Wisdom of Students: A Consistent Automatic Short Answer Grading Technique

Shourya Roy[†] **Sandipan Dandapat**[‡] **Ajay Nagesh**[*] **Y. Narahari**[§]
[†]Xerox Research Centre India, Bangalore,
[‡]Microsoft India, Hyderabad,
[*]University of Massachusetts, Amherst,
[§]Indian Institute of Science, Bangalore
shourya.roy@xerox.com, sadandap@microsoft.com, ajaynagesh@gmail.com, hari@csa.iisc.ernet.in

Abstract

Automatic short answer grading (ASAG) techniques are designed to automatically assess *short* answers written in natural language having a length of a few words to a few sentences. In this paper, we report an intriguing finding that the set of short answers to a question, collectively, share significant lexical commonalities. Based on this finding, we propose an unsupervised ASAG technique that only requires sequential pattern mining in the first step and an intuitive scoring process in the second step. We demonstrate, using multiple datasets, that the proposed technique effectively exploits *wisdom of students* to deliver comparable or better performance than prior ASAG techniques as well as distributional semantics-based approaches that require heavy training with a large corpus. Moreover, by virtue of being independent of instructor provided model answers, our technique offers *consistency* by overcoming the limitation of undesired variability in performance exhibited by existing unsupervised techniques.

1 Introduction

Automatic grading systems have been in practice in the educational domain for many years now, but primarily for *recognition* questions where students have to choose the correct answer from given options such as multiple choice questions (MCQs). Prior research has shown that such recognition questions are deficient as they do not capture multiple aspects of acquired knowledge such as reasoning and self-explanation (Wang et al., 2008). In contrast, *recall* questions that seek students' constructed answers in natural language have been found to be more effective in assessing their acquired knowledge. However, automating assessment of such answers is non-trivial owing to linguistic variations (a given answer could be articulated in different ways); subjective nature of assessment (multiple possible correct answers or no correct answer); lack of consistency in human rating (non-binary scoring on an ordinal scale within a range); etc. Consequently, this has remained a repetitive and tedious job for teaching instructors and is often seen as an overhead and non-rewarding. This paper is about a computational technique for automatically grading constructed student answers in natural language. In particular, we are interested in *short answers*: a few words to a few sentences long (everything in between fill-in-the-gap and essay type answers (Burrows et al., 2015)) and refer to the task as *Automatic Short Answer Grading* (ASAG). An example ASAG task is shown in Table 1.

Question	How are overloaded functions differentiated by the compiler?
Model Ans	Based on the function signature. When an overloaded function is called, the compiler will find the function whose signature is closest to the given function call.
Stud#1	It looks at the number, types, and order of arguments in the function call
Stud#2	By the number, and the types and order of the parameters.

Table 1: Example of question, model answer, and student answers from an undergraduate computer science course (Mohler and Mihalcea, 2009). These will be used as the running example through out the paper.

A large fraction of prior work in ASAG systems mostly comprises techniques that require extensive intervention from instructors (Roy et al., 2015; Burrows et al., 2015). In one group of such works, the instructors are expected to list key concepts (and their possible variations) to look for in student responses and grade them using *concept mapping* (Burstein et al., 1999; Leacock and Chodorow, 2003) and *information extraction*-based techniques (Bachman et al., 2002; Mitchell et al., 2002). These techniques are tedious and

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 178–187,
Varanasi, India. December 2016.

unlikely to generalize and moreover tend to lead to a large number of false negatives owing to unspecified linguistic variations. The other group of techniques requires instructors to grade a fraction of student answers (typically ranging from half to three-quarter) to train *supervised learning algorithms* (Sukkarieh et al., 2011; Madnani et al., 2013) as training data for building classification or regression models. Unlike many applications where generation of labelled data is a one-time exercise, ASAG does not fit into *train-once-and-apply-forever* model. Every ASAG task is unique and would require ongoing instructor involvement to create labelled data. Requirement of such ongoing involvement of instructors limits the benefits of automation in practical and real-life applications.

The other broad group of work (*corpus-based* and *document similarity*-based techniques) uses various document similarity measures to grade student answers. These techniques largely reduce the need for human involvement. They do not require instructors to list all possible variations of model answers; rather rely on the measures to assess similarity between student and model answers. Unlike supervised techniques there is no ongoing instructor involvement as providing model answer is a one time task that does not depend on size of the student population. However, these techniques suffer from multiple other shortcomings. First, there is no standardization around how model answers are written across datasets or even within a dataset. The model answer in Table 1 is more detailed and self-contained than the model answer ``Abstraction and reusability'' for another question ``What are the main advantages associated with object-oriented programming?'' from the same dataset. This immediately hints at the fact that the same measure is unlikely to work for both questions. Second, variations in model answers can affect the performance of ASAG technique significantly. Using another valid model answer for the question in Table 1 ``The compiler selects a proper function to execute based on number, types and order of arguments in the function call.'', causes correlation of ASAG scores with human scores vary significantly.[1] Finally, many similarity-based ASAG techniques require access to rich knowledge-bases which may not be available for all languages and all subject matters thereby limiting their applicability.

In this paper, we first report an intriguing finding that short answers to a question contain significant lexical overlap among them and such overlapping text are typically related to the correct answer to the question. We convert this finding into a technique assuming (and validating for datasets used) that such commonalities are characteristics of correct answers as typically there are fewer ways of expressing correct answers than incorrect ones.[2]

The proposed technique can be implemented in two steps. In the first step, we pose a variant of sequential pattern mining problem (Agrawal and Srikant, 1995) to identify sequential word patterns that are more common (than the rest of the patterns) among student answers. In the second step, based on our intuition driven hypothesis, that presence of such common patterns is indicative of correct answers, we deduce the scores using an intuitive scoring method (assigning weights to patterns by their length along with frequencies). The approach is truly unsupervised as it does not require human supervision in terms of pre-graded answers or manually crafted key concepts. Unlike similarity-based techniques, it does not suffer from non-standardization of model answers. Other than minimal pre-processing, we do not perform any feature engineering which is typical of ASAG solutions and thus our approach generalizes better. In the sequel, we use the words *approach*, *method* and *technique* synonymously.

Our contributions: The contributions and novelty of this work are summarized below.

- We report a novel and potentially surprising finding regarding the extent of lexical overlap between students' short answers to a question. Exploiting the finding, we propose a new ASAG technique which is completely unsupervised, consistent, and generalizable (§ 3).

[1] We will see evidence of such variations in Section 4.3.

[2] Before describing the technique, we acknowledge that this assumption may not be true for all questions. For example, occasionally instructors design difficult and tricky questions which mislead students to incorrect answers. Nonetheless, we empirically demonstrate that the proposed technique is comparable to existing document similarity based ASAG techniques on standard datasets and thereby providing a truly unsupervised strong baseline at the least. We provide additional discussion as future work in Section 5 towards leveraging the proposed technique for designing more practical ASAG techniques.

- We demonstrate with quantitative results on multiple datasets that the proposed technique delivers comparable or better performance than similarity-based ASAG techniques on various dimensions. In particular, it offers *consistency* by overcoming the limitation of performance variability exhibited by similarity-based techniques caused by switches to equivalent model answers(§ 4.2).
- We provide detailed qualitative analysis with examples from our datasets to portray how the proposed technique would work even under different scenarios such as when most student answers are not perfect or when there are multiple model answers (§ 4.4).
- We create and offer a new dataset on high-school English reading comprehension task in a Central Board of Secondary Education (CBSE) school in India. The dataset contains 14 questions answered by 58 students. The answers were graded by two human raters based on model answers and an optional scoring scheme (§ 4.1).

2 Prior Art

Two recently written survey papers by Roy et. al (2015) and Burrows et. al. (2015) provide comprehensive views of research in ASAG. Both of them have grouped prior research based on the types of approaches used as well as extent of human supervision needed. In this section, we review similarity-based ASAG techniques (e.g. lexical, knowledge-based, vector space etc.).

Similarity based ASAG techniques are premised on measuring similarity between model and student answers. Higher the similarity, higher the score a student answer receives and vice versa. Various types of similarity measures have been used in prior art of ASAG. Among the **lexical** measures, Evaluating Responses with BLEU (ERB) due to Pérez et al. (2004) is one of the earliest work. It adapted the most popular evaluation measure for machine translation, BLEU (Papineni et al., 2001) for ASAG with a set of Natural Language Processing (NLP) techniques such as stemming, closed-class word removal, etc. This work initially appeared as a part of an ASAG system, *Atenea* (Alfonseca and Pérez, 2004) and later as *Willow* (Pérez-Marín and Pascual-Nieto, 2011). Mohler and Mihalcea (2009) conducted a comparative study of different semantic similarity measures for ASAG including **knowledge-based** measures using Wordnet as well as **vector space-based** measures such as Latent Semantic Analysis (LSA) (Landauer et al., 1998) and Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2006). LSA has remained as a popular approach for ASAG and been applied in many variations (Graesser et al., 2000; Wiemer-Hastings and Zipitria, 2001; Kanejiya et al., 2003; Klein et al., 2011). Lexical and semantic measures have been combined to validate natural complementarity of syntax and semantics for ASAG tasks (Pérez et al., 2005). Combination of different string matching and overlap techniques were studied by Gütl on a small scale dataset (2008). Gomaa and Fahmy compared several lexical and corpus-based similarity algorithms (13 string-based and 4 corpus) and their combinations for grading answers in 0-5 scale (2012). Irrespective of the underlying similarity measure used, these techniques rely solely on the instructor provided model answer for scoring student answers. This central reliance on model answer leads to significant variation in performances even when the model answer is replaced by another equivalent model answer. Dzikovska et al. conducted a 5-way (non-ordinal scale) Student Response Analysis challenge as a part of SemEval-2013 (2013). However, the task had more emphasis on giving feedback on student answers possibly using textual entailment techniques.

3 Proposed Approach

In this section, we provide intuition and details of our proposed approach.

3.1 The Intuition: Wisdom of Students

In similarity-based ASAG techniques, every student answer is compared against the model answer **independently** to arrive at a score indicating the goodness of the student answer. These methods ignore the fact that student answers to a question, as a collection, is expected to share more lexical **commonalities** than any arbitrary collection of text snippets. While the extent of commonality varies, we observe such commonalities for almost all questions in the datasets we dealt with. An example is shown in "Sample Answers" to our running example question in Table 2.

We also empirically note that the correct stu-

dent answers are expected to contain *more* of these commonalities than incorrect ones. For our running example, words such as *argument*, *number*, *order*, *execute*, *type* are among the most frequent ones and are related to the correct answer to the question ("Sample Patterns" in Table 2). If we can identify these commonalities from student answers, then we wondered, can the same be used to score them as well? We hypothesize that the correct student answers are expected to contain *more* of these commonalities than incorrect ones and propose an intuitive scoring technique based on such commonalities.

Sample Answers	The number and type of its parameters.
	The compiler selects the proper functions to execute based on number, types and order of arguments in the function call.
	It selects the proper function to execute based on number, types and order of arguments in the function call.
	The compiler selects proper function to execute based on number, types and order of arguments in the function call.
	Is based on number, types, and order of arguments in the function call.
	Compiler selects proper function to execute based on number, types and order of arguments in the function call.
Sample Patterns	(number, type, order, argument, call)
	(select, proper, execut, base, number)
	(base, number, type, order)
	(execut, base, number, type)
	(proper, execut, base)

Table 2: Illustration: A few student answers and a selection of sequential patterns of stemmed words identified for the question and model answer shown in Table 1.

3.2 Proposed Technique

We model the task of finding commonalities from student answers in a manner similar to the *sequential pattern mining* problem (Agrawal and Srikant, 1995). Sequential patterns in the context of text has been used to capture non-contiguous sequence of words for classification and clustering (Jaillet et al., 2006). Prior work has reported that for such tasks sequential patterns have more reliable statistical properties than commonly used lexical features e.g. n-grams in NLP domain (Sebastiani, 2002). For short answers too, our observation was that sequential patterns are more statistically significant and less noisy than n-grams.

The following two steps, namely, *mining sequential patterns* and *scoring answers* are repeated for all questions by making two passes over all students answers.

Step 1: Mining Sequential Patterns

The objective of this step is to extract commonly occurring patterns and quantify the notion of commonalities using *support*:

1. A student answer (s_i) is converted to a sequence of words ($w_1^i, w_2^i, \ldots, w_n^i$) by removing stopwords and stemming content words to their respective base forms. We optionally remove words which appear in the corresponding question to avoid giving importance to *parrot answering*.

2. A sequential pattern (SP), p of length l, is a sequence of l tokens from s_i i.e. $p = w_{i_1}, w_{i_2}, \ldots, w_{i_l}$ such that $i_1 < i_2 < \ldots < i_l$.

3. Support of p is defined as $sup(p) = |\{s_i : p \in s_i\}|$ i.e. the number of student answers containing p. Connecting to our intuition, patterns with high support are commonalities among answers we are looking for. Lower part of Table 2 shows some of the common patterns obtained from all student answers for the question in Table 1.

4. While $sup(p)$ captures importance of a pattern, they ignore *lexical diversity* (Johansson, 2009) of student answers across questions. Consider two sequential patterns p_1 and p_2 with equal support t for two different questions q_1 and q_2. If we consider only support values then both p_1 and p_2 will contribute equally to score student answers. However, if student answers to q_1 are more *diverse* than q_2, then the fact that p_1 has a support of t is more significant than p_2 having the same support t. We quantify this factor by extending the well known measure *type-token ratio* (TTR) (Lieven, 1978) for lexical diversity. TTR values are in range [0,1] and higher values indicate higher diversity.

$$TTR(d) = \frac{\#\text{distinct patterns of length } d}{\#\text{patterns of length } d}$$

Step 2: Scoring Answers

In this step, we make a second pass over all answers and use the statistics gathered in the previous step to score them:

1. Again a student answer (s_i) is represented as a sequence of words ($w_1^i, w_2^i, \ldots, w_n^i$) by doing the same pre-processing.

2. Building on our intuition, score of s_i is a function of support of all patterns $p \in s_i$. We realize that longer patterns will have (much) less support but their presence are stronger indications than presence of shorter patterns. Hence score of s_i ($Sc(s_i)$) should be a function of support of all patterns p in s_i weighed by respective pattern length. We tried different weighing schemes and found *exponential* weighing ($sup(p)^{len(p)}$) to work the best with respect to correlation with human scores. This indeed matches with our intuition that longer patterns are significantly more important than shorter ones.

3. Subsequently, we use lexical diversity values to bring down contribution of a pattern p if there is less diversity among student answers by modifying $Sc(s_i)$ as below :

$$Sc(s_i) = \sum_{p \in s_i} sup(p)^{len(p)} \times TTR(len(p))$$

4. $Sc(s_i)$ values are normalized using min-max normalization and scaled by maximum obtainable marks in a question.

3.3 Discussion on the Proposed Technique

Step-1 of the proposed technique is similar to the classical sequential pattern mining algorithm but with a couple of differences. Firstly, in our context a common support threshold does not make sense as support values of patterns vary widely across questions depending on nature of answers expected as well as their difficulty levels. Secondly, typically in the literature *closed* and *maximal* frequent patterns are of interest as they subsume smaller length frequent patterns(Tan et al., 2006). However, we consider all patterns to score answers in step-2 but differentially weigh them based on their length.

The basic intuition of the proposed technique can be traced back to the concept based systems of early days of ASAG (Burstein et al., 1998; Nielsen et al., 2008) where instead of (or along with) model answers, a list of concepts were specified. ASAG systems scored student answers based on presence or absence of listed concepts or variations thereof. However, those systems needed the concepts to be specified by the instructor, whereas our endeavour is to identify the important concepts in an unsupervised manner. Recently Ramachandran et al. demonstrated effectiveness of using student answers to a question to extract patterns for ASAG (Ramachandran et al., 2015). They used a graph based approach to extract patterns from groups of questions and their answers towards constructing regular expression alike patterns. While they use a supervised approach using the extracted patterns as features, our approach is completely unsupervised - hence easier to test on new datasets and deploy in real life. Secondly, we opine that regular expression based features can be constraining towards generalization and real life usage for free text answers.

4 Experimental Results

4.1 Experimental Protocols

Datasets: The recent survey papers referred to in Section 2 noted that rarely any ASAG work reported results on multiple (standard) datasets (Burrows et al., 2015; Roy et al., 2015). They emphasized the need for sharing of datasets and structured evaluations on them. Towards that, we evaluated the proposed technique and compared against multiple similarity-based baseline techniques on three datasets:

- **CSD:** This is one of the earliest ASAG datasets consisting of 21 questions with 30 student answers evaluated each on a scale of 0-5 from an undergraduate computer science course (Mohler and Mihalcea, 2009). Student answers were independently evaluated by two annotators and automatic techniques are measured against their average.
- **X-CSD:** This is an extended version of CSD with 81 questions by the same authors (Mohler et al., 2011).
- **RCD:** We created a new dataset on a reading comprehension assignment for Standard-12 students in Central Board of Secondary Education (CBSE) in India. The dataset contains 14 questions answered by 58 students. The answers were graded by two expert human raters based on model answers, again on a scale of 0-5.

All datasets have less than (total number of questions $\times$ total number of students) answers as presumably some students did not answer some questions. We mark such missing entries as "No Answer" and corresponding groundtruth scores as zero.

Metrics: A wide variety of evaluation metrics has been used in the literature for measuring goodness of ASAG techniques. We use Pearson's r in this paper as it has been one of the most popular metrics though its appropriateness have been questioned (Mohler and Mihalcea, 2009). For every question we compute Pearson's r between groundtruth and predicted scores and average across all questions are reported.

4.2 Quantitative Results

We compare the proposed technique against word similarity based ASAG techniques briefly described in Section 2. The basic premise of word similarity based ASAG techniques is: higher the similarity between the model and a student answer, higher the score the latter receives. Given two texts, model answer M and a student answer S, we conduct standard pre-processing operations such as stopword removal and stemming. The score of S with respect to M is then defined as:

$$asym(M, S) = \frac{1}{k} \sum_{i=1}^{k} \max_{\mathbf{s}_j \in S} (sim(\mathbf{m}_i, \mathbf{s}_j)) \quad (1)$$

where $\mathbf{m}_i$ and $\mathbf{s}_j$ are pre-processed n-grams of M and S respectively and k is the number of n-grams in M. For $n = 1$, $\mathbf{m}_i$ and $\mathbf{s}_j$ are words of M and S; and k is the length of M with respect to number of words. $sim(.,.)$ is a textual similarity measure of one of the following types:

- **Lexical**: In this category, we consider lexical overlap (**LO**) between model and student answers. It is a simple baseline measure which looks for exact match for every content word (post pre-processing e.g. stopword removal and stemming).
- **Knowledge based**: These measures employ a background ontology to arrive at word level semantic similarity values based on various factors such as distance between two words, lowest common ancestor, etc. Mohler and Mihalcea (2009) compared eight different knowledge-based measures to compute similarities between words in the model and student answers using Wordnet (Miller, 1995). We select the two best performing measures from their work viz. shortest path (**SP**) and the measure proposed by Jiang and Conrath (**JCN**) (Jiang and Conrath, 1997).
- **Vector space based**: In this category, we have chosen one of the most popular measures of semantic similarity, namely, Latent Semantic Analysis (**LSA**) (Landauer et al., 1998) trained on a Wikipedia dump. We also consider the recently popular word2vec tool (**W2V**) (Mikolov et al.,) to obtain vector representation of words which are trained on 300 million words of Google news dataset and are of length 300. Both LSA and W2V build on several related ideas towards capturing importance of context to obtain vector representation of words e.g. the distributional hypothesis "Words will occur in similar contexts if and only if they have similar meanings" (Harris, 1968). Similarity between words is measured as the cosine distance between corresponding word vectors in the resultant vector space using the well known dot product formula.

4.3 Results

In this section first we present comparative performance of the proposed technique against word similarity based ASAG techniques. Secondly, we expose a vulnerability of word similarity based ASAG techniques owing to their sole reliance on instructor provided model answer and thereby bringing out another benefit of the proposed technique.

Performance with respect to instructor provided model answers: Table 3 shows comparative performances of the proposed technique against unsupervised ASAG techniques for CSD. For each question as well as aggregate across all questions, winners are emphasized. In aggregate, the proposed technique performs comparatively and better than LO and knowledge based measures (JCN and SP) but a few points worse than LSA and W2V. Secondly, it is evident that, no one technique consistently outperforms others. In fact, the proposed technique has more question wise *winners* than LO, SP, JCN, and LSA. For example, for (Q19) (which is our running example in this paper), it has much better performance than the rest supporting our argument that leveraging student answer corpus is effective. Finally, variations across questions are significant - while high correlation is achieved for Q1, Q15 etc. but they remained low for Q12 and Q16. Considering such high variations across different questions, it is unlikely that any one method would perform the best across all types of questions in a general setting.

Table 4 shows the overall performance of the

Q#	Proposed	LO	JCN	SP	LSA	W2V
1	0.56	0.79	0.84	0.81	0.78	**0.85**
2	**0.55**	0.42	0.34	0.40	0.27	0.35
3	0.32	0.33	0.27	0.46	0.51	**0.62**
4	0.75	0.81	0.81	**0.86**	0.80	0.81
5	0.69	0.68	0.68	0.83	**0.84**	0.73
6	0.67	**0.82**	0.75	0.76	0.73	0.78
7	**0.71**	0.49	0.49	0.50	0.64	0.63
8	0.68	**0.79**	0.73	0.74	0.67	**0.79**
9	**0.58**	0.55	0.57	0.54	0.43	**0.58**
10	0.67	0.73	0.67	0.50	0.66	**0.75**
11	0.54	0.43	0.49	0.61	0.52	**0.64**
12	**0.46**	0.14	0.14	-0.10	0.24	0.30
13	0.70	0.57	0.75	0.69	**0.76**	0.67
14	0.52	0.48	0.52	0.45	**0.68**	**0.68**
15	0.56	0.55	0.80	0.69	**0.95**	0.84
16	0.05	**0.30**	**0.30**	**0.30**	0.24	0.29
17	0.68	0.83	**0.84**	0.81	0.71	0.76
18	0.23	0.55	0.17	0.33	**0.61**	0.46
19	**0.51**	0.37	0.43	0.50	0.43	0.44
20	**0.69**	0.55	0.51	0.35	0.61	0.50
21	0.76	0.78	0.76	0.68	0.75	**0.80**
Agg.	0.57	0.57	0.56	0.56	0.61	**0.63**

Table 3: Question wise Pearson's r of the proposed technique against unsupervised ASAG techniques.

proposed technique against word similarity based ASAG techniques. The top row of the table shows the inter-annotator agreement (IAA) for all datasets. For CSD and XCSD, performance of the proposed technique is respectively comparable and better than IAA. For both the datasets, its performance is comparable or better than lexical and knowledge-based methods but about 0.06 worse than the vector space methods.

	CSD	XCSD	RCD
IAA	0.59	0.54	0.67
LO	0.57	0.67	**0.56**
JCN	0.56	0.65	0.55
SP	0.56	0.67	0.38
LSA	0.61	**0.73**	0.47
W2V	**0.63**	**0.73**	**0.56**
Proposed	0.57	0.67	0.41

Table 4: Comparison of Pearson's r of unsupervised ASAG techniques against the proposed technique.

Variation with changes in model answers: The word similarity based ASAG techniques suffer from a surprising shortcoming. Variation in model answers can significantly affect their performance. Consider another possible model answer of our running example "`based on number, types and and order of arguments in the function call.`" Replacing the instructor provided model answer with this is not expected to change human evaluation of student answers. However doing so make unsupervised ASAG techniques exhibit significant change in performance (Pearson's r for LO, JCN, SP, LSA and W2V get changed by 43%, 23%, 2%, 19% and 20% respectively). Towards systematically exploring this, for each question we select those student answers which were graded as perfect 5/5 by the instructor with respect to the model answer. We consider each of them (and the instructor provided model answer) as a model answer in turn and grade remaining student answers. Resulting variation is shown in Table 5 in terms of minimum and maximum Pearson's r obtained as well as standard deviation of r values.[3] Careful observation of Table 5 reveals that all word similarity based ASAG technique show variation in performance for almost all questions. In some cases correlation with human provided groundtruth scores goes to a low minimum (even negative) and very high maximum correlation. Standard deviation values indicate high degree of variation in Pearson's r. On the other hand, the proposed technique due to its independence with model answers, does not exhibit *any* fluctuation. It is interesting to note that for all questions the minimum correlation obtained by all word similarity based ASAG techniques is worse than corresponding correlation of the proposed technique. In fact, for multiple questions the proposed technique performs comparably to the maximum correlation obtained by the word similarity based ASAG techniques. While further studies will be required to understand the root causes of these variation, it is unlikely that any one method would work the best for all questions.

4.4 Qualitative Analysis

In this section, we provide qualitative analysis to address a few questions which an intrigued reader might have:

- **Proposed technique will work only if all/most student answers are perfect**: No - even if most student answers are imperfect

[3] Owing to space constraint we show numbers for one representative technique from lexical, knowledge-based and vector space categories but we note similar variations for other techniques too.

Qs.	LO			JC			W2V			Proposed
	Min	Max	SD	Min	Max	SD	Min	Max	SD	
1	0.45	0.79	0.13	0.46	0.84	0.15	0.44	0.85	0.13	0.56
2	0.42	0.75	0.14	0.28	0.75	0.19	0.35	0.80	0.19	0.55
3	0.23	0.41	0.06	0.22	0.50	0.11	0.41	0.62	0.07	0.32
4	0.39	0.97	0.16	0.39	0.97	0.16	0.40	0.99	0.17	0.75
5	0.43	0.68	0.09	0.43	0.68	0.09	0.62	0.79	0.05	0.69
6	0.58	0.82	0.08	0.53	0.75	0.07	0.67	0.81	0.05	0.67
7	0.33	0.66	0.08	0.37	0.67	0.08	0.54	0.73	0.06	0.71
8	0.30	0.79	0.16	0.34	0.74	0.13	0.20	0.79	0.20	0.68
9	0.42	0.61	0.05	0.45	0.66	0.06	0.41	0.63	0.06	0.58
10	0.62	0.73	0.04	0.51	0.67	0.06	0.62	0.81	0.07	0.67
11	0.10	0.51	0.10	0.00	0.58	0.11	0.28	0.64	0.08	0.54
12	0.01	0.59	0.18	0.06	0.63	0.19	0.23	0.69	0.15	0.46
13	0.48	0.62	0.04	0.64	0.75	0.03	0.63	0.75	0.04	0.70
14	0.48	0.57	0.04	0.52	0.52	0.00	0.54	0.68	0.07	0.52
15	0.07	0.82	0.54	0.39	0.91	0.17	0.21	0.84	0.16	0.56
16	-0.09	0.30	0.11	-0.09	0.30	0.10	-0.02	0.29	0.09	0.05
17	0.37	0.83	0.11	0.53	0.84	0.09	0.51	0.81	0.10	0.68
18	0.15	0.72	0.18	0.02	0.74	0.23	0.03	0.74	0.20	0.23
19	0.22	0.58	0.09	0.34	0.56	0.06	0.17	0.57	0.10	0.51
20	0.22	0.79	0.13	0.33	0.84	0.13	0.29	0.83	0.15	0.69
21	0.32	0.81	0.12	0.45	0.80	0.09	0.42	0.83	0.10	0.76

Table 5: Fluctuation in performance (minimum (Min), maximum (Max) and standard deviation (SD)) of ASAG techniques with different model answers. The proposed technique does not exhibit any fluctuation.

the proposed technique could work well. This is because partially correct answers contribute towards boosting up support values of common patterns. For example, only 3 out of 31 students got perfect $5/5$ in (Q2) of CSD: `` What stages in the software life cycle are influenced by the testing stage?''. In spite of that proposed method has significantly better correlation than all word similarity based ASAG techniques. It is obvious that the proposed technique would perform the best when all answers are correct in the same manner and worst if in an unlikely case all are wrong in the same manner.

- **This will not work if there are multiple correct answers**: As long as a large enough fraction of students have written a correct answer, the proposed technique would work. Among various types, example seeking questions would fall in this category. Consider Q4 of RCD: ``Give two examples of people who are most vulnerable to RSI.'' where proposed method (0.40) has a better correlation than W2V (0.26) and JC (0.33). In fact, similarity-based techniques would not work well, if correct examples are not semantically similar.
- **Instructors won't have any control on assessment**: True - while we do not see this as a major drawback, we are working on an extension which will offer teachers more control.

5 Conclusion and Future Work

In this paper, we present a novel and intuitive finding in ASAG and propose a truly unsupervised and simple technique. The proposed method intelligently exploits structure of the ASAG problem to offer comparable performance to knowledge-based measures which depend on human curated Wordnet built over years and vector space-based measures which are trained on an astronomically large corpora. While the proposed technique is based on the assumption of *wisdom of students*, we intend to work on validating its correctness in broader settings including non-English and Science subjects. We believe that the proposed technique would, at the least, serve as a strong baseline for future ASAG research. Noting the wide variation in performance of different measures across questions, we are working towards exploring to bring together the proposed technique with word similarity based techniques. Finally, we see ASAG as an important line of research with the growing popularity of Massive Online Open Courses (MOOCs) and their limited assessment capability based solely on recognition questions.

References

Rakesh Agrawal and Ramakrishnan Srikant. 1995. Mining Sequential Patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14.

Enrique Alfonseca and Diana Pérez. 2004. Automatic Assessment of Open Ended Questions with a Bleu-Inspired Algorithm and Shallow NLP. In *EsTAL*, volume 3230 of *Lecture Notes in Computer Science*, pages 25–35. Springer.

Lyle F. Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J. Pan, Chris Salvador, and Yasuyo Sawaki. 2002. A Reliable Approach to Automatic Assessment of Short Answer Free Responses. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 1998. *Computer Analysis of Essays*. Educational Testing Service.

Jill Burstein, Susanne Wolff, and Chi Lu. 1999. Using lexical semantic techniques to classify free-responses. In *Breadth and depth of semantic lexicons*, pages 227–244. Springer.

Myroslava O. Dzikovska, Rodney D. Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013. SemEval-2013 task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. Technical report, DTIC Document.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *AAAI*, pages 1301–1306. AAAI Press.

Wael H. Gomaa and Aly A. Fahmy. 2012. Short Answer Grading Using String Similarity and Corpus-based Similarity. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 3(11).

Arthur C. Graesser, Peter M. Wiemer-Hastings, Katja Wiemer-Hastings, Derek Harter, and Natalie K. Person. 2000. Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor. *Interactive Learning Environments*, 8(2):129–147.

Christian Gütl. 2008. Moving Towards a Fully Automatic Knowledge Assessment Tool. *International Journal of Emerging Technologies in Learning*, 3(1):1–11.

Zellig Harris. 1968. *Mathematical Structures of Language*. John Wiley and Son, New York.

Simon Jaillet, Anne Laurent, and Maguelonne Teisseire. 2006. Sequential Patterns for Text Categorization. *Intelligent Data Analysis*, 10(3):199–214.

Jay J. Jiang and David W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, pages 19–33.

Victoria Johansson. 2009. Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective. *Working Papers in Linguistics*, 53:61–79.

Dharmendra Kanejiya, Arun Kumar, and Surendra Prasad. 2003. Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA. In *Proceedings of the HLT-NAACL workshop on Building Educational Applications using Natural Language Processing*, pages 53–60.

Richard Klein, Angelo Kyrilov, and Mayya Tokman. 2011. Automated assessment of short free-text responses in computer science using latent semantic analysis. In *ITiCSE*, pages 158–162. ACM.

Thomas K Landauer, Peter W. Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated Scoring of Short-answer Questions. *Computers and the Humanities*, 37(4):389–405.

Elena VM Lieven. 1978. *Conversations Between Mothers and Young Children: Individual Differences and their Possible Implication for the Study of Language Learning*.

Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated Scoring of a Summary Writing Task Designed to Measure Reading Comprehension. In *Proceedings of the 8th workshop on innovative use of nlp for building educational applications*, pages 163–168.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards Robust Computerized Marking of Free-Text Responses. In *Proceedings of International Computer Aided Assessment Conference (CAA)*.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 567–575.

Michael Mohler, Razvan C. Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 752–762.

Rodney D. Nielsen, Jason Buckingham, Gary Knoll, Ben Marsh, and Leysia Palen. 2008. A Taxonomy of Questions for Question Generation. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical report, IBM Research Report.

Diana Pérez, Enrique Alfonseca, and Pilar Rodríguez. 2004. Application of the BLEU Method for Evaluating Free-text Answers in an E-learning Environment. In *LREC*. European Language Resources Association.

Diana Pérez, Alfio Gliozzo, Carlo Strapparava, Enrique Alfonseca, Pilar Rodriguez, and Bernardo Magnini. 2005. Automatic Assessment of Students' Free-Text Answers Underpinned by the Combination of a BLEU-Inspired Algorithm and Latent Semantic Analysis. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference (FLAIRS)*.

Diana Pérez-Marín and Ismael Pascual-Nieto. 2011. Willow: a system to automatically assess students free-text answers by using a combination of shallow NLP techniques. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):155–169.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying Patterns for Short Answer Scoring using Graph-based Lexico-semantic Text Matching. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106.

Shourya Roy, Y. Narahari, and Om D. Deshmukh. 2015. A Perspective on Computer Assisted Assessment Techniques for Short Free-Text Answers. In *Proceedings of the International Conference on Computer Assisted Assessment (CAA)*, pages 96–109. Springer.

Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Jana Z. Sukkarieh, Ali Mohammad-Djafari, Jean-Franc¸ois Bercher, and Pierre Bessie´re. 2011. Using a MaxEnt Classifier for the Automatic Content Scoring of Free-text Responses. In *Proceedings of the AIP Conference American Institute of Physics*, volume 1305, page 41.

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. 2006. *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston.

Hao-Chuan Wang, Chun-Yen Chang, and Tsai-Yen Li. 2008. Assessing Creative Problem-solving with Automated Text Grading. *Computers and Education*, 51(4):1450–1466.

P. Wiemer-Hastings and I. Zipitria. 2001. Rules for Syntax, Vectors for Semantics. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Erlbaum.