An Interactive Short Answer Grading System Based on Active Learning Models

Andrew Kwok-Fai Lui*
Dept of Electronic Eng and Computer Science
Hong Kong Metropolitan University
Hong Kong SAR China
0000-0003-4990-7570

Sin-chun Ng
School of Computing and Information Science
Anglia Ruskin University
Cambridge, UK
0000-0002-2972-530X

Stella Wing-Nga Cheung
Dept of Electronic Eng and Computer Science
Hong Kong Metropolitan University
Hong Kong SAR China

***Abstract*—Grading automation can improve learning experience with quick around-the-clock feedback and superior grading consistency. Obtaining annotated data for training short answer grading models is costly. Active learning has been proven an effective approach to build accurate models with few annotated data. This paper presents an active learning approach of short answer grading that comprises of a few novelties. The first is a specialized active learning formulation adapted to short answer grading principles. The second is a proposal to exploit human expertise in fine-tuning several active learning model parameters for adaptation to the specifics of each grading task. The third is an interactive short answer grading system that is designed for building better quality grading model by informing users with data visualizations. The prototype presented in the paper should provide a useful conceptual demonstration for real-life deployment of active learning for short answer grading and further research in an enhanced interactive form of active learning.**

***Keywords*—automated short answer grading, active learning, generalized uncertainty sampling, interactive grading system, graphical user interface**

I. Introduction

Short answer question is a popular assessment type testing the recall of relevant knowledge. To provide short answer questions in online learning, a scalable, consistent, and responsive grading technology is essential. Automated short answer grading (ASAG) is an educational technology that uses a computational process to grade answers. The most challenging problem in ASAG development is the modelling of grading decisions. An accurate grading model requires comprehensive knowledge of the correct reference answers, exceptions, and the differentiation of borderline answers. The machine learning approaches to building ASAG models acquire the knowledge in two steps. First, a sample of annotated short answers is collected as examples of grading decisions, and then the examples are generalized for grading decisions on unseen answers.

Grading an answer is the annotation of the answer with a class label such as correct, wrong, or some levels of partially correct. Human annotation is slow and costly, and grading automation has largely reduced the involvement of human graders. However, machine learning models are generally data hungry, and more annotated data there are the better are the models' performance. As a further reduction of the number of human annotations is desirable, a more effective approach of building grading models is needed.

The performance of machine learning grading models is as good as the annotated data used to train them. A good quality sample of annotated answers should be representative of major grading decisions and illuminating for borderline cases. In addition, the sample should be devoid of redundancy and obvious anomalous cases such as *"I don't know"*. Consider that there is a budget of human annotations, and a fixed number of answers will be annotated, it is potentially advantageous to optimize the sample of annotated answers. Query strategies refer to the algorithms used for selecting or sampling data for annotations. A range of algorithms based on approaches such as analysis of features of the unannotated and annotated data, the generalization error of the model, and others have been proposed and studied.

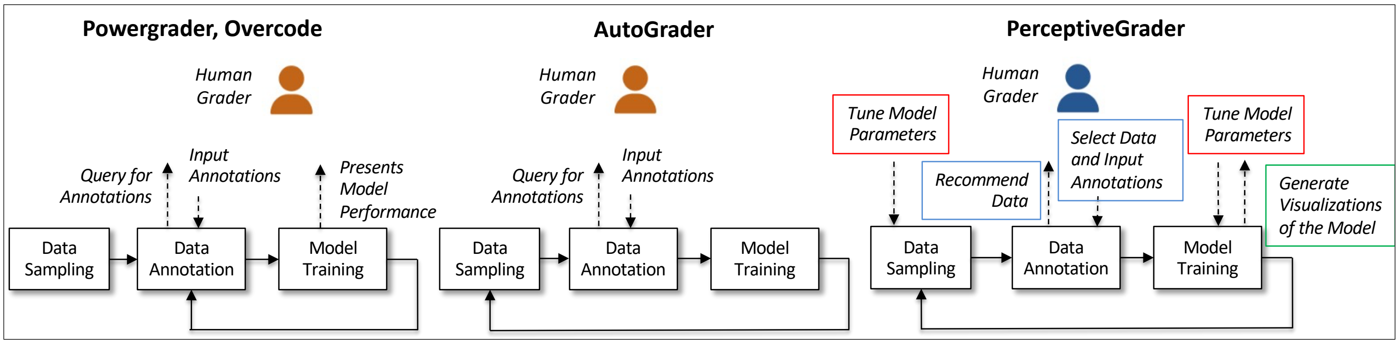


Fig. 1. An illustration of the unique features of PerceptiveGrader, in comparison to the existing work including PowerGrader, Overcode and AutoGrader.

Every ASAG task will operate with certain model parameters unique to the context of the question. For examples, parameters such as the largest semantic deviations between equivalent answers and the uncertainty levels for borderline answers may be different from one question to another. Tuning these model parameters may improve the model quality which may consequentially improve the effectiveness of query strategies. This paper proposes an interactive short answer grading system with provisions for human graders to adjust on the model parameters that have been initialized computationally. The system has two major functions, which include, first, answer sampling with interactive adjustments, and second, answer annotation. The two functions are integrated under a user interface augmented with visualizations of various aspects of the grading task for making informed decisions.

A. *A Review of Interactive Grading Systems*

Although human annotation is an essential task in building ASAG systems, the majority of existing ASAG studies assumed annotated dataset already available and factored out the costly task that would render grading automation infeasible. Few studies have discussed effective user interface design for both functions. *AutoGrader-Beta* is a short answer grading system that provides a question-wise view in addition to a student-view as an organization of the answers [7]. It provides a basic interface, consisting of the question text, answer text, grade options, and feedback options, for the input of annotation and feedback. Maintaining consistency between annotations is not easy without a suitable tool. *PowerGrader* is another interactive grading system that provides visualizations of similar answers and their annotations and functions for decision reversal and fine-tuning [1], [2]. *Overcode*, a grading system for computer programs, offers a view of near-duplicated answers to avoid redundancy and inconsistency in the annotation [4]. Many computer graders (e.g., [4], [18], [19]) are designed with a focus on visualization to assist the human annotations.

In the aspect of answer sampling and query strategies, *Overcode* and *PowerGrader* have included a sampling algorithm that recommends the most representative answers for the annotation. The user grader can annotate sampled answers and observe the updated grading model interactively. Based on the perceived performance of the grading model, the user makes the call to stop further annotations. *AutoGrader-Beta* have used a different sampling algorithm that recommends the most uncertain answers for the annotation. According to a preset budget, the system queries for annotations and updates the grading model. The sampling algorithms used in all the reviewed systems appear to have the model parameters tuned and fixed for all tasks.

This paper presents *PerceptiveGrader*, which is superior to the existing systems in several aspects. Fig. 1 illustrates the similarities and differences between *PreceptiveGraders* and the others. First, the user interface allows interactive tuning of model parameters so that some question specific settings may be incorporated into model building. Second, the user interface presents a list of recommended answers and imposes no restrictions on the annotation task before the re-training of model. For example, the user grader can decide to annotate more if the recommendations appear to be diverse or to annotate less if there are many redundancies. Third, the user interface offers a range of visualizations on various aspects of the models.

The remainder of this paper is structured as follows. Section II describes in detail the active learning formulation for short answer grading to be used in the proposed interactive grading system. Section III then explains the system design of PerceptiveGrader, including the operation procedure and the user interface design. The prototype implementation is used to show the feasibility of the proposed ideas and illustrate examples of the user interactions in model parameter tuning. Section IV presents the key findings in a preliminary study major observations from a system evaluation. Section V concludes the paper with suggestions of research directions.

II. Active Learning

A. *Machine Learning Approaches for Building ASAG Models*

The machine learning approaches for building ASAG models include supervised learning, unsupervised learning, and active learning. Fig. 2 illustrates their differences. The supervised learning approach assumes a pre-existing large sample of annotated data. Recent methods of this approach are often designed to exploit the large amount of data in learning latent powerful features with deep learning algorithms. The unsupervised learning approach uses cluster analysis to select a sample of highly representative data for annotation. The studies on *PowerGrader* showed that a reasonable quality grading model could be built from a small sample of annotated data [1-2].

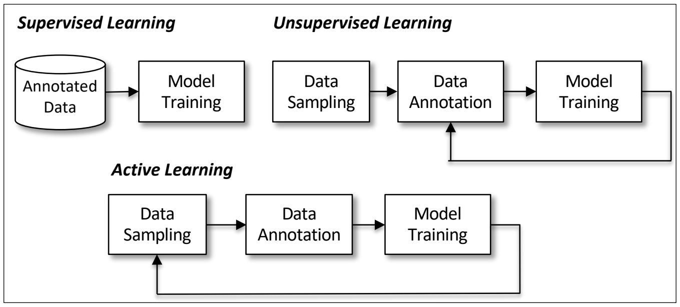
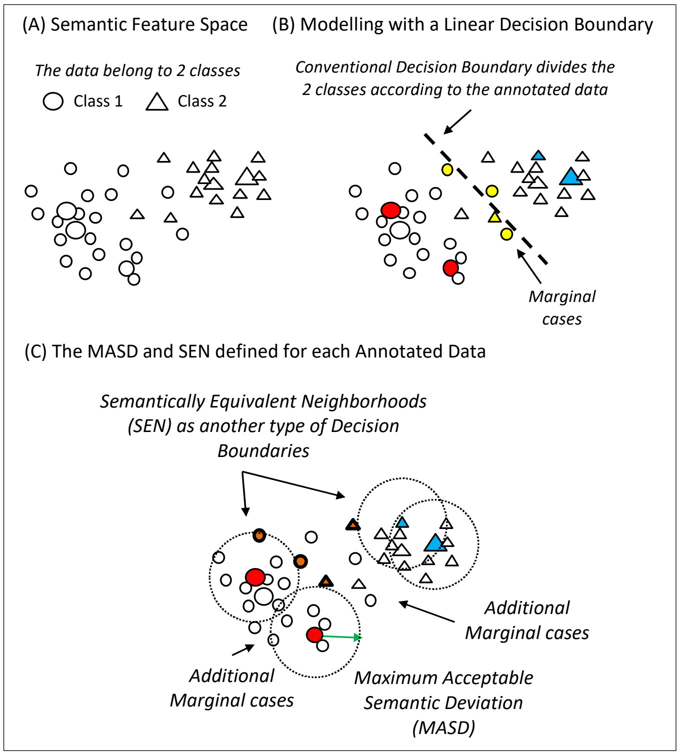


Fig. 2. Comparison between the approaches of active learning, supervised learning and unsupervised learning.

In the sampling of data for annotation, the active learning approach exploits features in the trained models in addition to features in the unannotated data. It enabled a more diversified sampling that includes particularly the uncertain and borderline cases. Many general formulations for data sampling have been proposed and evaluated. They are based on mainly three metrics namely representativeness, informativeness, and diversity. Representativeness refers to the ability of samples to match the features of unannotated data, and the centroids of clusters are examples of representative sample. Information refers to the ability of samples in reducing uncertainty errors, and the borderline data around the decision boundaries are examples of informative sample. Diversity refers to the mutual differences between samples. The three metrics has been proven effective in formulations for reducing the number of annotations for reaching the same accuracy in model training.

The active learning approach has been shown applicable in domains including computer vision [11], materials design [12], and credit card fraud detection [13]. Each application domain may vary in the definitions of representativeness, informativeness, and diversity. Accordingly, new formulations based on the three general metrics were found in the adaptations of active learning to the above domains. Studies in active learning for short answer grading are rarely found in the literature [5], [7], and in both studies general formulations were evaluated.



(A) Semantic Feature Space
(B) Modelling with a Linear Decision Boundary
The data belong to 2 classes
○ Class 1 △ Class 2
Conventional Decision Boundary divides the 2 classes according to the annotated data
Marginal cases
(C) The MASD and SEN defined for each Annotated Data
Semantically Equivalent Neighborhoods (SEN) as another type of Decision Boundaries
Additional Marginal cases
Additional Marginal cases
Maximum Acceptable Semantic Deviation (MASD)

Fig. 3. A binary classification grading model (left) is solved with a simple model using a linear decision boundary and borderline cases found along the boundary (middle). However, the maximum acceptable semantic deviation defines the semantic equivalent neighborhood and the data around and beyond the edges are also borderline cases (right).

B. *Specific Features in Short Answers*

To develop specialized active learning formulations for short answer grading, the specific features in short answers must be recognized and exploited.

1) Frequent Correct Answers. The content of many correct answers are similarly constructed. If there are more than one way to answer correctly, then they can be easily identified as clusters or modes of answers in a cluster analysis. The central answers in the cluster can represent the whole cluster.

2) Common Misconceptions. Similarly, some wrong answers may be similarly constructed that represent common misconceptions. Highly representative answers may also be identified as in the frequent correct answers.

3) Maximum Acceptable Semantic Deviation (MASD). Answers with a small amount of semantic deviation may be accepted as equivalent. Given an annotated answer, there exists a neigbourhood within which the unannotated answers are considered semantically equivalent. The MASD defines the maximum semantic deviation which is the radius of the neigbourhood.

4) Semantically Equivalent Neighbourhood (SEN). The semantically equivalent neighbourhood defined by the MASD is a specialized decision boundary that differentiates a known grade on one side and uncertain grade on the other side. This is different from the decision boundaries that divides answers of two grade classes.

5) Borderline Answers. The grades of borderline answers are highly uncertain. Generally the borderline cases are those with a large distance from the nearest annotated data and those roughly equdistant from two data annotated with different classes. The definition of SEN has redefined the marignal cases to be those beyond the boundaries of the neighbourhoods.

6) Anomalous Answers. The anomalous answers are those irrelevant to the questions such as *"I don't know"* and *"Too difficult"*. These answers are semantically very distant from the frequent correct answers and wrong answers. The minimum distance for identification of anomalous answers is defined as anomalous data trigger (ADT).

C. *Interactive Active Learning for Short Answer Grading*

This section proposes a formulation of active learning for short answer grading. Assume that there is a dataset of short answers S which have been mapped to a semantic hyperspace according to a text representation. The active learning formulation does not specify requirements on the text representation but generally expects a high-dimensional compact representation such as these based on distributional semantics.

The interactive active learning for short answer grading formulation is outlined in Algorithm 1 below. The user is given the final say on when to stop further annotation. The algorithm is in a cycle as depicted in the bottom of Fig. 2. In the algorithm, lines 1 to 3 are for pre-computation of essential model parameters and lookup table, lines 4 to 14 is the interactive cycle of data sampling (line 5 to 9), data annotation (line 10 to 13) and model update (line 14).

The representativeness of data is evaluated based on density peak clustering with a boundary for density computation according to the MASD [14]. The definition for the density is given in eq. (1) where $d\left(s_i, s_j\right)$ is the Euclidean distance between two answers in the semantic feature space, and W_d is the weight for scaling the representativeness formulated as density and other factors in the answer sampling for annotation.

$$Den_{s_i} = log_{W_d} \sum\nolimits_{s_j \in SEN} exp(-\frac{d\left(s_i, s_j\right)^2}{MASD^2}) \qquad (1)$$

The *IFGain* is the information gain in annotating an answer according to the current estimation of an entropy-based uncertainty and the expected entropy-based uncertainty after annotation (i.e., which should be 0). There is a specific *IFGain* for an answer in every subspace. The *IFGlobal* is the uncertainty of the class of an answer in different subspaces. The formulations for these are inspired by [15]. The grading value GV_{s_i} for every answer is computing with eq. (4), where W_G is the weighting parameter between local and global uncertainty.

$$GV_{s_i} = \left(\frac{\sum_{sp_p} IFGain_{s_i,p} \times Den_{s_i}}{n}\right) + W_G \times IFGlobal_{s_i} \quad (2)$$

Algorithm 1 Active Learning for Short Answer Grading

Input: A dataset of short answers $S = \{s_i\}_{i=1}^{n}$ where n is the number of answers, and the dataset contains annotated answers and unannotated answers represented as S^G and S^U respectively where $S = S^G \cap S^U$. A graded answer s^G has a manually given class $y \in Y$ such that $L(s^G) = y$ where Y is the set of possible grade classes.

Output: The grading model $F(S, Y): S \rightarrow Y$, so that every answer in the S^U can receive a grade class from the model.

```
Precompute MASD and ADT with unsupervised
learning from $S$
Precompute M subspaces based on random sampling
$\{sp_0, sp_1, \dots, sp_{M-1}\}$
Precompute $Den_{s_i,p}$for $s_i \in S$ and $p \in \{sp_{0..M-1}\}$
while the user is willing to improve the model do
    for $0 \leq p < M$ do
        Compute $IFGain_{s_i}$ for $s_i \in S$ in $sp_p$
    Compute $IFGlobal_{s_i}$ for $s_i \in S$
    Compute the grading value $GV_{s_i}$
    Rank answers in $S^U$  based on $GV_{s_i}$
    Recommend the top answers to the user
    while the user is willing to annotate answers do
        Annotate a selected answer $s_i \in S^U$and move
        $s_i$ to $S^G$ so that $S^G = S^G \cup \{s_i\}$
    Update the model $F(S, Y)$
```

The model parameters W_d and W_G are critical in the balance between exploration and exploitation [9]. Active learning is an iterative learning process about the feature space and the data distribution. Exploitation focusses on learning more from known data distribution and exploration focusses on learning more about unknown regions in the feature space.

D. *Tunable Model Parameters*

The interactive active learning formulation for short answer grading comprises several model parameters that determine the following characteristics.

1. The maximum semantic deviation between two answers considered as equivalent (MASD).
2. The minimum semantic deviation from the most representative answers of an answer considered as an anomaly and a wrong answer (ADT).
3. The balance between exploration and exploitation in answer sampling for annotation (W_d and W_G).
4. The balance between local exploration and global exploration (W_G).
5. The grading value for every answer (the user could choose to ignore certain recommendations).
6. The batch size, which is the number of annotations between two model updates (the user could choose to stop annotation).

The model parameters 1 and 2 have a default value computed with an unsupervised learning method on the unannotated data. The default values for parameters 3 and 4 are system preset. There is no default values for parameters 5 and 6 as the decisions are purely based on user interactions.

III. Perceptive Grader

A. *Overview*

PerceptiveGrader is client-server system as shown in Figure 3. The server is responsible for the execution of active learning models, and in addition the management of submitted answers and the encoding of short answers into semantic representations. Every ASAG task is treated as a project, which comprises of the annotated dataset, the adjusted model parameters, and the updated grading model. The submitted answers may be stored in a database or the file system. No submission functionality is included in the server, and instead allowing the server to interface with various existing submission or assessment systems. The client is executed in a web browser. It is responsible for user interactions and data visualization. Internet protocols connect the client and server. Fig. 4 describes graphically the system design of PerceptiveGrader.

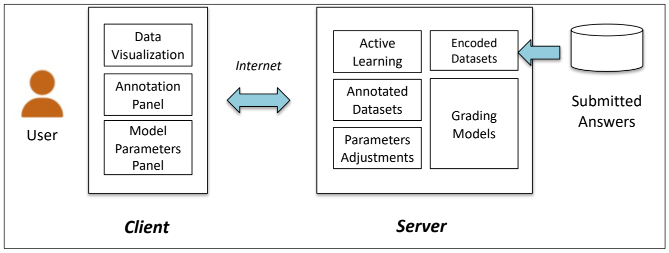


Fig. 4. The system design of PerceptiveGrader.

B. *System Design*

The model building can operate in a stream-based approach or a pool-based approach. The difference between the two approaches is on how decisions are made on the sampling of data for human annotations. The former samples each data in an incoming data stream and the latter samples a pool of collected data. In the stream-based approach, a grading project may start as soon as the first submissions arrive. A small quantity data is less likely to represent the eventual data distribution and the reliability of the generalized uncertainty measure is hampered. The pool-based approach can better exploit the cost-effectiveness of active learning. Starting with a reasonable data sample size provides the algorithm and the user a better representation of the eventual data population.

Fig. 5 illustrates a likely scenario of interactive building of grading models. The human grader may be needed again in the future when the grading model performance is found unacceptable after more answers are submitted.

C. Prototype Implementation

A prototype implementation was completed for proof of the concepts. The server side was developed with *flask* for developing web applications and the suite of *sklearn*, *numpy*, and *pandas* for the implementation of the active learning algorithm.

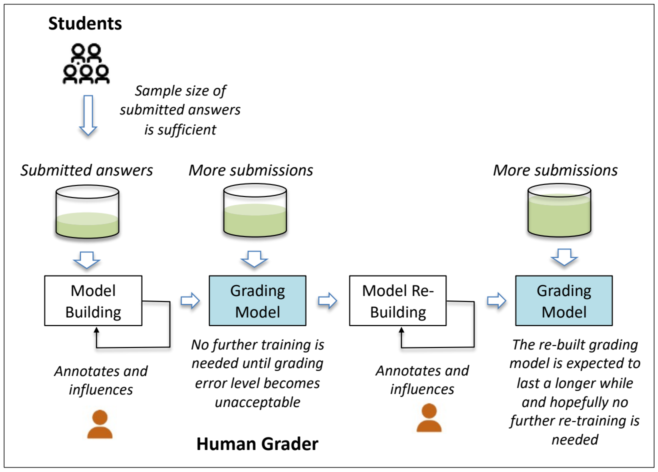

Students
Sample size of submitted answers is sufficient
Submitted answers
More submissions
More submissions
Model Building
Grading Model
Model Re-Building
Grading Model
Annotates and influences
No further training is needed until grading error level becomes unacceptable
Annotates and influences
The re-built grading model is expected to last a longer while and hopefully no further re-training is needed
Human Grader

Fig. 5. Multi-stage interactive building of grading model. The active learning algorithm works better after a reasonable sample size of submitted answers is received. The grading model resulting from the initial model building may last a while until the grading error becomes unacceptable.

D. ASAG Datasets for Demonstration

Three gold standard short answer datasets were used to illustrate the operation of *PerceptiveGrader*. The *USCIS* dataset [1] contains 20 sets of question and answer sampled from the United States Citizenship Examination. The *SCIENT* dataset [17] contains 197 question and answer sets of the Science domain. The *Hewlett Foundation* dataset [16] contains 10 sets of question and answers of which the mean length of answers is significantly longer than that from the *USCIS* dataset. Table I below summarizes the questions selected for the demonstration and their key characteristics. The differences in the number of answers and the mean length of answers can demonstrate the flexibility of the system.

E. The Main Client Panel

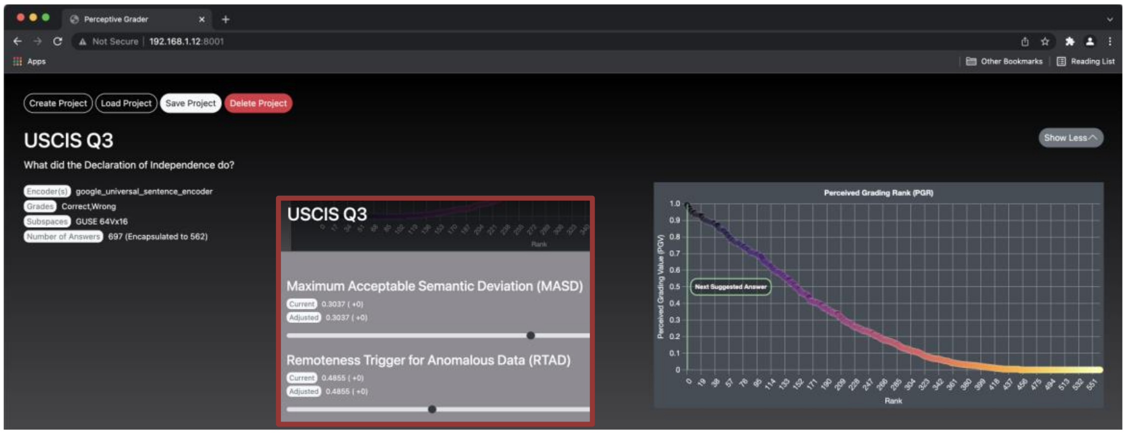
The layout of the main client panel is shown in Fig. 6. The main panel displays the key characteristics of the project, including the question text and some statistics and distributions of the answer set. The bottom part of the panel displays a 2D projection of the answers in the feature spaces. The most frequent answers and the anomalous answers have been removed from the projection. Clicking on the distribution panel will bring up the MASD and ADT parameter adjustment panel. The panel allows adjustments according to the interactions with the annotation panel.

TABLE I. The Questions Selected for Demonstration and the Key Characteristics of the Answer Sets

Datasets	Question	# Answers	Length
USCIS Q3	What did the Declaration of Independence do?	698	8 words
SCIENT EV_12b	Alice planted one radish seed in each of 5 separate pots … What is the range of tolerance for water for these radish seeds? Explain how you decided the range of tolerance.	100	13 words
Hewlett Q6	List and describe three processes used by cells to control the movement of substances across the cell membrane.	1797	50 words
Hewlett Q10	What is the effect of different lid colors on the air temperature inside a glass jar exposed to a lamp?	1640	60 words

F. The Annotation Panel and the Answer Panel

Fig. 7 shows the annotation panel which is located below the main client panel. The top ranked answers, according to the generalized uncertainty, are displayed on the first page. The other selected answers are displayed on next pages. There is a button for re-training the grading model, probably after the annotations of several answers.



USCIS Q3

What did the Declaration of Independence do?

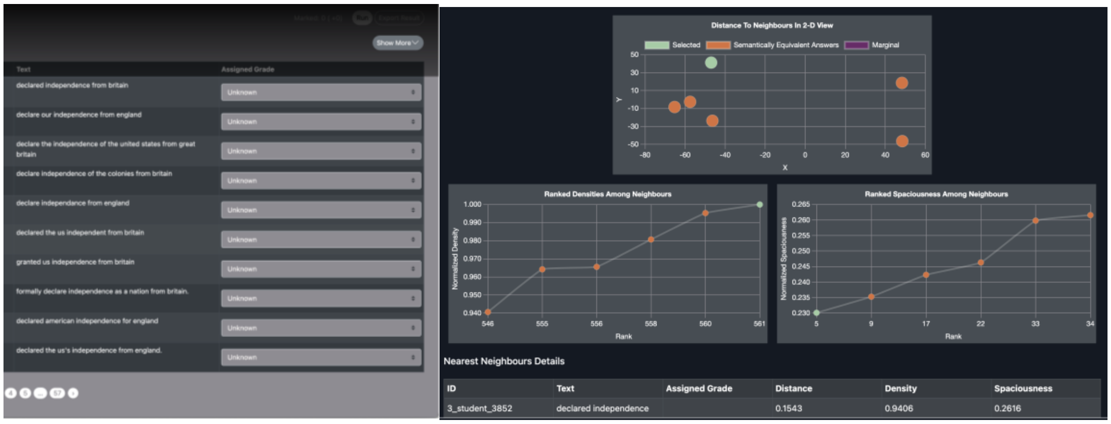
Encoder(s) google_universal_sentence_encoder
Grades Correct,Wrong
Subspaces GUSE 64Vx16
Number of Answers 697 (Encapsulated to 562)

USCIS Q3

Maximum Acceptable Semantic Deviation (MASD)
Current 0.3037 (+0)
Adjusted 0.3037 (+0)

Remoteness Trigger for Anomalous Data (RTAD)
Current 0.4855 (+0)
Adjusted 0.4855 (+0)

Fig. 6. The main client panel of PerceptiveGrader. The project screen displays the key characteristics of the question-and-answer set. The speciousness ranked data distribution is shown on the right, from which the default values of MASD and ADT are shown. The bottom shows a 2D projection of the data distribution.



Text	Assigned Grade
declared independence from britain	Unknown
declare our independence from england	Unknown
declare the independence of the united states from great britain	Unknown
declare independence of the colonies from britain	Unknown
declare independance from england	Unknown
declared the us independent from britain	Unknown
granted us independence from britain	Unknown
formally declare independence as a nation from britain.	Unknown
declared american independence for england	Unknown
declared the us's independence from england.	Unknown

Nearest Neighbours Details

ID	Text	Assigned Grade	Distance	Density	Spaciousness
3_student_3852	declared independence		0.1543	0.9406	0.2616

Fig. 7. The annotation panel (left) where the top ranked answers may be assessed and graded. Clicking on an answer row would bring up the answer panel (right) that details neighborhood of the answer.

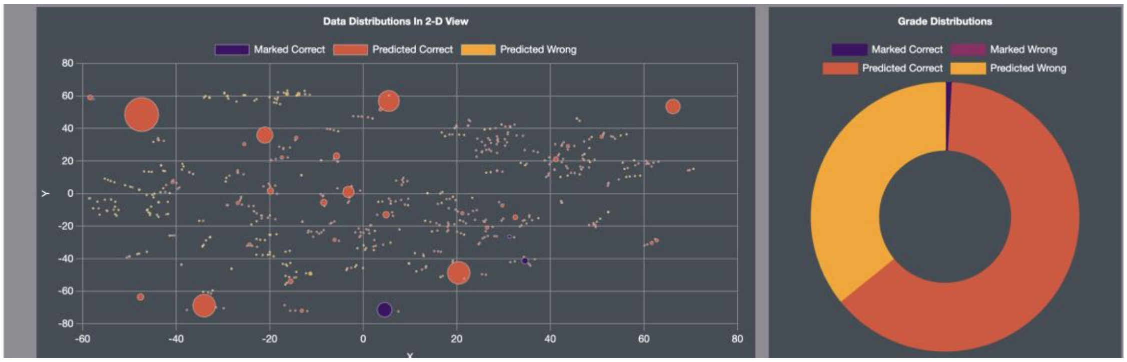


Fig. 8. The 2D projection of the semantic feature space illustrates the distributions of the answers and the current class distributions of the grading model (left). The grade distribution panel is shown on the side (right).

There are cases that human graders may perform better than state-of-the-art semantic encoding representations in recognizing semantic equivalence. In addition to evaluating the text content, distributions of similar answers and their annotations can help differentiate equivalent answers and borderline answers. The right-hand side of Fig. 7 shows the visualization of the neighborhood of a selected answer on the annotation panel.

G. Visualizations of the Grading Model

A good understanding of the status of the grading model is important. Human graders may refer to the data distributions and class distributions to make decisions on whether more or fewer annotations would make a difference. The 2D projections are produced by the t-SNE module [8]. The visualization panel provides several choices of projections that the graders can switch between. Fig. 8 illustrates the visualization panel and several clusters of correct answers. The rest consists of borderline answers and anomalous answers. The proportion of borderline answers is particularly high for this answer set.

IV. Results

A. Performance of the Active Learning Formulation

The performance of the active learning formulation for short answer grading, abbreviated to ALSAG, presented above is compared to random sampling as well as to representative-based sampling. The clustering algorithms of K-Means and Birch were selected as the method for the representative-based sampling.

Fig. 9 compares the grading accuracy of models built using our active learning formulations and the benchmarks. The horizontal axis indicates the percentage of the annotated answer in the datasets. Our formulation ALSAG performed better than the benchmark for all annotated percentages. Purposeful sampling was found better than random sampling. In addition, with 3% to 5% of the answers annotated, the active learning method could achieve nearly 90% accuracy.

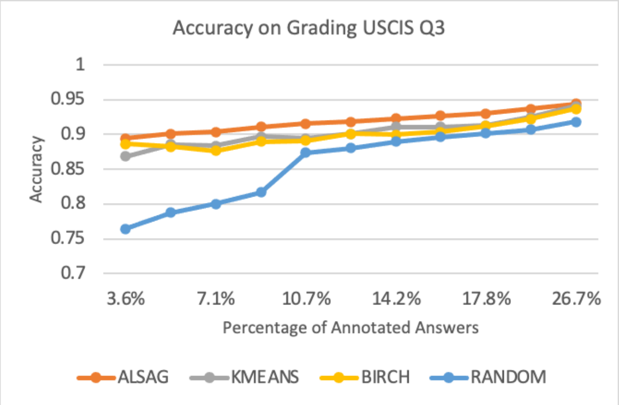


Fig. 9. Comparison of the grading accuracy of ALSAG to other benchmarks on the dataset of USCIS Q3.

The length of answers in the Hewlett foundation dataset is generally longer. Fig. 10 and Fig. 11 show that the grading accuracy of ALSAG models still performed better than the benchmark with Q6 and Q10 of the dataset.

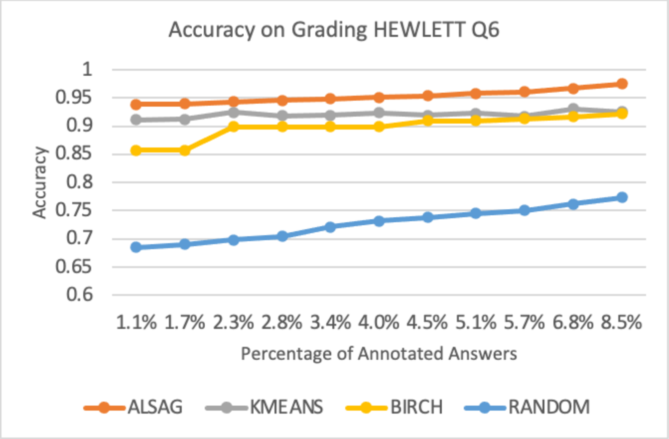


Fig. 10. Comparison of the grading accuracy of ALSAG to other benchmarks on the dataset of Hewlett Foundation Q6.

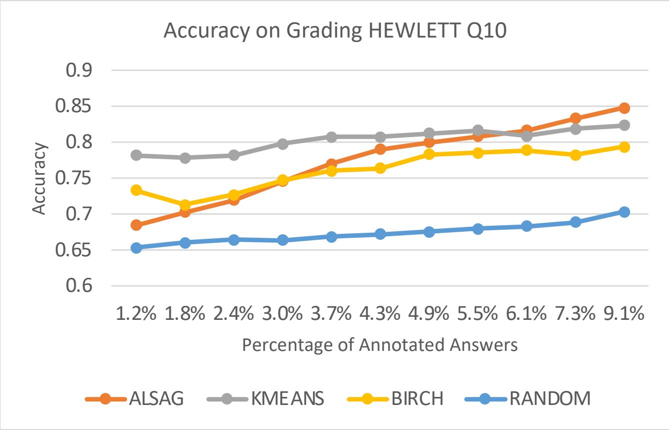


Fig. 11. Comparison of the grading accuracy of ALSAG to other benchmarks on the dataset of Hewlett Foundation Q6.

In depth evaluation of the performance of the active learning formulation is beyond the scope of this paper. The above results suggest that a balanced exploration and exploitation in our formulation should outperform just exploitation in the representative-based sampling methods, which is consistent with previous studies.

B. *Interactive Grading*

To illustrate the interactive tuning of active learning mode parameters, a simulated interactive grading session was carried out. A human grader was engaged and tasked to use *ProspectiveGrader* on the question *EV_12b* of the *SCIENT* dataset. This simulation focused on the adjustment of the rank of recommended answers based on data visualizations

Table II records the human grader's considerations of the recommended answers and the reasons of whether to follow the recommendation or to skip for the next recommendation. In the annotation #12 and #19, the human grader made references to the answer distribution and the neighborhood visualizations.

TABLE II. THE FIRST 20 DECISIONS MADE BY THE HUMAN GRADER ON THE RECOMMENDED ANSWERS

Annotation #	Decisions	Reasons
1-5	Followed	No reference
6	Followed	Predicted grade wrong
7	Skipped & Next	Predicted grade correct
8	Skipped & Next	Predicted grade correct
9-10	Followed	Predicted grade wrong
11	Skipped & Next	Predicted grade correct
12	Followed	Differentiate from neighbour
13	Skipped & Next	Predicted grade correct
14-15	Followed	Predicted grade wrong
16	Skipped & Next	Predicted grade correct
17	Followed	Predicted grade of neighbour wrong
18	Followed	Predicted grade wrong
19	Followed	Inconsistent between subspaces
20	Followed	Differentiate from neighbour

Fig. 12 compares the accuracy of the interactive grading to the original ALSAG model and other benchmarks. After 20 annotations, the interactive grading was found to perform better than ALSAG. It is a demonstration that the intelligence of the human grader could improve the sampling of answers for annotations.

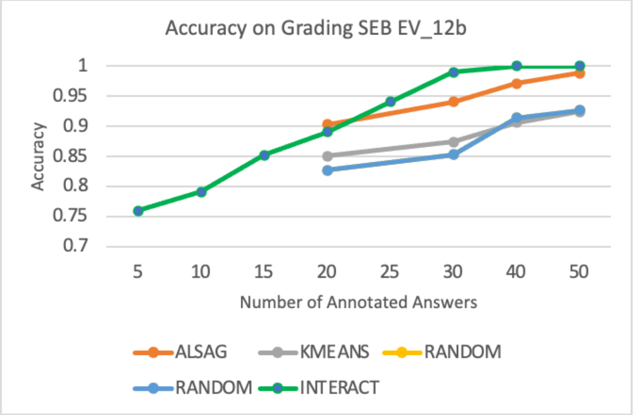


Fig. 12. Comparison of the grading accuracy of interactive grading (labeled as INTERACT) compared to ALSAG and other benchmarks.

V. Conclusion

A novel automated short answer grading system, known as *PerceptiveGrader*, has been presented. The core of the system is an active learning formulation designed for short answer grading. The formulation is the first that has taken into consideration of principles of short answer grading including acceptable semantic deviations and anomalous answers. The formulation has been evaluated with gold standard datasets and found outperform other benchmark methods. The same grading model accuracy can be achieved with fewer number of annotations.

Another novel feature of the system is the provision of interactive model parameter tuning. Every ASAG task may have a specific grading context and a corresponding optimal set of model parameters. The system resolves the issue with, first, unsupervised learning of the model parameters from the dataset, and second, interactive interface for fine-tuning the model parameters. The latter aims to exploit the external knowledge of human graders.

To enable the human graders making informed decisions, the system provided visualizations of the mode. The visualizations include global views of the distributions of answers, as well as local neighborhood views around selected answers. The knowledge acquired through interacting with the visualization enables the human graders to fine-tune model parameters and to estimate the performance of the updated models.

The prototype implementation of the grading system has served as a proof-of-concept of the novelties. The performance of the active learning formulation has been evaluated, and the prototype system has enabled hands-on testing of interactive fine-tuning of model parameters. A small-scale preliminary evaluation has been included in this paper. The quantitative and experiential findings will enable fixing the weaknesses and improving the performance.

In the next stage of development, a full-scale evaluation of both the active learning formulation and the interactive grading system will be carried out. For the former, the contribution of every representativeness, informativeness, and short answer feature to the data sampling accuracy is to be studied in detail. For the latter, the effectiveness of each visualization panel and the usability of the interface will be evaluated.

Short answer grading operates in a semantic feature space that is often defined by a pre-trained general sentence representation. Each ASAG task has a topical focus. Some subspaces are more relevant than other subspaces in the general feature space. A potential improvement to the active learning formulation is to add weights to the relevant subspaces. The significant differences in the semantics can be well represented.

Acknowledgment

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

[1] S. Basu, C. Jacobs, and L. Vanderwende, “Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading,” *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 391–402, Dec. 2013, doi: 10.1162/tacl_a_00236.

[2] M. Brooks, S. Basu, C. Jacobs, and L. Vanderwende, “Divide and correct: Using Clusters to Grade Short Answers at scale.,” in *Proceedings of First ACM Conference on Learning@Scale Conference*, 2014, pp. 89–98.

[3] B. Du *et al.*, “Exploring Representativeness and Informativeness for Active Learning,” *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 14–26, Jan. 2017, doi: 10.1109/tcyb.2015.2496974.

[4] E. L. Glassman, J. Scott, R. Singh, P. J. Guo, and R. C. Miller, “OverCode: Visualizing variation in student solutions to programming problems at scale,” *ACM Transactions on Computer-Human Interaction*, vol. 22, no. 2, pp. 1–35, Apr. 2015, doi: 10.1145/2699751.

[5] A. Horbach and A. Palmer, “Investigating active learning for short-answer scoring,” in *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*, 2016, pp. 301–311.

[6] S. Jayashankar and R. Sridaran, “Superlative model using word cloud for short answers evaluation in eLearning,” *Education and Information Technologies*, vol. 22, no. 5, pp. 2383–2402, Oct. 2016, doi: 10.1007/s10639-016-9547-0.

[7] J. Kishaan, M. Muthuraja, D. Nair, and P. G. Plöger, “Using Active Learning for Assisted Short Answer Grading,” presented at the ICML 2020 Workshop on Real World Experiment Design and Active Learning, Aug. 2020.

[8] L. V. D. Maaten and G. Hinton, “Visualizing Data Using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[9] T. Osugi, D. Kim, and S. Scott, “Balancing exploration and exploitation: A new algorithm for active machine learning,” in *Fifth IEEE International Conference on Data Mining*, Nov. 2005, pp. 8–16.

[10] S. Terman, “GroverCode: code canonicalization and clustering applied to grading,” Doctoral dissertation, Massachusetts Institute of Technology, 2016.

[11] Y. Siddiqui, J. Valentin, and M. Nießner, “Viewal: Active learning with viewpoint entropy for semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9433–9443.

[12] C. Kim, A. Chandrasekaran, A. Jha, and R. Ramprasad, “Active-learning and materials design: the example of high glass transition temperature polymers,” *MRS Communications*, vol. 9, no. 3, pp. 860–866, Sep. 2019, doi: 10.1557/mrc.2019.78.

[13] F. Carcillo, Y.-A. Le Borgne, O. Caelen, and G. Bontempi, “Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization,” *International Journal of Data Science and Analytics*, vol. 5, no. 4, pp. 285–300, Apr. 2018, doi: 10.1007/s41060-018-0116-z.

[14] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014, doi: 10.1126/science.1242072.

[15] Y. Shi, Z. Yu, W. Cao, C. L. P. Chen, H.-S. Wong, and G. Han, “Fast and Effective Active Clustering Ensemble Based on Density Peak,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3593–3607, Aug. 2021, doi: 10.1109/tnnls.2020.3015795.

[16] J. Peters, and P. Jankiewicz, The William and Flora Hewlett Foundation Automated Student Assessment Prize (ASAP). ASAP Short Answer Scoring Competition System Description. 2012. Downloaded from http://kaggle.com/asap-sas/

[17] M. O. Dzikovska et al., “Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge,” in *Proc. 7th Int. Workshop Semantic Eval.*, NAACL-HLT, Atlanta, GA, USA, Jun. 14–15, 2013, pp. 263–274.

[18] T. B. Nguyen and T. B. Pham, "The System of Classified and Auto-Mark Source Code for Students’ Exams," *International Journal of Information and Education Technology* vol. 6, no. 7, pp. 522-527, 2016.

[19] K. Hiroki and I. Ushio, "An Online Automated Scoring System for Java Programming Assignments," *International Journal of Information and Education Technology* vol. 6, no. 4, pp. 275-279, 2016.