DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5783

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

THEORETICAL ANSWER EVALUATION USING LSA, BLEU, WMD AND FUZZY LOGIC

Praful Mishra
K.J Somaiya Institute Of Engineering And Information Technology, Sion, Mumbai, India

Anmol Mishra
K.J Somaiya Institute Of Engineering And Information Technology, Sion, Mumbai, India

Aniket Bharti
K.J Somaiya Institute Of Engineering And Information Technology, Sion, Mumbai, India

Prof. Sarita Ambadekar
K.J Somaiya Institute Of Engineering And Information Technology, Sion, Mumbai, India

Abstract : Assessment of student answers to grade their overall understanding of a subject is a critical task. However grading can be monotonous and sometimes can be tedious task for the teachers. Automatic Grading can reduce tedium on teachers but it is complicated by free form student inputs. The main task of automatic grading system is to assign ordinal scores to student answers, based on "model" or ideal answers. Here we introduce a novel framework comprising of three building blocks Word Mover Distance(WMD)a statistical model Latent Semantic Analysis(LSA),Bilingual Evaluation Understudy(BLEU) and Fuzzy logic, a model based on degree of truth to output scores. In other words LSA is used to identify the semantic similarity between two concepts. Word Mover's Distance (WMD), uses vector encoding of words to calculate the minimum cumulative distance that words from a reference solution need to travel to match words from a student answer. This cumulative distance assess the distance between two documents in a meaningful way, even when they have no words in common. Fuzzy logic is a primitive model in this system which is used to output the final score based on inputs which are the outputs of LSA and WMD. This proposed method gives better precision, enhanced dependability of results, thus saving the effort and time of staff.

Keywords: Latent Semantic Analysis, Singular Value Decomposition, N-gram, Word Mover Distance, Fuzzy Logic.

I. INTRODUCTION

Theory based examinations are held periodically to assess students academically.The purpose of these assessment is to gain insight of student understanding and knowledge enhancement. However the manual evaluation of answers sometimes can be monotonous ,bias errors and tiring. To overcome these difficulties a faster and reliable method to evaluate answer is required. Natural Language Processing (NLP) is a technique in Artificial Intelligence that enables us to analyze and synthesize natural language. NLP is further divided into syntactic analysis and semantic analysis. Latter is used for analysing grammar and arrangement of words in such a manner that they show relationship among themselves. While former is used to extract meaning from text. We publish a paper that uses both syntactic and semantic methods to evaluate student's answer and allot them marks.

This paper proposes a algorithm to avoid this gruesome manual answer evaluation. The assessment of answers is done using a novel framework comprising of generative probabilistic technique and degree of truth techniques. Due to the freedom of input students write a particular sentence in various form.LSA[1] measures the semantic similarity of these answers with standard answer by finding out important topics in both, BLEU[2]avoids overrating of answer by LSA. WMD[4] measures the similarity of students answer with standard answers even if sentences in both are written different way but mean the same. LSA,BLEU,WMD along with soft computing technique Fuzzy Logic[3] gives the overall assessment of student and standard answer.

II. EXISTING SYSTEM

Attali and Burstein devloped E-rater (Electronic Essay Grade), that checks the writing style and the structure of the essays rather than the specific content. Text features like vocabulary level used, word occurring probability, correlation, word length and essay length were extracted and parsed using MSNLP (Microsoft Natural Language Processing) tool with training essays. These features are allotted weightage. Already graded essay are used for evaluating new essays, its features are compared to already graded essays. The strength is the agreement between the E-rater and human is above 97%. The weakness of E-rater is that it requires a number of manually scored training essays to score the answers.

C-rater (Concept rater) (Burstein et al., 2001; Yigal et al., 2008) was also developed by ETS (Educational Testing Service) and it is also called as content rater. The scoring is based on the content and concept. It uses natural language

processing techniques i.e., it uses lexical semantic techniques are used to build the scoring system. This system uses domain related, concept based data in evaluation.

In 2010, Cutrone and Chang in their research paper proposed a short answer evaluation method using natural language processing (NLP) techniques. This technique reduces the standard answer and student answer into its canonical form and compares them. Canonical forms of the standard and student answer were found using techniques like tokenization, stemming, morphological variation, etc. It can evaluate single-sentence answers only.

III. PROPOSED SYSTEM

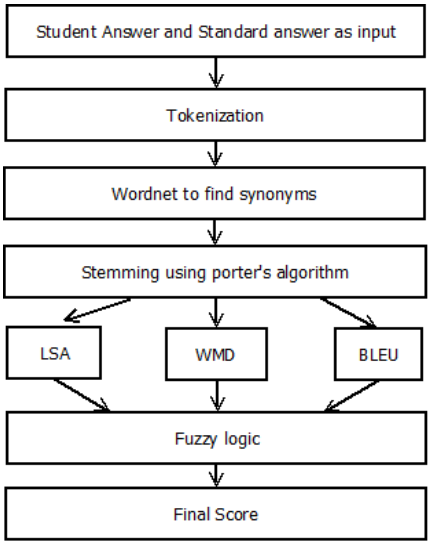

Fig. 1 High-level steps in subjective evaluation

Preprocessing

The first towards answer assessment is the preprocessing of student and model answer. The preprocessing methods such as tokenization, pruning and stemming are applied on the student answers to reduce the size of the answer and prepare the input for further processing.

1.Tokenization

Student and model answer both are tokenized before processing. Each sentence in student and model answer is broken down into words. All the punctuations and white spaces are removed and only words are stored and given for further processing.

2.Pruning

Pruning[1] is a preprocessing technique which removes the stop words such as “is”, “are”, “you”, “me”, etc.

3.Wordnet

It may be the case that a student may have used synonym of word in model answer. So we use wordnet to find all the synonyms of tokenized words using wordnet. From synonyms of each token we find the standard word used in model answer.

4.Stemming

Stemming[1] removes suffixes and prefixes attached with the terms. In this work, the pruned words are given to stemming process.

Answer Evaluation

1.Latent Semantic Analysis

Humans can easily understand hidden meaning between two sentences by comparing relations between words. LSA[1] is a statistical-algebraic technique for extracting and inferring contextual usage of words in documents. The document is nothing but text corpus. LSA evaluates documents to find out core meaning in that document. It creates a matrix of words occurring in each document. In this matrix rows represent terms and columns represents paragraph.

Singular Value Decomposition is applied to the matrix which decomposes it into three other matrix. By reducing dimension we reconstruct the original matrix again. This newly reconstructed matrix has different value in each rows as compared to old matrix. New value represent hidden semantics within the paragraph. Applying cosine similarity between two columns we get similarity measure between two paragraphs.

2.Bilingual Evaluation Understudy

The problem with LSA is that it tends to over grade answer that has repetitive number of keywords. Secondly LSA does not consider the word order for example: For LSA carbon dioxide and dioxide carbon are same thing. These drawbacks are overcome using BLEU[2]. The core idea of these algorithms is that the more similar a student’s answer is to the model answer the better it is, and, consequently, it will have a higher score. BLEU outputs a metric value ranging between 0 and 1 using N-gram. The frequency of each token in student answer along with total number of words in the student answer is calculated. One should take care that these values do not exceed the frequency of token in model answer and total number of words in model answer respectively. If any value exceeds then the corresponding value calculated from model answer and is then divided by frequency of each token in student answer and by the total number of words giving us a fraction value between 0 and 1 as a final score.

The procedure is as follows

1.Count number of N-grams from student answer appearing in model answer (up to value of N). If occurrence of N-gram in student answer is more than model answer then it is clipped to maximum frequency with which it appear in model answer

2.Combine scores of each value of N, as a weighted linear average (Geometric Mean).

3.Apply brevity factor as a penalty factor for short student answer.

3. Word Mover’s Distance

The measure of similarity between two block of text can be used as a good measure for evaluation of answers. Ideally statically based algorithm like LSA,BLEU etc can capture semantic relation between two documents .cosine similarity can easily captures this relation easily but it fails when two documents convey same meaning but using completely set of

words. So when two document have no word in common their euclidean distance would be maximum.

One way out of this conundrum is the word mover's distance (WMD)[4] that adapts the earth mover's distance to the space of documents. At a high abstraction, the WMD is the minimum distance required to transport the words from one document to another. We assume that we are given a word embedding matrix(word2vec).

We use Word Mover Distance (WMD) problem on a matrix of pairwise distances between each state vector of the model and student answers. If a word wi appears fi times in a document, its weight is calculated as

$$d_i = \frac{f_i}{\sum_{j=1}^{n} f_j} \qquad (1)$$

where n is the number of unique words in the document. The higher its weight, the more important the word is. The dissimilarity between word wi in student answer and word wj in model answer can be computed as

$$c(w_i, w_j) = \|x_i - x_j\|^2 \qquad (2)$$

where xi and xj are the embeddings of the words wi and wj , respectively.

Let D be and D0 be nBOW representations of student and model answer respectively. Let T ∈ R n×n be a flow matrix, where Ti j≥ 0 denotes how much the word wi in D has to "travel" to reach the word wj in D 0 , and n is the number of unique words appearing in D and D 0 . To transform D to D 0 entirely, we ensure that the complete flow from the word wi equals di and the incoming flow to the word wj equals d 0 j . The WMD is defined as the minimum cumulative cost needed to move all words from D to D 0 ,

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij}\, c(w_i, w_j) \qquad (3)$$

subject to

$$\sum_{j=1}^{n} T_{ij} = d_i, \forall i \in \{1, \ldots, n\}, \sum_{i=1}^{n} T_{ij} = d'_j, \forall j \in \{1, \ldots, n\}$$

The solution is achieved by finding Ti j that minimizes the expression in Equation 1.applied this to obtain nearest neighbors for document classification, i.e. k-NN classification .Therefore, WMD is a good choice for semantically evaluating a similarity between documents.

4.Fuzzy Logic

Scores given by LSA, BLEU and WMD are mapped to a single score using Fuzzy logic[3].Classical logic only permits conclusions which are either true or false. However, Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. It is based on the fuzzy set theory that allows the elements of a set to have varying degrees of membership, from a non-membership grade of 0 to a full membership of 100 per cent or grade 1.Fuzzy logic uses IF-THEN rule based approach and has tolerance towards imprecise data. These features make it highly applicable for evaluation system. Traditional crisp variable has value either 0 or 1 indicating completely false or completely true. However fuzzy variable can have truth value varying between 0 and 1 indicating their degree of membership. Each of the above algorithm i.e. LSA, BLEU, WMD represent degree of correlation between student answer and the model answer so they can be used as fuzzy variables. These are independent variable as they are pointing toward different aspect of similarity between model answer and student answer. The technique uses Mamdani's fuzzy inference mechanism. It has following stages.

Step 1: Fuzzification of inputs
Step 2: Application of fuzzy operators
Step 3: Application of the implication method.
Step 4: Aggregation of all outputs
Step 5: Defuzzification

The fuzzy logic model is designed with three input variables that are scores of LSA,BLEU and WMD with three membership functions (poor, average, and good) and one output variable (final) with three membership functions (low, medium, high).

The rules are as follows:

If(lsa['poor'] & bleu['poor'] &wmd['poor']) then final['low']
lsa['poor'] & bleu['poor'] &wmd['average'], final['medium']
lsa['poor'] & bleu['poor'] &wmd['good'], final['medium']
lsa['poor'] & bleu['average'] &wmd['poor'], final['low']
lsa['poor'] & bleu['average'] &wmd['average'], final['medium']
lsa['poor'] & bleu['average'] &wmd['good'], final['medium']

IV.RESULT

For the testing of novel technique and evaluation of answers we have created our own database of answers. The database consists of multiple brief answers of various students for various question with their standard answer. The hybrid technique was applied to the database for the assessment of the student's answers. The same answers were given to human evaluator. The marks generated by the system and by the systems are compared. The accuracy of results of novel technique varied between 0.71 and 0.85. A sample of the results generated is given in Table 1. The individual marks generated for sample answers of subject chemistry (Group 18 Element) are given along with human-assigned score.

Table 1 Sample results generated using hybrid technique
Question: -What are System Calls ?
Maximum Marks: - 5

Student's Answer	**SYSTEM SCORE**	**Teacher's Score**
It is an interface to the service available by operating system for the computer user. It acts as interface between processors and operating system. These are general routine code in C++ and C. Types-process control-file manipulation-device manipulation-communication-information maintenance.	2.592	3
The various calls generated during the execution of a program are called as system calls. Below the application level stands the GUI which provides an interface for user interface. The kernel handles all the system process , memory allocation , segmentation etc. It also undertake file and resource handling.	2.685	3
system call is the interface between the user and an operating system.It is the call for the function of the operating system .They are the runtime written in different languages like c,Javaetc..Types 1)Process control- when the process is getting executed in the main memory.	1.8	1

V. CONCLUSION

Most of the evaluation systems available in online, evaluate only the objective type answers. The proposed system evaluates the descriptive type answers of students. Input size of file is reduced using pruning and stemming. The assessment performance of the clustering is improved due to semantic method LSA used for text transformation.
For evaluation, the proposed method uses the semantic similarity between words in sentences. It provides more effective evaluation of the learning process. The proposed Assessment algorithm evaluating the descriptive type answers in O(n) time, for n number of answers. This system would be of great help for the academic institutions in reducing the work and time of evaluation and to speed up the publication of results.

REFERENCES

[1] Meena,K, Lawrance,R"Semantic Similarity Based Assessment of Descriptive Type Answers " in IEEE, 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16).
[2] KishorePapineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu " BLEU: a Method for Automatic Evaluation of Machine Translation" IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com
[3] Intelligent Fuzzy Decision Making for Subjective Answer Evaluation Using Utility Functions Ani Thomas; M. K. Kowar; Sanjay Sharma 2008 First International Conference on Emerging Trends in Engineering and Technology Year: 2008 Pages:587 - 591 IEEE Conference Publications.
[4] From Word Embeddings To Document Distances ;Matt J. Kusner ,Yu Sun,Nicholas I. Kolkin,Kilian Q. Weinberger. Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130