Journal of Computer Science and Applications.
ISSN 2231-1270 Volume 6, Number 1 (2014), pp. 1-10

http://www.irphouse.com

A Survey of Various Techniques used for Free Text Evaluation

Rashmi Gurung

Dept. of Computer Science & Engineering
Sikkim Manipal Institute of Technology, Majitar, Sikkim
rasjme@hotmail.com

ABSTRACT

Assessment is the process through which teacher's judge as to what and how much a student has learned. There exist various techniques of assessment of a learner, namely, subjective tests, objective tests, viva voce, etc. However, of all these techniques used, the subjective or the free text based tests are found to be the best mode of evaluation as it allows the evaluator to test all facets of knowledge acquired by the learner during the course and even beyond. While other techniques like Multiple Choice Questions (MCQ), matching pairs or viva–voce have their rightful places in the teaching–learning process, none, match the versatility and completeness of textual answer based tests, also known as subjective tests. Assessing student learning by automatic evaluation of the free text answers to descriptive questions is a very important task since it would be fast and free of human fatigue. However, the computational logistics, in terms of program complexity and time taken are very high.

Keywords: Free text, evaluation, stop words, feedback.

1. INTRODUCTION

Writing is the means through which one is able to express the thoughts and knowledge one has on a particular topic in ones' own words (free text). The method of determining the knowledge one possess on a particular topic is evaluation which can be performed through various methods like objective test, multiple choice test, subjective test, etc. All these methods of evaluation are based on the result obtained by the individual after having performed any of the above mentioned tests.

Interest in question answering has shifted from factoid questions to descriptive questions [1]. The advantage of descriptive tests over other methods of evaluation involves in the elimination of random guessing of answers since the student needs to

provide an elaborate answer on the question asked and not just choose a single answer out of the several possibilities as in case of multiple choice questions, thereby compelling the students to remember important keywords and form an answer in their own words.

The difficulty involved in the evaluation of free text answers is that each student will have their own answers and it is difficult to determine whether an answer is a good answer [2]. The computational challenge that this task poses is also immense. To find out the correctness, or the degree of it, it would be necessary to find out the meaning of the answer text. Going by the complexity of human mind and the various knowledge and creativity issues that guide sentence formation and knowledge expression, interpretation and extraction, the hardness level can easily be fathomed. The automatic evaluation of answers would not only assist the teachers in reducing their workloads but can also help the students, as they can identify their mistakes through the feedback provided by the system after determining their strong and weak areas [3].

The drawbacks for using computers as an assessing tool are the fear of computer failure [4] at the time of submission of the answers and lack of technical training provided to teachers and students. This paper, presents a survey of various techniques that have been used to the evaluation of free text answers. Section 2 contains the background knowledge that should be known prior to going ahead on answer evaluation. Section 3 illustrates the different techniques used for the evaluation of free text answers and Section 4 briefs the conclusion.

2. Background Knowledge

Since the task of evaluation of free text answers is a part of natural language processing (NLP), a few concepts are closely interlinked and easily confused as the other. The following are the tasks which are very similar to answer evaluation but are not the one and the same thing.

A. Firstly, the evaluation of free text answers is not one of the following:

1) Summary Maker:

The summary maker takes as input a source document or multiple source documents, processes it using a computer program that performs an automatic summarization and produces as output a summary of the document provided as an input. The summary generated could contain the important points which are present in the original document or could be based on the certain keywords provided by the user containing only the required information.

2) Emotion Identifier:

Emotional identifiers are systems which are capable of determining the emotional state of an individual by analyzing the expressions generated by a face and the voice generated during speech. The speech and facial expressions of an individual are collected using sensors. The emotional state of an individual is then recognized based on the analysis of the information (speech and facial expressions) gathered using

different algorithms namely Face Recognition Algorithms, Artificial Neural Networks, Hidden Markov Models [5].

3) Page Ranking:
Page Rank is an algorithm developed by Larry Page and Sergey Brin [6]. The page ranking algorithm is based on a graph consisting of nodes as web pages and edges as hyperlinks [6]. Each page has a numerical weight assigned to it which is referred to as Page Rank. The page rank of a page changes due to the links pointing to it. A page rank of a page will be the sum of the numerical weights of the pages pointing to it, but the weight provided to the page may vary depending upon the number of outgoing links the previous pages have.

B. The following techniques should be known before starting with the approaches used in the evaluation of free text answers.
1) Word Sense Disambiguation:
Some words possess multiple meanings leading to ambiguity. So it is up to the interpreter as to what inference is drawn from a sentence depending upon the use of the ambiguous word in the sentence. Word Sense Disambiguation is the method of determining the exact meaning of a word in a sentence [7]. This method is of importance to the evaluation of answer since the correct meaning of the words used in the answer should be known.

For example, the word "dice" can be interpreted differently based on its use in a sentence. The meaning of the word is determined based upon the other words in the sentence. In the sentence “Let’s play dice” implies the game sense while in the sentence “Peel the potatoes and dice them” implies cutting the potatoes into small cubes.

2) Part of Speech Tagging:
Part of Speech (POS) Tagging is the method of marking each word present in a sentence with a particular part of speech tag. POS Tagging is mainly performed to resolve ambiguity. POS tagging algorithms are of two types: rule based and stochastic [8]. As stated in [8] the Rule-based tagger uses a set of rules to assign tags to ambiguous words and Stochastic tagger resolves tagging ambiguities by using a tagged or an untagged corpus to determine the POS tags for a word. An example of rule-based tagging is Brill's tagger [9] and an example for stochastic tagging is a tagger using the Hidden Markov Model technique [8].

3) Predicates and arguments:
Role and Reference Grammar (RRG) developed by William Foley and Robert Van Valin. Jr. is based on predicates and their arguments where parsing is performed to identify verb, adjective, noun. This method extracts the meaning of a sentence by identifying the predicate which is a verb present in the sentence first and then the modifiers present in the sentence namely the article, adverbs, adjectives are identified [10].

3. Techniques used

The different techniques that are used for the evaluation of free text answers are discussed below.

A. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a mathematical technique [11] proposed by Launder et al. As the name suggests the technique is capable of extracting the meaning of words which are existing but hidden in a text by applying a decomposition method to determine the semantic similarity between the text submitted by the student and the reference text available in the system. This technique is also known as the bag-of-words approach [12]. The working of LSA is described as follows: In this method a list of sentences is provided as input by a student. Then from the list of sentences provided as input, certain words are extracted. The words extracted does not include stop words which are frequently occurring words like 'a', 'and', 'the', 'is', etc. After the extraction of words from the student text, a matrix is created consisting of rows as the words and columns as the sentences from the student text. Each element of the matrix represents the number of occurrence of the word against the sentences. Each element of the matrix is then converted to its log and divided by the entropy value computed using – p log p [11]. The matrix now obtained is subjected to singular value decomposition (SVD).

In SVD, the matrix A created consisting of m rows of words and n columns of sentences is subjected to decomposition into three matrices as stated below in [13]:

$$A_{m\times n} = U_{m\times m} \times S_{m\times n} \times V^T{}_{n\times n} \quad (1)$$

Where
$U^TU = I$
$V^TV = I$

Both U and V are orthogonal matrices and S is an identity matrix [13]. Then from the three new matrices obtained a specific number of dimensions are chosen to obtain a new matrix $\bar{A}$ which is compared with a similar matrix created for the reference text using LSA technique. The reduction process of the original matrix A to $\bar{A}$ in LSA helps in identifying the meaning of the word with respect to the context it is being used in a sentence. The results produced by SVD using the cosine between vectors closely matches with humans [14].

However, LSA has some limitations. The LSA technique takes sentences as input for extracting the meaning of words, thereby failing to take into consideration the order of the words [11] and tends to ignore stop words which also convey meaning to sentences. Although this limitation exists LSA is used in a number of systems that performs the free text evaluation of answers namely:

1) Apex:

Apex is a web based learning application [4] developed by Dessus et al [14]. The system provides a user friendly environment for the students and has a database where the questions and answers are stored. The teacher is assigned the task of marking the

course where he selects a topic and then a notion for the topic is selected [14]. The student after having selected a topic needs to enter an answer. The evaluation performed by Apex comprises of three modes: content, outline and coherence based evaluation [4].

In the content based mode of evaluation the answer submitted by the student is compared with the prior notion entered by the teacher earlier. Depending upon the existence of a notion in the text entered by the student a similarity score using LSA technique [11] ranging from -1 to 1 [14] is assigned for each notion.

In the outline based mode of evaluation, each paragraph of the text entered by the student is compared with the notion entered by the teachers, if the similarity score obtained using LSA technique is high then a notion similar to the paragraph is displayed else notion is not displayed.

The coherence based mode of evaluation is able to detect the missing linkages between sentences in a text and provide this information to the students.

2) Intelligent Essay Assessor:

Intelligent Essay Assessor developed by Landauer, Foltz and Laham is a web based application [15] where the student can submit their essays for evaluation. Unlike Apex, IEA does not require the teachers to select topics and their notions. IEA is provided the essays based on a particular topic to serve as a reference for evaluation purpose [15]. The essays of the student are then compared with the reference essays and a feedback is provided based on content, mechanics and style [4]. The content based module of evaluation uses the LSA technique to determine the semantic similarity between the essays. The next module which is the mechanics module is responsible for checking spelling and grammatical mistakes present in the student essay. The style module of evaluation takes into account the grammar and writing style of the student essay and uses the LSA technique for measuring the semantic similarity between the student essay and the reference essays. In addition to this IEA has the ability to provide information on essays which it considers to be: highly creative, out of topic, not following the standards of writing an essay [15].

The feedback provided by IEA is fast (within 20 sec [4]) and effective since it provides feedback to students on any topic that they might have skipped and also allows them the facility to make corrections and resubmit their essays. This application is of benefit for the students since they can get instant feedback on the essay that they have written which helps them to rectify their mistakes and hence improve their writing and thinking skills.

A modification to LSA has been proposed by Kanejiya et al. where an approach called Syntactically Enhanced LSA (SELSA) [12] has been described. As described earlier the LSA technique takes into consideration a matrix of order words × sentences but in this modification proposed the syntactic information of the sentence is taken into account the POS tag of the words present in the sentence hence called SELSA. For a word the POS tag of the previous word is taken into account so called prevtag [12]. In this technique a matrix of order prevtag words × sentences is created. The entropy, ε_{i_j} and the matrix element, $x_{i\_j,\ k}$ are computed as follows as stated in [12]:

$$\varepsilon_{i\_j} = -\frac{1}{\log K}\sum_{k=1}^{K}\frac{freq_{i\_j,k}}{t_{i\_j}}\log\frac{freq_{i\_j,k}}{t_{i\_j}} \quad (2)$$

$$x_{i\_j,k} = (1 - \varepsilon_{i\_j})\frac{freq_{i\_j,k}}{n_k} \quad (3)$$

Where:

$freq_{i\_j,\ k}$ implies the number of times the word w_i (where i is the vocabulary) with prevtag p_j (where j is the part of speech tagged vocabulary) appears in the document d_k (where k is the number of documents).

t_{i_j} implies the number of times the i_j^{th} word-prevtag pair appears in the document and is computed using formula 4 as stated in [12]:

$$t_{i\_j} = \sum_{k=1}^{K} freq_{i\_j,k} \quad (4)$$

The matrix x of order prevtag words × sentences is then subjected to SVD as in case of LSA. This decomposition technique filters out the noise and determines the semantic similarity between the prevtag words and sentences.

B. BiLingual Evaluation Understudy (BLUE) Algorithm:

The BLUE algorithm proposed by Papineni et al. is an n-gram co-occurrence scoring procedure [16]. Here n-gram implies a sequence of words which are used to perform comparison of two different texts. In this method the input sentence provided is translated by the machine and then n-gram matches between the machine translation and the reference translation is counted. The machine generated translation is considered to be better if the number of matches of the n-grams between the machine translated sentence and reference translations is high [16]. N-gram co-occurrence scoring is typically performed segment-by segment, where a segment is the minimum unit of translation coherence [17].

The working of the BLUE algorithm is as follows: The candidate provides an input sentence in one language which is then translated by the machine to another language. The BLUE algorithm then performs a comparison of the candidate translation with the reference translation. The following tasks are then performed as stated in [16]: Each word in the candidate translation is then compared for a match in the reference translations that are available in the system and is stored as count. Once a word of any reference translation matches with the candidate translation it cannot be taken into consideration again. The count for each word of the candidate translation against the reference translations is computed. Then the comparison of the counts of each word for all the reference translations is performed and the maximum count of each word is taken into consideration. The occurrence of each candidate word against the candidate translation is computed and then it is compared with the maximum count of each word computed earlier. The comparison 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) [16]. The MUP value for each of the candidate translation is computed where the number of words n used for determining MUP may range from 1 to 4.

The MUP value obtained for each value of n then needs to be summed using log since BLUE uses the average logarithm with uniform weights [16]. The following formula is then used to compute brevity penalty BP and BLUE as stated in [16] where 'c' is the length of the candidate translation and 'r' is the length of the reference translation:

$$BP = \begin{cases} 1 \; if \; c > r \\ e^{\left(1-\frac{r}{c}\right)} \; if \; c \leq r \end{cases} \quad (5)$$

$$BLUE = BP.\exp\left(\sum_{n=1}^{N} \frac{\log MUP(n)}{N}\right) \quad (6)$$

The limitations of BLUE are (i) Dependent on the reference translation [18] for evaluation purpose. (ii) Takes into consideration only the precision score ignoring recall [18] of words. Despite these disadvantages BLUE is used in the following system for evaluation of free text answers.

1) Atenea:

Perez et al. [19] developed Atenea for the evaluation of free text answers. The system 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 [19]. Since a machine translation is employed, the system is capable of evaluating answers in English as well as Spanish. Evaluation is thus possible irrespective of the language the student wishes to answer the question.

Atenea allows the user to choose a question. Based on the question the student is then expected to answer the question in any language preferred either English or Spanish. When a question is selected the corresponding reference answers are retrieved from the systems database. The answer provided by the student is then subjected to a number of natural language processing (NLP) techniques like stemming, word sense disambiguation [19], etc. The answer is then subjected to BLUE algorithm so that the machine translated answer can be compared with the reference answers present in the system. The algorithm then provides a score to the students for the answer they have submitted. Apart from evaluating the student answer in English or Spanish, the user interface of Atenea can be customized as per the students need. The feedback provided to the users may be either a basic feedback [19] providing just the score for the answer submitted or a detailed one highlighting the answer submitted if the words in the answer submitted matches with the reference answer [19].

C. Natural Language Processing (NLP) Techniques:

Artificial Intelligence (AI) is the science of developing intelligent entities which are made by humans and capable of simulating natural behavior so called artificial. These entities are considered to be intelligent since they are capable of acquiring knowledge from the environment and applying it. Natural language processing is a branch of AI which aims at building intelligent entities that is capable of performing task on the natural language that humans use. But the problem with processing human languages is the diversity in form and structure thereby leading to ambiguity. Hence to resolve this ambiguity different NLP techniques are used. The following systems use NLP

techniques for the evaluation of free text answers.

1) C-Rater:
C-Rater developed by Education Testing Service (ETS) uses NLP techniques for the evaluation of free text answers provided by the students. The following tasks are performed for evaluating the student answer as stated in [20]: The student answer is subjected to correction of spelling mistakes and grammatical errors prior to being processed. The answer is then subjected to POS tagging to resolve ambiguity. Phrases, predicates and relationship between the predicates [20] are extracted from the student answer by means of a feature extractor. The model answers available for evaluation of the student answer is also processed using the similar NLP tools. The processed model answer and student answer are then subjected to a matching algorithm called Goldmap which is a rule based pattern matching algorithm [20]. The algorithm produces a score which is provided as a feedback to the students for the concepts that they have stated in their answer.

2) Automark:
Automark developed by Mitchell et al [21] is a software system capable of evaluating free text answers provided by the students to subjective questions. The system allows the teacher to enter reference marking schemes [21] which serves as a reference for answer evaluation. Here also the student answer is expected to have spelling mistakes and typing errors. So, the system performs a preprocessing of the student answer to correct the spelling mistakes before performing other tasks. The following tasks are then performed as stated in [21]: the student answer is subjected to a sentence analyzer which identifies the phrases and relationship that exists between them. A pattern matching module is then used to determine if there exist any matches between the reference marking scheme provided by the teachers and the processed student answer. As in C-Rater the pattern matching module produces a score which is provided as a feedback to the students for the answer they have submitted.

4. CONCLUSION

This paper presents a survey of the various techniques that have been used for the evaluation of free text answers. The detail of each technique, their strengths, their limitations, systems in which the technique have been employed for the evaluation of free text answers has been discussed. The modification that has been made to the LSA evaluation technique has also been discussed. In all the systems discussed above it has been observed that in addition to evaluation of free text answers feedback to the students is also being provided. It is important to provide feedback during evaluation since evaluation is not just providing a score for the answer submitted but improving the learning process and helping the students by providing an effective feedback to the students.

References

[1] Lin J. and Fushman D.D., 2005, 'Automatically Evaluating Answers to Definition Questions', Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 931–938.

[2] Magnini B., Negri M., Prevete R. and Tanev H., 2002, 'Towards Automatic Evaluation of Question/Answering Systems', Third International Conference on Language Resources and Evaluation (LREC-2002) Proceedings, 128-134.

[3] 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.

[4] 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

[5] Proyecto Fin de Carrera, 2010, 'Face Recognition Algorithms'. Retrieved from:

[6] http://www.ehu.es/ccwintco/uploads/e/eb/PFC-IonMarques.pdf

[7] Page L., Brin S., Motwani R. and Winograd T., 1999, 'The PageRank citation ranking: bringing order to the web'. Retrieved from: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

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

[9] Dandapat S., Sarkar S. and Basu A., 2004, 'A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali', International Conference on Computational Intelligence, 169-172.

[10] Dao T. N. and Simpson T., 2005, 'Measuring Similarity between Sentences'. Retrieved from: http://trac.research.cc.gatech.edu/ccl/export/184/SecondMindProject/SM/SM. WordNet/Paper/WordNetDotNet_Semantic_Similarity.pdf

[11] Guest E. and Brown S., 2007, 'Using Role and Reference Grammar to Support Computer-Assisted Assessment of Free-Text Answers', Unpublished, Leeds Metropolitan University.

[12] Landauer T. K., Foltz P. W. and Laham D., 1998, 'An Introduction to Latent Semantic Analysis', Discourse processes 25, 2-3, 259-284.

[13] Kanejiya D., Kumar A. and Prasad S., 2003, '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, 2, 53-60.

[14] Singular Value Decomposition Tutorial. Retrieved from: http://web.mit.edu/be.400/www/SVD/Singualr_Value_Decomposition.htm

[15] Dessus P., Lemaire B. and Vernier A., 2000, 'Free-text Assessment in a Virtual Campus', Proceedings of the 3rd International Conference on Human-Learning Systems, 2-14.

[16] Foltz P. W., Laham D. and Landauer T.K., 1999, 'The Intelligent Essay Assessor: Applications to Educational Technology', Interactive Multimedia Electronic Journal of Computer-Enhanced Learning 1, 2. Retrieved from: http://imej.wfu.edu/articles/1999/2/04

[17] Papineni K., Roukos S., Ward T. and Zhu W., 2002, 'BLEU: a Method for Automatic Evaluation of Machine Translation', Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 311-318.

[18] Doddington G., 2002, 'Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics', Proceedings of the Second International Conference on Human Language Technology Research, 138-145.

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

[20] http://lrec-conf.org/proceedings/lrec2004/pdf/615.pdf

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

[22] Leacock C. and Chodorow M, 2009, 'C-rater: Automatic Content Scoring for Short Constructed Responses', Proceedings of the Twenty Second International FLAIRS Conference, 290-295.

[23] Mitchell T., Russell T., Broomhead P. and Aldridge, N., 2002, 'Towards Robust Computerized Marking of Free-Text Responses'. Retrieved from: https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/1884/1/Mitchell_t1.pdf