COGNITIVELY AIDED ZERO-SHOT AUTOMATIC ESSAY GRADING

Sandeep Mathias[1], Rudra Murthy[2], Diptesh Kanojia[1,3], and Pushpak Bhattacharyya[1]

[1]Indian Institute of Technology, Bombay

[2]IBM Research India

[3]IITB-Monash Research Academy

OUTLINE

- Definitions
- Related Work
- System Architecture
- Dataset
- Experiment Configuration
- Results
- Conclusion & Future Work

DEFINITIONS

- An **essay** is a piece of text written in response to a topic, called a **prompt**.
- **Automatic essay grading** (AEG) is using a machine to assign a score to the essay.
- **Zero-shot AEG** is training a system to grade essays without using any target prompt essays during the training process.
- **Cognitively Aided Zero-shot AEG** is using cognitive information to help in zero-shot AEG.

RELATED WORK

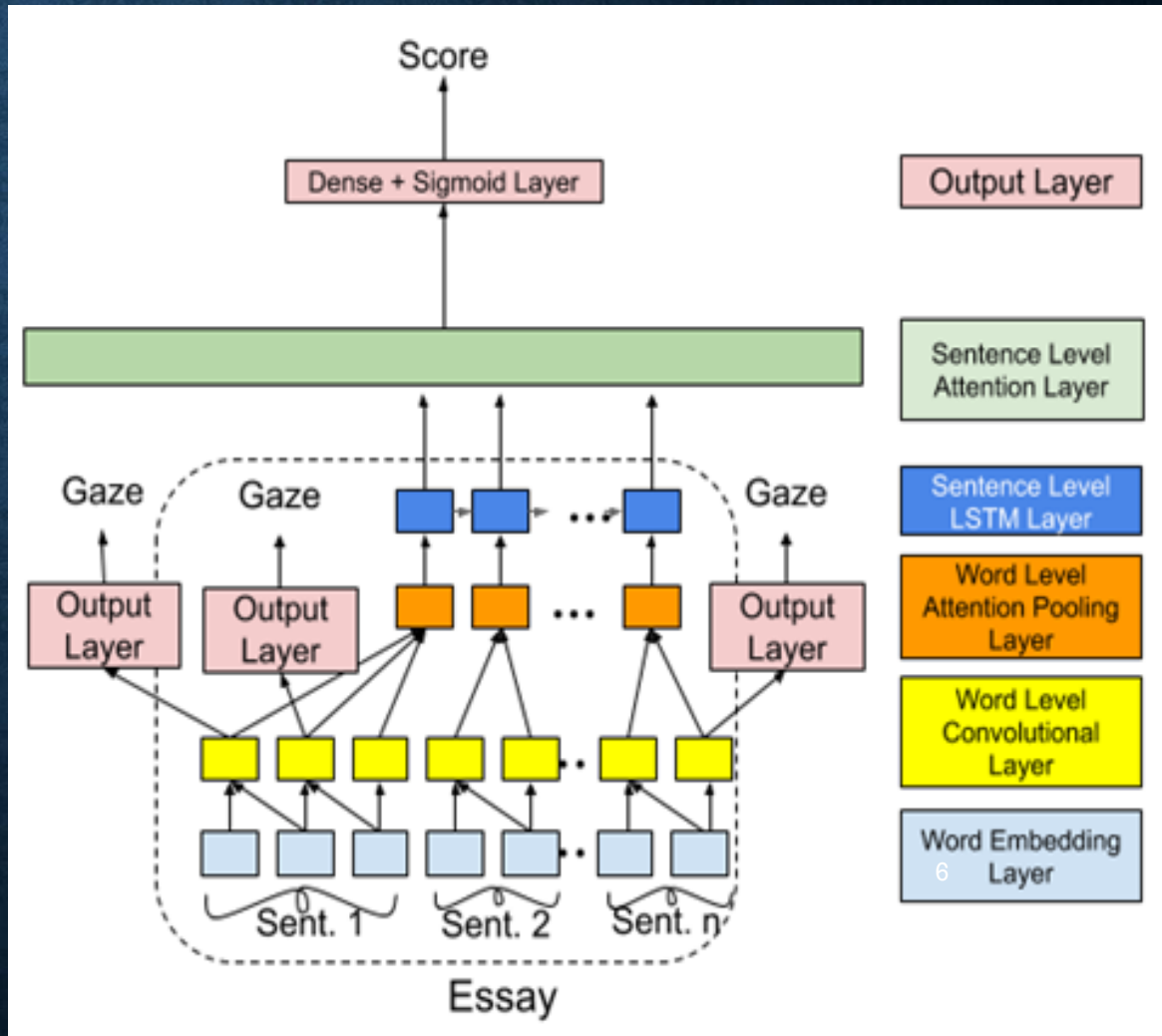
- Cross-Prompt AEG:
  - Machine learning approaches – Phandi et al. (2015), Cozma et al. (2018)
  - Deep learning approaches – Dong et al. (2016), Jin et al. (2018)
- Learning gaze behaviour:
  - Part-of-speech tagging: Barrett et al. (2016a), Barrett et al. (2016b)
  - Sentence Simplification: Klerke et al. (2016)
  - Readability: Singh et al. (2016), Gonzalez-Garduno and Sogaard (2018)
  - Sentiment Analysis: Mishra et al. (2018), Barrett et al. (2018), Long et al. (2019)
  - Named Entity Recognition: Hollenstein and Zhang (2019)
  - Automatic Essay Grading: Mathias et al. (2020)

METHOD

- Multi-task learning (MTL) is a machine learning paradigm where we use information from auxiliary tasks to aid in solving a primary task.
- We use gaze behaviour for very few essays (Mathias et al. 2020) and use that data to help in training a model to score other essays.
- In this framework, scoring the essay is the primary task, and learning the gaze behaviour is the auxiliary task.

SYSTEM ARCHITECTURE

- Embedding layer: Input is words, output is word embeddings.
- Word-level CNN & Attention Pooling layers: Input is the word-embeddings, output is the sentence representation.
- Sentence-level LSTM & Attention Pooling layers: Input is sentence representations, output is essay representation.
- Gaze behaviour is learnt at the word-level CNN layer.



ESSAY GRADING DATASET

Prompt ID	No. of Essays	Score Range	Avg. Word Count	Essay Type
Prompt 1	1783	2-12	350	Persuasive
Prompt 2	1800	1-6	350	Persuasive
Prompt 3	1726	0-3	150	Source-Dependent Response
Prompt 4	1770	0-3	150	Source-Dependent Response
Prompt 5	1805	0-4	150	Source-Dependent Response
Prompt 6	1800	0-4	150	Source-Dependent Response
Prompt 7	1569	0-30	250	Narrative
Prompt 8	723	0-60	650	Narrative
Total / Mean	12976	0-60	250	Multiple Types

GAZE BEHAVIOUR DATASET

- No. of annotators = 8
- Maximum words per essay = 250

Essay Set	Scored 0	Scored 1	Scored 2	Scored 3	Scored 4	Total
Prompt 3	2	4	5	1	-	12
Prompt 4	2	3	4	3	-	12
Prompt 5	2	1	3	5	1	12
Prompt 6	2	2	3	4	1	12
Total	**8**	**10**	**15**	**13**	**2**	**48**

NETWORK HYPERPARAMETERS

Layer	Hyperparameter	Value
Embedding Layer	Pre-trained Embedding	GloVe
	Embedding Dimensions	50
Word-level CNN	Kernel Size	5
	Filters	50
Sentence-level LSTM	Hidden Units	100
Network-wide	Batch Size	100
	Epochs	100
	Learning Rate	0.001
	Dropout Rate	0.5
	Momentum	0.9
Gaze Feature Weights	Dwell Time	0.05
	First Fixation Duration	0.05
	IsRegression	0.01
	Run Count	0.01
	Skip	0.1

EXPERIMENT CONFIGURATION

- 1 essay set (Target Essay Set) is used for testing. Remaining 7 essay sets are used for training and validation.
- Evaluation Method: Five-fold cross-validation
- Training and validation sets are split into 5 folds – 4 for training and 1 for validation.
- Evaluation metric: Quadratic Weighted Kappa (QWK)
- Testing set is evaluated on the best-performing model for the validation set.

RESULTS

Target Essay Set	No Gaze	Gaze
Prompt 1	0.319	**0.423***
Prompt 2	0.391	**0.439***
Prompt 3	0.508	**0.545***
Prompt 4	0.548	**0.626***
Prompt 5	0.548	**0.628***
Prompt 6	0.599	**0.600**
Prompt 7	0.362	**0.420***
Prompt 8	**0.316**	0.286
Mean QWK	0.449	**0.498***

ANALYSIS

- We observe that the gaze data is helping in improved performance of the AEG system, even though there are no target essay set essays present in training.
- The only prompt where this is not true is for Prompt 8 – which is a narrative prompt with much longer essays compared to the other 7 Prompts.

CONCLUSION & FUTURE WORK

- Using gaze behaviour, we are able to learn cognitive information which is useful for grading essays – even when we have no essays from that prompt as part of the training data.
  - We see an almost 5% improvement when using gaze behaviour as compared to when we don't.
- In the future, we plan to extend our work to grading essay traits – like content, organization, style, etc. – as well.

THANK YOU!

Questions?

REFERENCES

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Sogaard. 2016a. Weakly Spuervised Part-of-Speech Tagging Using Eye-Tracking Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579 – 584, Berlin, Germany. Association for Computational Linguistics.

Maria Barrett, Frank Keller, and Anders Sogaard. 2016b. Cross-Lingual Transfer of Correlations Between Parts of Speech and Gaze Features. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1330 – 1339, Osaka, Japan. The COLING 2016 Organizing Committee.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei and Anders Sogaard. 2018. Sequence Classification with Human Attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302 – 312, Brussels, Belgium. Association for Computational Linguistics.

Madalina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. Automated Essay Scoring with String Kernels and Word Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503 – 509, Melbourne, Australia. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic Features for Essay Scoring – An Empirical Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072 – 1077, Austin, Texas, USA. Association for Computational Linguistics.

Ana V Gonzalez-Garduno and Anders Sogaard. 2018. Learning to Predict Readability Using Eye-Movement Data from Natives and Learners. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5118 – 5124, New Orleans, Louisiana, USA. AAAI.

Nora Hollenstein and Ce Zhang. 2019. Entity Recognition at First Sight: Improving NER with Eye Movement Information. In *Proceedings of the 2019 Conference of the North-American Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1 – 10, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sigrid Klerke, Yoav Goldberg, and Anders Sogaard. 2016. Improving Sentence Compression by Learning to Predict Gaze. In *Proceedings of the 2016 Conference of the North-American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528 – 1533, San Diego, California, USA. Association for Computational Linguistics.

Yunfei Long, Rong Xiang, Qin Lu, Chu-Ren Huang, and Minglei Li. 2019. Improving Attention Model Based on Cognition Grounded Data for Sentiment Analysis. *IEEE Transactions of Affective Computing.*

Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Eyes are the Windows to the Soul: Predicting the Rating of Text Quality Using Gaze Behaviour. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2352 – 2362. Melbourne, Australia. Association for Computational Linguistics.

Sandeep Mathias, Rudra Murthy, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2020. Happy are those who Grade without Seeing: A Multi-task Learning Approach to Grade Essays Using Gaze Behaviour. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference in Natural Language Processing*, pages 858 – 872, Suzhou, China. Association for Computational Linguistics.

Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey. 2018. Cognition-Cognizant Sentiment Analysis with Multitask Subjectivity Summarization Based on Annotators' Gaze Behaviour. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5884 – 5891, New Orleans, Louisiana, USA. AAAI.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431 – 439, Lisbon, Portugal. Association for Computational Linguistics.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying Sentence Complexity Based on Eye-Tracking Measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee