# Automatic Free Text Answer Evaluation using Knowledge Network

Udit Kumar Chakraborty
Dept. of Comp. Sc. & Engg.,
Sikkim Manipal Institute of Technology
Sikkim, India

Sampa Das
Dept. of Comp. Sc. & Engg.,
Sikkim Manipal Institute of Technology,
Sikkim, India

## ABSTRACT
The recent developments in the field of Information and Communication Technology (ICT) has resulted in a major paradigm shift in pedagogy and teaching learning. No longer restricted to the four walls of a classroom, ubiquitous learning is the current trend and universities are gearing up for larger intakes. In the wake of such an upsurge in volume, automatic evaluation of student answers is the need of the hour. The limitations of close ended questions notwithstanding, the popularity of such types are on the rise due to implementation ease. This paper presents a technique to evaluate free text answers of students which will augment and popularize the use of open ended questions for evaluation in online learning.

**Keywords:** Knowledge Network, Free Text, Evaluation, Keywords, Expressions

## 1. INTRODUCTION
Teaching learning has always considered being a part of social development. The teaching learning paradigm or pedagogy however has changed continuously over the years, the latest trend being ubiquitous learning delivered right up to the learner, even while on the move. Though it still consists largely of delivery of Knowledge from the teacher to the learner, the mode has undergone huge transformation. The success of the learning process lies in the application of the gathered knowledge or the skill set estimated. The measurement of the success is through the process of evaluation which therefore becomes a very important aspect of the teaching learning process. The process of evaluation not only judges the learner but also decides upon the success of the learning process and contributes towards the evolution of teaching- learning.

In the present context, with the huge impact of the internet and the associated connectivity, teaching-learning has taken to the web with course delivery taking place across geographical barriers with the teacher and student not even meeting face to face.

Video and audio lectures, online study materials and massive online open classrooms being the standard modes of delivery and learning, evaluation has become even more important and complex. The issue of volume further complicates the problem as Universities now allow larger enrolments. Self-paced learning, open courseware, large volume and lack of personal interaction make evaluation more difficult than before and thus make the system prone to errors. It therefore becomes important to automate the process of evaluating learner response.

Various implementations of automated learner evaluation exist. There are two broad types of question patterns, namely open-ended and close-ended. Due to inherent difficulties in computing with natural languages, the paradigm of automated evaluation gradually drifted towards close-ended questions. However, there have been attempts at evaluating open-ended free text answers since these are the more sought after types owing to the requirements of learner ingenuity. There have been three major approaches, Natural Languages Processing (NLP) based which are computation heavy, keyword centric, which do not evaluate all aspects and the sense extraction based which is where most of the current efforts are concentrated.

The current paper proposes a Semantic Network inspired approach toward the evaluation of free-text answers. The flexibility that the system allows would be limited to the domain of knowledge supported by a Semantic Network. Presently the work is restricted to definition type questions only.

## 2. LITERATURE SURVEY
Question answering has steadily shifted from being inclined towards factoid questions to be popularly accepting descriptive questions (1) 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 (2). In the initial attempts surface feature based assessment (3) of the essay, number of punctuations, number of connectives, average word length etc. Extract to find the correlations between already graded essays and the essays to be graded on syntactic, rhetorical or topical content of the text. After initial syntactic parsing, the system brings out the rhetorical structure of the text based on sentences containing rhetorical arguments. A major drawback of this approach is that it does not consider the sense on the semantic content of the essays and as a result could be gamed by intelligent use of the correct surface features.

Latent Semantic Analysis (LSA) (4) proposed the extraction of word meaning present in a sentence. This technique extracts the word meaning after removal of stop words and builds a matrix to store the frequency of use of every under word to calculate the entropy. Subsequently, the similarity measure between two documents is computed using Single Value Decomposition (SVD).

The Intelligent Essay Assessor (IEA) developed by Foltz, Laham and Landauer, (5)], 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 LSA technique however has a number of drawbacks, the principle being inconsideration towards word sequences. This method overlooks the logic and semantic relational aspects in free-text which are the essence of natural language texts. 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 (6) .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 (6) .

The problems of LSA have been worked upon and reported with modifications in the syntactically Enhanced LSA (SELSA) (7). This approach 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.

Another popular technique followed by some implementations is the BiLingual Evaluation Understudy algorithm (BLUE) (8). An n-gram scoring method this algorithms compares the machine translated output with reference translations using word n-grams.

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 (9), and used for the evaluation of free text answers.

The partial coverage of the types of answers evaluated by the proposed techniques and the lack of universally accepted solution strategy has forced attempts at specific narrow requirements. Rein (10), proposed a system to help in evaluation of mathematical problems, while Lingling et al. (11)presented an approach for the automatic grading of code assignments. The work by Siddiqi, Harrison, Siddiqi, (12) through 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.

# 3. EVALUATING FREE TEXT ANSWERS

The work leading to the current paper tries to evaluate free text answers received as learners' response to questions in an e-Learning scenario. The scenario imposes certain restrictions on the nature of the text received which may be but are not limited to grammar errors and spelling errors. To keep the task simple and avoid distractions, the present paper considers the learner response free of such errors.

To evaluate an answer, the proposed system builds a Knowledge Network, which is a structure similar to a Semantic Network and acts as a connected graph which is a knowledge repository. Such unique repositories would exit for every question and would be constructed with fact and features extracted from model answers submitted by subject expects. The Knowledge Network would also be built by a subject expect.

A model Knowledge Network is shown in Fgure1, for the question "What is a computer?" The model answer to the question is "Computer is an electronic device that can store and manipulate data". The circular nodes in Figure1 are the keywords while the edges bring out the relation between a pair of keywords. The relations, represented by the edge are phases, are directional and are weighted. The weights of the edges and keywords are decided a priory by a subject expert.

The Knowledge Network is stored using a tabular structure which stores the keywords. The directional information is not stored in the tables and considered a default L & R.

**Table1: Tabular representation of sample knowledge representation.**

| KW1 | W | REL | W | KW2 | W |
|---|---|---|---|---|---|
| Computer | | Is an | | Electronics | |
| Computer | | Is an | | Device | |
| Computer | | Can | | Store data | |
| Computer | | Can | | Manipulate data | |
| Electronics | | - | | device | |
| Device | | That can | | Store data | |
| Device | | That can | | Manipulate data | |
| Store | | And | | Manipulate data | |
| Manipulate | | And | | Store data | |
| Store | | - | | Data | |
| Manipulate | | - | | data | |

The weight associated with the keywords and the associated expression, decided by the subject expert are computed in manners appropriate for the level of the learner and showed sum up to the total weight of the question.

The pre-processing part compressing of keyword and associated relation and assignment of weights being manually steps as shown in algorithm Answer–Evaluation

**Algorithm: Answer-Evaluation**

*Input* *Knowledge Network*
*Output* *Score of student*

***Begin***
***While*** *(! End-of-Answer)*
*Search keywords*

```
    If (keyword found)
        Populates List
    End If
End While
For (All- Pairs- Keywords)
    Search Expression
    If (Expression Found)
        Add Score
    End if
End For
Return Score
End
```

## 4. EXPERIMENTATION AND RESULT

The method proposed and detailed here in above was implemented and tested using a dataset of real students of first year engineering in the age group of 18-19 years. The group of 10 students were equally split by gender and the answers were doubly blind evaluated by human evaluators before putting them through the automated marking scheme. Table 2 lists the scores of the students along with the average score of the two human evaluators.

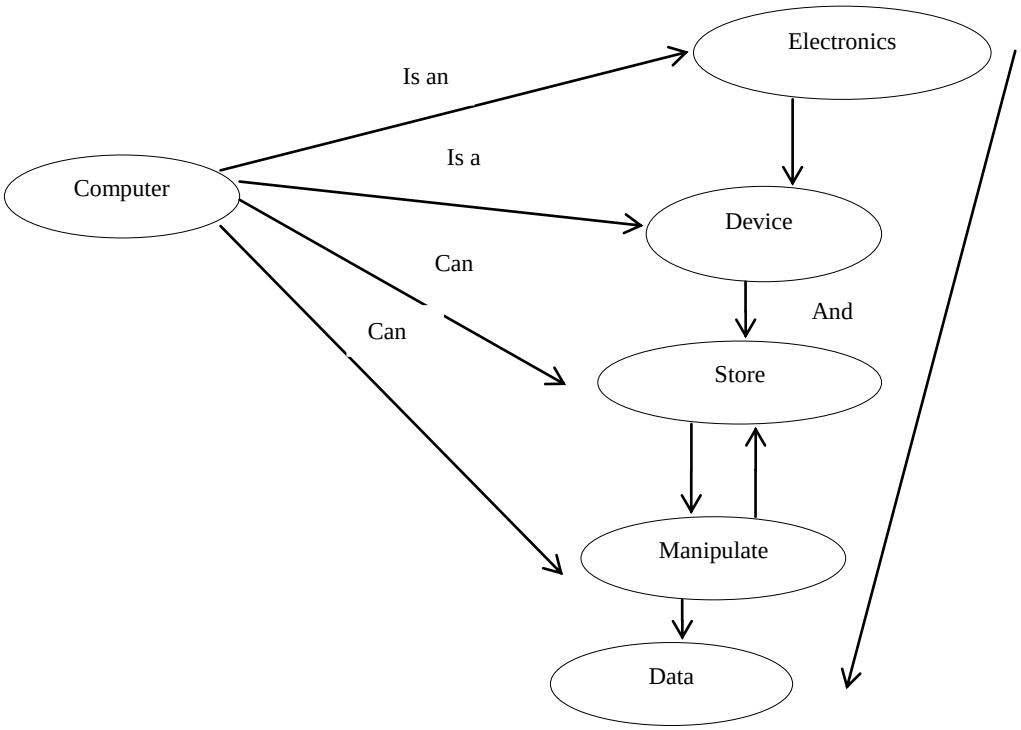


**Figure1: Knowledge Network representation of sample answer**

**Table 2: Comparative Scores**

| Id. | Human Evaluator (Average) | Machine Generated Score |
|---|---|---|
| 1 | 1.3 | 0.6 |
| 2 | 1.3 | 1.3 |
| 3 | 1.0 | 0.7 |
| 4 | 1.3 | 0.7 |
| 5 | 1.0 | 0.6 |
| 6 | 1.3 | 1.0 |
| 7 | 0 | 0.4 |
| 8 | 0 | 0.3 |
| 9 | 1.0 | 0.7 |
| 10 | 2.0 | 0.6 |

The scores returned by the automated evaluation system returns a Pearson correlation of 0.5383 when compared with the average score of two human evaluators.

## 5. LIMITATION AND FUTURE SCOPE

The present system considers building a Knowledge Network, much like the Semantic Networks, for every question that the learner is posed with. This is tedious and relies heavily on human expertise. The Knowledge Network has to be meticulously created considering wider knowledge and not be limited to one or more model answers merely, to significantly match with the human evaluators' world knowledge.

The current work has been tested for definition type questions only and has fairly limited scope. The complete Knowledge Network should consider the use and referencing of synonyms for keywords which the current implementation does not consider. It is also non-receptive towards variations in the expression of relations in a pair of keywords. This limits the application as the learners' are forced towards rote learning.

However, considering the basic skeletal model and the fact that it is an idea that has scope of refinement and scalability, it is acceptable as a prototype model. The results certainly show potential and prospects for growth.

## 6. REFERENCES

[1] Free-Text assesment in a virtual campus. Dessus, P, Lemaire, B and Vernier, A. 2000. K.Zreik(Ed.)proceedings of third Internatonal Conference on Human System Learning. pp. 61-76.

[2] The Imminence of grading essays by computer,47. Page, E.,. 1966. pp. 23-243.

[3] Computer Analysis of Essays. Burstein, J., et al., et al. 1998. NCME Symposium on Automated Scoring.

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

[5] The ntelligent Essay Assessor:Applications to Educational Technology. Foltz, P.W., Laham, D. and Landauer, T.K. 1991. Interactive Multimedia Educational Journal of Computer Enhanced Learning.On-line journal1(2).

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

[7] Automated Evaloutionof students' answers using syntactcally enhanced LSA. Kenejya, D., Kumar, A. and Prasad, S. 2003. Association for Computational Linguastics,Proceeding of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing-Vol. 2. pp. 53-60.

[8] BLUE:a Method for Auomatic Evaluation of Machine Translation. Papineni, K., et al., et al. 2002. Proceedings of the 40th Annual Meeting of Association for Computational Linguistics. pp. 311-318.

[9] Adaptng the automatic assesment of free-text answers to the students. Perez, D. and Alfonseca, E. Loughborough : s.n., 2005. Preceeding of 9th International Computer-Assisted Assessment.

[10] Prospects of automatic assessment of step-by-step solution n algebra. Rein, P. Washington,DC,USA : s.n., 2009. IEEE Computer Society. pp. 535-537.

[11] An assessment tool for assembly language programming. Lingling, M., et al., et al. 2008. IEEE,Proceeding of International Conference on Computer Science and Software Engineering,5. pp. 882-884.

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