Comparative Study of Techniques used for Automatic Evaluation of Free Text Answer

Ms. Paden Rinchen
Sikkim Manipal Institiute of Technology
East Sikkim, India
whoserpaden@gmail.com

Abstract **–**Evaluation in education allows for obtaining, organizing, and presenting information about how much and how well the student is learning. Assessment of free text answers have become a necessity , not only for today's education system but also to erase the load and errors of manual correction. This paper basically presents a comparative study of some of the techniques used to achieve the goal , the limitation of the respective methods.

Keywords: *E-learning, Free text, automatic evaluation, techniques.*

*****

I. INTRODUCTION

Evaluation of subjective or text based answers has been a hurdle in the development of e-learning systems since a very long time. Attempts using classical Natural Language Processing and/or Adaptive techniques have met with only limited success. If developed, the solution will not only redefine e-learning but open up newer avenues in Human Computer Interaction as well.

The problem related to evaluation of subjective answers is that each student has his/her own way of answering and it is difficult to determine the degree of correctness [1]. The assessment of the correctness of an answer involves the evaluation of grammar and knowledge woven using the conceived interpretation and creativity of a human mind. Human Evaluation, though slow and carrying drawbacks of human fatigue and bias is the only accepted method for evaluation of text based answers, as the intelligence of one human can be fathomed by another. However, with the development of communication and internet technologies, the reach and nature of education has changed with its spread across geographical, social and political boundaries with an exponential growth of intake volume. This has made the drawbacks of human evolution come out more glaring than ever before and interfere with the importance of automatic evaluation is being felt by all concerned. The automatic evaluation of answers would not only assist teachers in reducing their workloads but also help students as they can identify their mistakes through the feedback provided by the system after determining their strong and weak areas [2].

The only disadvantage of using computer as evaluation tool is the possibility of failure [3] at the time of submission of answers and lack of technical knowledge provided to the teachers and students

Attempts at developing automatic evaluation systems have been made by researchers at various points using different techniques, which have found some acceptance. Some of which are discussed below along with its pros and cons.

II. TECHNIQUES USED

A. Latent Semantic Analysis(LSA) is a mathematical technique[4] proposed by Laundauer, Foltz and Laham in 1998. It supports semantic similarity measurement between texts. This technique is capable of extracting hidden meanings in a text by applying decomposition method. It is also known as bag-of-words approach[5]. In this technique whenever a text is provided by the user, certain words are extracted. Stop words like 'a', 'and', ;the', ;is', etc are not included. Once the words are extracted from the input text, a matrix is created. The matrix stores the frequency of occurrence of each word against the sentences. Each element of the matrix is then converted to its log and its entropy value is calculated. A Singular Value Decomposition of the resulting matrix returns the cosine measure of similarity of two documents under review. This method takes one document as the standard and subsequently compares the similarity with the other document.

1, **Auto Tutor** (Graesser et al. , 1999)[6]was developed at Tutoring Research Group of University of Memphis. It is a fully automated computer tutor that assists students in learning subjects like hardware, operating systems and internet. It presents question and answers from a curriculum, tries to assess the learner's answer entered through the keyboard and formulates dialog moves that are sensitive to

the learners contributions. LSA forms a major mechanism that evaluates the quality of students answer. It was found that the performance of LSA in terms of evaluating answers was equivalent to an intermediate expert human evaluator.

2. **Apex** (Dessus et al. , 2000)[7] is a web based learning environment which rates the learners response in free text with reference to answers already stored in the system database. Once connected to the system, a student selects the topic or question with which his or her response is compared to the stored answer. The semantic similarity value is measured using LSA.

3. **Intelligent Essay Assessor** [8] developed by Landauer, Foltz and Laham is a web based application [6]. It is basically used for essay based evaluation. IEA provides the essays based on a particular topic which serves as a reference for evaluation purpose [8]. The essays of the student are then compared with the reference essays and a feedback is provided based on content, mechanics and style [9]. For evaluating the content based module, it uses LSA to determine the semantic similarity between the essays. Mechanics module is responsible for spelling check and grammatical errors. As the feedback is instant it provides facility to correct and resubmit their essays, thereby helping students to improve their writing and thinking skills.

4. SELSA(Syntactically Enhanced LSA) It is a modification proposed by Kanejiya, Kumar and Prasad(2003)[10]. It considers a word along with its context by taking it along with its adjacent words as a unit of knowledge representation. It overcomes the shortcomings of LSA as it considers word order, which is however limited to adjacent words only. The identified corpus is POS tagged and the matrix similar to LSA is populated. The difference lies in the rows of the matrix which consist of word-prevtag pairs in place of words only as in LSA.

B. BiLingual Evaluation Understudy(BLEU Algorithm)proposed by Papineni et al. [11] is an n-gram co-occurrence scoring procedure. The main idea of BLEU is to measure the translation closeness between a candidate translation and a set of reference translations with a numerical metric. n-implies a sequence of words which are used to perform comparision of two different texts. In this method the input sentence provided is translated by machine and then n-gram matches between the machine translation and reference translation is counted.

The user provides an input text in one language which is translated by the machine in another language. The BLEU algorithm performs the comparision of the input translation with the reference translation. Each word is then compared for a match in the reference translation that is available in the system and is stored as count. Once a word of any reference translation matches with the input translation it cannot be taken into consideration again. The count for each word of the input translation against the reference translation is performed and the maximum count of each word is taken into consideration. The occurrence of each input word against the input translation is computed and then it is compared with the maximum count of each word computed earlierThe comparision leads to the selection of minimum count value for each word. These minimum count values are then added and divided by the number of words present in the candidate translation. The value obtained is called the modified unigram precision(MUP)[12]. The MUP value for each of the candidate translation is computed where the number of words n used for determining MUP may range from 1to 4.

1. ERB (Evaluating Responses with BLEU)Perez, Alfonseca & Rodriguez 2004)[13**]** is a BLEU inspired algorithm with a set of NLP techniques. The core idea of this algorithm is the same as the BLEU, that more similar a student's answer(input answer)is to the teachers answer(reference answer), the better it is and consequently it will have a higher score.

BLEU uses MUP metric that clips the frequency of the n-gram. MUP is calculated for each value of n, which ranges from 1 to 4. A weighted sum of the logs of MUP's is performed. Lastly a penalization is applied to very shot answers, which might be incomplete, by multiplying the previous value by Brevity Penalty(BP)factor. The original algorithm is modified so that it takes into account not only the precision(the original BLEU score)but also the recall that is calculated by studying the percentage of the referernces that is covered by the student's answer using a modified Brevity Penalty(MBP)factor.

Equation 1 shows the final formula for calculating the score of an answer *a*. *n* represents the length of the *n*-grams and N is the highest value that *n* can take.

$$ERB_{score}(a)=MBP(a) * e \sum_{n=0}^{N} \frac{LOG(mup(n))}{N} \quad \text{(Eq. 1)}$$

2. ATENEA Perez et al. [14] developed Atenea for the assessment of subjective answers. The system is based on the BLEU algorithm and is capable of evaluating answers in English as well as Spanish. It is not only capable of evaluating the answers provided by students but also allows them to personalize the user interface as according to their requirements [14]. Evaluation is thus possible irrespective of the language the student wishes to answer the question.

Atenea allows the user to choose a question for which the student is at liberty of answering either in any language preferably English or Spanish. Whenever a question is selected the answer is retrieved from the database, so before the students answer is compared to the reference answer it is subjected to various natural language processing techniques like stemming, word sense disambiguation after which BLEU algorithm is implemented so that the users answer and reference answer can be compared. As a feedback it provides score for the answer submitted.

C. Natural Language Processing (NLP) Techniques

NLP has developed various techniques that are linguistically inspired i. e. test is syntactically parsed using information from a formal grammar and lexicon, the resulting answer is then interpreted semantically and used to extract information about the meaning. NLP maybe deep or shallow i. e. parsing every part of every sentence or parsing only certain parts or passages respectively. It also uses statistical means to disambiguate word senses or multiple parse of same sentence. It tends to focus on one document or text at a time because of which it can be rather expensive. It includes techniques like word stemming (removing suffixes), synonym normalization, part-of-speech (POS) and role determination etc (e. g. subject and object)

1. C-Rater was developed by Education Testing Service (ETS). It evaluates the students answer in the following method. The students answer is pre-processed to correction of spelling mistakes and grammatical errors. Then the answer is subjected to POS tagging to remove ambiguity. Phrases, predicated and relationship between the predicates [15] are extracted from the answer by means of feature extractor. The model answer available is now processed using similar NLP tools. The processed answer and the model answer are then subjected to a matching algorithm called Goldmap which is rule based pattern matching algorithm[15]. The algorithm provides a score which acts as a feedback to the students for the concepts that they have stated in their answer.

2. AutoMark was developed by Mitchell et al[16] to assess subjective answers. It allows the questionnaire to set reference marking schemes [13]. It allows the student the liberty of making spelling mistakes as the system auto corrects the spelling during the pre processing of the answer. The answer is then subjected to a sentence analyzer that identifies the phrases and relationship between them. A pattern matching module then identifies if there exist any similarity between the reference answer and the students answer. It also provides a score to the student for the answer submitted.

III. COMPARISION

S. No	Techniques	Systems Implemented on	Limitation
1	LSA	Auto Tutor	a) It does not use word order b)It ignores stop words which can change the meaning of the sentence
		Apex SELSA	
2	BLEU	ERB	a)Highly dependent on reference answers
3	NLP	C-Rater	a)Not fully reliable as it cannot determine the skills of student in writing and expressing

IV. CONCLUSION

This paper reviews the different techniques present for automatic evaluation of free text or subjective answers which has been considered as globally accepted solution to the automatic evaluation problem. Most of the approaches discussed are keyword oriented and do not consider the adjacent terms. Since in natural language processing even stop words like ‘a’, ’the’, ’and’, etc convey meaning to the sentence, the degree of correctness of an answer is still a fuzzy concept which has not been considered till now.

REFERENCES

[1] Magnini B. , Negri M. , Prevete R. and Tanev H. , *“Towards Automatic Evaluation of Question/Answering Systems”*, Third International Conference on Language Resources and Evaluation (LREC-2002) Proceedings, 2002.

[2] Chakraborty P. , 2012, ‘Developing an Intelligent Tutoring System for Assessing Students' Cognition and Evaluating Descriptive Type Answers’, International Journal of Modern Engineering Research, 2, 985-990.

[3] Perez D. , 2004, 'Automatic Evaluation of Users' Short Essays by using Statistical and Shallow Natural Language Processing Techniques', PhD diss. , Master's thesis, Universidad Autónoma de Madrid. Retrieved from: http://www. ii. uam. es/dperez/tea. pdf

[4] Landauer T. K. , Foltz P. W. and Laham D. , "An Introduction to Latent Semantic Analysis", Discourse processes 25, 2-3, 1998, pp. 259-284.

[5] Guest E. and Brown S. , "Using role and reference grammar to support computer-assisted assessment of free-text answers", Unpublished, Leeds Metropolitan University, 2007.

[6] Graesse r, Wiemer-Hastings , P. Wiemer-Hastings , R. Kreuz and Tutoring Research Group, 1999, Autotutor : A simulation of a human tutor. Journal of Cognitive System Research, 1:35-51.

[7] P. Dessus, B. Lemaire and A. Vernier. 2000. Free Text assessment in a virtual campus. In Proc. 3rd Int. Conf. on Human System Learning(CAPS'2000), Paris.

[8] Landauer T. K. , Foltz P. W. and Laham D. , "An Introduction to Latent Semantic Analysis", Discourse processes 25, 2-3, 1998, pp. 259-284.

[9] Ramakrishnan G. , Prithviraj B. P. , Deepa A. , Bhattacharyya P. and Chakrabarti S. , "Soft Word Sense Disambiguation", In Proceedings of GWC, 4, 2004. Retrieved from: http://hnk. ffzg. hr/bibl/gwc2004/pdf/88. pdf

[10] Kanejiya D. , Kumar A. and Prasad S. , "Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA", Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications using Natural Language Processing, Association for Computational Linguistics, 2003, pp. 53-60.

[11] Foltz P. W. , Laham D. and Landauer T. K. , "The Intelligent Essay Assessor: Applications to Educational Technology", Interactive Multimedia Electronic Journal of Computer-Enhanced Learning 1, 1999.

[12] Perez D. , Alfonseca E. , Rodríguez P. , "Application of the BLEU Method for Evaluating Free-text Answers in an E-learning Environment", LREC, 2004, pp. 1351-1354.

[13] Perez D. , Alfonseca E. , "Adapting the Automatic Assessment of Free-Text Answers to the Students", 2005. Retrieved from: https://dspace. lboro. ac. uk/dspacejspui/bitstream/2134/2000/1/PerezD_AlfonsecaE. pdf

[14] http://lrecconf. org/proceedings/lrec2004/pdf/615. pdf

[15] 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

[16] Perez D. , Alfonseca E. 2004. , "Automatic Assessment of short questions with a BLEU inspired algorithm and a shallow semantic representation. In Advances in Natural Language Processiing, volume 3230 of Lecture Notes in Computer Science, Springer Verlag. 25-35.

[17] http://lrecconf. org/proceedings/lrec2004/pdf/615. pdf