Toward Educator-focused Automated Essay Scoring Systems

Mike Hardy
Tennessee Department of Education
Michael.hardy@tn.gov

Abstract

This paper presents methods for improving automated essay scoring with techniques that address the computational trade-offs of self-attention and document length. To make Automated Essay Scoring (AES) more useful to practitioners, researchers must overcome the challenges of data and label availability, authentic and extended writing, domain scoring, prompt and source variety, and transfer learning. This paper addresses these challenges using neural network models by employing techniques that preserve essay length as an important feature without increasing model training costs. It introduces techniques for minimizing classification loss on ordinal labels using multi-objective learning, capturing semantic information across the entire essay using sentence embeddings to use transformer architecture across arbitrarily long documents, and automated hyperparameter generation based on prompt-corpus metadata, resulting in state-of-the-art results for some essay tasks.

1 Introduction

Assessments measuring student learning need to increase in validity—accurately measuring intended learning. While standardized, multiple-choice tests have increased statistical reliability, writing remains a more authentic measure of student learning. However, hand-scoring essays is both expensive and time-consuming, leading to increased emphasis in and subsequent leaps in Automated Essay Scoring (AES) in last two decades.

However, the market and research for AES has been driven primarily by companies who do extensive essay grading, and primarily focuses on producing a reliable holistic—or overall—scores (see Ke et al., 2019). While these scores serve the purposes of standardized testing, holistic scores are less useful to educators working to remediate immediate student needs, and, for educators in public schools, AES services are typically unavailable at all. The work of this paper is to refocus research in AES to be educator-focused and pedagogically-aligned by introducing methodologies and approaches that value the long-term needs of educators and students: teachers want a free, useful, and insightful automatic scorer for writing done by their students during the school year and students want meaningful instruction and feedback that improves their writing. Becoming more useful and insightful requires including domain-specific scores, in addition to holistic scores, that represent pedagogically aligned writing traits that match instructional needs of teachers and the learning needs of students.

This is a massive challenge, as that kind of AES would require zero-shot learning on novel prompts of arbitrary length, which may be source-dependent and which would need to be scored across various writing traits accounting for the writing development progression found by student age within K-12 schools. Meeting this challenge requires a re-focus of the current AES research trends to work toward that ultimate outcome.

To better serve educators and learners, some of the most immediate challenges for AES include data availability, authentic and extended writing, domain scoring, transfer learning, and prompt and source variety.

2 Related Work

This section will detail approaches to solving some of the challenges in public-educator-focused essay scoring.

2.1 Corpora and Labeling

Due to the need to protect the educational information of minors, there is very limited data available to the public. This provides a great advantage to companies to provide grading services, as they jointly share a monopoly on essay grading data. Only one dataset representing K12 students is publicly available: the Hewitt Foundation's ASAP-AES dataset[1] that was used in the 2012 Kaggle competition of the same name, shown in Table 1. This dataset, like nearly every other publicly available datasets for student essay scoring, is annotated holistically, with only two of the eight essays containing trait annotations available beyond holistic scoring, which were not all evaluated for the competition and offer different traits using different scoring and scaling criteria. While using the same data source is helpful for benchmark comparison, it significantly restricts the robustness of transferrable models, and increases the bias

[1] https://www.kaggle.com/c/asap-aes/

found within the sample when developing AES methodologies.

Additionally, the score ranges and grading criteria are different for different prompts, making it challenging to effectively address gaps across prompts. Due to computation restrictions or to avoid the challenges of unusual classifications, many researchers create models on only part of the ASAP set (Zhang et al., 2018; Nadeem, et al., 2019, Mayfield et al., 2020)

Nearly all existing AES systems were developed for holistic scoring, with trait-specific scoring not starting until 2004. Primarily using researcher-generated estimations, features, or ratings, studies on single traits such as organization (Persing et al., 2010; Taghipour, 2017), prompt adherence (Persing and Ng, 2014), thesis clarity (Persing and Ng, 2013; Ke et al., 2019), argument strength (Persing and Ng, 2015; Taghipour, 2017; Carlile et al., 2018), coherence (Somasundaran et al., 2014), and stance (Persing and Ng, 2016), and narrative quality (Somasundaran et al., 2018). Responding the challenge of obtaining trait-labeled datasets, Mathias et al. (2016), produced trait labels for the remaining six essays to create the ASAP++ dataset. These trait labels are not gold-standard (as they received ratings from only one reviewer), but their effort is a worthy contribution, as it is followed up with a study (Mathias et al., 2020) where they replicate certain SOTA models across multiple traits. Through its imperfections, this label augmentation heightens the need for scoring models can that support transfer learning, so when a gold-standard, trait-labeled dataset becomes available, we have algorithms that can fully take advantage of that learning.

Another noteworthy limitation of the ASAP dataset is that the essays have already been standardized to a degree, as they do not contain any paragraph information and have gone through significant preprocessing to remove named entities and most capitalized words. While this was a precaution to protect minors from potentially identifiable information, it limits the long-term ability of this dataset to produce more authentic models.

In short, AES desperately needs labeled data, and until that is available, on-going research needs to prioritize methods that extract more meaningful information from limited samples.

Essay	Grade	Len	Score Range	Essay Count	Essay Description
1	8	350	2-12	1785	Persuasive Letter about Technology Use
2	10	350	1-6	1800	Persuasive Essay about Library Censorship
3	10	150	0-3	1726	Source-dependent Analysis of Setting
4	10	150	0-3	1772	Source-dependent Analysis of Author's Purpose
5	8	150	0-4	1805	Source-dependent Analysis of Mood
6	10	150	0-4	1800	Source-dependent Demonstration of comprehension of Text
7	7	250	0-30	1730	Narrative about Patience
8	10	650	0-60	918	Narrative about Laughter

Table 1: Overview of the ASAP-AES dataset from the Hewitt Foundation including prompt, grade-level of authors, average length of essay in words, range of possible scores, number of essays per prompt, and essay description.

2.2 AES Approaches

AES methods have typically relied on handcrafted features (Larkey, 1998; Foltz et al., 1999; Attali and Burstein, 2006; Dikli, 2006; Wang and Brown, 2008; Chen and He, 2013; Somasundaran et al., 2014; Yannakoudakis et al., 2014; Phandi et al., 2015), a technique that can be labor-intensive, and, for domain scoring and source-dependent prompts, less-transferrable. This has made the creation of AES systems largely limited to assessment companies and has restricted the ability to capture meaning and content, which would be required for modeling challenging writing domains.

The extensive use of feature engineering (Zesch, Wojatzki, and Scholten-Akoun 2015) has been used to produce the most commonly cited baseline for the ASAP dataset, a model called the “Enhanced AI Scoring Engine” (EASE) which was the best open-source system from the Kaggle competition, achieving third place overall (Phandi et al., 2015).

Following this trend, the most successful publicly available models utilize each full essay but simultaneously reduce the ability to recover levels of meaning that would support natural language inference or question answering required for robust domain-scoring. Cozma et al (2018) recorded state-of-the-art results across the full ASAP dataset, but employed character-level 15-grams as string kernels and pretrained and grouped word embeddings, providing the only source of semantic information. While producing state of the art results, the model does not move practitioner-focused research meaningfully forward.

Extended writing can help reveal student learning in ways that short responses cannot. However, dealing with length and extracting meaningful inference has been challenging for researchers, and the approach has been to choose either utilizing the full essay length or trying to extract higher-level semantic features.

To reduce the expense feature-engineering, neural approaches to automated essay scoring began in 2016 using corpus-specific word embeddings with LSTMs (Alikaniotis et al., 2016; Mim et al., 2019) and with CNNs (Dong and Zhang, 2016; Dasgupta et al., 2018), and combined LSTM + CNN ensembles (Taghipour and Ng, 2016). The latter paper has become the most widely cited deep-learning AES approach and the model evaluation methods used are those still used by AES researchers, as the test dataset for the ASAP competition is not publicly available. Taghipour and Ng took one-hot word vectors as input, using a convolution layer to extract n-gram level dependencies before processing by an LSTM to capture longer distance dependencies. Using LSTMs has been very useful to capture the logical and semantic flow essays, culminating in Tay et al. (2018) implementing an LSTM model with an attention window across its hidden states, helping to address information loss across longer essays.

Transformer-based models have added great power to natural language processing tasks (Vaswani et al., 2017) by using attention instead of RNNs, eliminating temporal information loss. These breakthroughs have resulted in powerful language representations, most notably BERT (Devlin et al., 2019). AES has seen various BERT and XLNet implementations (Rodriguez et al, 2019, Yang et al.,2019, Mayfield and Black, 2020) including the current deep-learning methods SOTA (Liu et al., 2019). The method employed by Liu et al. (2019) team is very complex and computationally expensive, utilizing three separate implementations of BERT embeddings topped with LSTMs across these values and separate BERT values for the prompt, before being combined with a second model, which employs feature engineering (seemingly losing some of the purpose of neural models). Despite the extensive power of the model, it still did not outperform the model of Cozma et al. (2018).

The failure of BERT and the transformer revolution to revolutionize essay grading is due to the computational requirements of transformers and the restrictions of BERT. As Vaswani et al. (2017) note, self-attention mechanisms increase in computational complexity proportionally with the square of the length of the sequence and the dimensions of the model. BERT limits this computational expense by restricting its tokenized input to 512 characters, resulting in every BERT implementation truncating essays arbitrarily to fit the sequence length constraint. When students are demonstrating learning through writing, not all of their most important writing comes at the beginning of an essay nor can be decontextualized by restricting vocabulary. Until this paper, there has not yet been an AES model that employs self-attention without essay truncation across the entirety of the ASAP data set. AES research should explore models with more efficient self-attention of arbitrarily long documents like Beltagy, et al. (2020), which works to reduce the cost of the attention matrix computations, or begun by Nadeem et al. (2019) who implement a hierarchical attention model (Yang et al., 2016; Kowsari et al., 2017). Reducing computation or start-up costs to model training is important in supporting the objective of providing a free and open-source resource to teachers.

3 Contributions

This paper builds upon previous research to improve AES systems for practitioners, by employing techniques that use the full essay with self-attention without increasing computation by the square of the length. New contributions to AES using the ASAP dataset include:

- First application of ordinal classification structures (Chen and He, 2013) to deep learning models.
- First use multiple objective functions minimize classification loss essay scores
- First use of semantic sentence-level embeddings via the Universal Sentence Encoder (USE) (Cer et al., 2018) as the atomic vectors supplied to a deep network to capture semantic information across entire essays, allowing for processing of much longer documents.
- Automatic hyperparameter generation based on essay prompt and corpus metadata. While this data is used extensively in statistical language models and handcrafted features, it has not been applied to the task of automatically tuning hyperparameters.
- First use of multi-head attention (Vaswani et al., 2017) across sentence embeddings for AES.

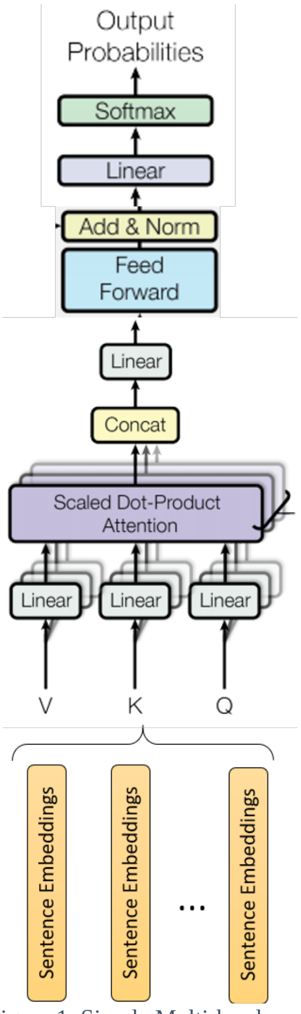


Figure 1: Simple Multi-head Attention Classifier with USE sentence embeddings.

4 Models

All experiments used USE sentence embeddings as a scalable and transferrable method for capturing complex content information from essays.

- **LSTM**: a simple LSTM model (Hochreiter et al., 1997), with a two dense layers prior to classification, was used with USE embedding inputs for comparison to baseline and for simple experimentation.
- **MHA**: Multi-head Attention, shown in figure 1, based on the

transformer model from Vaswani et al. (2017)[2], which inputs utilizing only a classifier vector from each essay in a single dense layer feed forward classifier and Luong multiplicative attention (Luong et al., 2015).

- **2MHA**: Two-layer Multi-head Attention, using the CLS vector for the single feed forward classifier.
- **MHA+BLSTM**: This model takes multiple layers of the MHA model and then uses a bidirectional LSTM to aggregate the attended features temporally prior to passing the values to a dense classifying network. This model intends to maximize the temporal information in longer essays by capturing features across all hidden outputs of the multi-head attention layers with both combined LSTM and feed forward networks.

5 Evaluation

Since the initial Kaggle competition, AES systems typically use Quadratic Weighted Kappa (QWK) on each essay set separately for evaluation, however, other measures include Pearson's correlation, Spearman's correlation, Kendall's Tau, and Cohen's Kappa. However, not all evaluation methods are practically meaningful. For example, the first LSTM model for AES (Alikaniotis *et al.*, 2016) used Cohen's Kappa across all essays and scoring systems, producing a score that does not actually "surpass similar state-of-the-art systems" in real-world results, as the paper claims. This is because essay sets that are on different scoring ranges cannot be evaluated without taking into consideration the actual possible earnable ratings. Since that time, other researchers (e.g. Dasgupta et al., 2018) have used that same metric.

Since (Taghipour and Ng, 2016), all analyses using this dataset have used both QWK and a 5-fold cross validation (60% train/20% dev/20% test) to evaluate models, since the test set used in the competition is not publicly available. As a baseline, Taghipour and Ng (2016) utilize Bayesian linear ridge regression (BLRR) as reported in (Phandi et al., 2015), as implemented in the EASE system. For comparability, this paper follows all of the same evaluation practices used with this dataset, with the exception that experiments were conducted that reduced the training and validation sets to test various models' abilities with even more limited data.

6 Experiments

In pursuit of educator-focused AES, several experiments were conducted to facilitate future research.

6.1 Setting of Hyperparameters

Aside from some values explicitly discussed below and in the interest of rapid transferability of models to novel scenarios, hyperparameters were chosen through defining relationships from known essay set metadata. Each hyperparameter that was tuned in this way is listed in Table 2 with the corresponding metadata values of the essay sets.

Hyperparameter	Metadata used to Define
Loss Weight (see 6.2)	# of classes
Dropout	# of observations, # of classes
Dimension of feed forward layer	# of classes
# of attention heads	# of classes, prompt type
Batch size	# of observations, # of classes
Epochs	# of observations, # of classes
Callback Patience	# of observations, # of classes

Table 2: Hyperparameter tuning based on AES-specific essay metadata.

Various experiments were conducted on the appropriate learning rate, however, the final learning rate used is defined using a custom schedule taken from Vaswani et al. (2017):

$$lrate = d_{model}^{-0.5} * min(step\_num^{-0.5}, step\_num * warmup\_steps^{-1.5})$$

6.2 Categorical and Regression Classification

While every other deep-learning AES study after (Taghipour and Ng, 2016) treats autograding as a regression problem, correlation may not be as meaningful for essay scoring with less than 6 labels. The use of regression as the only loss function for deep learning AES models unnecessarily restricts the amount of information contained in labels. For essays with many possible labels (like essays 7 and 8), regression was more powerful. However, state-of-the-art results were produced for narrower grading ranges by training for categorical loss.

To maximize the value of both approaches, models were trained with two objective functions: a classification loss (discussed in 6.3 below) and MeanSquaredError. The hyperparameter *P* representing the weighting of the classification loss (with *1-P* as the regression loss) is defined with a logistic function using the essay set metadata:

$$\mathrm{P_{classify}} = \frac{\mathbf{L}}{1+e^{k(n_c-\overline{n_c})}} + \mathrm{c}$$

where n_c is the number of classes in a given essay, $\overline{n_c}$ is the average number of classes across all essays, k is the slope of the logistic function (set to 0.5), and L and c are constants marking the limits of the weighting, set to 0.9 and 0.001 respectively.

6.3 Cross-Entropy vs Ordinal Classification Loss

While Categorical Cross Entropy (CCE) is reliable and powerful, the ordinal nature of essay scoring labels offers additional information about each label that can be used when calculating loss as a function related to the closeness to the actual label. A custom objective function based on

[2] Code was modified from Vaswani et al., (2017) found at www.tensorflow.org/tutorials/text/transformer

QWK[3] to include label smoothing was constructed and the comparison between it and CCE as the objective are shown in in Table 3 for the LSTM model.

Essay Set	CCE	Ordinal Loss
1	0.67	0.72
2	0.62	0.58
3	0.66	0.69
4	0.79	0.81
5	0.79	0.77
6	0.74	0.74
7	0.70	0.74
8	0.29	0.42
Avg	0.66	**0.68**

Table 2: Best QWK scores for each essay with the LSTM model (no label smoothing)

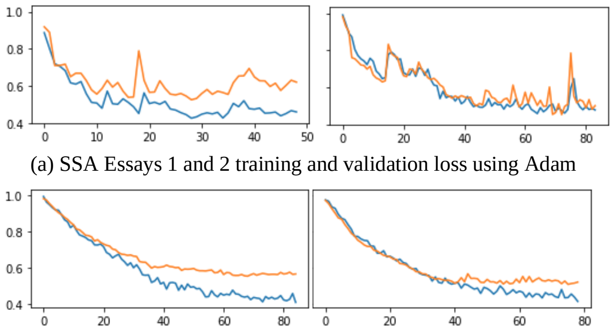
6.4 Embedding-sensitive Optimizer Selection

Selecting to represent all of the essay information through sentence-level embeddings requires maximizing the information preserved through embeddings. Small sample sizes and requiring high dropout necessitates an optimizer that provides greater numerical stability. Essays were tested using the `Adam` and `AdaMax` optimizers (Kingma et al., 2015).

The `AdaMax` algorithm implements `Adam` using the infinity norm as an approximation of L^2 norm for high dimensionality (such as embeddings), and results in a learning rate that maximizes the step taken based on the previous step and current gradient:

$$u_t = \max(\beta_2 \cdot u_{t-1}, |g_t|)$$

The difference in performance is illustrated in Fig. 2



(a) SSA Essays 1 and 2 training and validation loss using Adam

(b) SSA Essays 1 and 2 training and validation loss using AdaMax

Fig. 2. Adam vs AdaMax [α=0.001, β_1=0.9, and β_2=0.999]

6.5 Reduced Data Sets

Each of the experiments was conducted multiple times, using various portions of the already-reduced five-fold training and validation sets in order to estimate. While the results are not reported due to space constraints, the overall classification loss to the best performing models stayed above baseline at a 40% reduction in samples.

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	**0.74**	0.66
USE+2MTA	0.75	0.74	0.68	**0.83**	0.78	0.79	0.77	0.62	0.74
USE+2MTA+BLSTM	0.83	0.64	0.70	0.82	0.79	0.82	0.80	0.65	0.76

Table 3: Results reported in QWK for each essay set against current state-of-the-art and baselines.

[3] Code used to construct the classifier used the kappa_loss function from tensorflow add-ons: https://github.com/tensorflow/addons/blob/v0.10.0/tensorflow_addons/losses/kappa_loss.py

7 Results

The results of the models can be found in Table 3. Listed as the top state-of-the-art results is “Top Private Statistical NLP” which represents a composite of the highest scores across the key private companies participating in the ASAP challenge (Shermis, 2014). This composite, while it represents older data, is still significant, because it represents companies with access to significant amounts of data.

While the more complex attention model with LSTMs neared SOTA results, it had, along with all of the models, unusually poor performance on Essay 2. Investigation of losses varies greatly, but appears to be connected to the variety of approaches students took when interpreting “library censorship.”

However, SOTA performance was achieved on two different essay sets: a literary analysis of author’s purpose in a story arc and a narrative on laughter with with 60 possible labels. The latter demonstrates the power of using attention on sentence embedding, because it was able to capture meaning on the essay set that has half as many observations, is much longer on average, and has dramatically more possible labels than other essays sets. This essay best captures the complexity of creating educator-focused AES systems.

8 Conclusion

These modest results provide optimism for the use of semantic embeddings, transformer attention, and custom objective functions and hyperparameters as foundations for future research towards more meaningful educator-focused AES systems.

The work of this paper only scratches the surface of the tremendous work needed to move automated essay scoring from the realms of businesses and academic exercises into the hands of teachers and students. Much work is left to be done, but these results promise exciting possibilities for future work that could enhance the education experiences for public school students everywhere. Having free, accurate pedagogically-aligned automatic scoring options will lead to increases of more authentic assessments for improving student learning.

Acknowledgements

Thanks to Dr. Daniel Cer for guidance in interpretation of sentence embeddings and recommendations for future avenues of research.

References

[Alikaniotis *et al.*, 2016] Dimitrios Alikaniotis, Helen Yan- nakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proc. of ACL*, pages 715–725, 2016.

[Beltagy et al., 2020] Iz Beltagy, Matthew E. Peters, Arman Cohan. *Longformer: The Long-Document Transformer*. arXiv:2004.05150, 2020.

[Chen and He, 2013] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In *Proc. of EMNLP*, pages 1741–1752, 2013.

[Cer et al., 2018] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-C´espedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. *Universal Sentence Encoder*. arXiv:1803.11175v2

[Cozma *et al.*, 2018] Madalina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. Automated essay scoring with string kernels and word embeddings. In *Proc. of ACL*, pages 503–509, 2018.

[Dasgupta *et al.*, 2018] Tirthankar Dasgupta, Abir Naskar, Rupsa Saha and Lipika Dey. Augmenting Textual Qualitative Features in Deep Convolution Recurrent Neural Network for Automatic Essay Scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Ken- ton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understand- ing. In *Proc. of NAACL-HLT*, pages 4171–4186, 2019.

[Dong and Zhang, 2016] Fei Dong and Yue Zhang. Auto- matic features for essay scoring – An empirical study. In *Proc. of EMNLP*, pages 1072–1077, 2016.

[Dong *et al.*, 2017] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proc. of CoNLL*, pages 153–162, 2017.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Jin *et al.*, 2018] Cancan Jin, Ben He, Kai Hui, and Le Sun. TDNN: A two-stage deep neural network for prompt- independent automated essay scoring. In *Proc. of ACL*, pages 1088–1097, 2018.

[Ke *et al.*, 2018] Zixuan Ke, Winston Carlile, Nishant Gurra- padi, and Vincent Ng. Learning to give feedback: Model- ing attributes affecting argument persuasiveness in student essays. In *Proc. of IJCAI*, pages 4130–4136, 2018.

[Ke *et al.*, 2019] Zixuan Ke and Vincent Ng. Automated Essay Scoring: A Survey of the State of the Art. In *Proc. of IJCAI*, pp. 6300-6308, 2019.

[Kingma *et al.*, 2015] Diederik P. Kingma and, Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proc. of ICLR*, 2015.

[Kowsari et al., 2017] Kamran Kowsari, Donald E. Brownx, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E.

Barnes. *HDLTex: Hierarchical Deep Learning for Text Classification* arXiv:1709.08267v2

[Liu et al., 2019] Jiawei Liu, Yang Xu, and Yaguang Zhu. Automated Essay Scoring based on Two-Stage Learning. https://arxiv.org/abs/1901.07744. 2019.

[Luong et al., 2015] Minh-Thang Luong Hieu Pham Christopher D. Manning. *Effective Approaches to Attention-based Neural Machine Translation*, arXiv:1508.04025, 2015.

[Mayfield and Black, 2020] Elijah Mayfield and Alan W Black. Should You Fine-Tune BERT for Automated Essay Scoring?. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, 2020.

[Mathias et al, 2018] Sandeep Mathias and Pushpak Bhattacharyya, ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. In *Proc. of LREC*, 2018.

[Mathias et al, 2020] Sandeep Mathias and Pushpak Bhattacharyya, Can Neural Networks Automatically Score Essay Traits?. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, 2020.

[Mim et al., 2019] Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, Kentaro Inui. Unsupervised Learning of Discourse-Aware Text Representation for Essay Scoring. In *Proc. of ACL: Student Research Workshop*, pp.378-385, 2019.

[Nadeem *et al.*, 2019] Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated Essay Scoring with Discourse-Aware Neural Models. In *Proc 14th Workshop on Innovative Use of NLP for Educational Applications*, pages 484–493, 2019.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543, 2014.

[Persing and Ng, 2013] Isaac Persing and Vincent Ng. Mod- eling thesis clarity in student essays. In *Proc. of ACL*, pages 260–269, 2013.

[Persing and Ng, 2014] Isaac Persing and Vincent Ng. Mod- eling prompt adherence in student essays. In *Proc. of ACL*, pages 1534–1543, 2014.

[Persing and Ng, 2015] Isaac Persing and Vincent Ng. Mod- eling argument strength in student essays. In *Proc. of ACL- IJCNLP*, pages 543–552, 2015.

[Persing *et al.*, 2010] Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proc. of EMNLP*, pages 229–239, 2010.

[Phandi *et al.*, 2015] Peter Phandi, Kian Ming A Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proc. of EMNLP*, pages 431–439, 2015.

[Rodriguez *et al.*, 2019] Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. Language models and automated essay scoring. arXiv preprint arXiv:1909.09482. 2019.

[Shermis and Burstein, 2013] Mark D. Shermis and Jill C. Burstein. *Handbook of Automated Essay Evaluation: Cur- rent Applications and New Directions*. Routledge, New York, NY, 2013.

[Shermis, 2014] Mark D Shermis. State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. Assessing Writing, 20:53–76. 2014.

[Somasundaran *et al.*, 2014] Swapna Somasundaran, Jill Burstein, and Martin Chodorow. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proc. of COLING*, pages 950–961, 2014.

[Stab and Gurevych, 2014] Christian Stab and Iryna Gurevych. Annotating argument components and rela- tions in persuasive essays. In *Proc. of COLING*, pages 1501–1510, 2014.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proc. of NIPS*, pages 3104–3112, 2014.

[Taghipour and Ng, 2016] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proc. of EMNLP*, pp 1882–1891, 2016.

[Tay *et al.*, 2018] Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. SkipFlow: Incorporating neural coher- ence features for end-to-end automatic text scoring. In *Proc. of AAAI*, pages 5948–5955, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is all you need, arXiv:1706.03762, 2017.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480-1489, 2016.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages 5754–5764, 2019.

[Zesch *et al.*, 2015] Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. Task-independent features for au- tomated essay grading. In *Proc. of the BEA Workshop*, pages 224–232, 2015.