ACTA: Short-Answer Grading in High-Stakes Medical Exams

King Yiu Suen[1], **Victoria Yaneva**[1] **Le An Ha**[2] **Janet Mee**[1],

Yiyun Zhou[1], **Polina Harik**[1]

[1]National Board of Medical Examiners, Philadelphia, USA
{ksuen, vyaneva, jmee, yyzhou, pharik}@nbme.org
[2]University of Wolverhamton, UK
ha.l.a@wlv.ac.uk

Abstract

This paper presents the ACTA system, which performs automated short-answer grading in the domain of high-stakes medical exams. The system builds upon previous work on neural similarity-based grading approaches by applying these to the medical domain and utilizing contrastive learning as a means to optimize the similarity metric. ACTA is evaluated against three strong baselines and is developed in alignment with operational needs, where low-confidence responses are flagged for human review. Learning curves are explored to understand the effects of training data on performance. The results demonstrate that ACTA leads to substantially lower number of responses being flagged for human review, while maintaining high classification accuracy.

1 Introduction

Automated Short Answer Grading (ASAG) has been a longstanding educational application of NLP. The task of classifying the free-text responses to short-answer questions (SAQs) as *correct* or *incorrect* is made challenging by the fact that the same concept may be expressed in a myriad of different ways. The problem has received considerable attention, with several competitions organized on the topic such as a SemEval shared task by Dzikovska et al. (2013) or the ASAP 2 Kaggle competition[1].

Most broadly, the ASAG literature defines two scoring approaches: an *instance-based approach*, where a system is trained on a portion of the data and outputs a predicted score for a given new response, and a *similarity-based approach*, where each new response assumes the label of an annotated response it is matched to using some similarity metric (Bexte et al., 2022). In early work, pre-neural similarity-based approaches were shown to lag behind the less interpretable instance-based approaches (Sakaguchi et al., 2015). Since then, neural similarity-based approaches have shown increasing promise by learning response (or question-response) embeddings and matching the pairs using cosine similarity (e.g. Schneider et al. (2022)). Bexte et al. (2022) proposed that the similarity-based approach can be further improved if the similarity metric is appropriately optimized. In their work, a pretrained Sentence-BERT model (Reimers and Gurevych, 2019) is fine-tuned on answer pairs and then a k-nearest neighbors classifier is used to match a new response based on its similarity to the labeled ones. These advances have led to a considerable improvement over the instance-based approach not only in terms of accuracy, but also in terms of interpretability and the need for less annotated data for training.

In this study, we present the ACTA system (Analysis of Clinical Text for Assessment), where we build upon the work of Bexte et al. (2022) by exploring the use of contrastive learning (Chopra et al., 2005) as a way to optimize the performance of similarity-based approaches and by applying the approach to the clinical domain. The contributions of this paper are as follows:

- Exploration of the similarity-based ASAG approach in the clinical domain, which is characterized by a number of challenging idiosyncrasies such as complex terminology, extensive use of abbreviations, misspellings, etc.
- Comparison of the results to three baselines: majority class, a similarity-based approach without finetuning, and a previous scoring system designed for the clinical domain.
- System and evaluation design constructed in alignment with operational needs, where responses that do not satisfy a given confidence threshold are flagged for human review.
- Exploration of learning curves with various training set sizes, as well as experimentation with various confidence thresholds.

[1]https://www.kaggle.com/c/asap-sas

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 443–447

2 Data

We perform experiments on two datasets containing short free-text responses to clinical test items.

Set 1 consists of SHARP items (Short Answer Rationale Provision items) – an item format where examinees see a patient chart and are asked to provide a free-text response regarding the most likely diagnosis (e.g., "plantar fasciitis", "dermatomyositis"), most appropriate next steps (e.g., "Administer corticosteroids then do arterial biopsys"), causes (e.g., "Homocysteine and MMA levels in blood"), etc.[2] A total of 44 items were administered in a pilot involving 177 4th-year US medical students. Each student saw each item, resulting in a total of 7,788 responses (of which 2,807 were unique).

Set 2 consists of short-answer questions, which present a vignette[3] describing a clinical case. Similar to Set 1, the Set 2 responses included diagnoses, causes, and treatments, among other categories of responses. These items were administered to 8,162 US medical students as part of their Internal Medicine school subject exam. There were 71 Set 2 items, where each item was seen by an average of 176 examinees (SD = 12.620), resulting in a total of 12,508 free-text responses (5,696 unique).

Responses from both sets were scored as *correct* or *incorrect* by content experts (physicians and nurse practitioners) using a scoring rubric for each item. For Set 1, two subject matter experts scored the items together as part of developing scoring guidelines for future pilots (hence agreement statistics for independent scoring cannot be reported). Another group of physicians reviewed the scores and confirmed agreement with the scoring procedure. For Set 2, four judges scored the items. Kappa coefficients (based on unique responses) for the six possible pairs of judges ranged from 0.89 to 0.92, indicating strong agreement. Scoring resulted in 5,201 correct responses (66.78%) for Set 1 and 8,086 (64.64%) for Set 2.

3 Method

We use contrastive representation learning (Chopra et al., 2005) to encode responses into embedding vectors such that responses with the same score have similar embeddings and responses with different scores have very different ones. For any given two responses, the degree to which they are matched can then be measured by the cosine similiarty between their embedding vectors. Similar to Bexte et al. (2022), we use Sentence-BERT (a.k.a. SBERT) to derive the embeddings for each response, since the model introduces a modification of the pretrained BERT network that "reduces the effort for finding the most similar pair from 65 hours with BERT / RoBERTa to about 5 seconds" (Reimers and Gurevych, 2019).

First, we pair up every response with every other response for the same item. Each pair is assigned a label of 1 if both responses have the same score (both correct or both incorrect), 0 otherwise. For each pair, the two responses are passed to SBERT independently, producing two sentence embedding vectors (one for each response).

The contrastive loss encourages the model to minimize the embedding distance when responses have the same score, and maximize the distance otherwise. To do that, the cosine similarity and the cosine distance between the sentence embedding of the first response e_1 and the sentence embedding of the second response e_2 are defined as:

$$similarity(e_1, e_2) = \frac{e_1^{\mathsf{T}} \cdot e_2}{||e_1||||e_2||}$$

$$distance(e_1, e_2) = 1 - similarity(e_1, e_2)$$

Then, the contrastive loss is defined as

$$\mathcal{L}(e_1, e_2, label) = label \cdot (distance(e_1, e_2))^2 +$$

$$(1 - label) \cdot max(0, margin - distance(e_1, e_2))^2$$

where margin is a hyperparameter, defining the lower bound distance between responses with different scores. One advantage of contrastive loss over cosine similarity loss is that it goes to 0 for negative pairs when the distance is farther than the margin. When dissimilar inputs are sufficiently distant there is no more pressure on the model to keep pushing them apart, which could allow the model to focus on improving the most erroneous cases.

During inference, the trained model is used to compute the cosine similarity between the sentence embedding of the new response and the sentence embedding of every annotation (i.e., responses of the same item in the training set). If the highest

[2]Other aspects of the SHARP item format that refer to subsequent steps for measuring clinical reasoning are not described here.

[3]See Ha et al. (2020) for a detailed description of the use of vignette-based SAQs in medicine.

		Set 1 (SHARP items)						Set 2 (SAQs)			
	Training	**20**	**40**	**60**	**80**	**120**	**142**	**20%**	**40%**	**60%**	**80%**
INCITE	**F1**	.986	.986	.989	.984	.988	.989	.88	.9	.88	.882
	Unmatched	**488**	**442**	397	354	334	318	987	830	748	711
ACTA No Finetuning	**F1**	.998	.998	1.	1.	1.	1.	.999	.999	.999	.997
	Unmatched	623	523	463	429	385	368	970	835	743	684
ACTA Finetuned	**F1**	.995	.993	.977	.979	.982	.982	.991	.991	.978	.972
	Unmatched	545	443	**201**	**123**	**47**	**44**	**734**	**497**	**274**	**172**

Table 1: Results for a similarity threshold of .95, where "F1" indicates classification performance for all matched items and "Unmatched" indicates the number of items that need to go through human scoring. For Set 1, the training data size is measured in number of examinees whose data was used for training (e.g., the first 20 examinees, the first 40, etc.). In Set 2, it is measured as percentage of the full dataset. Note that for ACTA No Finetuning, the term "training set" refers to the subset of data used to identify the most similar instances for a given new response.

cosine similarity is less than a given threshold, the new response is labeled as *unmatched* and flagged for human rater review. Otherwise, the new response assumes the score of the annotation that it has the highest cosine similarity with. For detailed training parameters, see Appenidx A.

4 Experimental setup

Baselines: We compare the approach proposed in ACTA to three baselines: **a majority class baseline** (always predicting a *correct* response); **ACTA No finetuning** – a similarity-based approach using SBERT, where the model was not trained to optimize the similarity metric. We use *all-MiniLM-L6-v2*[4], which has been pretrained on 1B sentence pairs, as our backbone model for both SBERT-no-training and SBERT. Finally, **the INCITE system** (Sarker et al., 2019), which is specifically developed to score clinical text by capturing a variety of ways clinical concepts can be expressed. INCITE is a rule-based modular pipeline utilizing custom-built lexicons, which contain observed misspellings for medical concepts and non-standard expressions, as well as common concepts and abbreviations from online resources. The tool performs direct and fuzzy matching between a new response and an annotated response (or a lexicon variant of it) using a fixed or dynamic Levenshtein ratio threshold (in our case - .95). Full details about the INCITE system are available in Sarker et al. (2019).

Learning curves: We compare the approaches by experimenting with different training set sizes and evaluating on the same test set of 20% held-out data (1,5K responses for Set 1 and 2,5K for Set 2). This provides insight on an important practical consideration - how much training data is enough to train a reliable and accurate model (Heilman and Madnani, 2015). To emulate an operational scenario, the division of training and test sets (and the increase in training data) are based on the chronological order in which the responses were received.

Evaluation metrics Another practical consideration is to directly answer two questions of operational significance: "How accurate is the system for responses that it is able to score?" and "How many responses do human raters still need to score manually?". To address these, we present two separate metrics – *F1* for matched responses and *total number of unmatched responses* – as opposed to capturing the number of unmatched responses through the measure of Recall. This setup allows the selection of more strict or liberal thresholds depending on the intended use, e.g., high-stakes summative assessment where high precision is paramount vs. formative assessment, where there can be a trade off between precision and wider response coverage.

Thresholds: A conservative similarity threshold of .95 is selected apriori to ensure high confidence that the matched responses are scored correctly. All items below that threshold are considered unmatched and are sent for human scoring. We first present detailed results for this threshold. Next, we experiment with a variety of other thresholds and compare their effect on the two evaluation metrics.

5 Results

The majority class baseline was .79 for Set 1 and .794 for Set 2. The remaining results for a threshold of .95 are presented in Table 1. As can be seen, all three systems (INCITE, ACTA No finetuning, and ACTA Finetuned) achieve very high F1 scores for the responses they were able to match for Set 1 (lowest F1 was .977 for ACTA Finetuned and .984 for INCITE). For the much larger Set 2, we see a higher F1 score range of .97 - .99 for ACTA

[4]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

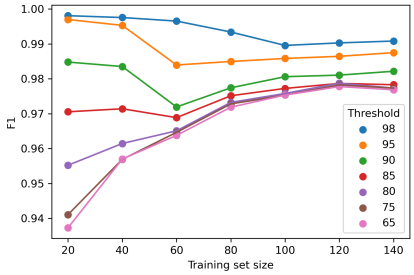

Figure 1: F1 score for Set 1 (SHARP items) as a function of similarity threshold and training set size.

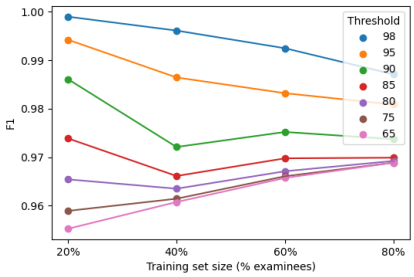

Figure 2: F1 score for Set 2 (SAQs) as a function of similarity threshold and training set size.

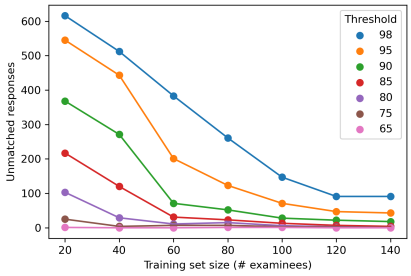

Figure 3: Number of unmatched responses for Set 1 (SHARP items) as a function of similarity threshold and training set size.

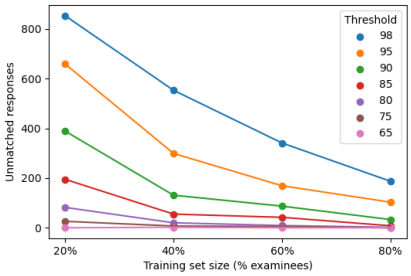

Figure 4: Number of unmatched responses for Set 2 (SAQ items) as a function of similarity threshold and training set size.

compared to .88 - .90 for INCITE. The F1 score remains high when evaluation is performed using 5-fold cross validation (not shown in the tables): the average ACTA Finetuned F1 across folds for Set 1 is .985 with an average number of unmatched responses across folds = 49.8. For Set 2 the F1 score is .98 with an average number of unmatched responses across folds = 88.8. Overall, the results suggest a consistently high level of confidence in ACTA's output for all matched responses.

When looking at the *unmatched* responses, we see dramatic differences between the three systems. When training on more than 40 examinees, *INCITE* and *ACTA No finetuning* have significantly more responses that require human review and increasing the amount of training data leads to small improvements. *ACTA Finetuned* leaves fewer unmatched responses and continuously improves with the addition of more training data. These results show the when finetuned using contrastive loss, ACTA can ultimately save more human effort than INCITE and that the gains increase with data size.

Next, we experiment with different matching thresholds by replacing the .95 value with a range of values: .98, .90, .85, .80, .75, .70, and .65. F1 remains high even with lower thresholds: For Set 1, the lowest F1 is .937 (threshold = .65 when training on data from 20 examinees). For Set 2 it is .95 for the same configuration (for detailed F1 results for each threshold, see figures 1 and 2). The number of unmatched responses, however, decreases significantly (see Figures 3 and 4) – there are either 0 or 1 unmatched responses in both sets across all training configurations for threshold .65. This shows that with more liberal thresholds, the need for human scoring almost disappears (except the need for continuous quality verification). Selecting the right trade-off between F1 and number of responses that need to undergo human review remains an operational decision.

6 Conclusion

This study showed that a similarity-based clinical ASAG system finetuned using contrastive loss outperforms the INCITE and ACTA No Finetuning baselines. Lowering the similarity threshold value significantly decreases the number of unmatched responses, while – contrary to expectation – the F1 score remains high at > .93 across conditions. The condition of weakest supervision – training on 20 examinees from Set 1 with a similarity threshold of .65 – shows that 880 annotated responses are

sufficient to score *all* 1.5K test set responses with F1 = .93. Similarly, when training on 20% of the data from Set 2 with threshold of .65, *all* 2.5K test set responses are scored with F1 = .95.

The evaluation setup allows operational experts to balance the confidence threshold with a minimum necessary F1 score, where items with more errors can have more stringent similarity thresholds and vice-versa. The threshold may also vary depending on intended use: formative exams may tolerate a lower F1 to gain wider coverage, while summative assessments may have stricter criteria.

In addition to its accuracy and wider coverage of responses, the interpretability of ACTA as a similarity-based system is an important advancement in clinical assessment compared to instance-based ASAG systems (e.g., Ha et al. (2020)). Interpretability holds special significance in the realm of automated scoring, as the value of the scores depends on the trust placed by various stakeholders (such as faculty, students, and residency selection programs, among others) in their fairness, reliability, and validity.

Like many other products, automated scoring tools are complex systems that have a significant impact not only because of their technical capabilities but also due to how they are used and the way their results are interpreted. Misusing these tools or interpreting their outputs incorrectly can lead to serious ethical issues. In a summative context, the models described in this article are intended to be used as hybrid systems, where human raters always review borderline cases. In a formative context, it is crucial to carefully examine the relationship between the use of the system and its impact on learning outcomes, as this is essential evidence for validity.

Next steps include exploration of the effects of different "gaming" strategies (e.g., intentionally providing generic instead of specific answers) and potential differential functioning across demographic groups. Notably, ACTA is intended as a hybrid system, where cases of examinees who perform near or below the passing standard are reviewed by human experts.

References

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring-how to make s-bert keep up with bert. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, NORTH TEXAS STATE UNIV DENTON.

Le Ha, Victoria Yaneva, Polina Harik, Ravi Pandian, Amy Morales, and Brian Clauser. 2020. Automated prediction of examinee proficiency from short-answer questions.

Michael Heilman and Nitin Madnani. 2015. The impact of training data on automated short answer scoring performance. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 81–85.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 1049–1054.

Abeed Sarker, Ari Z Klein, Janet Mee, Polina Harik, and Graciela Gonzalez-Hernandez. 2019. An interpretable natural language processing system for written medical examination assessment. *Journal of biomedical informatics*, 98:103268.

Johannes Schneider, Robin Richner, and Micha Riser. 2022. Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, pages 1–31.

A Appendix

batch_size = 32; log_every_n_step = 100; lr = 0.00002; margin = 0.5; max_length = 512; model_name_or_path = "sentence-transformers/all-MiniLM-L6-v2"; num_epochs = 1; num_training_participants = 142; num_workers = 8; threshold = 0.95; warmup_ratio = 0.1; weight_decay = 0.01