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

1. Introduction

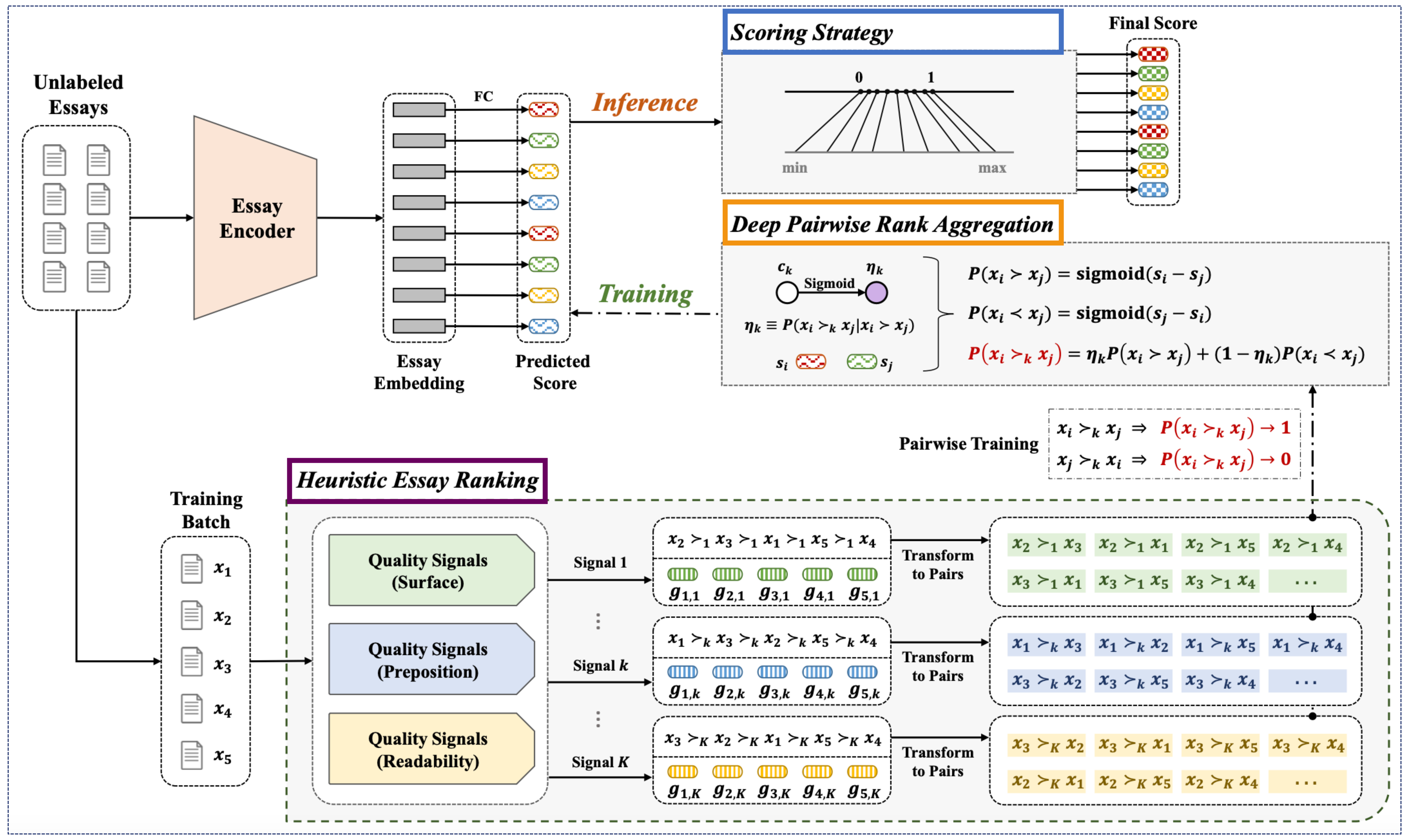
- Automated Essay Scoring (AES) aims to score writing quality of essays without human intervention.
- SOTA AES models are trained in a supervised way with large labeled corpora.
- Collecting a large volume of labeled essays is both time-consuming and labor-intensive.
- **Unsupervised AES** does not require groundtruth scores for training, and has potential in scientific research and practical applications.

2. Motivation

- Two existing unsupervised AES methods select one heuristic quality signal to train the models, but both of which achieve poor performance.
- A single heuristic quality signal can not fully describe the quality of essay.
- More heuristic quality signals should be introduced to bring stronger and more robust supervision.

3. ULRA Framework

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.*



◆ **Heuristic Essay Ranking (HER)** generates partial-order pairs through ranking essays according to heuristic quality signals.

◆ **Deep Pairwise Rank Aggregation (DPRA)** trains a neural AES model by aggregating the partial-order pairs derived from multiple quality signals into a unified supervision.

◆ **Scoring Strategy** transforms the predicted scores given by the neural AES model into the range of the pre-defined score set.

4. Experiments

Performance Comparison

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

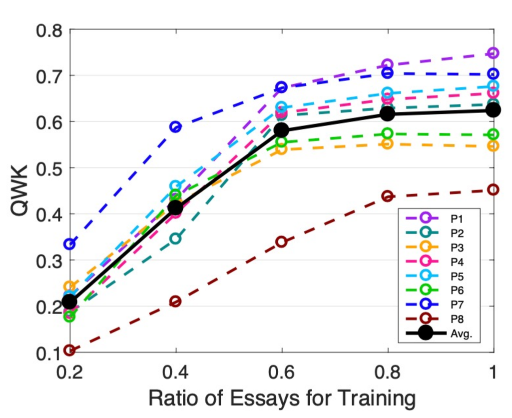
Setting	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
Supervised	BLRR (Phandi et al., 2015)	.761	.606	.621	.742	.784	.775	.730	.617	.705
	CNN-LSTM-Att (Dong et al., 2017)	.822	.682	.672	.814	.803	.811	.801	.705	.764
	TSLF (Liu et al., 2019)	.852	.736	.731	.801	.823	.792	.762	.684	.773
	HA-LSTM (Cao et al., 2020)	.828	.718	.711	.787	.808	.814	.786	.734	.773
	R^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**

Inductive Setting

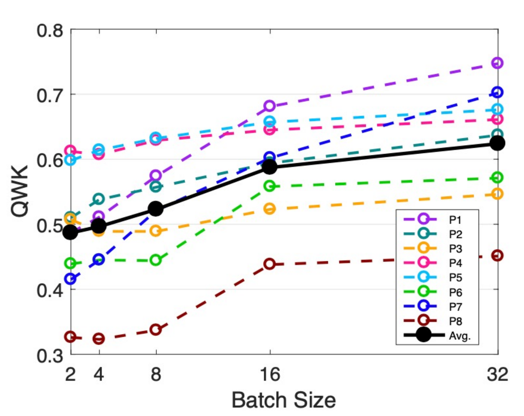
Ablation Study

	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
Full Model	**.757**	**.621**	**.547**	.628	**.664**	**.562**	**.694**	.450	**.615**
− **learnable** η_k (fix $\eta_k = 1$)	.702	.610	.504	.610	.651	.547	.610	.380	.577
− **pretrained neural model** (using CNN-LSTM-Att)	.634	.599	.501	.628	.411	.553	.641	.419	.548
− **pretrained neural model** (using HA-LSTM)	.653	.613	.513	.605	.600	.501	.615	.436	.567
− **neural model** (all s_i are set as learnable parameters)	.432	.481	.451	.519	.600	.450	.484	.213	.454
− **surface signals** (preposition & readability signals)	.714	.610	.419	.593	.623	.541	.585	**.451**	.567
− **preposition signals** (surface & readability signals)	.694	.584	.504	.613	.649	.515	.643	**.451**	.582
− **readability signals** (surface & preposition signals)	.712	.584	.471	.626	.631	.500	.683	.431	.580
− **preposition & readability signals** (only surface signals)	.672	.588	.543	.628	.597	.497	.612	.434	.571
− **surface & readability signals** (only preposition signals)	.691	.553	.441	.518	.483	.429	.677	.403	.524
− **surface & preposition signals** (only readability signals)	.654	.627	.464	.563	.598	.514	.661	.444	.566
w/ averaged signal as supervision	.524	.541	.501	.615	.646	.542	.545	.245	.520
w/ averaged output as prediction	.536	.542	.519	.621	.632	.561	.553	.270	.529
w/ aggregated signal as supervision	.548	.544	.531	.624	.648	.548	.562	.262	.533
w/ aggregated output as prediction	.573	.544	.530	**.629**	.649	.551	.566	.260	.538

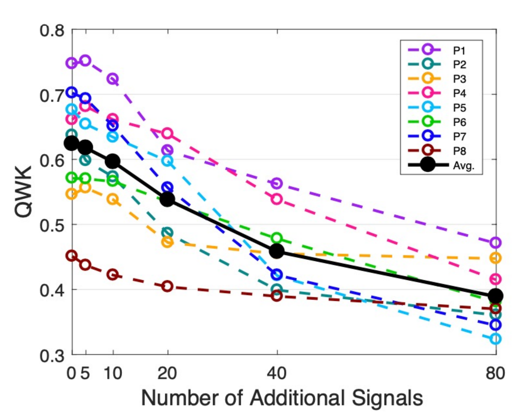
Model Analysis: Part I



More Unlabeled Essays.

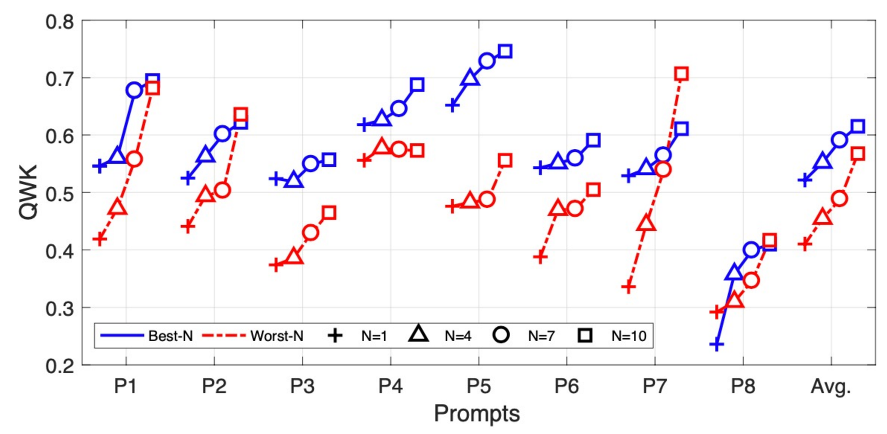


More Training Pairs.



Weak Signals.

Model Analysis: Part II



Effect of More Signals.

The results are reported by training with the N best or worst signals from the signal set.

		P1	P2	P3	P4	P5	P6	P7	P8	Avg.
T	G	.674	.789	.998	.999	.922	.897	.812	.585	.835
	O	.757	.621	.547	.628	.664	.562	.694	.450	.615
I	G	.635	.610	.567	.842	.713	.769	.717	.448	.663
	O	.759	.508	.608	.644	.711	.577	.661	.446	.614

Groundtruth as Signal.

Comparison the performance of applying ground-truth score as the quality signal (**G**) with that of applying 20 heuristic quality signals (**O**) under all 8 prompts of the ASAP dataset. **T** and **I** denote the transductive and inductive settings, respectively.

	P1	P2	P3	P4	P5	P6	P7	P8
Transductive	.7438	.6855	.6677	.7813	.5365	.6033	.8360	.8932
Inductive	.7442	.6659	.6052	.7994	.5681	.6259	.8254	.9007

Effect of Confidence Weights.

Spearman's correlation coefficient between the learned confidence weights and corresponding QWKs, which are calculated between groundtruth scores and the employed 20 signals under each prompt.

	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
G	.840	.693	.688	.730	.807	.704	.730	.610	.725
N	.545	.551	**.645**	**.729**	.736	.554	.601	.300	.583
T	.576	.595	.631	.727	**.742**	.553	.673	.346	.605
U	.543	.568	.632	.728	.730	.554	.586	.296	.580
O	**.757**	**.621**	.547	.628	.664	**.562**	**.694**	**.450**	**.615**

Effect of Different Scoring Strategies.

G, **N**, **T**, and **U** denote the scoring strategies based on the groundtruth, normal, triangle, and uniform distributions, respectively. **O** denotes our scoring strategy.

Code is available in https://github.com/tenvence/ulra.