Automatic short answer grading using rough concept clusters

Udit Kr. Chakraborty* and Debanjan Konar

Department of Computer Science and Engineering,
Sikkim Manipal Institute of Technology,
Sikkim (E), India
Email: udit.kc@gmail.com
Email: konar.debanjan@gmail.com
*Corresponding author

Samir Roy

Department of Computer Science and Engineering,
National Institute of Technical Teacher's Training and Research,
Salt Lake City, Kolkata, India
Email: samir.cst@gmail.com

Sankhayan Choudhury

Department of Computer Science and Engineering,
University of Calcutta,
Salt Lake City, Kolkata, India
Email: sankhayan@gmail.com

Abstract: Evaluation of text-based answers has stayed as a challenge for researchers in recent years and with the growing acceptance of e-learning system, a solution needs to be achieved fast. While assessing the knowledge content, correctness of expression and linguistic patterns are complex issues in themselves, a smaller answer may be evaluated using keyword matching only. The work proposed in this paper is aimed at evaluating smaller text answers, no longer than a single sentence using keyword matching. The proposed method agglomerates keywords from a group of model answers forming clusters of words. The evaluation process thereafter exploits the inherent roughness of the keyword clusters to evaluate a learners' response through comparison and keyword matching. The novelty in the proposed system lies in the usage of fuzzy membership functions along with rough set theory to evaluate the answers. Rigorous tests have been conducted on dataset built for the purpose returned good correlation values with the average of two human evaluators. The proposed system also fares better than latent semantic analysis (LSA) based and link grammar-based evaluation systems.

Keywords: text answer; single sentence; keyword; concept cluster; rough set; latent semantic analysis; LSA; link grammar.

Reference to this paper should be made as follows: Chakraborty, U.K., Konar, D., Roy, S. and Choudhury, S. (xxxx) 'Automatic short answer grading using rough concept clusters', *Int. J. Advanced Intelligence Paradigms*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Udit Kr. Chakraborty is an Associate Professor in the Department of Computer Science and Engineering at the Sikkim Manipal Institute of Technology, Sikkim (E), India. He has published a number of high quality research articles in journals of repute and national and international conferences. He has co-authored a book on soft computing and is involved in active research in the field of machine learning and education technology.

Debanjan Konar is an Assistant Professor in the Department of Computer Science and Engineering at the Sikkim Manipal Institute of Technology, Sikkim (E), India. He is currently pursuing his Doctoral Research from Indian Institute of Technology, Delhi and an active Researcher in the area of soft computing and quantum computing.

Samir Roy is a Professor of Computer Science and Engineering working with the National Institute of Technical Teacher's Training and Research, Kolkata, India. He has authored and co-authored a number of highly rated research papers and also published a book on soft computing. His areas of research include education technology, soft computing, and machine learning. His teaching interests include computer algorithms, data structure, artificial intelligence and formal languages and automata theory.

Sankhayan Choudhury is a Professor of Computer Science and Engineering at the University of Calcutta, Kolkata and has authored a large number of highly rated and cited research papers. His areas of interest include computer networks, distributed computing, soft computing and bioinformatics.

1 Introduction

In this era of information and communication technology (ICT)-based education, assessment of learners' answers under e-learning environment remains a daunting task. To reduce effort, most e-learning systems use objective type questions to evaluate students. Objective type questions, namely multiple choice, matching pairs or single word responses do not require understanding the text. Subjective answers, also otherwise referred to as free-text answers, on the other hand, allow students to express their views and support their ideas while answering the question (Pintso et al., 2010). The objectivity and quantifiability being reasons behind the popularity of objective type or close-ended questions, their limitations arise out of the fact that close-ended questions fail to check the learners' knowledge and understanding of the proof and theoretical aspects (Chang et al., 2007). Further, a learner during the test may be in a state of full knowledge, partial knowledge or absence of knowledge, partial misconception or full misconception. Close-ended 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 as it is also possible to score in such tests using pure guess work (Ngee et al., 2011).

On the other hand, the evaluation and grading of subjective questions is expensive, time consuming and prone to measurement errors. Errors may creep in due to faulty judgment, fatigue, bias or inconsistencies in the grading process. The complexities are increased manifold by large enrollments in universities due to the emergence of e-learning systems as ubiquitous education platforms. There also occur certain scenarios

where a student needs an assessment even when an instructor is not available (Griff and Matter, 2013). It is under such circumstances and for such reasons that automatic assessment of learners' response has become more relevant in recent times. Text-based responses may be short answers or essays, and each has to be evaluated differently. Normally, automatic short answer grading (ASAG) systems evaluate and mark a learners' response by comparing it with one or more correct answers (Mohler and Mihalcea, 2009). The growth of ASAG systems had been slow, largely due to the inherent computational complexities involved in processing natural languages. However, recent advances in computational linguistics have opened up the possibility of being able to automate the marking of free text responses typed into a computer without having to create systems that fully understand the answer (Pulman and Sukkarieh, 2005).

The current work presents a rough set inspired system that uses *hierarchical agglomerative clustering* for evaluation of users' textual response. The system has been designed keeping in view, the on-line examination systems under an e-learning environment. While definitions of short answers state that these vary in length from a single word to a couple of sentences, the current work obeys the following five point criteria. First, the question must require a response that recalls external knowledge instead of requiring the answer to be recognised from within the question. Second, the question must require a response given in natural language. Third, the answer length must be restricted to a single sentence. Fourth, the assessment of the responses should focus on the content instead of writing style. Fifth, the level of openness in open-ended versus close-ended responses should be restricted with an objective question design (Burrows et al., 2015). The technique is restricted to evaluate single sentence answers which are classified as text-explicit (TE) and do not require inference to be drawn between two or more text segments (Hermet et al., 2006). Single sentence answers have been considered as these are of pedagogical significance as these do not allow the learner the flexibility to answer elaborately.

The presented technique considers lexical relations brought out through the surface syntactic constituents of a cluster of reference answers through the use of keywords. It does not consider the word meaning recorded in any knowledge base or general ontology. A group of model or reference answers are required to be analysed and clustered prior to the assessment process. The learners' response is thereafter evaluated with respect to the cluster it most closely resembles.

The organisation of the article is as follows. A compact literature survey of the articles relevant to automated question answering is provided in Section 2. Section 3 covers the basic theoretic fundamentals of hierarchical agglomerative clustering and its necessary operations. Section 4 explains the foundations of rough sets necessary for the proposed work. Section 5 elucidates the primary objective of research work of this article. This section throws light into the pre-processing of model answers. In this paper, to evaluate learner responses', a rough inspired fuzzy membership function is proposed. The performance analysis of the proposed rough concept cluster-based approach reported in Section 6. In addition, the statistical correlation of the proposed technique with the existing techniques is incorporated. Section 7concludes this paper with new horizons of research in future.

2 Literature survey

Automated question answering has over the years evolved from being of multiple choice question type to factoid answers to be popularly accepting descriptive questions (Lin and Fushman, 2005). Two major types of approaches have been popularly taken, the first being free-text assessment based on surface features and later free-text assessment based on course content (Desus et al., 2000). There also are conceptual differences between the approaches. While some systems are tolerant towards sentence construction, grammar and spelling errors and are concentrated at extracting the sense of the answer, others are designed to be as close to the human evaluator (HE) as possible (Ziai et al., 2012).

The earliest attempts towards automated assessment of text in natural language using computation methods dates back to Page (1966) where the length of 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. This approach to essay grading was followed by e-rater (Burstein et al., 1998) which also extracts correlations between already graded essays and ungraded ones using about 60 surface features. Some more notable work on essay assessment includes *intelligent essay assessor (IEA)* (Foltz et al., 1991), which uses *latent semantic analysis (LSA)* (Landauer et al., 1998). Several systems built based on the LSA method to find the similarity between the student answers and the standard correct answer showed several problems correlated with the use of *LSA* (Omran et al., 2014), as *LSA* paid no attention to word order; and much less attention to sentence structure.

Automated short answer grading (ASAG) systems however are different from essay assessors as these compare students' responses to questions with manually defined target responses or answer keys in order to judge the appropriateness of the responses, or in order to automatically assign a grade (Ziai et al., 2012). There exist different approaches to building such systems to grade students depending on the perspective. While some are developed to check the knowledge acquired by the student, others check language learning and a third variant evaluates language comprehension. One of the earliest such system is *Apex* (Desus et al., 2000), a web-based learning application system which rates the learners' response in free text with reference to answers already stored in the system database. *Apex* uses *LSA* to measure the semantic similarity value and returns a textual evaluation on how closely a notion was covered from the score returned.

A prominent approach for *ASAG* has been computing the semantic similarity between the students answer and the model answer (Zhang et al., 2013). Such approaches have been used in *C-rater* (Leacock and Chodorow, 2003) and also by Mohler and Mihalcea (2009) and Mohler et al. (2011). *C-rater* works on model building from model answers through the extraction of key concepts using natural language processing (NLP) techniques. The students answer is graded based on the coverage of these key concepts (Pulman and Sukkarieh, 2005). These approaches, being corpus based, build a semantic space from the word distribution in the corpus of unannotated text and measure the belongingness of the student answer to that space (Chakraborty et al., 2014c). The most notable approach in this field is *explicit semantic analysis (ESA)* (Gabrilovich and Markovitch, 2009) which measures how strongly a given word in the input text associates with the semantic space created by model answers and can find semantic relatedness in spite of not having common words. However, polysemy is a major problem with semantic similarity-based approaches as the context of the word is not taken into consideration.

On the contrary, in the knowledge-based approach, the methods applied try to extract relations between words and encode them instead of just looking at the words as information carriers. While some proposed methods leveraged *Wordnet* (Budanitsky and Hirst, 2006), others relied on collaboratively built lexical resources (Hovy et al., 2013). While using such resources ease implementation, the volume is sometimes overkill and some approaches try to build a domain specific knowledge network for the purpose of evaluation (Chakraborty and Das, 2015). However, manually building and maintaining lexical or knowledge resources is time consuming and expensive and has been criticised for having limited domain (Zesch and Gurevych, 2010). This domain specific technique has been used in Chakraborty et al. (2014c) and Paden and Chakraborty (2015) with mixed results. Some recent approaches (Jadidinejad and Mahmoudi, 2014) have even harnessed Wikipedia to build an associative network of concepts for *ASAG*. However, since the acceptance of Wikipedia as a dependable knowledge source is debatable and would not be recommended by universities, the method is better applied for non-formal education.

Some *ASAG* systems like *Atenea* (Perez et al., 2005) and the one presented by Selvi and Banerjee (2010) have used an enhanced version of the *bilingual evaluation understudy algorithm (BLEU)* (Papineni et al., 2002), which is an n-gram scoring method that compares the machine translated output with reference translations using word n-grams. When used for *ASAG*, it assesses a text by computing a score based on explicit word-to-word match between the students' answer and teachers' answer. The students answer is evaluated based on whether it returns an exact match, stemmed word match, synonym match, acronym match, etc.

BLEU 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.

The lack of acceptance of a single method for *ASAG* has resulted in efforts from researchers to come up with appropriate solutions tailor made for given situations. Ros'e et al. (2003) proposed *Carmel ITC* to evaluate student responses and dealt with average response length of 48 words. While the average length of English sentences range from 20 to 22 words, *Carmel ITC* deals with average two sentence answers. This system was designed to perform text classification using naive Bayes text classifier and decision tree using a bag of words approach. In a 50-fold cross-validation experiment with one physics question, six classes and 126 student responses, hand-tagged by two annotators, *Carmel ITC* reaches an *F-measure* value of 0.85. An interesting artificial intelligence-based approach has been reported by Makatchev and VanLehn (2007), which uses first order predicate logic representations for target and learner responses and match them for similarity. Of the 16 semantic labels used, the highest configuration yielded an *F-measure* of 0.4974. In their work, Kumaran and Sankar (2015) present an ontology-based approach for short answer evaluation which returns a correlation of 0.8 with HE for 30 students on seven (7) questions. This approach used a *NLP*-based parser, subject extractor and ontology extractor for comparing the students answers with model answer. In spite of such volume of reported literature, 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 (Siddiqi et al., 2010) and commercially available systems still use

paraphrase matching or regular expressions (Gutierrez et al., 2010). A recent work by Mittal used a hybrid technique bringing in and amalgamating *LSA* and *BLEU* (Mittal and Devi, 2016). In their work, Sultan et al. (2016) used a semantic similarity-based technique to evaluate short answers. However, as the authors admit "There is, however, immense scope for improvement" (Sultan et al., 2016).

The *ASAG* technique presented in this paper uses a bag of words approach towards achievement of learner score for an answer. The initial model answer creation apart, the presented method is free of human supervision and since it considers a set of words, can be classified as a corpus-based technique which is lightweight since it does not need elaborate NLP. The use of clustering of similar answers, as used in the proposed technique is a popular approach, and has been used by other researchers like Jing (2015) and Basu et al. (2013).

3 Hierarchical agglomerative clustering

Hierarchical agglomerative clustering (Lior and Maimon, 2005) is a bottom-up clustering technique where each cluster has a hierarchy of clusters, as the name suggests. This technique starts with a single object cluster which iteratively grows through the agglomeration of the closest pair of clusters through some pre-defined similarity criteria thereby forming a nested structure, of clusters generated earlier, lying within those generated later. The small clusters that are generated through this process are helpful in data discovery and also produce an ordering of objects which is informative for data display. However, a major disadvantage is that an incorrectly grouped object cannot be relocated.

The process of hierarchical agglomerative clustering may be summed up through the following steps:

1 Begin
2 Assign each object to a separate cluster
3 Repeat until number of cluster reduced to one (01)
  3.1 Evaluate all pair-wise distances between clusters
  3.2 Build a distance matrix
  3.3 Search for pair with shortest distance
  3.4 Remove shortest distance pair from matrix and merge them
4 End

The work presented through this paper user hierarchical agglomerative clustering technique using *Jaccard coefficient* for clustering into knowledge clusters a group of model answers prepared by domain experts. The clustering is done based on keyword matches and continues iteratively till a pre-determined threshold is reached. The choice of *Jaccard coefficient* is due to the fact that it performs better when dealing with text data (Huang et al., 2008; Niwattanakul et al., 2013).

The *Jaccard similarity coefficient* is a statistical measure of the similarity and diversity of sample sets. It measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets, given by:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

It is evident that, $0 \leq J(A, B) \leq 1$, and if both the sets A and B are empty then, $J(A, B) = 1$.

Given the subjectivity of text-based answers, it is implied that most learners would answer differently. Each cluster veritably representing a group of seemingly not closely associated answers, grouped by virtue of common keyword usage is subsequently used to assess a learners response as correct or incorrect depending on whether it associates with a given cluster or not.

4 Rough set theory

Proposed by Pawlak (1982) *rough set theory* is a mathematical approach to deal with imperfect data. The theory has found vast applications in the areas of artificial intelligence and cognitive science, especially in the fields of machine learning, knowledge acquisition, decision analysis, knowledge discovery, inductive reasoning, etc.

The underlying system for constructing rough sets is the set algebra, $(2U, \neg, \cap, \cup)$. An element A of $2U$ represents a non-vague concept. When such a non-vague concept is viewed with respect to elements of the quotient set U/R, i.e., the equivalence classes of, it becomes vague and uncertain (Yao, 1998).

The theory is based on the approximation concept and lower/upper approximation of a set. Given any subset of attributes R, any concept $X \subseteq U$ can be defined approximately by employing two crisp sets called lower and upper approximations. If $X \subseteq U$ and $A \subseteq R$, the upper and lower approximations can be defined as:

$$\overline{A}X = \{x \in U : [x]_A \cap X \neq \phi \tag{2}$$

$$\underline{A}X = \{x \in U : [x]_A \subseteq X\} \tag{3}$$

The upper approximation contains all elements which possibly belong to X, while the lower approximation consists of all elements that definitely belong to X. The difference between these two approximations constitutes the boundary region of the rough set.

Rough set theory further defines *indiscernibility* or *indistinguishability* between elements of a set. Considering that P is a set of attributes and $P \subseteq A$, where A is the set of all attributes, two elements x and y are *P-indiscernible* or *indistinguishable* over the attribute set P, denoted as $IND(P)$, if:

$$IND(P) = \{(x, y) \in U \mid \forall a \in P, a(x) = a(y) \tag{4}$$

Rough sets can also be characterised in terms of accuracy of approximation, defined by:

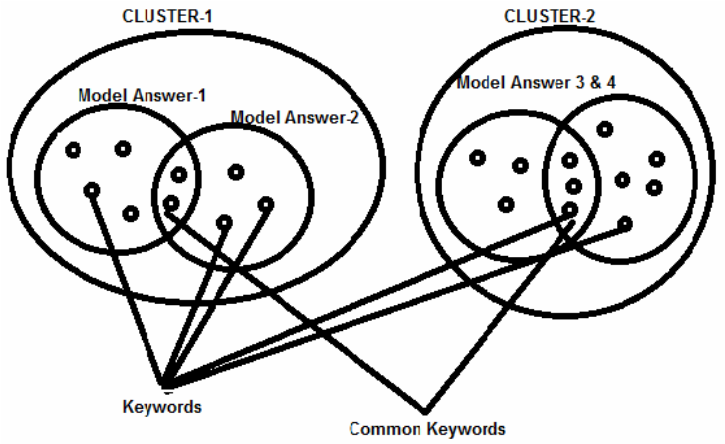
$$\alpha_A(X) \frac{\underline{A}X}{\overline{A}X} \tag{5}$$

The initial clusters formed from the model answers are inherently rough in nature. This can be substantiated by exploring the fact that a cluster would contain model answers with all or some matching keywords with similarity measure higher than a given

threshold. However, all model answers have the same outcome, i.e., all model answers have been accepted as completely correct. So, the scenario finally settles down to one or more clusters, each having answers which are indiscernible over certain keywords and not indiscernible over certain others and yet have the same outcome, which is total acceptance.

Figure 1 depicts the scenario that arises out of clustering the model answers all of which are correct. Since the process starts with a single answer, the cluster formed will have an intersection area containing keywords which would be forming the core of the answer or the lower approximation. All other keywords, which do not belong to the intersection, would form the upper approximation of the cluster.

Figure 1 Cluster formation



5 Proposed work

The proposed system as presented in this paper evaluates text-based answers and is part of a larger setup for e-learning as shown in Figure 2. The platform is centered on two key entities, namely the 'learner entity' and the 'coach'. All other modules assist these two to communicate effectively, leading the learner entity through better learning experience. The automated evaluation process interacts with the learner, his/her performance records and returns assessment to the coach, who may be human or an automated system which guides the learner through the corrective/betterment path.

The process of evaluation ideally depends on factors like, learners' behaviour, preference, interaction with environment and learning history which would influence the hardness level of the questions. The current paper however, concentrates only on the evaluation part without delving into the influencing factors as mentioned. The process of evaluation of a learners' response under the proposed system is a two staged process. During the first or the pre-processing stage, the model answers prepared by subject experts are collected and clustered into concept clusters. These concept clusters are inherently rough as they contain multiple model answers having different keywords yet the same outcome, as these are model answers and are representative of complete

knowledge with respect to the domain under consideration. This inherent roughness is exploited in associating the learners' responses with the concept clusters based on lower and upper bound matches, through knowledge mining. The entire process of evaluation is figuratively summarised through Figure 3.

Figure 2 The e-learning platform

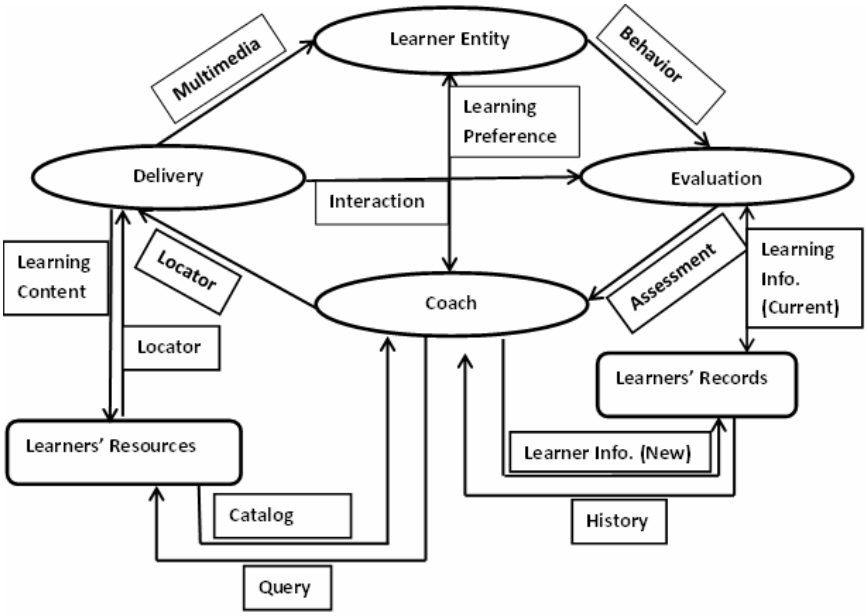
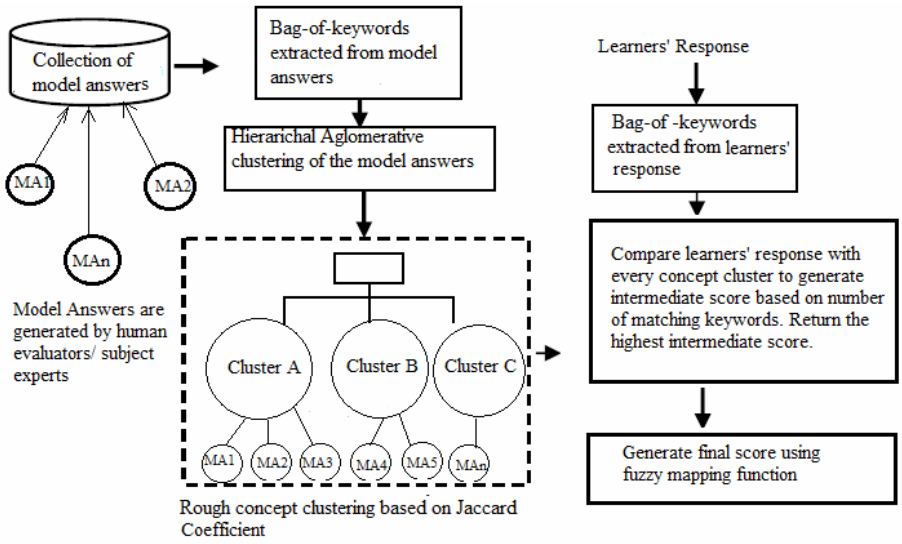


Figure 3 The structure of the proposed ASAG system



5.1 Pre-processing

The pre-processing phase (Chiu et al., 2007) essentially trims the answer down to a keyword only group through a process of removal of punctuations, numerals and stop words. The algorithm *pre-process* is used for this purpose:

```
Algorithm pre-process
Input: an array of strings A[n], where each A[i] is an answer.
Output: the array A[n] with each A[i] processed.

1   Begin
2   For i=1 to n do
    2.1   While (!End of string A[i]) do
          2.1.1   Remove all punctuations
          2.1.2   Remove all numerals
          2.1.3   Remove all stop words
          2.1.4   Convert all words to singular
          2.1.5   Remove repeated words
    2.2   End while
3   End for
4   End
```

Prior to the initial concept clustering, every model answer is pre-processed to return a bag-of-words. The same process is repeated during actual evaluation of the learners' response where the student answers are pre-processed before actual evaluation.

An example of extraction of bag-of-keywords from model correct answers is shown as an example for the given code segment for the question, *'What will be the output of the following code?'*

The sample answers considered for the example are listed as Table 1.

```
main()
{
    int x;
    float y;
    for(x=0; ;x++)
    {   if(x ≤ 2)
        y = 10/x;
        printf("\%f", y);
    }
}
```

The stages in the extraction process of representative keywords of a particular answer in pre-processing are shown in Table 2.

Table 1 Sample answers

Sl. no.	*Answer*
1	After use loop, the program cannot execute normally.
2	When use loop, it shows syntax error message.
3	When calculate division, the program cannot execute normally.
4	When use if statement, the result produced by the program is an error.
5	When use if statement it shows syntax error message.
6	Division by zero gives syntax error.
7	When division operation will be performed it shows syntax error.

Table 2 Representative keyword extraction

Steps	*KW1*	*KW2*	*KW3*	*KW4*	*KW5*	*KW6*	*KW7*	*KW8*	*KW9*
Segment an answer	After	Use	Loop	,	The	Program	Cannot	Execute	Normally
Delete punctuations	After	Use	Loop		The	Program	Cannot	Execute	Normally
Delete articles	After	Use	Loop			Program	Cannot	Execute	Normally
Delete stop words		Use	Loop			Program	Cannot	Execute	Normally

5.2 Concept clustering

Once all model answers have been stripped off punctuations, numerals and stop words, the individual bags of words are clustered using *hierarchical agglomerative clustering* using a *Jaccard coefficient* value of 0.5 as a threshold. This results in two model answers being put in the same concept cluster only if the similarity value is higher than the threshold. A high value has been decided as the model answers are to the same question and all are correct, and answers fall in the same cluster only if half or more of the keywords among those are common. A simple example is considered to illustrate the idea of concept clustering.

Given a question, 'What is an array?' and six responses as given in Table 3 along with the keywords after stripping the stop words and plural to singular conversion, the *Jaccard coefficient* is calculated and listed in Table 4. The emboldened values are the ones crossing the threshold level and are therefore clustered as shown in Figure 4.

Figure 4 Sample concept clusters

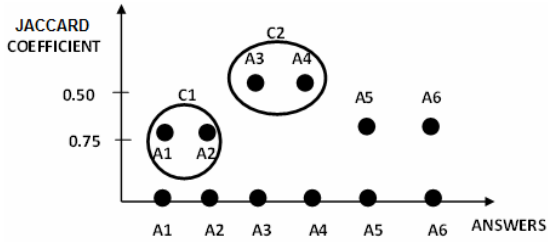


Table 3 Sample answer and extracted keywords

Sl. no.	*Answer*	*Keywords*
1	An array is a set of ordered doubles representing values and their indices.	Array, set, ordered, double, representing, value, and, index
2	An array is a set of values and index doubles.	Array, set, value, and, index, double
3	An array is the collection of homogeneous elements stored in continuous memory locations.	Array, collection, homogeneous, element, stored, in, continuous, memory, location
4	Array is a collection of variables of similar data types stored in contiguous memory locations.	Array, collection, variable, similar, data type, stored, in, continuous, memory, location
5	It is a derived data structure in which continuous memory allocation takes place with same data types.	Derived, data, structure, continuous, memory, allocation, take, place, same, data type
6	Arrays are derived data types which contain elements of same data type in continuous memory locations.	Array, derived, contain, element, same, data type, continuous, memory, location

5.3 Evaluating learners' responses

Post pre-processing and concept clustering with the model answers, the system is ready to evaluate the learners' responses. Each response is accepted as a text answer one sentence long and is pre-processed to remove the punctuations, numerals and stop words. As a result, every answer is converted to a bag-of-words, every word being a keyword. Each such bag-of-words is thereafter compared with every concept cluster to find out the number of matched keywords. An intermediate score for the learners' response is generated considering the concept cluster showing the maximum number of common keywords, using expressions (6) or (7) depending the roughness of the cluster. The algorithm *Evaluate* depicts the entire process.

$$y = \frac{\left(|\,\overline{A}X\,| - |\,\underline{AX}\,|\right) \times \textit{Number of keywords matched with concept cluster}}{\left(|\,\overline{A}X\,| - |\,\underline{AX}\,| + 1\right) \times \left(\overline{A}X - 1\right)} \tag{6}$$

$$y = \frac{\textit{Number of keywords matched with concept cluster}}{\textit{Total number of keywords in cluster}} \tag{7}$$

Expression (6) is used if the cluster returning maximum number of keyword matches is rough, i.e., if:

$$\overline{A}X \neq \underline{A}X \tag{8}$$

The expression for the intermediate score takes care of both high and low accuracy of approximation of roughness with high and low keyword match counts. Considering marginal roughness, i.e., when the lower approximation would contain only one keyword (worst case) in the cluster, the expression (6) becomes:

$$y = \frac{Number\ of\ keywords\ matched\ with\ concept\ cluster}{\left(|\overline{A}X| - |\underline{AX}| + 1\right)}$$
$$y = \frac{Number\ of\ keywords\ matched\ with\ concept\ cluster}{|\overline{A}|\,X} \tag{9}$$

Algorithm evaluate

Input: An array of strings A[n], where each A[i] is an answer.

Data structures: A[n] – an array of strings; B[n] – an array of strings (used to store keywords from each answer); C[m] – an array of strings to store the clusters; K[m][n] – A 2D array to store the number of matched keywords

Output: The intermediate score 'y'

```
1   Begin
2   For i = 1 to n
      2.1   B[i] = Pre-process(A[i])
3   End for
4   For i= 1 to n
      4.1   For j = 1 to m
              4.1.1   K[i][j] = matched_keywords(B[i], C[j])
      4.2   End for
5   End for
6   For i = 1 to n
      6.1   For j = 1 to m
              6.1.1   N = j if( max(K[i][j]))
      6.2   End for
7   End for
8   If C[N] (ĀX ≠ A̲X)
      8.1   y = ((|ĀX| − |A̲X|) × Kij) / ((|ĀX| − |A̲X| + 1) × (ĀX − 1))
9   Else
      9.1   y = Number of keywords matched with concept cluster / Total number of keywords in cluster
10  End If
11  End
```

Expression (9) would return a value of y very close to 1 for a high number of matched keywords and a very low value for a low number of matched keywords. On the other hand, if the cluster is rough with the difference between lower and upper approximation *cardinalities* being high (m), then for k matching keywords, y becomes:

$$y = \frac{m \times k}{(m+1)(k+1)} \tag{10}$$

The proposed system therefore makes it difficult to score high for crispier concept answers, or concepts that are well defined and relatively easier to score on answers where the concepts are not as well defined.

A similar approach using intelligent fuzzy evaluation functions have also been used for single word response (answer) evaluation and reported by Chakraborty et al. (2014a). The interested reader may refer to Chakraborty et al. (2014a) for more details on similar evaluation functions.

The value of y, which is the *intermediate score*, is thereafter converted to the final score to return a value between *0* and *1*. This exercise becomes essential as HEs score students discreetly and the machine score should resemble the HE as closely as possible. The final score is generated using the expression given as (11):

$$s = \mu(y) = \begin{cases} 0 & \text{if } y < 0.25 \\ 0.25 & \text{if } 0.25 \le y < 0.4 \\ 0.5 & \text{if } 0.4 \le y < 0.6 \\ 1 & \text{if } y \ge 0.6 \end{cases} \tag{11}$$

6 Experimentation and results

The proposed model was tested on dataset generated for the purpose, since standard datasets like http://web.eecs.umich.edu/mihalcea/downloads.html deal with short text responses which are not restricted to single sentences. Further, the choice of model answers being important in the whole process, the non-availability of these in the above standard dataset required normalisation of scores based on identification of the highest scoring answer as model answer. This was considered non-compliant with the concept coverage issue as the best student response might not be the complete response equivalent to a model.

Rigorous experiments were carried out on several single sentence answers collected from the students' responses. The proposed system was implemented and tested with 281 single sentence answers in English with an average length of 16 words per answer. The answers were collected from actual examinations on data structures, conducted for four different groups having 81, 78, 47 and 75 students respectively. The students were given questions as part of class tests, in groups and asked to type in their responses which were gathered on a data server before being processed separately. The human evaluated scores of these tests contributed to the students' grade at the end of the semester. The students were in the age group 17–19 years and pursuing an UG program in engineering. This data has been made available at https://www.researchgate.net/publication/281845137_Data_SingleSentence. Each group was given a different question to answer and for every question the teachers were asked to prepare 12 model answers matching their expectations from the students. The answers were blind evaluated by two HEs and the average score taken as the score of the HE for comparison with the system generated score. Table 4 lists the four questions.

Most automated evaluating systems available work on text-based answers which are longer than a single sentence answer. The dataset presented here was evaluated using *LSA* which is generally used to evaluate essays but has also been used on short text answers. The intermediate score and the final score of the proposed method along with the HE's

score for ten possible answers against the question 'What is an array?' are shown in Table 5. Table 6 lists the set of model answers for the question. The correlation values returned with the average score of two HEs are listed in Tables 7 and 8. The values returned by *LSA* are indicative of the fact that the method does not work well when dealing with short answers as is the case for which the proposed system has been developed. It reveals a shortcoming of *LSA* as reported in Hastings et al. (2001) that it does not work well with text having length less than 200 words.

Table 4 List of questions

Sl. no.	*Question*
1	Define data structures
2	How is continuous memory location assignment in arrays advantageous?
3	What is the problem with array representation of queues?
4	What is an array?

Table 5 Comparative scores

Sl. no.	*Learners' response*	*Human evaluators' score*	*Proposed method*	
			Intermediate score (y)	*Final score (S)*
1	A set of index doubles	0	0.292	0.25
2	An array is a collection of data type in contiguous memory locations.	0	0.53	0.5
3	An array is a regular grouping of data of same kind.	0	0.343	0.25
4	An array is a set of an element in an order.	0	0.343	0.25
5	An array is a set of build and index.	0	0.292	0.25
6	A data structure that stores homogeneous type elements	0.25	0.571	0.5
7	An array is a collection of set values in contiguous memory locations.	0.25	0.571	0.5
8	Homogeneous set of data having continuous memory allocation.	0.5	0.25	0.25
9	Finite ordered collection of homogeneous data elements.	0.5	0.417	0.5
10	An array is defined as a set of value and index pairs.	1	1	1

Results are better than *LSA* for the above mentioned four questions on the same 281 student data where returned when tested with the software system proposed in Chakraborty et al. (2014b), which uses a *link grammar*-based approach to discover and seek relations between words. However, both *LSA* and the *link grammar*-based system ideally being designed for longer answers are more inclined in seeking out language

features which makes them computationally heavy, which is an added advantage for the current system.

The performance analysis of the proposed system reveals that it returns lower correlation with HEs for question number 4 which is a definition type question. The same was explored in the previous section wherein mention was made about difficult scoring with crispier answers. Evidently, HEs using their intelligence judge the learners' knowledge state better for deviations from a definition which is not possible using an automated approach.

The scores returned by the proposed system and the average score of two HEs have been plotted for questions 1, 2, 3 and 4 and shown as Figures 5, 6, 7 and 8, respectively. In the plots, the X-axis represents the student ID's while the Y-axis represents the score line.

Table 6 Model answers

Sl. no.	*Model answer*
1	An array is defined as a set of value and index pairs.
2	An array classically is a set of indices and values associated with it while from the programming point of view is a homogeneous collection of similar data types stored continuously.
3	An array is a collection of elements of similar data type arranged in contiguous memory locations.
4	An array is a collection of elements of similar data type values arranged in contiguous manner.
5	An array is a collection of homogeneous data elements in memory in continuous memory locations.
6	An array is a collection of similar type of data elements having contiguous memory location.
7	An array is a collection of values and its indices.
8	An array is a collection of variables having the same name and same data type stored in contiguous memory locations.
9	An array is a linear data structure which can store similar types of data in contiguous memory locations.
10	An array is a derived data type which stores data of same type in contiguous memory locations.
11	An array is a set of ordered doubles representing values and their indices.
12	An array is a set of value, index doubles.

Table 7 Comparative Pearson correlation values

Question no.	*Pearson correlation with average of two human evaluators'*		
	Proposed system	*LSA-based system*	*Link grammar-based system*
1	0.9124	0.10748	0.59634
2	0.69709	0.20039	0.40129
3	0.83883	0.14693	0.49937
4	0.59591	0.22396	0.38833

Table 8 Spearman correlation values

Question no.	*Spearman correlation with average of two human evaluators*
1	0.8529
2	0.3955
3	0.6529
4	0.4801

Figure 5 Plot showing scores for question 1 (see online version for colours)

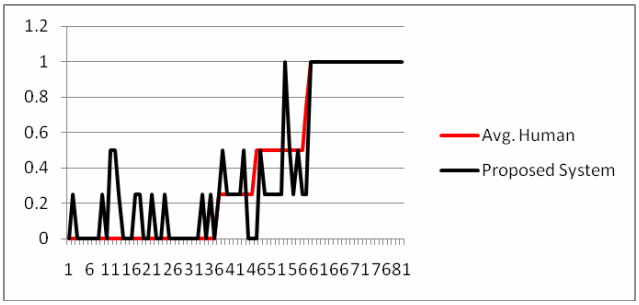


Figure 6 Plot showing scores for question 2 (see online version for colours)

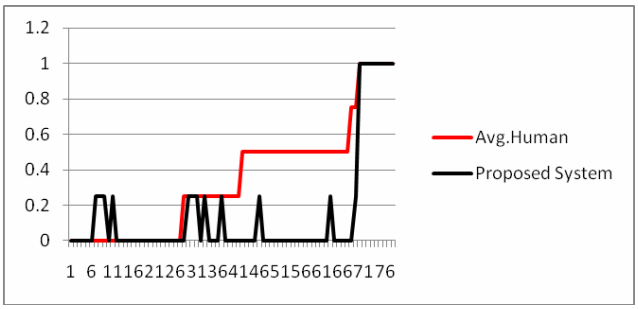


Figure 7 Plot showing scores for question 3 (see online version for colours)

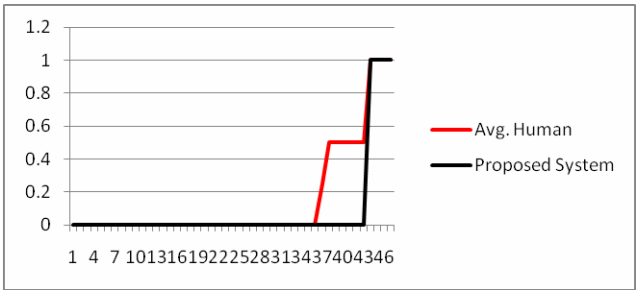
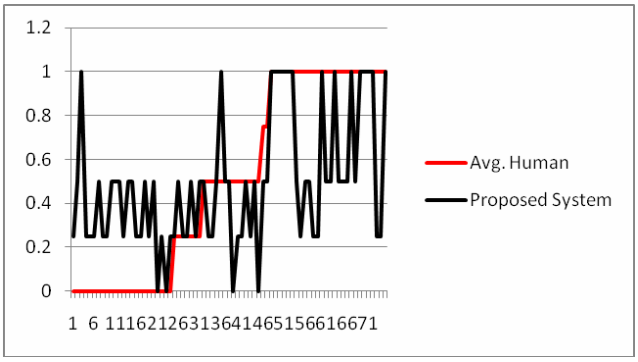


Figure 8 Plot showing scores for question 4 (see online version for colours)



7 Conclusions

Various approaches have been proposed for the automated evaluation of free text answers, the popular ones being *LSA*, *SELSA*, *BLEU* or a combine of *LSA* and *BLEU*. None however has been popularly accepted for high stake tests. The popular approaches have been guided through the use of NLP techniques or surface features. However, the inherent vagueness or roughness, that marks natural language expressions, have not been explored or reported in literature. The proposed technique uses a rough set-based approach to evaluate text-based answers through concept clustering.

The problem under consideration is a subset of the much larger field of free text answer evaluation by virtue of the limitation it imposes on the length of the answer. As short answer question are commonly used in examinations to assess the basic knowledge and understanding of a topic before more in-depth assessment questions are asked, these serve a specific purpose. Single sentence answers are of further pedagogical significance as these do not allow the learner the flexibility to answer elaborately. Since the lengths of the answers considered are short, the possibilities of discovery of linguistic features are limited. This not only justifies the use of a keyword only approach but also considerably reduces the computational complexity involved in the process. The use of rough set theory, hitherto before unexplored in evaluation, is a further addition to the novelty of the work. Roughness of the concept clusters is representative of the inherent roughness of natural language expressions which is exploited in the proposed method. The expressions for intermediate score generation bring out with sufficient accuracy the proximity of a learners' response to the concept cluster which is further modified to resemble a human generated score.

An advantage of the proposed method is that the number of model answers used for concept cluster formation may be varied to enhance the accuracy of the system. If the learners are expected to stick to a few basic answers, then the number of model answers may be reduced thereby reducing or even eliminating any roughness in the concept cluster. An increase in the number of model answers increases the ability of the evaluation system to allow for creative expressions on the part of the learner which adds further value to the proposed system. However, it is worth adding that the correlation values depend on the scoring pattern of the HE, and since different persons have different

evaluating styles, the same may vary based on the choice of the evaluator. The choice of model answers is also important in deciding the accuracy of the results. However, since the method allows for modelling using more than one model answers and finally knowledge discovery is based on the concept clusters, the effect of a not so accurate model answer would be mitigated.

The work carried out using a dataset comprising of 281 answers in all having answers' to four different questions returned scores showing improved performance over the existing systems and establish the proposed approach as a valid and viable solution strategy to evaluate single sentence text answers. The correlation values with the average of two HEs, as returned by the proposed system also prove the merit of the approach.

References

Basu, S., Jacobs, C. and Vanderwende, L. (2013) 'Powergrading: a clustering approach to amplify human effort for short answer grading', *TACL*, Vol. 1, pp.391–402.

Budanitsky, A. and Hirst, G. (2006) 'Evaluating wordnet-based measures of lexical semantic relatedness', *Computational Linguistics*, March, Vol. 32, No. 1, pp.13–47.

Burrows, S., Gurevych, I. and Stein, B. (2015) 'The eras and trends of automatic short answer grading', *International Journal of Artificial Intelligence in Education*, Vol. 25, No. 1, pp.60–117, IOS Press.

Burstein, J., Kukich, K., Wolff, S., Lu, C. and Chodorow, M. (1998) *Computer Analysis of Essays* [online] https://www.ets.org/Media/Research/pdf/erater_ncmefinal.pdf (accessed 13 December 2015).

Chakraborty, U., Konar, D., Roy, S. and Choudhury, S. (2014a) 'Intelligent fuzzy spelling evaluator for e-learning systems', *Education and Information Technology*, Vol. 21, pp.171–184, Springer Science+Business Media, New York, DOI: 10.1007/s10639-014-9314-z.

Chakraborty, U.K., Gurung, R. and Roy, S. (2014b) 'Semantic similarity based approach for automatic evaluation of free text answers using link grammar', *Proceedings of 2014 IEEE Sixth International Conference on Technology for Education*, pp.218–221.

Chakraborty, U.K., Roy, S. and Choudhury, S. (2014c) 'A novel semantic similarity based technique for computer assisted automatic evaluation of textual answers', *Advanced Computing and Informatics Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics*, Kolkata, Vol. 1, pp.393–402.

Chakraborty, U.K. and Das, S. (2015) 'Automatic free text answer evaluation using knowledge network', *International Journal of Computer Applications*, Vol. 117, No. 3, pp.5–8.

Chang, S.H., Lin, P.C. and Lin, Z.C. (2007) 'Measures of partial knowledge and unexpected responses in multiple-choice tests', *Educational Technology and Society*, Vol. 10, No. 4, pp.95–109.

Chiu, D.Y., Chen, P.S. and Pan, Y.C. (2007) 'Dynamic FAQ retrieval with rough set theory', *International Journal of Computer Science and Network Security*, Vol. 7, No. 8, pp.204–211.

Desus, P., Lemaire, B. and Vernier, A. (2000) 'Free-text assessment in a virtual campus', *Proceedings of Third International Conference on Human System Learning*, pp.61–76.

Foltz, P.W., Laham, D. and Landauer, T.K. (1991) 'The intelligent essay assessor: applications to educational technology', *Interactive Multimedia Electronic Journal of Computer Enhanced Learning*, Vol. 1, No. 2, pp.939–944.

Gabrilovich, E. and Markovitch, S. (2009) 'Wikipedia based semantic interpretation for natural language processing', *Journal of Artificial Intelligence Research*, Vol. 34, pp.443–498.

Griff, E.R. and Matter, S.F. (2013) 'Evaluation of an adaptive online learning system', *British Journal of Educational Technology*, Vol. 44, No. 1, pp.170–176.

Gutierrez, I., Kloos, C. and Crespo, R. (2010) 'Assessing assessment formats: the current picture', in *IEEE Education Engineering (EDUCON*, pp.1233–1238, DOI: 10.1109/EDUCON.2010.5492384.

Hastings, W., Moore, J.D. and Stenning, K. (2001) 'Rules for syntax, vectors for semantics', *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pp.1112–1117.

Hermet, M., Szpakowicz, S., Duquette, L. and Leuven, S.N. (2006) 'Automated analysis of students' free-text answers for computer-assisted assessment', *Proceedings of TAL and ALAO Workshop*, pp.835–845.

Hovy, E., Navigli, R. and Ponzetto, S.P. (2013) 'Collaboratively built semi-structured content and artificial intelligence', *Artificial Intelligence*, Vol. 194, pp.2–27.

Huang, A., Holland, J., Nicholas, A. and Brignoli, D. (2008) 'Similarity measures for text document clustering', *Proceedings of New Zealand Computer Science Research Student Conference*, January 2008, pp. 49–56.

Jadidinejad, A.H. and Mahmoudi, F. (2014) 'Unsupervised short answer grading using spreading activation over an associative network of concepts', *Canadian Journal of Information and Library Science*, December 2014, Vol. 38, No. 4, pp.287–303.

Jing, S. (2015) *Automatic Grading of Short Answers for MOOC via Semi-supervised Document Clustering* [online] http://www.educationaldatamining.org/EDM2015/uploads/papers/paper_225.pdf (accessed 2 November 2015).

Kumaran, V.S. and Sankar, A. (2015) 'Towards an automated system for short-answer assessment using ontology mapping', *International Arab Journal of e-Technology*, January 2015, Vol. 4, No. 1, pp.17–24.

Landauer, T.K., Foltz, P.W. and Laham, D. (1998) 'An introduction to latent semantic analysis', *Discourse Processes*, Vol. 25, Nos. 2/3, pp.259–284.

Leacock, C. and Chodorow, M. (2003) 'C-rater: automated scoring of short answer questions', *Computers and Humanities*, Vol. 37, No. 4, pp.389–405.

Lin, J. and Fushman, D.D. (2005) 'Automatically evaluating answers to definition questions', *Proceedings of Human Language Technology Conference and Conference of Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp.931–938.

Lior, R. and Maimon, O. (2005) 'Clustering methods', *Data Mining and Knowledge Discovery Handbook*, Vol. 4, pp.321–352, Springer, USA.

Makatchev, M. and VanLehn, K. (2007) 'Combining bayesian networks and formal reasoning for semantic classification of student utterances', *Proceedings of the International Conference on AI in Education (AIED)*, Los Angeles.

Mittal, H. and Devi, M.S. (2016) 'Computerized evaluation of subjective answers using hybrid technique', in Saini, H.S., Sayal, R. and Rawat, S.S. (Eds.): *Innovations in Computer Science and Engineering, Edition: 1*, Springer-Verlag, Singapur.

Mohler, M. and Mihalcea, R. (2009) 'Text-to-text semantic similarity for automatic short answer grading', *Proceedings of 12th Conference of the European Chapter of the ACL*, Athens.

Mohler, M., Bunescu, R. and Mihalcea, R. (2011) 'Learning to grade short answer questions using semantic similarity measures and dependency graph alignments', *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp.752–762.

Ngee, P. et al. (2011) 'Guessing partial knowledge and misconceptions in multiple choice tests', *Educational Technology and Society*, Vol. 14, No. 4, pp.99–110.

Niwattanakul, S., Singthongchai, J., Naenudorn, E. and Wanapu, S. (2013) 'Using of Jaccard coefficient for keywords similarity', *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Vol. 1.

Omran, B., Muftah, A., Aziz, A. and Juzaiddin, M. (2014) 'Syntactically enhanced LSA methods in automatic essay grading systems for short answers', *Proceedings of the 3rd International Conference on Computer Engineering and Mathematical Sciences (ICCEMS 2014)*, pp.412–417.

Paden, R. and Chakraborty, U.K. (2015) 'Open ended tool for self- paced learning', *International Journal of Hybrid Information Technology*, Vol. 8, No. 8, pp.201–208.

Page, E. (1966) 'The imminence of grading essays by computer', *The Phi Delta Kappan*, Vol. 47, pp.238–243.

Papineni, K., Roukos, S., Ward, T. and Zhu, W-J. (2002) 'BLEU: a method for automatic evaluation of machine translation', *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics*, pp.311–318.

Pawlak, Z. (1982) 'Rough sets', *International Journal of Computer and Information Science*, January 1982, Vol. 11, No. 5, pp.341–356.

Perez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodriguez, P. and Magnini, B. (2005) 'Automatic assessment of students\textquoteright~free-text answers underpinned by the combination of a BLEU-inspired algorithm and latent semantic analysis', *Proceedings of FLAIRS*.

Pintso, M., Doucet, A.V. and Fernandez-Ramos, A. (2010) 'Measuring students-information skills through concept mapping', *Journal of Information Science*, Vol. 36, No. 4, pp.464–480.

Pulman, S.G. and Sukkarieh, J.Z. (2005) 'Automatic short answer marking', *Proceedings of the Second Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics*, Michigan, pp.9–16.

Ros'e, C.P., Roque, A., Bhembe, D. and VanLehn, K. (2013) 'A hybrid approach to content analysis for automatic essay grading', *Proceedings of the 2003 Conference of the North American Technology Chapter of the Association of Computational Linguistics on Human Language*, Vol. 2, pp.88–90.

Selvi, P. and Banerjee, A.K. (2010) *Automatic Short Answer Grading System (ASAGS)*, arXiv preprint arXiv: 1011.1742.

Siddiqi, R., Harrison, J. and Siddiqi, R. (2010) 'Improving teaching and learning through automated short-answer marking', *IEEE Transactions on Learning Technologies*, Vol. 3, No. 3, pp.237–249.

Sultan, M.A., Salazar, C. and Sumner, T. (2016) 'Fast and easy short answer grading with high accuracy', *Proceedings of NAACL-HLT*, San Diego, California, pp.1070–1075.

Yao, Y.Y. (1998) 'A comparative study of fuzzy sets and rough sets', *Information Sciences*, Vol. 109, Nos. 1–4, pp.227–242.

Zesch, T. and Gurevych, T. (2010) 'Wisdom of crowds versus wisdom of linguists-measuring the semantic relatedness of words', *Natural Language Engineering*, Vol. 16, No. 1, pp.25–29.

Zhang, Z., Gentile, A.L. and Ciravegna, F. (2013) 'Recent advances in methods of lexical semantic relatedness: a survey', *Natural Language Engineering*, Vol. 19, no. 4, pp.411–479.

Ziai, R., Ott, N. and Meurers, D. (2012) 'Short answer assessment-establishing links between research strands', *Proceedings of The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pp.190–200.