Journal of Computer Science

Original Research Paper

High Precision Latent Semantic Evaluation for Descriptive Answer Assessment

[1]Amarjeet Kaur and [2]M. Sasi Kumar

[1]Computer Science and Technology, SNDT Women's University, UMIT, Mumbai, India
[2]Research and Development, Centre for Development of Advanced Computing, Mumbai, India

Article history
Received: 05-05-2018
Revised: 09-10-2018
Accepted: 16-10-2018

Corresponding Author:
Amarjeet Kaur
Computer Science and Technology, SNDT Women's University, UMIT, Mumbai, India
Email: dhariwal.amarjeet@gmail.com

Abstract: This paper proposes an approach to evaluate student's descriptive answers, using comparison-based approach in which student's answer is compared with the standard answer. The standard answers contains domain specific knowledge as per the category (how, why, what, etc.) of questions asked in the examination. Several state-of-art claims that LSA correlates with the human assessor's way of evaluation. With this as background, we investigated evaluation of students' descriptive answer using Latent Semantic Analysis (LSA). In the course of research, it was discovered that standard LSA has limitations like: LSA research usually involves heterogeneous text (text from various domains) which may include irrelevant terms that are highly susceptible to noisy, missing and inconsistent data. We propose a new technique inspired by LSA, denoted as "High Precision Latent Semantic Evaluation" (HPLSE), LSA has been modified to overcome some of the limitations; this has also increased precision. By using the proposed technique (HPLSE), for the same datasets, average score difference and standard deviation between a human assessor and computer assessor has reduced and the Pearson correlation coefficient (r) has increased considerably. The new technique has been discussed and demonstrates on various problem classes.

Keywords: Latent Semantic Analysis, Descriptive Answer, Assessment, Dimension Reduction, Feature Extraction, Evaluation

Introduction

The current system of manual evaluation has some limitations due to which it becomes important to automate the descriptive answers evaluation. It has been noticed that different assessors give different marks to the same response. Additionally, it takes a lot of assessors to evaluate large number of answer sheets.

Evaluation of objective answer is an easy task and well supported by many systems, but descriptive answers evaluation is still an open problem. Various student essays evaluation systems have been under development since 1960s. A National network of US universities supported the development of system to grade essays for thousands of high school students' essays. It scores essays by processing number of essays on the same topic, each scored by two or more human assessors. In 1960s computer technology was not stable enough or accessible enough to expand into large scale.

Some of the systems, such as, Intelligent Essay Assessor, State of essence, Summary Street, Apex, Autotutor and Select-a Kibitzer; though differing in subject domain and the similarities, all are LSA-based. All such systems claim that LSA correlates with the human assessors.

This was one of the motivations of looking at LSA for our research. LSA is a statistical natural language processing (NLP) method for inferring meaning from a text. It was developed by researchers at Bellcore as an information retrieval technique (Deerwester *et al*., 1990) in the late 1980s. LSA provided an advantage over keyword-based methods, which could induce associative meanings of the query (Deerwester *et al*., 1990) rather than relying on exact matches. LSA uses linear algebra techniques to learn the conceptual correlations for a collection of text.

Most of the systems mentioned above and further in section 2, are comparison based in which student's response is compared with standard answer/essay. In a broad view, all such systems have three major modules: student answer representation, standard answer or reference answer representation and the comparison unit. Available systems are useful for essay grading and short answer grading systems, but descriptive answer evaluation system is still an open research issue. Our approach is also comparison based.

After experimenting with LSA for evaluation of students' descriptive answers for various categories of questions, it has been observed that, there is a significant gap between the assessment by human assessor and the results of our computer assessor.

LSA, in general, can be considered as an excellent information retrieval technique, but for this specific task of assessment of students' descriptive answers, the results are not satisfactory. The reason can be that some of the basic features of the technique are not suitable for this problem. LSA has been modified to overcome some of these issues/limitations and the proposed technique, denoted as 'High Precision Latent Semantic Evaluation' (HPLSE) has been used for automation of descriptive answer evaluation process, with much better results.

The paper is organized as follows: Section 2 explains research works related to the field of automation of descriptive answer evaluation using LSA and list of LSA modifications done so far. Section 3 explains methodology used to determine the semantic similarity between two texts. Section 4 proposes a new technique denoted as High Precision Latent Semantic Evaluation (HPLSE) and its implementation with the results. Section 5 lists several issues, conclusion and areas of improvement that future studies will address.

Literature Review

In this section, research work related to the field of Descriptive Answer Assessment (DAA) has been discussed. Methods and techniques implemented so far for automation of DAA process are discussed. Details of LSA technique with the kind of modifications has been done are also mentioned.

Several state-of-the-art short answer graders require manually designed patterns which have to be matched with the student's response; if matched, implies correct response. One of the information extraction-based system (Sukkarieh *et al*., 2005) is developed by the Oxford University to fulfil the need of the University of Cambridge Local Examinations Syndicate (UCLES) as many of the UCLES exam questions are short answers questions. In this system, hand crafted patterns are filtered from the training datasets by human experts and the student responses were matched with these patterns.

Research Work Relevant to the Field of Automation of DAA

Considerable work has been done in the area of using LSA to evaluate essays and to provide content-based feedback, but evaluating descriptive answers is still an open problem.

A text similarity approach was taken in (Kumar and Dey, 2013), for grading short answers without any human interventions unlike previous work. Texts from student answer are compared with the texts of standard answer by applying similarity measures. The standard answer is expanded with the topper part (best matched answers) of the students in next iteration, to increase the adequacy of the standard/reference answer. This issue has been already raised in the introduction section of this report as it's an important aspect of automation of descriptive answer evaluation process as well.

Instead of matching the student's textual answer with the textual patterns in the training dataset, this approach (Da Silva *et al*., 2012) adopted a model in which, the comparisons of student's cognitive structure (concept maps) with reference ontology was used. For comparing the student's concept map with the reference concept map, an alignment tool (COMA++) has been used. The alignment technique for learning assessment is used for the identification of entities with the same meaning i.e. checking the semantic similarity between two entities even when the two strings are not identical.

Online tools that support managing of online assessments such as Moodle and Zoho are based on string matching technique for short answers but long answer evaluation is still handled manually by most of the systems. Some of the approaches are based on keyword matching, sequence matching, quantitative analysis, fuzzy system, rule based system which provides some solution for online assessment of answer sheets, but the general descriptive answer evaluation is still an open problem.

Research Work Related to LSA Technique

LSA, initially proposed as a text search technique, gradually was used to deal with natural language processing tasks like content analysis, document summarization, semantic analysis and patent analysis etc. An improvement to LSA was introduced as Probabilistic Latent Semantic Analysis (PLSA), but according to researchers, in PLSA number of parameters grows linearly with the size of corpus. This leads to problems of overfitting (Zhu and Li, 2012). Another problem with the model is that it is not clear how to assign probability to a document outside of the training set.

Improvements to PLSA lead to LDA (Latent Dirichlet Allocation). Researchers (Zhu and Li, 2012) claimed that LDA provides more intuitive topic model but it has evidently much lower precision values for any case of given parameters and thus the LSA is a better choice for comparative summarization.

This research work (Martínez-Huertas *et al*., 2018) focuses on automatic essay evaluation, specifically on automatic assessment of student's summaries using traditional LSA and inbuilt rubric (a novel LSA). Two conditions are analyzed using inbuilt rubric method: few vs. many lexical descriptors required to accommodate expert rubric and weighted vs. non-weighted method. The weighted method is intended to penalize for irrelevant terms/excess number of terms written by the students. But practically, in DAA negative marking for irrelevant terms are not acceptable. So use of weighted and non-weighted method doesn't contribute much in universities DAA system. Pearson correlation between human expert judgment and inbuilt rubric is 0.79 which is better than

traditional LSA (r = 0.67). A general corpus has been used for training purpose, if we use domain specific corpus then it can increase performance of the inbuilt rubric.

Leonhard and Dai (2009) proposed a topic based multi-document summarization method based on Probabilistic Latent Semantic Analysis (PLSA), in which sentences and queries are represented as probability distributions over latent topics. In this work, researchers have primarily focused on investigating the capability of PLSA approach to model documents from various topics. Researchers evaluated three similarity measures in this approach: The symmetric Kullback-Leibler (KL) divergence, the Jensen-Shannon (JS) divergence and the cosine similarity. They combine query-focused features and thematic sentence features into an overall sentence score. Both PLSA and LSA approaches are implemented for the same data samples but the performance improvements are not significant at $p < 0.05$.

The mathematical technique, Singular value Decomposition (SVD) is applied in LSA for dimension reduction and to eliminate noisy information. In one of the research work (Fallucchi and Zanzotto, 2009), analyses of the effect of SVD feature selection with respect to the baseline are explored. Manual feature selections are compared with the SVD feature selection for validation. It has been concluded that SVD feature selection shows improvement in its performance, but still needs to explore some issues such as: (1) whether SVD feature selection has an positive effect in syntactic features space or not? (2) Are SVD Feature selection is better in comparison with other unsupervised feature selection models in case of probability taxonomy learning?

SVD has also been used to encrypt images (El Abbadi *et al*., 2014) and the decrypted images are close to the original one. SVD can be used for text encryption. The encryption and decryption time of images using SVD is also very promising.

SVD shows improvement in many areas of research and capable of solving various research problems. Many of the research work across various fields exploit LSA, but empirical evidences require more investigations.

Research Work Related to DAA using LSA

LSA, in general, can be considered as an excellent information retrieval technique, but for this specific task of assessment of students' descriptive answers, the results are not satisfactory. After experimenting with LSA for evaluation of students' descriptive answers for various categories of questions, it has been observed that, there is a significant gap between the assessment by human assessor and the results of our computer assessor. The reason can be that some of the basic features of the technique are not suitable for this problem. LSA has been modified to overcome some of these issues/limitations and the proposed technique, denoted as "High Precision Latent Semantic Evaluation" (HPLSE) has been used for automation of descriptive answer evaluation process, with much better results.

Researchers (dos Santos and Favero, 2015), have used LSA for automatic evaluation of written answers where LSA pre-processes the answers using unigrams and bigrams of words. Use of n-gram (n = 1, 2) technique has improved the accuracy of the system. This idea of using n-gram technique instead of traditional Bag of Words technique can be adopted in future work. In this research work, reference answer is considered as a first document in term-document matrix and student answers as the other document. The accuracy of the system is 78.5%.

Researchers (Anirudh *et al*., 2016) have proposed a score recommendation system that works well for descriptive answers with smaller amount of variations from the assessor's perspectives. The method used in this system does not rely on any kind of domain specific corpus. Evaluation score had been calculated on the basis of analysis of student's answers against an answer key. In further work, the feature computations can be improved with the domain specific corpus and can further enhance the accuracy of a system.

Researchers (Thomas *et al*., 2015), have also used LSA for automatic answer assessment and the proposed system assesses the descriptive answers by comparing it with the ideal answer using LSA, positional indexing and spell checking. A word-document matrix is created, where words are collected from the submitted student answers and student descriptive answer are considered as a document. The relevant keywords with the index position are given by the teacher. The order of keywords written by students is compared with the keyword order of ideal answer using positional indexing. Cohen's Kappa method is used to get the strength of agreement between teacher and tool. The results obtained by experimenting with the three different students' datasets are 0.64, 0.73 and 0.61 kappa score. The system fails to handle the cases where most of the students give wrong answers.

The review so far shows that various methods and techniques have been implemented to solve the research problem of automation of descriptive answer evaluation process. Techniques and method such as graphical representation of student answer using LSA, textual representation using PLSA and LSA have been tried. Most of the systems mentioned above have used comparison-based approach, in which students' descriptive answer are compared with standard answer/essay.

The Pearson correlation between human assessor and system are in the range {0.6-0.78.5}, which definitely needs some improvement with the capability of handling exceptional cases. The exceptional cases like when most of the students have written wrong answers in the examination or out of scope answers. There are many such cases which should be handled first to bring descriptive answer evaluation system in a practical field. Available systems are useful for essay grading and short answer grading systems, but descriptive answer evaluation system is still an open research issue.

DOI: 10.3844/jcssp.2018.1293.1302

Methodology

This section covers the overall proposed approach. The proposed system is a comparison-based evaluation system, in which the students' descriptive answer would be compared with the standard descriptive answer.

Generally, a descriptive answer having more than one sentence has a complex structure. A teacher evaluating such an answer looks for a collection of information in the answer as per the category of question asked in the examination. Where and how to find these points depends on the category of question. For example, in "How" type of question answer written in various steps in a process is usually expected in a standard order; altering the steps may change the outcome. But, the list of points in an answer written for "what" and "why" type of questions, permits a more flexible ordering. Keeping this in mind, we attempted to analyze the questions usually asked in examinations and identify categories based on structural and property similarity (detailed description mentioned in Table 1). We briefly discuss our attempt in this regard here.

Syntactic Structure of Descriptive Answers

Various categories of questions are asked by teachers in the exam question paper of universities/institutions. Examples of which are explain, describe, what, why, how, justify, define, elaborate, short notes, comparison based, etc. Some categories of question like, "draw" and "calculate" are excluded from the list because diagrams and mathematical expressions are out of the scope of this research work. After exploring different exam question papers of the universities, a list of categories of questions are formed. Below mentioned table include things expected to be covered in the answer. Analysis of various categories of questions gives clarity about the syntactic structure of descriptive answers and evaluation parameters. The length of the descriptive answer can vary from a phrase to a sentence to a page or multiple pages.

Table 1: Various categories of questions

Sr. No.	Question category	What is to be covered in answer	Syntactic Structure of answer
1	DEFINE	A verbal description of the meaning of some general term	_
2	WHAT	Meaning of the term	_
3	DESCRIBE	It will list some of the properties or feature of a thing/term	Set of text listing some of the features of a concept
4	EXPLAIN	It will relate the thing to a larger context, thereby making it more "understandable"	Some pre and post set of text with the key answer containing detailed information about the topic.
5	WHY	To give a reason for some event	Justifying the reason with the set of sentences containing some pre and post supporting statements.
6	HOW	Steps involved in the process (*in sequential manner)	In a procedure, sentences are dependent on its previous one, in a block(i.e. set of statements)
7	WHEN	Declare the instance	Instance would be the key in a block with the supporting sentences.
8	DISCUSS	It involves examining the various reasons for and against some topic (to make use of some background information surrounding that topic)	Started with some introductory statements (pre-key answer) and then the key answer contain clarification about its usefulness and pitfalls.
9	DISTINGUISH/COMPARE/DIFFERENCE	It involves describing two or more things, emphasizing those aspects where the things are similar or different	Two or more blocks discussing about the differences or similarity based on some features or parameters
10	EVALUATE	It will make use of some criteria for deciding whether one thing is better or worse than another.	_
11	IDENTIFY	It involves recognition skills	_
12	USE/APPLICATIONS	Where to implement	_
13	GIVE/STATE/WRITE/MENTION	Refer to something briefly and without going into detail	Set of text precisely on a topic.
14	LIST/NAME	A number of connected items or names	Set of texts based on same background concept, not necessarily in a sequential manner.
15	CLASSIFY	To arrange or organize according to class or category	Form a cluster based on some common feature.
16	ANALYSE	To discover or reveal a concept/thing through some examination	_

DOI: 10.3844/jcssp.2018.1293.1302

After classifying the questions from exam question paper into different categories, samples of exam answer sheets were collected from the universities/institution. The descriptive exam answer sheets of engineering students which were related to the subjects of computer engineering stream like distributed computing system, artificial intelligence etc. were collected.

Samples of exam answer sheets were evaluated from five different assessors, to check for variation in marks allocation. These, already assessed answer sheets of students, are analyzed with the support of other teachers/assessors in order to understand the psychology of assessor and the way he/she allocate/deduct marks for the answers written by the students in the examination.

Some assumptions have been taken for simplifying the task of automation, such as, sentences in student's descriptive answer are assumed to be grammatically correct, with no spelling mistakes. Only textual answers are considered. Diagrams and mathematical expressions are out of the scope of this research work.

Proposed Approach

The proposed comparison-based approach determines the similarity between student and standard descriptive answer. Broadly, the system has three major modules: the standard answer representation, the student answer representation and a comparison unit (as shown in Fig. 1).

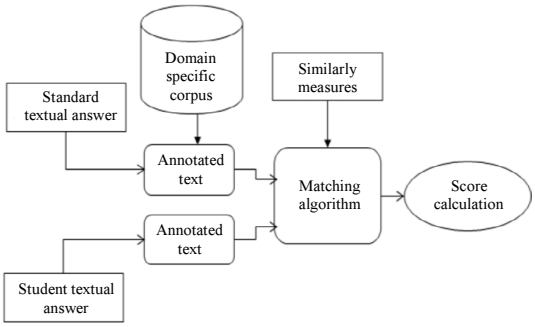

Fig. 1: The broad approach

Standard textual answer: The standard textual answer is the precise answer written by a domain expert or a teacher.

Student textual answer: The samples of student textual answer used for this experiment are free-form text and are in a range of 5-6 grammatically correct English sentences (approx. 80-100 words).

Domain specific corpus: The domain specific corpus includes data from various e-resources and textbooks related to that domain.

Steps to create domain specific corpus:

Step1: Collect domain related textual data from various e-resources and textbooks.

Step2: Analyze textual data by calculating frequency count of all the unigrams, bigrams and trigrams occurred in the domain related textual data using text analyzers (http://online-utility.org/text/analyser.jsp) it's a Free software utility which allows to find out the most frequent phrases and frequencies of it. Non-English language texts are supported. It also counts number of words, characters, sentences and syllables and calculates lexical density.

Step4: If frequency of a keyword is beyond threshold level {Hfv -high frequency value keywords (generally list of stop words) and Lfv - low frequency value (rare keywords don't contribute in defining meaning of a concept)} then it should be discarded from the corpus (please refer Fig. 2).

Step5: After filtrations, use this domain specific corpus for HPLSE algorithm.

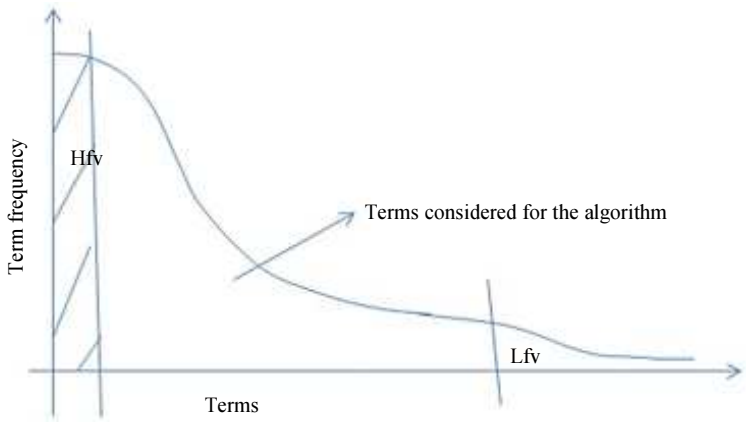

Fig. 2: Generation of domain specific corpus

DOI: 10.3844/jcssp.2018.1293.1302

Matching algorithm: Description about the matching algorithm using HPLSE technique is explained in section 4 of this research paper. In the below mentioned block diagram shown in Fig. 3, LSA and HPLSE techniques are explained.

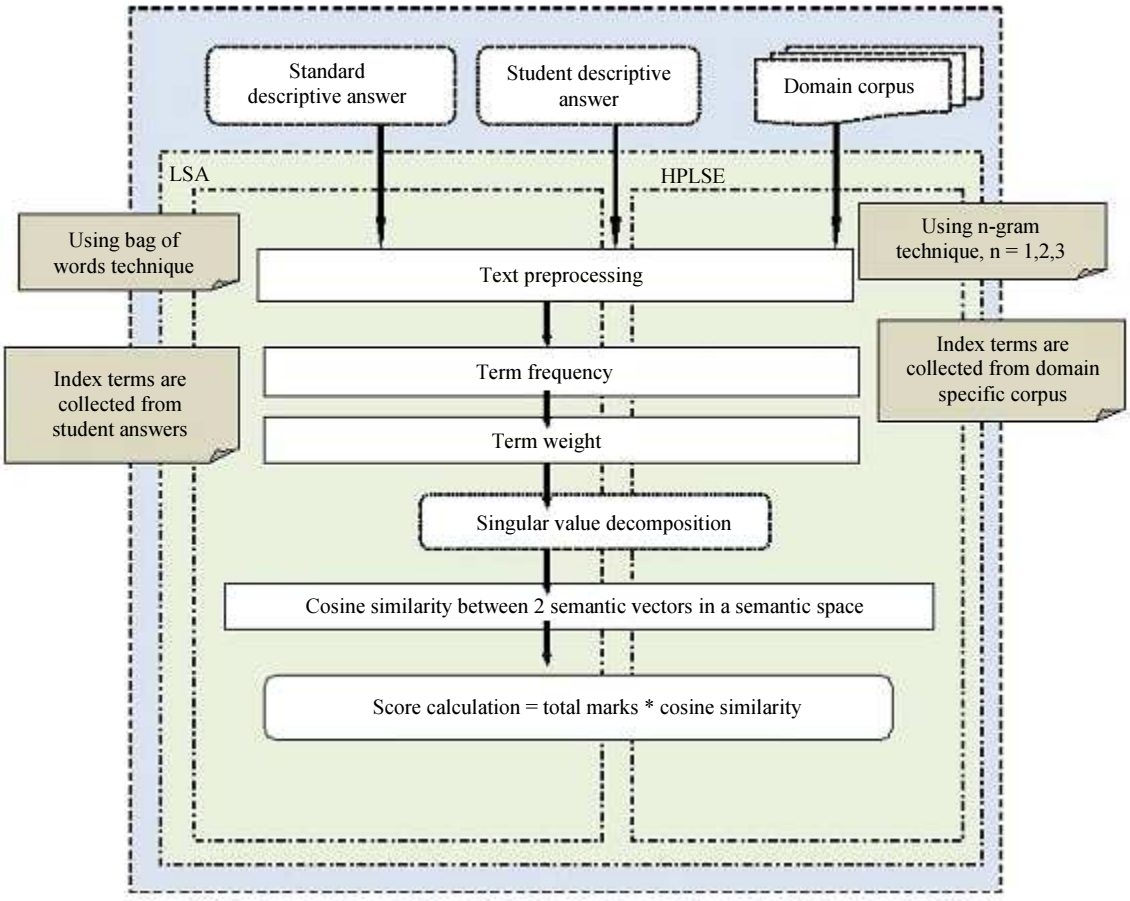

Fig. 3: Block diagram-comparing LSA and HPLSE technique

The basic differences between LSA and HPLSE are discussed (as mentioned in Table 2) as follows:

Table 2: LSA v/s HPLSE

S no.	LSA	HPLSE
1	Bag of Words technique is used. Assumption: Each word meant only one concept and each concept was described only by one word and words are assumed to have only one meaning.	N-gram technique where N=1,2,3 Unigram, bigram and trigram are collected from the corpus.
2	Similarity through co-occurrences of words across the documents.	Similarity through terms from domain specific corpus.
3	Relevance check through query matrix where, query matrix is formulated using various algorithms**.**	Relevance check through standard answer given by human expert.
4	LSA research usually involves heterogeneous text, general corpus (corpora size>20k words, 20k passages). Heterogeneous text may include irrelevant terms, highly susceptible to noisy, missing and inconsistent data.	Consider narrow domain or domain specific corpus consequently, reduces polysemy
5	LSA has high recall but less precision. The precision declines because of spurious co-occurrences.	High recall and high precision

DOI: 10.3844/jcssp.2018.1293.1302

HPLSE (High Precision Latent Semantic Evaluation) Technique

A modified algorithm (A modified version of LSA) has been introduced as High Precision Latent Semantic Evaluation (HPLSE).

HPLSE is a technique in natural language processing derived from LSA, for finding the semantic similarity between the students' descriptive answers and the standard answer.

In contrast with LSA, index words (rows in term-document matrix) are collected from domain specific corpus and not from the documents pool or paragraphs. This modification expected to increase the precision and recall of HPLSE for evaluating student descriptive answers.

In LSA, the frequently used words in the answer become part of index terms pool. So when a large number of students would have written wrong descriptive answers, the irrelevant index words would become part of the index terms pool, resulting into false outcomes. Such problems have been rectified using HPLSE.

High Precision Latent Semantic Evaluation (Modified version of LSA)

The steps of applying HPLSE for automated assessment of descriptive answer are:

Step 1: HPLSE begins with the construction of a term-document matrix *X*. Determine the unigram, bigram and trigram from the domain specific corpus collected from various e-resources and place in the rows of term-document matrix, *X*.

Step 2: Consider all unigrams, bigrams, trigrams from domain specific corpus as rows in a term-document matrix and all students' descriptive answers as documents.

In a term-document matrix (X), each unigram, bigram and trigram are represented by a row (i = 1,2,3……m) and each student descriptive answers is represented by a column (j = 1,2,3……n), with each matrix cell, initially representing the number of times (term frequency, tf_{ij}) the associated term appears in the student descriptive answer.

Step 3: Construct a t x 1 query matrix q, by considering terms from standard descriptive answer and calculating term frequency of each term as a cell value of query matrix.

Step 4: Weight each entry tf_{ij} in X using TF-IDF (Term Frequency- Inverse Document Frequency) weight function. Weight function is used to determine the importance of the each term. The new weighted matrix is X_w.

Step 5: Singular value decomposition (SVD) is applied on the matrix X_w to decompose matrix X_w into three other matrices, an ***m*** by ***r*** term-concept vector matrix (U), an ***r*** by ***r*** singular values matrix (S), ***r*** by ***n*** concept-document vector matrix (V^T), which satisfy the following relations:

$$X_{w'} = USV^T \tag{1}$$

Step 6: Choose an optimum dimension k to reduce $X_{w'}$:

$$X_{w'k} = U_k S_k V_k^T \tag{2}$$

where, U_k and VT_k matrices, define the term and document vector spaces. Dimension reduction is used to keep the important information, while reducing the noisy data from the dataset by setting less important dimensions to zero.

A rule of thumb for finding out the optimum value of k is to retain enough singular values to make up to 90% of the energy in S, i.e., the sum of the squares of the retained singular values should be at least 90% of the sum of the squares of all the singular values (http://infolab.stanford.edu/~ullman/mmds/ch11.pdf).

To find out the optimum dimension value (k), use this empirical formula:

$$\frac{\sum_{i=1}^{r-j} S_{ii}}{\sum_{i=1}^{r} S_{ii}} \geq 0.9 \; then \; remove \; S_{(r-j+1)}$$

where, j = 1, 2, 3…r r = number of singular values in the singular matrix S.

Number of optimum dimension is typically on the order 100 to 300 dimensions in LSA, but HPLSE also works well in less number of dimensions.

Step 7: Compute,

$$V' = S_k V_k^T \tag{3}$$

V', the reduced weighted frequency documents.

Step 8: Compute the query matrix q':

$$q' = q^T U_k \left(1 / S_k\right) \tag{4}$$

Step 9: Compute the cosine similarity,

$$Cos\left(q', V'\right) = \left(q'.V'\right) / \left(|\, q' \,|\, x \,|\, V' \,|\right) \tag{5}$$

where, Cos (q', V') -> similarity match between student answer and standard descriptive answer.

Step 10: Marks awarded by computer assessor (CA):

$$CA\,score = total\,marks\,x\,Cos\left(q', V'\right) \quad (6)$$

Implementation

The data for this experiment consisted of student's answers (1440 samples) in electronic form. The samples of student descriptive answer used for this experiment are free-form text and are in a range of 5-6 grammatically correct English sentences (approx. 80-100 words). Same set of student descriptive answers are used for both the techniques - LSA and HPLSE.

The general steps of HPLSE technique are implemented using python 2.7 and the steps are: (Please refer section 4.1 for detailed description of each step of algorithm).

Step 1: Consider all the students' descriptive answers as documents in term-document matrix (*for classification of documents*) and a standard answer as a query matrix.

Step 2: Determine the unigram, bigrams and trigrams from the domain specific corpus collected from various e-resources (please refer section 4.1) and place in the rows of term-document matrix.

Step 3: Create the frequency count matrix.

Step 4: Modify the frequency count matrix, by applying TF-IDF (Deerwester *et al*., 1990) weight function to each cell of the term-document matrix.

Step 5: Apply Singular value decomposition; decompose the term-document matrix.

Step 6: Dimension reduction (to reduce the noisy data).

Step 7: Calculate cosine similarity between two vectors.

Comparing Results of LSA and HPLSE Technique

Three performance measures are used to analyse the efficiency of the technique such as average score difference (ASD), Standard Deviation (SD) and a Pearson Correlation (PC) between computer assessor (marks calculated using LSA and HPLSE technique) and HA (as shown in Table 3-5).

Table 3: Performance analysis of LSA and HPLSE technique by comparing average score difference (HA-CA)

S no.	Marks allocated	Sample size	Domain	Question category	LSA technique (ASD)	HPLSE technique (ASD)	LSA % difference	HPLSE % difference
1	5	320	Computers	Why	1.73	0.83	34.6	16.6
2	5	320	Computers	What	2.36	0.56	47.2	11.2
3	1	400	Electronics	What	0.23	0.11	23.0	11.0
4	2	400	Electronics	Write	0.76	0.48	38.0	24.0

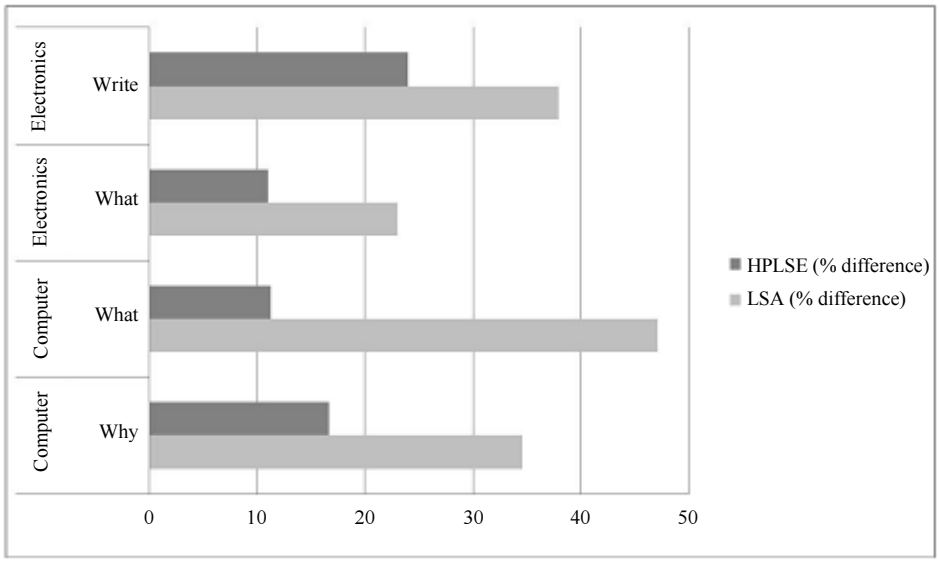

Fig. 4: Average score Difference (HA-CA), for different category and domain of questions

Table 4: Performance analysis of LSA and HPLSE technique, by comparing its standard deviation

S no.	Marks allocated	Sample size	Domain	Category of question	LSA Technique SD	HPLSE Technique SD
1	5	320	Computers	Why	1.07	0.61
2	5	320	Computers	What	1.45	0.38
3	1	400	Electronics	What	0.17	0.14
4	2	400	Electronics	Write	0.64	0.36

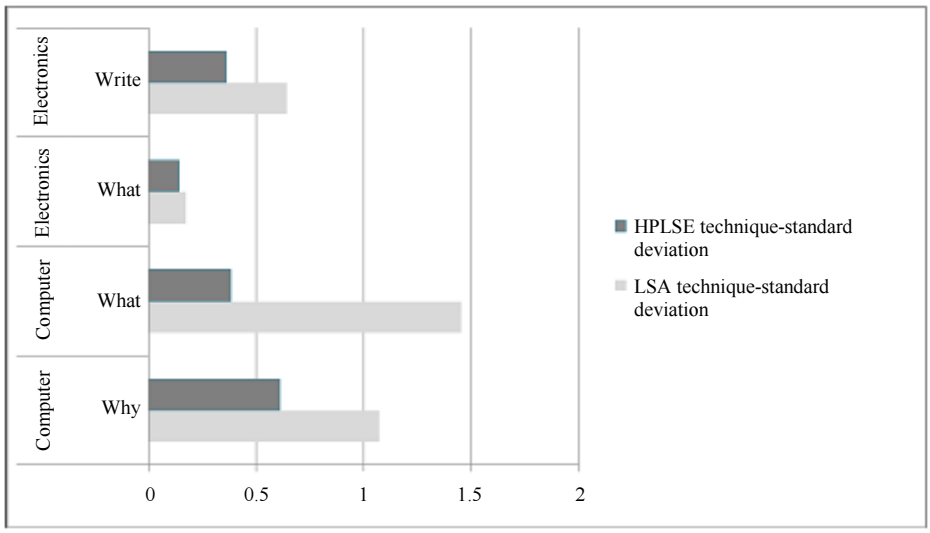

Fig. 5: Standard deviation, for different category and domain of questions

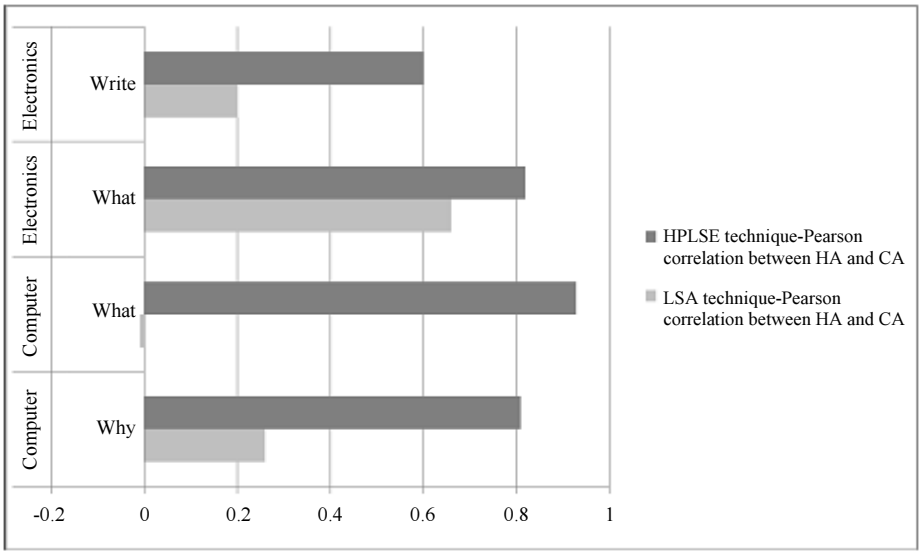

Fig. 6: Pearson Correlation between HA and CA, for different category and domain of questions

Table 5: Performance analysis of LSA and HPLSE technique, by comparing its Pearson Correlation

S no.	Marks allocated	Sample size	Domain	Question category	HPLSE Technique (PC between HA and CA)	LSA Technique (PC between HA and CA)
1	5	320	Computers	Why	0.81	0.26
2	5	320	Computers	What	0.93	-0.01
3	1	400	Electronics	What	0.82	0.66
4	2	400	Electronics	Write	0.60	0.20

The results shown in Fig. 4-6 signify a significant improvement in HPLSE performance over LSA.

Conclusion and Future Scope

In this research work, automation of descriptive answer evaluation process has been tried and initially standard LSA was used for the same. But, the results were not satisfactory. Some modifications were incorporated, keeping in mind the context of automated assessment of descriptive answers. This modification has been introduced as HPLSE (High Precision Latent Semantic Evaluation) and the results revealed a significant improvement over LSA. By using HPLSE technique for the same datasets, average score difference and standard deviation between a human and

computer assessor has reduced (please refer Fig. 4 and 5) and the Pearson correlation coefficient (r) has increased considerably (please refer Fig. 6). The reasons for improvements are:

- Ability of a system to retrieve the relevant and reject the irrelevant phrases from the student's descriptive answer
- Precise relevance check according to human assessor perception provides high precision
- Pruning of extra terms from the corpus, reducing polysemy

In Future studies, categories of question like How, compare, draw, evaluation of mathematical expression may be tried.

Acknowledgement

We would like to thank Dr. Sanjay S. Pawar, Principal, Usha Mittal Institute of Technology, SNDT Women's University, Mumbai, for guidance and support.

Funding Information

This research was supported by The Department of Science and Technology, Ministry of Science and Technology, under Women Scientist Scheme (WOS-A) – SR/WOS-A/ET-1064/2014(G).

Author's Contributions

Amarjeet Kaur and Dr. M. Sasikumar: Conception, Design, Data collection/processing, Data analysis/interpretations, Literature review and Writer.

Dr. M. Sasikumar: Supervision/Mentor and Critical review.

Ethics

We confirm that we have read guidance on competing interests and will assure that none of the authors have any competing interests in this manuscript entitled "High Precision Latent Semantic Evaluation for Descriptive Answer Assessment".

References

Anirudh, K., S. Sachin, A.R. Deshpande, D. Jawahar and S. Gowri, 2016. A score recommendation system towards automating assessment in professional courses. Proceedings of the IEEE 8th International Conference on Technology for Education, Dec. 2-4, IEEE Xplore Press. Mumbai, India, pp: 140-143. DOI: 10.1109/T4E.2016.036

Da Silva, A.A., N. Padilha, S. Siqueira, F. Baiao and K. Revoredo, 2012. Using concept maps and ontology alignment for learning assessment. IEEE Technol. Eng. Educ., 7: 33-40.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman, 1990. Indexing by latent semantic analysis. J. Am. Society Inform. Sci., 41: 391-407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

dos Santos, J.C.A. and E.L. Favero, 2015. Practical use of a Latent Semantic Analysis (LSA) model for automatic evaluation of written answers. J. Brazilian Comput. Society, 21: 21-21. DOI: 10.1186/s13173-015-0039-7

El Abbadi, N.K., A. Mohamad and M.A. Hameed, 2014. Image encryption based on singular value decomposition. J. Comput. Sci., 10: 1222-1230. DOI: 10.3844/jcssp.2014.1222.1230

Fallucchi, F. and F.M. Zanzotto, 2009. Singular value decomposition for feature selection in taxonomy learning, Proceedings of the International Conference on Recent Advances in Natural Language Processing, (NLP' 09), Borovets, Bulgaria, pp: 82-87.

http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

http://online-utility.org/text/analyser.jsp

Kumar, N. and L. Dey, 2013. Automatic quality assessment of documents with application to essay grading. Proceedings of the 12th Mexican International Conference on Artificial Intelligence, (CAI' 13).

Leonhard, H. and L.T. Dai, 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. Proceedings of the Recent Advances in Natural Language Processing, (NLP' 09), Borovets, Bulgaria, pp: 144-149.

Martínez-Huertas, J.Á., O. Jastrzebska, A. Mencu, J. Moraleda and R. Olmos *et al*., 2018. Analyzing two automatic assessment LSA methods (Inbuilt Rubric Vs. Golden Summary) in summaries extracted from expository texts. Psicologia Educativa, 24: 85-92.

Sukkarieh, J.Z., S.G. Pulman and N. Raikes, 2005. Auto-marking 2: An update on the UCLES Oxford University research into using computational linguistics to score short free text responses. Proceedings of the 30th Annual Conference of the International Association for Educational Assessment, (AEA' 05).

Thomas, N.T., K. Ashwini and B. Kamal, 2015. Automatic answer assessment in LMS using latent semantic analysis. Proc. Comput. Sci., 58: 257-264. DOI: 10.1016/j.procs.2015.08.019

Zhu, T. and K. Li, 2012. The similarity measure based on LDA for automatic summarization. Proc. Eng., 29: 2944-2949. DOI: 10.1016/j.proeng.2012.01.419