AES-Multi Scoring

A Project Report in partial fulfillment of the degree

Bachelor of Technology

in

Computer Science & Engineering

By

19K41A05G3	**MOHD AMAAN**
19K41A05H5	**V.ASHISH**

Under the Guidance of
D. Ramesh

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
S R ENGINEERING COLLEGE(A), ANANTHASAGAR, WARANGAL
(Affiliated to JNTUH, Accredited by NBA)

November 2022

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled "**AES-Multi Scoring**" is a record of bonafide work carried out by the students Mohd Amaan, Vanguri Ashish, bearing Roll No(s) 19K41A05G3, 19K41A05H5 during the academic year 2022-2023 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science Engineering** by the Jawaharlal Nehru Technological University, Hyderabad.

Supervisor **Head of the Department**

External Examiner

ABSTRACT

In recent years, pre-trained models have become dominant in most natural language processing (NLP) tasks. However, in the area of Automated Essay Scoring (AES), pre-trained models such as BERT have not been properly used to outperform other deep learning models such as LSTM. In this paper, we introduce a novel multi-scale essay representation for BERT that can be jointly learned. We also employ multiple losses and transfer learning from out-of-domain essays to further improve the performance. Experiment results show that our approach derives much benefit from joint learning of multi-scale essay representation and obtains almost the state-of-the- art result among all deep learning models in the ASAP1 task. Our multi-scale essay representation also generalizes well to Common- Lit Readability Prize (CRP2) data set, which suggests that the novel text representation proposed in this paper may be a new and effective choice for long-text tasks

Table of Contents

1.INTRODUCTION

AES is a valuable task, which can promote the development of automated assessment and help teachers reduce the heavy burden of assessment. With the rise of online education in recent years, more and more researchers begin to pay attention to this field. AES systems typically consist of two modules, which are essay representation and essay scoring module. The essay representation module extracts features to represent an essay and the essay scoring module rates the essay with the extracted features. When a teacher rates an essay, the scores are often affected by multiple signals from different granularity levels, such as token level, sentence level, paragraph level and etc. For example, the features may include the numbers of words, the essay structure, the master degree of vocabulary and syntactic complexity, etc. These features come from different scales of the essay. This inspires us to extract multi-scale features from the essays which represent multi-level characteristics of the essays. Most of the deep neural networks AES systems use LSTM or CNN. Some researchers attempt to use BERT in their AES systems but fail to outperform other deep neural networks methods. We believe previous approaches using BERT for AES suffer from at least three limitations. First, the pre-trained models are usually trained on sentence level, but fail to learn enough knowledge of essays. Second, the AES training data is usually quite limited for direct fine tuning of the pre-trained models in order to learn better representation of essays. Last but not least, mean squared error is commonly used in the AES task as the loss function. However, the distribution of the sample population and the sorting properties between samples are also important issues to be considered when designing the loss functions as they imitate the psychological process of teachers rating essays. Different optimizations can also bring diversity to the final overall score distribution and contribute to the effectiveness of ensemble learning. Due to COVID 19 outbreak, an online educational system has become inevitable. In the present scenario, almost all the educational institutions ranging from schools to colleges adapt the online education system. The assessment plays a significant role in measuring the learning ability of the student. Most automated evaluation is available for multiple choice questions, but assessing short and essay answers remain a challenge. The education system is changing its shift to online-mode, like conducting computer-based exams and automatic evaluation. It is a crucial application related to the education domain, which uses natural language processing (NLP) and Machine Learning techniques. The evaluation of essays is impossible with simple programming languages and simple techniques like pattern matching and language processing. Here the problem is for a single question, we will get more responses from students with a different explanation. So, we need to evaluate all the answers concerning the question. This Multi-Scoring will evaluate the response as scoring, word choice and organization. To address the aforementioned issues and limitations, we introduce joint learning of multi-scale essay representation into the AES task with BERT, which outperforms the state of the art deep learning models based on LSTM (Dong et al., 2017; Tay et al., 2018). We propose to explicitly model more effective representations by extracting multi- scale features as well as leveraging the knowledge learned from numerous sentence data. As the training data is limited, we also employ transfer learning from out-of-domain essays which is inspired by. To introduce the diversity of essay scoring distribution, we combine two other loss functions with MSE. When training our

model with multiple losses and transfer learning using R- Drop (Liang et al., 2021), we almost achieve the state-of-the-art result among all deep learning models. The source code of prediction module with a trained model for ASAP's prompt 8 is publicly available.

2.RELATED WORK

The dominant approaches in AES can be grouped into three categories: traditional AES, deep neural networks AES and pre-training AES. Traditional AES usually uses regression or ranking systems with complicated hand- crafted features to rate an essay. These handcrafted features are based on the prior knowledge of linguists. Therefore, they can achieve good performance even with small amounts of data. Deep-Learning Neural Networks AES has made great progress and achieved comparable results with traditional AES recently.

While the handcrafted features are complicated to implement and careful manual design makes these features less portable, deep neural networks such as LSTM or CNN can automatically discover and learn com- plex features of essays, which makes AES an end-to-end task. Saving much time to design features, deep neural networks can transfer well among different AES tasks. By combining traditional and deep neural network approaches, AES can even obtain a better result, which benefits from both representations. Pretraining AES uses the pre-trained language model as the initial essay representation module and fine tune the model on the essay training set. Though the pre-trained methods have achieved the state of the art performance in most NLP tasks, most of them fail to show an advantage over other deep learning methods in AES task. As far as we know, the work from Cao et al. (2020) and Yang et al. (2020) are the only two pre-training approaches which surpass the other deep learning methods. Their improvement mainly comes from the training optimization. Cao et al. (2020) employ two self-supervised tasks and domain adversarial training, while Yang et al. (2020) combine regression and ranking to train their model.

3.DESIGN

3.1 REQUIREMENT SPECIFICATION (S/W & H/W)

Hardware Requirements

- ✓ **System** : Pentium 4, Intel Core i3, i5, i7 and 2GHz Minimum
- ✓ **RAM** **:**4GB or above
- ✓ **Hard Disk** **:**10GB or above
- ✓ **Input** **:**Keyboard and Mouse
- ✓ **Output** **:**Monitor or PC

Software Requirements

- ✓ **OS** **:**Windows 8 or Higher Versions
- ✓ **Platform** **:** Jupiter Notebook
- ✓ **Program Language** **:** Python

4.APPROACH

The AES task is defined as following:

Given an essay with n words X = xi n, we need to output one score y as a result of measuring the level of this essay.

Quadratic weighted Kappa (QWK) (Cohen, 1968) metric is commonly used to evaluate AES systems by researchers, which measures the agreement between the scoring results of two raters.

4.1Multi-scale Essay Representation:

We obtain the multi-scale essay representation from three scales: token-scale, segment-scale and document-scale.

Token-scale and Document-scale Input We apply one pre-trained BERT (Devlin et al., 2019) model for token-scale and document-scale essay representations. The BERT tokenizer is used to split the essay into a token sequence T1 = [t1, t2, ...tn], where ti is the token and n is the number of the tokens in the essay. The token we mentioned in this paper all refer to Word Piece, which is obtained by the sub-word tokenization algorithm used for BERT. We construct a new sequence T2 from T1 as following. L is set to 510, which is the max sequence length supported by BERT. We construct a new sequence T2 from T1 as following. L is set to 510, which is the max sequence length supported by BERT vector [hi,1, hi,2, ..., hi,d] representing the ith sequence output, and hi,j is the jth element in hi. Segment scale Assuming the segment-scale value set is K = [k1, k2, ...ki, ..., kS], where S is the number of segment scales we want to ex- plore, and ki is the ith segment-scale in K. Given a token sequence T1 = [t1, t2, tn] for an essay, we obtain the segment scale essay representation corresponding to scale ki as follows:

1.We define np as the maximum number of to- kens corresponding to each essay prompt p. We truncate the token sequence to np tokens if the essay length is longer than np, otherwise we pad [PAD] to the sequence to reach the length np.

2.Divide the token sequence into m = [n /k |

T2 = [CLS]+[t1, t2, .., tL]+[SEP] n > L

[CLS]+T1+[SEP] n = L

[CLS]+T1+[PAD]∗(L − n)+[SEP] n < L

segments and each segment is of length ki except for the last segment, which is similar to the work of (Mulyar et al., 2019).

The final input representation are the sum of the token embeddings, the segmentation embeddings and the position embeddings. A detailed description can be found in the work of BERT (Devlin et al., 2019).

Document scale. The document scale representation is obtained by the [CLS] output of the BERT model. As the [CLS] output aggregates the whole sequence representation, it attempts to extract the essay information from the most global granularity. Token-scale As the BERT model is pre-trained by Masked Language Modeling (Devlin et al., 2019), the sequence outputs can capture the con- text information to represent each token. An essay often consists of hundreds of tokens, thus RNN is not the proper choice to combine all the token information due to the gradients vanishing problem. Instead, we utilize a max-pooling operation to all

the sequence outputs and obtain the combined token-scale essay representation. Specifically, the max-pooling layer generates a dimensional vector W = [w1, w2, ..., wj, ..., wd] and the element
wj is computed as below:
wj = max{h1,j, h2,j, ..., hn,j}
where d is the hidden size of the BERT model. As we use the pre-trained BERT model bert base uncased4, the hidden size d is 768. All the n sequence outputs of the BERT model are annotated as [h , h , ..., h , ..., h], where h is a dimensional

3.Input each of the m segment tokens into the BERT model, and get m segment representation vectors from the [CLS] output.
4.Use an LSTM model to process the sequence of m segment representations, followed by attention pooling operation on the hidden states of the LSTM output to obtain the segment- scale essay representation corresponding to scale ki.
The LSTM cell units process the sequence of segment representations and generate the hidden states as follows:
it = σ(Qi · st + Ui · ht−1 + bi) ft = σ(Qf · st + Uf · ht−1 + bf)
cˆt = tanh(Qc · st + Uc · ht−1 + bc) ct = it ◦ cˆt + ft ◦ ct−1
ot = σ(Qo · st + Uo · ht−1 + bo) ht = ot ◦ tanh(ct)
where st is the tth segment representation from BERT [CLS] output and ht is the tth hidden state generated from LSTM. Qi, Qf , Qc, Qo, Ui, Uf , Uc and Uo are weight matrices, and bi, bf , bc, and bo are bias vectors.
The attention pooling operation we use is similar to the work of (Dong et al., 2017), which is defined as follows:
αˆt = tanh(Qa · ht + ba)

1 2 i n i
eqa·αˆt αt =4j
eqa·αˆj
o = Σt αt
· ht

5.MODEL ARCHITECTURE

We apply one BERT model to obtain the document scale and token-scale essay representation. The concatenation of them is input into a dense regression layer which predicts the score corresponding to the document-scale and token-scale. For each segment-scale k with number of segments m, we apply another BERT model to get m CLS outputs, and apply an LSTM model followed by an attention layer to get the segment-scale representation. We input the segment-scale representation into another dense regression layer to get the score corresponding to segment-scale k. The final score is obtained by adding the scores of all S segment- scales and the score of the document-scale and token-scale.

6. DATA EVALUATION

ASAP data set is widely used in the AES task, which contains eight different prompts. A detaileddescription can be seen. For each prompt,the Word Piece length indicates the smallest number which is bigger than the length of 90% of the essays in terms of Word Piece number. We evaluate the scoring performance using QWK on ASAP data set, which is the official metric in the ASAP competition. Following previous work, we adopt 5-fold cross validation with 60/20/20 split for train, develop and test sets.

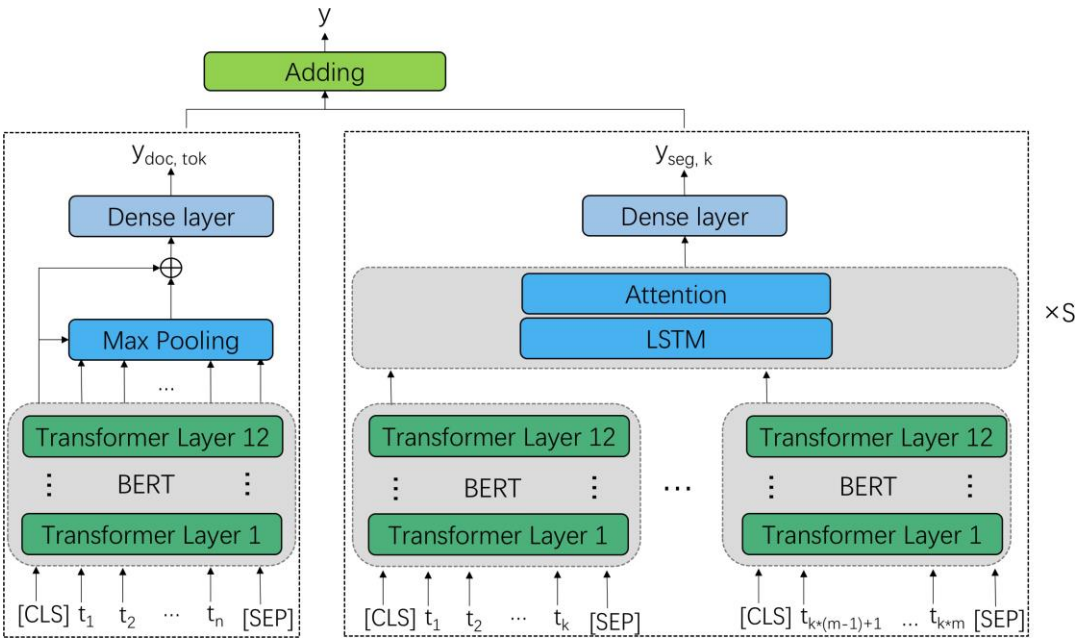

Figure1:Multi-Scale Essay Representation

7.DATASETS AND ATTRIBUTES

For this project the dataset is customized dataset. It consists of the response which need to be evaluated for reviewer1,2, Word choice and Organization. The dataset has 512 responses with in score of 0 to 5.

	iD	Response	Reviewer-	Reviewer-	word choi	Organization
2	1	An operat	4	4	3	1
3	1	An operat	5	5	2	3
4	1	Collection	2	1	1	1
5	1	It is an int	2	1	1	0
6	1	An operat	3	2	2	1
7	1	It is a platf	1	1	1	1
8	1	An operat	5	5	3	3
9	1	software v	3	2	2	1
10	1	Operating	4	4	2	1
11	1	An operat	4	4	2	2
12	1	Operating	2	2	2	1
13	1	An operat	2	2	1	1
14	1	It is the int	2	2	1	1
15	1	An operat	3	3	1	1
16	1	An	5	3	3	0
17	1	An operat	5	5	3	2
18	1	An operat	5	5	3	2
19	1	An operat	4	4	3	3
20	1	An operat	3	3	3	3
21	1	Operating	4	4	3	2

8.RESULTS

This AES-Multi-Scoring will evaluate the response as scoring, word choice and organization.

9.CONCLUSION AND FUTURE SCOPE

we propose a novel multi-scale essay representation approach based on pre-trained language model, and employ multiple losses and transfer learning for AES task. We almost obtain the state of the art result among deep learning models. In addition, we show multi-scale representation has a significant advantage when dealing with long texts. One of the future directions could be exploring soft multi-scale representation. Introducing linguistic knowledge to segment at a more reasonable scale may bring further improvement.

10.REFERENCES

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neu- ral networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Lin- guistics, pages 715–725.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. The Journal of Technol- ogy, Learning, and Assessment, 4(3).

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In arXiv: Computation and Language.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In SIGIR ’20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information, pages 1011–1020.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In Proceedings of the 2013 Conference on Empiri- cal Methods in Natural Language Processing, pages 1741–1752.

J Cohen. 1968. Weighted kappa: nominal scale agree- ment provision for scaled disagreement or partial credit. Psychological bulletin, 70(4):213–220.

Ma˘da˘lina Cozma, Andrei M. Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Compu- tational Linguistics.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural net- work for automatic essay scoring. In Proceedings of the 5th Workshop on Natural Language Process- ing Techniques for Educational Applications, pages 93–102.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understand- ing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Com- putational Linguistics: Human Language Technolo- gies, pages 4171–4186.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In Proceed- ings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1072–1077.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention- based recurrent convolutional neural network for au- tomatic essay scoring. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 153–162.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coher- ence modeling for adversarially crafted input. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Compu- tational Linguistics: Human Language Technolog, pages 263–271.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: A two-stage deep neural network for prompt- independent automated essay scoring. In Proceed- ings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 1088–1097.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In 3rd International Conference for Learning Representations.

Leah S. Larkey. 1998. Automatic essay grading using text categorization techniques. In SIGIR '98 Proceedings of the 21st annual international ACM SI- GIR conference on Research and development in information retrieva, pages 90–95.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie- Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In Advances in Neural Information Processing Systems, pages 10890–10905.

Sandeep Mathias, Rudra Murthy, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2020. Happy are those who grade without seeing: A multi- task learning approach to grade essays using gaze behaviour. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 858–872.

Elijah Mayfield and Alan W Black. 2020. 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.

Panitan Muangkammuen and Fumiyo Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. In Proceedings of the 1st Con- ference of the Asia-Pacific Chapter of the Associa- tion for Computational Linguistics and the 10th In- ternational Joint Conference on Natural Language Processing: Student Research Workshop, pages 116– 123.

Andriy Mulyar, Elliot Schumacher, Masoud Rouhizadeh, and Mark Dredze. 2019. Pheno- typing of clinical notes with improved document classification models using contextualized neural

language models. In 33rd Conference on Neural Information Processing Systems (NeurIPS).

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 431–439.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 13745–13753.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and automated essay scoring. In arXiv: Computation and Lan- guage.

Lawrence M. Rudner and Tahung Liang. 2002. Au- tomated essay scoring using bayes' theorem. The

Journal of Technology, Learning, and Assessment, 1(2):3–21.

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. 2020. Multi-stage pre- training for automated chinese essay scoring. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6723–6733.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In Proceed- ings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1882–1891.

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow:incorporating neural coherence features for end-to-end automatic text scoring. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pages 5948–5955.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating hand- crafted features. In Proceedings of the 28th Inter- national Conference on Computational Linguistics, pages 6077–6088.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xu- anjing Huang. 2018. Automatic essay scoring incor- porating rating schema via reinforcement learning. In Proceedings of the 2018 Conference on Empiri- cal Methods in Natural Language Processing, pages 791–797.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. En- hancing 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.