A framework for effectively utilising human grading input in automated short answer grading

Andrew Kwok-Fai Lui*,
Sin-Chun Ng and
Stella Wing-Nga Cheung

School of Science and Technology,
Hong Kong Metropolitan University,
Hong Kong SAR, China
Email: andrew.lui@computer.org
Email: sin.ng@aru.ac.uk
Email: wncheung@hkmu.edu.hk
*Corresponding author

Abstract: Short answer questions are effective for recall knowledge assessment. Grading a large amount of short answers is costly and time consuming. To apply short answer questions on MOOCs platforms, the issues of scalability and responsiveness must be addressed. Automated grading uses a computing process and a machine learning grading model to classify answers into correct, wrong, and other levels of correctness. The divide-and-grade approach is proven effective in reducing the annotation effort needed for the learning the grading model. This paper presents an improvement on the divide-and-grade approach that is designed to increase the utility of human actions. A novel short answer grading framework is proposed that addresses the selection of impactful answers for grading, the injection of the ground-truth grades for steering towards purer final clusters, and the final grade assignments. Experiment results indicate the grading quality can be improved with the same level of human actions.

Keywords: automated short answer grading; clustering; semi-supervised clustering; MOOCs; automated grading.

Reference to this paper should be made as follows: Lui, A.K-F., Ng, S-C. and Cheung, S.W-N. (2022) 'A framework for effectively utilising human grading input in automated short answer grading', *Int. J. Mobile Learning and Organisation*, Vol. 16, No. 3, pp.266–286.

Biographical notes: Andrew Kwok-Fai Lui received the PhD degree from The Australian National University, Canberra, ACT, Australia, in 1998. He is currently a Professor with the Department of Technology, Hong Kong Metropolitan University. His current research interests include computational intelligence, traffic modelling, evolutionary computation, and computer science education.

Sin-Chun Ng received the BSc degree (Hons) in Information Technology and the PhD degree from the Department of Electronic Engineering, City University of Hong Kong, in 1990 and 2000, respectively. She was a Professor with the School of Science and Technology, Hong Kong Metropolitan University when this work was delivered. Her research interests include evolutionary computation, neural networks, multimedia technology, and educational technology.

Stella Wing-Nga Cheung received the BSc degree (Hons) in computing from the Open University of Hong Kong in 2013 and the MSc degree with Distinction from the Department of Computer Science, City University of Hong Kong in 2014. She is currently a Research Assistant with the School of Science and Technology, Hong Kong Metropolitan University.

This paper is a revised and expanded version of a paper entitled 'Effective Use of Small Amount of Graded Answers for Automated Grading of Short Answers' presented at 'International Conference in Open and Innovative Education', Hong Kong SAR, July 2020.

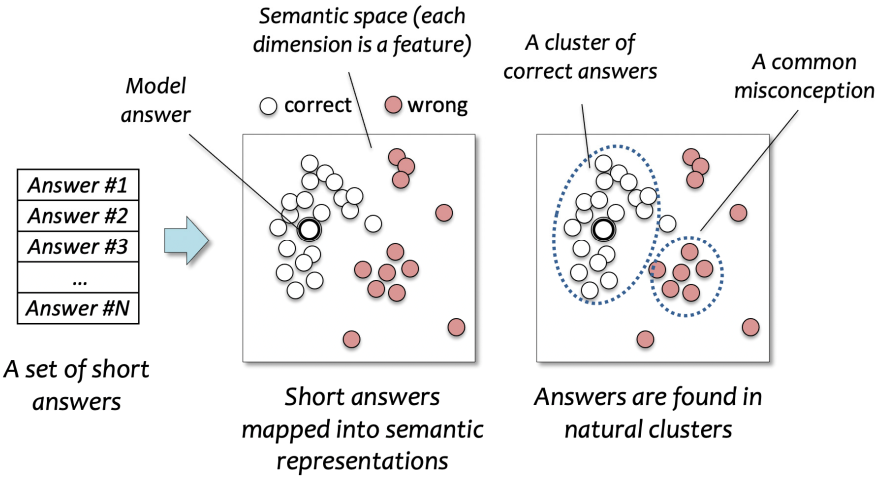
1 Introduction

Short answer questions are commonly used for knowledge assessment in mobile and online learning. These questions are designed to solicit short text answers that represent the knowledge sought by the questions. In addition to multiple choice questions, short answer questions were found also a preferred form of mobile assessment (Bogdanović et al., 2014). They are able to accurately measure the recall of knowledge in a specific topic. As students are required to compose short answers with their own writing, luck plays a lesser role and the latent thinking process is better revealed. The free text nature also allows for assessment of partial correctness and awarding of partial credits. Text composition on a mobile device is now commonplace as instant messaging has become a part of daily life. Many people are adept at input text answers using on-screen keyboard or voice-to-text conversion technology (Ha and Zhang, 2019).

In the past few years, more and more learners were adding to the already popular Massive Open Online Course (MOOC) platforms (Shah, 2019). Computer-graded multiple-choice questions are the default question type for formative assessment. Short answer questions would contribute significantly to the online learning experience, but the concern in grading effort must first be resolved before a widespread use becomes feasible. The current solutions such as peer grading are inadequate in the consistency and accuracy. Automated grading with artificial intelligence technologies can substantially reduce manual grading cost, provide superior consistency compared to peer grading, and shorten the turn-around time in large scale educational operations (Yu et al., 2017). A set of short answers often contains highly similar answers, which can be modelled and used to remove redundancies in grading effort.

Short answer questions are designed to elicit a short textual response from students whose knowledge of a topic is to be measured. Each question comes with one or more model answers that delineate the questioned knowledge in a specific manner. In general, responses found to be semantically sufficiently similar to a model answer are considered as correct. Other responses may be simply considered as wrong. In attempting a question, a specific and concise response is sought as students recall relevant knowledge and manifest the knowledge in writing. The correct answers are usually very similar to each other and also the model answer. Some wrong answers are also very similar to each other if there are common misconceptions among the students. In addition to these clusters of common answers, there are also anomalous answers made up arbitrarily by some clueless students. Figure 1 illustrates these natural aggregations among short answer sets.

Figure 1 The specific nature of short answers tends to induce sets of similar answers aggregating around model answers and common misconceptions

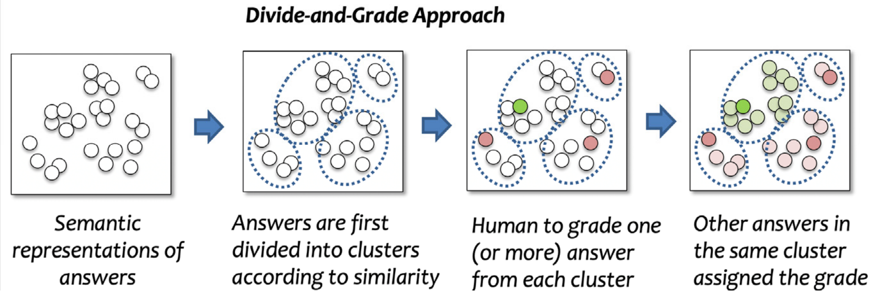

Semantic space (each dimension is a feature)
Model answer
○ correct
● wrong
A cluster of correct answers
A common misconception
Answer #1
Answer #2
Answer #3
...
Answer #N
A set of short answers
Short answers mapped into semantic representations
Answers are found in natural clusters

1.1 Automated short answer grading

Automated grading uses a computing process and a grading model to divide answers into correct, wrong, and other levels of correctness. An automated grader can differentiate the semantics of short answers and label the answers according to a sample of ground-truth graded answers. The definition of correct and wrong is inherently subjective. The input of human graders is necessary for specifying the grading model. The amount of human engagement is an important consideration from the cost perspective. The benefit of automated grading is greatly diminished if substantial human grading input is needed.

Among the existing Automated Short Answer Grading (ASAG) techniques, the divide-and-grade approach (Basu et al., 2013; Brooks et al., 2014; Zesch et al., 2015; Horbach and Pinkal, 2018) is designed to make effective use of human grading input and to reduce human engagement. The approach recognises that some grading actions can worth more than others, depending on how impactful the answers are. For example, in a cluster of highly similar answers, these answers should have the same ground-truth grade. The first answer to be graded in the cluster is the most impactful, because the ground-truth grade may be propagated to the whole cluster. Grading the other answers may be considered at best for validation and at worst redundant. The divide-and-grade approach utilises cluster analysis to identify the natural clusters of common answers, and then select one or two answers from every cluster for grading. The selected answers are grading impactful, and the number of grading actions is greatly reduced. Figure 2 illustrates how the divide-and-grade approach can efficiently complete the grading of an answer set.

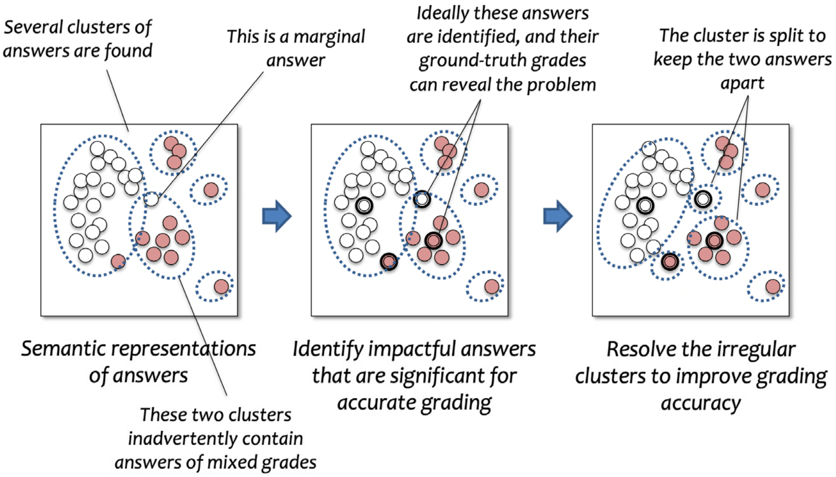
Figure 2 The divide-and-grade approach of ASAG can greatly reduce human grading effort through rounding up similar answers first such that one or two grading actions would be sufficient to determine the grade of the whole clusters


Divide-and-Grade Approach
Semantic representations of answers
Answers are first divided into clusters according to similarity
Human to grade one (or more) answer from each cluster
Other answers in the same cluster assigned the grade

1.2 Aim and structure of the paper

The aim of this paper is to present an improvement for the divide-and-grade approach that further enhances the effective utilisation of human grading actions. The cluster analysis stage plays a pivotal role in the identification of the natural clusters in answer sets and the impactful answers for grading. The performance often suffers from noise (e.g. misspellings), anomalous answers and marginal answers (i.e. answers that are borderline correct or wrong). The resulting clusters may contain answers of mixed grades. The external knowledge of ground-truth grades should help the cluster analysis resolve these irregularities and improve the grading accuracy, as illustrated in Figure 3. The paper will propose a novel cluster analysis framework that can make more effective use of human grading actions by utilising them both in the cluster analysis stage and the grading stage. A number of formulations for identifying the impactful answers for the cluster analysis will also be proposed and evaluated.

Figure 3 Improve the divide-and-grade approach with identifying answers that are significant, positively or negatively, to grading accuracy, and then to obtain their ground-truth grades through human grading, and finally to utilise them in steering the cluster analysis


Several clusters of answers are found
This is a marginal answer
Ideally these answers are identified, and their ground-truth grades can reveal the problem
The cluster is split to keep the two answers apart
Semantic representations of answers
Identify impactful answers that are significant for accurate grading
Resolve the irregular clusters to improve grading accuracy
These two clusters inadvertently contain answers of mixed grades

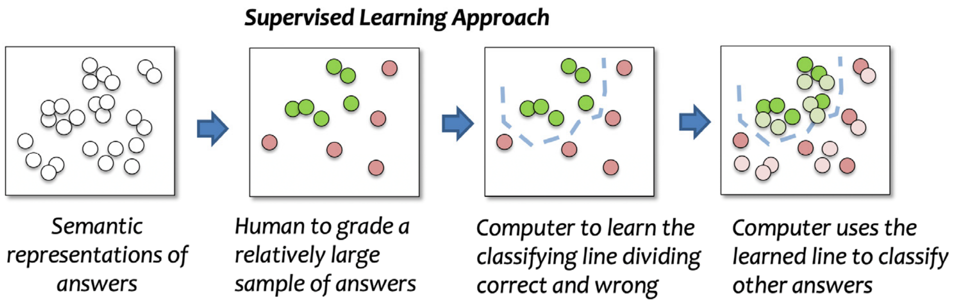
The remainder of the paper is structured as follows. The next section will review the rationale, the operation, the strengths and weaknesses of the divide-and-grade approach, and in particular discuss the value of human grading input. The section will conclude with a research question that the rest of the paper will address. Section 3 will describe the new automated short answer clustering framework and discuss the designs and the expected values of the key components. Section 4 and 5 will describe the experimental results and the significance of the findings. The paper is then concluded with a summary and suggestions for further research.

2 Literature review

ASAG uses a computing process and a grading model to differentiate student answers into correct, wrong, and other levels of correctness. The grading model contains specifications of correct and wrong answers with respect to their semantics or meanings. The mainstream ASAG research often apply the technology of supervised machine learning to augment the grading model. A sample of ground-truth graded answers is prepared and then used to train the grading model so that the model can generalise the specification and predict the correctness of unseen answers. Figure 4 illustrates these steps in the supervised learning approach. A larger training sample generally leads to better grading accuracy. Examples of this supervised learning approach from the literature include CoMiC-EN (Meurers et al., 2011; Mohler and Mihalcea, 2009; Mohler et al., 2011), and ETS (Heilman and Madnani, 2013). Burrows et al. (2015) conducted a comprehensive review of this approach and highlighted some of its strengths and weaknesses.

A major weakness of the supervised learning approach is ineffective utilisation of human grading input. The large training sample usually comes from grading a collection of real submissions. Some grading actions are unavoidably redundant, and the effort is wasted. A filter may be applied to catch the lexicographically equal answers, but other highly similar answers still require human input. In addition, each question needs a specialised grading model, and the grading model built for one question cannot be transferred to another. For large-scale deployment of automated grading, another approach that will make effective and yet sparing use of human grading input is desired.

Figure 4 An illustration of the supervised learning approach of ASAG. White represents unknown grade. Green and red represent two classes of ground-truth grades from which the classifying line is learned. The lighter green and red are the grades given by the classifier to unseen answers


Supervised Learning Approach
Semantic representations of answers
Human to grade a relatively large sample of answers
Computer to learn the classifying line dividing correct and wrong
Computer uses the learned line to classify other answers

2.1 The divide-and-grade approach and misclustering errors

The divide-and-grade approach of ASAG facilitates the selection and the grading of impactful answers through cluster analysis. Basu et al. (2013) was the first, to the best of our knowledge, to propose a divide-and-grade method to short answer grading. A clustering algorithm was employed to identify clusters and sub-clusters of similar answers, and then a human grader was engaged for assessing the answers in every cluster and grading them. Experiments showed that a significant reduction of human grading actions and therefore each action was more effectively utilised. Another finding was that the strengths of cluster analysis and human grading could be better utilised in an interactive setting. Brooks et al. (2014) extended the work and developed an interactive grading system that had integrated clustering with feedback tasks. Feedback is an important and effective instructional means. A useful side effect of identifying clusters of answers is to generate feedback efficiently at scale, because all answers in the same cluster can receive the same feedback.

A pure cluster contains answers of the same ground-truth grade (e.g. correct and wrong). A single grading action is sufficient for grading a pure cluster, but grading errors occur if the cluster is impure. In practice, cluster analysis has no certain means to prevent clusters with mixed ground-truth grades. The semantics of answers are used for estimating the ground-truth grade differences and such estimating is never perfect. Basu et al. (2013) and Brooks et al. (2014) addressed this problem of misclustering grading errors with an interactive corrective method. More than one grading action is suggested for each cluster. The human grader should examine the answers of a cluster one by one until a decision could be made that either the cluster appeared to be pure, or the cluster looked to have many impurities and all answers have to be human graded. The sub-clustering notion suggested in Basu et al. (2013) was a ploy to hopefully break down an impure cluster into smaller sub-clusters with the hope that some sub-clusters would be pure, and if some sub-clusters could still be impure, their smaller size would reduce the overall grading actions. This interactive corrective method needs additional human grading actions as a human grader makes informed decisions between more reliable grading and less grading effort.

2.2 More effective use of grading actions for divide-and-grade

A more promising method to alleviate the problem of misclustering grading errors is to take preventive measures in the cluster analysis stage and to facilitate the output of purer clusters. If a cluster is known to be impure during cluster analysis, then actions may be taken to resolve the issue. Knowledge of some ground-truth grades should therefore help guide cluster analysis to find purer clusters. For example, knowing that a correct and a wrong ground-truth grades are found in the same cluster is useful to indicate the cluster is undesirable. The model answers of a question are essentially ground-truth grades that can reduce grading actions (Basu et al., 2013) and can serve as the seeds for correct clusters (Zhang et al., 2016).

The preventive measures require gathering ground-truth grades and making them available for cluster analysis. In addition to model answers, more ground-truth grades are obtained simply by engaging a human grader earlier. These ground-truth grades will be exploited in the cluster analysis as well as the final grading of clusters. The grading actions are more effectively utilised. Research on the preventive method only arose recently resulting in very few published papers based on semi-supervised machine learning (Wang et al., 2019; Horbach and Pinkal, 2018). A factor to make the preventive measures effective is the selection of impactful answers for the early grading. The notion of impactful is however multifaceted. An impactful answer may be one that best represents the grade of a cluster. Another impactful answer may be a marginal answer and therefore an uncertainty between clusters. The ground-truth of the former leads to more accurate grading and the latter help cluster analysis decide cluster boundary. More understanding about the selection of impactful answer is needed for the divide-and-conquer approach.

2.4 Cluster analysis for the divide-and-grade approach

The objectives of cluster analysis is to produce clusters of similar answers. A number of off-the-shelf clustering algorithms have been employed and evaluated for the divide-and-grade approach including centroid-based methods (Basu et al., 2013), hierarchical connection-based methods (Zhang et al., 2016) and density-based methods (Cai et al., 2016). Each of these algorithms perform well generally, but poor in adapting to answer sets with too few or too many aggregations, aggregations with varying densities, too many outliers, and other distribution issues that could emerge.

A suitable clustering algorithm for the divide-and-grade approach should satisfy a number of challenging and competing technical requirements. First, it should allow the injection of external knowledge in the form of ground-truth grades and utilise the knowledge effectively in the clustering process. Second, it should enable the identification of more impactful answers. Third, it should adapt to the preference of human graders through producing optimal solutions of different number of clusters.

The framework of multi-objective evolutionary clustering considers cluster analysis as a multiple objective optimisation problem. A multi-objective optimisation algorithm is capable of producing a set of optimal solutions according to two or more objective functions. An evolutionary algorithm can search for optimal solutions based on biological mechanism of reproduction of a population of solutions and natural selection of better solutions. The framework should work effectively for the divide-and-grade approach. The objective functions offer a flexible structure for indicating favourable or unfavourable clustering due to external knowledge and other preferences such as cluster numbers. The objective functions may be updated during the evolutionary process to enable interactive steering of the clustering process. Handl and Knowles (2007) proposed an efficient multi-objective evolutionary clustering algorithm and Garza-Fabre et al. (2017) made further improvements in efficiency and scalability.

2.5 Summary

It was found that the divide-and-grade approach can make human grading actions effectively utilised and therefore reduce grading effort compared to the mainstream supervised learning approach. The key to the success of the approach is the cluster analysis finding pure clusters in the answers. The injection of ground-truth grades of a small sample of answers into cluster analysis appears to be useful to improve the purity of clusters. The human grading actions are more effectively utilised because each action is used for cluster analysis as well as the final grading.

The research question is how to design a divide-and-grade framework that can effectively select and utilise a small graded sample for both cluster analysis and the final grading. To answer the question, a framework based on multi-objective evolutionary clustering will be proposed and evaluated. The framework will allow the investigation of the selection of impactful answers for grading and the injection of ground-truth grades into cluster analysis.

3 A novel divide-and-conquer short answer grading framework

The novel divide-and-grade short answer grading framework (NDSAGF) consists of several components.

- A model of semantic representations of short answers.
- A multi-objective evolutionary short answer clustering algorithm.
- A method of using ground-truth grades to improve clustering.
- A method of selecting impactful answers.
- A method of grading the answers in clusters.

To cater for the intended audience of this paper, the technical content is kept to a minimal and pitched at a conceptual level.

3.1 Semantic representations of short answers

A computable semantic representation of short answers is needed for the input to the divide-and-grade procedure. The representation consists of a set of features each of which captures a certain semantic aspect in the short answers. The early days of ASAG research relied on manually engineered features. The statistical features such as word frequencies consider only the words in student answers, model answers and questions, and enriched with techniques such as TF-IDF (Basu et al., 2013) and Latent Semantic Analysis (LSA) (Mohler and Mihalcea, 2009). The corpus-based features, such as part-of-speech, named entities, and textual entailment (Mohler et al., 2011), are based on external knowledge about the semantics of words and sentences such as a large labelled corpus (Gomaa and Fahmy, 2012) or a lexical resource (Vii et al., 2019). Any semantic aspect may be significant for a particular short answer question, and a rich semantic presentation is preferred. An effective approach is to aggregate a large number of manually engineered features in a scattergun approach with the hope that some combinations may produce favourable outcomes. Examples of this approach include stacking (Heilman and Madnani, 2013), ensemble (Roy et al., 2016), and feature selection through an ablation study (Sahu and Bhowmick, 2019).

The most effective model for mapping short answers into their semantic representations is based on deep-learning distributional semantics (Boleda, 2020). The semantic features are computer extracted from analysing a gigantic corpus. For examples, word embedding models and sentence embedding models can discover rich and compact semantic features. The semantics of a short answer can be represented as an aggregation of the semantics of its words and sentences. Sakaguchi et al. (2015) developed similarity features based on a pre-trained word embedding model and likewise Menini et al. (2019)

used a pre-trained sentence embedding for ASAG. The most recent models, such as Google's Universal Sentence Encoder (Cer et al., 2018), are superior in the richness of features and the versatility in different domains.

3.2 The evolutionary short answer clustering algorithm

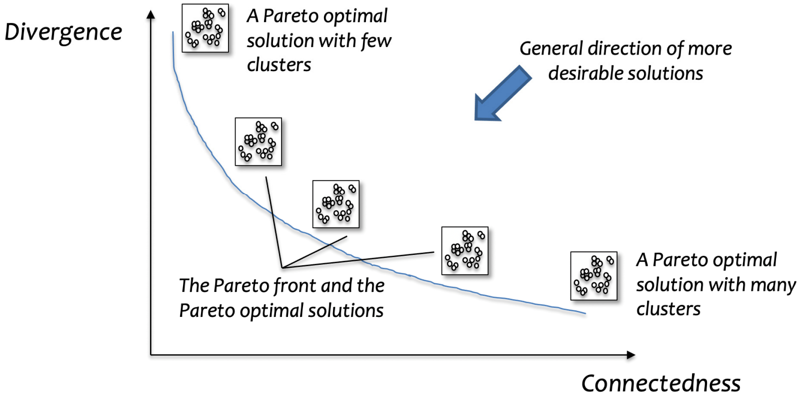
There are numerous ways of dividing a set of short answers into clusters. An optimisation formulation of clustering states that the best clustering solution optimises one or more measures of desirable clusters. In addition to the common measures of good clusters including compactness within same clusters and separation between clusters, new measures based on knowledge of ground-truth grades will favour the selection of purer clusters as the optimal solutions. The injection of ground-truth grades allows estimation of cluster purity and other desirable clustering qualities.

The proposed short answer clustering framework is based on the multi-objective evolutionary clustering algorithm developed by Garza-Fabre et al. (2017) and originated from Handl and Knowles (2007). The evolutionary approach can find optimal solutions through evolving a population of potential solutions. The poor solutions are gradually replaced with good solutions through the quasi-evolutionary operations of mutations and crossovers. The optimal solutions, according to optimising more than one objective function, are a set of Pareto optimal solutions which mean no objective function can be improved without degrading some other objectives. The aforementioned measures of short answer clustering have conflicting objectives. The multi-objective evolutionary clustering algorithm can resolve the conflict and bring practical benefits that will be described in Figure 5 and in the following paragraphs. Only the relevant key points of the algorithm are outlined and the details of the algorithm can be found in the references.

- Graph-based encoding. Each clustering solution is represented by a set of connecting edges between short answers. Two connected answers are considered to be in the same cluster.
- Edge initialisation with minimum spanning tree. A minimum spanning tree (MST) is made up of a subset of edges that connect all answers with the minimum total distance. The MST is used to determine the initial connections for the solutions as it provides a near-optimal sub-structure.
- Cluster initialisation with removal of interesting edges. An interesting edge is likely to be a connection between two clusters. An edge's interestingness is estimated from both its length and whether the two nodes are both the more distant nodes of each other. The initial clusters are formed from their removal.
- Evolutionary operators. The mutation operator randomly cuts an edge in a solution and replaces it with another edge randomly selected from the most similar answers of the originator of the removed edge. The crossover operator forms a new solution from edges selected from two best solutions.
- Objective functions. The two basic objective functions are the cluster *compactness* and the cluster *connectedness*. The former favours solutions with more clusters. The latter is a more calculation efficient version of the between cluster separation measure, and it favours solutions with fewer clusters.

- Solutions with different cluster numbers. Due to the nature of the two objective functions, the Pareto solution set consists of solutions of different cluster numbers. Human graders can select one according to their grading budgets.

Figure 5 The multi-objective evolutionary clustering algorithm will find a number of Pareto optimal solutions and the cluster numbers of these solutions range from few to many. The two dimensions represent the two objective functions, in which divergence is the opposite of compactness


Divergence
A Pareto optimal solution with few clusters
General direction of more desirable solutions
The Pareto front and the Pareto optimal solutions
A Pareto optimal solution with many clusters
Connectedness

3.3 *Using ground-truth grades to prevent impure clusters*

Answers with different ground-truth grades should not be in the same clusters. For example, answer pairs consisting of a correct answer and a wrong answer are considered as no-link pairs. These no-link pairs can be used to invalidate clusters in the evolutionary clustering algorithm. The following lists some possible schemes.

- Algebraic scheme. Each no-link violation should incur a penalty in the objective functions.
- Set-theoretic scheme. Any no-link violation should incur a penalty in the objective functions.
- Termination scheme. A solution has a non-link violation in a cluster should be terminated.

For the first two schemes, the penalty may be in the form of an additional term in the basic objective functions or as the third objective function.

3.4 *Selecting impactful answers for grading*

For every cluster two types of answers are considered impactful for grading. The first type represents the grade of the whole cluster. The answer nearest to the centroid of the cluster should offer the most reliable representation. The second type represents the marginal answers that misclustering is more likely. The answers near the boundary of the cluster should be the most uncertain answer that may attach to a nearby cluster instead. These three ways may be used to select the impactful boundary answer.

- The answer furthest from the centroid.
- The answer connected by the most interestingness edge.
- The answer connected by the longest edge.

The identification of impactful answers of a cluster has a practical issue on the most suitable moment.

- At the end of cluster analysis. The clusters are at the optimal quality and the identification of impactful answers is most reliable, but the ground-truth grades of the impactful answers cannot contribute to prevention of impure clusters.
- At the start of the cluster analysis. The clusters come from initialisation and are sub-optimal. The identification of impactful answers is less reliable but their ground-truth grades can make more impact on the cluster analysis stage.

The midway alternative is to identify and grade the impactful answers during cluster analysis. The selection impactful answer is somewhat reliable and the ground-truth can still influence the later generations of the evolutionary clustering algorithm.

3.5 Final grading of clusters

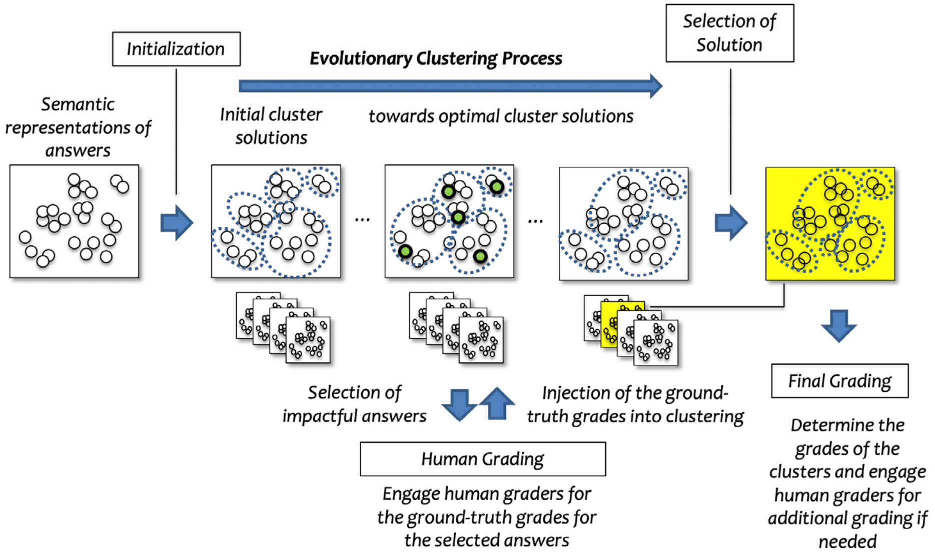
The final grading stage begins after the completion of the cluster analysis stage. In the original divide-and-grade approach, one or more answers are selected from each cluster for human grading. In the proposed short answer clustering framework, there are new steps regarding the grading stage. The first new step is the human selection of one Pareto optimal solution from the evolutionary clustering algorithm. It is assumed that the human grader has a preference on number of grading actions and accordingly select the solution with the preferred number of clusters. In addition, the human grader can make reference to external measures such as the gap statistics (Tibshirani et al., 2001) for an optimal number of clusters.

The second new step is the exploitation of previously graded impactful answers. The clusters of the selected Pareto solution are examined one by one. The target is to have at least one ground-truth grade per cluster, or for better reliability two or more ground-truth grades. A human grader is engaged to provide the additional ground-truth grades for a cluster if needed, and the more impactful answers should be selected.

At the end, if the one or more ground-truth grades of a cluster are the same, that same grade is used for the whole cluster. If there is an inconsistency, that is a correct and a wrong answer, the cluster is split into two smaller clusters according to the nearer ground-truth answer.

Figure 6 describes how the components of the novel short answer grading framework are working together.

Figure 6 The novel divide-and-grade short answer grading framework based on an evolutionary clustering algorithm and augmented with injection of ground-truth answers


Initialization
Evolutionary Clustering Process
Selection of Solution
Semantic representations of answers
Initial cluster solutions
towards optimal cluster solutions
...
...
Selection of impactful answers
Injection of the ground-truth grades into clustering
Human Grading
Engage human graders for the ground-truth grades for the selected answers
Final Grading
Determine the grades of the clusters and engage human graders for additional grading if needed

4 Method

To answer the research question, a number of experiments have been carried out to evaluate the proposed novel short answer clustering framework. A prototype implementation of the framework has been developed with the programming language of Python and several data mining modules including Sklearn. In addition, a framework for evolutionary computing called Platypus was used as the basis of multi-objective evolutionary optimisation.

4.1 Data

The Powergrading dataset (Basu et al., 2013) was selected for evaluation. The dataset consists of 20 questions selected from the United States Citizenship Examination. The answers have been 3-way manually annotated (correct, unsure, and wrong) by 3 independent human markers, and the inter-annotator agreement was found to be satisfactorily. For the purpose of this study, the unsure and the wrong grades were recoded to wrong, as opposed to the correct grade.

In the dataset, larger sets of answers were found for question 1 to question 8, with each set has nearly 700 answers. The answer sets of these 8 questions have different average length that could give different challenges to automated grading. These 8 questions were selected in this study and their question text and statistics of their answer sets are shown in Table 1. The selected answer sets were converted into 512-dimension semantic representations using a pre-trained Google's Universal Sentence Encoder TF2.0 Saved Model V4. In other words, each short answer is represented by 512 dense semantic features.

Table 1 The eight questions used for the evaluation study in this paper and the statistics of their answer sets

	Question	*# Correct answers*	*# Wrong answers*	*Average length*
Q1	What are the first ten amendments to the US Constitution called?	652	47	3 words
Q2	What is one right or freedom from the First Amendment?	613	85	3 words
Q3	What did the Declaration of Independence do?	567	138	8 words
Q4	What is the economic system in the USA?	561	137	2 words
Q5	Name one of the three branches of the US government.	657	41	2 words
Q6	Who or what makes federal (na- tional) laws in the USA?	415	283	3 words
Q7	Why do some states have more Representatives than other states?	650	48	6 words
Q8	If both the President and the Vice-President can no longer serve, who becomes President?	415	283	4 words

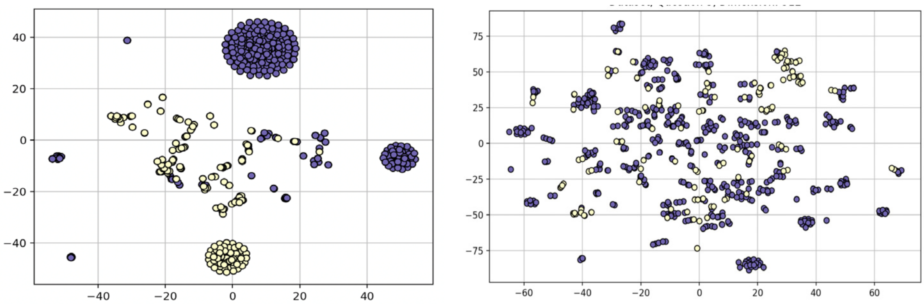
The answer sets of these questions have differences in the forms of answer aggregations and outlier distributions. For example, many correct answers in Q1 are in very few large clusters and most cluster analysis algorithms can pick them out. Table 2 shows that there are 7 lexicographical versions of correct answers for Q1 and only insignificant differences separating them. On the other hand, there are many versions of wrong answers. The marginal answers consist of misspelling cases and conceptually marginal cases and their acceptance may depend on individual human graders.

Table 2 Number of lexicographical versions of the answer sets of Q1

Ground-truth grades	*# Lexicographical versions*	*Examples*
Correct	7	“The bill of rights”, “Bills of rights”, “Bill of rights”, “Bills of right”
Wrong	35	“i don’t know”, “freedom of speech.”, “peoples rights.”, “forgot..”
Marginal	7	“bill of righta”, “bill of rigthes”, “us constitution”, “constitutional rights”

Question 3 and question 8 demonstrated two characteristic compositions of answer modes and outliers. Figure 7 shows the distributions of answers of Q3 and Q8 in the semantic space defined by the model. It is clear that the Q8 answer set has a couple of large correct clusters, a larger wrong cluster and a number of small clusters and outliers. The Q3 answer set has many small clusters and analysis lot of outliers. The two questions pose different challenges to the identification and selection of impactful answers.

Figure 7 The answer set distribution in the semantic space (projected to two-dimensional plane for visualisation) of two selected questions of the Powergrading dataset. The left graph comes from Q8 and the right graph from Q3. Dark dots indicate ground-truth correct answers and light dots are wrong answers. The aggregation of answers is clear from the plots. The Q8 answer set contains a few large clusters. The Q3 answer set contains a lot more smaller clusters



5 Results and discussions

This section reports the findings of the evaluations, structured to illustrate the improvement on grading accuracy resulting from the use of ground-truth grades in the clustering algorithm, to show effective ways of selecting the impactful answers, and to paint a holistic view of the grading performance on different answer sets.

5.1 Utilisation of ground-truth grades for improved clustering

The injection of ground-truth grades into the clustering algorithm should lead to fewer impure clusters and better grading accuracy. A cluster found to have both correct and wrong answers should incur a penalty. This experiment was carried out to compare several schemes of implementing penalty in the objective functions. The settings of the experiment are summarised in the following.

- The selection of impactful answer and injection of ground-truth grades at half-point of the evolutionary clustering generation.
- The two impactful answers were selected from the centroid and the borderline answer with the highest connectedness.
- At the final grading, at least two ground-truth grades were used per cluster and a cluster was divided if a violation was found.

The two measures used for expressing grading accuracy are the accuracy and the F1-score, and they are defined in equations (1) and (2), where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F1\ score = \frac{TP}{TP + \frac{1}{2}\left(FP + FN\right)} \tag{2}$$

The F1 score is the harmonic mean between recall and precision and it is a better metric for imbalanced class distribution that exists in our datasets (i.e. correct answers are much more than wrong answers). The experiments were run with different random seeds and the mean of several runs was used after removal of outliers. For the multi-objective evolutionary clustering algorithm, the population size (i.e. the number of solutions) was set to 100 and the number of generations was set to 120.

The experimental results are summarised in Table 3. For each scheme, the cases of fewer preferred cluster of 30 and more preferred clusters of 50 were evaluated.

Table 3 Comparison of grading accuracies given different schemes of injecting ground-truth grades into the multi-objective evolutionary clustering algorithm

		Q3			*Q8*		
	Preferred cluster #	*# Grading actions*	*Accuracy*	*F1 score*	*# Grading actions*	*Accuracy*	*F1 score*
Evolutionary clustering baseline	30	46	0.8082	0.8849	47	0.9776	0.9778
Algebraic scheme (penalty per violation in the basic objectives)	30	38	0.8302	0.9055	63	0.9909	0.9924
	50	64	0.8766	0.9293	71	0.9909	0.9924
Algebraic scheme (penalty per violation as third objective)	30	65	0.8718	0.9249	40	0.9871	0.9893
	50	88	0.8867	0.9340	43	0.9785	0.9820
Termination scheme	30	Failed to converge					
	50	Failed to converge					
Set-theoretic scheme (fixed penalty in the basic objectives)	30	72	**0.8824**	0.9316	63	**0.9914**	0.9928
	50	148	**0.9053**	0.9431	73	**0.9933**	0.9944
Set-theoretic scheme (fixed penalty as third objective)	30	48	0.8766	0.9293	62	0.9847	0.9873
	50	114	0.8580	0.9105	67	0.9862	0.9884

The results show that injection of ground-truth grades at half-point of the clustering algorithm improved grading accuracy. The improvement is more significant for Q3, as the clusters distribution of the answer set in Q8 appears to be less ambiguous and the baseline already achieves a good accuracy. Among the three schemes evaluated, both algebraic and set-theoretic work well and achieve a similar performance. However, it was found that the termination scheme often made the evolutional clustering algorithm unable to continue. As the termination scheme would remove all old and new solutions with any violation, and the clustering algorithm would sometimes have no remaining solution to work on.

In this experiment, each cluster should have at least two ground-truth grades. There are a few reasons why the final grading actions are not always twice as many as the preferred cluster numbers. First, the actual cluster numbers were often not the preferred cluster numbers. The clustering algorithm is a stochastic process. Second, the ground-truth grades were likely to be re-distributed among the final clusters, and additional grading actions were required for clusters with fewer than two ground-truth grades. Third, some final clusters had only one member and no second grading action was needed.

The significance of achieving a higher grading accuracy is that the human effort spent on comparable grading actions is worth more. In the actual application of this approach, the human grader is expected to work with the automated grading system interactively. The early engagement of human grading in the clustering process could better steer the answer clusters to comply with the ground-truth.

5.2 Selection of impactful answers

For more effective utilisation of each human grading actions, the answers selected should maximise the benefits of knowing their ground-truth grades. This experiment was carried out to investigate different definitions of an impactful answer and to compare the grading performance when each one was adopted.

- The centroid of each cluster.
- The centroid of each cluster plus one borderline answer estimated based on maximum connectedness. A high connectedness means that many of the neighbours of an answer are in another cluster.
- The centroid of each cluster plus two borderline answers estimated based on maximum connectedness. This serves to demonstrate the impact of more grading actions per cluster.
- The centroid of each cluster plus one borderline answer estimated based on maximum interestingness in one of its edges in the MST. A high interestingness means that an answer has a comparatively long but essential edge.
- Randomly selected answers from each cluster, which served as another baseline for comparison.

Note that only one answer can be selected from clusters with just a single member regardless of the definition. The settings of the experiment are summarised below. The experimental results are summarised in Table 4.

- The selection of impactful answer and injection of ground-truth grades at half-point of the evolutionary clustering generation.
- The algebraic scheme that imposes 0.1 penalty as the third objective function was used.
- At the final grading, at least two ground-truth grades were used per cluster and a cluster was divided if a violation was found.

For each definition, the cases of fewer preferred cluster of 30 and more preferred clusters of 50 were evaluated.

The results show that a purposeful selection of impactful answers gave better performance compared to a random selection. The combination of centroids and borderline answers looked to offer more than just the centroids. Clearly, the additional grading actions would help anyway but the grading of second borderline answers did not always lead to better performance. The highest connectedness rather than the highest interestingness indicated a better representation of the borderline or the marginal answers for clusters.

Some answers are found more impactful than others. The selection methods can be integrated with interactive grading for building an ordered list of raw answers starting from the most impactful ones. The list can guide the interactive grading procedure and allow human graders to make informed decision on whether more grading actions are desired.

Table 4 Comparison of grading accuracies given different methods of selecting the impactful answers for additional human grading and the injection of ground-truth grades

		Q3			*Q8*		
	Preferred cluster #	*# Actual clusters*	*# Grading actions*	*Accuracy*	*# Actual clusters*	*# Grading actions*	*Accuracy*
Evolutionary clustering baseline	30	36	46	0.8082	31	47	0.9778
Centroid only	30	32	35	0.8714	30	40	0.9871
	50	50	56	0.8709	36	43	0.9785
Centroid + Borderline (highest connectedness)	30	43	64	**0.8723**	33	60	0.9842
	50	67	100	**0.8943**	40	65	0.9842
Centroid + 2 Borderline (highest connectedness)	30	58	88	0.8651	36	70	**0.9909**
	50	97	134	0.8924	48	77	**0.9890**
Centroid + Borderline (highest interestingness)	30	53	71	0.8250	32	56	0.9795
	50	53	86	0.8800	36	63	0.9842
Random selection	30	36	51	0.7781	32	53	0.9833
	50	58	84	0.8183	40	51	0.9871

5.3 Overall grading performance

To demonstrate the robustness of the framework for answer sets with different distributions in the semantic space, this experiment compares the performance improved in each of the questions due to the framework. The experimental settings used were the best combination of impactful answer selection methods and ground-truth grades injection schemes. Table 5 compares the performance of the framework on each of the answer datasets from the 8 questions. The set-theoretic scheme based on centroid + borderline (highest connectedness) was the setting of the experiment.

Table 5 Grading performance of the novel divide-and-grade short answer grading framework on datasets of different distribution characteristics

		Accuracy							
	Preferred Clusters #	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*
Baseline	30	0.9102	0.8042	0.8082	0.8959	0.9866	0.9097	0.9309	0.9776
Baseline	50	0.9102	0.9632	0.8843	0.9470	0.9799	**0.9589**	0.9026	0.9799
NDSAGF	30	**0.9967**	**0.9938**	**0.8824**	**0.9642**	**0.9895**	**0.9847**	**0.9661**	**0.9914**
NDSAGF	50	**0.9967**	**0.9933**	**0.9053**	**0.9542**	**0.9871**	0.9546	**0.9776**	**0.9933**

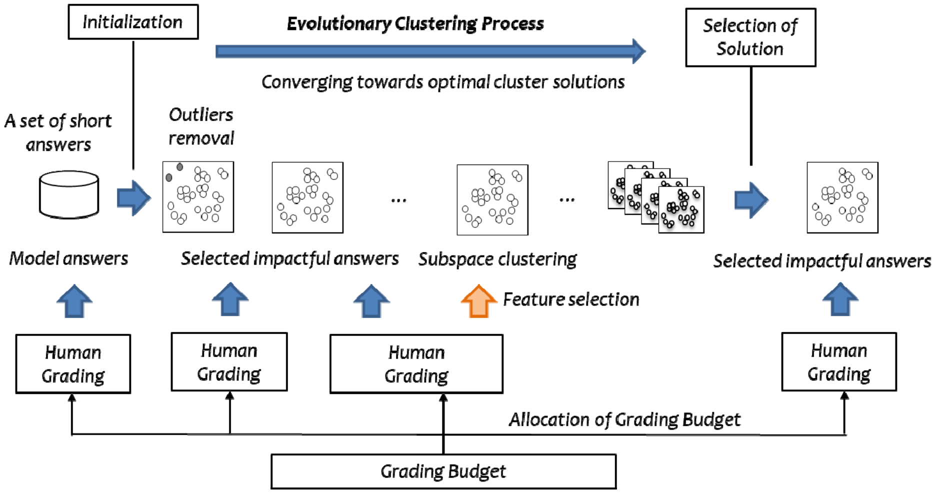
The results show a robust performance of the framework in all the answer sets for all the selected questions.

6 Conclusion

The divide-and-grade approach of ASAG is an attractive alternative to the supervised learning approach due to the potential saving of human grading effort. The performance of cluster analysis is key to the success of the approach. The clustering of short answers is prone to misclustering due to imperfect semantic representation, noisy input and marginal cases. The investigation described in this paper attempted to improve the divide-and-grade approach through the injection of ground-truth grades into the cluster analysis stage. A framework for divide-and-conquer based short answer grading was described and evaluated. The experimental results indicated that suitable ways of ground-truth grade injection could reduce grading errors due to misclustering.

The principle of the divide-and-grade approach is based on effective utilisation of human grading actions. This paper has made two major contributions. The first is to demonstrate formulations for factoring in ground-truth grades effectively in short answer clustering. This is in effect turning a multi-objective evolutionary clustering algorithm into a semi-supervised variant. The second is to demonstrate the importance of identifying the more impactful answers and the relative effectiveness of the ground-truth grades of impactful answers over other answers. Human grading actions should be applied to where they matter most.

Figure 8 Human grading actions and the resulting ground-truth grades may further improve the divide-and-grade approach through several schemes for future work that include better utilising the model answers, multi-stage injection of ground-truth grades, feature selection according to known differentiation between correct and wrong answers, and systematic allocation of grading budget



After a human grading action has turned out a ground-truth grade, the next consideration is to make better utilisation of it. Figure 8 outlines several schemes that should improve the divide-and-grade approach. The following paragraphs describe these schemes that should be investigated in future work.

Ground-truth grades were obtained and injected into clustering algorithm at the half-point of evolution in this investigation. As discussed in Section 3, there are pros and cons if the action is taken at other time-points. More experiments are required for better understanding of the trade-off.

A correct-wrong pair of ground-truth grades is needed for identifying irregular clusters. A correct answer alone, for example a model answer, may still prove to be useful. If all model answers are known, then answers that are sufficient far away from the model answers are likely to be wrong. The outlier factor (Liu et al., 2018) may serve as an additional condition to identify global anomalous answers written by clueless students. These anomalous answers may be removed from cluster analysis to save computation and improve cluster quality. The reliability of this method needs to come from an empirical evaluation.

The framework facilitates human engagement at various stages of the divide-and-grade process. Setting a budget number of grading actions should make the framework friendlier to human graders. A strategy for allocation of the grading budget between the possible injection points is becoming essential for further optimising the utilisation of human engagement.

The set of answers with ground-truth grades, especially the marginal answers, may be able to infer the significant semantic features for grade differentiation. As in subspace clustering (Deng et al., 2016), the clustering algorithm may be made to exploit the relevance of semantic features in the adaptive identification of clusters.

Automated short answer grading uses a grading model to differentiate correct answers from wrong answers. The supervised learning approach needs a large sample of human pre-graded answers to build a good model. The divide-and-conquer approach, of which this work is based on, does not need pre-grading and thus substantially reduce human effort. The grading model is learned from the distribution of raw answers. The proposed algorithm recognises that the grading model can improve with a small sample of pre-graded answers, which indicates certain pairs of answers should be in separate clusters. The findings showed that a relatively small additional human effort can increase the accuracy of the grading model.

Acknowledgement

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E10/19).

References

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

Bogdanović, Z., Barać, D., Jovanić, B., Popović, S. and Radenković, B. (2014) 'Evaluation of mobile assessment in a learning management system', *British Journal of Educational Technology*, Vol. 45, No. 2, pp.231–244.

Boleda, G. (2020) 'Distributional semantics and linguistic theory', *Annual Review of Linguistics*.

Brooks, M., Basu, S., Jacobs, C. and Vanderwende, L. (2014) 'Divide and correct: using clusters to grade short answers at scale', *Proceedings of the first ACM conference on Learning@ scale conference*, pp.89–98.

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.

Cai, Z., Gong, Y., Qiu, Q., Hu, X. and Graesser, A. (2016) 'Making autotutor agents smarter: Autotutor answer clustering and iterative script authoring', *International Conference on Intelligent Virtual Agents*, Springer, Cham, pp.438–441.

Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. and Sung, Y.H. (2018) 'Universal sentence encoder', *arXiv preprint arXiv:1803.11175*.

Deng, Z., Choi, K.S., Jiang, Y., Wang, J. and Wang, S. (2016) 'A survey on soft subspace clustering', *Information sciences*, Vol. 348, pp.84–106.

Garza-Fabre, M., Handl, J. and Knowles, J. (2017) 'An improved and more scalable evolutionary approach to multiobjective clustering', *IEEE Transactions on Evolutionary Computation*, Vol. 22, No. 4, pp.515–535.

Gomaa, W.H. and Fahmy, A.A. (2012) 'Short answer grading using string similarity and corpus-based similarity', *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 3, No. 11.

Ha, L.S. and Zhang, C. (2019) 'Are computers better than smartphones for web survey responses?' *Online Information Review*.

Handl, J. and Knowles, J. (2007) 'An evolutionary approach to multiobjective clustering', *IEEE transactions on Evolutionary Computation*, Vol. 11, No. 1, pp.56–76.

Heilman, M. and Madnani, N. (2013) 'ETS: Domain adaptation and stacking for short answer scoring', *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp.275–279).

Horbach, A. and Pinkal, M. (2018) 'Semi-supervised clustering for short answer scoring', *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Liu, H., Li, J., Wu, Y. and Fu, Y. (2018) 'Clustering with outlier removal', *arXiv preprint arXiv:1801.01899*.

Meurers, D., Ziai, R., Ott, N. and Bailey, S.M. (2011) 'Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions', *International Journal of Continuing Engineering Education and Life Long Learning*, Vol. 21, No. 4, pp.355–369.

Menini, S., Tonelli, S., De Gasperis, G., Vittorini, P. (2019) 'Automated short answer grading: a simple solution for a difficult task', *Sesta Conferenza Italiana di Linguistica Computazionale (CLiC-it 2019)*.

Mohler, M. and Mihalcea, R. (2009) 'Text-to-text semantic similarity for automatic short answer grading', *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp.567–575).

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* (pp.752–762).

Roy, S., Bhatt, H.S. and Narahari, Y. (2016) 'Transfer learning for automatic short answer grading', *Proceedings of the Twenty-second European Conference on Artificial Intelligence* (pp.1622–1623).

Sahu, A. and Bhowmick, P.K. (2019) 'Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance', *IEEE Transactions on Learning Technologies*, Vol. 13, No. 1, pp.77–90.

Sakaguchi, K., Heilman, M. and Madnani, N. (2015) 'Effective feature integration for automated short answer scoring', *Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies* (pp.1049–1054).

Shah, D. (2019) *By The Numbers: MOOCs in 2019*. https://www.classcentral.com /report/mooc-stats-2019 (accessed on 31 July 2020).

Tibshirani, R., Walther, G. and Hastie, T. (2001) 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 2, pp.411–423.

Vii, S., Tayal, D. and Jain, A. (2019) 'A fuzzy WordNet graph based approach to find key terms for students short answer evaluation', *2019 4th international conference on internet of things: Smart innovation and usages (IoT-SIU)*(pp.1–6). IEEE.

Wang, Y., Fang, H., Jin, Q. and Ma, J. (2019) 'SSPA: an effective semi-supervised peer assessment method for large scale MOOCs', *Interactive Learning Environments*, pp.1–19.

Yu, H., Miao, C., Leung, C. and White, T.J. (2017) 'Towards AI-powered personalization in MOOC learning', *NPJ Science of Learning*, Vol. 2, No. 1, pp.1–5.

Zesch, T., Heilman, M. and Cahill, A. (2015) 'Reducing annotation efforts in supervised short answer scoring', *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp.124–132).

Zhang, Y., Shah, R. and Chi, M. (2016) 'Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading', *International Educational Data Mining Society*.