http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (6): 1018-1022, 2006
ISSN 1812-5638

Online Testing and Assessment of Textual Responses using NOVEL

[1]P. Selvi and [2]N.P. Gopalan
[1]Department of Computer Science and Engineering,
[2]Department of Compter Applications,
National Institute of Technology,
Tiruchirapalli, India

Abstract: Testing and assessment remain an integral part of instructional systems design for traditional classroom based courses as well as online training courses. The goal of testing is to determine if learning objectives have been accomplished. Formative evaluation using online testing helps students assess their level of knowledge of course material. In addition it gives the instructor a better idea of what students understand as well as the concepts that still need clarification. This study proposed NOVEL algorithm, a method to assess the short essays written by students. It is a statistical method for inferring meaning from a text. We conclude that, although it is only applicable to a restricted category of questions, NOVEL attains better results than other keyword-based procedures. Its simplicity and language-independence makes it is a good candidate to be combined with other well-studied computer assessment scoring procedures. This study explains how NOVEL works and explains how NOVEL is particularly suited to web-assessment.

Key words: Online assessment, web assessment, e-learning, automatic assessment, electronic marking, computerized evaluation

INTRODUCTION

Online assessment is the type of computer-delivered assessment. It allows computer delivered tests to be provided at remote locations, using either the internet or company intranets. Tests held on central computer can be delivered to the candidate online; the candidate answers the questions and the responses are sent to the central web server.

Online assessment, in relation to e-Learning, is considered to be the weak link within the concept as Internet tests are seen as still lacking performance assessments that support learning. The primary difficulties with assessment through e-learning are; delivery and communication; the lack of proximity and body language for feedback; the lack of instructor perception and control over the learning environment; difficulty of authentication and privacy and the lack of informal interaction (Garrison and Anderson, 2003). Due to this nature of the online classroom, certain aspects of curriculum are difficult to assess.

With the development of online assessment has come the shift from testing to assessment. Assessment focuses on the integration of instruction, learning and assessment (Jochems *et al*., 2004). The term performance learning has evolved, referring to the level perceived as worthwhile and relevant to the learner and gives the learner the ability to use acquired skills and knowledge together (Jochems *et al*., 2004).

Assessment is directly linked to effective teaching and learning as it typically displays understanding and achievement. It integrates assessment and instruction and should ideally diagnose any misconceptions or challenges the student is facing. Assessment in an online environment should function to; communicate achievement status for students; provide self-evaluation to the learner, identify student placement for education paths or programs; motivate the learner and evaluate the effectiveness of instruction programs (Garrison and Anderson, 2003). Assessment should occur throughout the entire online course and provide the learner with frequent and comprehensive formative feedback, an important aspect in shaping fundamental learning. It may be production focused, behaviour focused or extended written response to ensure that the learner applies critical thinking skills in the process (Jochems *et al*., 2004).

With a move towards delivering more teaching online, online assessment is a logical step. Online assessment also allows integration with existing web-based materials such as audio/video. This allows for examinations using this material.

Corresponding Author: P. Selvi, Department of Computer Science and Engineering, National Institute of Technology, Tiruchirapalli, India Tel: 044-22781196

Peer assessment activities enable students to criticise or comment on content related criteria in relation to other students' work. This means of assessing supports the development of continuous self reflection of both behaviour and learning and decreases teacher workload, ultimately promoting social interaction, individual accountability and positive interdependency (Jochems *et al*., 2004).

The interest in the development and in use of Computer-based Assessment Systems (CbAS) has grown exponentially in the last few years, due to the increasing in the number of students attending universities and the possibilities provided by e-learning approaches to asynchronous and ubiquitous education. CAA of free-text answers is a long-standing problem that has attracted interest from the research community since the sixties (Page, 1966) and has not been fully solved yet. On the other hand, the success of e-learning and the advances in other areas such as Information Extraction (IE) and Natural Language Processing (NLP) have made CAA of free-text answers a flourishing research line in the last few years. A computer can examine and analyze answers in much more detail than a teacher, as it is totally free from myths, false beliefs and value biases (Streeter *et al*., 2003). In the literature, several techniques have been used to tackle this problem providing increasingly better results. They can be grouped into five main categories: statistical, NLP, IE, clustering and integrated-approaches (Valenti *et al*., 2003).

This study presents an application of the NOVEL algorithm (Selvi and Gopalan, 2006) for the assessment of students' textual responses in online assessment. We argue that this procedure, although it does not attain a correlation high enough to be used as a stand-alone assessment tool, improves other existing keyword-based procedures and it is a good candidate for replacing them in existing applications. It keeps all their advantages (language independence and simplicity) and produces better results.

THE NOVEL (SYNTACTICALLY ENHANCED BLEU) METHOD

The NOVEL method (Sclvi and Gopalan, 2006) was proposed to assess the short essays written by students. Its robustness stems from the fact that it works with several reference texts (human-made answer key), against which it compares the candidate text. The pseudocode has given in the following box.

1. For the text values of 1 to N (typically from 1 to 4),

Calculate the percentage of n-grams
based on the primitive and composite fearures.

$$\%\text{of n-grams } (P_n) = \frac{\sum_i \left\{\begin{array}{l}\text{the number of n-grams in segment i ,in the candidate text ,}\\ \text{with a matching reference coocurence in segment i}\end{array}\right\}}{\sum_i \left\{\begin{array}{l}\text{the number of n-grams in segment i ,}\\ \text{in the candidate text being evaluated}\end{array}\right\}}$$

2. Combine the makes obtained for each value of N, as a weighted average.
3. Apply a brevity factor to penalize the short candidate texts.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e(1-r/c) & \text{if } c \le r \end{cases}$$

Where:
c = the total number of words in the candidate text.
R = the total number of words in the reference text.

$$NOVEL = BP.\exp\left(\sum_{n=1}^{N} w_n \log(p_n)\right)$$

$$Score = \exp\left\{\sum_{n=1}^{N} W_n \log(P_n) - \max\left[\frac{L^*_{ref}}{L_{sys}} - 1, 0\right]\right\}$$

Where:
$w_n = N^{-1}$
N - 1
L^*_{ref} = the number of words in the reference text that is closest in length to the candidate text being scords.

L_{sys} = the number of words in the candidate text being scored.

BLEU (Papineni *et al*., 2002) is one of the methods for automatic evaluation of translation quality. It uses the ratio of co-occurring n-grams between a translation and single or multiple reference sentences. But this algorithm attempts to go beyond the simple n-gram matching to account for the lexical and syntactical variations between the system and the reference outputs.

It can be seen from this pseudocode that NOVEL is not only a keyword matching method between pairs of texts. It takes into account several other factors that make it more robust:

- This method computes the semantic similarity between Candidate text and expert text based on the primitive and composite features. The features that compare single terms from two sentences, called 'primitive' features and those that match word pairs to word pairs, called 'composite' features.
- Our approaches consider a number of linguistic approaches to text analysis and are based on both single words and simplex noun phrases. Each of these morphological, syntactic and semantic features has several variations. Thus, we consider the following possible matches between text units:

- Noun phrases (NP) Matching
- WordNet synsets (WnSyn)
- Semantic Verb similarity
- Matching of Proper nouns, names, organizations.
- Ordering
- Distance

- It calculates the length of the text in comparison with the lengths of reference texts. This is because the candidate text should be similar to the reference texts (if the translation has been well done). Therefore, the fact that the candidate text is shorter than the reference texts is considered an indicative of a poor quality translation and thus, BLEU penalizes it with a Brevity Penalty factor that lowers the score.
- The measure of similarity can be considered as a precision value that calculates how many of the n-grams from the candidate appear in the reference texts. This value has been modified, as the number of occurrences of an n-gram in the candidate text is clipped at the maximum number of occurrences it has in the reference texts. Therefore, an n-gram that is repeated very often in the candidate text will not increment the score if it only appears a few times in the references.

The final score is the result of the weighted sum of the logarithms of the different values of the precision, for n varying from 1 to 4. It is not interesting to try higher values of n since coincidences longer than four-grams are very unusual.

NOVEL's output is always a number between 0 and 1. This value indicates how similar the candidate and reference texts are. In fact, the closer the value is to 1, the more similar they are Papineni *et al.* (2001) report a correlation above 96% when comparing NOVEL's scores with the human-made scores. This algorithm has also been applied to evaluate assessment of open-ended questions (Selvi and Gopalan, 2006).

NOVEL IN ON-LINE ASSESSMENT

Overview: Online assessment is a form of computer based assessment where the assessment is delivered across the internet or a local intranet. The student can take the assessment itself through an internet browser.

The key features of online assessment system's are follows:

- The creation of question to be delivered across the internet. The range of question types that can be created vary from one teacher to another teacher. So, the questions are designed with little or no knowledge of HTML. i.e. using the web browser as on interface.
- The storage of questions in a database. This will allow the question designer to build up a question bank.
- The delivery of questions to the students across the Internet or intranet.
- The Online Assessment System uses TCP/IP protocol for all of the communication between the application and the customer.
- Some form of feedback and or scoring to be delivered back to the student either during or after the assessment
- NOVEL algorithm is used for assess the student textual responses.

In the case of assessment of student answers, we can consider that the students' responses are the candidate text and the teacher can write a set of correct answers (with a different word choice) to be taken as references (Weiner, 2004).

EXPERIMENTATION

For evaluation purposes, we have built seven different benchmark data from real exams. The seven sets, which include more than 1000 answers altogether, are described in Table 1. Two different human judges, who also wrote the reference texts, scored all the answers manually. The instructions they received were to score each answer in a scale between 0 and a maximum score (e.g., 1 or 10) and to write two or three reference answers for each question. We are currently transcribing other three sets corresponding to three more questions, but we still have only a few answers for each.

The evaluation datasets are given in the Table 1. The Columns indicate the set number; number of candidate texts (NC), their mean length (MC), number of reference texts (NR), their mean length (MR), question type (Def = definitions; A/D = advantages/disadvantages; Y/N = justified Yes/No) and a short description (OS = Operating System exam question; OOP = Object-Oriented Programming exam question).

Table 1: Evaluation datasets

SET	NC	MC	NR	MR	TYPE	Desc
1	38	67	4	130	Def.	OS
2	79	51	3	42	Def.	OOP
3	96	44	4	30	Def.	OS
4	11	81	4	64	Def.	OS
5	143	48	7	27	A/D	OS
6	295	56	8	55	A/D	OS
7	117	127	5	71	Y/N	OS
8	117	166	3	186	A/D	OS
9	14	118	3	108	Y/N	OS
10	14	116	3	105	Def.	OS

We have classified the ten questions in three distinct categories:

- Definitions and descriptions, e.g., What is an operative system, Describe how to encapsulate a class in C++.
- Advantages and disadvantages, e.g., Enumerate the advantages and disadvantages of the token ring algorithm.
- Yes/No question, e.g., Is RPC appropriate for a chat server? (Justify your answer).

All the answers were marked manually by at least two teachers, allowing for intermediate scores if the answer was only partially correct. For instance, if the maximum score for a given question is defined as 1.5, then teachers may mark it with intermediate values, such as 0, 0.25, 0.5, 0.6, etc.

The discourse structure of the answer is different for each of these kinds. Definitions (and small descriptions) are the simplest ones. In the case of enumerations of advantages and disadvantages of something, students can structure the answer in many ways and an N gram-based procedure is not expected to identify mistakes such as citing something which is an advantage as a disadvantage.

Experiment performed: The algorithm has been evaluated, for each of the data sets, by comparing the N-grams from the student answers against the references and obtaining the NOVEL score for each candidate. The correlation value between the automatic scores and the judges' scores has been taken as the indicator of the goodness of this procedure.

We have varied the following parameters of NOVEL:

- The number of reference texts used in the evaluation process.
- The length (N) of the maximum N-gram to look for coincidences in the reference texts.
- The brevity penalty factor to penalize short answers.

RESULTS AND DISCUSSION

Number of reference texts: As said before, NOVEL is very sensitive to the number and the quality of the reference texts available (Selvi and Gopalan, 2006). In our experiment, we have varied the number of references from 1 up to the maximum number available for each question (Table 1, column NR). Table 2 shows the results for each of the data sets. As can be seen, in general, the results improved with the number of references, although in some cases the addition of a new reference having words in common with wrong answers decreased the accuracy of the marks.

Table 2: Scores of BLEU for a varying number of reference texts

Data sets	One	Two	Three	Four	Five	Six	Seven
1	0.42	0.62	0.73				
2	0.31	0.36	0.38	0.41			
3	0.39	0.41	0.32	0.31			
4	0.40	0.42	0.43	0.51	0.53		
5	0.35	0.36	0.41	0.39	0.42	0.46	
6	0.06	0.08	0.07	0.05	0.04	0.06	0.09
7	0.13	0.23	0.33	0.41			
8	0.42	0.44	0.51				
9	0.71	0.81	0.72				
10	0.72	0.80	0.85				

Table 3: Scores of BLUE for a different choice of the type of N-grams chosen in the comparison from only unigrams (in the first column) to N-grams with lengths from 1 to 4 (in the last column)

Data sets	Uni-gram	Bi-gram	Tri-gram	Four-gram
1	0.56	0.57	0.54	0.50
2	0.48	0.47	0.49	0.46
3	0.24	0.25	0.27	0.24
4	0.59	0.73	0.62	0.55
5	0.40	0.45	0.52	0.41
6	0.08	0.06	0.05	0.03
7	0.25	0.32	0.28	0.22
8	0.28	0.26	0.25	0.23
9	0.81	0.72	0.65	0.60
10	0.93	0.91	0.85	0.82

Results for different N-grams: We have varied the maximum size of the N-grams taken into consideration from 1 to 4. As Table 3 shows, in most cases, the addition of bigrams and trigrams improves the resulting correlation, which indicates that collocations are useful in the evaluation. Given that the students' answers are completely unrestricted, in most of the cases they do not contain enough sequences of four consecutive words from the references and therefore using four-grams affects the results negatively.

Brevity penalty: The NOVEL procedure is basically a precision score, as it calculates the percentage of N-grams from the candidate answer which appears in any reference, but it does not check the recall of the answer. Therefore, it applies a brevity penalty so as to penalize very short answers which do not convey the complete information. In its original form (Papineni *et al*., 2001), BLEU compares the length of the candidate answer against the length of the reference with the most similar length. We have also tested the following brevity penalty:

- For each reference text, calculate the percentage of its words covered in the candidate.
- BP = Add up all the percentages.
- Multiply the basic NOVEL score and BP.

Comparison with other methods: We have implemented three other scoring algorithms as baseline:

Key words: Consisting in calculating the proportion of words which appear in any of the reference texts.

Table 4: Comparison of NOVEL with three other methods. Because of the kind of evaluation performed, those with very few answers couldn't be evaluated with VSM

Data sets	NOVEL	BLEU	Key words	VSM
1	0.73	0.58	0.07	0.31
2	0.41	0.36	0.23	0.09
3	0.41	0.36	0.19	0.24
4	0.53	0.82	0.57	-----
5	0.46	0.41	0.57	0.52
6	0.09	0.02	0.05	0.05
7	0.41	0.21	0.32	0.17
8	0.51	0.41	0.22	0.17
9	0.81	0.73	0.24	----
10	0.85	0.75	0.09	----

VSM: Using a vectorial representation of the answers. In this case, we cannot implement it with reference texts, but by calculating similarities between answers. We have done a five-fold cross-evaluation, in which 20% of the candidate texts are taken as training set for calculating tf.idf weights for each term. The rest of the answers are assigned the score of the text in the training set which is most similar to it. The results obtained are listed in Table 4. The improvement using NOVEL is significant with 0.95 confidence.

CONCLUSIONS

We have described here an application of the NOVEL algorithm for online evaluation of student answers with a shallow procedure.

The main advantages of this approach are that

- It is very simple to program (just a few hours).
- It is language-independent, as the only processing done to the text is tokenization.
- It can be integrated with other techniques and resources, such as thesauri, deep parsers etc., or it could be used in substitution of other keyword-based procedures in more complex systems.

On the other hand, as has sometimes been noted, NOVEL is very dependent on the choice of the reference texts, so that leaves a high responsibility for the professors, who have to write them. Secondly, it is not suitable for all kinds of questions, such as those where the order of the sentences is important or, as we have seen, for an enumeration of advantages and disadvantages. Further processing would be necessary for scoring these questions.

We observe that, for evaluating student answers, a new brevity penalty that measures directly the recall of the student answer improves the results and makes it less necessary to use higher-order N-grams.

As can be seen, NOVEL clearly outperforms other keyword-based algorithms. Although it is not directly comparable to VSM, given that the evaluation procedure is different, the results hint that it has produced a better correlation.

Therefore, we conclude that the current version of the system could be effectively used in web environment as a help to teachers who want to double check the scores they are giving, as well as for students who want to solve more practicals than the ones they receive in the classroom. Nevertheless, we do not recommend the use of this version as a stand-alone application.

The following are some ideas for future work:

- Automatise the production of the reference texts.
- Perform the evaluation against yet more existing systems.
- Explore how to extend the procedure with other linguistic processing modules.

REFERENCES

Garrison, D. and T. Anderson, 2003. E-Learning in the 21st Century. London: Routledge Falmer.

Jochems, W., J. Van Merrienboer and R. Koper, 2004. Integrated E-Learning: Implications for Pedagogy, Technology and Organisation. London: Routledge Falmer.

Page, E., 1966. The imminence of grading essays by computer. Phi Delta Kappan.

Papineni, K., S. Roukos, T. Ward and W. Zhu, 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

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

Valenti, S., F. Neri and A. Cucchiarelli, 2003. An overview of current research on automated essay grading. J. Inform. Technol. Edu., 2: 319-330.

Weiner, J.A., 2004. Web-based assessment: Issues and applications in personnel selection. June 22, 2004 IPMAAC 28th Annual Conference on Personnel Assessment.