system); run the same sample through the marker and perform regression analysis, which tries to get a best fit between the grades given by the marker and those resulting from the combination of the parameters; upload the resulting data to the server.

PERFORMANCE: The PS-ME is still a work in progress. Thus, no performance indicators are available yet.

Research Issues

In 1996, Page introduced a distinction between grading essays for content and for style, where the former refers loosely to what an essay says, while the latter to "syntax and mechanics and diction and other aspects of the way it is said" (Page, 1996). Some of the systems discussed in this paper evaluate essays primarily either for content (IEA, ETS I, C-Rater) or for style (PEG). Finally, some of the systems evaluate essays taking in account both aspects (BETSY, SEAR, Automark, PS-ME).

Another dimension that may be used to classify automated essay marking systems depends on the approach adopted for assessing style and/or content.

According to Page, the intrinsic variables of interest for grading the style of essays, i.e. fluency, diction, grammar and punctuation, cannot be measured directly but can be evaluated through possible correlates (proxes). For example, fluency "was correlated with the prox of the number of words" (Page, 1994).

Therefore, automated essay marking platforms may be classified according to the approach adopted for measuring content and style. We will adopt the term "Rating Simulation" for systems measuring intrinsic variables of interest either for content or style through proxes, and "Master Analysis" for systems measuring the actual dimensions (Williams, 2001).

The two coordinates discussed above have been summarized in Table 1.

	Rating Simulation	Master Analysis
Content	IEA, BETSY, IEMS, SEAR	ETS I, E-Rater, C-Rater, Automark, PS-ME
Style	PEG, BETSY, IEMS, SEAR	E-Rater, Automark, PS-ME

Table 1 - Automated essay grading systems' classification

Thus, for instance IEA grades essays for content adopting proxes, while PEG uses proxes for assessing style issues of essays. Therefore, the two systems have been put in the first row of Table 1. Whenever a system grades essays both for style and content is inserted in both columns of the table (as f.i. BETSY, IEMS, SEAR, E-RATER, Automark and PS-ME). A first conclusion that can be drawn from Table 1 is that most of the late systems developed are aimed to grade essays both for style and for content.

The most common problems encountered in the research on automated essay grading are the absence both of a good standard to calibrate human marks and of a clear set of rules for selecting master texts. The issue is made clear by Table 2, which lists all the platforms described in the paper, along with the underlying model, the claimed performance and the test bed adopted.

As a first remark, it must be noted that seven out of ten systems are based on the use of Natural Language Processing tools, which in some cases are complemented with statistical based approaches. This seems an interesting point for researchers interested in the development of new tools for automated essay grading.

As it appears immediately from Table 2, three different criteria have been reported to measure the performances of the systems: accuracy of the results (ACC), multiple regression correlation (CORR) and percentage of agreement between grades produced by the systems and grades assigned by human expert (AGREEM). A first conclusion that could be raised is that in order to really compare the performance of the systems some sort of unified measure should be defined.

Four recognizable sources of error in the computerized marking have been identified by most authors: failure to correctly identify misspelled or incorrectly used words, failure to properly analyze the sentence structure, failure to identify an incorrect qualification, omission of a mark scheme template. What is unclear is how the systems based on statistical approaches could face these issues. However, "the problem of incorrect qualification requires some innovative thinking" (Christie, 2002). According to Hearst (2000), pronoun and other co reference resolution tools are essential because anaphora abound in students' free-form responses. Therefore, the described drawbacks can be reduced through improvements in the sentence analyzers, spell checkers and semantic processors: thus, adopting NLP technologies should allow most of these issues to be solved.

The current research in automated text categorization field may provide useful results.

System	Technique	Performance			Test Bed
		ACC	CORR	AGREEM	
PEG	Statistical		87		
IEA	Algebra/NLP			85-91	GMAT essay
ETS I	NLP	93-96			
E-Rater	Statistical/NLP			87-94	GMAT essay
C-Rater	NLP			80	
BETSY	Bayesian Text Classification/Statistical	80			462 essays for training 80 essays as dataset
IEMS	Indextron		80		Essays from 85 students
Automark	NLP		93-96		National Curriculum Assessment of science
SEAR	NLP	-	-		
PS-ME	NLP	-	-		

Table 2 – A comparison of the systems described in section 2

Text categorization is the problem of assigning predefined categories to free text document. The idea of automated essay grading based on text categorization techniques, text complexity features and linear regression methods was first explored by Larkey (1998). The underlying idea of this approach relies on training of binary classifiers to distinguish "good" from "bad" essays and on using the scores produced by the classifiers to rank essays and assign grades to them. Several standard text categorization techniques are used to fulfill this goal: first, independent Bayesian classifiers allow assigning probabilities to documents estimating the likelihood that they belong to specific classes; then, an analysis of the occurrence of certain words in the documents is carried out and a k-nearest neighbor technique is used to find those essays closest to a sample of human graded essays; finally, eleven text complexity features are used to assess the style of the essays. Larkey conducted a number of regression trials, using different combinations of components. She also used a number of essay sets, including essays on social studies, where content was the primary interest and essay on general opinion where style was the main criteria for assessment.

A growing number of statistical learning methods have been applied to solve the problem of automated text categorization in the last few years, including regression models, nearest neighbor classifiers, Bayes belief networks, decision trees, rule learning algorithms, neural networks and inductive learning systems (Ying, 1997). This growing number of available methods is raising the need for cross method evaluation. In fact, the performance of a classifier strongly depends on the choice of data used for evaluation. Thus, comparing categorization methods without analyzing collection differences, and drawing conclusions based on the results of flawed experiments raise questions about the validity of some published evaluations. These problems need to be addressed to clarify the confusion among researchers and to prevent the repetition of similar mistakes. Integrating results from different evaluations into a global comparison by evaluating one or more baseline classifiers on multiple collections, by generating the performance of other classifiers using a common baseline classifier, and by analyzing collection biases based on variation of several baseline classifiers, has been shown to be possible by Yang (1998).

But the most relevant problem in the field of automated essay grading is the difficulty of obtaining a large corpus of essays (Christie, 2003; Larkey, 2003) each with its own grade on which experts agree. Such a collection, along with the definition of common performance evaluation criteria, could be used as a test bed for a homogeneous comparison of different automated grading systems. Moreover, these text sources can be used to apply to automated essay grading the machine learning algorithms well known in NLP research field, which consist of two steps: a training phase, in which the grading rules are acquired using various algorithms, and a testing phase, in which the rules gathered in the first step are used to determine the most probable grade for a particular essay. The weakness of these methods is the lack of a widely available collection of documents, because their performances are strongly affected by the size of the collection (Cucchiarelli, Faggioli, & Velardi, 2000). A larger set of documents will enable the acquisition of a larger set of rules during the training phase, thus a higher accuracy in grading.

Thus our research will be focused on the identification of a set of parameters that may be used to rate the performances of the automated essay grading systems in an unambiguous way; at the same time, we will try to promote the discussion among the researchers in the field, for the creation of a very large corpus of essays that may become a reference for everyone interested in automated essay grading.

References

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

Burstein, J., Kukich, K., Wolff, S., Chi, L., & Chodorow M. (1998). Enriching automated essay scoring using discourse marking. *Proceedings of the Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics, Montreal, Canada*.

Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essay and short answers. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

Christie, J. R. (1999). Automated essay marking-for both style and content. In M. Danson (Ed.), *Proceedings of the Third Annual Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

Christie, J. R. (2003). Email communication with author. 14th April.

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? *2nd. Conference on Language Resources and Evaluation (LREC), Athens, Greece*.

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 for Information Science*, *41*(6), 391-407.

de Oliveira, P.C.F., Ahmad, K., & Gillam, L. (2002). A financial news summarization system based on lexical cohesion. *Proceedings of the International Conference on Terminology and Knowledge Engineering, Nancy, France.*

Grondlund, N. E. (1985). *Measurement and evaluation in teaching*. New York: Macmillan.

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15*(5), 22-37, IEEE CS Press.

Honan, W. (1999, January 27). High tech comes to the classroom: Machines that grade essay. *New York Times*.

Jerrams-Smith, J., Soh, V., & Callear D. (2001). Bridging gaps in computerized assessment of texts. *Proceedings of the International Conference on Advanced Learning Technologies, 139-140*, IEEE.

Laham, D. & Foltz, P. W. (2000). The intelligent essay assessor. In T.K. Landauer (Ed.), *IEEE Intelligent Systems, 2000*.

Landauer, T. K., Foltz, P. W., & Laham D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*. Retreived from http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. In *Proceedings of the 21st ACM/SIGIR (SIGIR-98), 90-96*. ACM.

Larkey, L. S. (2003). Email communication with author. 15th April.

Mason, O. & Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

Ming, P.Y., Mikhailov, A.A., & Kuan, T.L. (2000). Intelligent essay marking system. In C. Cheers (Ed.), *Learners Together, Feb. 2000*, NgeeANN Polytechnic, Singapore. http://ipdweb.np.edu.sg/lt/feb00/intelligent_essay_marking.pdf

Mitchell, T., Russel, T., Broomhead, P., & Aldridge N. (2002). Towards robust computerized marking of free-text responses. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughboroug University, Loughborouh, UK.

Page, E.B. (1996). Grading essay by computer: Why the controversy? Handout for *NCME Invited Symposium.*

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

Palmer, J., Williams, R., & Dreher H. (2002). Automated essay grading system applied to a first year university subject- How can we do it better. *Proceedings of the Informing Science and IT Education (InSITE) Conference, Cork, Ireland*, 1221-1229.

Rudner, L.M. & Liang, T. (2002). Automated essay scoring using Bayes' Theorem. *The Journal of Technology, Learning and Assessment, 1*(2), 3-21.

Thompson, C. (2001). Can computers understand the meaning of words? Maybe, in the new of latent semantic analysis. *ROB Magazine*. Retrieved from http://www.vector7.com/client_sites/ROB_preview/html/thompson.html

Valenti, S., Cucchiarelli, A., & Panti M. (2000). Web based assessment of student learning. In A. Aggarwal (Ed.), *Web-based Learning & Teaching Technologies, Opportunities and Challenges*, 175-197. Idea Group Publishing.

Valenti, S., Cucchiarelli, A., & Panti, M. (2002). Computer based assessment systems evaluation via the ISO9126 quality model. *Journal of Information Technology Education, 1* (3), 157-175.

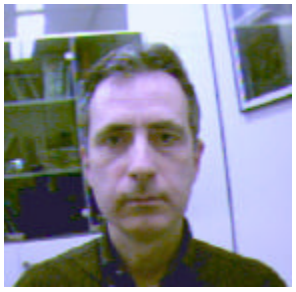
Whittington, D. & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, UK.

Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In A. Hermann & M.M. Kulski (eds). *Expanding Horizons in Teaching and Learning*. Proceedings of the 10th Annual Teaching and Learning Forum, Perth: Curtin University of Technology.

Yang, Y. (1997). An evaluation of statistical approaches to text categorization. *Technical Report CMU-CS*-97-127, School of Computer Science, Carnegie Mellon University, April 1997.

Biographies

Salvatore (Sal) Valenti is senior researcher at the Università Politecnica delle Marche, Italy. He has been a member of several research projects funded by the Ministry of Instruction, University and Research (MIUR), by the National Research Council (CNR) and by the European Community. His research activities are in the fields of Computer Based Assessment and on Distance Learning. He is Board member of the JITE. He is serving as reviewer for Educational Technology & Society and for Current Issues in Education. He has been appointed chair of the track on "Information Technology Education" at the 2004 International Conference of the International Resources Management Association. He is author of more than 60 papers published on books, journals and proceedings of international conferences.



Alessandro Cucchiarelli is senior researcher at the Università Politecnica delle Marche, Italy. His main research interests are focused on Automatic Evaluation of Software and on NLP techniques applied to Information Extraction. He has been also involved in research activities on Models and Tools for Cooperative Distributed Information Systems, Requirement Engineering and Robotics. He has been a member of groups working in several research projects funded by the Ministry of Research, the National Research Council and the European Union (Cost13, ECRAN). He is author of more than 70 papers published on books, journals and proceedings of international conferences.

Francesca Neri is a doctoral student at the Università Politecnica delle Marche, Italy. She received her Laurea Degree in Electronic Engineering from Università Politecnica delle Marche. Her research interests include natural language processing, machine learning and information extraction.