Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring

Tomoya Mizumoto[1,2*] **Hiroki Ouchi**[1,2] **Yoriko Isobe**[1] **Paul Reisert**[1,2]
Ryo Nagata[3,1] **Satoshi Sekine**[1] **Kentaro Inui**[2,1]
[1]RIKEN AIP, [2]Tohoku University, [3]Konan University
{tomoya.mizumoto, hiroki.ouchi, yoriko.isobe, paul.reisert, satoshi.sekine}@riken.jp,
nagata-bea2019@ml.hyogo-u.ac.jp., inui@ecei.tohoku.ac.jp

Abstract

This paper provides an analytical assessment of student short answer responses with a view to potential benefits in pedagogical contexts. We first propose and formalize two novel analytical assessment tasks: analytic score prediction and justification identification, and then provide the first dataset created for analytic short answer scoring research. Subsequently, we present a neural baseline model and report our extensive empirical results to demonstrate how our dataset can be used to explore new and intriguing technical challenges in short answer scoring. The dataset is publicly available for research purposes.

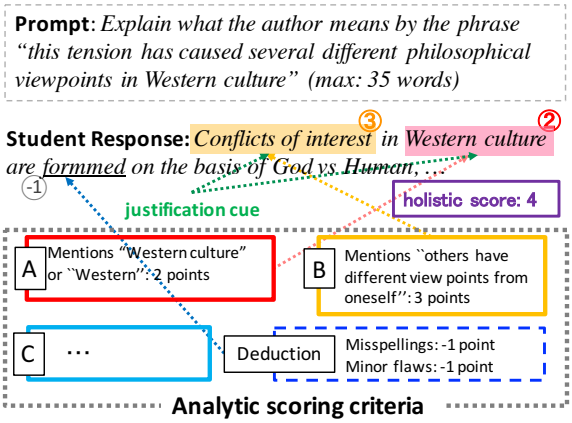

Figure 1: Example of short answer scoring with several analytic criteria.

1 Introduction

Short answer scoring (SAS) is the task of assessing short, written, free-text student responses to a given prompt. Typically, a prompt is a text which either elicits recall of information that was given in a reading passage, asks for a summary of a reading passage, or asks students to draw on knowledge they already have. The task is to assess the responses based on context and writing quality, in accordance with the criteria prespecified for each assessment by a *scoring rubric*. Automation of this process has the potential to significantly reduce the workload of human raters and has attracted a considerable amount of attention from both academia and industry (Riordan et al., 2017; Zhao et al., 2017; Sultan et al., 2016; Heilman and Madnani, 2015; Pulman and Sukkarieh, 2005; Leacock and Chodorow, 2003; Vigilante, 1999, etc.).

It should be emphasized that, in admissions tests and other tests, such as writing proficiency tests, large groups of students receive and respond to the exact same set of problems, for which rubrics have been prepared in advance. In other words, rubrics are normally available in the SAS setting as they are in preset paper assignments. Additionally, at least a small amount of training data is also available because responses are scored by human raters in any case.

This paper examines the issue of analytical assessment of short answer responses. Typically, in a short answer setting, a scoring rubric comprises multiple analytic criteria, each of which stipulates different aspects of the conditions necessary for a response to receive points, and the overall score (referred to as the *holistic score*) of a given student response is determined by some predefined function (e.g., summation) of the score gained for each analytic criterion (*analytic score*).

Consider the example illustrated in Figure 1, where a student response is assessed according to multiple analytic scoring rubrics (denoted by A, B, C, etc.). The response gains two points for analytic criterion A (denoted by the red circled "2") and three points for B (yellow circled "3"), and the holistic score is given by the total of the analytic scores ($+2$ for A, $+3$ for B, and -1 for the mis-

*Current affiliation: Future Corporation, mizumoto.tomoya.mh7@is.naist.jp

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 316–325
Florence, Italy, August 2, 2019.

spelling).

Assessing student responses by analytic scores as well as holistic scores is essential in pedagogical contexts because (i) for teachers, analytic scores are useful for a precise assessment of student proficiency, and (ii) for students, analytic scores can be used as informative feedback indicating what has been achieved and what remains to be learned next. To the best of our knowledge, however, no prior study on automatic SAS has ever addressed this task.

Motivated by this background, we propose and formalize two analytical assessment tasks of SAS, (i) *analytic score prediction* and (ii) *justification identification*. Analytic score prediction is the task of predicting the analytic score for each analytic scoring criterion, whereas justification identification is the task of identifying the *justification cue* for each analytic score. By justification cue, we refer to the segment of the response (a subsequence of words) that causes the response to be awarded points in the analytic score. In Figure 1, for example, the phrase *Western culture* is identified as a justification for criterion A, whereas the phrase *Conflicts of interest* is a justification for criterion B. Justification cues not only explain the model's prediction but also help students learn how to improve their responses.

One crucial issue in addressing such analytical assessment tasks is the lack of data. The datasets that are presently available for SAS research (Mohler et al., 2011; ASAP-SAS; Dzikovska et al., 2013; Basu et al., 2013, etc.) are all accompanied by annotations of holistic scores alone. In this study, we developed a new dataset with annotated analytic scores and justification cues as well as holistic scores. The dataset contains 2,100 sample student responses for each of six distinct reading comprehension test prompts, collected from commercial achievement tests for Japanese high school students. The dataset is publicly available for research purposes.[1]

SAS requires content-based, prompt-specific rubrics, which means that one needs to create a labeled dataset to train a model for each given prompt. This nature of the task raises the issue of how one can reduce the required labelling costs while achieving sufficient performance. This challenge is even more critical in analytical assessment because annotating student responses with analytic scores and justification cues tends to be much more costly than when only holistic scores are used. This study explores several situations with limited amounts of analytic scores and justification cues as well as large numbers of holistic scores. We show that analytical assessment performance for analytic score prediction and justification identification can be improved by compensating for a lack of data with these different types of annotations.

The contributions of this work are three-fold. First, we propose and formalize two analytical assessment tasks: analytic score prediction and justification identification. Second, we have created the first dataset for analytic SAS and released it for research. Third, we present a neural baseline model and report some of the empirical results to demonstrate how our dataset can be used to address new amd intriguing technical challenges in SAS.

2 Task

2.1 Analytic criteria

We assume that each prompt is provided with a scoring rubric which comprises several (typically two to four) analytic criteria. Each analytic criterion stipulates the conditions under which a student response will gain an analytic score, typically in the form of "if it includes the content $\langle \ldots \rangle$, the response gains x points. "

A response may lose a few points owing to misspellings or other minor flaws (referred to as *deductions*). We also regard the criteria for such deductions as special analytic scoring rubrics which are allotted negative points.

The holistic (total) score of a response is assumed to be the sum of all the item scores including the deductions.

2.2 Analytic score prediction

Analytic score prediction is the task of predicting the score of a given student response for each analytic criterion. Given a student response that consists of T words $w_{1:T} = (w_1, \cdots, w_T)$, the goal is to predict the analytic score $y^{(m)} \in \mathbb{R}$ for each criterion $m \in \mathcal{M}$, where $\mathcal{M}$ is a given set of analytic criteria.

As an evaluation metric, we use quadratic weighted kappa (QWK) (Cohen, 1968), which is commonly used in the SAS literature.

[1] https://aip-nlu.gitlab.io/resources/sas-japanese

2.3 Justification identification

Justification identification is the task of identifying a justification cue in a given student response for each analytic score. A justification cue is the segment of a response that causes that response to gain points in the analytic score. For a content-based criterion (i.e., a criterion of the form "if it includes the content $\langle \ldots \rangle$, the response gains x points"), the fragment that explicitly expresses the required content is a justification cue. Justification cues not only explain the model's prediction but also help students learn how to improve their responses.

Formally, given a student response $w_{1:T} = (w_1, \cdots, w_T)$, the goal is to identify the phrase $w_{i:j}^{(m)} = (w_i, \cdots, w_j)$, where $1 \leq i \leq j \leq T$, for each criterion m. As an evaluation metric, we use precision, recall and F1 scores based on the overlaps between gold-standard (henceforth "gold") and predicted justification cues (phrases). Consider the following example.

A carbon filament was used.
[gold]
[pred]

Here, the gold justification is *A carbon filament*, and the predicted one is *filament was*. The number of true positives (TP) is 1 (*filament*), that of false positives (FP) is 1 (*was*), and that of false negatives (FN) is 2 (*A carbon*). Thus we can calculate the precision, $1/(1+1) = 0.50$, and the recall, $1/(1+2) = 0.33$. F1 score is then $2 \times 0.50 \times 0.33/(0.50+0.33) = 0.398$.

3 Dataset

This section provides an overview of our dataset.

3.1 Original dataset

Table 1 shows the statistics of our dataset. The dataset consists of six prompts and 2,100 student responses for each prompt. Those prompts and their rubrics were collected from commercial achievement tests provided by a long-standing leading education company, where problems and rubrics are carefully generated by professional experts. All the prompts are for reading comprehension tests and are of the type that requires recall of information that has been given (either explicitly or implicitly) in a reading passage.

Responses (6 prompts $\times$ 2,100 responses) were originally annotated with holistic scores by professional raters employed by the education company (not by those employed for this research). Before the scoring, the raters were carefully instructed about the rubrics and conducted a trial annotation on the same sample response set for calibration.

3.2 Analytical assessment annotation

Each prompt in this dataset has three or four analytic criteria. The stipulation of each criterion is provided in the rubric. However, the responses in the dataset were originally annotated only with holistic scores and not with analytic scores. This is often the case in the real-world answer scoring business because (i) the manual annotation of individual analytic scores tends to be very costly, and (ii) proficient human assessors can efficiently grade a student response with a holistic score taking analytic scores into account "implicitly". Accordingly, we employed expert annotators and conducted additional annotation of all the responses with analytic scores and justification cues.

Before instructing the annotators to work on the dataset, we first investigated the difficulty of annotation. For each prompt, we randomly sampled 100 responses from the 2,100 responses and used them to train and calibrate the annotators. During this calibration process, we instructed the annotators to identify analytic scores so that, for each given student response, the sum of the analytic scores would be equal to the holistic score given in the original dataset. Then, using 100 additional exclusively sampled responses, we measured the inter-annotator agreement.

Table 2 shows the inter-annotator agreement of analytic scores for each prompt in Kappa (Cohen, 1960) and QWK. The results are reasonably high. This means that the annotation of analytic scores is not too difficult for expert human annotators. Given this observation, the remaining 1,900 responses for each prompt were annotated by a single annotator with self-double checking. To avoid inconsistency across annotators, we assigned all 1,900 responses to each prompt to the same annotator. Furthermore, if an annotator was not confident about scoring a given response, the annotator was instructed to discuss the response with person who prepared the the exam to reach a consensus. As a result, we obtained 12,600 student responses (6 prompts $\times$ 2,100 responses) with ana-

	Q1	Q2	Q3	Q4	Q5	Q6
Max holistic score	16	12	12	15	15	14
Average holistic score	6.8	4.0	5.3	5.5	4.6	5.5
Standard deviation	3.5	1.8	2.1	2.7	2.6	3.1
# analytic criteria	4	4	4	3	3	3
length (char.) limit	70	50	60	70	70	60
Average length (char.)	62.86	45.15	54.13	65.53	64.83	55.44

Table 1: Statistics of our dataset.

	Q1	Q2	Q3	Q4	Q5	Q6	Ave.
Kappa	.93	.92	.79	.70	.83	.82	.84
QWK	.96	.94	.76	.84	.82	.90	.87

Table 2: Inter-annotator agreement of analytic scores in Kappa (Cohen, 1960) and Quadratic Weighted Kappa (QWK) (Cohen, 1968). The scores are calculated by averaging the agreement scores for each analytic criterion.

lytic scores and justification cues for each prompt.

In the future, we intend to extend the dataset by adding a wider variety of prompts. In fact, we have already started the annotation for three additional prompts and plan to extend the dataset to a far larger scale. However, our current dataset is already as large as the biggest existing datasets available for SAS research (ASAP-SAS), and furthermore, no prior dataset has been annotated with analytical assessment.

4 A Neural Baseline Model

The goal of the rest of the paper is to demonstrate how our dataset can be used to address intriguing but as yet unexplored challenges in analytic SAS. To this end, we first present our neural network baseline model in this section and then report some of the experimental results withwe have obtained using the model in the next section.

4.1 Overall architecture

Figure 2 illustrates the overall architecture of our baseline model. The idea is three-fold: (i) build a distinct model of analytic score prediction for each analytic criterion based on Riordan et al. (2017)'s model for holistic SAS, (ii) train the analytic score prediction models jointly with the holistic score prediction model, and (iii) use supervised attention for justification identification.

The model includes $|\mathcal{M}|$ analytic score models and an addition layer. First, the input student response $w_{1:T} = (w_1, w_2, \cdots, w_T)$ is mapped to word embeddings. Second, these embeddings are

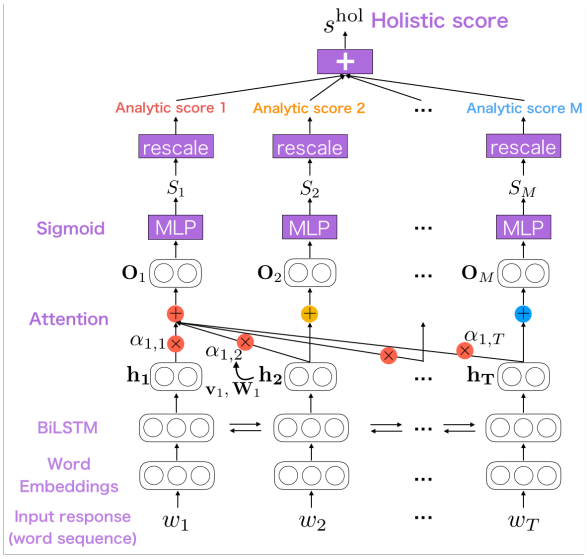

Figure 2: Overview of the baseline model for analytic short answer scoring.

fed to the BiLSTM layer. Third, through an attention mechanism associated with each analytic criterion $m \in \mathcal{M}$, an analytic scoring model outputs the analytic score s_m. Finally, the addition layer sums up the analytic scores to calculate the holistic score s_{hol},

Formally, the holistic score s_{hol} is calculated by summing all the analytic scores $\{s_m \mid m \in \mathcal{M}\}$.

$$s_{\text{hol}} = \max\Big(\sum_{m \in \mathcal{M}} \text{rescale}(s_m), 0\Big) \ , \quad (1)$$

$$s_m = f_m(w_{1:T}) \ . \quad (2)$$

Here, we use $\max(\cdot, 0)$ to prevent negative scoring in the event that no scoring rubric criteria are met, misspellings, and other minor flaws. The function "rescale$(\cdot)$" scales the analytic score back to the original score range. As Equation 3 in Section 4.2 shows, we use the sigmoid function to compute each analytic score. This means that each analytic score takes a value from 0 to 1, i.e., $s_m \in [0, 1]$. We thus re-scale the 0-1 ranged score to the original scaled score. Consider a case in which the analytic scoring model outputs $s_m = 0.7$ for an analytic criterion assigned 3 points. The rescale func-

tion multiplies 3 by the score $s_m = 0.7$, and the resulting score is 2.1. This score of 2.1 is then rounded off, and 2 is summed into the holistic score.

One advantage of this architecture is that the connection between the holistic and analytic scoring models enables the loss of the holistic score to back-propagate to the analytic scoring models. This means that without analytic score annotations, each analytic scoring model can still be trained with holistic score signals.

4.2 Analytic scoring model

Each analytic scoring model f_m in Equation 2 is defined as follows:

$$f_m(w) = \text{sigmoid}(\mathbf{w}_m \cdot \mathbf{o}_m + b_m) \ , \quad (3)$$

, where $\mathbf{w}_m$ is a parameter vector and b_m is a bias value. An attention vector $\mathbf{o}_m$ is calculated by an attention mechanism, i.e., $\mathbf{o}_m = f_m^{\text{att}}(\mathbf{h}_{1:T})$, where a sequence of the hidden states $\mathbf{h}_{1:T} = (\mathbf{h}_1, \cdots, \mathbf{h}_T)$ is output by a BiLSTM layer.

As mentioned above, owing to the use of the sigmoid function, each analytic score takes a value from 0 to 1, i.e., $s_m \in [0, 1]$. In the training phase, we also scale gold analytic scores to match the scale. In the evaluation phase, the predicted scores are re-scaled back to their original range.

4.3 Attention mechanism

An attention mechanism f_m^{att} is defined as follows:

$$f_m^{\text{att}}(\mathbf{h}_{1:T}) = \sum_{t=1}^{T} \alpha_{m,t} \mathbf{h}_t \quad (4)$$

An attention value $\alpha_{m,t}$ denotes the importance weight, which represents relative importance of the t-th word for predicting analytic score s_m.

4.4 Justification identification method

The attention mechanism is used not only for analytic score prediction but also for justification identification. Specifically, based on the attention scores α, we extract a set of justification cues $\mathcal{C}$.

$$\alpha_{\text{max}} = \max_{t=1,\cdots,T} \alpha_t \ ,$$
$$\mathcal{C} = \{t \in [1, T] \mid \alpha_{\text{max}} - \alpha_t < \beta\} \ .$$

Here, we first calculate the maximum attention score α_{max} among all the attention scores. We then extract the word indices t if the difference between the maximum score α_{max} and its score α_t is less than the threshold β. As a result, we can obtain a set of justification cues $\mathcal{C}$. The threshold β is a hyperparameter, which is selected by using the development set.

4.5 Training

Training with analytic scores. To train each analytic scoring model, we minimize the mean squared error (MSE) as the loss function,

$$\frac{1}{N} \sum_{n=1}^{N} \sum_{i \in \mathcal{I}^{(n)}} (s_i^{(n)} - \hat{s}_i^{(n)})^2 \ , \quad (5)$$

where N is the number of training instances, and $s_i^{(n)}$ and $\hat{s}_i^{(n)}$ are the predicted score and gold score, respectively.

Training with holistic scores. To train the whole network on holistic score annotations, we minimize the MSE calculated with gold and predicted holistic scores (Equation 1) as follows:

$$\frac{1}{N} \sum_{n=1}^{N} (s_{\text{hol}}^{(n)} - \hat{s}_{\text{hol}}^{(n)})^2 \ , \quad (6)$$

where N is the number of training instances, and $s_{\text{hol}}^{(n)}$ and $\hat{s}_{\text{hol}}^{(n)}$ are the predicted score and gold score, respectively.

Supervised attention. We further train the attention mechanism for each criterion in a supervised manner, called supervised attention (Mi et al., 2016; Liu et al., 2016; Kamigaito et al., 2017). In supervised attention, attention is learned from the difference between the span where the attention is focused and the given gold signal of a justification cue. Following a previous study by Liu et al. (2016), we add a soft constraint method to obtain the following objective function:

$$\sum_{i \in \mathcal{I}} \{ \frac{1}{N} \sum_{n=1}^{N} (s_i^{(n)} - \hat{s}_i^{(n)})^2 + \frac{\lambda}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} (\alpha_{i,t}^{(n)} - \hat{\alpha}_{i,t}^{(n)})^2 \} \quad (7)$$

where $\alpha_{i,t}^{(n)}$ denotes an attention weight, $\hat{\alpha}_{i,t}$ is the supervision of attention that corresponds to the justification cue annotated by human assessors, and $\lambda > 0$ is a hyper-parameter. If the t-th word is part of a gold justification cue (e.g., the phrase

"Western culture" in Figure 1), $\hat{\alpha}_{i,t}^{(n)}$ is 1, otherwise it is 0.

If an analytic score is zero, all the attention weights $\{\hat{\alpha}_{i,t}^{(n)}\}_{t=1}^{T}$ take zero values. To solve this problem, we explicitly encode the information that there is no justification cue by appending a dummy token to an input sequence. Specifically, we add $\hat{\alpha}_{i,T+1}$ to $\{\hat{\alpha}_{i,t}\}_{t=1}^{T}$ and set its value to 1 if an analytic score is zero and to 0 otherwise.

5 Experiments

5.1 Settings

Dataset We first split our dataset into three subsets for each prompt: 1,600 responses for training, 250 responses for development, and 250 responses for testing. To tokenize the response texts, we employed an off-the-shelf morphological analyzer, MeCab 0.98 (Kudo et al., 2004), with default settings.

Implementation We implemented the neural baseline model with Keras and TensorFlow. The code will be made publicly available at *an anonymous URL* once the paper is accepted. We chose the same hyperparameters and training settings as in Riordan et al. (2017)'s holistic scoring model.

SVR Baseline We also implemented another simpler baseline model based on the support vector regression model (SVR) following Sakaguchi et al. (2015) to provide sparse feature-based baseline results. We adopted the feature set proposed by Sakaguchi et al. (2015), which includes word 1-gram, word 2-gram, and predicate-argument structure features[2]. We used KNP 4.16 (Kawahara and Kurohashi, 2006) to extract Japanese predicate-argument structure features.

5.2 Experimental scenarios

As argued in Sections 1 and 3.2, one crucial issue in analytic SAS is that the annotation of analytic scores and justification cues is far more expensive than holistic score annotation. One of our primary concerns, therefore, is finding ways to reduce the required labeling costs while achieving sufficient performance. To explore this issue, we consider three experimental scenarios:

[2]We excluded response length and character n-gram features because the performance was worse on the development set.

	Q1	Q2	Q3	Q4	Q5	Q6	Ave.
Analytic/Justification: 25							
SVR	.55	.60	.20	.54	.58	.45	.486
NN base	.60	.62	.19	.58	.64	.47	.516
+just.	.74	.73	.29	.64	.74	.53	.610
+hol.	.94	.84	.48	.72	.86	.75	.764
Analytic/Justification: 50							
SVR	.69	.73	.29	.64	.68	.56	.596
NN base	.77	.78	.29	.68	.72	.59	.638
+just.	.83	.85	.38	.71	.78	.64	.700
+hol.	.95	.93	.59	.71	.87	.79	.806
Analytic/Justification: 100							
SVR	.77	.80	.35	.72	.73	.66	.670
NN base	.87	.84	.40	.74	.79	.67	.719
+just.	.90	.88	.52	.76	.81	.72	.767
+hol.	.96	.93	.67	.81	.87	.82	.844
Analytic/Justification: 200							
SVR	.85	.87	.44	.77	.78	.71	.735
NN base	.92	.91	.57	.78	.83	.76	.794
+just.	.95	.92	.65	.80	.84	.78	.822
+hol.	.97	.94	.72	.82	.88	.83	.859
human	.96	.94	.76	.84	.82	.90	.873

Table 3: Performance in QWK for analytic score prediction. "SVR" denotes the SVR baseline model described in Section 5.1. "NN base", "+just. ", and "+hol. " denote the models trained in the three hypothetical situations, Situations (i) to (iii), described in Section 5.2., respectively.

Scenario (i): Basic setting (analytic score signals only) The first scenario assumes that we only have analytic scores annotated to a small set of responses. Thus we can train a model on these annotations for each task. We consider this scenario as our baseline scenario. We refer to the model for this scenario as "NN base."

Scenario (ii): (i) + justification signals In addition to the analytic score annotations, the second scenario assumes that we have justification cues annotated to the same set of responses. We can thus train a model on both the analytic score and justification annotations.

Scenario (iii): (ii) + holistic score signals In addition to the analytic scores and justification cues, the third scenario assumes that we have holistic scores annotated to a relatively large set of responses. In addition to implementing supervised learning, we can train models in a weakly supervised manner using holistic scores.

All the reported results are the average of ten distinct trials with the use of ten different random seeds.

5.3 Analytic score prediction

Scenario (i) Table 3 shows the results of each model. Here we vary the numbers of analytic scores and justification cues used for training each model. "Analytic/Justification: N" denotes that we used $N \in \{25, 50, 100, 200\}$ analytic scores and justification cues, respectively.[3] In all the settings, the base analytic scoring model (NN base) consistently outperformed the SVR. Also, compared with human performance, the analytic scoring models yields reasonably strong results.

Scenario (ii) Here, we are interested in the effects of gold justification signals on analytic score prediction. In Table 3, "+just." denotes the models trained on N analytic scores and the same number of justification signals. Comparing the base model (NN base) with the justification-added model (+just.), we observed that gold justification signals consistently improved the base model in all the settings. This result reveals that gold justification signals are useful for analytic score prediction.

Scenario (iii) Another issue is the effects of holistic score signals on analytic score prediction. In Table 3, "+hol." denotes the models trained on N analytic score signals, N justification signals, and 1,600 holistic scores signals. Comparing the justification-added model (+just) with the holistic-score-added model (+hol.), we observed that extra holistic score signals contributed to further performance improvement. This result suggests that holistic score signals are useful for analytic score prediction.

Summary These results suggest that our scenarios (ii) and (iii) are both worth considering in order to improve the performance of analytic score prediction. Note that the gains achieved by incorporating scenarios (ii) and (iii) are both statistically significant ($p < 0.01$ by a paired bootstrap test (Koehn, 2004)). Specifically, the performance of the "+just." model was significantly better than that of the "NN base" model for all the prompts. The performance of the "+hol." model was also significantly better than that of the "+just." model for all the prompts.

	Prec.	Rec.	F1
NN base (100)	.332	.491	.349
+just. (100)	.837	.703	.758
+hol.	.807	.692	.738

Table 4: Performance of justification identification.

5.4 Justification identification

Scenario (i) Table 4 shows the results for justification identification. The "NN base" model is trained on analytic scores of 100 responses. This means that we used no justification signals for training. Nevertheless, the model was able to identify some phrases that appeared in the training responses frequently and that were strongly associated with analytic scores (e.g., the phrase "Western culture" in Figure 1). This result suggests that, although this model's performance was not very strong, some useful information relevant to justification identification can be exploited from the analytic score signals alone.

Scenario (ii) In Table 4, "+just." denotes the model trained on analytic scores as well as the justification cues of 100 responses. Naturally, the model's performance was drastically improved when we fed it the gold justification signals (0.349 to 0.758 in F1).

Scenario (iii) In Table 4, "+hol." denotes the model trained on 100 analytic score signals, 100 justification signals, and 1,600 holistic score signals. Interestingly, the model's performance was not improved by the incorporation of the extra holistic score signals (0.758 vs. 0.738 in F1). This is in contrast to the case of analytic score prediction task, which was improved by the extra holistic score signals. A more in-depth analysis of this matter is needed, but our findings do raise the non-trivial question of which architecture is optimal to maximize the gain that results from including justification identification from holistic score signals.

Additional analysis Another interesting question deals with how well the accuracy of analytic score prediction correlates with the accuracy of justification identification. We observed that the neural baseline models showed strong performance for justification identification. These results raise the simple question of whether the sys-

[3] Since our dataset is entirely annotated with analytic scores, one could conduct experiments with more training signals.

tem is able to correctly predict the analytic scores for each response with the same high performance seen in justification identification. To answer this question, we created two subgroups from among the responses to Q3[4]: (i) responses with higher precision ($>$ .70) and (ii) those with lower precision ($<$ .50) on the justification identification task. We then calculated the QWK for each of these groups. We obtained QWK values of 0.835 and 0.182 (averaged across all the criteria) for responses with higher and lower precision, respectively. This strong correlation between analytic scoring and justification empirically indicates the feasibility of simultaneously pursuing the two analytical assessment tasks because one benefits from the other.

5.5 Holistic score prediction

Our dataset can, of course, be used to conduct experiments on holistic SAS as well. One unique advantage of our dataset is that it contains analytic scores and justification cues, and thus one can draw more profound insights using these new types of annotations. For example, we can investigate the effects of analytic score signals on holistic score prediction.

Table 5 shows the results for holistic score prediction. The first thing to note here is the comparison between the SVR model and the "hol." model trained on only the holistic score signals. We can observe that the "SVR" model consistently outperformed the "hol." model, that the difference in their performance was smaller with a larger training set, and that the two models have nearly comparable QWK (0.848 vs. 0.844) for $n = 1600$. The second issue is the comparison between the "hol." model and the "analytic" model trained on only the analytic score signals. In all the settings, the "analytic" model considerably outperformed the "hol." model. This indicates that analytic score signals are very informative for training a holistic score prediction model as well. The third issue is the comparison between the "NN base" model and the "+just." model trained on both the analytic score and justification signals. We can observe that using justification signals as well as analytic score signals for training further boosts the performance at holistic score prediction, particularly when the training set is smaller.

[4]To simplify the analysis, we selected Q3, which exhibited the lowest performance.

n	100	200	400	800	1600
SVR (n)	.724	.772	.810	.832	.848
hol. (n)	.671	.733	.782	.815	.844
NN base (n)	.738	.803	.841	.869	.891
+just. (n)	.776	.827	.856	.876	.892

Table 5: The performances of holistic score prediction. n denotes the number of training instances (responses). "hol. (n)" denotes the model trained with n holistic score signals only. "NN base (n)" denotes the model trained with the analytic score signals of n responses. "+just. (n)" denotes the model trained with both analytic scores and justification signals of n responses.

Summary These results imply that, when only a limited number of responses is available for training a holistic scoring model, it may well be worth annotating them with analytic scores and justification cues as well as with holistic scores. Note that this findings regarding the correlation between holistic and analytic score predictions has never previously been reported in the context of SAS. Our dataset containing analytic score and justification annotations opens up several potential directions of research in the field of SAS.

6 Related Work

Short answer scoring Previous research on SAS has solely focused on holistic score prediction. We believe that this is partly because, to date, the publicly available datasets for SAS have contained holistic scores only (Mohler et al., 2011; Dzikovska et al., 2012, 2013; ASAP-SAS) . To the best of our knowledge, our dataset is the first to provide both annotated analytic scores and their justification cues.

Analytical assessment Analytical assessment has been studied in the context of automated essay scoring (Persing and Ng, 2016, 2015, etc.). The analytic criteria adopted in essay scoring tend to be more general, e.g., organization, clarity, and argument strength. In contrast, analytic criteria in SAS are typically *prompt-specific* as in our examples in Figure 1. Thus, the analytic criteria need to be learned by the model separately for each individual prompt. It is an interesting open question whether the insights gained from essay scoring research can be applicable to analytic SAS research.

Interpretability of neural models In recent years, the interpretability of neural models has received widespread attention. Some research on in-

terpretability has been conducted in the image processing field (Bach et al., 2015; Shrikumar et al., 2017). In NLP, researchers have attempted to interpret the model by analyzing the focus of attention of neural networks (Ghader and Monz, 2017; Vinyals et al., 2015). In these previous studies, however, the attention was qualitatively rather than quantitatively analyzed. In contrast, we quantitatively evaluated the justifications by examining the extent to which justification cues correspond to the span on which the system focuses to predict the analytic score. To the best of our knowledge, this is the first evaluation of the performance of justifications (i.e., interpretability) in SAS.

7 Conclusion

In this paper, we have examined analytical assessment for SAS. We proposed and formalized two analytic tasks: (i) analytic score prediction and (ii) justification identification. For these tasks, we developed a new dataset with analytic score and justification cue annotations. We then designed a neural model that predicts analytic scores simultaneously with a holistic score and trained the model with only a small number of analytic score signals and a larger number of holistic score signals. Through our extensive experiments, we have provided intriguing research scenarios and questions on the correlations between analytic and holistic scores.

One interesting line of future research is the possibility of developing datasets in other languages. It is worth examining scoring models in multilingual settings, although we plan to start by creating and releasing an English-language dataset. Another line of future research could include the development of more sophisticated models. In this paper, analytic scoring models calculate scores independently, yet there are some interdependencies between analytic score criteria. Accordingly, we plan to develop a model that incorporates this interdependency.

Acknowledgements

We thank Tomoya Okubo for helping to obtain the data for Japanese short answer scoring and Takamiya Gakuen Yoyogi Seminar for providing the data.

References

ASAP-SAS. 2012. Scoring short answer essays. ASAP short answer scoring competition system description.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one*, 10(7).

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological bulletin*, 70(4):213–220.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 200–210.

Myroslava O. Dzikovska, Rodney D. Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval)*, pages 263–274.

Hamidreza Ghader and Christof Monz. 2017. What does Attention in Neural Machine Translation Pay Attention to? In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 30–39.

Michael Heilman and Nitin Madnani. 2015. The Impact of Training Data on Automated Short Answer Scoring Performance. In *Proceedings of the 10th Workshop on Building Educational Applications Using NLP (BEA)*, pages 81–85.

Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2017. Supervised Attention for Sequence-to-Sequence Constituency Parsing. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 7–12.

Daisuke Kawahara and Sadao Kurohashi. 2006. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. In *Proceedings of Human Language Technologies: The 2006 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 176–183.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 230–237.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated Scoring of Short-Answer Questions. *Computer and the Humanities*, 37:389–405.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural Machine Translation with Supervised Attention. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 3093–3102.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised Attentions for Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2283–2288.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 752–762.

Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 543–552.

Isaac Persing and Vincent Ng. 2016. Modeling Stance in Student Essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2174–2184.

Stephen G. Pulman and Jana Z. Sukkarieh. 2005. Automatic Short Answer Marking. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP (BEA)*, pages 9–16.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating Neural Architectures for Short Answer Scoring. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP (BEA)*, pages 159–168.

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 (NAACL-HLT)*, pages 1049–1054.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the Thirty-fourth International Conference on Machine Learning (ICML)*, pages 3145–3153.

Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and Easy Short Answer Grading with High Accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1070–1075.

Richard Vigilante. 1999. Online Computer Scoring of Constructed-response Questions. *Journal of information technology impact*, 1(2):57–62.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar As a Foreign Language. In *Proceedings of Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2773–2781.

Siyuan Zhao, Yaqiong Zhang, Xiaolu Xiong, Anthony Botelho, and Neil T. Heffernan. 2017. A Memory-Augmented Neural Model for Automated Grading. In *Proceedings of Fourth ACM Conference on Learning @ Scale*, pages 189–192.