Abstract - The use of language in free text answers provides an excellent means for assessment, but it is hard work to mark this type of material. The automatic assessment of open questions can be applied successfully to carefully written literal text explicit and literal text implicit questions, both categories very frequent in Engineering. Existing automatic marking systems exhibit results comparable to manual marking, but most of them fall short of educators' expectations. Teachers like to keep control of marking, judge by themselves unusual answers, correct their own model answer on the fly. Hence, Computer Assisted Assessment (CAA) is preferred to fully automatic assessment. The inherent ambiguity of natural language and the difficulties of knowledge representation add to the constraints imposed by teachers' requirements. However, there is still space for a system which is both useful and practicable. This paper reviews the types of questions apt for automatic marking and their usefulness in Engineering, examines some representative existing automatic marking systems, states requirements of teachers, reviews state of the art possibilities and puts forward some ideas for the design of a computer assisted open question marking system based on a controlled language for answer writing and knowledge representation comparison for assessment.

Index Terms - Assessment, e-Assessment, Open questions, Free-text answers.

INTRODUCTION

Assessment is hard work in all its stages. Automatic assessment can be successfully carried out on closed questions such as [1] [2]:

- Multiple Choice, with one or more correct answers.
- True or False.
- Short Answer, a word or simple phrase from a list.
- Numerical, an exact number or a number within a range.
- Matching, a two-column concept matching question.
- Calculated Questions: random values are placed into an equation, result differs each time the question is seen.

Though useful to check whether students have grasped the essentials, closed questions offer poor information on the student's ability to actively manage and apply their recently acquired knowledge. Automatic marking of closed questions does not require much extra work, but preparing closed questions tests remains a tedious, exacting, time consuming chore which must, moreover, be carefully undertaken.

Assessment of Engineering knowledge can be achieved, to a great extent, by problem solving and calculation. However, calculations are of little use if not oriented by the right concepts and ideas. Open questions are a common way to get an insight into the conceptual maturity a student may have achieved after a learning period. In open questions the student is asked to produce a piece of text using his/her own choice of words and form of expression (free-text answers). Questions can be very specific, with only a phrase or two for a correct answer, or require the student to write a brief essay on the subject.

Even for short answer questions, the marking of open questions tests proves much harder than that of closed questions tests. The automatic marking of free text answers is still a field of research, but recently some proposals have been tried out with promising results. The types of questions most useful for Engineering Education, being mostly of a factual type, are apt for automatic marking.

Teachers keep reluctant to the use of free text automatic assessment systems [3]. Besides the difficulties of learning to use a new tool, and choosing one in the first place, some educative goals may not have been adequately addressed so far. A brilliant talk or piece of writing full of meaning to a human audience may be marked as complete nonsense by an automatic system. Not any type of question may be automatically assessed. A statistical probability will not be acceptable as a mark unless endorsed by a human teacher who has seen the answers by himself. Automatic marking depends on reference material not always available, or a carefully constructed reference answer. Teachers would not like to lose contact with their students.

Computer Assisted Assessment (CAA), instead of purely automatic marking, is closer to teachers' and students' expectations of a reliable assessment system.

This article starts by stating the role open questions may play in Engineering Education, points out the main actors' expectations and educative requirements for automatic open question assessment, examines some representative existing systems, goes through present day possibilities and limitations, advocates for CAA instead of completely automated marking, reviews the types of questions apt for automatic marking, states the requirements of a CAA system to be educationally satisfactory and puts forward some ideas for the design of one such system based on the use of controlled language and knowledge representation.
ROLE OF OPEN QUESTIONS IN ENGINEERING EDUCATION

Being strongly based in Mathematics and Logics, knowledge of Engineering subjects is most frequently assessed by problem solving and calculations. Closed questions may be used cost effectively to verify comprehension, within their known limitations. Generally, asking students to construct something provides better assessment than asking them to select among a small set of alternatives [4]. The use of open questions in Engineering Education may provide benefits in at least two dimensions of assessment: conceptual maturity and ability of communication.

The construction of Engineering solutions rarely starts by a calculation. Mathematics is the means of expression, but devising a solution remains essentially a qualitative process. Being able to select concepts from reality (or a question), search for related concepts in the student's mental body of knowledge, establish adequate relations, is closer to the creative process involved in the application of Engineering knowledge.

Students of Engineering, and graduated engineers, are generally not recognized as good writers or speakers, notwithstanding specific efforts present in modern syllabuses. Few Engineering students are aware of the number of reports, forms and written documents their professional life will ask them to produce. Ability to clearly express ideas, choose the adequate words, combine them in a bunch of readable sentences cannot but help students become familiar with the objective, concise, matter of fact style Engineering writing demands.

Assessment by open questions, eliciting free-text answers, may help fill the gap between closed questions and calculations on one side, and report writing on the other. Though training in report writing is included in most undergraduate courses one way or other, writing small pieces of free-text in an exam will certainly do no harm.

EDUCATIVE EXPECTATIONS

In Education, the different actors have different expectations concerning automatic assessment:

- institutions look for a reliable, cost effective system.
- teachers try to lighten the burden of their work without resigning their own judgment of students' assessment work.
- students expect a fair system capable of reflecting their true knowledge and capabilities, with results available in a reasonably short time.
- developers attempt to restrict the generic, intractable problem of natural language understanding into a computationally feasible task, imposing as few limitations as possible on the other actors' expectations.

A successful system should be able to satisfy all these expectations to some degree.

ASSESSMENT SYSTEM REQUIREMENTS

In open answer assessment several ambitious educative values come into play.

- **The language as a means of expression.** In written or oral work, a student is expected to exhibit a mastering of language adequate to the subject in question: a clear expression, richness of vocabulary, adequate use of sentence structure, a measured use of simile and metaphor, concision without awkwardness, a style adequate to the subject, some ambiguity to pique the reader's imagination, a cold determinism to report facts. Richness of language is that of life itself. In Engineering, however, clearness of expression, precision in choice of words, conciseness and matter-of-fact style are probably the most outstanding virtues. Mastering of language will help or hinder an engineer's career to a greater extent than is usually seen.

- **Original ideas.** No true educator would accept to let go inadvertently a student's original idea. Creative, original ideas are in the core of human progress. It is not known how much originality is being quenched by our present assessment systems, each day more structured.

- **The value of concept.** A human educator can easily differentiate a badly expressed correct concept from an erroneous concept, thus rewarding the idea and penalizing the expression.

- **Error correcting.** An assessment instance is also an educative instance: a student's misunderstanding can be detected and corrected during the assessment process.

- **Teaching or assessment correction.** The presence of an error, specially if recurrent, may indicate a flaw in the assessment or the teaching process.

A first limit emerges on automatic assessment. It would not be possible to assess automatically:

- the use of language and expressiveness of an answer, in the general case. In Engineering, where technical and factual matters demand a uniform, strict style, this is not as severe a limitation as in subjects where a more narrative style is required or possible.

- literary, philosophical, religious or creative areas where originality in style and expression may be close to the essence of the subject. Again, this is not a limitation for Engineering.

- well conceived but wrongly expressed ideas. Engineering is as prone to this type of error as any other subject.

- automatic or assisted assessment would be generally limited to the factual aspects of a subject, in the best case to opinions or pieces of criticism of wide acceptance, already present in the assessment reference material. Not a limitation for Engineering.

Most teachers would require several facilities from an automatic or assisted assessment system:
• see by themselves uncommon answers, were they original in their concepts or simply wrong.
• be able to change a mark given automatically, to reflect the teacher’s valuation of the answer.
• be able to change on the fly the marking criteria on coming to a yet unregistered error or a possibly correct answer not included in the model answer, without be compelled to backward revise all the marking.

These reasonable, desirable features rule out, in education, a purely automatic assessment system, at least in the present state of the art. A CAA system would carry most of the marking burden, automatically marking or suggesting marks for the most common situations, but keeping the active role of the teacher, drawing his/her attention to the infrequent answers, were they erroneous or original. The system would learn from the teacher's judgment by including a new type of error or a new correct or partially correct answer, modifying the model answer accordingly [5].

THE CHALLENGE OF OPEN QUESTIONS AUTOMATIC ASSESSMENT

The automatic assessment of open questions as such, with no limitations, is an intractable problem in the present state of the art. Natural language is inherently ambiguous, both in word meaning and discourse; the same ideas can be expressed with very different words and syntax; cross reference among sentences may be difficult to infer; "knowledge of the world" is frequently required to understand the question and provide an answer. In the last decades, Natural Language Processing (NLP), a field of Artificial Intelligence (AI), has achieved outstanding results by the use of Statistics in the analysis of large collections of annotated texts (corpus) [6].

Considering open questions addressing summative assessment, where emphasis is on content and not on style or other formal properties, the automatic marking of open questions to the level of accuracy required in education is recognized as an AI-complete problem [4]. Such problems are intractable with present day means and knowledge.

Even trivial free text answers may prove very difficult to mark. A question on probability where the correct answer could be expressed as 0.18, 18% or the calculation leading to the result, produced 117 different answers in 144, when answers were treated as text strings [4].

Not all questions are so demanding, but other problems persist. Mispellings of correct answers is very common, and can be dealt with accepting words not far from a certain "edit distance". Edit distance can be defined as the minimum number of character substitutions required to convert a word into another. Accepting words at an edit distance of 2 reduces considerably the number of answers which must be seen. In short words, however, even an edit distance of 2 may be too much [5] [4].

A more difficult problem to deal with is that of context-dependent synonyms, the collection of words that can be accepted as reasonable substitutes of the ones mostly accepted within a particular field of knowledge. Context synonyms are very frequent; a number of them may turn up even in a simple question. An attempt to predict them may prove not worth the effort; in the end context synonyms must be dealt with when they turn up in the answers, by human intervention [4].

TYPES OF ASSESSMENT AND QUESTIONS

Assessment may be summative or formative. Summative assessment attempts to verify if the intended educative goals have been achieved, typically in the form of an exam or similar instance. This type of assessment has also diagnostic value, since it helps detect deficiencies in the learning or the teaching. Formative assessment aims at producing achievements on the part of the students; it is a bidirectional process between the student and the teacher to enlarge, recognize and respond to learning [7] [8].

Summative assessment may be attempted by means of CAA. Formative assessment is a complex, interactive process; the computer can only provide a modest support [6].

Questions can be classified in two dimensions: cognitive and formal.

In the cognitive dimension three categories are recognized, each one associated to a different form of comprehension: literal, interpreted and critical [9] [6]:

• in literal questions meaning is directly conveyed in the words and expression.
• interpreted questions require to elaborate on the words to get the meaning, which must be inferred from the text.
• critical questions demand careful judgment or judicious evaluation on the part of the student, sometimes according to his own notions on the subject.

Only literal questions can be put to automatic marking with present day means. Literal comprehension questions deal with definitions and causal relations in texts [9]. Most questions related to Engineering subjects fall into this category.

In the formal dimension questions can be classified as text explicit, text implicit and “script implicit”:

• in text explicit questions the answer is right in the texts available to the students.
• in text implicit questions the student is required to search for information and make links and inferences across the text.
• in script implicit questions the student is required to draw on his own knowledge, besides any text, to build an answer. This last category requires the student to perform an inference involving his/her "knowledge of the world"; these questions are very difficult to mark automatically [9].

Knowledge of Engineering subjects can be assessed to a great extent with text explicit or text implicit questions. These types of questions can be treated by recovering, at
least partially, the necessary fragments from one or several sentences in the answer text, and comparing to a reference answer, a reference text or a knowledge structure.

Considering only literal text explicit and literal text implicit questions automatic evaluation of open questions becomes possible [9] [6].

THE ASSESSMENT PROCESS

This section provides a short revision of the assessment process considering the possibilities and limitations of CAA.

Preparation.

• question types. Knowledge of the type of question and its possible answers simplifies its preparation and enables for automatic marking.
• question validation. Tests should be proposed to other teachers or assistants to ensure adequacy of the questions and their expected answers.

Application.

• the assessment instance. Students answer the questions on their computers, in the school premises or remotely.
• feedback to student. CAA does not allow for immediate feedback to the student; a teacher must go through the answers, with the help of the system. For immediate feedback closed answers offer a better alternative.

Marking.

• from reference texts. There is a variety of ways in which a collection of reference texts can be used in automatic marking. Atypical answers are presented to the teacher, who marks them indicating correct or incorrect concepts and relationships. The CAA system "learns" from the marked answer. As the marking goes on less and less atypical answers are found, as the system registers more of the possible variants.
• by dynamic construction of the model answer. The answers are presented to the teacher, who marks correct and incorrect concepts and relations. The system registers one or both. On finding similar situations the system marks new answers according to previously registered correct or incorrect answers.
• marking revision. Exam results very near to the minimum score required for approval should be revised by the teacher. Ideally there should be very few, since the types of questions which can be automatically marked are little prone to differences in criteria.

AUTOMATIC ASSESSMENT SYSTEMS PROPOSALS

Most open questions automatic assessment systems rely on some kind of reference material against which students' answers are compared. Reference material may be classified in different, sometimes overlapping, categories.

Training set. a number of correct answers or answers manually marked by teachers.

Session F2B

• Project Essay Grade (PEG), a pioneer work of the '60s, obtains several values considering formal aspects of the answer. A correspondence between formal aspects and the quality of the answer is assumed [10].
• Educational Testing Service I (ETS I). Looks for some formal properties in fragments of 15 to 20 words. In a test instance 200 texts out of 378 were used to build the reference model [10].
• ETS II, a later version, led to E-rater, a system based on Statistics and NLP. New answers are compared to a set of training answers using two different similarity measures [10] [3].
• C-rater reduces the number of training texts focusing on specific information that must be present in a correct answer [3].
• Paperless School free-text Marking Engine (PS-ME) makes use of NLP starting from 30 texts manually marked by teachers; both right and wrong answers are included [3].
• Atenea uses NLP and statistics to compare a small set of answers written by several teachers. Marks short answer questions using the ERB module (Evaluating Responses with BLEU) and NLP [11].
• Pullman and Sukkarieh [12] cautiously limit their work to short answers ranging from a few words to five lines, typical of factual sciences. They started using knowledge extraction by pattern search, but question preparation proved to be hard work, leading them to try machine learning techniques on a set of training texts.

Set of features: features defined by the teachers, expressed in text or some structured form, which must be present in correct answers.

• Lackey System, uses text categorization techniques [3].
• Automated Text Marker (ATM) searches for previously identified concepts and dependencies by means of information extraction [3].
• Auto Mark uses NLP to analyze and compare text against predefined templates filled in by teachers. Schema Extract Analyze and Report (SEAR) and Auto-marking also use as reference material templates generated by teachers [3].

Text or set of texts: a model answer or a text on the subject or a set of texts representative of a language or subject area (corpus). This reference material may exist by itself, it is not necessarily prepared by the teachers.

• Latent Semantic Analysis (LSA) is based on a representation of the contextual use of words. The reference text or "model answer" may be a textbook or an answer written by a teacher [10]. Intelligent Essay Assessor (IEA) obtains an LSA model from training texts [3]. LSA was tested with a set of 1929 students' answers and 142.580 texts in English on general subjects [11].

• The original ERB (Evaluating Responses with BLEU) proposal used the BLEU (Bi Lingual Evaluation
Understudy) algorithm to mark answers against a set of reference texts. [13].

Comparison against a semantic structure representation of the correct answer.  
- Huntington and Hunt employ automatic translation techniques to draw from each answer a semantic representation which is then compared to another semantic representation obtained from a model answer [10].  
- Intelligent Assessment Technologies allows to mark short answers against a template created by means of a graphics tool [14].

Criticism

Most automatic marking systems show high scores of correctness when compared to manual marking. This is not surprising if training texts or teachers criteria show a relatively high degree of divergence, which is usually the case. Moreover, some comparisons are based on form, style or other highly subjective factors [4].

Pulman and Sukkarieh's very careful work on an automatic marking system developed at Oxford University and tested at Cambridge University attains 90% coincidence with manual marking of short textual answers of factual content (1-mark answers) when teachers agree on the answer. However, they honestly report some unexplained discrepancies on some particular questions. On more complex answers (2-mark answers), the system made more mistakes and results differed widely against human markers [15].

Several weak points have been detected in purely automatic marking systems: the need for training material, the difficulty of marking even trivial questions, the handling of orthographic or expression errors, the difficult problem of interpreting context dependent synonyms, the lack of certainty in statistical methods, among others [4] [9].

The Case for Computer Assisted Assessment

Automatic marking of open text answers has shown outstanding achievements by processing statistically large collections of text. Is is not surprising to find coincidence with manual marking as high as 95%. Besides possible objections to the validity of the comparison because of the potential diversity of teachers' answers or marking, no student will be satisfied to get a mark which is "95% correct": it would mean to accept a student among twenty has been wrongly marked. Perhaps human marking is not better than that, but the student will more easily accept human judgment than a machine mark, or at least is more in the habit of doing so. Perhaps in personnel selection or other areas this error might not be of major consequence, but in Education every effort must be done to avoid it, offering the student as much warranties as possible, specially if systems are known to be not fully reliable.

CAA or a variant called HCC (Human Computer Collaborative) assessment, have been proposed as alternatives to purely automatic systems [5]. CAA deliberately calls for human intervention or "moderation" in the marking process [14]. Human intervention makes the whole process more reliable and versatile, offers increased warranties and is closer to educators' expectations, at the same time lightening the burden of routine work without losing too much contact with the students.

CAA allows to go without training material, enabling the dynamic creation of a knowledge representation or model answer during the marking process, showing the teacher several answers to the same question for comparison and judgment [5], or extracting from selected texts carefully worded text implicit or text explicit questions [6]. Sound knowledge of the type of question proposed enables and simplifies the handling of the answer [6].

Requirements for an ideal system

A computer assisted software for open questions assessment would, ideally, comply with the following requirements:

- preferably not require the generation of reference material. Hard work will almost certainly be required to achieve the qualities of accuracy, determinism and completeness so as to be reliable as correct answer references.
- preferably not be based on manually marked answers. Many teachers do no like to propose the same questions in different exams, the subject changes, the course contents change, questions become obsolete, new textbooks differ from former ones, students interchange information among them.
- be friendly, easy to use, when devising questions and model answers, perhaps through graphics of concepts and relations. An example of such a system, based on templates, is given by Mitchell et al. [14].
- transparent handling of lemmatization, syntax and other language aspects by means of adequate NLP techniques, in an effort to recognize similar semantic content besides differences in language expression.
- effective handling of context synonyms within the subject area, were it by the use of specific dictionaries or the ability to create them, preferably dynamically during the marking process [14].
- show the teacher all atypical answers for his direct assessment, allowing him/her to eventually include them in the model answer template or structure; do not show answers conveying the same knowledge and differing only in their language expression.
- show any answer at the teacher's request. Many teachers would like to see several answers at the same time for the sake of comparison [4].
- automatic backward marking when including a new right or wrong answer in the model template. This feature is unavoidable if the model answer is to be dynamically built during the marking process.

Many other desirable features come to mind. Only the most significant have been included, attending to the quality...
of the assessment, ease of use, clearness in presentation, teachers' acceptance. Even within the perhaps limited set of features outlined it may be very difficult to fully accomplish such a wish list.

A CAA system for open question assessment based on answers written in a controlled language is being developed by the authors. Key components of the proposed system are a controlled language of general use or specific for an area, an intelligent editor friendly enough to allow easy answer writing using the controlled language, a knowledge extraction routine and the implementation of a distance measurement to compare knowledge structures. The strength of the proposal lies on the fact that controlled language text can be successfully treated by NLP techniques, since the ambiguities and uncertainties inherent to natural language text have been resolved at the moment of writing the answer. The editor puts forward and guides the student to the writing of a simple, direct, unambiguous answer, from which a knowledge representation structure can be obtained. Knowledge and techniques developed around the semantic web provide a variety of tools and structures both for the comparison of such structures as well as to dynamically adapt them to new answers.

CONCLUSIONS

Engineering Education can profit from open questions assessment to get insight into the conceptual maturity of students' knowledge and their communications abilities. Educators' expectations make CAA preferable to the purely automatic assessment of open questions. CAA of open questions becomes possible if question types are limited to those apt for treatment, all of these of interest for Engineering. An ideal system must not depend on training material, must be reliable and easy to use, must allow for teacher intervention, adapt to new answers or criteria, backward mark already seen answers. Controlled language allows for unambiguous, NLP treatable answers; the semantic web provides tools for knowledge extraction and comparison; techniques adopted or adapted from both these fields provide a promising way for a CAA system of open questions automatic assessment.

ACKNOWLEDGMENTS

This work was carried on with the help of the Coordinating Action 508AC0341 SOLITE - Software Libre en Teleformación, CYTED Program.

REFERENCES

[11] Pérez, Diana; Gliozzo, Alfonso; Strapparava, Carlo; Alfonseca, Enrique; Rodríguez, Pilar; Magnini, Bernardo. Automatic Assessment of Students' free-text Answers underpinned by the Combination of a BLEU-inspired algorithm and Latent Semantic Analysis. American Association for Artificial Intelligence

AUTHOR INFORMATION

Victor Gonzalez-Barbome, Assistant Professor, Universidad de la Republica, Uruguay, vagonbar@fing.edu.uy.

Martin Llamas-Nistal, Associate Professor, Universidad de Vigo, Spain, Martin.Llamas@det.uvigo.es