Toward Educator-focused Automated Essay Scoring Systems

Mike Hardy

Progress of Automated Essay Scoring

1998 E-Rater (ETS)
- Statistical NLP begins

2014 ASAP-AES Competition (Hewitt)
- Competition and study still serve as human and private baseline for all automatic essay grading with public data set of 8 different essay prompts
- Quadratic Weighted Kappa established as evaluation metric

2016 Neural Networks Introduced
- Alikaniotis et al: Poor evaluation metric selected
- Taghipour and Ng: generally accepted evaluation methodology, outperformed humans

2017 Fancier NN: LSTM-CNN-Attention
- Dong et al: new SOTA,
- Zhang et al: first NN attempt at source dependency

2018 Non-neural Method at Public SOTA
- Cozma et al: char n-grams + super-bag-of-word-embeddings
- Tay et al: efficiency through LSTM attention aggregation

2019 BERT and Transformers
- Liu et al: Current SOTA for all 8 essays→ three parallel BERT models + LSTM + custom features + etc + etc

2020 Attempting at Domain Transfer
- Mayfield et al: BERT on some of the essays for transfer learning

Problems and Challenges

- Small amounts of labeled public data: Field still dominated by large testing companies ($$)
  - Baselines from 2014 private company competition results still strong (near SOTA)
  - Advantages: Astronomical amounts of data, many engineers and feature creators, years of experience
- Essays are longer than tweets
  - Heavy Compute (all models make concessions)
    - Transformer/Attention = $O(\text{sequence\_length}^2 \times \text{embedding\_size})$
  - Not full essay: BERT truncate, w2v
    - Unfortunately, this is not how students write—they don’t capture all their ideas in non-stop words or in the first half of an essay.

Evaluation Criteria and Baseline

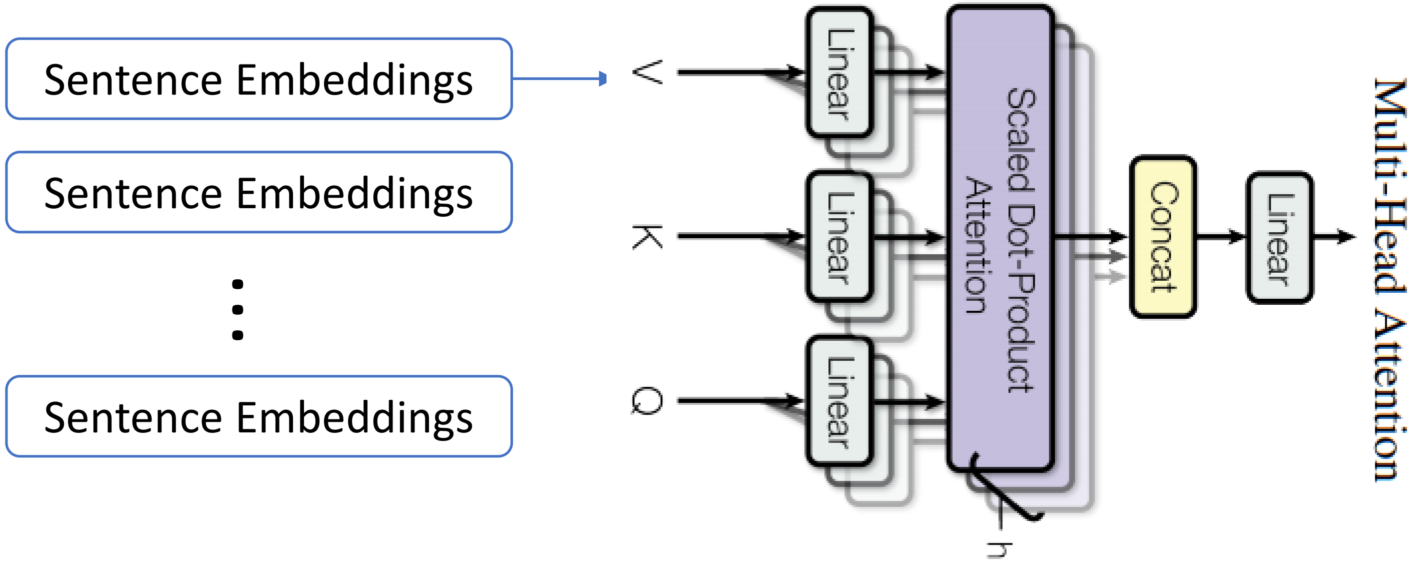
- Average QWK:
  - A measure of agreement for ordinal data relating distances
- Taghipour re-established validity, using 5-fold CV with separate test set
- 3rd place from the public competition is widely used as baseline.

Prompt	Grade Level	Len	Score Range	Train Size	Dev Size	Test Size	Description
1	8	350	2-12	1190	298	298	Persuasive Letter about Technology Use
2	10	350	1-6	1200	300	300	Persuasive Essay about Library Censorship
3	10	150	0-3	1151	288	288	Literary Analysis of Setting (Source 1)
4	10	150	0-3	1181	295	295	Analysis of Author's Purpose (Source 2)
5	8	150	0-4	1203	301	301	Analysis of Mood (Source 3)
6	10	150	0-4	1200	300	300	Demonstration of comprehension of Text (Source 4)
7	7	250	0-30	1153	288	288	Narrative about Patience
8	10	650	0-60	612	153	153	Narrative about Laughter

Quick Note on Evaluation

- The purpose of essay grading is to provide the appropriate grade to the appropriate essay.
- Alikaniotis (2016) calculated various average measures of agreement, across all essays, rather than for each essay individually before taking the average.
  - Cohen's –κ / QWK across every test? 0.989! SOTA!
    - SOTA from using word2vec pretrained BOW embeddings on a 2-layer bidirectional LSTM trained for 100 epochs.
  - …except…tests aren't getting the correct score. It is a mathematical exercise, not actually solving the problem.
    - Difference in range of possible scores does not scale appropriately. It favors essays with larger scales.

Base Model Design



Novel Solutions to the Length Challenge

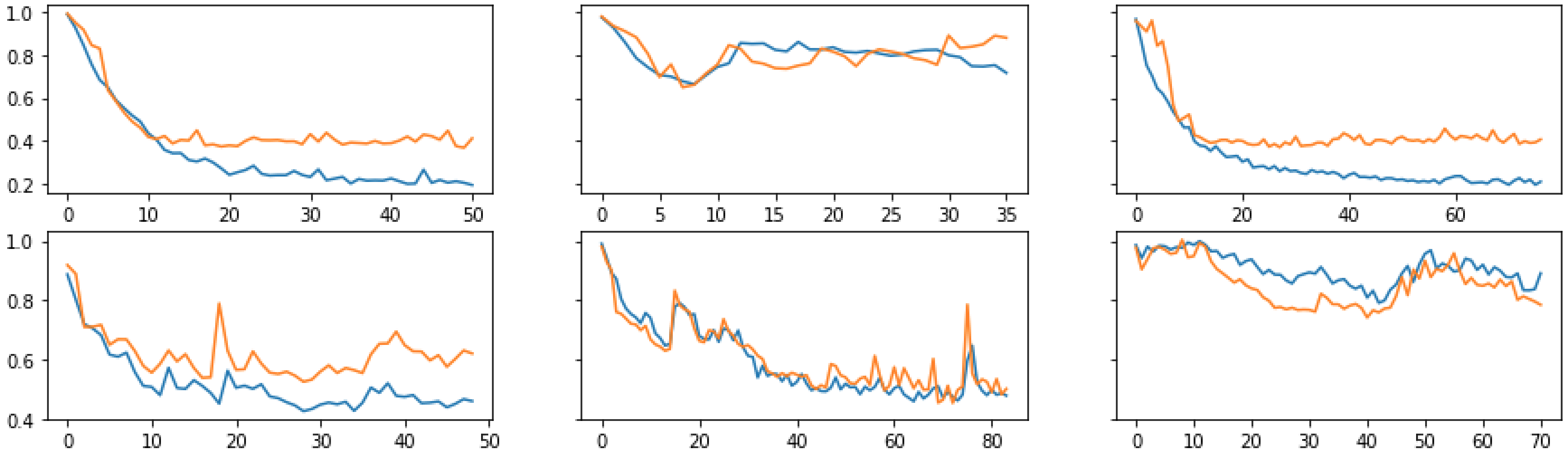
- Sentence Embeddings:
  - Use USE to capture semantic representations of all the sentences.
- Reduce NN compute needs:
  - Multi-head Attention Decoder, instead of BERT
  - Other studies cited “8 GPUs” “8000 epochs”, etc.
    - 1 GPU, I can run an epoch in max 7 seconds. Avg = 2 second
- Adapting hyperparameters given nature of essay set
  - Essays with more essay classes can handle more complex layers
- Custom Loss function for ordinal data

Optimizers for Sentence Embeddings

Adam

#1

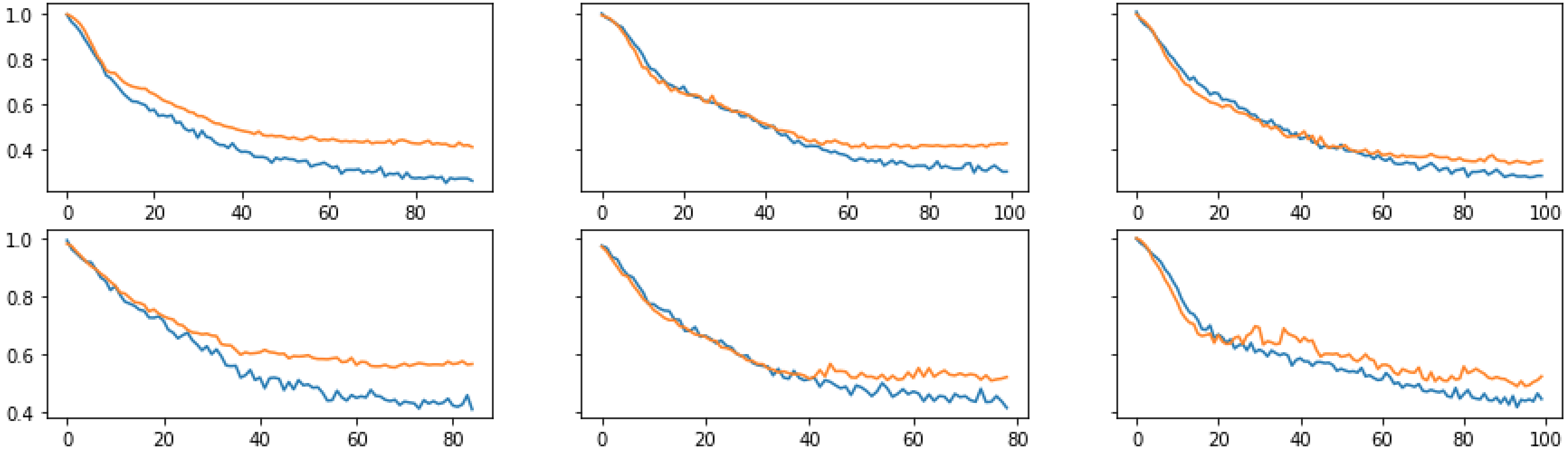
#2



Adamax

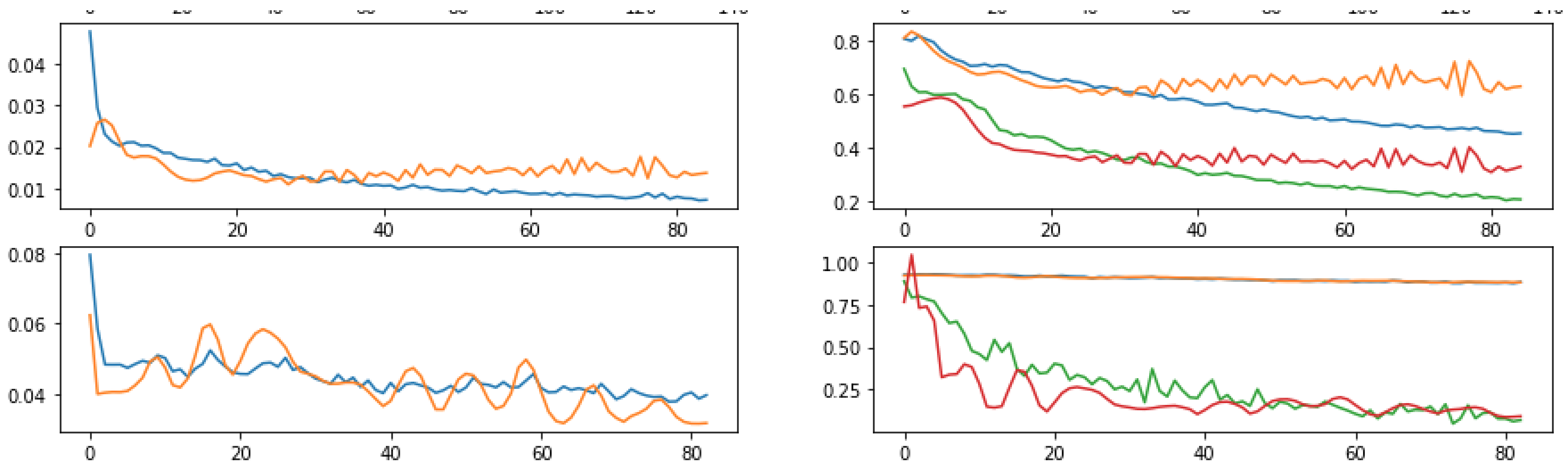
#1

#2



Objective Functions

- For essays with few categories:
  - SparseCategoricalCrossEntropy < custom QWK loss
- For essays with many categories
  - Regression using MeanSquaredError
- Best overall:
  - Custom combination of both objectives (grading and alignment), but much fickler.



Baselines and SOTA

Model	1	2	3	4	5	6	7	8	Avg
LSTM (Alikaniotis)	0.47	0.28	0.50	0.58	0.51	0.50	0.67	0.25	0.47
Multi-stage DNN (Jin)	0.77	0.69	0.63	0.76	0.74	0.68	0.63	0.57	0.69
EASE (baseline)	0.76	0.61	0.62	0.74	0.78	0.78	0.73	0.62	0.71
CNN-RNN (Dasgupta)	0.80	0.63	0.71	0.71	0.80	0.83	0.82	0.70	0.75
Human Raters (Shermis 2014)	0.73	0.80	0.76	0.77	0.85	0.74	0.72	0.61	0.75
LSTM+CNN Ensemble (Taghipour)	0.82	0.69	0.69	0.81	0.81	0.82	0.81	0.64	0.76
LSTM+CNN+Attention (Dong)	0.82	0.68	0.67	0.81	0.80	0.81	0.80	0.71	0.76
BERT Multistage Ensemble (Liu)	0.85	0.74	0.73	0.80	0.82	0.79	0.76	0.68	0.77
String Kernel / Word Embeds (Cozma)	0.85	0.73	0.68	0.83	0.83	0.83	0.80	0.73	0.79
Top Private Statistical NLP (Shermis 2014)	0.82	0.74	0.75	0.82	0.83	0.81	0.84	0.73	0.79
USE+LSTM	0.72	0.58	0.69	0.81	0.77	0.74	0.74	0.42	0.68
USE+ MTA	0.74	0.48	0.54	0.74	0.68	0.64	0.72	<u>0.74</u>	0.66
USE+2MTA	0.75	<u>0.74</u>	0.68	<u>0.83</u>	0.78	0.79	0.77	0.62	0.74
USE+2MTA+BLSTM	<u>0.83</u>	0.64	<u>0.70</u>	0.82	<u>0.79</u>	<u>0.82</u>	<u>0.80</u>	0.65	<u>0.76</u>

**As re-implemented by Jin et al*

Future Work

- Model that surpasses human ability on all essays
- Trait scoring (especially content-based scoring), using semantic work
- Transfer learning (identify traits trainable across prompts)
- Transferrable source- and prompt-dependence (co-attention)
- Multi-task (all in one multitasker)
- Zero shot source-dependent abilities
- Offer as a free, open-source service
- Make teachers more effective