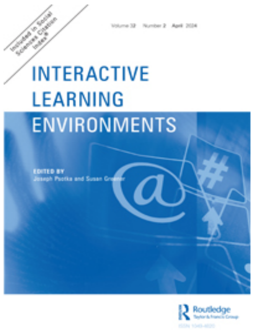

Interactive Learning Environments

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/nile20

Automated short answer grading with computer-assisted grading example acquisition based on active learning

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

To cite this article: Andrew Kwok-Fai Lui, Sin-Chun Ng & Stella Wing-Nga Cheung (05 Dec 2022): Automated short answer grading with computer-assisted grading example acquisition based on active learning, Interactive Learning Environments, DOI: 10.1080/10494820.2022.2137530

To link to this article: https://doi.org/10.1080/10494820.2022.2137530

Published online: 05 Dec 2022.

INTERACTIVE LEARNING ENVIRONMENTS
https://doi.org/10.1080/10494820.2022.2137530

Automated short answer grading with computer-assisted grading example acquisition based on active learning

Andrew Kwok-Fai Lui [a], Sin-Chun Ng [b] and Stella Wing-Nga Cheung [a]

[a]Department of Electronic Engineering and Computer Science, Hong Kong Metropolitan University, Hong Kong, People's Republic of China; [b]School of Computing and Information Science, Anglia Ruskin University, Cambridge, UK

ABSTRACT
The technology of automated short answer grading (ASAG) can efficiently process answers according to human-prepared grading examples. Computer-assisted acquisition of grading examples uses a computer algorithm to sample real student responses for potentially good examples. The process is critical for optimizing the grading accuracy of machine learning models given a budget of human effort and the appeal of ASAG to online learning providers. This paper presents a novel method called short answer grading with active learning (SAGAL) that features a unified formulation comprising the heuristics for identifying potentially optimal examples of representative answers, borderline answers, and anomalous answers. The method is based on active learning, which iteratively samples good examples and queries for annotation to increase the sampling accuracy. SAGAL has been evaluated with three different public datasets of distinctive characteristics. The results show that the resulting models generally outperform the baseline semi-supervised learning methods on the same number of grading examples.

ARTICLE HISTORY
Received 14 May 2022
Accepted 11 October 2022

KEYWORDS
Short answer grading; active learning; classification; autograding; cost-effectiveness

Introduction

The assessment of student learning can be achieved by asking many types of questions, including short answer questions. Short answer questions expect constructed natural language responses that recall specific facts (Burrows et al., 2015; Livingston, 2009) or express subjective opinions and justifications (Zhang et al., 2022). The global embrace of online learning has stimulated strong demand for fast, scalable, and around-the-clock assessment services. As a result, automated grading has become a critical enabling technology. Many automated short answer grading (ASAG) methods can be found in the literature, and those based on machine learning use a computation model to grade answers according to a human-prepared set of grading examples. A grading example is a human-annotated answer, representing the grading decision on the answer. The example set demonstrates how each short answer question should be graded. The size of an example set is a reflection of the open-ended-ness of the question. In practice, a few examples are sufficient to describe the expected answers in closed-ended questions, while substantially more examples are needed in assessing subjective opinions (Zhang et al., 2022).

The construction of machine learning models is a procedure of learning from annotated examples so that the models are able to generalize and to give proper output for unseen input data. The capacity in generalization is useful for the assessment of freely constructed short answers. A

CONTACT Andrew Kwok-Fai Lui alui@hkmu.edu.hk

major challenge in many machine learning applications is the acquisition of annotated examples. The annotated training datasets assumed in earlier ASAG works (e.g. Mohler et al., 2011) in fact must be prepared from scratch in real deployment of ASAG tasks. Therefore costly human participation is required in the preparation work, in addition the number of annotated examples is subject to budget limitation. Machine learning researchers have proven that properly crafted examples can make a significant difference in model performance, and it can be achieved with optimization of example acquisition (Lewis & Gale, 1994).

Computer-assisted acquisition of grading examples uses a computer algorithm to sample received student responses for potentially good examples. Sourcing examples from the received responses can avoid the concept drift problem, which refers to the coverage mismatch between hand-crafted examples and real responses (Marton & Säljö, 1976). In addition, algorithmic sampling is superior to random sampling because, first, filters to prevent redundancy can be added, and second, heuristics estimating the goodness of examples can be applied. The human grader is allowed to focus on the annotation work and can be spared from shifting through the real responses for good examples.

In the machine learning literature, two heuristics have emerged as useful goodness measures, namely representativeness and informativeness (Gu et al., 2020; Huang et al., 2014). A representative example can represent a significant proportion of the dataset due to their similarity. An informative example can reduce the uncertainty in the model, such as the decision boundaries and the anomalies. Therefore, a response with high representativeness or high informativeness is considered to be potentially good example. Assessment experts have also made similar recommendations on how to select good examples for short answer marking schemes. Reference answers and borderline answers are generally considered as essential for inclusion in marking schemes (Ahmed & Pollitt, 2011; Butcher & Jordan, 2010; Livingston, 2009). The reference answers should match most of the correct responses and therefore they are highly representative. The borderline answers are highly informative because they can delineate marginal responses. Figure 1 illustrates a sample marking scheme.

In the ASAG literature, however, the understanding on the computer-assisted acquisition of grading examples is limited and disjointed. A stream of works proposed using cluster analysis to find representative examples (Brooks et al., 2014; Marvaniya et al., 2018; Zesch et al., 2015). The findings demonstrated an improvement on the annotated example set and reduction in human annotation effort. However, these works failed to address the marginal responses and the uncertainty of the cluster edges. Horbach and Palmer (2016) evaluated several uncertainty-based sampling methods and were the first to propose using active learning to enhance the sampling. Active learning is a machine learning approach that iteratively acquires annotated data and updates the model (Saar-Tsechansky & Provost, 2004; Cohn et al., 1994), such that more marginal responses can be

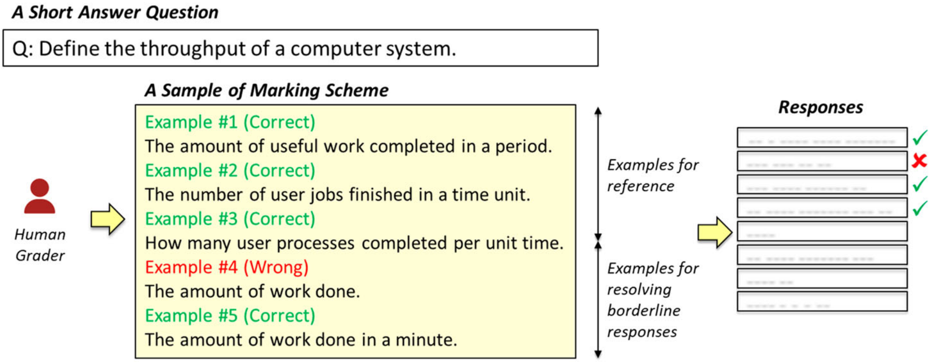
A Short Answer Question
Q: Define the throughput of a computer system.
A Sample of Marking Scheme
Example #1 (Correct)
The amount of useful work completed in a period.
Example #2 (Correct)
The number of user jobs finished in a time unit.
Example #3 (Correct)
How many user processes completed per unit time.
Example #4 (Wrong)
The amount of work done.
Example #5 (Correct)
The amount of work done in a minute.
Human Grader
Examples for reference
Examples for resolving borderline responses
Responses
✓
✗
✓
✓

Figure 1. A short answer question and a sample marking scheme for the question.

identified. A unified formulation that incorporates the reference answers, the borderline answers, and other potentially optimized examples are lacking. In addition, the current works have not drawn from the recent development in the integration of active learning and clustering (Kumar & Gupta, 2020; Shi et al., 2020), which can provide a suitable basis for better methods of preparing optimized sets of grading examples.

This paper investigates computer-assisted acquisition of grading examples for training short answer grading models. Short answer grading with active learning (SAGAL) is proposed to address the desire for optimized sets of grading examples. The key contributions of this paper are summarized below.

- The first comprehensive investigation on ASAG from the perspective of grading example preparation.
- The integration of the following three research areas in machine learning, including density peak clustering (Rodriguez & Laio, 2014), the ascription tree (Shi et al., 2020), and one-class classification (Khan & Madden, 2014), into the study of ASAG methods.
- The unification of the identification of reference answers and borderline answers, as well as redundancy avoidance, into one formulation.
- The comprehensive evaluation of the proposed SAGAL method using three gold-standard ASAG datasets.

Literature review

Essentials of short answer grading models

Short answer grading is a task of matching student responses with enumerable facts and mentions (Livingston, 2009). The expected responses can be the results of external knowledge recall (Burrows et al., 2015). Expressing subject opinions is also enabled if there is a finite set of opinions (Zhang et al., 2022). The task also involves understanding the syntactic and semantic variations of a natural language response and in particular, determination of the decision boundaries between correct and wrong. The role of the grading examples is to enumerate the expectations and specify the decision boundaries (Ahmed & Pollitt, 2011; Butcher & Jordan, 2010).

Short answer grading model is a classifier that divides student responses into groups of different grades. The classifier operates in a semantic feature space in which received student responses are mapped and classified according to the semantic features. The role of the semantic feature space is to extract the semantics and to support a computable vector representation. Many techniques based on hand-engineered language features and corpus-based features have been evaluated (Burrows et al., 2015; Galhardi & Brancher, 2018). Recent studies indicated that deep feature learning from a gigantic corpus can offer a compact and rich semantic feature space that brings superior performance in grading tasks (Gaddipati et al., 2020; Sawatzki et al., 2022).

A proper generalization of grading examples depends on the form of and the placement of decision boundaries. In multi-class classification models, the decision boundaries are in the form of a smooth hypersurface separating two or more classes. A well-learned hypersurface decision boundary generally requires a large number of annotated examples. The generalization is prone to error if the decision boundary lacks support of nearby annotated examples. In one-class classification, the decision boundaries are in the form of a hypersphere of a certain radius that separate data of the main class and the unknown class (Khan & Madden, 2014). The radius of the hypersphere encodes the maximum acceptable deviation from the annotated example at the centre and provides a semantic support to the decision boundary. A small number of annotated examples is enough to learn reliable decision boundaries of this form. Note that the radius can be first set to a default or estimated value and then adjusted to the newly added annotated examples. Both forms of decision boundaries are relevant to short answer grading models due to the variability of the example set.

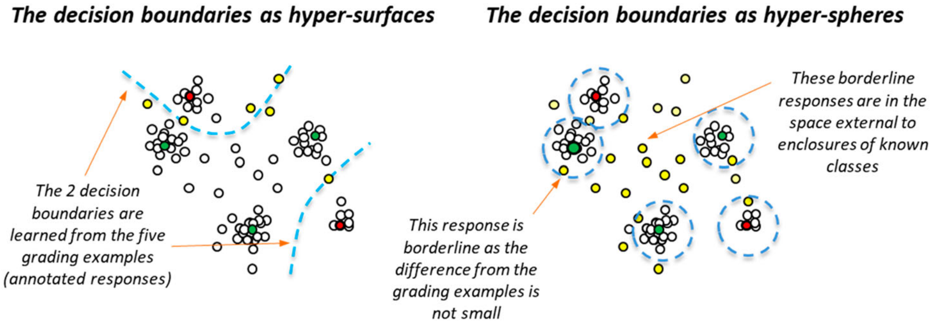
The decision boundaries as hyper-surfaces
The decision boundaries as hyper-spheres
The 2 decision boundaries are learned from the five grading examples (annotated responses)
This response is borderline as the difference from the grading examples is not small
These borderline responses are in the space external to enclosures of known classes

Figure 2. The responses, as hollow and solid circles, embedded in a feature space. On the left the decision boundaries are in the form of hypersurfaces and on the right the decision boundaries are in the form of hyper-spheres.

Figure 2 illustrates the significance of the forms of decision boundaries in a short answer grading model.

Computer-assisted acquisition of grading examples

In computer-assisted acquisition of training examples, the options are representative and uncertainty sampling. The former identifies examples from the mode of each data cluster, and the latter identifies examples based on the class uncertainty in the model (Lewis & Catlett, 1994; Lewis & Gale, 1994). The two sampling formulations can be combined to further increase the accuracy in uncertainty sampling (Xu et al., 2003).

In considering the machine learning approaches for building ASAG models, the options are supervised learning, semi-supervised learning, and active learning. Figure 3 illustrates their similarities and differences. The supervised learning approach, including the deep learning approach, is capable of exploiting large annotated data, if available, to train very accurate models. The semi-supervised learning approach attempts to minimize human annotation effort with representation sampling. The stream of studies initiated by Basu et al. (2013) have demonstrated that applying cluster analysis on student responses can effectively identify an optimized set of grading examples and support the same accuracy with much fewer examples (Brooks et al., 2014; Horbach & Palmer, 2016; Marvaniya et al., 2018; Zesch et al., 2015).

The active learning approach is an enhanced version of the semi-supervised approach (Saar-Tsechansky & Provost, 2004; Cohn et al., 1994). The key feature is the cycle of updating the grading model, sampling new grading examples, and annotation. By improving the accuracy of the grading model, the example sampling can exploit maximum knowledge from the human annotator (Du et al., 2017; Kumar & Gupta, 2020).

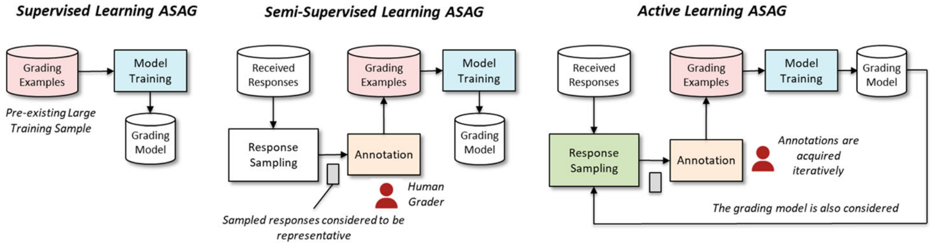

Figure 3. Comparison between the three machine learning based ASAG configurations with regard to how annotated data are acquired and exploited in model training and data evaluation.

Potential of active learning in the acquisition of grading examples

This section discusses the potential of active learning in the acquisition of grading examples. Similar to the prior work by Horbach and Palmer (2016) and Kishaan et al. (2020), the cycle in active learning comprises the stage of model training, response sampling, and annotation.

Batch size and batch sampling

The batch size is the number of examples sampled and annotated in each cycle. A large batch size reduces the number of iterations and model training time, but the sampled data in each batch may be less optimal. According to the batch size, the top-ranked student responses are selected according to the favourable metrics such as representativeness, grading uncertainty, and diversity. The final rank order may be determined by a unified formulation of the metrics or a batch selection strategy (Englhardt et al., 2020).

Diversity and anomalies

The diversity of a sample of grading examples is a measure of the semantic dissimilarity between the examples. Redundancy is harmful to optimized example acquisition. Ensuring the examples are kept apart in the semantic space is an effective heuristic for the avoidance of redundancy (Cardoso et al., 2017).

Anomalies represent the data that diverges significantly from the majority of the data. In short answer grading, anomalous responses refer to those completely irrelevant to the expectations, such as “I don't know” or “not sure.” Anomalous responses can be handled by eliminating them before the start of active learning, or considering them as wrong answers. The former method can help speed up active learning but the latter method is useful for an initial estimation of the decision boundaries in the grading model. In the first few batches of grading examples, an absence of wrong examples is possible, and in this case, the anomalous responses can offer some wrong examples.

Sampling responses according to representativeness

A response that represents many semantically similar responses and their grade is a representative response or a response with high representativeness. Its annotation enables the generalization of its grade to similar responses (Basu et al., 2013). Cluster analysis is the most used method in active learning for finding representative responses and ensuring a level of diversity as a desirable side-effect. Good clusters are characterized by their compactness and separability.

Sampling borderline responses according to grade uncertainty

A borderline response is indicated by its high level of grade uncertainty (Lewis & Gale, 1994). The value of its annotation is in refining or revising the decision boundaries. In the machine learning literature, the corresponding metric is known as informativeness, which refers to the amount of information gained with the annotation.

Borderline responses are those located close to decision boundaries. Their identification is more challenging in grading models without explicit decision boundaries. One method to compute the informativeness is based on the entropy of the grade of nearby responses in the semantic space. A locality with mixed grades can indicate that borderline responses can be found there.

Semantic diversity with a model ensemble

Another effective perspective of the borderline concept is based on an ensemble of feature spaces. Each of the feature spaces comes from random subsampling of the original space and represents a certain semantic perspective on the response set. A grade is uncertain if the models in the ensemble put the response to different sides of the decision boundaries (Barr et al., 2014).

Active clustering with density peak and efficient model updating

Shi et al. (2020) proposed an active clustering ensemble algorithm that provides the foundation for the development of the SAGAL method. It considers both representativeness and informativeness in a unified sampling method. In addition, the informativeness has a local component measuring the uncertainty with respect to the local neighbourhood and a global component for that in the ensemble feature space. The algorithm includes a fast update strategy of annotation propagation in model updating.

Density peak clustering

The active clustering ensemble algorithm uses density peak clustering to identify the potential representative data (Nguyen & Smeulders, 2004). Density peak clustering computes the local density of every data and then analyses the density peaks defined as the highest density locally. The local density can be estimated based on the number of data in the neighbourhood or the inverse of the mean distance to the nearest neighbours.

Tree structure for fast label propagation

The active clustering ensemble algorithm uses a tree structure called the ascription tree to support the fast model update strategy. A similar tree structure is also proposed by Chen et al. (2020). These tree structures, which are often pre-computed, connect all data from the highest density to the lowest density. The tree structure facilitates fast updates of the class labels through propagation along the edges.

Short answer grading with active learning

This section describes the definition and the training of the SAGAL model.

The grading model

The grading model is a function $F(S, Y):S \rightarrow Y$ that maps a semantic feature vector of a response to a grade label, where Y is the set of possible labels, $S = \{s_i\}_{i=1}^{n}$ is a set of n number of student responses to a short answer question, and every response s_i is a feature vector defined as $s_i = (x_1, \cdots, x_m)$ where x_j is one of the m semantic features. The responses are divided into the grading example set S^G and the unannotated response set S^U, where $S = S^G \cup S^U$, $S^G \cap S^U \equiv \emptyset$. The grading examples have a grade from annotation and the grades of unannotated responses are assigned by the model.

The model assigns grades to the responses in S^U based on the density gradient. Using a tree structure called the response tree, the annotated grades at density peaks propagate to the unannotated responses at lower-density locations. The density is estimated from the number of responses in a local neighbourhood of radius ε and it is computed with the following equation based on a cut-off Gaussian kernel, where the response set S_i^N is within the ε-neighbourhood, and d_{ij} is the semantic distance between two responses s_i and s_j. The cut-off Gaussian kernel may not be the generally most effective (Hou & Pelillo, 2016) but it suits the shape of ε-neighbourhoods particularly.

$$\rho(s_i) = \sum_{j \in S_i^N, i \neq j} exp\left(-\frac{d_{ij}^2}{\varepsilon^2}\right) \tag{1}$$

The response tree comprises parent–child connections that span the response set S. The parent of every response is the nearest higher peak, with the exception of the highest peak, which is the root of the tree. After the response tree is pre-constructed, it is utilized in the grade propagation process, which simply involves every unannotated response inheriting the grade from the parent. Figure 4 illustrates the structure of the response tree.

Anomalous responses

The anomalous responses are identified during the grade propagation in the response tree. A break-off distance α is defined such that a response is considered anomalous if the distance from the nearest ancestor grading example is greater than α. In SAGAL, the anomalous responses are assumed to be wrong answers.

Majority voting based on model ensemble

The grading model uses an ensemble of grading sub-models, each of which is based on a random sub-sampling of the semantic features for improved model performance (Shi et al., 2020). The final assigned grade is based on the majority vote.

Model initialization and training

The details of the initialization phase and the training phase are described below and illustrated step-by-step in Figure 5.

The initialization phase

The initialization phase involves randomly generating the feature subspaces for the model ensemble, embedding the responses to the subspaces, and for every subspace pre-computing several key parameters, including the local density of all responses, the MASD ε and the grade inheritance break-off distance α, and the response tree. The following shows the details of the computation.

- Randomly sub-sample a compact sentence embedding of *nf* number of features into *nd* number of *m*-dimensional subspaces, and then map the response set to every subspace.
- For every subspace, compute the pairwise semantic distances between all responses and rank them from shortest to longest. The MASD ε is learned from the rank order list, which is used to indicate the representative compactness of each response set (Shi et al., 2020). Figure 6 shows a graphical illustration of the rank order list. SAGAL uses the 50 percentile as the estimation of ε.
- Determine the break-off distance α, either as a pre-determined multiplier of ε, or the point of maximum curvature between the 70 and 100 percentiles in the rank order list.
- For every subspace, compute the local density $\rho(s_i)$ for every response using Equation 1. Construct the response tree based on the computed local densities.

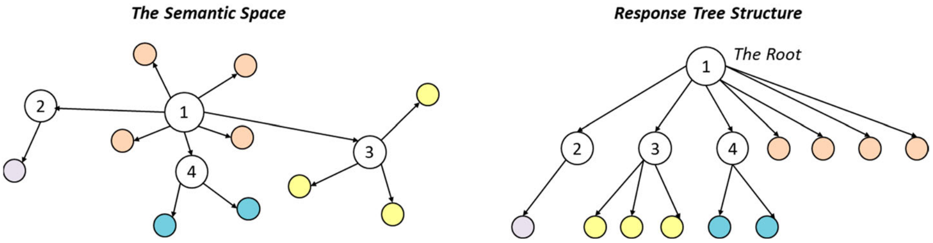

Figure 4. A semantic space with some embedded responses is shown on the left. The size of the circles represent the local density. The highest density peak is response #1, which is the root of the response tree shown on the right. Every other response has its nearest higher density response as the parent.

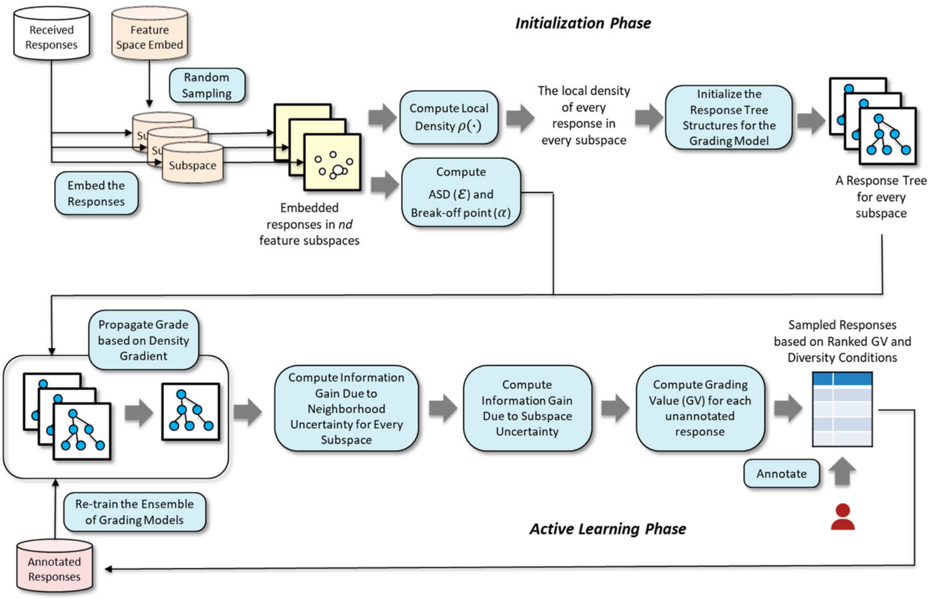
Initialization Phase
Received Responses
Feature Space Embed
Random Sampling
Su
Su
Subspace
Embed the Responses
Embedded responses in nd feature subspaces
Compute Local Density ρ(·)
Compute ASD (ℰ) and Break-off point (α)
The local density of every response in every subspace
Initialize the Response Tree Structures for the Grading Model
A Response Tree for every subspace
Propagate Grade based on Density Gradient
Compute Information Gain Due to Neighborhood Uncertainty for Every Subspace
Compute Information Gain Due to Subspace Uncertainty
Compute Grading Value (GV) for each unannotated response
Sampled Responses based on Ranked GV and Diversity Conditions
Annotate
Re-train the Ensemble of Grading Models
Annotated Responses
Active Learning Phase

Figure 5. A step-by-step illustration of the SAGAL algorithm.

The training phase

The training phase embarks on cycles of sampling examples according to the grading values, querying for annotation, and updating the grading model.

The example sampling is based on the grading value (GV_i) of the response s_i, which is defined in Equation (2) below, where sp_z is the z subspace of the ensemble SSP, H_N and H_S are the uncertainty measurements for the neighbourhood and for the ensemble respectively.

$$GV_i = \frac{1}{nd} \sum_{sp_z \in SSP} ((H_N(s_i,\ z)) \times \rho(s_i, z)) + H_S(s_i) \tag{2}$$

$$H_S(s_i) = -\sum_{k=1}^{|Y|} p_i^k log_2 p_i^k \tag{3}$$

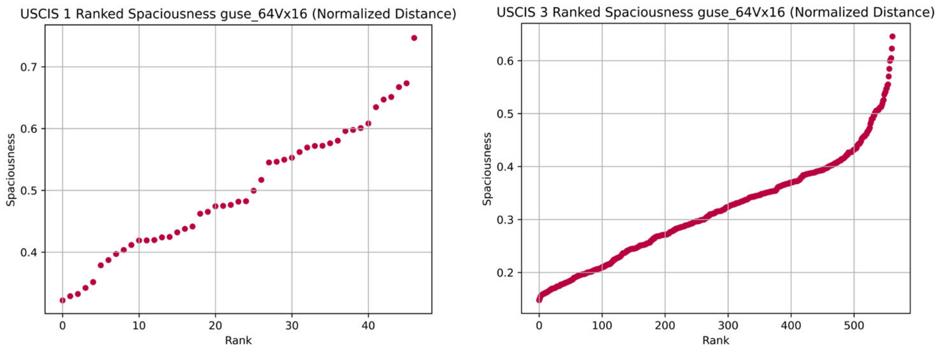

Figure 6. Rank ordered list the pairwise distances of all responses in the datasets of USCIS Q1 and Q3 (from left to right) from Basu et al. (2013). A turning-up point is noted near the high rank range, which indicates the anomalies in the response sets.

$$H_N(s_i) = -\sum_{k=1}^{|Y|} p_i^k log_2 p_i^k + \frac{1}{|Y|} \sum_{j=1}^{|Y|} P\left(G_N(s_i) \mid \ddot{y} \right) \tag{4}$$

$$G_N(s_i) = -\sum_{k=1}^{|Y|} \underset{i}{\check{p}}\, k \; log_2 \; \underset{i}{\check{p}}\, k \tag{5}$$

$$\underset{i}{\check{p}}\, k \; = p_i^k \pm (p_i^k \times t^k) \quad \text{if } \overset{=}{\ddot{y}}\, k \text{ or otherwise} \tag{6}$$

The ensemble uncertainty H_S is based on the votes from every subspace grading model, which is illustrated in Figure 7 and defined in Equation (3), where p_i^k is the probability of the response s_i given the grade k among the subspaces.

The subspace neighbourhood uncertainty H_N is based on the diversity of the assigned grades within the ε-neighbourhood. Normally an annotated response would remove the uncertainty associated with self and the other nearby responses within the neighbourhood. However, the presence of other annotated responses in the same neighbourhood complicates the issue. The post-annotated estimation of the grade diversity of the annotated responses t^k of grade class k is used instead. This estimation is conditional on the actual grade y to be annotated and this is factored in Equations (4) to (6).

After the components of GV is computed, the values are min–max normalized to the range to [0, 1] as in Hou and Cui (2017). Note that the local density $\rho(s_i)$ is multiplied to the uncertainty due to the neighbourhood to emphasize the amplification effect of human annotation. Finally, the GV is computed and the unannotated responses are ranked accordingly. The diversity in the sampling of potential examples is a rule that the sampled responses should not be within the ε-neighbourhood of another sampled response (Englhardt et al., 2020). A batch of selected responses is then sent to a human for annotation.

Stopping conditions

The grading model should progressively improve as the training cycle continues. The representativeness metric increases the number of responses within the MASD of a grading example. The informativeness metric makes corrections to the decision boundaries. The marginal gain decreases as more of the semantic space has been explored and exploited. There is a point that the example acquisition effort is considered not worthwhile for the model improvement. This decision is subject to human perception.

In practice, the exhaustion of the annotation budget is considered a stopping condition. A budget of assessment outlay supports effective planning and management of online courses. The training

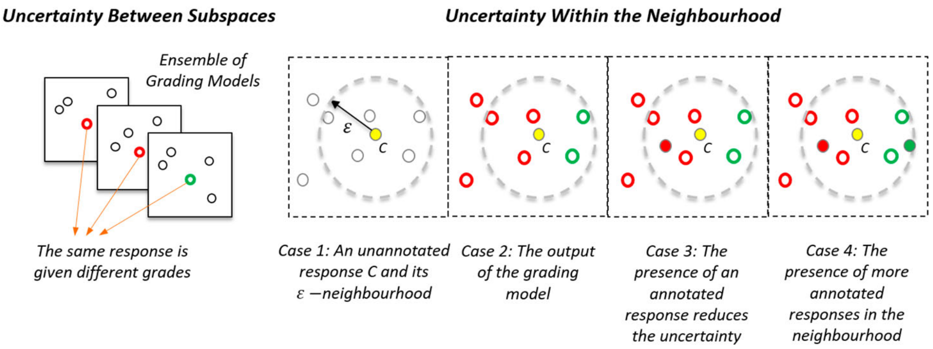
Uncertainty Between Subspaces
Uncertainty Within the Neighbourhood
Ensemble of Grading Models
ε
C
C
C
C
The same response is given different grades
Case 1: An unannotated response C and its ε −neighbourhood
Case 2: The output of the grading model
Case 3: The presence of an annotated response reduces the uncertainty
Case 4: The presence of more annotated responses in the neighbourhood

Figure 7. Comparison of the two types of borderline responses due to uncertainty between subspace grading models (left) and due to uncertainty within the ε-neighbourhood (right) respectively.

cycle can be restarted in the future if the performance validation is not satisfactory or there is a budget increase.

Results

This section reports the experiments for evaluating the performance of SAGAL and the key findings.

The datasets and performance metrics

Three publicly available datasets were selected due to their varieties in the subject domains, the sample size, and the lengths of responses. See the footnotes for the URLs to download them. Their key statistics are summarized in Tables 1–3.

- The USCIS dataset (Basu et al., 2013) contains 20 questions sampled from the United States Citizenship Examination. High specific answers are asked.
- The SciEntsBank dataset (Dzikovska et al., 2013) is large and covers many disciplines in science. Many questions are case studies in the science domain and are looking for facts, explanations, and reasoning.
- The Hewlett Foundation (HF) dataset (Peters & Jankiewicz, 2012) consists of more open-ended questions. The average length of the responses is the longest among the datasets.

The proposed SAGAL was compared to the following baselines.

- Random sampling. The queries sent to the human annotator were selected randomly from S^U. This is a common baseline used in similar work (Horbach & Palmer, 2016).
- Representative sampling. The response selection is based on the clustering algorithm *KMeans*, which maximizes the separation between clusters and minimizes the variance within clusters. This representativeness-based measurement has been used in other computer-assisted grading example acquisition works (Basu et al., 2013; Brooks et al., 2014; Zesch et al., 2015).
- Local representative sampling. The response selection is based on the clustering algorithm *Birch*, which finds smaller local clusters before merging them into larger clusters.

Table 1. Key statistics of the selected response sets from USCIS used in the evaluation.

USCIS Datasets	# Responses	# Correct	# Wrong	Average length (# words)
Q1	699	652	47	3
Q2	698	613	85	3
Q3	705	567	138	8
Q4	698	561	137	2
Q5	698	657	41	2
Q6	698	415	283	3
Q7	698	650	48	6
Q8	698	415	283	4

Table 2. Key statistics of the selected response sets from SciEntsBank-SEB2 used in the evaluation.

SCIEND-SEB2 Datasets	# Responses	# Correct	# Wrong	Average Length (# words)
DAMAGED_BULB_SWITCH_Q	100	64	28	7
DESCRIBE_GAP_LOCATE_PROCEDURE_Q	99	43	64	8
EV_12b	100	31	69	20
EV_25	101	7	94	16
HB_24b1	100	36	64	7
HB_35	100	20	80	14
HYBRID_BURNED_OUT_EXPLAIN_Q2	76	41	35	11
WA_52b	100	22	78	14

Table 3. Key statistics of the selected response sets from HF used in the evaluation.

HF Datasets	# Responses	# Correct	# Wrong	Average length (# words)
HF Q2	1278	784	494	59
HF Q5	1795	76	1719	26
HF Q6	1797	122	1675	24
HF Q7	1799	419	1380	41
HF Q8	1799	777	1022	54
HF Q10	1640	580	1060	42

The effectiveness in selecting the important examples is evaluated based on the accuracy of the grading model against an increasing number of budget of human annotations. If an algorithm is effective, the accuracy would be higher than others on the same grading budget.

In the following, several minor variants of SAGAL were evaluated against the baseline using the three groups of response sets. Due to the stochastic nature of subspace subsampling, the reported figures are the mean of 10 repetitions.

Performance on the USCIS datasets: small sets with freely constructed responses

Figure 8 compares the performance of SAGAL with the baseline using the USCIS response sets. In addition to the original SAGAL, a variant with fixed ε-neighbourhood size was also included in the evaluation and the decision boundary was fixed. The number of human annotations ranged from 20 to 150, which in these datasets, represented from around 3% to 21% of the size of the response sets.

SAGAL was found to generally perform better than the baselines. The random sampling method was found to be the worst as in previous studies (e.g. Horbach & Palmer, 2016). The questions in the USCIS dataset, particularly Q1, asked for very short and highly specific answers. Just a few annotations was enough to attain perfect accuracy due to its extremely specific nature, as there were only 49 lexicographical versions found in the response set. Only a few borderline responses were found, rendering active learning irrelevant. Mainly Q3 and Q7 offered some room for constructing responses different from the reference answers and therefore bred more borderline cases.

Performance on the SciEntsBank datasets: small sets with freely constructed responses

Figure 9 compares the performance of SAGAL with the baseline using the SciEntsBank response sets. As the upper limit was close to 100% annotation proportion, the interesting range was near to the low end. SAGAL outperformed the baselines in all the response sets if there were 20% or more annotations. The open-ended nature of these response sets invited constructed responses resulting in many borderline cases. Among the two SAGAL variants, the variant with fixed ε-neighbourhood size performed better. The SciEntsBank response sets were small and did not provide significant information for determining the neighbourhood size.

Performance on the HF datasets: large sets with open-ended responses

Figure 10 compares the performance of SAGAL with the baseline using the HF response sets. The same range of annotation budgets had been used in Figure 8 to Figure 10, but they represented different proportions of the response set sizes. HF Q2 and HF Q6 enabled the SAGAL variants to perform better than the baselines, but the other response sets appeared to be challenging to SAGAL.

However, due to the larger response sets, the upper limit of 150 annotations actually represented only around 8% to 10% of the response sets in the HF dataset. If the upper bound, based on the proportion of annotations, was increased to a comparable level of around 25%, then SAGAL was found better than the baselines, as shown in Figure 11. The mean length of responses for the four selected response sets was significantly longer than the others, meaning that the variations in the responses would be the greatest and most challenging to overcome.

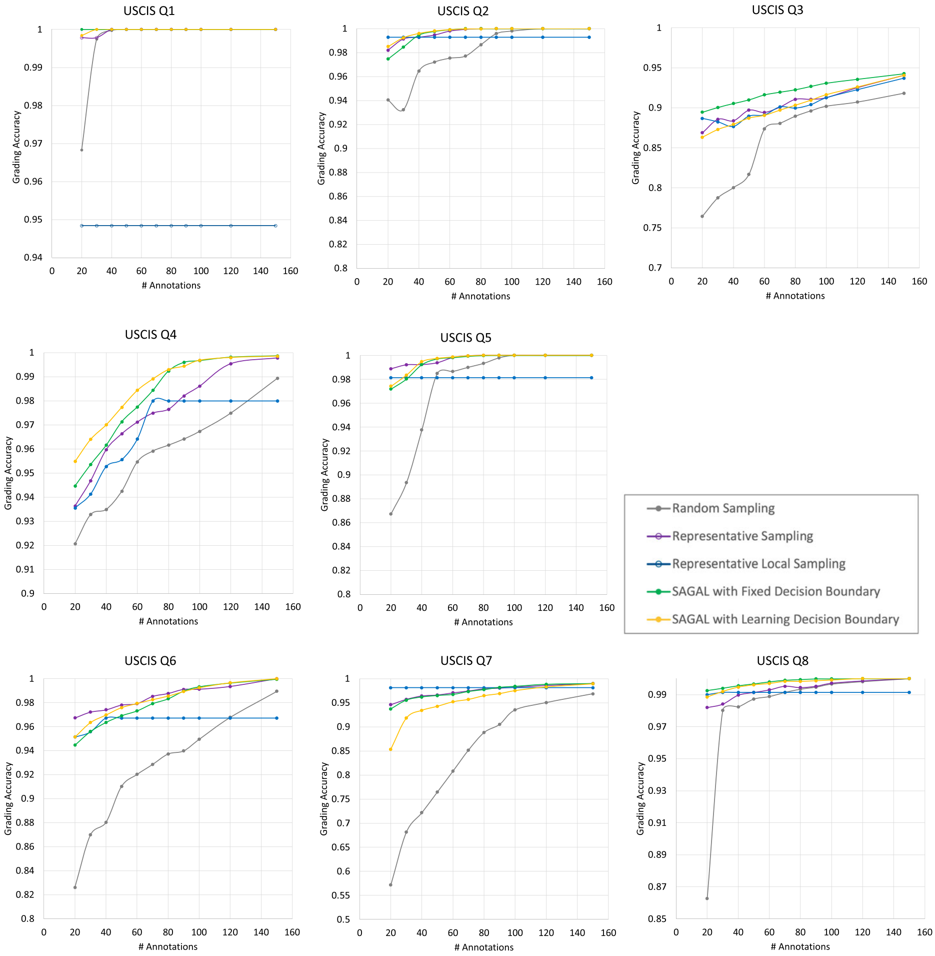

Figure 8. The performance of two variants of SAGAL and the baselines are compared based on the 8 response sets in the USCIS dataset. The SAGAL variants are shown in the lighter colours of green and yellow.

In all four cases, the final version of SAGAL with learned ε-neighbourhood size performed better. Comparatively, the response sets were significantly larger in the HF dataset for the unsupervised learning to perform for the initial guess of the neighbourhood size.

Anomalous response handling

Figure 11 highlights the performance of the anomalous response handling in SAGAL. A SAGAL variant without anomalous response detection was created and evaluated. The variant performed poorly, especially when the number of annotations was small. In one case, the variant outperformed one of the two SAGAL variants when the percentage of annotation was more than 14%. Generally, the anomalous response handling had a positive effect but the effect diminished as the number of annotations increases.

Discussion

The findings provided a strong support for the active learning approach of computer-assisted grading example acquisition. Given the same grading budget, SAGAL could generally perform

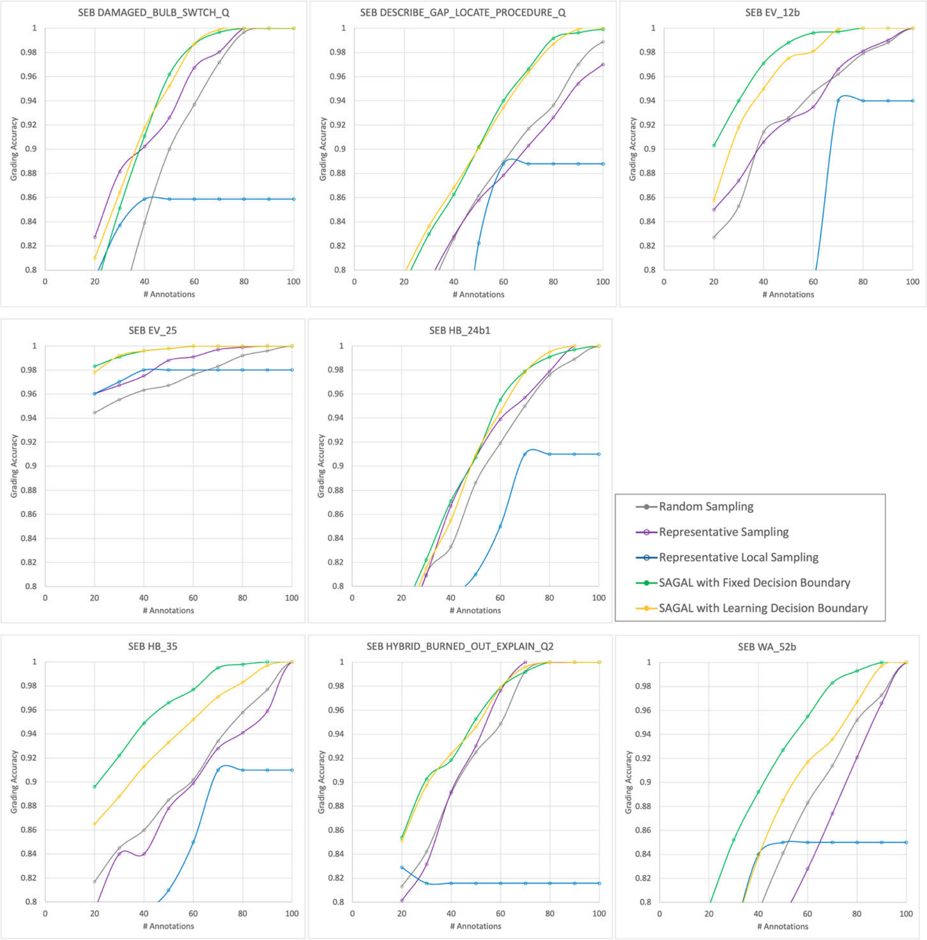

Figure 9. The performance of two variants of SAGAL and the baselines are compared based on the 8 response sets in the SciEntsBank dataset.

better than random sampling and representative sampling. The former result indicated the necessity of optimized sampling of grading examples, and the latter result suggested that the uncertainty sampling should be used together with the representative sampling (Horbach & Palmer, 2016).

The unified formulation of example sampling was considered a desirable feature of SAGAL. It was found to be flexible and extensible. The formulation considered the representativeness and informativeness metrics and also exploited anomalous responses and diversity in the sampling.

Different ASAG tasks posed different challenges to SAGAL as indicated in the varying performance compared to the baselines. For example, the difference between SAGAL and the representative sampling baselines was marginal for most of the USCIS datasets, suggesting that mainly the representativeness metric was useful. The questions of these datasets, as displayed in Basu et al. (2013), mostly expected very specific and simple facts. The responses were in fact on average 2–8 word long as shown in Table 1 and did not offer much room for minor variations that led to borderline responses. At around 40 annotations, which was equivalent to 6% of the response sample, both SAGAL and representative sampling achieved 95% accuracy in all but two of the datasets. This represented a significant saving in grading effort as concluded by Basu et al. (2013), but the caveat of the homogeneous nature of the datasets should be noted.

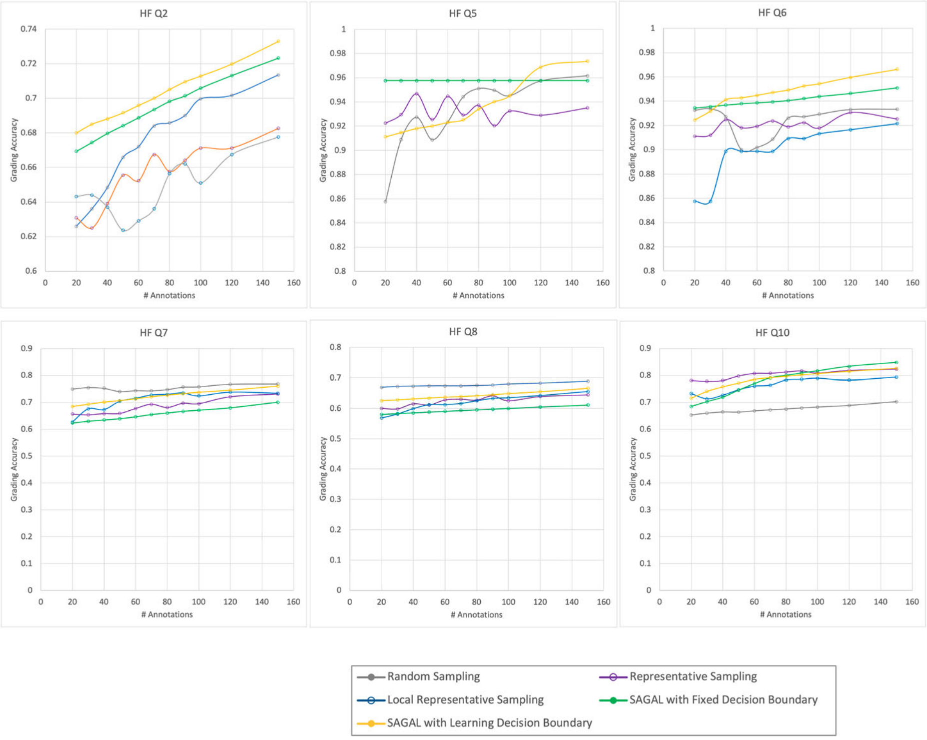

Figure 10. The performance of two variants of SAGAL and the baselines are compared based on the 6 response sets in the HF dataset.

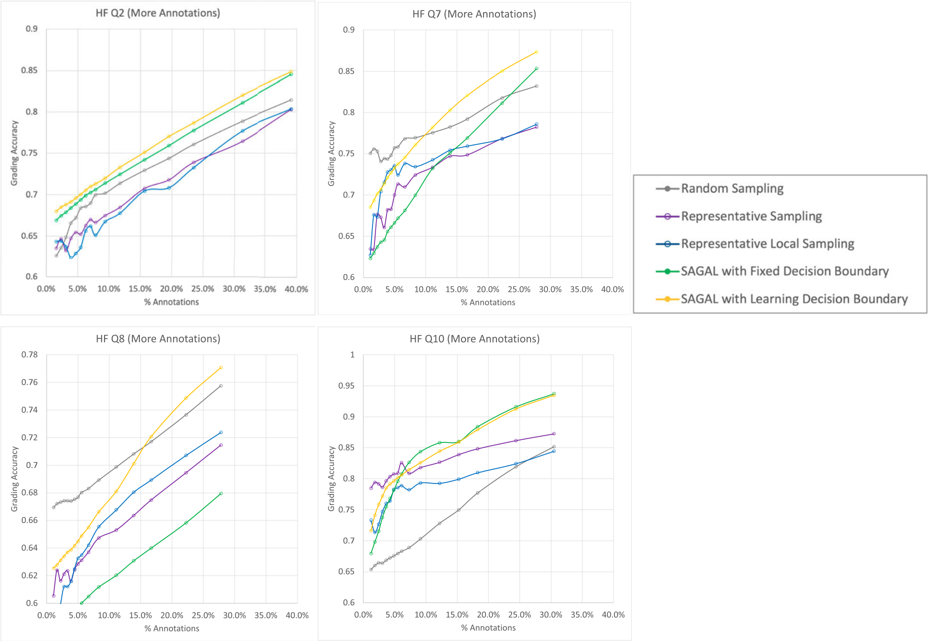

Figure 11. The performance of the two SAGAL variants on 4 selected HF response sets.

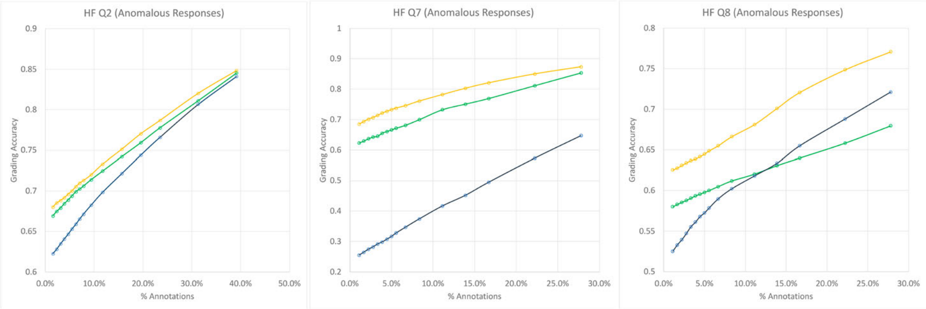

Figure 12. The performance of SAGAL without anomalous response handling (shown in dark blue) based on Q2, Q7, and Q8 response sets of the HF dataset.

SAGAL had a significant performance gain over the baselines in the ASAG tasks involving heterogeneous responses. In Figure 9, the findings from the SciEntsBank datasets showed a clear gap between SAGAL and the baselines from 20 annotations (equivalent to 20% of the sample), meaning that SAGAL was already working on the uncertainty in the decision boundaries after finding the major cluster peaks. Similarly as evidenced in Figures 10 and 11, SAGAL experienced difficulties with the many uncertain responses of the HF datasets at around grading budget of 20% of the sample. The heterogeneity of the two families of datasets highlighted the strength of SAGAL over other computer-assisted example acquisition methods.

The HF datasets came from open-ended questions and offered few and insignificant structures for exploration. The performance of the representative sampling was similar to that of random sampling. SAGAL experienced a long cold-start period (Barata et al., 2021), as it required an annotation level at around 10% to 20% of the sample to locate the decision boundaries and to resolve the uncertain responses. SAGAL was found to recover from the cold-start problem effectively even in the most heterogeneous datasets. Active learning facilitated accurate identification of borderline responses and it therefore worked better in grading tasks expecting many borderline responses.

SAGAL handled class imbalance datasets, such as USCIS Q5 and SciEntBank EV_25, quite well. Labelling the anomalous responses as wrong answers helped alleviate over-representation of correct answers in the representative samples. Reasonable decision boundaries could established in the cold-start stage in SAGAL models (Figure 12).

In addition, the ε-neighbourhood defined by the hyper-spherical decision boundaries was also useful to initialize the decision boundaries in the cold-start stage. The radius ε was initially computed from ranked inter-response distance and that of the individual decision boundary was updated with the grading model. The ε-neighbourhood with learning facilities worked well in datasets with large samples, such as the HF datasets. A fixed radius was a conservative choice for the datasets of sample

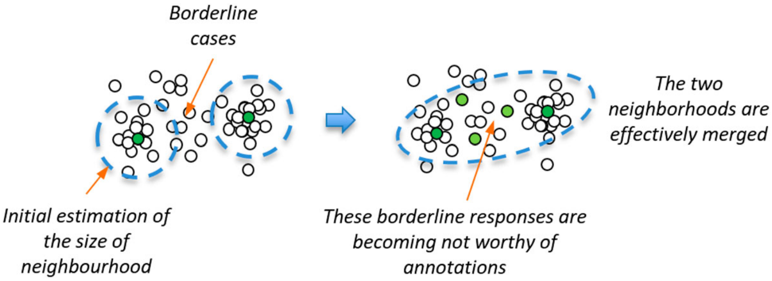
Borderline cases
Initial estimation of the size of neighbourhood
These borderline responses are becoming not worthy of annotations
The two neighborhoods are effectively merged

Figure 13. The diminished importance of the initial estimation of the size of ε-neighbourhood and the effective merger of neighbourhoods.

size inadequate for making reasonable estimation. The importance of the initial estimation of the neighbourhood size diminished as the grading model is more knowledgeable. The annotations of borderline responses gradually redefined the decision boundaries and even observed that two one-class enclosures merged into a large one as illustrated in Figure 13.

Conclusion

The need for an improved computer-assisted grading example acquisition method for ASAG prompted the work described in the paper. An investigation of the types of potential examples valuable to the grading model and the relevant formulations for their identification was carried out. Guided by a rigorous experimental design, SAGAL, the solution proposed in the paper, was designed and implemented and then evaluated on three publicly available ASAG datasets. SAGAL was inspired by active learning, an approach in machine learning suitable for model building with small training samples. SAGAL was found to be generally better than the alternative example acquisition strategies based on semi-supervised learning. The findings have increased understanding of computer-assisted grading example acquisition and improved the appeal of ASAG in real-world applications.

Directions for further investigation are suggested as follows. A key finding of this work was the influence of the characteristics of datasets on grading model performance. Osugi et al. (2005) proposed an active learning algorithm that could dynamically adjust the importance between representativeness and informativeness according to data distribution. The algorithm was based on prioritizing two strategies of example sampling, namely, exploration and exploitation. The former aims to cover as much of the semantic feature space as possible through identifying the representative responses, while the latter targets at increasing the information on the uncertain parts of the space. The priority of the two strategies should change with situations. For example, the grading model at the early stage of active learning is mostly uncertain and exploration is a more suitable strategy, and exploitation is needed more for more widely distributed response sets. The adaptive emphasis on either one depending on the active learning stage and the characteristics of the response sets is worthwhile to investigate further. Another proposed direction is to extend SAGAL from the current 2-way grading to multi-way grading and other ordinal scales. An enriched definition of borderline responses is essential for differentiating minor and major grade separation (Zhang et al., 2019).

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).

(1) SciEntsBank dataset downloaded from https://github.com/dbbrandt/short_answer_granding_capstone_project
(2) HF dataset downloaded from https://www.kaggle.com/c/asap-sas
(3) USCIS dataset downloaded from https://www.microsoft.com/en-us/download/details.aspx?id=52397

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by UGC Research Grants Council [grant number UGC/FDS16/E10/19].

Notes on contributors

Andrew Kwok-Fai Lui received the PhD degree from The Australian National University, Canberra, ACT, Australia, in 1998. He is currently a Professor at the Department of Electronic Engineering and Computer Science, 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 is currently a Senior Lecture with the School of Computing and Information Science at Anglia Ruskin University, UK. 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 was a Research Assistant with the School of Science and Technology, Hong Kong Metropolitan University when this work was carried out.

ORCID

Andrew Kwok-Fai Lui http://orcid.org/0000-0003-4990-7570

References

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 259–278. https://doi.org/10.1080/0969594X.2010.546775

Barata, R., Leite, M., Pacheco, R., Sampaio, M. O., Ascensão, J. T., & Bizarro, P. (2021). Active learning for imbalanced data under cold start. *Proceedings of the Second ACM International Conference on AI in Finance*. pp. 1–9. Association for Computing Machinery.

Barr, J. R., Bowyer, K. W., & Flynn, P. J. (2014). Framework for active clustering with ensembles. *IEEE Transactions on Information Forensics and Security*, *9*(11), 1986–2001. https://doi.org/10.1109/tifs.2014.2359369

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402. https://doi.org/10.1162/tacl_a_00236

Brooks, M., Basu, S., Jacobs, C., & Vanderwende, L. (2014, March 4-5). Divide and correct: Using clusters to grade short answers at scale. *Proceedings of First ACM Conference on Learning@Scale Conference. First ACM Conference on Learning@Scale Conference*. ACM, pp. 89–98.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, *55*(2), 489–499. https://doi.org/10.1016/j.compedu.2010.02.012

Cardoso, T. N., Silva, R. M., Canuto, S., Moro, M. M., & Gonçalves, M. A. (2017). Ranked batchmode active learning. *Information Sciences*, *379*, 313–337. https://doi.org/10.1016/j.ins.2016.10.037

Chen, Y., Hu, X., Fan, W., Shen, L., Zhang, Z., Liu, X., Du, J., Li, H., Chen, Y., & Li, H. (2020). Fast density peak clustering for large scale data based on kNN. *KnowledgeBased Systems*, *187*, 104824. https://doi.org/10.1016/j.knosys.2019.06.032

Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2), 201–221.

Du, B., Wang, Z., Zhang, L., Zhang, L., Liu, W., Shen, J., & Tao, D. (2017). Exploring representativeness and informativeness for active learning. *IEEE Transactions on Cybernetics*, *47*(1), 14–26. https://doi.org/10.1109/tcyb.2015.2496974

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., & Dang, H. T. (2013). *Semeval2013 Task 7: The joint student response analysis and 8th recognizing textual entailment challenge*. North Texas State University.

Englhardt, A., Trittenbach, H., Vetter, D., & Böhm, K. (2020, May 5-8). Finding the sweet spot: Batch selection for OneClass active learning. *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, pp. 118–126.

Gaddipati, S. K., Nair, D., & Plöger, P. G. (2020). Comparative evaluation of pretrained transfer learning models on automatic short answer grading. *arXiv preprint arXiv:2009.01303*.

Galhardi, L. B., & Brancher, J. D. (2018). Machine learning approach for automatic short answer grading: A systematic review. *Ibero-American Conference on Artificial Intelligence*. pp. 380–391.

Gu, B., Zhai, Z., Deng, C., & Huang, H. (2020). Efficient active learning by querying discriminative and representative samples and fully exploiting unlabeled data. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9), 4111–4122.

Huang, S. J., Jin, R., & Zhou, Z. H. (2014). Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10), 1936–1949.

Horbach, A., & Palmer, A. (2016, Jun 16). Investigating active learning for short-answer scoring. *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*. pp. 301–311.

Hou, J., & Cui, H. (2017, May 16-18). Density normalization in density peak based clustering. *International Workshop on GraphBased Representations in Pattern Recognition*. Springer, pp. 187–196.

Hou, J., & Pelillo, M. (2016, Dec 4-8). A new density kernel in density peak based clustering. *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 468–473.

Khan, S. S., & Madden, M. G. (2014). Oneclass classification: Taxonomy of study and review of techniques. *The Knowledge Engineering Review*, *29*(3), 345–374. https://doi.org/10.1017/S026988891300043X

Kishaan, J., Muthuraja, M., Nair, D., & Plöger, P. G. (2020, Jul 18). Using active learning for assisted short answer grading. *ICML 2020 Workshop on Real World Experiment Design and Active Learning*..

Kumar, P., & Gupta, A. (2020). Active learning query strategies for classification, regression, and clustering: A survey. *Journal of Computer Science and Technology*, *35*(4), 913–945. https://doi.org/10.1007/s11390-020-9487-4

Lewis, D. D., & Catlett, J. (1994, Jul 10-13). Heterogeneous uncertainty sampling for supervised learning. *Proceedings of the 11th International Conference on Machine Learning*. Elsevier, pp. 148–156.

Lewis, D. D., & Gale, W. A. (1994, Aug 1). A sequential algorithm for training text classifiers. *Proceedings of Annual Conference of ACM Special Interest Group in Information Retrieval (SIGIR'94)*. Springer, pp. 3–12.

Livingston, S. A. (2009). *Constructed-response test questions: Why we use them; how we score them. R&D connections. Number 11*. *Educational Testing Service*. ERIC.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I – outcome and process. *British Journal of Educational Psychology*, *46*(1), 4–11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Marvaniya, S., Saha, S., Dhamecha, T. I., Foltz, P., Sindhgatta, R., & Sengupta, B. (2018, Oct 22-26). Creating scoring rubric from representative student answers for improved short answer grading. *27th ACM International Conference on Information and Knowledge Management*, pp. 993–1002.

Mohler, M., Bunescu, R., & 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.

Nguyen, H. T., & Smeulders, A. (2004, Jul 4-8). Active learning using pre-clustering. *Proceedings of the Twenty-first International Conference on Machine Learning*. pp. 79–86.

Osugi, T., Kim, D., & Scott, S. (2005, Nov 27-30). Balancing exploration and exploitation: A new algorithm for active machine learning. *Proceedings of Fifth IEEE International Conference on Data Mining (ICDM'05)*. p. 8.

Peters, J., & Jankiewicz, P. (2012). *The William and Flora Hewlett Foundation Automated Student Assessment Prize (ASAP). ASAP Short Answer Scoring Competition System Description*. [online] Kaggle. Retrieved 31 Jan. 2022. http://kaggle.com/asap-sas/

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*(6191), 1492–1496.

SaarTsechansky, M., & Provost, F. (2004). Active sampling for class probability estimation and ranking. *Machine Learning*, *54*(2), 153–178.

Sawatzki, J., Schlippe, T., & BennerWickner, M. (2022). Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. In Eric C. K. Cheng, Rekha B. Koul, Tianchong Wang, & Xinguo Yu (Eds.), *Artificial intelligence in education: Emerging technologies, models and applications* (pp. 65–75). Springer.

Shi, Y., Yu, Z., Cao, W., Philip, C. C. L., Wong, H.-S., & Han, G. (2020). Fast and effective active clustering ensemble based on density peak. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3593–3607. https://doi.org/10.1109/TNNLS.2020.3015795

Xu, Z., Yu, K., Tresp, V., Xu, X., & Wang, J. (2003, Apr 14-16). Representative sampling for text classification using support vector machines. *European Conference on Information Retrieval*. Springer, pp. 393–407.

Zesch, T., Heilman, M., & Cahill, A. (2015, Jun 4). 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, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2022). An automatic shortanswer grading model for semiopenended questions. *Interactive Learning Environments*, *30*(1), 177–190. https://doi.org/10.1080/10494820.2019.1648300

Zhang, Y., Cheung, Y., & Tan, K. C. (2019). A unified entropy based distance metric for ordinal and nominal attribute data clustering. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(1), 39–52. https://doi.org/10.1109/TNNLS.2019.2899381