Survey of Techniques Used for Answer Evaluation using Semantic Network

Sampa Das
Dept. of Comp. Sc. & Engg.
Sikkim Manipal Institute Of Technology
Sikkim, India
E-mail:sampa.sit@gmail.com

***Abstract*—** Automatic Evaluation of Free Text Answers' have become a necessity, not only for better acceptance of online learning, but also to handle the pressure of assessment of a large number of students' responses in a fatigue free pedagogically correct method in traditional learning environments. This work is aimed at developing a model to evaluate free text answers of students based on the semantic similarity it has with the model answers prepared by teachers. The model answers are prepared prior to the evaluation process and through a process of dynamic semantic network building, a model is prepared which is used in evaluation. The proposed technique should allow the flexibility of comparing a student's answer with two or more model answers and finally evaluating it against the model answer it most closely resembles.

***Keywords*—** *E-learning, automatic evaluation, semantic network.*

*****

I. INTRODUCTION

Teaching-learning has been an important aspect of societal behavior of man since time immemorial. Over the years, the teaching-learning paradigm has undergone huge metamorphosis from the gurukul system of ancient India to the present day distance learning systems. The pedagogy/andragogy process, still involve interactions between two major actors, namely instructor/supervisor/teacher and the learner, where the teacher is expected to have assimilated the facts, ideas and the underlying concepts of a topic to be able to propagate it so as to enable the learner to assimilate the same. However, the setup in which they interact is not the same anymore and for all current developments must be considered dynamic.

Whether the assimilation at the learners' end has been fruitful has then to be investigated and evaluated. Evaluation therefore is the marker to decide the success of the learning endeavor. It is thus the decision box, which determines the performance of the teaching-learning process enacted with respect to the stated and pre-decided outcomes of the course. The outcome provides both the teacher and the learner the framework to plan the progress. The importance of evaluation towards effective teaching-learning needs no further arguments in support, and we know that it decides not only the amount of learning but also contributes towards the refinement of the learning process.

In the current context of technology aided education and the immense pedagogical transformations that has come along with it, the evaluation aspect of teaching-learning has been further complicated. Due to the lack of personal interaction between the teacher and learner, the human element of continuous evaluation during learning is not possible anymore. The issue of volume further augments the complexity, as the human evaluator would now have to deal with a larger number of learner responses in comparison to classical teaching-learning. The much required intelligent features of knowledge, impartiality and benevolence of the human evaluator is thus put through a lot of stress occurring out of fatigue. It therefore becomes necessary to find an automated solution to the problem of evaluating learner response.

Automated evaluation of learners' response has been thought over, experimented and implemented using various platforms built for the purpose. Due to the inherent computational difficulties in Natural Language Processing and the associated implementation complexities, the trend had shifted towards close ended question systems. These are straightforward and do not need any Natural Language Processing (NLP) based techniques or algorithms. This type of tests are however not fully reliable for evaluation of fulfillment of learning outcomes as these cannot determine the skills of students in writing and expressing. While their popularity may be credited to the objectivity and quantifiability, they have their limitations. It is difficult to check the learners' knowledge and understanding of the proof and theoretical aspects [1]. Further, a learner during the test may be in a state of full knowledge, partial knowledge, absence of knowledge, partial misconception or full misconception. Multiple choice questions not only fail to credit students for partial knowledge but also may credit answers even if the learner is in state of absence of knowledge or partial or full misconception because it is also possible to score in such tests using pure guess work [2]

The claim of higher efficacy of open ended questions over close ended ones is still debatable. However, what goes without contention is that a wider range of the learners' ability can be tested using open ended questions [3].

The problem with the evaluation of open ended questions lies in the variation in answering and presentation that each learner may adapt. Since the learner presents his response in his own words and style, the same question will have as many different answers as the number of learners' answering

due to the richness in form and structure of natural languages.

The salient features of open ended questions can be listed as [4]:

- No fixed method
- No fixed answer or many possible answers
- Solved in different ways and at different levels

To evaluate such answers, the degree of correctness of each response has to be evaluated, which requires sense extraction. This presents an immense computational challenge as it requires implementing knowledge extraction based on semantics and context. The closest approach may be to evaluate with respect to one or a couple of model answers made available. Even then, it has to be based on finding the semantic similarity between the correct model and the learner response. To augment the difficulty there exists the comparisons drawn between the human evaluator and the machine evaluation. The score returned by the human evaluator would be fuzzy, within a range of the minimum and the maximum scores permissible, and it would be influenced by the evaluators understanding, knowledge and benevolence. It must also be considered that no two individuals would grade an answer in exactly the same manner as the quality of evaluation of essay type answer books involving multiple evaluators for courses with large number of enrolments is likely to be affected due to heterogeneity in experience, expertise and maturity of evaluators [5].

The purpose of this work is to build an automated system that evaluates the free text responses of the learner to questions which requires the answer to be constructed rather than memorized and written verbatim ranging from a single sentence to four-five sentences. Attempts made towards the accomplishment of this task by esteemed researchers form considerable literature. However a solution acceptable to all and fit for all types of questions requiring free text responses has not yet been developed and this problem has prevented automated marking systems from being used in high-stake short-answer marking [6]. The popular approaches consider either a keyword centric or n-gram based methodologies, with pre-processing resulting in removal of stop words from the text to be evaluated. The approach being presented is in deviation from the mentioned ideas and considers not only the keywords but also the relations they have using a dynamic scalable semantic network. Unlike n-grams technique, the number of words before and after a keyword is not fixed and varies depending on the occurrence of the next keyword.

II. LITERATURE SERVEY

Question answering has steadily shifted from being inclined towards factoid questions to be popularly accepting descriptive questions [7] and attempts at developing automated systems for practical usage have also met with some success. There has largely been two broad approaches to this, the first being free-text assessment based on surface features and later free-text assessment based on course content [8].

The earliest attempts towards surface feature based assessment was reported in [9] where the length of the essay, number of punctuations, number of connectives, average word length etc. of an essay were used to find the correlations between already graded essays and the essays to be graded. The e-rater,[10] also extracts correlations between already graded essays and ungraded ones using about sixty surface features similar to [9]. Each of these features relies on syntactic, rhetorical or topical content of the text. After initial syntactic parsing, the system identifies the rhetorical structure of the essay depending on sentences containing rhetorical arguments. e-rater reports a correlation of 0.8 with two human evaluators. A major drawback of this approach is that it does not consider the semantic content of the essays and as a result an apparently unrelated answer having the right mix of surface features could be returned with a good score.

The limitations of free-text assessment based on surface features is overcome to an acceptable degree by the Latent Semantic Analysis (LSA) technique [11], which is a mathematical technique for extracting the meaning of words that are present in a sentence or a passage. LSA works on extraction of index words from documents after removal of stop words and subsequent matrix building. The matrix stores the frequency of occurrence of each index word against each title, which is used to calculate the entropy of each of the index words. A Single Value Decomposition of the resulting matrix returns the cosine measure of the similarity of two documents under review. This technique takes one document as a standard and finds the similarity of the other document with it using the LSA method.

The Intelligent Essay Assessor (IEA) developed by Foltz, Laham and Landauer, [12], uses the LSA technique to assess essays of learners' and has been used for online evaluation. The essays apart from being graded based on the similarity of content with respect to one or more reference essays are also evaluated for grammar and spelling. The system claims to be capable of assessing the amount of knowledge a student has through the automatic evaluation of essays submitted by the students and the grades generated highly correlate with that of human assessors. The feedback that IEA generates
is also helpful for students to enable them to find out and correct their mistakes, which makes it particularly suitable for the e-Learning scenario.

Dessus, Lemaire and Vernier [8], developed a web-based learning application system by the name of Apex which rates the learners' response in free text with reference to answers already stored in the system database. The task of the teacher is to identify the topic and notion in a text, with which the learners' response in then compared. The semantic similarity value is measured using LSA and it denotes the knowledge of the student on the selected topic. The system returns textual evaluation based on how closely a notion was covered from the score returned.

Syntactically Enhanced LSA (SELSA), a modification of LSA proposed by Kanejiya, Kumar and Prasad [13] considers a word along with its context by taking it along

with its adjacent words as a unit of knowledge representation. The SELSA approach overcomes the shortcoming of LSA as it considers the word order, which however is limited to the 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 the words only as in LSA.

Popular as it may be, LSA has its share of drawbacks too. A principle disadvantage of this approach is that it does not take into consideration the word order. This bag of words approach overlooks the logic and semantic relations that are reflected in text and are so important in evaluation. Even a simple example may show that LSA does not recover the optimal semantic factors intended in the pedagogical example used in many LSA publications [14]. The computational complexity involved in LSA is also large as the size of the matrices grows with the number of documents that are taken as references. As it may be understood, during evaluation it would be inappropriate to compare the learners' response with only one model response. It is also seen that LSA does not scale up well. As the document space grows, it gets more and more difficult for LSA to recover the set of semantic factors for optimal results [14].

Another popular technique followed by some implementations is the Bilingual Evaluation Understudy algorithm (BLUE) [15]. The algorithm, which is an n-gram scoring method, compares the machine translated output with reference translations using word n-grams. N-gram co-occurrence scoring is typically performed segment-by segment, where a segment is the minimum unit of translation coherence [16] and the co-occurrence between the machine translation and reference translations is computed for each segment before summing. A higher matching n-gram between the reference translation and the text under study is considered better.

BLUE however has some shortcomings due to the facts that:

- It is overly dependent on the reference texts, whose choice therefore becomes a key factor in determining the success of the method.
- Since the basis is n-gram occurrence, this method is not suitable for all types of questions.

In spite of its drawbacks, BLUE has been used in the Atenea system [17], and used for the evaluation of free text answers.

The partial acceptability of the discussed techniques and the evolved systems that work on those principles being a cause of concern attempts have been made to develop methods and systems that can work on specific requirements of a particular setup. Rein [18], proposed a system to help in evaluation of mathematical problems, while Mu et al. [19] presented an approach for the automatic grading of code assignments. The work by Siddiqi, Harrison, Siddiqi, [5] presents a system called Indus Marker which takes up a particular subject and effectively influences the teaching learning process. It is designed for factual answers in Object Oriented Programming which have a crisp boundary separating the right and wrong answers. Indus Marker is based on structure matching similar to the LSA or BLUE and compares the learners' answer to a predefined answer.

Fig 1. Schematic diagram of the models perception of an answer

III. PROBLEM DEFINATION

The evaluation of essay type answers is a complex task even for the human evaluators. This complexity arises due to subjectivity in judgment about correctness and quality of the content of the essay [16]. The correct grading of such answers is time consuming and due to fluctuations in the application of evaluation criteria on the same answer by different learners' discrepancies naturally creep in. Such discrepancies, resulting out of mostly human factors which are difficult to find and measure are higher when the number of answers is more. To add to the confusion, there exist only few very broad guidelines about how to evaluate a free text based response. The interpretation of the available guidelines is largely subjective in nature and the degree of adherence varies depending on the scenario. A human evaluator evaluating an answer of average quality after evaluating a set of answers of much lower quality may award more marks to the average answer than he would have normally awarded. The same answer, if evaluated after evaluating a set of high quality answers would be graded inferior than what it would have been graded normally [20].

Since questions requiring free text responses are open ended, the learner needs to create an answer for every individual question, making the answers free of any generic structure. It is usually standard practice that evaluators look for coverage of points in an answer. If all points have been covered, with supporting facts and details following the correct writing conventions, the response is graded highest. Evaluators normally set their own rubrics for grading of answers and some may be lenient while others strict.

The current work is built upon the understanding that, an answer to a question is a collection of key points and facts with supporting logical word expressions. A point or fact presented is identified through keywords while the logical word expressions which augment the fact are brought out by the group of words appearing before and after a keyword. These groups of words connect the keywords and bring out the logical as well as factual sense of the answer. The consideration of the relational expressions between keywords is a deviation from the popularly accepted nugget approach, which considers only the keywords as building blocks. In the current work, the choice of the relational expression is not based on the popular n-gram technique but the occurrence of the next keyword. It is also worth mention that unlike other natural language processing approaches, the stop words are not removed from a response as these are considered to be important information carriers. Fig.1 presents the idea of how the answers are perceived by the model [21], which considered the same problem but solved it using a different approach. In their work, the authors used a keyword and its associated pre and post-expressions with fixed four dimensional sense association and extraction in the form of logic, count, certainty and part-of words. The

current work expands the possibilities of sense extraction by allowing the user to have a dynamic semantic network built for every model answer which can have different relations embedded without being limited by the fixed bracket of word and sense types.

IV. PROPOSED METHODOLOGY

The problem that the proposed automated system handles is stated as: *"Given a question Q, its model text based answer M_A and a learner response S_A, the system should be able to evaluate S_A on a scale of [0, 1] with respect to M_A."*

Since the learners' response is in free text and written using a natural language, the evaluation of the same would need knowledge of the subject concerned and the language at least up to the competence level of the learner. Since knowledge representation and language learning are complex issues, the proposed model evaluates the learner response with respect to a model answer framed by a teacher who is a subject expert. The scale of reference is thereby greatly reduced. Formally, the model answer is represented as:

$$M_A = \{KWP, *\} \quad \text{.....(i)}$$

where KWP is a set of keyword phrases and * represents the concatenation operator.

$$KWP = \{KWP_1, KWP_2, ...KWP_n\} \quad \text{.....(ii)}$$

where each KWP_i is a keyword phrase, where a keyword phrase is a keyword along with its associated expressions which may appear on either side of the keywords and links one keyword or concept with another.

$$KWP_i = \{PrP_i, Kw_i, PoP_i\} \quad \text{.....(iii)}$$

where KW_i is the particular keyword and PrP_i and PoP_i are the pre and post expressions associated with KW_i.

$$PrE = \{PrP_1, PrP_2, ..., PrP_n\} \quad \text{.....(iv)}$$

where PrE is the set of all pre-expressions

$$PoE = \{PoP_1, PoP_2, ..., PoP_n\} \quad \text{.....(v)}$$

where PoE is the list of all post-expressions

$$KW = \{KW_1, KW_2, ..., Kw_n\} \quad \text{.....(vi)}$$

KW is the list of all keywords

$$L\_KS = \{L\_KS_1, L\_KS_2, ..., L\_KS_n\} \quad \text{.....(vii)}$$

L_KS is the list of keyword synonyms

The idea is to build a semantic network for every model answer using the concept words and their associated relations with other concept words. To evaluate the students answer, the same has to be compared with the semantic network of the model answer(s) to find an appropriate weighted response from the system.

A sample answer and its model dynamic semantic network is as discussed:

Model Answer: A computer is an electronic device that can store and manipulate data

Fig2: Dynamic Semantic Network for Model Answer
:

The dynamicity lies in the proposition that the relations are not limited to predetermined or stored dictionary values and can vary from one interpretation to another. It is also worth mention here that a concept word can also represent a relation in this dynamic model which is not the case in the classical model of semantic networks.

Table 1. Tabular Representation of Dynamic Network

Other	Keyword	Relation	Keyword	Other
A	Computer	Is an	Electronic Device	
A	Computer	Can	Store	Data
A	Computer	Can	Manipulate	Data
A	Computer	Can Store and Manipulate	Data	

The tabular representation clearly brings out how the keywords 'Store', 'Manipulate' and 'Data' may have different interpretations in different contexts or even in the same answer. Depending on the interpretation or representation of the student the answer has to be evaluated with respect to the model created by the teacher and the student awarded marks.

REFERENCES

[1] Chang, S-H., Lin, P-C & Lin, Z-C., (2007) Measures of partial knowledge and unexpected responses in multiple-choice tests, *Educational Technology & Society*, *10(4)*, 95–109.

[2] Ngee, P., Lau, K., Lau, S.H., Hong, K.S. & Usop, H.,(2011), Guessing, Partial Knowledge, and Misconceptions in Multiple-Choice Tests, Educational Technology & Society, 14 (4), 99–110.

[3] Chakraborty, U.K. & Roy. S., (2010), Neural Network Based Intelligent Analysis of Learners' Response for an e-Learning Environment, Proceedings of *2nd International Conference on Education and Information Technology (ICETC), Vol. 2.* pp. 333-337.

[4] Yee, F.P., National Institute of Education, Singapore (2002), Using short open-ended mathematics questions to promote thinking and understanding, Retrieved from httegy for detection of inconsistency in evaluation of essay type answers, Education and Information Technologies, 19(4), 899-912

[5] Siddiqi, R., Harrison, J. & Siddiqi, R., (2010), Improving Teaching and Learning through Automated Short-Answer Marking, IEEE Transactions on Learning Technologies, 3(3), 237-249.

[6] Lin, J. & 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.

[7] Dessus, P., Lemaire, B. & Vernier, A., (2000), Free-text assessment in a virtual campus, in K. Zreik(Ed.), Proceedings of Third International Conference on Human System Learning 61-76.

[8] Page, E., (1966), The imminence of grading essays by computer, Phi Delta Kappa, 47, 238-243.

[9] Burstein, J., Kukich, K., .Wolff, S., Lu, C. & Chodorow, M., (1998), Computer Analysis of Essays, NCME Symposium on Automated Scoring.

[10] Landauer, T.K., Foltz, P.W. & Laham, D., (1998), An Introduction to Latent Semantic Analysis, Discourse Processes 25(2&3), 259-284.

[11] Foltz, P.W., Laham, D. & Landauer, T.K., (1991), The

Intelligent Essay Assessor : Applications to Educational Technology, Interactive Multimedia Education Journal of Computer Enhanced Learning. On-line journal 1(2).

[12] Kanejiya, D., Kumar, A. & Prasad, S., (2003), Automatic evaluation of students' answers using syntactically enhanced LSA, Association for Computational Linguistics, Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications using Natural Language Processing-Vol. 2, 53-60.

[13] Hoenkamp, E., (2011), Trading Spaces: On the Lore and Limitations of Latent Semantic Analysis, Giambattista Amati & Fabio Crestani,(Ed.), 'ICTIR' , Springer (LNCS), 40-51.

[14] Papineni, K., Roukos, S., Ward, T., Zhu, W., (2002), BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of Association for Computational Linguistics, 311-318.

[15] Doddington, G., (2002), Automatic evaluation of mp://math.unipa.it/~grim/SiFoong.PDF

[16] Shukla, A. & Chaudhary, B.D. (2013), A strategy for detection of inconsistency in evaluation of essay type answers. Education and Information Technologies, Springer, 19(4), 899-912.

[17] 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

[18] Rein, P., (2009), Prospects of automatic assessment of step-by-step solutions in algebra, IEEE Computer Society, Washington, DC, USA. 535-537.

[19] Lingling, M., Xiaojie, Q., Zhihong, Z., Gang, Z. & Ying, X., (2008), An assessment tool for assembly language programming, IEEE, Proceedings of International Conference on Computer Science and Software Engineering, 5., 882-884 .

[20] Escudeiro, N., Escudeiro, P., Cruz, A., (2011), Semi-Automatic Grading of Students' Answers Written in Free Text, The Electronic Journal of e-Learning 9(1), 15-22. Available online at: www.ejel.org.

[21] Chakraborty, U.K., Roy, S. & Choudhury, S., (2014), A Novel Semantic Similarity based Technique for Computer Assisted Automatic Evaluation of Textual Answers, , ICACNI 2014, M. K. Kundu et al. (Eds.), Advanced Computing, Networking and Informatics – Vol. 1, Smart Innovation, Systems and Technologies 27, Springer.

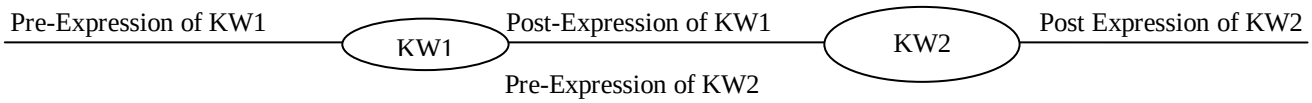

Fig.1. Schematic diagram of the models perception of an answer [21]

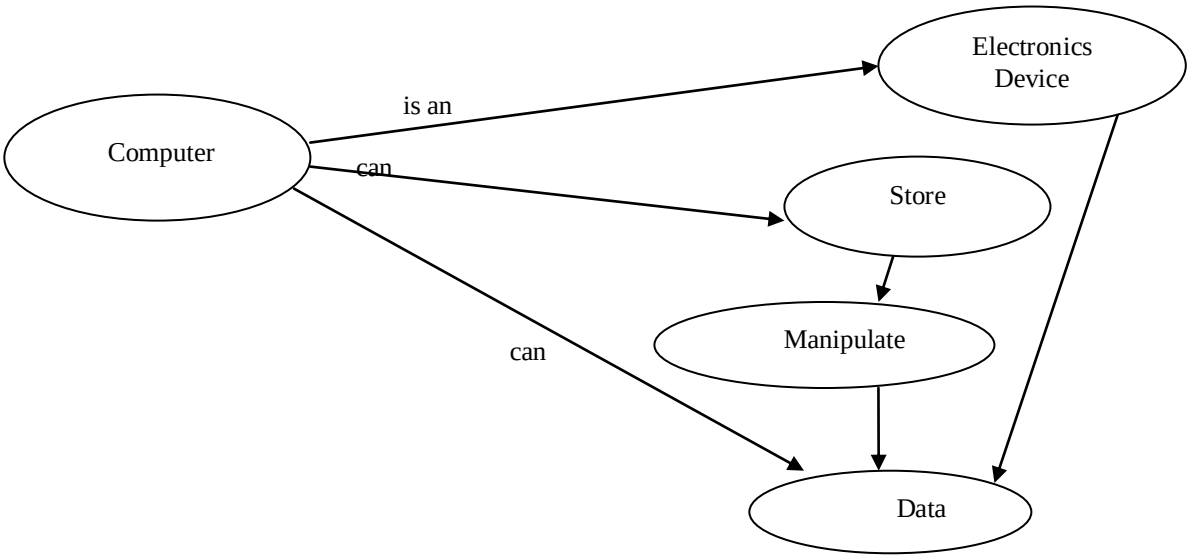

Fig 2: Dynamic Semantic Network for Model Answer