国际人工智能会议
AAAI 2021 论文北京预讲会

Automated Cross-prompt Scoring of Essay Traits

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, Jiajun Chen

Introduction

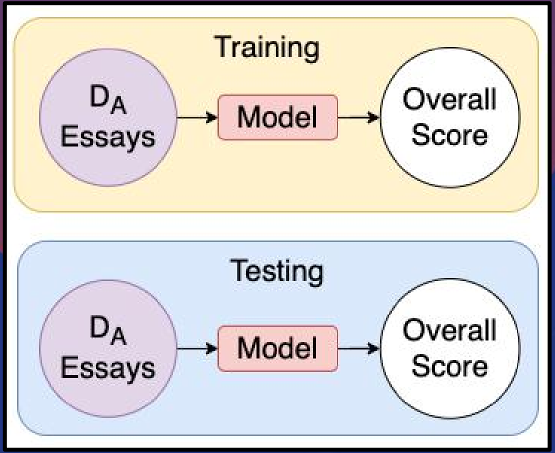
What is Automated Essay Scoring?

Automated Essay Scoring is the task of using computation to assess the quality of a written composition.

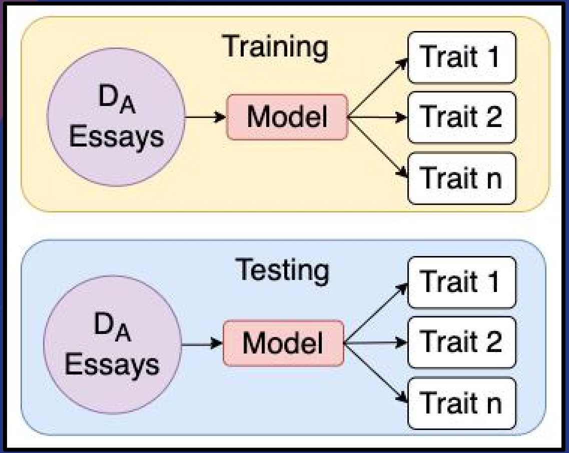
Why is Automated Essay Scoring important?

- Manually grading essays is time consuming and expensive
- Students can obtain instant feedback
- Teacher bias towards students can be mitigated

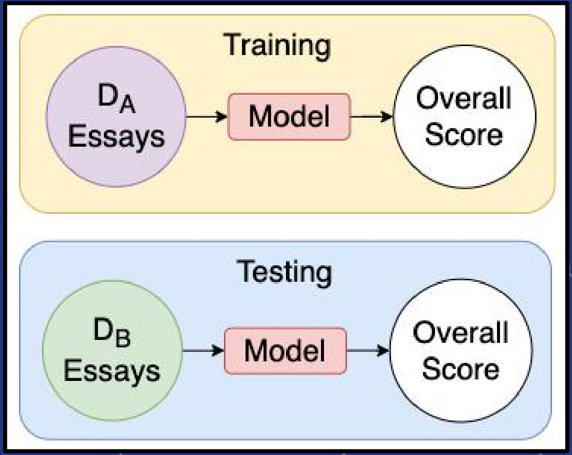
Introduction – Current Research



1. Prompt-specific Holistic Essay Scoring



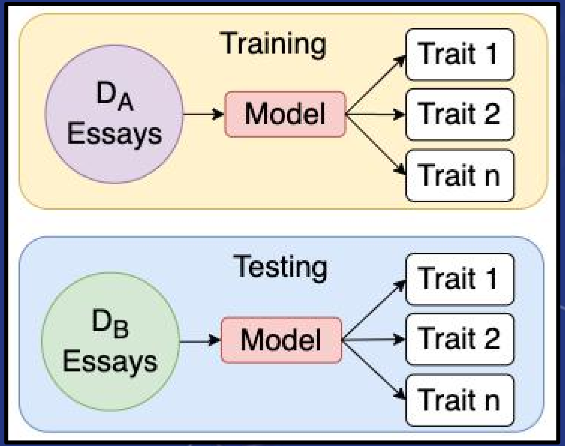
2. Prompt-specific Essay Trait Scoring



3. Cross-prompt Holistic Essay Scoring

Motivation

- Obtaining pre-graded essays for the target prompt is expensive and often unrealistic.
- Overall score is insufficient – in order to improve their writing, students require feedback regarding different aspects of their writing.
- For real-world applications, being able to perform well in cross-prompt setting and being able to provide feedback for multiple aspects of writing are both vital capabilities.



Cross-prompt Essay Trait Scoring

Challenges

- Partial trait coverage: Each essay set has its own set of relevant traits. Leads to low-resource situation for certain traits if they are present in only a few essay sets.
- Inter-trait relatedness: Certain traits are highly related to other traits. E.g. if an essay performs well for the word choice trait, it is likely that it will also possess good quality regarding its use of conventions.

Approach

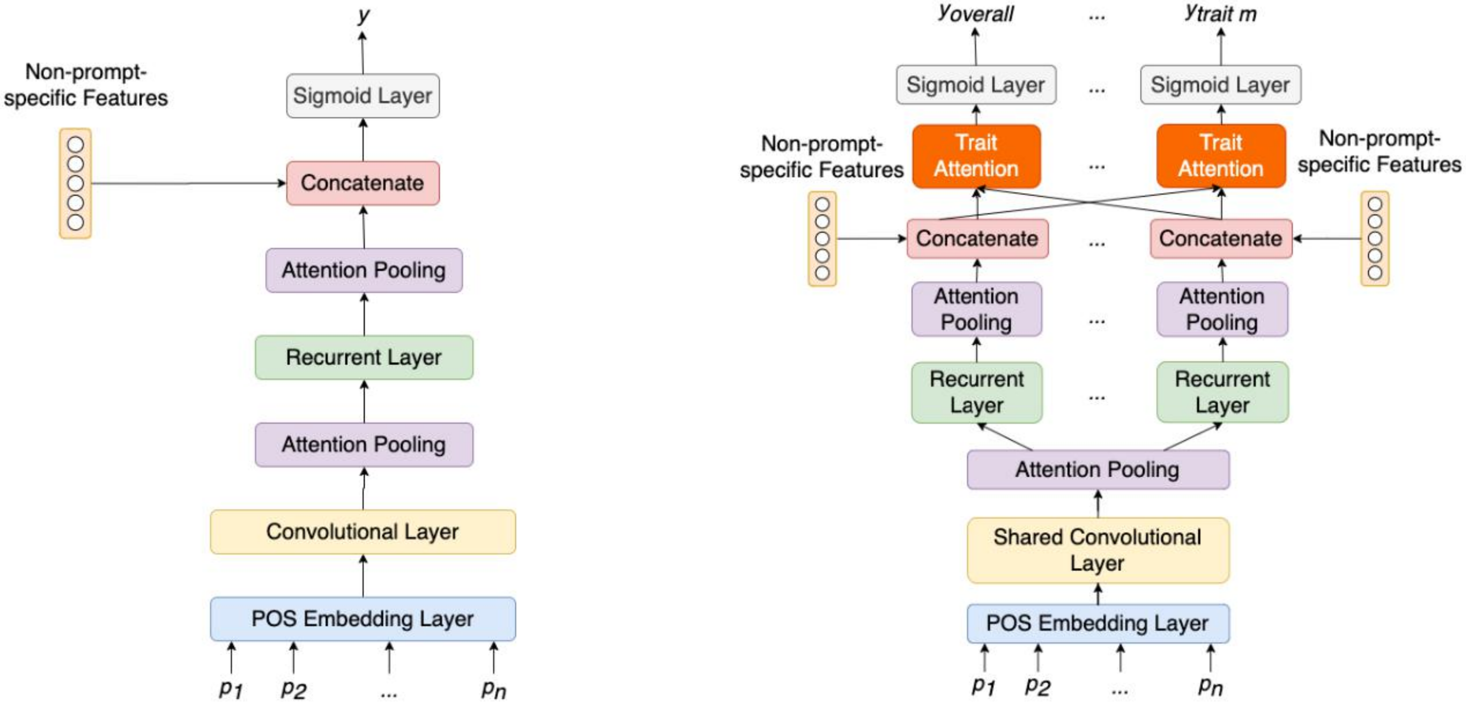
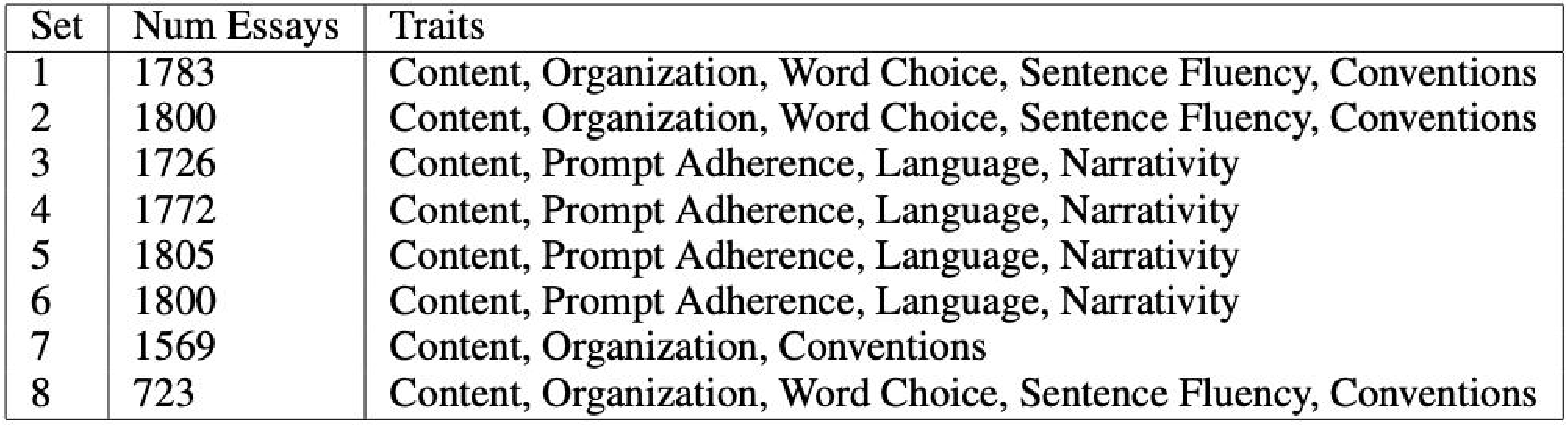

y
Non-prompt-specific Features
Sigmoid Layer
Concatenate
Attention Pooling
Recurrent Layer
Attention Pooling
Convolutional Layer
POS Embedding Layer
p_1 p_2 ... p_n
$y_{overall}$ ... $y_{trait\ m}$
Sigmoid Layer ... Sigmoid Layer
Trait Attention ... Trait Attention
Non-prompt-specific Features
Non-prompt-specific Features
Concatenate ... Concatenate
Attention Pooling ... Attention Pooling
Recurrent Layer ... Recurrent Layer
Attention Pooling
Shared Convolutional Layer
POS Embedding Layer
p_1 p_2 ... p_n

Figure 2: Architecture of both PAES (Ridley et al. 2020) on the left and our proposed CTS model (right)

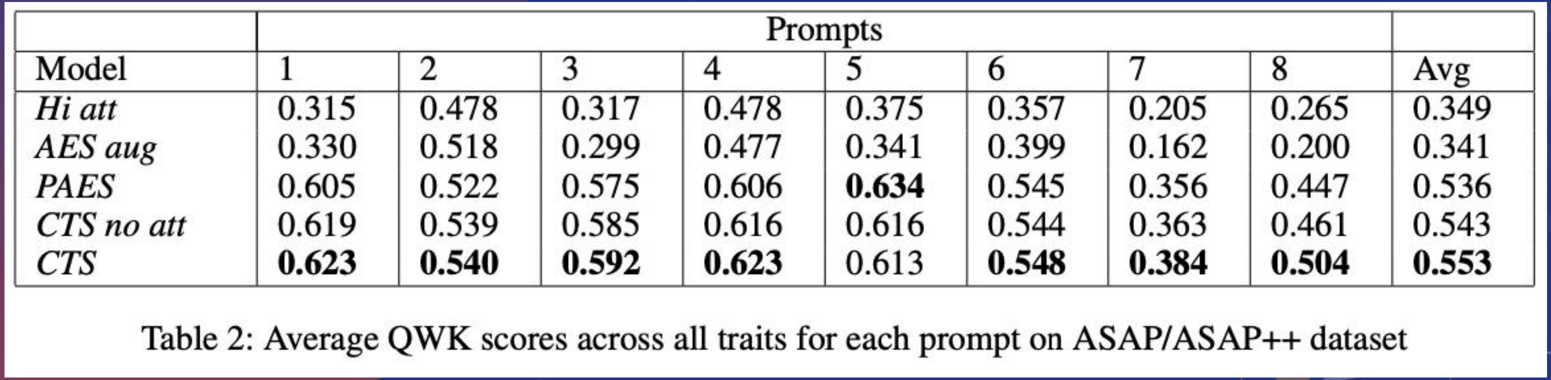
Experiments



Set	Num Essays	Traits
1	1783	Content, Organization, Word Choice, Sentence Fluency, Conventions
2	1800	Content, Organization, Word Choice, Sentence Fluency, Conventions
3	1726	Content, Prompt Adherence, Language, Narrativity
4	1772	Content, Prompt Adherence, Language, Narrativity
5	1805	Content, Prompt Adherence, Language, Narrativity
6	1800	Content, Prompt Adherence, Language, Narrativity
7	1569	Content, Organization, Conventions
8	723	Content, Organization, Word Choice, Sentence Fluency, Conventions

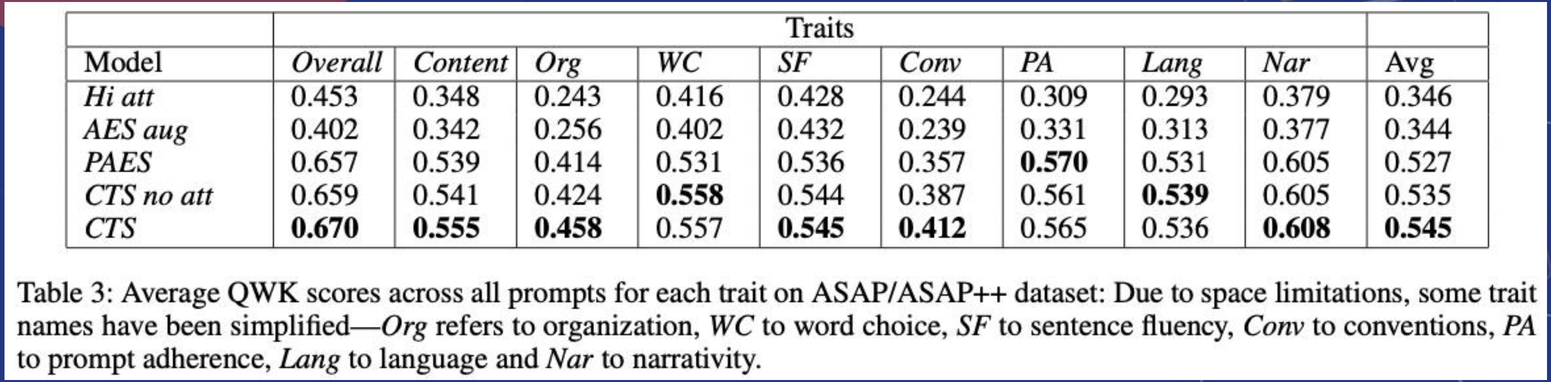
Table 1: ASAP and ASAP++ dataset traits

Experiments



	Prompts								
Model	1	2	3	4	5	6	7	8	Avg
Hi att	0.315	0.478	0.317	0.478	0.375	0.357	0.205	0.265	0.349
AES aug	0.330	0.518	0.299	0.477	0.341	0.399	0.162	0.200	0.341
PAES	0.605	0.522	0.575	0.606	**0.634**	0.545	0.356	0.447	0.536
CTS no att	0.619	0.539	0.585	0.616	0.616	0.544	0.363	0.461	0.543
CTS	**0.623**	**0.540**	**0.592**	**0.623**	0.613	**0.548**	**0.384**	**0.504**	**0.553**

Table 2: Average QWK scores across all traits for each prompt on ASAP/ASAP++ dataset



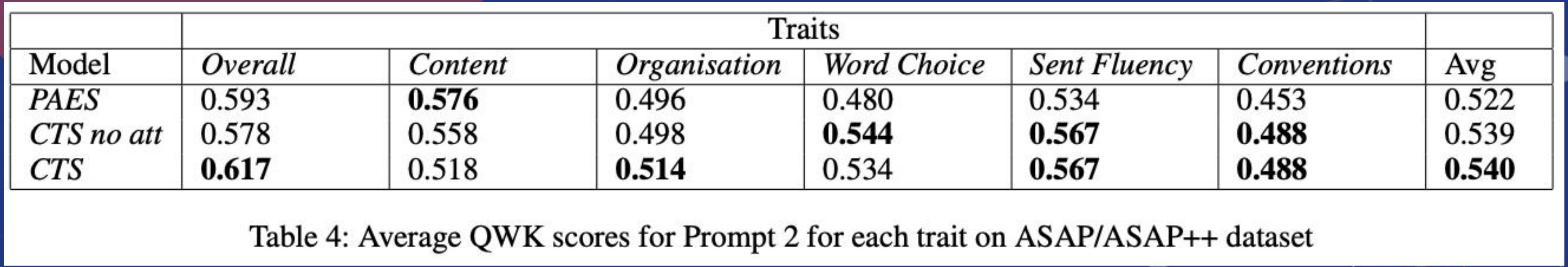
	Traits									
Model	*Overall*	*Content*	*Org*	*WC*	*SF*	*Conv*	*PA*	*Lang*	*Nar*	Avg
Hi att	0.453	0.348	0.243	0.416	0.428	0.244	0.309	0.293	0.379	0.346
AES aug	0.402	0.342	0.256	0.402	0.432	0.239	0.331	0.313	0.377	0.344
PAES	0.657	0.539	0.414	0.531	0.536	0.357	**0.570**	0.531	0.605	0.527
CTS no att	0.659	0.541	0.424	**0.558**	0.544	0.387	0.561	**0.539**	0.605	0.535
CTS	**0.670**	**0.555**	**0.458**	0.557	**0.545**	**0.412**	0.565	0.536	**0.608**	**0.545**

Table 3: Average QWK scores across all prompts for each trait on ASAP/ASAP++ dataset: Due to space limitations, some trait names have been simplified—*Org* refers to organization, *WC* to word choice, *SF* to sentence fluency, *Conv* to conventions, *PA* to prompt adherence, *Lang* to language and *Nar* to narrativity.

Experiments

Effect of Trait Sample Size

- Word Choice and Sentence Fluency only present in two other prompts.
- They are therefore underrepresented in training data.



	Traits						
Model	*Overall*	*Content*	*Organisation*	*Word Choice*	*Sent Fluency*	*Conventions*	Avg
PAES	0.593	**0.576**	0.496	0.480	0.534	0.453	0.522
CTS no att	0.578	0.558	0.498	**0.544**	**0.567**	**0.488**	0.539
CTS	**0.617**	0.518	**0.514**	0.534	**0.567**	**0.488**	**0.540**

Table 4: Average QWK scores for Prompt 2 for each trait on ASAP/ASAP++ dataset

Experiments

Effect of Trait-Attention

- Similar attention weights observed when predicting overall score.
- Higher attention weights observed for relevant traits when predicting specific traits.

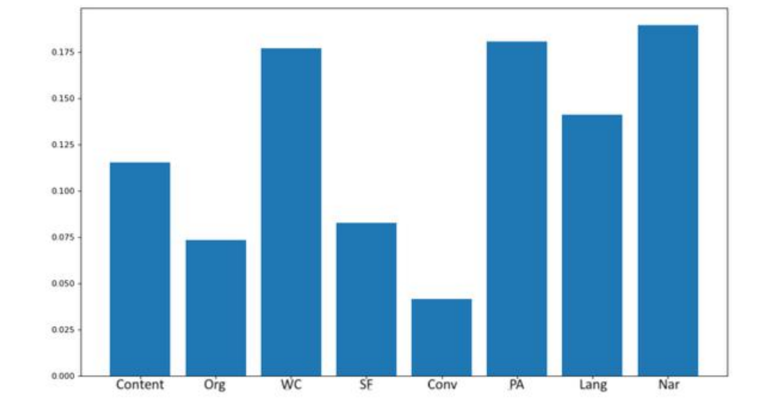


Figure 3: Attention weights for all traits when predicting *overall* score for Prompt 3

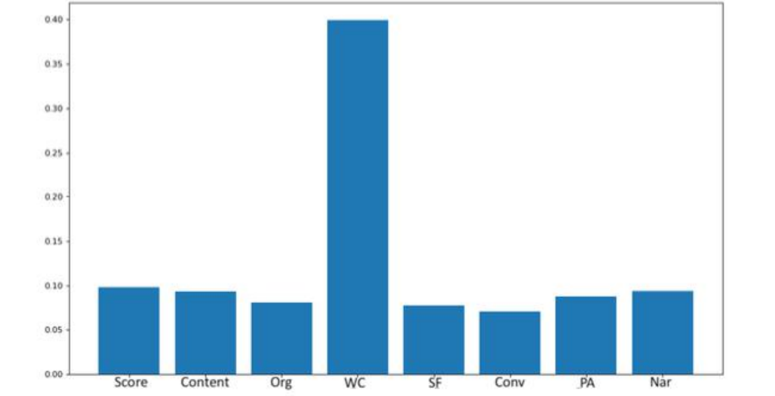


Figure 4: Attention weights for all traits when predicting the *language* score for Prompt 3

Conclusions

- We introduce a new task Automated Cross-prompt Scoring of Essay Traits to integrate two vital components of effective real-world AES systems.
- We devise a multi-task approach to mediate the issue of limited training caused by partial trait coverage.
- We make explicit use of inter-trait relationships through the use of a trait-attention mechanism.

THANKS

2020.12.19