Automated essay scoring: A review of the field

Paraskevas Lagakis
Software and Interactive Technologies (SWITCH) Lab,
Computer Science Department
Aristotle University of Thessaloniki
Greece
plagakis@csd.auth.gr

Stavros Demetriadis
Software and Interactive Technologies (SWITCH) Lab,
Computer Science Department
Aristotle University of Thessaloniki
Greece
sdemetri@csd.auth.gr

2021 International Conference on Computer, Information and Telecommunication Systems (CITS) | 978-1-6654-4913-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/CITS52676.2021.9618476

***Abstract*—This paper critically reviews the recently published scientific literature on the task of Automated Essay Scoring (AES), by examining the various systems and approaches used. Automated Essay Scoring, which is the process of automating the evaluation of answers to open-ended questions, most usually in educational settings, by utilizing NLP techniques, gathers an increasing amount of interest, due to its potential applications both commercially and as part of the learn-ing process. The focus of this paper is to analyze the most popular approaches in recently published AES systems, categorize the systems with respect to certain characteristics in their design, the datasets that they use and the evaluation schemas that are used to evaluate them, and finally discuss the recent trends and challenges of the field of AES.**

Keywords—Automated Essay Scoring, Automated Essay Grading, Automated Essay Evaluation, Natural Language Processing

I. INTRODUCTION

AES was initially introduced as a concept in 1966 by E. B. Page with his work on computer aided grading systems. Currently, AES is considered one of the most prominent academic applications of Natural Language Processing (NLP), where artificial intelligence is used to score any written document. Most AES researchers were attracted to perfecting the system to rate the quality of any essay with a single score. Holistic scoring was popular as a research field for two reasons. The main one being the abundancy of manual holistic scores available online that created fertile ground for holistic scoring systems to "learn" from a massive public database. Secondly, holistic scoring mechanisms were in high commercial demand. Creating a system that can automatically score thousands of essays in a short amount of time can save a lot of human effort, especially for institutions that provide standardized proficiency test certifications such as the Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT).

1) Survey methodology

This paper's goal is to document an overview of the AES systems literature, outlining both the milestones of AES from the creation of the field in 1966 as well as current trends, practices and developments in modern AES systems. While there are books [1, 2] and recent articles [3] published, that summarize the evolution of AES systems and the corresponding research up until then, the recent developments in the use of transformer models that have been applied in AES systems, have been rapid. For that reason, this paper has been written to also include the recent AES systems that use transformer models to provide state-of-the-art results.

To research the literature, we used Google Scholar with the search terms "Automated Essay Scoring", "AES" and "Automated Essay Evaluation". For the purposes of this paper, any systems that implemented automated evaluation for items other than essays (e.g. coding exercises or selected-response questions [4]) were not included in the review.

The literature of AES systems has been reviewed and a number of their distinguishing characteristics have been identified and analyzed, like the model of the system and the features used, the most common datasets used to train the models, the scoring methods and whether they are holistic or dimensional, and finally their evaluation metrics.

II. AES SYSTEMS

The next step is to use 4 characteristics to distinguish the existing AES systems; the methods they use in regards to features and the model they use for training; the scoring methods they use to evaluate essays; the datasets used for training and finally the metrics they used for evaluating their efficacy.

A. Features & Models

Most AES systems in literature are categorized in regards to the features they use, into systems that use feature-engineered models with handcrafted features, and systems with neural approaches, that can extract features automatically. The handcrafted features are easier to grasp and understand, and can be modified to better enable evaluating specific aspects of an essay. Also, this specificity can be used to provide feedback to the author of the essay.

1) Handcrafted features

Early AES systems where developed with handcrafted features in mind. For example the pioneering Project Essay Grader TM (PEG) from Page [5], which is considered to be the first AES system that established the field, used an approach of supervised learning with linear regression, and made use of features based on length, such as the average length of words in the text and the length of the text, as well as lexical features, like features that encode the number of occurrences of a particular punctuation in a text. Some of the other well-known systems that were developed early on, with handcrafted features approaches, include IEA, E-rater, IntelliMetric and BETSY. Table I presents some common handcrafted features [6, 3].

TABLE I. COMMON HANDCRAFTED FEATURES

Feature	***Description***
Length-based	Numbers of characters, words, sentences, and punctuation symbols. Average word lengths.
Syntactic	Numbers of nouns, verbs, adverbs, adjectives, and conjunctions. Parse tree depth. Grammatical error rates.
Word-based	Numbers of useful n-grams and stemmed n-grams. Numbers of spelling errors, sentiment words, and modals.
Readability	Numbers of difficult words and syllables.
Semantics	Semantic similarity based on latent semantic. Histogram-based features
Argumentation	Numbers of claims and premises. Argument tree depths.

978-1-6654-4913-7/21/$31.00 ©2021 IEEE

Feature	*Description*
Prompt-relevance	Number of words in essays for a prompt.

Lexical features have also been used together with syntactic features like parse trees [7] in their system, which was based on a ranking approach with the use of LambdaMART, used parse tree depth as a representation of the syntactic complexity of the sentences in a text.

Other broadly used handcrafted features have to do with predefined dictionaries, that contain lists of words separated in certain categories based on semantic, syntactic, grammatical or sentiment criteria. For instance, features are computed based on lists containing discourse connectives, correctly spelled words, sentiment words, and modals [8–12], as the presence of certain categories of words in an essay could reveal a writer's ability to organize her ideas, compose a cohesive and coherent response to the prompt, and master standard English.

[13, 8], which used a ranking approach using SVMs, also made use of features extracted from the grammatical error rates of a text. To compute the rates, manually annotated errors and error types were used.

2) Automated feature extraction

The other category of AES systems consists of models that erase the need for feature engineering, by following neural approaches. Such models are [14–19].

For instance, [14] is such a model, that automatically creates the features needed, through a model that has as an input the one-hot vectors of the words of the essay that needs to be evaluated, which is then fed to a convolution layer that extracts n-gram level features. These features then are also passed through a recurrent layer of a Long-Short Term Memory (LSTM) network that extracts a second vector of features. The n-gram level features refer to the local relations between words whereas the second vector refers to the long-distance relations between words in the essay. Together they act as input to a third dense layer that outputs the final holistic essay score. [19] also use an LSTM based approach, but add another layer to try and extract coherence features of an essay.

There are similar approaches like [15] that instead of one-hot word vectors use word-embeddings, that are vector representations of words that, through training, show semantic similarities between them. In [15], the proposal was made to use score-specific word embeddings (SSWEs) instead of training the embeddings to find similarities in unannotated corpora, as was the case before.

Other automated feature extraction models, like [18], proposed the use of the structure of an essay in two stages, the word-level and the sentence-level, with each one acting as input on their own convolution layer. The word-level layer has each sentence in one-hot vectors as word input and extracts n-gram features without the consideration of the other sentences. Then, after a pooling layer, these features are combined into a vector that represents the whole sentence, and these vectors act as input to the sentence-level layer. [18] use an attention mechanism to improve the above mentioned approach.

Both these two general approaches that were analyzed above, the models that use handcrafted features and the ones that automate feature extraction through neural approaches, have their advantages and disadvantages. Neural approaches with automatic feature engineering are valuable at extracting semantic features that are otherwise very hard to model by hand, but in order to be effective, they require large datasets with hand-annotated data that can be difficult to acquire, as we will see in the 2.3 Section, especially for languages other than English. Also, these approaches are more appropriate when we need to extract a holistic score of an essay, whereas approaches with hand-crafted features can be better used in dimension-specific AES systems. Hand-crafted features like grammatical and spelling errors are always going to be very useful to the effectiveness of an AES system, and are more transparent in how they work, and can provide useful feedback to the author of an essay. That's why recent AES systems try to incorporate both these worlds and create hybrid approaches that can integrate the advantages of the two different aspects of feature engineering.

3) Transformer-based models

One of the current trends in NLP research in general, is to use transformer-based models with fine-tuning in tasks having to do with supervised learning, with more traditional linear models losing popularity.

More specifically, the use of deep neural networks has been widely popularized in the NLP community in recent years, and in particular, the use of the transformer architecture that was introduced in 2017 [20], and established by BERT (Bidirectional Encoder Representations from Transformers) [21] has been the current trend, replacing older recurrent neural network (RNN) models such as the long short-term memory (LSTM).

Since the Transformer model enables better parallelization during training, it offers the option of training on datasets larger in size than before. Transformer models make use of those huge datasets of existing general text data, such as Wikipedia Corpus and Common Crawl, to pretrain multilayer neural networks with context-sensitive meaning of, and relations between, words, such as BERT and GPT (Generative Pre-trained Transformer). The models are then fine-tuned to a specific new labeled dataset and used for a variety of tasks with the most common among them being classification or structured prediction tasks.

As a result, in the field of AES research, that can be considered as a set of NPL problems, transformer models are an appropriate model for consideration. However, current efforts of using transformer models in AES systems provide similar performance to classical models, but at a significant additional cost in computational resources. Research [22, 23] suggests though, that transformers have unique characteristics that may provide benefits over classical methods, in order to justify the extra computational costs of running them. [24] also suggests that researchers should be encouraged to use hand-crafted features along with deep encoded features for better results. Finally, research attempts [25] have been made to limit the computational requirements of such models, without compromising significantly in the evaluation results.

One of the AES systems that incorporates transformers is the Two-Stage Learning Framework (TSLF) [26]. The model used in this system has two components, one utilizing the pre-trained BERT model to derive sentence embeddings (using the un-cased model with 12-layer, 768-hidden, 12-heads and 110M parameters) which are then fed as input to an RNN, while the second component uses hand-crafted features. This

hybrid approach has provided some promising results for TSLF. The BERT sentence representations are used to learn an essay score, a prompt-relevance score and a "coherence" score, trained on original and permuted essays. Document representations from the neural network and the hand crafted features are then used together in a gradient-boosting decision tree to predict the final essay score.

[27] demonstrate the relevance of discourse-aware structures and discourse-related pretraining, in the development and performance of a neural network AES system. Their system revolves around two approaches of neural models to map a written essay into a vector, which is used with ordinal regression to evaluate the essay holistically. Both models are based on LSTM, and incorporate document structure. The first one is HAN [28] and the second one is Bidirectional context with attention (BCA) [29], which extends the previous by checking for dependencies between sentences. A vector for the context similarities of the previous and the next sentence words is computed for each separate word of a sentence, and the final representation of each word takes those context vectors into account, making this model discourse aware. Finally, they used pretraining with the BERT embeddings and made the hypothesis that the next sentence prediction task would identify discourse coherence aspects of the essay.

Another AES system that utilizes the BERT model is R2BERT [24] , which proposes a new method called multi-loss for the task of fine-tuning BERT models in the usage of AES systems, and combines regression and ranking models in these tasks, with their experimental results indicating improvements in performance by using the multi-loss approach.

[30] is a recently published attempt at a system that uses a deep neural network (DNN) framework combined with item response theory (IRT) [31], with the goal being to avoid the bias of the human raters. This approach can be especially useful in low or medium stakes tests, where the training data might be of lower quality. The framework tests the efficacy of IRT combined with both a more "traditional" AES model of convolution neural networks with LSTM, and also with the more recent approach of using BERT.

[32] also tried a hybrid AES system that uses DNNs together with hand-crafted essay-level features, again trying multiple methods, like LSTM & BERT, with BERT being the one with the best experimental results.

B. Scoring Method

As mentioned in the Introduction, until 2004 the main task for AES systems was to aid to holistic scoring. Although the importance of holistic scoring is indisputable, AES systems could potentially have a wider impact if they could help students im-prove their essay writing skills while providing them with some sort of feedback. A mere low score could not possibly help a student understand the reason for scoring weakly neither would it show the areas where the student could improve and how. In light of this deficiency, researchers have begun anew to work on scoring specific aspects of text quality such as coherence [33, 34], technical mistakes, and relevance to prompt [35, 36]. AES systems that offer didactic feedback along multiple dimensions of essay quality have also been developed in the last 20 years, for example Criterion [37].

There are many aspects or dimensions of an essay that a machine can accurately evaluate, but most of them refer to the technical aspects of an essay, like the use of correct spelling and grammar, punctuation and word usage. However, the aspects of an essay that refer to its content quality and are an important part of a human grader evaluation, are still very difficult to be effectively estimated by an AES system. For example, even the most elaborate AES systems currently have trouble accurately evaluating aspects like coherence and persuasiveness of an essay, or an argument's clarity, and that is due to two main reasons: the first is the fact that such aspects are very hard to model, and the second has to do with the lack of datasets that provide human-graded dimension-specific data, especially compared with those that provide holistic scoring. These kinds of qualitative dimensions of an essay are still under research.

Dimension-specific scoring as a task, was developed at a later date, which, as shown in more detail in Section 3, examined the organization [38], thesis clarity [39], argument persuasiveness [40, 41], relevance to prompt [36], and coherence [34] of any given essay.

C. Datasets

In this section, 5 of the most broadly used English datasets that are known for training AES mechanisms will be presented. A limited number of AES datasets exist in other languages but these will not be analyzed any further in this paper.

The first corpus is the Cambridge Learner Corpus-First Certificate in English (CLC-FCE) test that provides to all examiners not only the final holistic score of their essay, but also a manually labeled grammatical error detection, such as unfitting tense [13]. This makes it possible to build systems not only for holistic scoring but also for morphological error detection and correction that might as well provide some sort of feedback to the author. Nonetheless, the CLC-FCE test system provides a fairly limited amount of essays per prompt which makes it troublesome to build prompt-specific engines of high efficiency (i.e., systems that "learn" from the same prompt).

TABLE II. ASAP DATASET

Prompt	*No. of Essays*	*Scoring range*	*Avg. essay length*
1	1783	2 – 12	350 words
2	1800	1 – 6	350 words
3	1726	0 – 3	150 words
4	1772	0 – 3	150 words
5	1805	0 – 4	150 words
6	1800	0 – 4	150 words
7	1569	0 – 30	250 words
8	723	0 – 60	650 words

The second corpus is the Automated Student Assessment Prize (ASAP) competition on essay scoring that was first released as part of a Kaggle competition in 2012 and sponsored by the Hewlett Foundation. In comparison to the CLC-FCE, this corpus was used extensively for holistic scoring. The reason for its wide usage was not only the massive amount of total number of essays written but also the

number of essays per prompt that summed up to 3000. Such a generous database makes it easier to build high-performance prompt-specific systems. This corpus however, has two main drawbacks that make it less reliable in some cases. One of them being that the score limits are not universal for all prompts, making it difficult to train a model on various prompts. Secondly, the essays have crucial alterations from the original scripts. For example, paragraphs are not distinguished from one another and capitalized letters are being erased.

The TOEFL11 corpus contains essays from the Test of English as a Foreign Language (TOEFL) exam [42]. These essays are equally distributed over 8 prompts and just as the title predisposes, there are 11 native languages spoken by the authors. The corpus was initially created for the Native Language Identification (NLI) task, to determine an author's native language based on their writings in a second language. However, there were only three proficiency levels defined by the corpus: Low, Medium and High. There has been an attempt to train AES systems on these three labels, applying them in the model as if they were holistic scores. Yet, the fundamental hypothesis that an essay's quality can be depicted by the language proficiency of its author is open to doubt.

An issue that affects the evolution of dimension-specific essay scoring research, involves the inadequacy of manually annotated corpora with dimension-specific scores. In order to overcome this issue, effort has been made to manually annotate multiple essays from the International Corpus of Learner English (ICLE) [43] along various dimensions of essay quality, such as (1) Organization, which applies to the level of organization throughout an essay [38]; (2) Thesis Clarity, which applies to how evidently the thesis of an author is explained within his/her essay [39]; (3) Prompt Adherence, which applies to how relevant an essay's substance is to the prompt for which it was written [36]; and (4) Argument Persuasiveness, which applies to the effort of the author to make a convincing argument [40].

The final corpus related to dimension-specific scoring that will be mentioned in this paper, is the Argument Annotated Essays (AAE) [44]. The corpus consists of essays captured from essayforum2; a site that provides feedback to students willing to upgrade their persuasive written abilities in order to claim higher scores in tests and exams. All 402 essays taken from the site were evaluated according to their persuasive structure.

D. Evaluation Metrics

Finally, this section refers to the evaluation metrics that are widely used in AES systems. The Quadratic weighted Kappa (QWK) is the go-to metric for most current AES systems, and is what we used in Table 3 to compare current state-of-the-art AES systems that were mentioned above and utilize the transformer models. It's an agreement metric with a range from 0 to 1. Other metrics that are often used are error metrics such as Mean Absolute Error (MAE) and Mean Square Error (MSE) and correlation metrics such as Pearson's Correlation Coefficient (PCC) and Spearman's Correlation Coefficient (SCC).

[45] analyzes the advantages and disadvantages for a variety of metrics in AES systems, including what we mentioned here. Finally, the usual evaluation schemas in AES systems are the in-domain evaluation, where an AES system is trained and evaluated on the same prompt and the average of all prompts is the final metric of its performance, and a cross-domain evaluation, where a system is trained and evaluated on different prompts (usually adopted by AES systems that perform transfer learning).

TABLE III. PERFORMANCE OF STATE-OF-THE-ART AES SYSTEMS THAT UTILIZE TRANSFORMER MODELS

System	***Dataset***	***QWK***
TSLF - (Liu et al., 2019)	ASAP – Prompt 1	0.852
	ASAP – Prompt 2	0.736
	ASAP – Prompt 3	0.731
	ASAP – Prompt 4	0.801
	ASAP – Prompt 5	0.823
	ASAP – Prompt 6	0.792
	ASAP – Prompt 7	0.762
	ASAP – Prompt 8	0.684
(Nadeem et al., 2019)	TOEFL – Split 1	0.729
	TOEFL – Split 2	0.715
	ASAP – Prompt 1	0.840
	ASAP – Prompt 2	0.711
R^2BERT - (Yang et al., 2020)	ASAP – Prompt 1	0.817
	ASAP – Prompt 2	0.719
	ASAP – Prompt 3	0.698
	ASAP – Prompt 4	0.845
	ASAP – Prompt 5	0.841
	ASAP – Prompt 6	0.847
	ASAP – Prompt 7	0.839
	ASAP – Prompt 8	0.744
BERT + Essay-level features - (Uto et al., 2021)	ASAP – Prompt 1	0.852
	ASAP – Prompt 2	0.651
	ASAP – Prompt 3	0.804
	ASAP – Prompt 4	0.888

III. RESULTS

In this section, we briefly present the results of the transformer based systems that have been published lately. AES systems with more traditional approaches that were mentioned here, be it systems that utilize hand-crafted features or neural network systems with automated feature extraction of previous years have already been thoroughly presented in [3] so they will not be part of our evaluation here.

The results that are presented in Table III are referring to 4 novel AES systems that use transformers to some extent. To the best of our knowledge, the combination of BERT and hand-crafted features for holistic scoring of essays, provides the best results and forms the current state-of-the-art.

These conclusions are in agreement with what we mentioned before, that the use of hand-crafted features still plays a crucial role for AES systems and will continue to do so for the foreseeable future, but can be combined with novel

approaches to fine-tune the results, and that hybrid systems that correctly utilize them still provide the best results.

IV. Summary

In summary, we examined that the field of AES has, since its inception more than 50 years ago, made a lot of progress, following the developments in the more general research field of NLP. Be that as it may, as we mentioned in previous chapters, the core fundamental approaches of AES, like hand-crafted features used in AES systems, are still valid today, and should not be ignored in future work, but rather they should be utilized to fine-tune and further improve the results of more advanced and novel approaches, like neural models and transformers.

The latest trend of using transformer models for a variety of NLP tasks, and the fact that transformers have rapidly been widely adopted as the model of choice for NLP problems, have not left AES without a mark. On the contrary, most of the recent research work has been focused on how to best utilize transformers for evaluating essays, with promising results when combined with more traditional approaches. Thus we believe that a promising direction for future AES systems is to first use transformers to pretrain deep bidirectional representations from the abundance of unlabeled data, and then fine-tune the results with the scoring output layer, while also expanding the initial input with hand-crafted features.

Another interesting topic that is still under-researched is exploring the scoring of different aspects of an essay. This topic includes a variety of sub-directions that present research interest, like data annotation with respect to specific essay aspects in contrast with the more often used holistic scoring, or the provision of automated feedback on an essay, which is of great value for more educational applications of AES. Especially as technological developments, remote working and latest changes in our living conditions like the COVID pandemic call for remote learning platforms like MOOCs that can support massive amounts of students, the idea that scalable feedback-providing automated solutions can be utilized to help improve such platforms, is of great interest, and aspect-specific AES systems can provide a more explainable and transparent way of evaluating written text and providing feedback to the author.

References

[1] (2003) Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US

[2] (2013) Handbook of automated essay evaluation: Current applications and new directions. Routledge/Taylor & Francis Group, New York, NY, US

[3] Ke Z, Ng V (2019) Automated essay scoring: A survey of the state of the art. IJCAI International Joint Conference on Artificial Intelligence August 2019:6300–6308.

[4] Isaacs T, Zara C, Herbert G, Coombs SJ, Smith C (2013) Key Concepts in Educational Assessment. SAGE

[5] Page EB (1966) The Imminence of... Grading Essays by Computer. The Phi Delta Kappan 47:238–243

[6] Phandi P, Chai KMA, Ng HT (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. Association for Computational Linguistics, Lisbon, Portugal, pp 431–439

[7] Chen H, He B (2013) Automated Essay Scoring by Maximizing Human-Machine Agreement. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Seattle, Washington, USA, pp 1741–1752

[8] Yannakoudakis H, Briscoe T, Alexopoulou D (2012) Automating second language acquisition research: integrating information visualisation and machine learning. pp 35–43

[9] Farra N, Somasundaran S, Burstein J (2015) Scoring Persuasive Essays Using Opinions and their Targets. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Denver, Colorado, pp 64–74

[10] McNamara DS, Crossley SA, Roscoe RD, Allen LK, Dai J (2015) A hierarchical classification approach to automated essay scoring. Assessing Writing 23:35–59.

[11] Cummins R, Zhang M, Briscoe T (2016) Constrained multi-task learning for automated essay scoring. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers 2:789–799.

[12] Amorim E, Cançado M, Veloso A (2018) Automated essay scoring in the presence of biased ratings. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1:229–237.

[13] Yannakoudakis H, Briscoe T, Medlock B (2011) A New Dataset and Method for Automatically Grading ESOL Texts. pp 180–189

[14] Taghipour K, Ng HT (2016) A neural approach to automated essay scoring. EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings 1882–1891.

[15] Alikaniotis D, Yannakoudakis H, Rei M (2016) Automatic text scoring using neural networks. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers 2:715–725.

[16] Farag Y, Yannakoudakis H, Briscoe T (2018) Neural automated essay scoring and coherence modeling for adversarially crafted input. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1:263–271.

[17] Jin C, He B, Hui K, Sun L (2018) TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 1:1088–1097.

[18] Dong F, Zhang Y, Yang J (2017) Attention-based recurrent convolutional neural network for automatic essay scoring. CoNLL 2017 - 21st Conference on Com-putational Natural Language Learning, Proceedings 153–162.

[19] Tay Y, Phan M, Tuan L, Hui S (2017) SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring

[20] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Lukasz, Polosukhin I (2017) Attention is all you need. Advances in Neural Information Processing Systems 2017-December:5999–6009

[21] Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs]

[22] Mayfield E, Black AW (2020) Should You Fine-Tune BERT for Automated Essay Scoring? 151–162.

[23] Rodriguez PU, Jafari A, Ormerod CM (2019) Language models and Automated Essay Scoring. arXiv

[24] Yang R, Cao J, Wen Z, Wu Y, He X (2020) Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. 1560–1569.

[25] Ormerod CM, Malhotra A, Jafari A (2021) Automated essay scoring using efficient transformer-based language models. 1–11

[26] Liu J, Xu Y, Zhu Y (2019) Automated Essay Scoring based on Two-Stage Learning. 1–7

[27] Nadeem F, Nguyen H, Liu Y, Ostendorf M (2019) Automated Essay Scoring with Discourse-Aware Neural Models. 484–493.

[28] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (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. Association for Computational Linguistics, San Diego, California, pp 1480–1489

[29] Nadeem F, Ostendorf M (2018) Estimating Linguistic Complexity for Science Texts. In: Proceedings of the Thirteenth Workshop on

Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, New Orleans, Louisiana, pp 45–55

[30] Uto M, Okano M (2020) Robust neural automated essay scoring using item response theory. Springer International Publishing

[31] Lord FM (1980) Applications of Item Response Theory To Practical Testing Problems, 1st edition. Lawrence Erlbaum Associates, Hillsdale, N.J

[32] Uto M, Xie Y, Ueno M (2021) Neural Automated Essay Scoring Incorporating Handcrafted Features. 6077–6088.

[33] Higgins D, Burstein J, Marcu D, Gentile C (2004) Evaluating Multiple Aspects of Coherence in Student Essays. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. Association for Computational Linguistics, Boston, Massachusetts, USA, pp 185–192

[34] Somasundaran S, Burstein J, Chodorow M (2014) Lexical Chaining for Measuring Discourse Coherence Quality in Test-taker Essays. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp 950–961

[35] Louis A, Higgins D (2010) Off-topic essay detection using short prompt texts. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Los Angeles, California, pp 92–95

[36] Persing I, Ng V (2014) Modeling Prompt Adherence in Student Essays. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, pp 1534–1543

[37] Burstein J, Chodorow M, Leacock C (2004) Automated Essay Evaluation: The Criterion Online Writing Service. AIMag 25:27–27.

[38] Persing I, Davis A, Ng V (2010) Modeling Organization in Student Essays. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Cambridge, MA, pp 229–239

[39] Persing I, Ng V (2013) Modeling Thesis Clarity in Student Essays. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Sofia, Bulgaria, pp 260–269

[40] Persing I, Ng V (2015) Modeling Argument Strength in Student Essays. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Beijing, China, pp 543–552

[41] Ke Z, Carlile W, Gurrapadi N, Ng V (2018) Learning to give feedback: Modeling attributes affecting Argument persuasiveness in student essays. IJCAI International Joint Conference on Artificial Intelligence, July 2018:4130–4136.

[42] Blanchard D, Tetreault J, Higgins D, Cahill A, Chodorow M (2013) TOEFL11: A CORPUS OF NON-NATIVE ENGLISH. ETS Research Report Series 2013:i–15.

[43] Granger S, Dagneaux E, Meunier F, Paquot M (2009) International Corpus of Learner English. Version 2. Handbook and CD-ROM

[44] Stab C, Gurevych I (2014) Annotating Argument Components and Relations in Persuasive Essays. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp 1501–1510

[45] Yannakoudakis H, Cummins R (2015) Evaluating the performance of Automated Text Scoring systems. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Denver, Colorado, pp 213–223