Aggregating Multiple Heuristic Signals as Supervision for Unsupervised Automated Essay Scoring

Cong Wang, Zhiwei Jiang, Yafeng Yin, Zifeng Cheng, Shiping Ge, Qing Gu
State Key Laboratory for Novel Software Technology, Nanjing University, China

Unsupervised Automated Essay Scoring

- Automated Essay Scoring (AES) aims to score writing quality of essays without human intervention.
- SOTA AES models are trained in a supervised way with large labeled corpora, comprising essays and their groundtruth quality scores.
- Collecting labeled essays is time-consuming and labor-intensive.

- **Unsupervised AES** does not require groundtruth scores for training, and has potential in scientific research and practical applications.

Motivation

- Chen et al. use ***number of unique term*** as initial score, and iteratively propagate scores to other essays in the same cluster. ☹
- Zhang and Litman use ***word count*** as weak supervision to train AES model. ☹

- **A single quality signal** cannot comprehensively describe the quality of essay.
- **More quality signals** bring stronger and more robust supervision. ☺

* **[Chen et al., 2010]** Yen-Yu Chen, Chien-Liang Liu, Tao-Hsing Chang, and Chia-Hoang Lee. 2010. An unsupervised automated essay scoring system. IEEE Computer Architecture Letters, 25(05):61–67.
* **[Zhang and Litman, 2021]** Haoran Zhang and Diane Litman. 2021. Essay quality signals as weak supervision for source-based essay scoring. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 85–96.

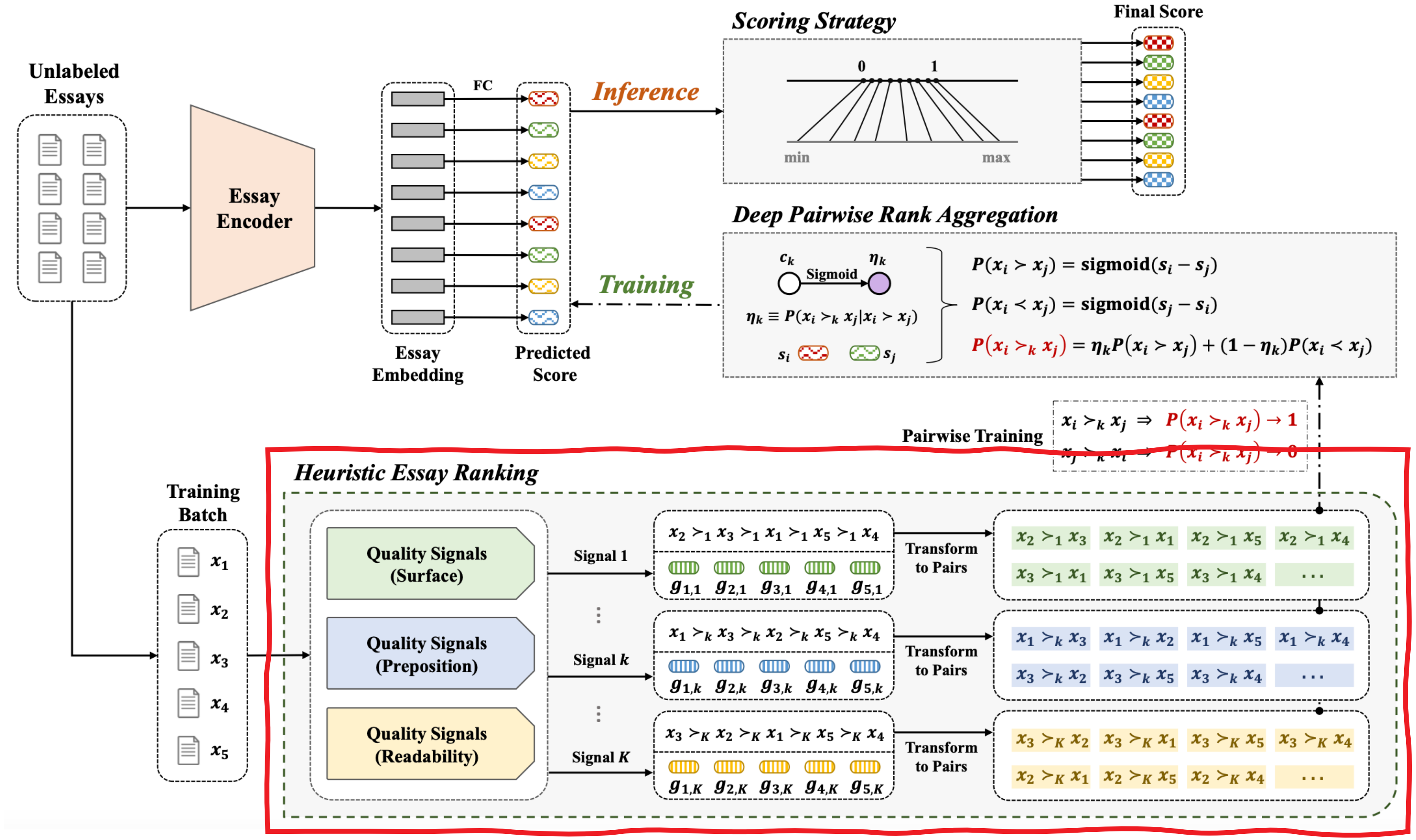
Our Method

A novel framework for Unsupervised AES by Learning from Rank Aggregation

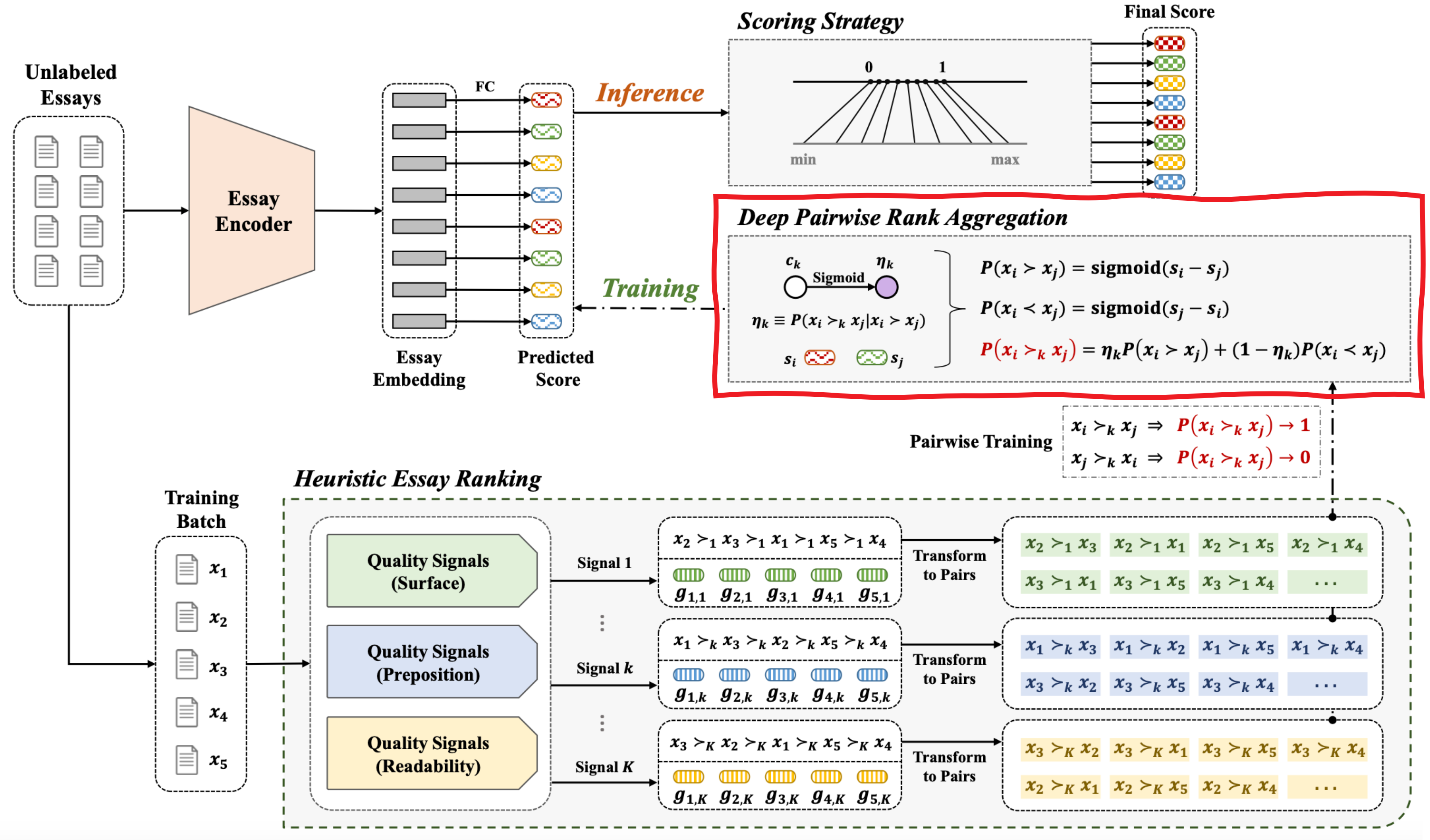
(ULRA)

Core idea *is to introduce multiple heuristic quality signals as pseudo-groundtruth, and then train a neural AES model by learning from the aggregation of them.*

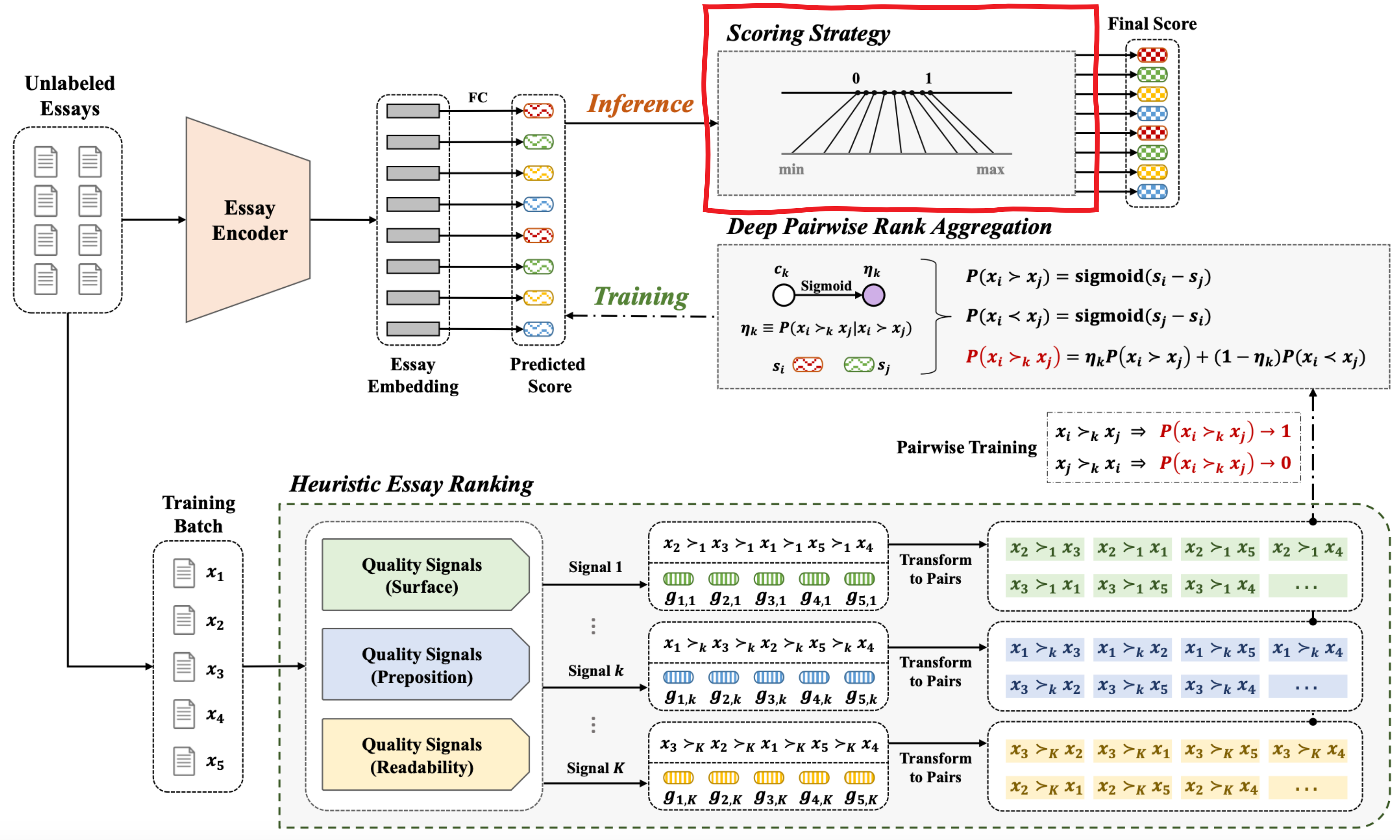
Our Method / HER



Our Method / DPRA



Our Method / Scoring Strategy



Experiments

Setting	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
One-Shot	**TGOD** (Jiang et al., 2021)	.772	.581	.690	.725	.776	.691	.766	.505	.688
Unsupervised	**Mean of the 20 quality signals**	.283	.333	.234	.353	.253	.206	.189	.264	.264
	Maximum of the 20 quality signals	.469	.536	.394	.471	.375	.323	.295	.458	.415
	Signal Clustering (Chen et al., 2010)	.355	.386	.370	.446	.509	.425	.428	.334	.407
	Signal Clustering w/ averaged signal as supervision	.393	.408	.383	.480	.500	.425	.470	.354	.427
	Signal Clustering w/ averaged output as prediction	.405	.413	.384	.498	.509	.435	.473	.370	.436
	Signal Clustering w/ aggregated signal as supervision	.359	.425	.404	.466	.535	.461	.465	.371	.436
	Signal Clustering w/ aggregated output as prediction	.363	.419	.397	.467	.544	.464	.467	.379	.438
	Signal Regression (Zhang and Litman, 2021)	.224	.321	.264	.404	.301	.441	.292	.353	.325
	Signal Regression w/ averaged signal as supervision	.232	.326	.271	.415	.303	.451	.304	.368	.334
	Signal Regression w/ averaged output as prediction	.249	.342	.289	.430	.311	.470	.316	.374	.348
	Signal Regression w/ aggregated signal as supervision	.246	.342	.263	.434	.309	.454	.304	.349	.338
	Signal Regression w/ aggregated output as prediction	.256	.344	.284	.451	.333	.496	.341	.345	.356
	Signal Aggregation (Chen et al., 2013)	.435	.480	.454	.608	.452	.439	.489	.218	.455
	ULRA (Ours)	**.757**	**.621**	**.547**	**.628**	**.664**	**.562**	**.694**	**.450**	**.615**

Transductive

Setting	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
				.621	.742	.784	.775	.730	.617	.705
				.672	.814	.803	.811	.801	.705	.764
				.731	.801	.823	.792	.762	.684	.773
				.711	.787	.808	.814	.786	.734	.773
	R^2BERT (Yang et al., 2020)	.817	.719	.698	.845	.841	.847	.839	.744	.794
	(Uto et al., 2020)	.852	.651	.804	.888	.885	.817	.864	.645	.801
Cross-Prompt	**CNN-LSTM-Att** (Dong et al., 2017)	.592	.553	.666	.680	.690	.656	.640	.565	.630
	HA-LSTM (Cao et al., 2020)	.633	.545	.685	.683	.729	.629	.281	.436	.578
	BERT (Cao et al., 2020)	.661	.669	.651	.698	.709	.599	.725	.574	.661
Unsupervised	**Mean of the 20 quality signals**	.320	.408	.285	.419	.262	.296	.305	.272	.320
	Maximum of the 20 quality signals	.511	.606	.420	.549	.368	.464	.427	.444	.474
	Signal Regression (Zhang and Litman, 2021)	.244	.309	.216	.338	.234	.189	.151	.247	.241
	Signal Regression w/ averaged signal as supervision	.253	.328	.219	.355	.247	.183	.162	.248	.249
	Signal Regression w/ averaged output as prediction	.269	.341	.213	.364	.239	.193	.180	.248	.256
	Signal Regression w/ aggregated signal as supervision	.252	.314	.239	.351	.246	.198	.167	.271	.255
	Signal Regression w/ aggregated output as prediction	.258	.319	.250	.365	.248	.216	.191	.300	.268
	ULRA (Ours)	**.759**	**.508**	**.608**	**.644**	**.711**	**.577**	**.661**	**.446**	**.614**

Experiments

Setting	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
One-Shot	**TGOD** (Jiang et al., 2021)	.772	.581	.690	.725	.776	.691	.766	.505	.688
Unsupervised	**Mean of the 20 quality signals**	.283	.333	.234	.353	.253	.206	.189	.264	.264
	Maximum of the 20 quality signals	.469	.536	.394	.471	.375	.323	.295	.458	.415
	Signal Clustering (Chen et al., 2010)	.355	.386	.370	.446	.509	.425	.428	.334	.407
	Signal Clustering w/ averaged signal as supervision	.393	.408	.383	.480	.500	.425	.470	.354	.427
	Signal Clustering w/ averaged output as prediction	.405	.413	.384	.498	.509	.435	.473	.370	.436
	Signal Clustering w/ aggregated signal as supervision	.359	.425	.404	.466	.535	.461	.465	.371	.436
	Signal Clustering w/ aggregated output as prediction	.363	.419	.397	.467	.544	.464	.467	.379	.438
	Signal Regression (Zhang and Litman, 2021)	.224	.321	.264	.404	.301	.441	.292	.353	.325
	Signal Regression w/ averaged signal as supervision	.232	.326	.271	.415	.303	.451	.304	.368	.334
	Signal Regression w/ averaged output as prediction	.249	.342	.289	.430	.311	.470	.316	.374	.348
	Signal Regression w/ aggregated signal									
	Signal Regression w/ aggregated outpu									
	Signal Aggregation (Chen et al., 2013)									
	ULRA (Ours)									

Inductive

Setting	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
Supervised	**BLRR** (Phandi et al., 2015)	.761	.606	.621	.742	.784	.775	.730	.617	.705
	CNN-LSTM-Att (Dong et al., 2017)	.822	.682	.672	.814	.803	.811	.801	.705	.764
	TSLF (Liu et al., 2019)	.852	.736	.731	.801	.823	.792	.762	.684	.773
	HA-LSTM (Cao et al., 2020)	.828	.718	.711	.787	.808	.814	.786	.734	.773
	R^{2}BERT (Yang et al., 2020)	.817	.719	.698	.845	.841	.847	.839	.744	.794
	(Uto et al., 2020)	.852	.651	.804	.888	.885	.817	.864	.645	.801
Cross-Prompt	**CNN-LSTM-Att** (Dong et al., 2017)	.592	.553	.666	.680	.690	.656	.640	.565	.630
	HA-LSTM (Cao et al., 2020)	.633	.545	.685	.683	.729	.629	.281	.436	.578
	BERT (Cao et al., 2020)	.661	.669	.651	.698	.709	.599	.725	.574	.661
Unsupervised	**Mean of the 20 quality signals**	.320	.408	.285	.419	.262	.296	.305	.272	.320
	Maximum of the 20 quality signals	.511	.606	.420	.549	.368	.464	.427	.444	.474
	Signal Regression (Zhang and Litman, 2021)	.244	.309	.216	.338	.234	.189	.151	.247	.241
	Signal Regression w/ averaged signal as supervision	.253	.328	.219	.355	.247	.183	.162	.248	.249
	Signal Regression w/ averaged output as prediction	.269	.341	.213	.364	.239	.193	.180	.248	.256
	Signal Regression w/ aggregated signal as supervision	.252	.314	.239	.351	.246	.198	.167	.271	.255
	Signal Regression w/ aggregated output as prediction	.258	.319	.250	.365	.248	.216	.191	.300	.268
	ULRA (Ours)	**.759**	**.508**	**.608**	**.644**	**.711**	**.577**	**.661**	**.446**	**.614**

Conclusion

- We aim to perform essay scoring under the unsupervised setting.
- We propose ULRA to train a neural AES model by aggregating the partial-order knowledge contained in multiple heuristic quality signals.
- To address the conflicts among different signals and get a unified supervision, we design a deep pairwise rank aggregation loss for model training.
- Experimental results demonstrate the effectiveness of ULRA for unsupervised essay scoring.

THANKS!