Automated Chinese Essay Scoring from Multiple Traits

Yaqiong He, Feng Jiang, Xiaomin Chu and **Peifeng Li***
School of Computer Science and Technology, Soochow University, Suzhou, China
{20204227070,fjiang}@stu.suda.edu.cn
{xmchu,pfli}@suda.edu.cn

Abstract

Automatic Essay Scoring (AES) is the task of using the computer to evaluate the quality of essays automatically. Current research on AES focuses on scoring the overall quality or single trait of prompt-specific essays. However, users expect to obtain not only the overall score but also the instant feedback from different traits to help improve their writing. Therefore, we first annotate a multi-trait dataset ACEA including 1220 argumentative essays from four traits, i.e., the essay organization, topic, logic, and language. And then we design a Hierarchical Multi-task Trait Scorer (HMTS) to evaluate the quality of writing by modeling these four traits. Moreover, we propose an inter-sequence attention mechanism to enhance information interaction between different tasks and design the trait-specific features for various tasks in AES. The experimental results on ACEA show that our HMTS can effectively score essays from multiple traits, outperforming several strong baselines.

1 Introduction

Automatic Essay Scoring (AES) is a task of using the computer to evaluate the quality of students' essays. An efficient AES system can bring many benefits to the field of education, including avoiding the influence of subjective factors on essay scoring, greatly reducing teachers' workload, and providing instant feedback for students' written essays.

Early work treated AES as a classification (Larkey, 1998; Rudner and Liang, 2002), regression (Attali and Burstein, 2005; Phandi et al., 2015), or ranking problem (Yannakoudakis et al., 2011; Chen and He, 2013), addressing AES by supervised learning. Recently, with the progress of deep learning, the end-to-end models based on neural networks (e.g., LSTM-based models or the combination of LSTMs and CNNs) have achieved impressive results (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Tay et al., 2018).

In consideration for more feedback in AES systems, some studies began to focus on scoring specific traits of essays, such as word choice, content (Page, 1966), organization (Persing et al., 2010; Song et al., 2020a), thesis clarity (Persing and Ng, 2013), prompt adherence (Persing and Ng, 2014), argument persistent (Ke et al., 2018), style (Page, 1966), etc. Most of the above studies focused on scoring prompt-specific essays. As previous researchers had pointed out (Jin et al., 2018; Ridley et al., 2020), it was difficult and expensive to access ample target-prompt essays in the real-world AES system. Therefore, some researchers on English AES had began to explore cross-prompt AES (Ridley et al., 2020; Phandi et al., 2015; Jin et al., 2018). Their methods utilized non-target-prompt essays to train models and score target-prompt essays.

In reality, people usually judge the essay from different perspectives to give the final overall holistic score. Apart from (Ridley et al., 2021; Kumar et al., 2021), few studies had combined the overall score with the feature-specific scores. Therefore, to make AES more interpretable and verify whether the information learned in the traits contributes to the learning of the overall score, we use multi-task learning to evaluate the overall holistic score from multiple traits and the score of each trait.

This paper focuses on scoring Chinese essays and their traits. It is worth noting that most of the previous studies were conducted on the English-language dataset ASAP[1], which contains eight different collections of essays with specific prompts. Unlike that, our Chinese dataset is thematically diverse and does not contain prompts, so it is more challenging than the English task.

There are some issues between English and Chinese for the overall score and the traits scores. First, they have different expression problems. The

*Corresponding author

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

Proceedings of the 29th International Conference on Computational Linguistics, pages 3007–3016
October 12–17, 2022.

most significant differences are that Chinese is a parataxis language while English is hypotaxis. In the task of English AES, word usage is the most important factor in English expression (Ridley et al., 2021). English text mainly forms semantic representation by combining the similarity between words. While Chinese essays pay more attention to sentence-level and paragraph-level expression. Second, Chinese and English essays have different evaluation emphases that we should evaluate the essay quality from different perspectives. For example, for text expression, the evaluation criteria for English essays are using language correctly and high-level vocabulary, while the evaluation criteria for Chinese essays are using beautiful sentences and smooth expressions.

Therefore, our task of scoring Chinese essays and their traits exhibits two challenges. First, we need to select appropriate traits to express Chinese essays and design the proper model to combine the different traits to prompt the performance of overall scoring. Second, a good model is needed to fully express the deep semantics of Chinese essays, which should have a good generalization ability to perform well in different trait tasks.

In this paper, we design a Hierarchical Multi-task Trait Scorer (HMTS) for the task of automated Chinese essay scoring. To cope with the first challenge, we created a dataset named "Automatic Chinese Essay Assessment (ACEA)" extended from the dataset used in Song et al. (2020b), which contains 1220 essays, the overall score of each essay, and the scores of the four traits (i.e., topic, organization, logic, and language). And then we utilize the multi-task learning framework to learn the trait levels. Specially, to better combine different trait expressions, we propose an inter-sequence attention mechanism, which can integrate the representations of different trait tasks. We believe that different trait tasks can complement each other. For example, an essay that scores well on logic will also score well on language performance. Besides, since different tasks have different evaluation criteria, we also obtain trait-specific features to enhance the representation of the specific trait tasks.

To cope with the second challenge, we utilize the hierarchical model to obtain the shared sentences and paragraph representation, and essay representation independent for each task. We use the XL-Net as the sentence encoder, which is a pre-trained model trained on large-scale datasets and can get better text semantic representation. In a word, the contributions of this paper are as follows:

- We are the first to introduce the multi-trait scores to Chinese AES, and grade our ACEA dataset with four traits, i.e., organization, topic, logic and language.
- We design a hierarchical multi-task trait scorer HMTS, which can superior express the semantics of Chinese essays and give the overall score as well as the trait scores for essays without prompt.
- We use the inter-sequence attention mechanism to fuse the representation of different traits to share the contribution of the other trait tasks, and use the trait-specific features to enhance the representation of each trait task.

2 Related Work

In this section, we first introduce the prompt-specific and cross-prompt holistic scoring methods and then describe the prompt-specific and multi-task trait scoring methods. Finally, we briefly introduced the task of Chinese AES.

2.1 Holistic Scoring

Prompt-Specific Holistic Scoring For the overall essay score, early studies mainly leveraged regression, classification or ranking algorithms combined with manual features (Page, 1966; Attali and Burstein, 2005; Larkey, 1998; Rudner and Liang, 2002; Yannakoudakis et al., 2011; Chen and He, 2013; Miltsakaki and Kukich, 2004; Burstein et al., 1998). In recent years, many studies (Taghipour and Ng, 2016; Cummins et al., 2016; Alikaniotis et al., 2016; Dong and Zhang, 2016; Dong et al., 2017; Tay et al., 2018) had shown that neural network methods can capture the deep semantics of essays more effectively. For example, Dong et al. (2017) used the attention mechanism to obtain the contribution of words and sentences to the essay. Tay et al. (2018) measured similarity between adjacent sentences to model coherence to evaluate text quality.

Cross-Prompt Holistic Scoring In fact, it's unrealistic to get ample essays with a specific prompt. A good AES should be able to provide feedback to essays on any prompt. Therefore, some studies on English focused on cross-prompt scoring. Song et al. (2020b), Phandi et al. (2015) and Cummins

et al. (2016) trained the model with large quantities of non-target-prompt and a small quantity of target-prompt essays and performed transfer learning to score target-prompt essays. However, these methods still need essays with target prompts. Some recent work explored cross-prompt scoring without target prompts. Jin et al. (2018) applied a two-stage approach, in which the first stage utilized the prompt-independent features to award pseudo labels to target prompt essays. Then in the second stage, the pseudo-labeled essays were used as training data to award the final score. Ridley et al. (2020) applied a single-stage neural network-based method that utilizes a set of general features to award scores to target-prompt essays.

2.2 Trait Scoring

Prompt-Specific Trait Scoring In addition to the overall score, students also expect to get feedback from different aspects through AES, which can assist students' growth.

In the last decade or so, some studies focused on scoring the single essay trait. For example, Song et al. (2020a) proposed a hierarchical multi-task learning model to evaluate the organization quality by learning the sentence and paragraph function. Ke et al. (2018) modeled argument persistence and the attributes of those arguments in student essays. Some more recent studies had adapted a leading prompt-specific holistic scoring method to output the score for different traits. Mathias and Bhattacharyya (2020) adapted some leading approaches for prompt-specific persistent scoring to the task of trait scoring. Hussein et al. (2020) also adapted a leading prompt-specific holistic scoring method, employing a multi-task architecture to output the overall score and scores for various traits simultaneously.

Multi-task Trait Scoring To obtain more feedback, some studies used multi-task learning to score essay traits and the essay itself. Kumar et al. (2021) proposed a prompt-specific method and introduced the multi-task learning (MTL) method to essay scoring, where scoring the overall essay score is the primary task, and scoring the essay traits is the auxiliary task. Ridley et al. (2021) provided a new approach named Automated Cross-prompt Scoring of Essay Traits, which can give the holistic score as well as the scores for different traits in the cross-prompt setting. The drawback of these studies is that their methods are not suitable for Chinese AES and also are prompt-specific. Due to the different evaluation criteria of Chinese and English essays, traits selection should be different.

2.3 Chinese AES

Research on Chinese AES is later than that of English AES and only a few works focused on this task. Among them, Yang and Cao (2012) applied LDA (Latent Semantic Analysis) to score Chinese essays. Fu et al. (2018) improved the accuracy of Chinese AES by identifying beautiful sentences and taking them as literary features. Song et al. (2020a) evaluated the organizational score of high school argumentative essays, and Song et al. (2020b) explored cross-prompt holistic scoring on four different essay sets, with articles in each dataset responding to a different prompt. However, their dataset is not publicly available.

3 ACEA Dataset

Following the college entrance examination essay scoring standard, we selected four traits of topic, organization, logic, and language for our multi-task learning, which can help evaluate the level of essay quality.

3.1 Traits Selection

For the essay topic, whether the essay conforms to the topic's meaning and whether the center is prominent is one of the criteria to evaluate the essay's grade based on the essay evaluation criteria of the college entrance examination.

For the essay organization, Song et al. (2020a) showed that organization is a critical aspect of Chinese writing. A well-organized article should have a clear structure to accurately and logically develop ideas. Chen (2016) had set up a set of solutions to evaluate the quality of the Chinese essays, taking the organization as one of the evaluation criteria.

For the essay logic, Yan (2019) showed that the intention of the essay should be profound and reflect the correct and positive thought and value orientation, and be consistent with the mainstream of social development and the spirit of the times. Therefore, we take logic as one of the traits, which takes whether the logic is clear and the content is profound as the evaluation standard.

For the essay language, Liu (2015) and Liu et al. (2016) used the rhetorical recognition of parallelism metaphor to express the literary grace of the article. Gong (2016) conducted in-depth research

Trait	Evaluation Criteria
Topic Grades	**Bad**: The topic is not clear; the material can't describe the topic; the center is unclear. **Medium**: The essay has a clear topic and center; the material can basically describe the topic. **Great**: The content revolves around the topic; the material is rich enough to express the topic; the center is prominent.
Organization Grades (Song et al., 2020a)	**Bad**: The essay is poorly structured. It is incomplete or misses key discourse elements. **Medium**: The essay is well structured and complete, but could be further improved. **Great**: The essay is fairly well structured and the organization is very clear and logical.
Language Grades	**Bad**: It is plain language or the sentence pattern is single. **Medium**: Some languages are beautiful; the overall language is fluent; there are changes in sentence patterns. **Great**: Beautiful and fluent language or flexible sentence pattern.
Logic Grades	**Bad**: Unclear logic; The discussion is not profound. **Medium**: The logic is clear and the discussion is not profound or logic and discussion are not too bad. **Great**: Clear logic and profound discussion.

Table 1: Trait grades and their evaluation criteria.

on rhetorical devices in Chinese essays and realized an AES system based on the figure of speech recognition, which showed that the introduction of rhetorical devices has produced good results for the system scoring. Therefore, we take the language as one of the traits. Language evaluation criteria include the use of beautiful sentences and the fluency of sentences.

3.2 Traits Grades

We evaluate the quality of topic, organization, language, and logic of the essay with three grades. The evaluation criteria are shown in Table 1.

3.3 Data Annotation

Our dataset contains 1220 argumentative essays published by Song et al. (2020a), which were written by senior high school students and selected from LeLeKetang[2]. These essays contain various topics and do not have specific prompts. The overall scores we used come from the LeLeKetang website and the organization scores have been given by Song et al. (2020a). We asked two annotators who are Chinese teachers in high school to assign other three trait grades for each essay.

For the other three trait scores, we counted the consistency rate of annotation. According to kappa calculation, the internal annotation consistency of topic, logic and language were 0.87, 0.9 and 0.88 respectively. A third annotator was introduced to discuss and analyze essays with disagreed annotations to get the final decision. Table 2 shows the statistics of the dataset.

[2]http://www.leleketang.com/zuowen/.

Trait	Bad	Medium	Great
Organization	245	670	305
Topic	209	549	462
Logic	302	596	322
Language	259	648	313
Overall	250	645	325

Table 2: Essay distribution on trait and grade.

4 HMTS Model

We employ a hierarchical model to learn the semantic representation of texts and reinforce the expression of each trait task through multi-task learning. For our hierarchical neural model, we utilize XLNet (Yang et al., 2019) to learn the sentence expressions and LSTM to learn the paragraph and essay expressions, which can simulate the coherence between the sentence and paragraph sequences. We also use the attention pooling to encode sentences, which aims to capture more relevant sentences that contribute to the paragraph semantic representation, thereby further expressing the quality of an article.

For our multi-task framework, unlike the English multi-task AES designed by Ridley et al. (2021) and Kumar et al. (2021), we choose different traits for Chinese trait scoring, and our test dataset contains essays with multiple prompts rather than a single prompt. To share information between different traits more clearly, we propose the inter-sequence attention mechanism to help each trait pay attention to the information contained in the other traits. We also introduce the trait-specific features to represent the trait quality.

Our HMTS model includes two stages and the

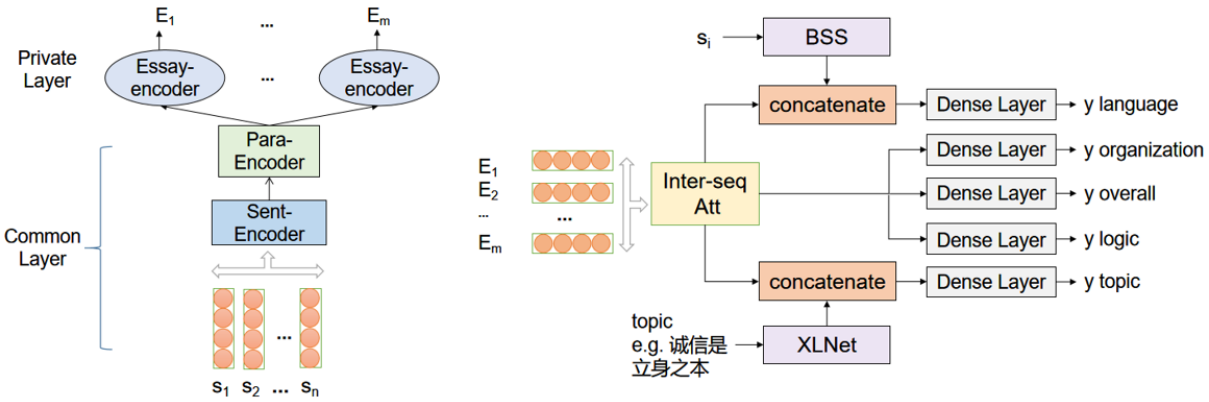

Figure 1: Architecture of our HMTS. The left part shows the sentence encoder and paragraph encoder in the common layer, and the essay encoder in the private layer. The right part contains the inter-sequence attention and trait-specific features in the private layer, in which BSS (Beautiful Sentence Scorer) is used to extract language-specific feature.

details are shown in Figure 1. HMTS includes two layers: 1) Common Layer: this layer aims to learn the useful low-level representations for all tasks, including the sentence and paragraph representation; 2) Privates Layer: this high-level layer is employed to learn more task-specific representations, including essay representation, the inter-sequence mechanism, and trait-specific features. High levels in multi-task architecture can represent more complex information (Sanh et al., 2019).

4.1 Common Layer

The parameters in the lower layer of our model HMTS are shared, so as to enable the sharing of information relevant to all trait tasks. Specifically, the lower layer includes the sentence encoder and the paragraph encoder.

Sentence Encoder A sequence of words $\boldsymbol{s}_i$= $\{w_1,w_2,...,w_m\}$ is modeled with a XLNet encoder, and the vector representation of the last word is used as the final sentence representation $\boldsymbol{s}_{e_i}$.

Paragraph Encoder Firstly, BiLSTM is used to encode the paragraph with the sentence sequence $\{\boldsymbol{s}_{e_i}, \boldsymbol{s}_{e_{i+1}}, ..., \boldsymbol{s}_{e_{i+n-1}}\}$ as follows, where $\boldsymbol{h}_i \in \mathbb{R}^{512}$ is the hidden representation at the time-step i.

$$\boldsymbol{h}_i = \text{BiLSTM}(\boldsymbol{s}_{e_i}, \boldsymbol{h}_{i-1}) \tag{1}$$

A common attention pooling layer is then applied to the hidden representations to learn the paragraph representation as follows.

$$\boldsymbol{e}_i = \tanh(\boldsymbol{W}_e \cdot \boldsymbol{h}_i + \boldsymbol{b}_e) \tag{2}$$

$$\boldsymbol{a}_i = \frac{exp(\boldsymbol{w}_a \cdot \boldsymbol{e}_i)}{\sum_{j=1}^{n} exp(\boldsymbol{w}_a \cdot \boldsymbol{e}_j)} \tag{3}$$

$$\boldsymbol{P} = \sum_{i=1}^{n} \boldsymbol{a}_i \cdot \boldsymbol{h}_i \tag{4}$$

where $\boldsymbol{W}_e$ and $\boldsymbol{W}_a$ are weight matrix and vector respectively, $\boldsymbol{b}_e$ is the bias vector, $\boldsymbol{e}_i$ and $\boldsymbol{a}_i$ are the attention vector and attention weight of the i-th sentence, n denotes the number of sentences in a paragraph. $\boldsymbol{P} \in \mathbb{R}^{512}$ is the final paragraph representation.

4.2 Private Layer

There are M (M=5) tasks in total (four trait tasks and one overall task), and each component in the private layer has M separate copies.

Essay Encoder We encode the paragraph sequence with BiLSTM as follows and the mean of hidden layers is used as essay representation.

$$\boldsymbol{H}_i^j = \text{BiLSTM}(\boldsymbol{P}_i^j, \boldsymbol{H}_{i-1}^j) \tag{5}$$

$$\boldsymbol{E}_j = \text{mean}(\boldsymbol{H}_1^j, \boldsymbol{H}_2^j, ..., \boldsymbol{H}_m^j) \tag{6}$$

where for the j-th task, $\boldsymbol{P}_i^j$ is the paragraph representation of the i-th paragraph, $\boldsymbol{H}_i^j$ is the hidden representation at the time-step i, and $\boldsymbol{E}_j$ is the essay representation.

Inter-sequence Attention Mechanism To obtain the contribution of the other tasks to the current trait task, we introduce the inter-sequence attention mechanism as follows.

$$\boldsymbol{ET} = \{\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3, \boldsymbol{E}_4, \boldsymbol{E}_5\} \tag{7}$$

$$\boldsymbol{E} = \tanh(\boldsymbol{ET}) \tag{8}$$

$$\boldsymbol{a} = \text{softmax}(\boldsymbol{v}^T \cdot \boldsymbol{E}) \tag{9}$$

$$\boldsymbol{R}_t = \tanh(\boldsymbol{ET} \cdot \boldsymbol{a}^T) \tag{10}$$

where $\boldsymbol{T} = \{\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3, \boldsymbol{E}_4, \boldsymbol{E}_5\}$ is the concatenation of the essay representation of each task. We initialize a training parameter $\boldsymbol{v} \in \mathbb{R}^{512}$ to obtain the final trait fusion $\boldsymbol{R}_t \in \mathbb{R}^{512}$ representation.

Task-specific Features To obtain a final representation for each task, we introduce the task-specific features to enhance the task representation.

Topic-task feature Since our dataset lacks prompts, we use the essay title as the topic task feature. Specifically, we encode the essay title with XLNet and concatenate the last word representation with the topic representation as the final topic-task essay representation as follows.

$$\boldsymbol{f}_{top} = \text{XLNet}(\boldsymbol{s}_t) \tag{11}$$

$$\boldsymbol{E}_{top} = \text{Dense}(\text{concatenate}(\boldsymbol{f}_{top}, \boldsymbol{R}_t)) \tag{12}$$

where $\boldsymbol{s}_t$ is the word sequence of the essay topic, $\boldsymbol{f}_{top}$ is the topic feature representation, and $\boldsymbol{E}_{top}$ is the final topic-task representation.

Language-task feature According to the language level evaluation rules in Subsection 3.2, beautiful and fluent language is an important influence for the language level. Therefore, we trained a model to predict whether sentences are beautiful or not. We selected 8840 underlined sentences from LeLeKetang as beautiful sentences, of which underlined sentences are beautiful sentences annotated on the website and 6714 bad sentences from the low-level essays. We split the dataset into the training dataset (80%) and testing dataset (20%) and used the XLNet classification model to encode the sentence and fed the last word representation into a dense layer for prediction, and the results are Yes and No.

We used the pre-trained beautiful sentence scorer (BSS) to predict sentences in each essay. All prediction results were concatenated with the language representation of the essay as the final language-task essay representation as follows.

$$\boldsymbol{s}'_{ij} = \text{BSS}(\boldsymbol{s}_{ij}) \tag{13}$$

$$\boldsymbol{f}_{lang} = \text{concatenate}(\boldsymbol{s}'_{i1}, \boldsymbol{s}'_{i2}, ..., \boldsymbol{s}'_{in}) \tag{14}$$

$$\boldsymbol{E}_{lang} = \text{concatenate}(\boldsymbol{f}_{lang}, \boldsymbol{R}_t) \tag{15}$$

where $\boldsymbol{s}_{ij}$ is the j-th sentence of the i-th essay, $\boldsymbol{s}'_{ij}$ is prediction result, $\boldsymbol{f}_{lang}$ is the language-task feature, and $\boldsymbol{E}_{lang}$ is the final language-task representation.

4.3 Prediction

Finally, a dense layer with a tanh activation function was applied to the representation of each task as follows.

$$\boldsymbol{y}_j = \tanh(\boldsymbol{W}_y^j \cdot \boldsymbol{E}_j + \boldsymbol{b}_y) \tag{16}$$

where $\boldsymbol{E}_j$ is the representation of the j-th trait task. For the organization-task, the logic task and the overall score, $\boldsymbol{E}_j = \boldsymbol{R}_t$; for the language task, $\boldsymbol{E}_j = \boldsymbol{E}_{lang}$; for the topic task, $\boldsymbol{E}_j = \boldsymbol{E}_{top}$. y_j is the predicted score for the j-th task, $tanh$ is the activation function, $\boldsymbol{W}_y^j$ is a weight vector, and $\boldsymbol{b}_y$ is a bias.

4.4 Loss

To train the parameters, we minimized the binary cross-entropy loss function. Given N essays and M tasks, the loss was calculated as follows.

$$L = -\frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{c=1}^{T} (y_{ijc} \cdot \log(P_{ijc})) \tag{17}$$

where T denotes the number of grades. If the golden grade of the sample i equals to c, y_{ijc} = 1; otherwise y_{ijc} = 0. P_{ijc} denotes the prediction probability that the sample i belongs to the grade c. where T denotes the number of grades. If the golden grade of the sample i in the task j equals to c, y_{ijc} = 1; otherwise y_{ijc} = 0. P_{ijc} denotes the prediction probability that the sample i in the task j belongs to the grade c.

5 Experimentation

In this section, we first introduce the experimental settings and then report the experimental results. Finally, we analyze the results on different aspects.

5.1 Experimental Settings

We conduct the experiments on our dataset ACEA described in Section 3. The average number of paragraphs and sentences per essay is 8 and 28, and the figure of words per sentence is 21. The maximum number of sentences and paragraphs in an essay (n and m) is set to 50 and 20, and the maximum number of sentences in a paragraph (n_p) is set to 20. The sentences and paragraphs shorter or longer than the limitations are padded or truncated.

We split our dataset into five folds as Song et al. (2020a). Cross-validation is conducted and the average performance is reported. Each fold has a similar distribution over essay grades. During

Parameter	Value
Embedding size	768
Dimension of BiLSTM hidden state	256
Batch size	32
Learning rate	0.0001
Optimizer	Adam

Table 3: Hyper-parameters and their values.

training, 10% of the training data was selected randomly as the validation set to find the optimal hyper-parameters. The detailed settings of hyper-parameters are listed in Table 3.

To evaluate the performance of our model, we use the Quadratic Weighted Kappa (QWK) metric, which has been widely adopted in both holistic essay scoring and essay trait scoring research (Ridley et al., 2020; Mathias and Bhattacharyya, 2020; Hussein et al., 2020) and is designed to measure the level of agreement between two raters.

5.2 Baselines

To verify the effectiveness of our HMTS, we conduct six strong baselines for comparison as follows.

CNN_LSTM_att (Dong et al., 2017): This model is a leading overall scoring model for specific prompt, which treated input essays as sentence-document hierarchies.

MTL (Kumar et al., 2021): This model uses CNN_LSTM_att to obtain the essay representation and concatenated it with the predicted trait scores to score the essay holistically. We use the multi-task architecture to train the model on each trait individually.

Song2020 (Song et al., 2020a): This model is proposed for organization evaluation of Chinese argumentative student essays, utilizing a hierarchical multi-task approach for joint discourse element identification and organization evaluation. We use a single-task architecture to train the model on each trait individually.

XLNet (Yang et al., 2019): This model uses the pre-trained model XLNet to obtain the paragraph representation, and then concatenate them and feed them into the dense layer to predict the essay grade.

5.3 Experimental Results

We reported the results of our experiments using single-task and multi-task models. Table 4 shows the performance comparison of the baselines and HMTS on the dataset ACEA.

We can find out that CNN_LSTM_att and MTL both perform poorly. There are two reasons for this result. First, CNN_LSTM_att and MTL are models designed for essays with specific prompts, while our dataset does not have specific prompts. Second, these models are designed for English AES and pay more attention to the contribution of words to sentences and essays, while Chinese AES pays more attention to sentence expression. In addition, we can find out that XLNet alone has achieved good performance, which shows the effectiveness of using the pre-trained model to obtain semantic information.

In comparison with the above single-task models, Song2020 achieves the best overall QWK. However, its QWK is still lower than our HMTS, since it only considers the organization element. Compared with the best baselines Song2020 on sing-task and XLNet on multi-task, our HMTS significantly improves overall QWK by 3.07 and 9.28, respectively. This result shows that our HMTS on four traits is effective for the overall essay scoring.

Table 4 also shows the performance comparison of the baselines and HMTS on each trait task on ACEA and we also can find that our HMTS achieves the best QWK on all four traits, i.e., organization, topic, logic and language. These results verify the fact that our HMTS also is effective on scoring traits. Besides, our model was significantly superior to all baselines with a p-value<0.05.

5.4 Analysis on Multi-task Learning

In Table 4, compared with the single-task HMTS, our multi-task HMTS significantly improves overall QWK by 5.15, and this result indicates the effectiveness of the multi-task framework on essay scoring. This is due to the fact that the single-task HMTS can not fully utilize the information contained in the article, while our multi-task HMTS can share the information between different traits, which contributes to the final performance of each trait task. Comparing the performance gap (1.59) between single-task and multi-task XLNet, this gap (5.15) between our HMTS is much larger. This also verifies the effectiveness of the multi-task framework of our HMTS.

In comparison with the single-task HMTS, our multi-task HMTS also achieves better QWK on three traits, i.e., organization (+2.03), topic (+6.52) and logic (+2.17), while it achieves comparable performance on the language trait (-1.16).

Task Type	Model	Overall	Organization	Topic	Logic	Language
Single-task	CNN_LSTM_att	40.13	39.0	37.50	40.23	52.58
	XLNet	52.13	52.47	48.84	52.55	56.68
	Song2020	54.78	55.76	46.23	52.84	50.28
	HMTS	**57.85**	**59.11**	**49.75**	**58.44**	**60.71**
Multi-task	CNN_LSTM_att	41.58	39.25	39.26	41.64	50.85
	MTL	43.59	41.50	38.82	42.70	52.51
	XLNet	53.72	52.76	49.52	53.30	56.75
	HMTS	**63.00**	**61.14**	**56.27**	**60.61**	**59.55**

Table 4: Performance comparison (QWK) of the baselines and HMTS on ACEA.

Model	QWK
HMTS	63.00
w/o Organization	-0.98
w/o Topic	-1.91
w/o Logic	-2.39
w/o Language	-0.91

Table 5: Ablation Experiments.

To gain some insight into the effectiveness of each trait, we conducted the ablation experiments as shown in Table 5, which remove one trait from HMTS one by one. We can find out the fact that the performance of the model HMTS will decline if we remove one trait task from the multi-task framework, which can reflect the contribution of each task to the overall score. We find out that the result decreases the most after removing the logic evaluation, which shows that logic is the most important trait that contributes to the overall score, compared with the organization, topic and language. The reason is that the essay scoring standard of college entrance examination is divided into the basic level and development level, and whether the logic of the essay is clear or whether the discussion is profound belongs to the development level, is the key factor to distinguish the excellent essay from other levels.

5.5 Analysis on Inter-sequence Attention

Table 6 shows the effectiveness of the trait-specific feature and the inter-sequence attention. In Table 6, HMTS (w/o inter-att) is the simplified version of HMTS without the inter-sequence attention mechanism. We can find out that HMTS outperforms HMTS (w/o inter-att) on QWK on all trait tasks. These results indicate that the inter-sequence attention can improve the overall score and traits score effectively, especially the topic trait with the highest improvement (+3.08).

5.6 Analysis on Task-specific Features

The use of well-designed trait-specific features is important for each trait task, because different traits have different evaluation criteria. In Table 6, HMTS (w/o trait-feat) is the simplified version of our HMTS without the task-specific features. Compared HMTS (w/o trait-feat) with HMTS, we can find out that the additional topic-task and language-task features improve the results of all trait tasks, especially the topic trait. This proved the effectiveness of the language-task and topic-task features.

We also evaluated the effectiveness of the beautiful sentence scorer (BSS), The F1 score is 0.91 and this indicates that our BSS can accurately identify beautiful sentences. These sentences are usually fluent in the language, flexible in sentence patterns, and good at using rhetoric and famous aphorisms.

5.7 Error Analysis

In analyzing the experimental results, we find that most essays with recognition errors were recognized as their adjacent grades. That is, the essays with the bad and great grades tend to be misclassified as the grade medium, and those with the medium grade tend to be misclassified as the grade great or bad. Especially, our model HMTS tends to wrongly assign a lower grade to an essay, which accounts for 62.9%, 71.0%, 65.8%, 55.5% and 54.1% on the overall score and the scores on the traits organization, logic, topic and language.

The accuracy of our HMTS identifying the essays with the great, medium, and bad grades are 58.0, 72.9 and 64.0, respectively. As we mentioned above, our HMTS has the underestimation trend and the accuracy on the bad essays is better than those on great essays. In addition, since the medium grade is the majority class, it achieves the best accuracy.

Model	Overall	Organization	Topic	Logic	Language
HMTS	63.00	61.14	56.27	60.61	59.55
w/o trait-fea	-0.56	-0.34	-1.41	-0.59	-0.29
w/o inter-att	-1.78	-1.06	-3.08	-0.59	-1.84
w/o inter-att&trait-fea	-2.07	-1.22	-3.44	-0.81	-1.96

Table 6: The effectiveness of the trait-specific feature and the inter-sequence attention.

6 Conclusion

We first annotate a multi-trait essay scoring dataset ACEA from the topic, organization, logic, and language aspects. And then we propose a multi-task learning framework HMTS for the Chinese AES task. Moreover, we propose an inter-sequence attention mechanism to enhance information interaction between the different trait tasks and design the trait-specific features for various trait tasks. The experimental results show that our HMTS can effectively score essays from multiple traits, outperforming several strong models. In the future, we will focus on incorporating the traits task into the overall task in a more effective framework.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 61836007, 62276177 and 62006167), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*.

Yigal Attali and Jill Burstein. 2005. Automated essay scoring with e-rater v. 2.0 (ets rr-04-45). *Educational Testing Service, Princeton, NJ*, pages 2–20.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. Automated scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 206–210.

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

Yile Chen. 2016. Research on key technology of chinese automated essay scoring based on regression analysis. Master's thesis, Harbin Institute of Technology.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 789–799, Berlin, Germany. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

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

Ruiji Fu, Dong Wang, Shijin Wang, Guoping Hu, and Ting Liu. 2018. Elegart sentence recognition for automated essay scoring. *Journal of Chinese Information Processing, Volume 32*, pages 88–97.

Jiefu Gong. 2016. The design and implementation of the rhetoric recognition system oriented chinese essay review. Master's thesis, Harbin Institute of Technology.

Mohamed Abdellatif Hussein, Hesham A., and Mohammad Nassef. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications, Volume 11*, pages 287–293.

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

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4130–4136.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Many hands make

light work: Using essay traits to automatically score essays. *arXiv preprint arXiv:2102.00781*.

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

Mingyang Liu. 2015. Research on the key technology of the automatic scoring of the college entrance examination essay. Master's thesis, Harbin Institute of Technology.

Mingyang Liu, Bing Qin, and Ting Liu. 2016. Automated chinese composition scoring based on the literary feature. *Intelligent Computer and Applications*.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering, Volume 10*, pages 25–55.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan, Volume 47*, pages 238–243.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 229–239.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

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, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence, Volume 35*, pages 13745–13753.

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 preprint arXiv:2008.01441*.

Lawrence M Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence, Volume 33*, pages 6949–6956.

Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu. 2020a. Hierarchical multi-task learning for organization evaluation of argumentative student essays. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3875–3881.

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. 2020b. 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 *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Yi Tay, Minh 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 AAAI conference on artificial intelligence, Volume 32*.

Dejvn Yan. 2019. Some strategies for writing a good composition for the college entrance examination,. *Middle School Chinese*, page 2.

Chen Yang and Yiwei Cao. 2012. Present situation and prospect of automatic composition scoring. *Language Teaching in Middle School*, pages 78–80.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 1–11. Curran Associates, Inc.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.