Automated Essay Scoring Using Discourse External Knowledge

Nisrine Ait Khayi[1], Vasile Rus[2]
[1] Independent Researcher
[2] University of Memphis
nisrine.khayi@mailfence.com, vrus@memphis.edu

Abstract

The Automated Essay Scoring (AES) task is an important NLP research problem given its significance for the education ecosystem. Recently, researchers started to apply a hybrid approach to this task. This hybrid approach incorporates into a deep learning model expert features that assess a particular dimension of the essay. Motivated by these successes, we propose to automatically assess essays using a hybrid approach that relies on external discourse knowledge. Our proposed model consists of using transformer-based embeddings to generate semantic representations of essays. Then, we incorporate several discourse features into these representations. Finally, we apply a linear classifier to generate the final score. To evaluate the effectiveness of this approach, we have conducted extensive experiments using the Automated Student Assessment Prize dataset (ASAP). The performance of the proposed model has been evaluated using the Quadratic Weighted Kappa (QWK) metric. The experimental results demonstrate the effectiveness of this approach in comparison with several existing solutions in literature.

1 Introduction

Automated Essay Scoring (AES) is an important educational application in Natural Language Processing (NLP). It consists of evaluating and grading the quality of written natural language essays using machine learning. By enhancing the AES systems, SDG 4 on "Quality Education" will benefit from this. Most of the research work done in the AES area is based on a holistic approach that assigns a single score to an essay [Taghipour et al.,2016; Dong and Zhang, 2016, 2017; Zhang et al.,2018, 2020]. This approach has been criticized for not being able to identify which aspects of the essays are weak and need improvement. To overcome this drawback and provide constructive feedback to learners, several researchers started to score a particular dimension of the essay such as organization [Taghipouret et al. 2017; Mathias et al. 2018; Song et al.,2020], sentence clarity [Ke et al, 2019], argument strength [Taghipour et al.,2017], style [Mathias et al., 2018] and narrative quality [Somasundaran et al.,2018]. However, little attention has been paid to evaluate the discourse aspect (i.e., conceptual structure) of the essay. Two important aspects of discourse have been emphasized so far in literature: coherence and cohesion. A coherent essay logically connected ideas that make sense to a reader. Cohesion refers to the presence or absence of linguistic cues in the text that allow the reader to make connections between the ideas in the text. Examples of these cues include conjunctions such as discourse indicators (DIs) (e.g., "because" and "for example"), coreference (e.g., "he" and "they"), substitution, ellipsis, etc. For example, I was born in Glasgow. Glasgow is the largest city in Scotland., is an example of a high cohesion text due to the explicit argument overlap related to the noun Glasgow. On the other hand, the following text, I was born in Glasgow. It is very nice in Scotland., has low cohesion as there are no explicit connections between the two sentences. One needs to make an inference linking Glasgow and Scotland in order to connect the two sentences. While cohesion defines the texture that keeps a text together (in the sense defined by Halliday and Hasan [1976]), coherence defines the overall structure and meaning of the text, i.e. the discourse. In other words, cohesion is the fabric while coherence is the outfit. According to a school of thought in Cognitive Science, the coherence, i.e. the outfit in our metaphor, is reader-dependent whereas cohesion is a property of the text, i.e., it is reader-agnostic. That is, different readers could see different outfits depending on their background relative to the text they are reading [Rus and Niraula, 2012]. To assess the discourse aspect of essays, an effective AES approach is needed.

AES task has been approached using two major methods: features engineering-based approach and neural approach. The first consists of predicting the score of an essay using handcrafted features (e.g., spelling errors, length of essay, etc.) and a simple regression model on top of that [Amorim et al.,2018; Nguyen and Litman,2018]. Although this approach has interpretability and explainability advantages, the features extraction process is tedious and expensive to achieve a high scoring accuracy. To alleviate this drawback, researchers have applied extensively deep learning models to obviate the need for features engineering [Yang et al., 2020;

Mayfield and Black, 2020; Rodriguez et al., 2019; Hirao et al.,2020; Nadeem et al., 2019]. These two approaches can be considered complementary since the handcrafted features can capture aspects that the neural approach may or may not account for and if they do, it is not clear how to know it. To get the benefit of both approaches, several researchers recently proposed a hybrid approach which consists of incorporating expert features into the neural approach [Uto et al., 2020; Liu et al. 2019; Ridley et al., 2020] The obtained results demonstrated the effectiveness of this hybrid approach which outperforms the traditional AES approaches with a significant margin.

Motivated by the successes of the hybrid approach for the AES task, we propose to add an external discourse knowledge source to evaluate essays. Our hybrid model is composed of many important components. Giving an essay, we apply transformer-based embeddings to generate its semantic representation. Next, we concatenate these embeddings with several handcrafted discourse features. First, we select some discourse features that have not been explored in the AES task such as the number of lexical chains [Morris and Hirst, 1991]. Then we select another set of discourse features from the Coh-Metrix tool [McNamara et al.,2014] that do not correlate with the first set. Finally, we feed the concatenation vector to a linear layer to predict the score. To interpret the generated transformer-based embedding features we computed their correlation with the Coh-Metrix features.

2 Related Work

To take the benefit of the AES feature engineering approach and the AES neural approach, researchers recently proposed a hybrid approach. Dasgupta and colleagues [2018] highlighted the limitations of current deep neural networks such as LSTM and CNN in identifying interconnection between the different factors involved in assessing the quality of a text. To overcome this drawback and enhance the performance on the AES task, the authors proposed a deep Convolutional Recurrent Neural Network (RNN) that processes a sequence of several handcrafted features such as lexical diversity, informativeness, cohesion, and well-formedness. Their experimental results showed that this hybrid approach can achieve state-of-the-art results. Uto and colleagues [2020] criticized the increased complexity of the previous framework because it applied RNN on the handcrafted features which can negatively affect training time. As a remediation of this issue, Uto and colleagues proposed to apply a Deep Neural Network (DNN) on essays to generate a distributed representation, then concatenate it with a handcrafted features-based vector (e.g., readability features, lexical features, syntactic features, etc.). Finally, they feed the merged vector to a linear layer to predict the final score. The authors proposed two types of DNN: 1 - a recurrent-based model such as LSTM and 2 - a transformer-based model such as BERT. This approach can be applied on other DNN- AES models easily without increasing the model complexity and it improves the performance prediction. Adopting the same approach, Liu and colleagues [Liu et al., 2019] proposed a Two-Stage Learning Framework (TSLF) which integrates both encoded features using DNN and handcrafted features. In a first stage, they proposed an LSTM based model to compute three different scores:1- a semantic score, 2- a coherence score, and 3 – a prompt-relevant score. In the second stage, the three scores are concatenated with handcrafted features (e.g., grammar errors, essay length in words and characters, vocabulary size, etc.) and then fed to a boosting tree model to predict the overall score. The experimental results demonstrated the effectiveness and robustness of the TSLF framework which outperformed many strong baselines such as CNN and LSTM on the five-eight prompts of the ASAP dataset. Ridley and colleagues [Ridley et al., 2020] highlighted the problem of cross-prompt AES for the scenario where there are no labeled target-prompt essays available for training. To alleviate this issue, Ridley and colleagues proposed a neural network combined with traditional linguistic features, avoiding the need for pseudo-labeling, the need for abundant unlabeled target-prompt essays, and the need for suitable distribution of quality in the target-prompt essays. This approach avoids overfitting to the non-target-prompt essays. Their experimental results showed that the proposed method yields state-of-the-art results. Based on those reported successes of the hybrid approach, we adopt it in this work with the additional innovation of using external discourse knowledge sources, as described briefly next. First, we extract embeddings of essays using pretrained language models. Second, we concatenate these generated embeddings with handcrafted discourse features derived from lexical chains, the Coh-Metrix tool, and others. Finally, we feed the distributed vector to a linear layer to predict the overall score.

3 Proposed Model

This section presents a more detailed description of the proposed hybrid model combined with discourse features derived from external, fully automated sources.

3.1 Essay Embeddings

The most recent pretrained language models leverage dense semantic vectors with high dimensionality. This causes sparsity issues when the text is short as it is the case of the ASAP dataset that has an average of 275 words. To overcome this drawback, we applied the compressing sentence model introduced by Zhao and colleagues [Zhao *et al.*,2022] for prompts:1-7. The authors proposed Homomorphic Projective Distillation (HPD) to learn compressed sentence embeddings by augmenting a small Transformer encoder model with a Principal Component Analysis (PCA) reduction module that reduces the dimensionality of the sentence embedding. This model proved to be effective, especially in the semantic textual similarity task with a Spearman's correlation performance gain of 2.7-4.5. Spearman's correlation is a non-

parametric measure that assesses how well is the relationship between two variables. Since prompt8 has the highest average length, we used XLNET [Yang *et al.* 2019] to generate semantic vectors of the corresponding essays. The main reason for choosing XLNET is that it includes a segment recurrence mechanism that reuses hidden states from previous segments. XLNET is a generalized autoregressive (AR) pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order.

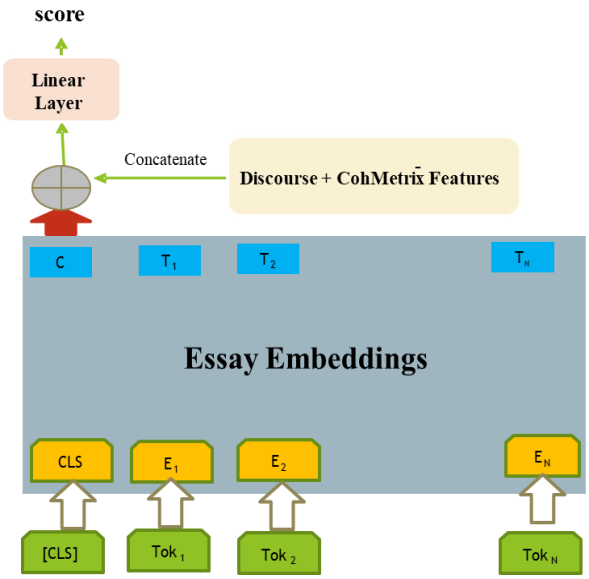

Figure 1: Model Architecture

3.2 Feature Extraction

Given an input essay of N tokens [t_1, t_2, · · ·, t_N], each token is transformed to its embedding and passed to the compressing model and XLNET. Then we collect the output of the [CLS] token and use it as a text representation of the essay. Then, we concatenate the resulting H vector with a set of discourse handcrafted features.

3.3 Discourse Features

This component encodes the discourse features of essays. It is derived from lexical chains. To be noted, this is the first time the lexical chains are explored in the AES task. Lexical chains represent sequences of semantically related words in a text. They have been used as an indicator of text cohesion [Morris et al., 1991]. Intuitively, an essay that contains many lexical chains, especially ones where the beginning and end of the chain cover a large span of the essay, tends to be more cohesive [Somasundaran et al. 2014]. The essay's scoring can be improved by incorporating lexical chains since cohesion is an important dimension of essay quality. The reason is that cohesion correlates positively with similarity. So, these lexical chains can be used to augment the essays' embeddings. Lexical chains are computed as follows: First, we create a list with all the relations of each noun. Then, we compute the lexical chain between each noun and their relation and apply a threshold of similarity between each word. Finally, we prune the lexical chain, deleting the chains that are weaker with just a few words. In this work, we consider the following lexical chains features:

- Average chain size: the average size of all the generated lexical chains.
- Number of varied chains: the number of chains that have distinct words.
- Number of large chains: the number of chains of a minimum size of 4 words.
- Number of large, varied chains: the number of chains of a minimum size of 4 distinct words.
- Percentage of large chains: the percentage of large chains in the total number of chains.
- Percentage of large varied chains: the percentage of the chains with a minimum size of 4 distinct words.

Additionally, we consider grammatical errors as a discourse measure [Burstein et al., 2013]. This feature addresses errors in grammar that could interfere with a reader's ability to construct meaning. We also consider word unigrams, bigrams, and trigrams as they encode discourse information about essays [Ke et al., 2019]. For instance, the bigram" people is" suggests ungrammaticality, the use of discourse connectives (e.g.," moreover"," however"). The key advantage of using n-grams as features is that they are language-independent as they can be applied to a new language with no additional effort.

3.4 Coh-Metrix Features

Coh-Metrix is a language analysis tool that assesses texts via cohesion, coherence, and readability. It provides 110 metrics that are classified into 11 groups:

1. *Descriptive;* used to check patterns in the text such as number of paragraphs, sentences, and words.
2. *Text Easability Principal Component Scores;* provide an evaluation of the text ease based on the linguistic characteristics of the text. They are also aligned with theories of text and discourse comprehension.
3. *Referential cohesion,* which assesses the number of cohesion relations that a human reader could do based on the propositions and sentences of the text.
4. *Latent Semantic Analysis,* which measures the semantic overlap between sentences and paragraphs. The scores range from 0 (low cohesion) to 1 (high cohesion).
5. *Lexical Diversity*, which measures the type to token ratios to infer the level of cohesion.
6. *Connectives*, an index which counts the incidence of connectives in a text.
7. *Situation Model*, which has been used in discourse processing to refer to the mental representation of a text involving much more than explicit words.
8. *Syntactic Complexity*, which is about syntactically

analyzing sentences and word density.

9. *Syntactic Pattern Density*, which assesses the incidence of different types of syntactic patterns in the texts.
10. *Word Information*, which measures the word type density in the text.
11. *Readability*, which assesses the text readability with formulae such as Flesch Reading Ease and Flesch-Kincaid Grade Level [Graesser *et al.,* 2005].

3.5 Linear Layer

We treat the AES task as a regression task. We use the following scoring function to map H to a scalar value by the ReLU activation function.

$$y = ReLU\ (w.\ H + b) \qquad (1)$$

where w is the weight vector, b is the bias and y is the final score.

4 Experiments and Results

To evaluate the performance of our proposed model, we have conducted several experiments using the Automated Student Assessment Prize (ASAP) dataset.

4.1 ASAP Dataset

The Automated Student Assessment Prize dataset consists of eight prompts for different topics, including narrative essays where the prompts require to narrate a story, source-dependent response essays where the writer responds to a question and argumentative essays where the writer has to convince the reader. In total, there are nearly 13,000 essays in the dataset. Each of the essays were written by high school students, grades 7 to 10. Each essay is assigned a score given by the instructors. Each of the eight prompts has its unique characteristics and different score resolution methods. The score range of each prompt thus differs. The following table displays key statistics of the dataset.

Prompt	# of Essays	Avg. Len	Ranges
1	1783	350	2-12
2	1800	350	1-6
3	1726	150	0-3
4	1772	150	0-3
5	1805	150	0-4
6	1800	150	0-4
7	1569	250	0-30
8	723	650	0-60

Table 1. Statistics of the ASAP dataset; Ranges means the score ranges

4.2 Experimental Settings

We performed all our experiments using a Tesla K80 GPU and a total of 12 GB of RAM. The model was implemented using the Hugging Face's library. We used the base version of the compressing model for prompt 1-7, and XLNET for prompt 8. The maximum sequence length of the compressing model is changed per prompt. The Adam optimizer [Kingma et al., 2014] with a learning rate of 1e-5 was used and the gradients were clipped to 1.0 to prevent exploding gradients. We evaluated our model using the Quadratic Weighted Kappa (QWK). About 80% of the data was used for training and 20% for testing. Each experiment was repeated several times, and we selected the best model. Before running the main experiment, we run the Coh-Metrix tool on the ASAP data to extract 110 features. Then, we compute the correlation between these extracted features and the discourse features (e.g., number of lexical chains, unigrams, bigrams, etc.) described previously to discard the correlated ones. We consider a p-value with a threshold of 0.8. Then we preprocess the text essays by removing the usernames, Nan values, punctuation and stop words. We normalize promt8 scores and we scale all the features prior to training and testing.

4.3 Evaluation Metric

Similar to prior works, we use the Quadratic Weighted Kappa (QWK) metric to evaluate the performance of our proposed method. It measures the agreement between calculated scores and gold/human expert ones. First, we compute the weight matrix following this formula:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \qquad (2)$$

where i and j are the golden scores and calculated scores respectively and N is the number of possible ratings.

Second, we compute the QWK score following this formula:

$$k = \frac{\sum_{i,j} W_{i,j} O_i}{\sum_{i,j} W_{i,j} E_{i,j}} \qquad (3)$$

where Oi,j is the number of essays which obtained a rating i by a human annotator and a rating j by the AES system. And the matrix E is calculated as the outer product of histogram vectors of the two ratings. The matrix E is then normalized by the sum of elements in E.

4.4 Experiments Results

Table 2 displays the empirical results of our proposed model on the ASAP data as well as the results of other models that have been reported in the literature. We show the QWK scores for each prompt and then we average the scores. As shown in the table, our proposed model outperforms several existing AES systems, in terms of the average score of the Quadratic Weighted Kappa, such as the baselines LSTM, CNN, and logistic regression. It also outperforms BERT with an improvement of 3%. The results also demonstrate that incorporating external knowledge into deep learning models increases the average performance significantly. For

example, adding features to the LSTM increases the average QWK score from 0.55% to 0.72%. We can observe the same performance improvement when we added the discourse features to our proposed model with 4% in the average QWK score. Furthermore, our proposed model reached the highest scores for prompt6 and prompt8 with QWK scores of 0.81% and 0.79% respectively. This result aligns with the results of Mathias and Bhattacharya [Mathias and Bhattacharya, 2018] that cohesion and coherence are the most important features of prompt6 and prompt8.

Model	P1	P2	P3	P4	P5	P6	P7	P8	Avg
LSTM	0.37	0.40	0.51	0.77	0.76	0.76	0.63	0.17	0.55
CNN	0.80	0.65	0.63	0.76	0.75	0.76	0.75	0.68	0.72
XLNET	0.77	**0.68**	0.69	0.80	0.78	0.79	0.78	0.62	**0.74**
BERT	**0.82**	0.39	**0.76**	**0.88**	**0.87**	0.58	0**.81**	0.54	0.71
Log R	**0.82**	0.64	0.66	0.70	0.78	0.67	0.72	0.60	0.70
LSTM+F	0.80	0.62	0.60	0.77	0.77	0.77	0.76	0.64	0.72
XHPD	0.70	0.62	0.64	0.75	0.72	0.71	0.71	0.73	0.70
XHPDF	0.76	0.63	0.70	0.78	0.73	**0.81**	0.78	0**.79**	**0.74**

Table 2. Experiments Results

Encoded Embeddings Interpretation

To understand the nature of the encoded embeddings generated from the XLNET and the compression sentence model, we generated the encoded embeddings of different dimensions using the testing ASAP dataset which consists of 270 text essays. Then, using the Pearson Coefficient, we computed a correlation score between every feature of each embedding matrix (compressed embeddings matrix for prompts 1-7 and XLNET embeddings matrix for prompt 8) and the Coh-Metrix features extracted from the same dataset. We considered a threshold of p-value of 0.7. We found that every encoded feature correlates with at least one Coh-Metrix feature or 15 features at most. This explains the discourse nature of the XLNET and compressing embeddings.

Conclusion

We presented a hybrid model that incorporates external discourse knowledge sources, which have not been explored extensively in the literature, for the AES task. Our proposed method consists of extracting semantic features from essays using pretrained language models and concatenating them with external features capturing the discourse aspect of essays in terms of cohesion and coherence. To evaluate the performance of our proposed approach, we have conducted several experiments. The experimental results reinforce the effectiveness of incorporating an external knowledge into deep learning models yielding competitive results for the AES task.

In the future, we are planning to overcome the shortcomings of XLNET in processing longer sequences than 512 tokens by using other transformers leveraging higher maximum length (e.g., "BigBird"). Moreover, we are planning to assess the discourse of essays in a manner consistent with human expert raters and explore other types of essays with more accuracy and logical consistency to improve the generalization capability of the model.

References

[Amorim *et al*.,2018] Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated Essay Scoring in the Presence of Biased Ratings. *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*New Orleans, Louisiana: Association for Computational Linguistics

[Burstein et al., 2013] Jill Burstein, Joel Tetreault, and Martin Chodorow. 2013. Holistic Discourse Coherence Annotation for Noisy Essay Writing. Dialogue and Discourse, 4(2).

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

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

[Dong *et al.*,2017] Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-Based Recurrent Convolutional Neural Network for Automatic Essay Scoring. *In Proceedings of the Conference on Computational Natural Language Learning*. Vancouver, Canada: Association for Computational Linguistics.

[Graesser *et al.*,2005] Arthur C. Graesser and Sarah Petschonek. 2005. Automated Systems that Analyze Text and Discourse: QUAID, Coh-Metrix, and AutoTutor. Advancing Health Outcomes Research Methods and Clinical Applications, McLean, VA: Degnon Associates.

[Halliday and Hasan, 1976] Michael Halliday and Ruqaiya Hasan. 1976. Cohesion in English London: Longman.

[Ke and Ng, 2019] Zixuan Ke and Vincent Ng.2019. Automated Essay Scoring: A Survey of the State of the Art. *In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.*

[Ke et al.,2019] Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and

Vincent Ng. 2019. Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

[Kingma et al.,2014] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980.

[Liu *et al*.,2019] Jiawei Liu, Yang Xu, and Yaguang Zhu. 2019. Automated Essay Scoring Based on Two-Stage Learning. *ArXiv, abs/1901.07744.*

[Mathias *et al*.,2018] Sandeep Mathias and Pushpak Bhattacharyya. 2018. Thank "Goodness"! A Way to Measure Style in Student Essays. *In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications.* Association for Computational Linguistics.

[Mayfield and Black,2020] Elijah Mayfield and Alan W Black. 2020. Should You Fine-Tune BERT for Automated Essay Scoring?. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Seattle, WA: Association for Computational Linguistics.

[Mathias and Bhattacharya, 2018] Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. *In Proceedings of the International Conference on Language Resources and Evaluation.*

[McNamara et al., 2014] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press.

[Morris and Hirst,1991] Jane Morris and Graeme Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. Computational Linguistics 17(1).

[Nguyen and Litman,2018] Huy V. Nguyen and Diane J. Litman. 2018. Argument Mining for Improving the Automated Scoring of Persuasive Essays. *In Proceedings of the AAI Conference on Artificial Intelligence*, New Orleans, Louisiana: Association for the Advancement of Artificial Intelligence.

[Ridley et al.,2020] Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt Agnostic Essay Scorer: A Domain Generalization Approach to Cross-Prompt Automated Essay Scoring. *ArXiv, abs/2008.01441*.

[Rodriguez et al.,2019] Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and Automated Essay Scoring. *ArXiv, abs/1909.09482.*

[Rus and Niraula, 2012] Rus, Vasile, and Nobal Niraula.2012. "Automated Detection of Local Coherence in Short Argumentative Essays based on Centering Theory." *Computational Linguistics and Intelligent Text Processing: 13th International Conference,* CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I 13. Springer Berlin Heidelberg, 2012.

[Somasundaran et al.,2018] Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards Evaluating Narrative Quality in Student Writing. Computational Linguistics 6(0).

[Song et al., 2020] Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu.2020. Hierarchical Multi-task Learning for Organization Evaluation of Argumentative Student Essays. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20,* 3875–38. International Joint Conferences on Artificial Intelligence Organization.

[Taghipour *et al.*, 2016] Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 1882–1891. Austin, Texas: Association for Computational Linguistics.

[Uto et al.,2020] Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural Automated Essay Scoring Incorporating Handcrafted Features. *In Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Yang et al.,2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNET: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32.

[Yang et al.,2020] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing Automated Essay Scoring Performance via Fine-Tuning Pre-trained Language Models with Combination of Regression and Ranking. *In Findings of the Association for Computational Linguistics: EMNLP 2020,* pages 1560–1569, Online. Association for Computational Linguistics.

[Zhang et al.,2018] Haoran Zhang and Diane Litman. 2018. Co- Attention Based Neural Network for Source-Dependent Essay Scoring. *In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Application*s, 399–409. New Orleans, Louisiana: Association for Computational

Linguistics.

[Zhang et al.,2020] Haoran Zhang and Diane Litman. 2020. Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8569–8584. Seattle, WA, USA (Virtual Event): Association for Computational Linguistics.

[Zhao et al.,2022] Xuandong Zhao, Zhiguo Yu, Ming Wu, and Lei Li. 2022.Compressing Sentence Representation for Semantic Retrieval via Homomorphic Projective Distillation. *arXiv preprint arXiv:2203.07687.*

.